Mobilités urbaines et données en ligne pour l'étude des maladies vectorielles à Delhi (Inde) et Bangkok (Thaïlande)

Alexandre Cebeillac

▶ **To cite this version:**

Alexandre Cebeillac. Mobilités urbaines et données en ligne pour l'étude des maladies vectorielles à Delhi (Inde) et Bangkok (Thaïlande). Géographie. Normandie Université, 2018. Français. NNT : 2018NORMR137 . tel-02089908v2

HAL Id: tel-02089908

https://theses.hal.science/tel-02089908v2

Submitted on 4 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité Géographie

Préparée au sein de l'Université de Rouen

Mobilités urbaines et données en ligne

pour l'étude des maladies vectorielles à Delhi (Inde) et Bangkok (Thaïlande)

Présentée et soutenue par
Alexandre CEBEILLAC

Thèse soutenue publiquement le 17 octobre 2018 devant le jury composé de		
Mme Lena SANDERS	HDR - Directrice de Recherche - UMR Géographie-Cités 8504 – CNRS	Rapporteur
M. Thomas THÉVENIN	Professeur - UMR THEMA 6049 - Université de Bourgogne	Rapporteur
Mme Isabelle THOMAS	Professeure - Directrice de Recherche FNRS - CORE - Université Catholique de Louvain	Examinateur
M. Didier JOSSELIN	Professeur - Directeur de Recherche CNRS - UMR ESPACE 7300 - Université d'Avignon	Examinateur
Mme Sonia CHARDONNEL	Chargée de Recherche - UMR PACTE 5194 - Université Grenoble Alpes	Examinateur
M. Thomas LOUAIL	Chargé de Recherche - UMR Géographie-Cités 8504 – CNRS	Examinateur
M. Éric DAUDÉ	HDR - Chargé de Recherche - UMR IDEES 6266 Rouen	Codirecteur de thèse
M. Alain VAGUET	HDR - MCF - UMR IDEES-6266 - Université de Rouen	Directeur de thèse

Thèse dirigée par Alain Vaguet et Eric Daudé, laboratoire UMR IDEES 6266

UNIVERSITÉ DE ROUEN NORMANDIE

ÉCOLE DOCTORALE 556 HSRT

idees

Alexandre Cebeillac : *Mobilités urbaines et données en ligne*, 2018

Encadrement :
Alain Vaguet (directeur)
Éric Daudé (co-encadrant)

Établissement :
Université de Rouen, Département de Géographie

Laboratoire de rattachement :
UMR IDEES-6266 CNRS

Financements :
Cette thèse a fait l'objet d'une allocation de mobilité internationale financée par l'INSHS du CNRS.

'Cause he gets up in the morning
And he goes to work at nine,
And he comes back home at five-thirty,
Gets the same train every time.
The Kinks (1965)

J'vais là où la vie me mène,
Là où mes pieds me trainent.
Fonky Familly (2001)

Every move you make
Every bond you break
Every step you take
I'll be watching you.
The Police (1983)

Résumé

Des maladies vectorielles émergentes, comme la dengue, aggravent les crises de santé publique dans les mégapoles asiatiques de Bangkok (Thaïlande) et Delhi (Inde). Les liens entre les moustiques et l'environnement urbain ont été documentés mais la compréhension des mobilités humaines, en tant qu'élément primordial de diffusion des virus, reste un objet de recherche d'intérêt général à développer.

En l'absence de données institutionnelles adaptées, notre recherche s'est d'abord orientée vers des enquêtes de terrain, puis sur la collecte, le traitement, la comparaison et la critique de données provenant d'acteurs majeurs d'Internet (*Twitter*, *Facebook*, *Google*, *Microsoft*). Leur potentiel varie selon les zones géographiques, mais elles permettent d'éclairer l'organisation et la structure des villes étudiées. De plus, elles font ressortir les temporalités et les interactions intra-urbaines. Toutefois, il semble encore difficilement envisageable de se passer de connaissances acquises *in situ*. En utilisant le concept d'espace d'activité, nous proposons une méthode permettant de produire des agendas individuels synthétiques, générés à partir de données Twitter et d'enquêtes de terrains. Il s'agit là d'une première étape dans l'élaboration d'un modèle de mobilité individu-centré à base d'agents.

Mots clés : Mobilité urbaine - Espace d'activité - Traces numériques & données massives - Dengue - Génération d'agendas - Delhi - Bangkok

Abstract

Emerging vector-borne diseases such as dengue intensify public health crises in the Asian megacities of Bangkok (Thailand) and Delhi (India). The links between mosquitoes and the urban environment are well documented, but our understanding of human movement, as a key element of virus spreading, has yet to be fully explored as a research subject.

Given the paucity in adequate or available institutional data, our research first focused on field surveys, and then *on* the collection, comparison and critique of data collected from major Internet platforms (*Twitter*, *Facebook*, *Google*, *Microsoft*). Their potential varies from one geographical area to another, still they shed light on the organization and structure of the studied cities. Moreover, they highlight intra-urban interactions and timeframes. However, carrying out such studies without knowledge acquired from the field seems unadapted. Using the concept of activity space, we propose a method to produce individual synthetic agendas, generated from *Twitter* data and field surveys. This is a first step in the development of an agent-based model of individual mobility.

Keywords : Urban mobility - Activity Space- Digital footprint - Dengue - Agenda-based modelling - Delhi - Bangkok

Publications & valorisation

Revues à comité de lecture

Cebeillac, A., Daudé, É., Vaguet, A., en relecture. « Spatial discontinuities, health and mobility ». *Revue Internationale de Géomatique.*

Rault, Y.-M., Mathew, S., Cebeillac, A., 2018. « The social dynamics of india's shopping malls ». *Bulletin de l'Association des Géographes Français.* 43 60.

Cebeillac, A., Huraux, T., Daudé, É., 2017. « Where ? When ? and how often ? What can we learn about daily urban mobilities from Twitter data and google map in Bangkok (Thailand) and what are the perspectives for dengues studies ? ». *Netcom, Mobilités et (r)évolutions numériques,* https://journals.openedition.org/netcom/2725

Cebeillac, A., Rault, Y.-M., 2016. « Contribution of geotagged Twitter data in the study of a social group's activity space. The case of the upper middle class in Delhi, India ». *Netcom, Réseaux sociaux et territoires.* 231 248. http://netcom.revues.org/2529.

Telle, O., Vaguet, A., Yadav, N.K., Lefebvre, B., Daudé, E., Paul, R.E., Cebeillac, A., Nagpal, B.N., 2016. « The Spread of Dengue in an Endemic Urban Milieu The Case of Delhi, India ». *PLOS ONE 11, e0146539.* https ://doi.org/10.1371/journal.pone.0146539

Chapitre d'ouvrage

Cebeillac, A., Le Bigot, B., à paraître. « Couplage entre enquête ethnographique et traces numériques : application aux mobilités quotidiennes d'un quartier de Bangkok », in : Meissonier, J., Vincent, S., Rabaud, M., Kaufmann, V. (Eds.), *Hybridation des méthodes d'analyse des comportements de mobilité.*

Communications

Cebeillac, A., Daudé, E., Misslin, R., Vaguet, A., « Apport de la génération d'environnements synthétiques et des données mobilités issues de traces numériques à la compréhension des discontinuités spatio-temporelles des épidémies de dengue à Bangkok (Thaïlande) ». Colloque Penser avec les discontinuités Arras, juin 2018.

Cebeillac A., « Couplages d'enquêtes de terrain et de traces numériques : Une approche mixte des mobilités à Delhi ». Séminaire AJEI, MSH-Paris Saclay, mai 2018.

Cebeillac A., Daudé É., Vaguet A., « Discontinuités spatiales, santé et mobiltiés. Analyses et typologies des Google POI et des Tweets pour caractériser les structures spatiales et les dynamiques d'attractivités de Bangkok (Thaïlande) », Actes du colloque SAGéo, Conférence internationale de Géomatique et Analyse Spatiale. Novembre 2017. (Prix du meilleur article). https://hal.archives-ouvertes.fr/hal-01649148/document

Huraux, T., R. Misslin, A. Cebeillac, É. Daudé et A. Vaguet . « Modélisation de l'impact des îlots de chaleur urbains sur les dynamiques de population d'Aedes aegypti, vecteur de la dengue et du virus Zika ». Actes du colloque SAGéo, Conférence internationale de Géomatique et Analyse Spatiale. Novembre 2017. https://halshs.archives-ouvertes.fr/halshs-01650033/document

Cebeillac. A. « Quel potentiel des données Twitter et Facebook dans le cadre d'études sur les mobiltiés quotidiennes ? Exemples à Bangkok et Delhi ». Ateliers Sageo Crowdsourcing et information géographique volontaire. Novembre 2017

Daudé, É., Cebeillac, A., Huraux, T., Maneerat, S., Misslin, R., Vaguet, A. « How Spatial Simulation Models Can Improve Interdisciplinary Researches In Health Geography ? MO3, An Agent - Based Model Of The Dengue Complex Pathogenic System ». 17th International Medical Geography Symposium, Angers, juillet 2017.

Misslin, R., Huraux, T., Maneerat, S., Cebeillac, A., Daudé, É., Vaguet, A. « Effect of urban heat island on the spatio-temporal distribution of Aedes aegypti, vector of dengue and Zika viruses : an agent-based simulation approach ». 17th International Medical Geography Symposium, Angers, juillet 2017.

Cebeillac, A., Daudé, É., Vaguet, A. « Using Twitter data for dengue epidemic modeling in Thailand ». 17th International Medical Geography Symposium, Angers, juillet 2017.

Cebeillac., A. « Twitter et Mobilités : quelques exemples à Delhi et Bangkok ». Séminaire UMR IDEES, juin 2017.

Huraux, T., Cebeillac, A., Misslin, R., et Daudé, É, « MOMOS : modélisation à base d'agents des mobilités quotidiennes en milieu urbain pour la simulation spatiale ». TheoQuant, mai 2017

Cebeillac, A., Daudé É., Huraux T. « Où et Quand ? Que nous apprennent les données Twitter et Google Map sur les mobilités quotidiennes de Bangkok (Thaïlande), coll. A., 8-9 novembre 2016, 15e colloque du GT Mobilités Spatiales, Fluidité Sociale (MSFS), Mobilités et (R)évolutions numériques, Champs-sur-Marne. https://msfs2016.sciencesconf.org/111848

Cebeillac, A., Rault, Y.-M., « Analyser les espaces de vie : quelle complémentarité de l'étude ethnographique et des données Twitter ? Exemple des espaces de vie des couches supérieures à Delhi ». coll. A., 8-9 novembre 2016, 15e colloque du GT Mobilités Spatiales, Fluidité Sociale (MSFS), Mobilités et Conference numériques, Champs-sur-Marne. https://msfs2016.sciencesconf.org/111849

Rault Y.-M., Cebeillac, A. « Luxury, middle classes' new horizon ? A mixed-method approach of luxury consumer's lifestyles and spatial practices in Indian metro cities ». 18emes ateliers Jeunes Chercheurs de l'AJEI. Pondichéry, Mars 2016.

Cebeillac, A., Parihar, R., « Cartography and the treatment of spatial data ». AJEI 2015, Varanasi

Daudé, E., Maneerat, S., Cebeillac, A., Misslin, R., « Dengue as a complex system : case studies in Delhi, Bangkok and "artifical cities" ». CSH, Delhi, 2014

Cebeillac, A., « Open source cartography ». AJEI 2014, New Delhi

Médias

Decryptageo, novembre 2017. « L'avenir de l'épidémiologie passera-t-il par Google et Twitter ? ». http://decryptageo.fr/lavenir-de-lepidemiologie-passera-t-il-par-google-et-twitter/

Le Monde, 7 avril 2015. « Contourner la censure, un jeu d'enfant pour les internautes turcs ». https://www.lemonde.fr/pixels/article/2015/04/07/contourner-la-censure-un-jeu-d-enfant-pour-les-internautes-turcs_4610829_4408996.html

Remerciements

Je tiens tout d'abord à remercier Lena SANDERS et Thomas THÉVENIN qui ont pris le temps de rapporter cette thèse, ainsi que les examinateurs, Isabelle THOMAS, Didier JOSSELIN, Sonia CHARDONNEL et Thomas LOUAIL. Je suis très honoré que vous ayez accepté de faire partie du jury et de relire et de discuter ce travail de recherche.

Je tiens évidemment à remercier chaleureusement Alain VAGUET et Éric DAUDÉ de m'avoir fait confiance et laissé une grande autonomie de travail. Vous avez formé un tandem parfait et complémentaire, et vous avez été toujours disponibles pour me permettre de faire avancer mes questionnements, partager des idées et des concepts et m'aider à me recentrer.

Je voudrais aussi à remercier Somsakun "Tookta" MANEERAT, Renaud MISSLIN et Thomas HURAUX, avec qui j'ai eu le plaisir de travailler. Je remercie également Olivier TELLE et Bertrand LEFEBVRE, qui m'ont énormément aidé en début de thèse en m'apportant des éléments de compréhensions de Delhi et Bangkok. Je remercie aussi Rick Paul qui m'a beaucoup éclairé sur l'épidémiologie.

I also d'like to thank Shankare GOWDA, for the survey we did in the heat in Maviya Nagar, for his professionalism, and for all his comments on Indian' society. Shabash.

Je remercie également mon laboratoire d'accueil en France, l'UMR IDEES de Rouen, notamment mon premier directeur, Michel BUSSI, ainsi que Sophie DE RUFFRAY et Damase MOURALIS. Je remercie aussi Nathalie DUVAL, Sophie DE PEINDRAY et Catherine GODARD qui m'ont permis de réaliser toutes mes démarches administratives, et de comprendre un peu mieux le fonctionnement de ce système complexe qu'est l'administration. Je souhaite aussi remercier Françoise CHARLES, Correspondante Informatique et Liberté (CIL) de l'université Rouen, qui m'a aidé à constituer un dossier pour la CNIL.

Je souhaite aussi remercier le Centre de Sciences Humaines (CSH) de New Delhi, ou j'ai passé plus de deux ans. Merci à Basudeb CHAUDHURI, Pushpa, Bruno DORIN, Jules NAUDET, Odile HENRY, Arnaud KABA, Xavier HOUDOY, Lorraine HOLHER, Bérénice BON, Vandana SOLANKI, Jean-Thomas MARTELLI, et tout le personnel.

Je voudrais aussi remercier Guillaume GAGNEROT et Aurélien CAS, informaticiens au CSH, et Gérald FOLIOT d'Humanum qui ont mis en place des serveurs me permettant de stocker mes données. Je remercie aussi Sébastien REY-COYREHOURCQ qui m'a donné quelques coups de main lorsque je n'arrivais pas à lancer des traitements à distance.

Je remercie les personnes qui ont développé les différentes librairies de l'écosystème de R, notamment Roger BIVAND pour les aspects cartographiques et Hadley WICKHAM pour ggplot2. Je remercie aussi les personnes actives sur les sites Stackoverflow et R-blogger, sources d'idées, d'inspirations et d'astuces pour faire tourner des codes sous R. Je remercie les personnes qui ont développé writer2latex, qui nous a fait gagner beaucoup de temps pour l'édition.

Je remercie aussi les membres de l'Association des Jeunes Études Indiennes (AJEI). Merci à Bérénice Girard avec qui nous avons organisé les ateliers de 2014 à la Jawaharlal Nehru University de Delhi, et merci à Avinash Mishra du Centre for Informal Sector & Labour Studies de nous avoir accueillis. Je remercie aussi les organisateurs des ateliers de 2015, Fabien Provost et Adrien Bouzard et ceux de 2018 Mathieu Ferry et Suneha Seetahul.

Je remercie aussi ma mère et mon père, pour leur soutien et leur amour pendant la garde partagée de ces derniers mois. Maman, t'étais une super coloc'. Je remercie les autres membres de ma famille, Julien, Matthieu, Pauline, Anouk, Cloé, Lulu, Aurore et Benoît.

Je remercie les différentes personnes qui m'ont hébergé lors de mes divers déplacements, pour des durées plus ou moins longues, et particulièrement la famille Hinfray (merci François et Ulrike), la famille Vaguet, les Oudin, Renaud, Remi, Cam' et Ju, et aussi toutes les personnes qui m'ont laissé un coin de canapé ces dernières années.

Je souhaite également remercier les différents doctorants ou jeunes chercheurs, avec qui j'ai pu discuter, échanger, faire des papiers ou envisager des collaborations et qui m'ont surtout permis de réaliser l'importance de croiser des études quantitatives et qualitatives. Je remercie donc Yves-Marie Rault, Brenda Le Bigot, Sébastien Michiels, Floriane Bolazzi, Aurélien Reys, Maxime Courant et Paolo Chevalier.

Je remercie Charlotte, Éloise, Rémi, Seb, Amélia, Éric et Alain qui ont participé à la relecture de ce long manuscrit.

La liste des choses pour lesquels je devrais remercier Rémi de Bercegol est très longue, je vais donc faire court. Merci Rémi d'avoir été un formidable colocataire, un guide remarquable lorsque je faisais mes premiers pas à Delhi et de m'héberger quand je viens sur Paris. Merci d'avoir relu une partie de la thèse et d'apporter des commentaires éclairés. Merci pour ces vacances mémorables à Shimla et à Arcachon.

Je remercie Sébastien Michiels qui a été mon compagnon de route lors de la fin de la thèse. C'était vraiment super de pouvoir discuter de nos recherches, mais aussi de faire de la musique et des bandes dessinées avec le p'tit Arsène. Mais tout cela ne fait que commencer.

Je remercie aussi mes amis en France ou en Inde. Merci donc (sans ordre précis) à Mael, Stevenn, Priyanka, Marion, Nico, Célia, Ju, Camille, Raul, Quintu, Antoine, Pinard, Mister Kapoor, Sébastien, Fab', Amélia, Camy, Arnaud, Rémi, Alok Kumar, Dorian Pathé, Lucie, Somdev, Babaï, Nebu, Solène, Raul, Martin, Florian, Anand, Manmeet Kaur, Khaliq, Arsène, Juliette, Elo, Cécile "Chaf", Camy, Julia, Yasnee, Guillaume, Floriane, Yves-Marie, Fannie, Aardra, Rahul, Gayatri, Raya, Kartik, Quentin, Julien Rox, Julia, Gerland, Senjuti, Jayant, Mathilde, Gina, Carinne, Arka, Maya, Xavier, Pablo, Somdev's mom, et tous ceux que j'ai oublié.

Je remercie évidemment Charlotte, pour tout.

I would especially like to thank Alok Kumar, Main Man G, Shupak, Toni "el dinosir",

PM Jayant Dhoom, le vieux Moine et le Roi pêcheur for these very studious musical weekends in Malviya Nagar. A break in the city uncontrolled noise, to make our own noise that we barely controlled. I would like to thank the Simian Recording label who wanted, surprisingly, to produce the resulting album.

Je remercie aussi France Culture et France Inter, pour toutes les émissions de grande qualité que j'ai pu écouter en baladodiffusion. Vive le service public. j'aimerais adresser un remerciement tout particulier à Axel Villard, dont l'écoute d'une de ces chroniques pour *la tête au carré* durant une sieste nous a insufflé l'idée d'utiliser des données de *Twitter*.

Finalement, je remercie les différents auteurs et musiciens qui m'ont accompagné et épaulé pendant cette thèse. Merci notamment à Apollinaire, Jimi Hendrix, Sonic Youth, Winshluss, le cirque Rouage, Pixies, Larcenet, Beastie Boys, Blur, Svinkels, Enki Bilal, Naïve New Beaters, Nirvana, The Beatles, The Doors, Baudelaire, Gorillaz, Relâche, Fat White Family, Park Circus, Gad, Fabcaro, PuppetMastaz, Boobalavoine et les autres.

Table des matières

Résumé . . . vii
Abstract . . . viii
Publications & valorisation . . . ix
Remerciements . . . xiii

Table des matières xix

Introduction générale 1

Partie A : De la prise en compte des mobilités dans l'étude des épidémies de dengue 11

Chapitre I : Extension du domaine de la dengue . . . 13
1 Épidémiologie de la Dengue . . . 13
2 Mobilités et propagation des épidémies . . . 26
3 La dengue en milieu urbain, sous l'angle de la complexité . . . 35

Chapitre II : L'épidémie et le territoire : Le fardeau de la dengue à Delhi et Bangkok . . . 41
1 Delhi, une mégapole faite de ruptures . . . 42
2 Les discontinuités de Bangkok . . . 54

Chapitre III : Prendre en compte les mobilités dans la modélisation des arboviroses : la possibilité de modèles . . . 79
1 Des débuts des modèles compartimentaux à la prise en compte du vecteur de la dengue . . . 80
2 Modélisation d'épidémies et mobilité humaine . . . 84
3 Ontologie d'un modèle de mobilité urbaine . . . 89

Partie B : Les traces numériques : de «nouvelles» données pour aborder les mobilités 103

Chapitre IV : L'abondance des traces numériques géolocalisées . . . 107
1 Des données protocolaires... . . . 110
2 Création et collecte de données sur les réseaux sociaux . . . 116
3 Collecte massive et identité numérique . . . 124
4 L'appareil législatif : la CNIL . . . 129

Chapitre V : « Nouvelles données » et mobilités urbaines : état de l'art . . . 139
1 Représentativité des différents jeux de données . . . 141
2 À la recherche de lois sur les mobilités individuelles . . . 147
3 Analyse des interactions spatiales . . . 151
4 Modélisation à base d'agents . . . 157

5 Des approches centrées sur les temporalités des activités 162
6 Autres études en contexte urbain 167
7 Utilisation en contexte d'épidémies maladies infectieuses 169
Chapitre VI : Traces numériques à Delhi et Bangkok 179
1 Données *Twitter* à Delhi et Bangkok 180
2 *Check-in Facebook* à Bangkok 211
Partie C : Approche mixte des mobilités à Delhi 229
Chapitre VII : Des enquêtes de terrain pour appréhender les mobilités à Delhi et poser les bases d'un modèle à base d'agents . 233
1 Malviya Nagar et ses alentours, un bon laboratoire pour aborder les mobilités à Delhi 233
2 Présentation de l'enquête de terrain 242
3 Des données terrain à des agendas individuels 269
Chapitre VIII : Traces numériques et espaces d'activités : analyse des mobilités et génération d'agents à Delhi 291
1 Vers une meilleure connaissance de l'échantillon 293
2 Des espaces d'activités discrets à des agendas individuels continus . . 310
3 Générer des agendas de synthèses 322
4 Discussion des résultats 331
Partie D : Mobilités et activités à Bangkok 347
Chapitre IX : Temporalité des activités à Bangkok 351
1 Différentes facettes des activités commerciales 351
2 Apports des données *Google* et *OSM* à la cartographie de l'utilisation du sol à Bangkok 357
3 Fréquentations temporelles 374
4 Définir l'utilisation du sol en fonction les profils temporels des traces numériques 379
Chapitre X : Les mobilités à Bangkok : Variations sur le thème des données et des méthodes 387
1 Le rythme de la ville 388
2 Les interactions dans la ville : données, méthodes et nuances 406
Chapitre XI : Génération d'agendas individuels : Premières notes d'un modèle à base d'agents 423
1 De données épisodiques à des agendas individuels continus 424
2 Génération d'agendas 442
3 Comment affecter des localisations ? 461
Conclusion générale 473

Bibliographie 485
Liste des figures . . . 521
Liste des tableaux . . . 531
Liste des acronymes . . . 532
Annexes 535
Annexe A Un modèle méta-population fermé . . . 536
Annexe B Analyses de nos traces numériques . . . 545
Annexe C D'autres données personnelles géolocalisées . . . 558
Annexe D Dossier CNIL Twitter . . . 560
Annexe E Complément d'information sur les *POI Google* . . . 563
Annexe F Regroupement des catégories de lieux *Facebook* . . . 567
Annexe G Recherche de certains points dans des bases de données externes. . . . 569
Annexe H Informations supplémentaires sur les traitements des données Twitter. . . . 572
Annexe I Questionnaire de terrain . . . 575
Annexe J Graphiques supplémentaires pour le chapitre 7. . . . 578
Annexe K Interprétation des classes de la classification de l'utilisation du sol à Bangkok (chapitre 9). . . . 579
Annexe L Graphiques supplémentaires pour le chapitre 11 . . . 581
Liste des figures (Annexe) . . . 584
Liste des tableaux (Annexe) . . . 584
Table des matières détaillée 585

Introduction générale

Des citadins mobiles dans des villes de plus en plus grandes

Les villes n'ont jamais été aussi attractives. Plus de la moitié de la population mondiale vit aujourd'hui en zone urbaine [1] et cette part n'a jamais cessé de croitre. Si on dénombrait 31 villes de plus de 5 millions d'habitants en 2000, elles sont actuellement 47 et regroupent plus d'un demi milliard de personnes [2]. On observe alors des phénomènes d'étalement ou de densification urbaine, parfois concomitants. Lorsque l'afflux de population dépasse les compétences et/ou budgets des planificateurs urbains, les systèmes de transport peuvent s'avérer sous-dimensionnés et non optimisés, n'ayant pas su être adaptés à l'évolution démographique. Si la pression foncière est trop importante, nous pouvons observer dans certains cas l'apparition de quartiers informels parfois insalubres, en particulier dans les pays du Sud, créant un terreau propice au développement de certaines maladies. En plus de ces considérations sociales et de la qualité de vie, cette tendance à la concentration des populations en zone urbaine implique aussi une augmentation des flux humains quotidiens au sein des différentes agglomérations.

Si les individus n'ont pas tous la même propension à se déplacer, du fait de contraintes socio-économiques (coût de la mobilité) ou de mode de vie et de perception (Kaufmann *et al.,* 2004), l'écrasante majorité se déplace quotidiennement, que cela soit pour se rendre à son travail, rendre visite à des amis, aller chercher les enfants à l'école, pratiquer une activité de consommation (faire les courses) ou simplement se promener. Ces différentes formes de mobilités urbaines, qu'elles soient routinières ou ponctuelles (Horton and Reynolds, 1971), participent à la création des espaces urbains, qui accueillent plus ou moins de personnes, selon leur niveau d'attractivité auprès de différents groupes socio-économiques et des moments de la journée. Les citadins se croisent donc, parfois dans des espaces assez réduits (métro, marché, etc.), sans forcément se toucher ou se connaître [3].

Quand les mobilités humaines contribuent à la propagation des épidémies de dengue

Les contacts directs ne sont pas nécessaires pour propager certaines maladies. Si le virus de la grippe se transmet d'un humain à l'autre par voie aérienne, d'autres sont transmis par l'intermédiaire de vecteurs, en général des insectes piqueurs. C'est par exemple le cas des virus de la dengue, du chikungunya ou du Zika, qui se propagent lorsqu'un moustique du genre *Aedes* (*aegypti* ou *albopictus)* porteur du virus pique et contamine un humain sain. En retour, lorsqu'une personne contaminée est piquée par un moustique sain du même genre, ce dernier

1. http://www.un.org/fr/deve opment/desa/news/popu at on/wor d urban zat on prospects.htm
2. http://www.un.org/en/deve opment/desa/popu at on/pub cat ons/pdf/urban zat on/the wor ds c t es n 2016 data book et.pdf
3. Certaines de ces personnes que nous apercevons fréquemment sans pour autant les connaître, notamment dans les transports en commun, sont qualifiables "d'individus familiers" (L. Sun *et al.,* 2013).

peut à son tour être contaminé. Le malade, ou hôte, transmet donc la maladie de manière indirecte, par l'intermède d'un moustique (le vecteur). On parle alors de maladie infectieuse à transmission vectorielle. Selon les études, entre 50 et 390 millions de personnes seraient contaminées chaque année par la dengue, maladie potentiellement mortelle, et principalement dans les villes des zones intertropicales, où le vecteur est présent (Bhatt *et al.,* 2013 ; Halstead, 2007).

S'il est admis que les moustiques ne se déplacent pas naturellement sur de longues distances, en général largement inférieures à 1 kilomètre (Maneerat et Daudé, 2016 ; Morlan et Hayes, 1958 ; Reiter *et al.,* 1995), ce n'est pas le cas des Hommes, qui de par leurs mobilités quotidiennes peuvent se déplacer dans différents secteurs d'une ville à des distances plus ou moins importantes de leur domicile (González et al., 2008 ; Song *et al.,* 2010a), et ainsi contribuer à la dissémination de la maladie (Reiner *et al.,* 2014 ; Stoddard *et al.,* 2013, 2009). Dès lors, la compréhension de la propagation de la dengue en milieu urbain passe par l'analyse des mobilités humaines (Barmak *et al.,* 2016 ; Daudé *et al.,* 2015 ; Perkins *et al.,* 2014 ; Telle *et al.,* 2016).

Comment étudier les mobilités quotidiennes ?

Mais il est difficile de comprendre les mobilités quotidiennes dans une ville (Banos et Thévenin, 2011), car ces dernières forment un système dont l'étude requiert la mobilisation d'outils fournis notamment par la géographie, la sociologie et la démographie (Bassand et Brulhardt, 1983), mais également l'informatique et la modélisation (Barbosa *et al.,* 2018). Si l'on considère que les mobilités quotidiennes à l'échelle d'une ville sont le produit de l'ensemble des mobilités individuelles qui s'influencent entre elles, une approche individu-centrée semble dès lors appropriée. Mais cette dernière nécessite la collecte de données permettant de constituer un échantillon, selon des approches qualitatives, quantitatives et si possible hybrides (Cebeillac et Le Bigot, à paraître ; Cebeillac et Rault, 2016).

Car les mobilités peuvent être étudiées sous différents angles. Dans son ouvrage « le sens du mouvement », Alain Berthoz décline le terme polysémique de "sens" selon ces trois significations : "direction", "signification" ou encore "perception" (Berthoz, 1997). Nous pouvons ainsi faire le rapprochement avec la notion de déplacement, qui induit des faits observables (mouvement), dotés d'une finalité (se rendre quelque part), avec une représentation du monde (perception et émotions personnelles). Si les outils de la sociologie et de la géographie humaine permettent grâce à une approche qualitative d'analyser simultanément ces trois aspects, ces disciplines ont aussi l'humilité d'accepter que dans la plupart des cas, leurs études ne peuvent être exhaustives et généralisables.

La notion de mouvement peut aussi être étudiée par des outils issus de la physique

des particules ou encore de l'analyse spatiale, définissant des lois et des relations à partir d'observations empiriques (*e.g.* Simini *et al.,* 2012 ; Song *et al.,* 2010a). Mais sont alors écartés les versants de signification et de perception des mobilités. De plus si cette approche quantitative a l'ambition de décrire aux mieux les mobilités d'une grande mégapole notamment via des modèles et des simulations, elle nécessite un échantillon représentatif et/ou de taille importante.

Ces deux approches paraissent complémentaires et reposent toutes deux sur la collecte de données. Alors que les matériaux de bases étaient traditionnellement recueillis *in situ*, via des enquêtes de terrain de plus ou moins grande ampleur, le développement et la généralisation progressive de l'utilisation de moyens de communication en itinérance implique, pour des raisons de protocoles ou d'usages, la création d'un nouveau genre de données : les traces numériques géolocalisées. Ces dernières, de natures, de volumes et de résolutions spatio-temporelles très différentes permettent depuis quelques années de conforter la pertinence d'une approche quantitative (Barbosa *et al.,* 2018), souvent au détriment des contextes socio-géographiques pourtant inhérents aux mobilités individuelles.

Les mobilités individuelles par l'espace d'activité

Le concept *d'espace d'activité* introduit par (Hägerstrand, 1970) fournit un cadre théorique pertinent pour l'analyse et la modélisation des mobilités quotidiennes (Banos et Thévenin, 2011). Cette approche part du principe qu'un individu se déplace pour réaliser une activité dans un lieu donné. D'un point de vue quantitatif, elle permet alors de prendre en compte deux des trois aspects fondamentaux des mobilités, à savoir les temporalités des visites de lieux (ou trajectoires spatio-temporelles), associées à un motif de déplacement (soit la réalisation d'une activité). Chaque individu possède alors un emploi du temps et réalise ces différentes séquences d'activités dans des lieux répartis dans l'espace. Ceci autorise à la fois des études de domaines qui relèvent de la physique et de l'analyse spatiale, notamment sur divers paramètres de dispersions, des fréquences de visites et de retours, etc. et aussi une approche plus sociale, avec des potentiels de mobilités différents selon les individus, en étudiant, par exemple le nombre de lieux fréquentés et les activités effectuées, probablement révélateurs de leur capital économique et culturel.

La flexibilité du concept d'espace d'activité permet également d'utiliser des données collectées *in situ* qui contiennent des informations sur les horaires de fréquentation de certains types de lieux (Banos *et al.,* 2005) (*e.g.* des enquêtes de terrain spécialement conçues pour aller dans ce sens, ou des enquêtes ménages déplacements), ou encore des données collectées *ex situ (i.e.* traces numériques géolocalisées), pour peu que l'on arrive à associer une activité à une localisation datée (Cebeillac *et al.,* 2017). La mobilisation du concept d'espace d'activité permet donc théoriquement de regrouper des considérations qualitatives et quantitatives au sein d'une même recherche.

Comment articuler mobilités et système denguien ?

L'étude de ces mobilités quotidiennes devient encore plus inter-disciplinaire lorsque l'on considère les aspects épidémiologiques, notamment pour mieux comprendre les mécanismes de propagation de maladies vectorielles comme la dengue. Le système des mobilités humaines vient alors s'ajouter à un système épidémiologique déjà complexe, où interagissent moustiques et environnement (Daudé *et al.,* 2015). Mais paradoxalement, la disposition et l'utilisation d'interfaces et de concepts communs et/ou compatibles permettant d'articuler mobilité humaine et épidémiologie entraîne une réduction des approches possibles. Les mobilités humaines qui s'inscrivent en zone urbaine peuvent être vues comme relativement indépendantes des conditions environnementales nécessaires au développement des moustiques, et peuvent donc être étudiées comme un sous-système autonome. Mais les moustiques du genre *Aedes*, surtout *aegypti*, dépendent de la présence d'Hommes pour assurer un repas sanguin indispensable à leur cycle de reproduction. Il faut donc établir des passerelles entre ces différents sous-systèmes.

Cette thèse s'inscrit dans le cadre du programme AEDESS [4]et du projet DENFREE [5], et est développée en parallèle à d'autres travaux, plus spécifiques à l'écologie et au déplacement d'*Aedes aegypti* (Maneerat, 2016 ; Misslin, 2017). Les travaux de Somsakun Maneerat sur le développement et le déplacement d'*Aedes aegypti* sont intimement liés à ceux de Renaud Misslin sur les aspects environnementaux favorables au développement du vecteur. La perception et les déplacements du moustique dans son environnement impliquent l'utilisation d'une échelle de travail extrêmement fine. L'absence de données directement observables et quantifiables a motivé une approche par la simulation à base d'agents qui permet la génération de moustiques de synthèse mobiles (Maneerat et Daudé, 2016), interagissant directement avec un environnement synthétique calibré notamment à partir de données climatiques et météorologiques et centré sur les besoins élémentaires du moustique : gîtes de pontes potentiels, présence de nectar ou encore de zones de repos (Misslin, 2017).

Concernant la modélisation des mobilités quotidiennes, le concept d'espace d'activité est tout à fait compatible avec la génération d'individus de synthèse mobiles dans le cadre d'une modélisation à base d'agent (Banos, 2013). Ce concept est aussi adapté pour étudier les relations et conséquences sur la santé lors de l'exposition d'un individu dans un environnement donné (Perchoux *et al.,* 2013), ce qui paraît pertinent dans le contexte de maladies vectorielles.

Les outils mis à disposition par les systèmes multi-agents (SMA) permettent de décomposer le système complexe qu'est la dengue en trois sous-systèmes plus simples qui

4. Analysis et Spatial Simulation of Dengue Emergence, financé par l'Agence Nationale de la Recherche (ANR). http://www.agence nat ona e recherche.fr/Project ANR 10 CEPL 0004

5. Financé par le Programme cadre pour la recherche et le développement technologique n°7 (FP7) https://www.up2europe.eu/european/projects/dengue research framework for res st ng ep dem cs n europe 16257.htm .

interagissent entre eux : les mobilités humaines, l'environnement et le moustique (Daudé, 2017). Si les données utilisées pour créer et calibrer de tels modèles sont appropriées au système denguien local, il est alors possible de créer des villes virtuelles, sortes de laboratoire permettant de mieux comprendre les phénomènes d'émergences et les processus de propagation d'épidémies dans des villes et sociétés réelles.

Les zones d'études choisies sont Bangkok (Thaïlande) et Delhi (Inde). Il s'agit de deux mégapoles[6] de respectivement 9,4 et 26,4 millions d'habitants en 2016[7], où la dengue est endémique, c'est-à-dire qu'elle y circule chaque année. Le choix de ces villes est aussi motivé par la présence de partenariats avec des organismes de santé locaux, comme le National Institute for Malaria Research (NIMR) à Delhi ou la Bangkok Metropolitan Adminstration (BMA) à Bangkok, qui nous ont fourni les données épidémiologiques (cas de dengues recensés).

Objectifs de la thèse

L'objectif de cette thèse peut se formuler très simplement : il s'agit d'analyser et de modéliser les mobilités humaines dans des mégapoles, ce qui permettrait par une mise en relation avec des modèles environnementaux de mieux comprendre la propagation intra-urbaine de maladies vectorielles comme la dengue.

Il est bien entendu inatteignable.

Néanmoins, nous espérons ici apporter quelques contributions, notamment sur la collecte, le traitement, la critique, la visualisation et l'implémentation de données de diverses natures dans l'étude et la compréhension des mobilités urbaines et dans un modèle à base d'agents.

Nous souhaitons mobiliser autant que possible une approche mixte, combinant à la fois des données, concepts et méthodes provenant d'approches qualitatives et quantitatives. Cela nous conduit d'une part à développer des compétences théoriques et techniques propres à chaque approche, mais également à entretenir un dialogue avec différents champs disciplinaires, et tout particulièrement la sociologie, l'informatique, la physique et l'épidémiologie.

Structure de la thèse

Cette thèse abordera de nombreux thèmes, allant des mobilités quotidiennes, du système de la dengue, des pratiques de modélisation ou encore des traces numériques et des différents enjeux qu'elles induisent. Si la pluridisciplinarité tend peu à peu à se généraliser au sein de la recherche, il est peu probable que tous ces sujets et concepts associés soient maîtrisés par

6. Nous considérons ici Bangkok comme une mégapole, même si les Nations Unis définissent un seuil à 10 millions d'habitants.

7. http://www.un.org/en/deve opment/desa/popu at on/pub cat ons/pdf/urban zat on/the wor ds c t es n 2016 data book et.pdf

tout un chacun. Nous prendrons donc le temps de présenter et de contextualiser chacun de ces thèmes, invitant le lecteur à passer les sections très générales qu'il maîtrise déjà. Ce travail se déroule dans deux villes très différentes que sont Delhi et Bangkok, que nous décrirons sur le plan géographique et sociétal, permettant une meilleure compréhension des particularités et des enjeux locaux de ces mégapoles.

Ces différentes thématiques, complexes et enchâssées les unes dans les autres, ont rendu difficile l'élaboration d'un plan linéaire. Cette thèse a été construite autour de chapitres relativement autonomes, pensés comme des papiers plus ou moins indépendants, mais reliés par le fil conducteur des mobilités urbaines.

Ce travail est divisé en quatre parties. Les parties A & B, principalement théoriques, serviront à orienter les travaux plus appliqués réalisés dans les parties C & D, dédiées respectivement à Delhi et Bangkok (figure 1). Nous avons fait le choix d'inclure dans les parties A & B, des chapitres descriptifs qui concernent nos zones d'études, afin d'apporter progressivement des éléments de contextualisation.

- La première partie traitera du lien étroit entre la dengue et les mobilités humaines :
    - Le premier chapitre consistera en une présentation générale du système de la dengue et de l'extension géographique de la maladie, le tout mis en parallèle avec les divers composants et définitions des mobilités humaines. Ces observations et constats nous inciteront à aborder la question sous l'angle de la complexité.
    - Le second chapitre fera office de présentation générale des mégapoles de Delhi et Bangkok, en insistant sur les structures urbaines, sociales et les mobilités au regard de la situation de la dengue. Si l'approche descriptive employée permet de pointer le rôle de ces déplacements dans la propagation de la maladie, ceux-ci ne sont nullement quantifiés à ce stade et n'expliquent en rien le système de la dengue.
    - La modélisation, qui peut permettre de déverrouiller la situation sera abordée dans le chapitre 3. Ce dernier débutera par une rapide revue bibliographique des différents modèles épidémiologiques pouvant prendre en compte les mobilités. Nous discuterons aussi des divers facteurs qui influencent les potentiels de mobilités. Puis nous présenterons plus en détail le concept d'espace d'activité qui nous permettra de formaliser une ontologie d'un modèle à base d'agents mobiles.
- Pour atteindre un tel objectif, il convient de collecter des données et utiliser des méthodes adaptées, problématiques dont la partie B fait l'objet. Les enquêtes de terrain ou les données institutionnelles (enquêtes ménage-déplacement) sont utilisées classiquement pour étudier les mobilités urbaines ou quotidiennes (*e.g.* Banos

et Thévenin, 2011 ; Commenges, 2013 etc.). Mais l'avènement et la progressive généralisation des technologies de communication en itinérance, impliquent la création d'un grand nombre de traces numériques géolocalisées. Cette partie questionnera leur utilisation dans l'étude et la modélisation des mobilités.

  - Le chapitre 4 reviendra sur la genèse de ces données, et sur leur impact sur nos sociétés contemporaines, notamment vis-à-vis des répercussions sur la vie privée et le rapport au consentement.

  - Nous effectuerons ensuite dans le chapitre 5 un état de l'art sur l'utilisation des traces numériques géolocalisées dans les études de mobilités urbaines,

  - Le chapitre 6 présentera les traces numériques individuelles que nous avons pu collecter à Delhi et à Bangkok, issues de la plateforme *Twitter*, sous forme d'études comparatives. Nous définirons pour chaque individu des espaces de vie, soit les différents lieux que ces personnes fréquentent à des temporalités variables, mais sans prendre en compte pour l'instant l'activité réalisée. Nous mobiliserons également pour Bangkok des données agrégées spatialement à des types d'établissements provenant du réseau social *Facebook*. Ce chapitre comporte également une portée méthodologique et un soin particulier sera apporté à la description des avantages, des limites et surtout des biais potentiels de ces différents jeux de données.

- Cette prise de recul est nécessaire pour la suite du travail, où nous utiliserons les grands volumes de données collectées pour analyser les mobilités à Delhi et Bangkok. Nous compléterons la description de ces villes faites dans le chapitre 2, par l'utilisation de données provenant d'Internet. La partie C se focalisera sur Delhi et interrogera notamment l'intérêt de combiner des données collectées sur le terrain à des données issues de *Twitter*.

  - Le chapitre 7 présentera les résultats d'une étude de terrain réalisée dans un quartier du sud de Delhi. Un protocole simple a permis la collecte d'informations sur les espaces d'activités d'une centaine de personnes. À partir d'une approche probabiliste, nous avons pu reconstituer des agendas continus dans le temps.

  - Dans le chapitre 8, nous compléterons dans un premier temps les espaces d'activités des utilisateurs de *Twitter* à Delhi, en associant pour chaque lieu fréquenté par un individu une activité potentiellement réalisée. Pour cela, nous mobiliserons des données issues d'*OpenStreetMap*, une plateforme cartographique libre, et de *Google Maps*. Nous passerons ensuite de données épisodiques à des agendas continus dans le temps, selon une approche similaire à celle développée dans le chapitre 7. Nous

proposerons ensuite une méthode permettant de générer des agendas de synthèse, applicable à la fois à des données collectées sur le terrain et aux données enregistrées sur la plateforme *Twitter*.

- La partie D traitera des mobilités à Bangkok, uniquement à partir de données récoltées sur Internet.

    - Nous présenterons les différentes formes d'organisations commerciales de Bangkok dans le chapitre 9. L'utilisation de données issues du service cartographique de *Google* et *d'OpenStreetMap* nous a permis de proposer différentes méthodes permettant d'obtenir une couche d'utilisation du sol. Cette dernière, couplée aux données *Twitter* ou *Facebook* permet d'apprécier les temporalités des activités dans la ville.

    - Le chapitre 10 s'intéressera aux mobilités globales dans Bangkok. Nous aborderons la difficulté de se déplacer dans la ville, très souvent congestionnée. Ceci est explicable en grande partie par le caractère principalement mono-centrique de la capitale Thaï et de l'importance des flux de type centre / périphérie. Néanmoins, une analyse plus fine passant par la construction de matrices origine-destination permettra de mettre aussi en avant l'importance des déplacements locaux et la délimitation de zones de mobilités fonctionnelles.

    - Nous traiterons enfin dans le chapitre 11 les mobilités d'un point de vue individuel. Les plus grands volumes de données et le bon niveau de précision de notre couche d'utilisation du sol nous autorise à reprendre et à améliorer (et simplifier) les méthodes de reconstitution et de génération d'agendas préconisées dans le chapitre 8. Les résultats obtenus sont ici très cohérents et encourageants, car les profils de fréquentation de la plupart des activités sont reproduits plutôt fidèlement. Néanmoins, des enquêtes de terrains ou des données institutionnelles désagrégées auraient pu être d'un grand soutien. Nous discuterons finalement de quelques pistes d'amélioration et d'affectation de localisation.

D'un point de vue idéologique, ce travail a été réalisé en n'utilisant que des outils informatiques libres et gratuits, allant du système d'exploitation (Ubuntu), aux logiciels de traitements statistiques et géographiques (R et qGis). Nous mettrons d'ailleurs prochainement à disposition les codes qui nous paraissent pertinents sous forme d'une librairie R. Aucune donnée n'a été achetée à des tiers et l'auteur a effectué lui-même tous les traitements qui seront présentés ici (de la collecte à l'analyse), sauf indications contraires.

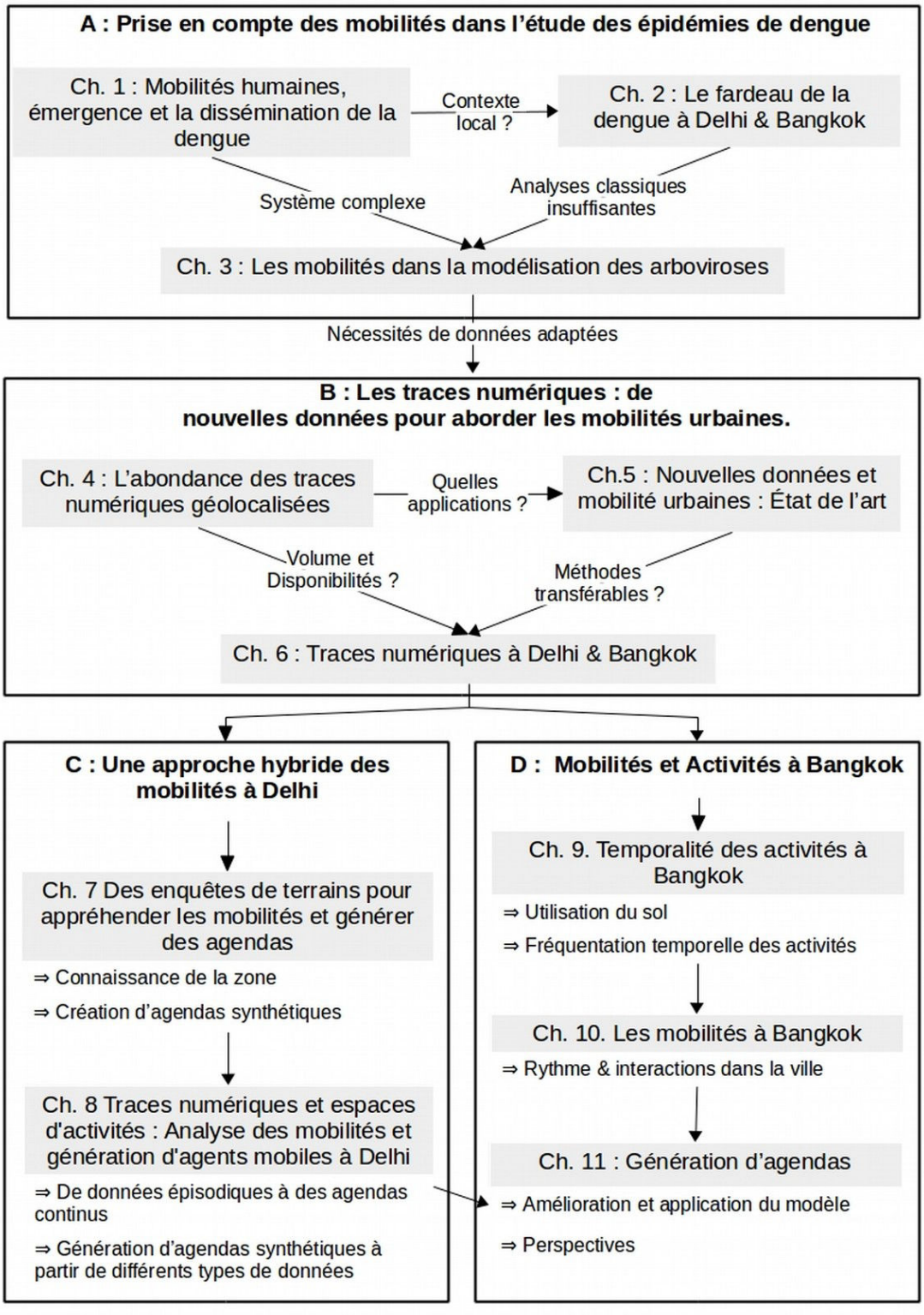


FIGURE 1 Plan de la thèse

Partie A:

De la prise en compte des mobilités dans l'étude des épidémies de dengue

Chapitre I:
Extension du domaine de la dengue

« Toutes les mairies du monde luttent contre toi, mais toi tu es toujours vivant. Elles emploient l'insecticide, l'acide, mais tu es invincible. On a inventé les bombes pour tuer l'humanité, mais les moustiques on n'arrive pas. Tu es dans les mares, tu es dans les jarres et tu es dans les bars (. . .) Moustique, ah moustique ! Tu es un salaud ! », Casimir Zao (1988).

1 Épidémiologie de la Dengue

1.1 Généralités sur la Dengue

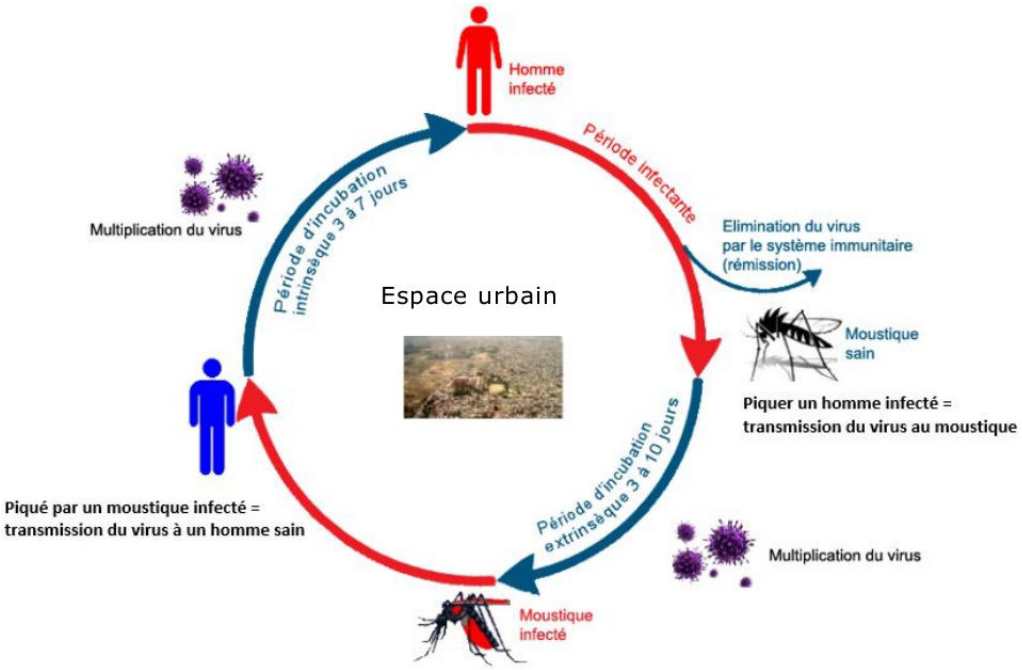


FIGURE 2 Cycle de transmission de la dengue, d'après S. Maneerat (2015)

La dengue est un virus de la famille des *Flaviviridae*, du genre Flavivirus, comme la fièvre jaune, le chikungunya ou le Zika (Gubler, 2014 ; Lindenbach et al., 2007). C'est une maladie infectieuse vectorielle, c'est-à-dire transmise par l'intermédiaire d'un vecteur, en l'occurrence des moustiques[8] du genre *Aedes*[9] : principalement *Aedes aegypti* et A*edes albopictus*[10] (Brancoft, 1906 ; Graham, 1903 ; Gubler, 2014). Seule la femelle pique l'Homme, car elle a besoin de sang humain pour son cycle de reproduction. Cette maladie toucherait, selon estimations, entre

8. Ce mode transmission par des arthropodes la classe comme une d'arbovirose.

9. Signifiait « déplaisant » en grec, mais « maison » ou « temple » en latin (dictionnaire collins). Les deux origines semblent cohérentes, *Aedes* étant relativement « déplaisant » et vie prés de l'homme et dans ces maisons.

10. Du latin *albo* « blanc » et *pictus* « coloré ». Également appelé moustique tigre du fait de ces rayures blanches.

50 et 100 millions de personnes par an, principalement en zone intertropicale dont 500 000 seraient admis à l'hôpital (Halstead, 2007). Le virus est adapté à la fois à ses vecteurs, et à ses hôtes (principalement les humains), ce qui signifie qu'un moustique infecté peut contaminer un humain sain, et inversement, une personne malade peut transmettre le virus au moustique sain qui le pique (figure 2) (Gubler, 1997 ; Lindenbach *et al.,* 2007).

Les moustiques femelles *Aedes* piquent en journée, principalement le matin et le soir, lorsque les Hommes ne sont pas forcément à leur domicile. Ainsi, les moustiquaires imprégnées, pierre angulaire de la lutte contre le paludisme[11] (Darriet, 2014), sont d'une efficacité réduite contre la dengue. Le virus peut également se transmettre d'Homme à Homme, par l'intermédiaire de contacts sanguins directs, comme lors de transfusion de sang contaminé ou la réutilisation des mêmes outils entre deux opérations chirurgicales (Wagner *et al.,* 2004), même si ce mode de transmission est relativement anecdotique (Wilder-Smith, 2014).

1.1.1 Aspects cliniques

Il existe 4 différents sérotypes[12], ou souches, soit des variations du virus : DENV1, DENV2, DENV3, DENV4 (Gubler, 1997). Ces multiples souches sont suffisamment différentes génétiquement pour ne pas conférer d'immunité croisée durable (Gubler, 2014 ; Iglesias *et al.,* 2014). Par exemple, une personne affectée par DENV1 sera, une fois guérie, *a priori* immunisée contre ce sérotype, mais pourra après quelques semaines être contaminée par les autres souches (Adams et al., 2006 ; Reich *et al.,* 2013 ; Salje *et al.,* 2012). La variété des souches peut s'expliquer entre autres par le fait qu'il s'agit d'un virus qui se réplique par ARN, ce qui implique plus d'erreurs lors des réplications que pour les virus à ADN (Bennett, 2014). Il est donc génétiquement moins stable et sa propension à muter est plus importante (Bennett, 2014 ; Iglesias *et al.,* 2014).

Après qu'un individu soit infecté, s'écoule une période d'incubation de 2 à 7 jours. La dengue se manifeste ensuite par de fortes fièvres, souvent accompagnées d'au moins deux des symptômes suivants : maux de tête, douleurs oculaires, articulaires ou musculaires, éruptions cutanées, manifestations hémorragiques dues à une baisse du niveau de plaquette (Institut Pasteur, (Telle, 2011 ; The Trinh et Wills, 2014). Ces premiers symptômes sont assez proches de ceux du paludisme, du Zika et du chikungunya, d'où la possible confusion (Telle, 2011). Pourtant un grand nombre de personnes contaminées ne développent pas de symptômes, ou alors de manière très atténuée (Chastel, 2012 ; Duong *et al.,* 2015). Elles sont asymptomatiques[13] mais continuent à jouer un rôle dans la diffusion de la maladie (Chastel, 2012 ; Duong *et al.,* 2015). Bien qu'elle soit difficile à quantifier pour la simple et bonne raison que ces personnes ne sont

11. L'anophèle, vecteur du paludisme pique principalement la nuit.
12. Bien que la découverte récente d'un nouveau sérotype DENV5 soit en cours de confirmation (Normile, 2013).
13. Où porteurs sains.

pas prises en charge cliniquement et donc non comptabilisées par les systèmes de santé, cette part de porteur sain représenterait environ les trois quarts des personnes infectées (Duong *et al.,* 2015 ; Grange *et al.,* 2014). Ces personnes asymptomatiques (ou aux symptômes très atténués) seraient impliquées dans près de 90 % des transmissions de dengue (ten Bosch *et al.,* 2018).

À l'opposé, dans certains cas (˜1 %), la maladie peut évoluer vers une forme plus sévère, la dengue hémorragique où la fièvre persiste et les hémorragies internes se multiplient du fait d'une chute du taux de plaquettes dans le sang. La maladie peut encore évoluer vers la dengue avec syndrome de choc et être mortelle si les soins appropriés ne sont pas apportés (*e.g.* en général une transfusion sanguine pour pallier le manque de plaquettes).

Le fait que certaines personnes soient asymptomatiques tandis que d'autres développent des formes sévères de la maladie pourrait s'expliquer par les spécificités génétiques de leur système immunitaire (Coffey *et al.,* 2009), ou par l'âge (Huy *et al.,* 2013). Il a été également observé que dans de nombreux cas, une seconde infection (soit une contamination ultérieure par une autre souche du virus) entraîne des conséquences en général plus graves pour la santé (notamment plus de cas de dengue hémorragique) qu'une primo-infection (Halstead *et al.,* 1970, p. 197, 1969). Une hypothèse est qu'au lieu de rentrer en compétition, les différentes souches virales se renforcent mutuellement, selon le principe d'« antibody-dependant enhancement » (ADE), où des anticorps facilitent l'entrée des nouvelles souches de virus dans les cellules de l'hôte (Dejnirattisai *et al.,* 2010 ; Halstead et O'rourke, 1977).

Après plus de 20 ans de recherches, un vaccin candidat contre les quatre souches de la dengue a été testé en phase 2b [14] en Thaïlande en 2009 et affichait une efficacité de 30 %, avec des différences entre les souches (Halstead, 2012 ; Sabchareon *et al.,* 2012). Une campagne de phase 3 de plus grande envergure a été ensuite menée pendant 3 ans sur un échantillon 35000 enfants entre 2 et 16 ans en Asie du sud-est et en Amérique du Sud. Il en ressort une efficacité globale de 60 %, avec des différences entre les souches et les tranches d'âge (Hadinegoro *et al.,* 2015). Le vaccin permet notamment de réduire de manière significative le nombre d'hospitalisations chez les plus de 9 ans et ce pendant au moins 4 ans (Gailhardou *et al.,* 2016 ; Hadinegoro *et al.,* 2015). Le vaccin, fabriqué par Sanofi-Pasteur sous le nom de *Dengvaxia*® a été recommandé par l'OMS en avril 2016 était autorisé dans 11 pays en juillet 2017 [15], dont la Thaïlande et le Brésil. Cela dit, il apparaît que le vaccin est plus efficace chez les personnes déjà contaminées par une souche du virus, et qu'il accroît les risques d'hospitalisation dus à la dengue chez les personnes initialement séronégatives (Aguiar *et al.,* 2016), notamment chez les 2-5 ans trois ans après leur vaccination (Halstead, 2016). Une hypothèse (non vérifiée pour l'instant) serait que la vaccination entraîne une forme de primo-infection asymptomatique. Une contamination ultérieure qui survient malgré la vaccination s'apparenterait alors à une

14. Soit une étude pilote sur un échantillon restreint.

15. www.sanofipasteur.com/en/art c es/first dengue vacc ne approved n more than 10 countr es.aspx visité en juillet 2017, mais hors ligne au moins depuis juillet 2018.

seconde infection, en général plus sévère. Les campagnes des vaccinations ont d'ailleurs été interrompues aux Philippines fin 2017 [16].

En l'absence de vaccin efficace sans effets secondaires notoires, la lutte anti-vectorielle est le meilleur moyen de réduire le fardeau que présente cette maladie. Cela passe notamment par la compréhension des mécanismes de développement du moustique, au regard des conditions environnementales.

1.1.2 Écologie des vecteurs

Le cycle de reproduction des moustiques se divise en deux stades : le stade aquatique, sous forme d'œufs, de larves puis de nymphes, et le stade aérien avec le moustique à proprement parler (figure 3).

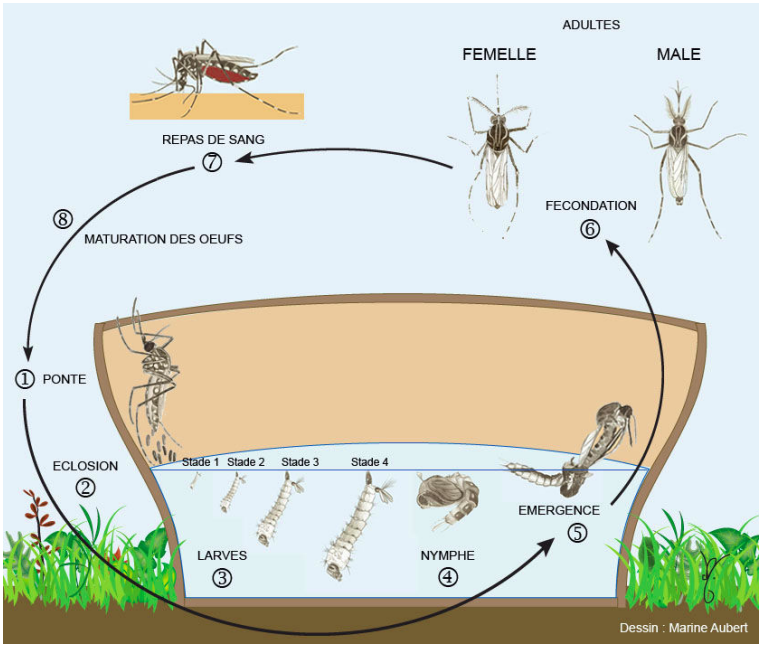


FIGURE 3 Cycle de vie du moustique d'après Laurent Guillaumot et Marine Aubert Source : Institut Pasteur Nouvelle-Calédonie, lien : http://www.institutpasteur.nc/les-moustiques-et-la-dengue/

À l'âge adulte, la femelle *Aedes* s'accouple avec un mâle et crée une « spermathèque » qui lui permettra de pondre plusieurs fois à partir d'un seul accouplement. Cette hypothèse largement répandue est cependant remise en cause après que des cas de polyandrie aient été observés (Degner et Harrington, 2016). Quoi qu'il en soit, après l'accouplement, la femelle *Aedes* cherche ensuite de la nourriture (nectar chez les plantes à fleurs) et des humains à piquer

16. https://www.emonde.fr/economie/article/2018/02/06/vaccination-contre-la-dengue-la-justice-philippine-attaque-sanofi_5252422_3234.html

pour assurer un repas sanguin [17] indispensable au cycle gonotrophique [18]. Après maturation des œufs, les femelles *Aedes* pondent préférentiellement dans de petits réservoirs d'eaux claires (Morrison *et al.,* 2004). De fait, les conteneurs créés par l'homme, tels que les coupelles de pots de fleurs, ou n'importe quel objet abandonné contenant de l'eau (pneus usés, assiettes cassées, etc.) sont pour elles d'excellents gîtes larvaires (Chadee, 2009 ; Reiter et Sprenger, 1987 ; Tun-Lin *et al.,* 1995) La non-connexion à un réseau d'adduction ou le manque d'eau en saison sèche conduit souvent des habitants à stocker de l'eau claire dans des réservoirs à proximité ou dans leur domicile, ce qui augmente le nombre de lieux de pontes pour les moustiques (WHO, 2012). De fait, la présence d'un grand nombre de moustiques *Aedes* est souvent associée à des critères sociaux économiques et de développement. Mais bien qu'ils pullulent en général dans les quartiers défavorisés (Åström *et al.,* 2012), cette relation n'est pas toujours vérifiée (Ríos-Velásquez *et al.,* 2007). Par exemple, à Delhi, de nombreux moustiques sont présents dans les quartiers les plus riches, du fait de la présence de jardins fleuris et de *coolers* [19], créant autant de gîtes larvaires potentiels (Telle, 2011 ; Telle *et al.,* 2016) Mais de manière générale, la prolifération d'*Aedes*, notamment *aegypti*, en milieu urbain dans les pays en voie de développement est favorisée par une urbanisation non maîtrisée (Gubler, 1998). La transformation de l'environnement local par l'Homme offre donc aux moustiques du genre *Aedes* des lieux propices à la ponte et les précipitations interviennent surtout dans la recharge en eau de ces derniers.

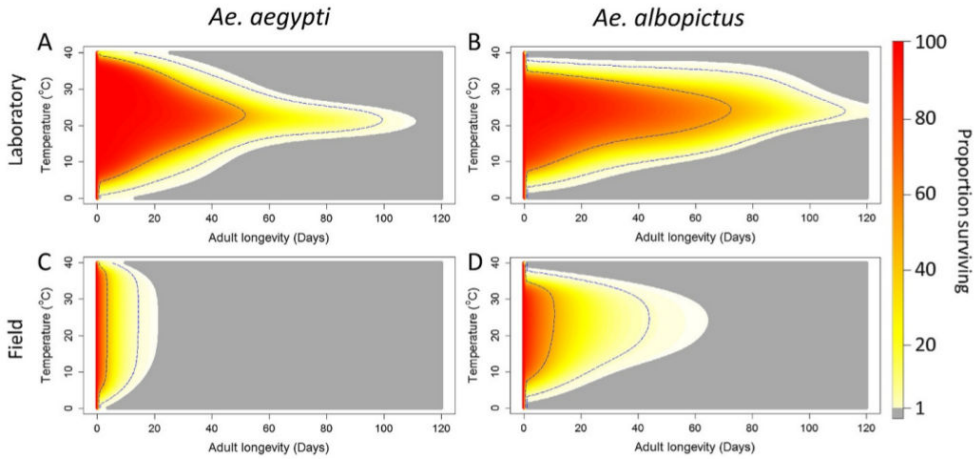


FIGURE 4 Probabilité de survies des femelles *Aedes* en fonction de la température, selon des conditions de laboratoires ou de terrain. D'après (Brady *et al.,* 2013).

Si les réservoirs en eau sont indispensables à l'accomplissement d'un cycle de reproduction du moustique, la température, comme pour toutes les espèces, joue un rôle primordial dans leur fonctionnement biologique et la température de confort d'*Ae. albopictus* et d'*Ae.aegypti* est

17. Une femelle Aedes peut piquer plusieurs personnes lors d'un même repas sanguin.
18. Le développement des œufs.
19. Il s'agit d'un système de refroidissement consistant en un bac rempli d'eau et d'un système de ventilation.

comprise entre 20 et 30 °C (Brady *et al.,* 2013). Mais bien qu'*Ae. aegypti* puisse survivre dans un plus grand spectre de température qu'*Ae. albopictus*, son taux de survie et sa longévité est globalement inférieure (Brady *et al.,* 2013) (figure 4). Leurs œufs résistent au dessèchement et au froid et peuvent donc rester dormant pendant plusieurs mois, attendant les conditions propices pour se développer. Mais les œufs d'*Ae. albopictus* peuvent théoriquement survivre à des températures plus extrêmes que ceux d'*Ae. aegypti*, jusqu'à des moyennes minimales de -2 °C (Kobayashi *et al.,* 2002), leur conférant un avantage pour se développer dans des zones plus septentrionales (Rodhain, 1995).

En plus de leur biologie, leur comportement fait qu'ils occupent des niches écologiques différentes (Eisen et Moore, 2013). *Ae. albopictus* privilégie en effet les environnements plus ruraux (Tsuda *et al.,* 2006), tandis qu'*Ae. aegypti* est maintenant parfaitement adapté aux environnements urbains (Brown *et al.,* 2014 ; Powell et Tabachnick, 2013). Les îlots de chaleur urbains entraînent en effet des températures plus élevées et plus stables que dans les campagnes (Misslin *et al.,* 2017). Mais les deux espèces sont parfois en compétition dans certaines régions du monde, et *Ae. albopictus* est par exemple en train de remplacer *Ae. aegypti* dans le sud-est des États-Unis (Benedict *et al.,* 2007 ; Juliano *et al.,* 2004). D'un point de vue virologique, ces deux moustiques n'ont pas la même propension à propager la dengue. *Ae. albopictus*, originaire d'Asie (Gubler, 2014), serait le vecteur originel de la dengue (Smith, 1956), mais il est moins « performant » qu'*Ae. aegypti* (originaire d'Afrique) pour transmettre le virus à l'Homme (Guzmán et Kouri, 2002). D'autres moustiques du genre *Aedes*, comme l'espèce endémique du Pacifique *Aedes polynesiensis*, peuvent également transmettre la dengue, mais elle est biologiquement moins performante qu'*Ae. aegypti* (Tsuda *et al.,* 2006)Ainsi, dans les régions où *albopictus* tend à remplacer *aegypti*, le risque de transmission de la dengue à l'Homme se réduit.

1.1.3 Distribution spatiale

Le développement des moustiques du genre *Aedes* est conditionné par la présence d'eau claire, et des températures relativement élevées, ce qui explique pourquoi on les retrouve essentiellement en zone intertropicale. La probabilité de distribution d'*Aedes albopictus* et *aegypti* a été estimée et simulée à l'échelle mondiale, à partir de plus de 40 000 occurrences dans des articles scientifiques et de recherche de terrains (Kraemer *et al.,* 2015a), le tout combiné avec des données climatiques et environnementales (Kraemer *et al.,* 2015b) (figure 5 et 6). Les contributions des variables utilisées pour leur modélisation montrent que la température est le paramètre principal pour les deux espèces, avec une contribution supérieure pour *aegypti* (54 %) que pour *albopictus* (46 %). Les variables de précipitations et de végétation contribuent plus à la distribution spatiale *d'albopictus*, tandis que le niveau d'urbanicité[20] est plus élevé

20. La propension à se développer dans des zones urbaines.

chez *aegypti* (Kraemer *et al.*, 2015b).

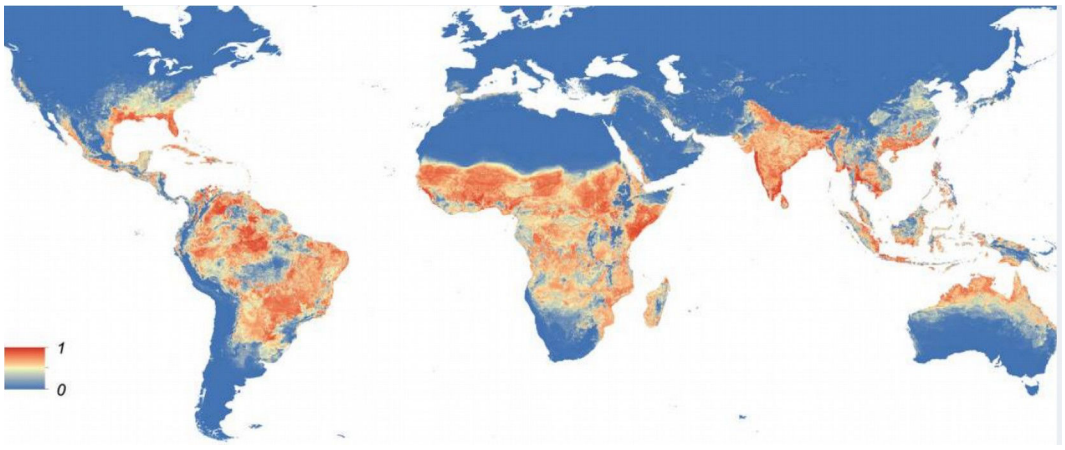


FIGURE 5 Probabilité d'occurrence d'*Aedes aegypti*, d'après Kraemer *et al.* (2015b).

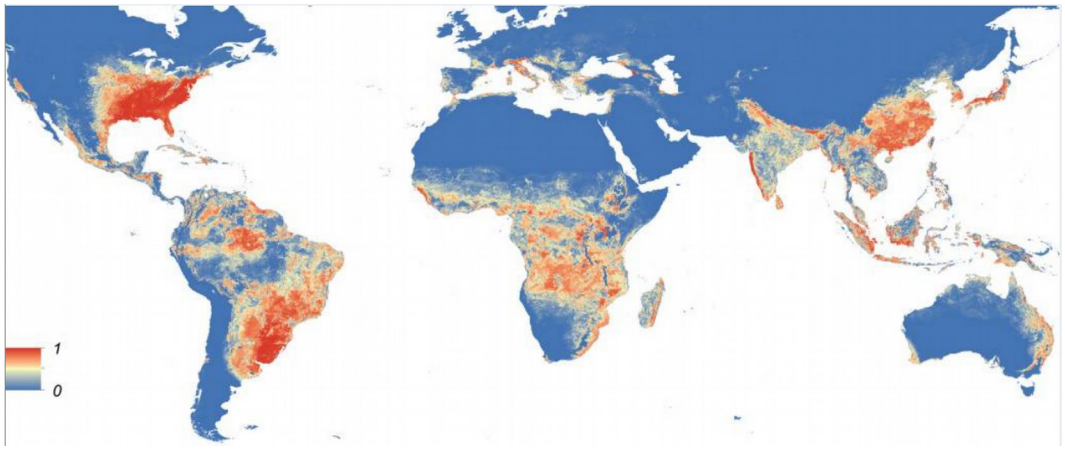


FIGURE 6 Probabilité d'occurrence d'*Aedes albopictus*, d'après Kraemer *et al.* (2015b).

Ces deux figures ci-dessus montrent que les aires de répartition *d'albopictus* et *d'aegypti* se superposent presque, avec un plus grand potentiel de développement en zones intertropicales pour *Ae. aegypti* et une amplitude latitudinale plus élevée pour *Ae. albopictus*. La prochaine section présentera l'évolution de ces aires de répartition au cours du temps.

1.2 Une petite histoire de la dengue

> « La géographie n'est autre chose que l'histoire dans l'espace, de même que l'histoire est la géographie dans le temps » d'Élisée Reclus, *l'Homme et la Terre* (1905-1908).

1.2.1 Evolution sémantique

L'étymologie des noms donnés aux maladies est un bon indicateur concernant les connaissances et croyances passées sur ces dernières. Par exemple le terme « Malaria », vient du latin *Mal'Aria*, qui signifie « mauvais air », laissant suggérer une contamination par un milieu défavorable à l'Homme, conformément à la théorie des miasmes (Hempelmann et Krafts, 2013). Le nom "Paludisme", vient du latin « palus », qui désigne des marais[21], ce qui lie directement l'environnement à la maladie. Le terme de « Dengue » viendrait quant à lui du Swahili « Kidinga pepo » qui fut le nom donné aux épidémies ayant sévi en 1823 et 1870 à Zanzibar et sur la côte Est-Africaine[22](Christie, 1872). Le mot *kindiga* (crampes) associé à *pepo* (esprit maléfique), signifierait littéralement « douleurs similaires à des crampes, produites par les forces d'un esprit maléfique[23] » (*ibid.)*.. Les différents noms donnés à la maladie au cours du temps sont d'ailleurs très souvent liés aux symptômes : « *Coup de Barre* » dans les Antilles en 1635, le « mal de genoux » en Égypte et « *fièvre des os* » à Jakarta en 1779, la « *scarlatine rhumatique* », la « *fièvre des os brisés* » à Philadelphie en 1780, la « Fièvre éphémère » à Calcutta en 1824, « fièvre de trois ou de sept jours » en Inde en 1909 puis « fièvre de cinq jours » en Indonésie dans les années 60 (Gubler, 2014).

Mais les premières traces écrites de symptômes compatibles avec ceux de la dengue sont bien antérieures. Elles datent de 992 et ont été retrouvées en Chine dans une « encyclopédie des symptômes des maladies et des remèdes » et fait référence à des articles plus anciens, l'un datant de 610 et un autre de la dynastie des Chin (entre 265 et 420 après J.C) (Gubler, 2006 ; Nobuchi, 1979). Elle y était alors nommée « water poison » (Gubler, 2006) faisant donc référence à l'environnement propice à la déclaration de la maladie. Cela dit, d'un point de vue prévention, nommer une maladie en fonction du type d'environnement dans laquelle elle semble pouvoir se transmettre chez l'Homme est assez pertinent, car les zones à éviter sont alors directement connues par les populations. Nommer une maladie en fonction de ces symptômes facilite certes le diagnostic médical, mais ne véhicule pas de connaissances, même inexactes, sur les modes de contamination.

1.2.2 Le moustique comme vecteur de maladies

Quoi qu'il en soit, la première épidémie de dengue à être scientifiquement décrite et documentée est celle survenue à Philadelphie en 1780 (Gubler, 2006 ; Rush, 1789 ; Wilder-

21. Il est d'ailleurs encore employé en tant que toponyme en Gironde, pour décrire des zones marécageuses en bordure de l'estuaire, de la Dordogne ou de la Garonne. Nous pouvons citer par exemple le lieu dit de "Clos Palu" à Saint Vincent de Paul, le chemin de la Palu à Tauriac, ou encore le quai de Paludate, à Bordeaux.

22. « Those inhabitants of Zanzibar who were pretty well advanced in life at once recognised it as a disease which was epidmic on the coast of Africa about forty eight years ago, and which was tehn called « kindinga pepo ».

23. « Cramp like pains, produced through the agency of an evil spirit »

Smith *et al.,* 2013). Rush y décrit notamment des fièvres malignes et contagieuses qu'il nomme « Bilious remitting fever ». Mais bien que les moustiques aient été suspectés de tout temps de transmettre des maladies (Service, 1978), aucun lien de cause à effet n'a été prouvé à l'époque entre cette épidémie et la présence de moustiques. Après une épidémie aux symptômes similaires survenue à nouveau à Philadelphie entre 1802 et 1803[24], Stubbins Ffith prouva par une série d'auto-expériences[25] que cette fièvre n'était pas directement contagieuse (Ffirth, 1804).

Il faudra attendre 1877 pour que Patrick Manson prouve que les moustiques peuvent transmettre des pathogènes, en isolant le ver responsable de la filariose (*Wucheriria bancrofli*) (Smith *et al.,* 2012). Trois ans plus tard, en 1880, Charles Louis Alphonse Laveran, alors médecin à Constantine, observe que le sang des personnes infectées par la malaria contient un parasite[26] (Bacaër, 2011 ; Smith *et al.,* 2012). Manson suggère que ce parasite puisse être transmis par un moustique, même s'il pense que l'infection vers l'homme ne se fait par piqûres mais via de l'eau contaminée par ces moustiques (Bacaër, 2011). Ronald Ross découvre en 1897, dans les Indes britanniques, que les moustiques du genre anophèle[27] transmettent le parasite responsable de la malaria à des oiseaux par leurs piqûres[28] (Ross, 1897). Dans la foulée, une équipe de parasitologues italiens menée par Grassi montrent que le parasite se transmet de la même manière chez l'homme (Bastianelli *et al.,* 1898) et décrivent le cycle complet du parasite (Grassi *et al.,* 1899).

Les travaux sur la dengue s'inscrivent dans la lignée de ceux sur la malaria et la fièvre jaune et en 1903 une étude suggère qu'elle se transmet par des moustiques (Graham, 1903). Le rôle d'*Aedes* est finalement démontré par Thomas Bancroft en 1906 (Brancoft, 1906 ; Smith *et al.,* 2012).

1.2.3 Isolement du Virus

Il faudra attendre 1943 pour que le virus de la dengue soit enfin isolé[29] par une équipe japonaise (Kimura et Hotta, 1944), mais ces résultats ne traversèrent les frontières que des années plus tard (Gubler, 2014). Sabin et son équipe isolèrent en 1945 deux souches de virus, l'une provenant de militaires stationnés à Hawaii qu'ils nommèrent DENV-1, l'autre provenant de militaires en poste en Nouvelle-Guinée, DENV-2 (Gubler, 2014 ; Sabin, 1952). Deux autres sérotypes, DENV-3 et DENV-4, furent isolés après une épidémie de Dengue à Manille en 1956 (Gubler, 2014 ; Hammon *et al.,* 1960). Une autre souche DENV-5 aurait été découverte en 2013

24. Qui s'avéra être la fièvre jaune.
25. Impliquant notamment l'application de vomi de personnes infectées sur des plaies ouvertes et dans les yeux, l'inhalation de ces vapeurs ou encore l'ingestion par voie orale.
26. Il obtiendra (tardivement) le prix Nobel de médecine pour cette découverte en 1907.
27. Du grec *anofelí*, signifiant « inutile ».
28. Il obtiendra lui aussi le prix Nobel de médecine, mais en 1902 et sera le précurseur de la modélisation mathématique des épidémies (chapitre 3)
29. Qui s'avèrera être DENV 1.

(Mustafa *et al.*, 2015 ; Normile, 2013) mais aucun article descriptif de l'étude phylogénétique n'a été publié pour l'heure par l'équipe impliquée.

1.2.4 Évolution des cas de Dengues dans le monde

La figure 7, montre la localisation (parfois approximative[30]) d'épidémies aux symptômes similaires à ceux de la Dengue et survenues entre 1635 et 1950 d'après les études rapportées par Gubler (2014). Elle met en avant que des épidémies ont été observées sur tous les continents, en zone intertropicales et parfois plus au nord, qu'elles touchent simultanément de plus en plus de zones au cours du temps et que les épidémies se rapprochent temporellement.

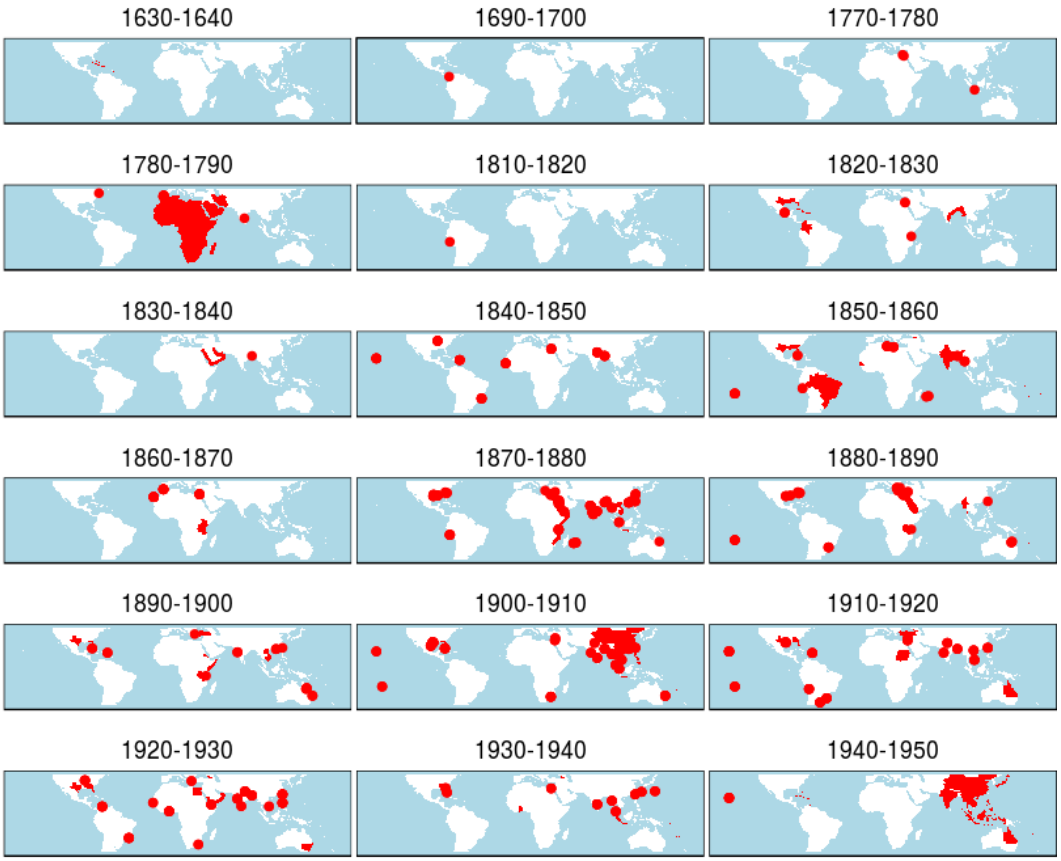


FIGURE 7 Zones ayant connu des épidémies aux symptômes similaires à ceux de la dengue (rouge) entre 1635 et 1950. Cartographié d'après Gubler (2014).

L'évolution des épidémies de dengue depuis les années 1960 est visible sur la figure 8 ci-après, réalisée à partir des travaux de (Messina *et al.*, 2014) pour la répartition des cas de dengue et de (Kraemer *et al.*, 2015a) pour celle des vecteurs. Elle montre tout d'abord

30. Des zones peuvent paraître très touchées, comme l'Afrique entre 1780 et 1790 ou le Brésil entre 1850 et 1860. Il s'agit d'artefacts liés aux approximations de la localisation de certaines épidémies, où seuls le continent ou le pays sont référencés.

que les épidémies de dengue sont récurrentes, voire endémiques en Asie du Sud-Est depuis les années 1960, tandis qu'elles débutent sur les côtes sud- américaines au milieu des années 1980 et touchent maintenant tout le continent. Nous observons aussi les impacts des politiques de démoustications, visibles dans le sud-est des États-Unis dans les années 60. Les moustiques y prolifèrent à nouveau à partir des années 1985, tandis que la colonisation du sud de l'Europe par *Albopictus* est amorcée à partir des années 1990. Cet ensemble de cartes montre aussi les systèmes de surveillances. En effet, le nombre de cas de dengue et de moustiques *Aedes* enregistrés en Afrique paraît bien faible par rapport à des zones où les conditions climatiques sont similaires.

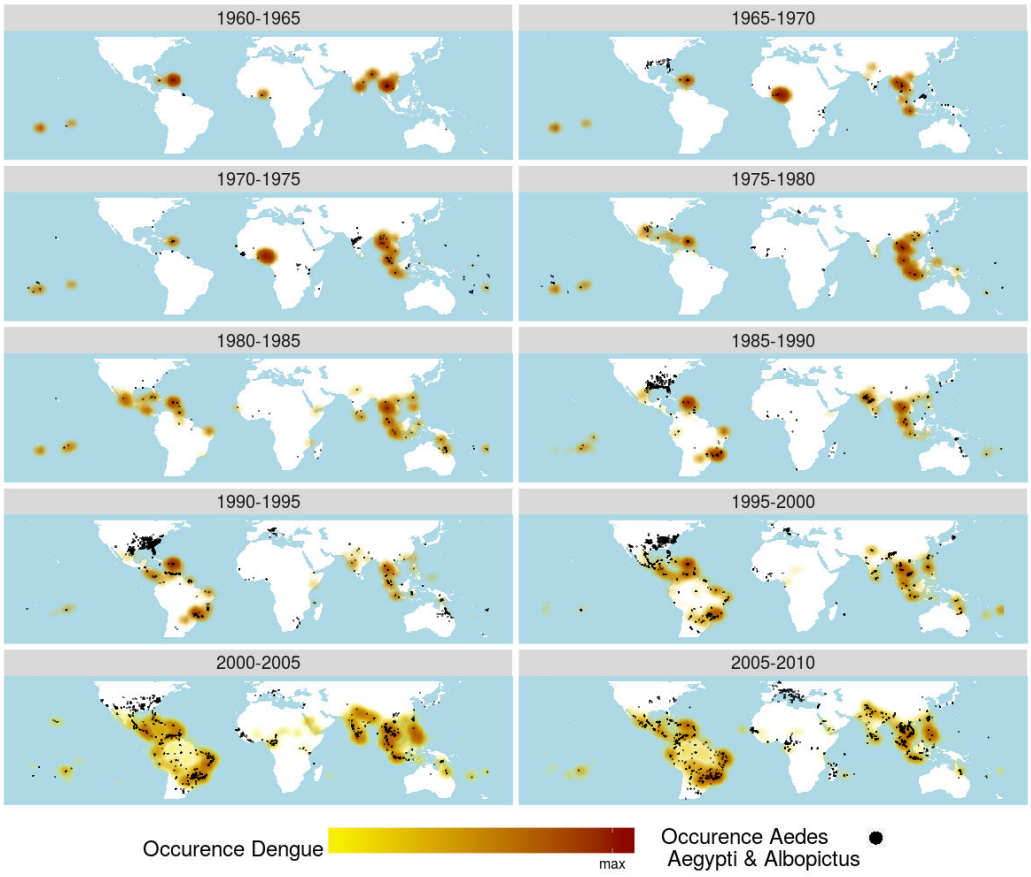


FIGURE 8 Évolution des occurrences de dengue et de moustiques *Aedes Aegypti* et *Albopictus* entre 1960 et 2010. D'après Messina *et al.* (2014) pour l'occurrence des cas de dengue et Kraemer *et al.* (2015a) pour l'occurrence des moustiques.

Évolution en Asie du Sud et du Sud-Est

La figure 9 est un zoom de la précédente, centrée sur l'Asie du Sud et du Sud-Est. Elle montre que la dengue est bien implantée à Bangkok depuis les années 1960, et que des cas sont maintenant enregistrés dans tout le pays. Nous pouvons observer pour l'Inde que des

cas de dengue sont enregistrés dans les années 1960-1970 sans que des vecteurs ne soient répertoriés. Aucune personne n'aurait été infectée par la maladie entre 1970 et 1980 alors que des moustiques sont repérés dans l'est et le centre du pays. Ceci peut s'expliquer soit par un système de surveillance déficient, soit par des mesures anti-vectorielles efficaces[31], soit par le fait que la maladie ou le moustique n'aient pas réussi à s'implanter durablement à l'époque. Il s'agit probablement d'un mélange de ces trois hypothèses. À partir des années 1980, des cas de dengue sont répertoriés à Delhi et dans le Kerala. Les grandes villes sont touchées (Delhi, Mumbai, Kolkata, Chennai, Bangalore) depuis les années 1990 et la dengue s'étend aujourd'hui dans la plaine du Gange, bassin de plus de 300 millions d'habitants (census 2011).

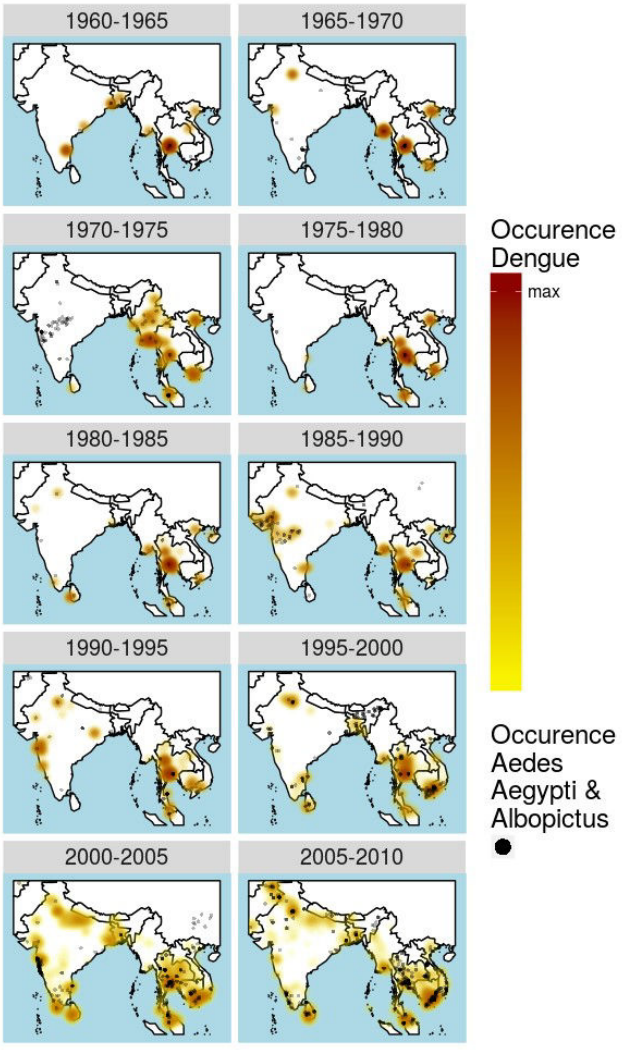


FIGURE 9 Évolution des occurrences de dengue et de moustiques *Aedes Aegypti* et *Albopictus* entre 1960 et 2010 en Asie. D'après Messina *et al.* (2014) pour l'occurrence des cas de dengue et Kraemer *et al.* (2015a) pour l'occurrence des moustiques.

31. L'Inde utilise du DDT depuis les années 1955.

Ces différentes cartes permettent de rendre compte d'un processus de dissémination et d'intensification des épidémies de dengue dans le monde, et particulièrement en Asie du Sud et du Sud-Est. Cette constante évolution va à l'encontre des observations faites depuis 1885 et la vaccination, où le nombre de personnes infectées et le nombre de décès liés à des maladies infectieuses suivaient une tendance globale à la baisse.

1.3 La dengue, une maladie émergente

En effet, les travaux de Pasteur et de Koch à la fin du XXe siècle avaient débouché sur la logique « une maladie, un agent pathogène », qu'il soit d'origine bactérienne, parasitaire, ou virale. Les approches d'éradication / réduction des agents pathogènes suivent alors une double stratégie. À l'échelle microbienne, l'étude des agents infectieux devrait rendre possible la création d'un vaccin. À l'échelle de l'environnement, les politiques hygiénistes devraient limiter les opportunités de prolifération de ces agents. Suite à ces politiques, les maladies infectieuses étaient désormais moins mortelles et moins fréquentes que les maladies non infectieuses (OMS) du moins dans les pays développés et ce jusqu'aux années 1970.

Mais l'apparition du sida à la fin des années 80 ou la recrudescence de la tuberculose et le développement de maladies vectorielles comme la dengue à partir des années 70 ont nécessité l'adoption de nouveaux paradigmes. Morse pose le concept de maladie émergente et ré-émergentes en 1990 (Morse et Schluedeberg, 1990). Il définit plus précisément ces termes en 1995, dans un papier paru dans le premier numéro du journal « Emerging Infectious Diseases ». Une maladie émergente ou ré-émergente est alors « une infection apparue récemment dans une population ou qui existait déjà, mais dont l'incidence ou l'emprise géographique augmentent rapidement »[32](Morse, 1995). Au regard des précédentes sections, la dengue rentre donc parfaitement dans cette définition. En reprenant ce concept, 335 évènements d'émergence ou réémergence de maladies infectieuses ont été dénombrés entre 1940 et 2004 (Jones *et al.*, 2008)

Le concept d'émergence, selon Lederberg[33], nécessite d'étendre la réflexion sur les facteurs et causes de l'apparition ou ré-apparition de maladies, plutôt que de rester dans la logique de classification des maladies en fonction de leurs agents pathogènes (Washer, 2010). Lederberg a ainsi établi une liste des principaux facteurs conduisant à l'émergence/ réémergence, constituant une nouvelle clé de lecture globale (Lederberg *et al.*, 1992) :

Les populations humaines et leurs comportements

La technologie et l'industrie

32. « Emerging infectious diseases can be defined as infections that have newly appeared in a population or have existed but are rapidly increasing in incidence or geographic range ». Elle reste aujourd'hui la définition choisie par l'OMS http://www.who. nt/top cs/emerg ng d seases/en/

33. Prix Nobel de physiologie en 1958.

Le développement économique et l'utilisation du sol

Les migrations internationales et le commerce

L'adaptation microbienne

Les déficits dans les mesures de santé publique

Parmi ces facteurs responsables de l'émergence, nous avons déjà évoqué pour la dengue le rôle des populations humaines et notamment dans la création de gîtes en eau favorable au moustique , la transformation de l'environnement par l'urbanisation, l'adaptation microbienne par le caractère évolutif des virus de la dengue à ARN. Il va de soi que le changement climatique, qui tend à une augmentation des températures moyennes globales serait également favorable à la prolifération des moustiques du genre *aedes* ces derniers étant adaptés à des températures plutôt élevées(Brady *et al.,* 2013). Nous ne traiterons pas davantage ces facteurs, déjà largement évoqués dans d'autres travaux (*e.g. (Maneerat, 2016 ; Misslin, 2017 ; Telle, 2011)*. Les sections suivantes vont se focaliser sur le rôle des mobilités humaines dans l'émergence et la propagation de la dengue, tant au niveau du vecteur ou des hôtes, qu'à l'échelle globale ou locale. Nous en profiterons pour revenir sur quelques définitions.

2 Mobilités et propagation des épidémies

2.1 Définition des mobilités spatiales

Le concept de mobilité revêt bien des significations dans les sciences humaines. Nous ne traiterons pas ici de la mobilité sociale chère aux sociologues. Nous nous focaliserons sur les mobilités spatiales, définies par (Brulhardt et Bassand, 1981) comme le concept qui : « recouvre tout déplacement de population dans l'espace physique, quelle que soit la durée et la distance du déplacement, les moyens utilisés, leurs causes et leurs conséquences ». Il peut également s'agir du « caractère de ce qui peut se mouvoir, changer de place, de position (s'oppose à immobilité)» (Brunet *et al.,* 1992) et appliqué ici aux moustiques et surtout aux Hommes. Kaufmann, (2012a) définit quant à lui la mobilité comme « l'intention, puis la réalisation d'un franchissement de l'espace géographique impliquant un changement social ». Nous ne sommes pas spécialement convaincus par la pertinence de la notion de « changement social », dans cette définition, notamment lors de déplacements de la vie de tous les jours, à l'échelle d'une journée.

Certains auteurs comme (Sheller et Urry, 2016, 2006) s'inscrivent dans le paradigme du « mobility turn », et proposent d'aborder les mobilités selon une approche globale, prenant en compte tout type de déplacements, même immatériels. D'autres comme (Cresswell, 2006) considèrent tous les mouvements effectués par les humains dans l'espace comme étant une

forme de mobilité, chacun de ces mouvements étant une construction sociale (Kaufmann, 2012a). Nous reviendrons peut-être dans quelques années sur ces positions, mais nous estimons pour l'instant que ces différentes approches englobantes n'apportent pas un cadre conceptuel pertinent pour nos travaux et complexifient inutilement la notion de mobilité. Nous en resterons donc à la définition de (Brulhardt et Bassand, 1981).

On distingue classiquement en sciences humaines quatre grands types de mobilité spatiale : les mobilités résidentielles, les migrations, les voyages et la mobilité quotidienne (Kaufmann *et al.,* 2004). Cette classification sous-entend des portées et des temporalités différentes. S'ajoute ensuite le caractère réversible ou non de ces déplacements, à savoir s'il y a un retour à l'état initial, soit le point de départ (Kaufmann *et al.,* 2004). Les différents types de déplacements / mobilités sont synthétisés dans le tableau 1, avec l'ajout des notions de cycle et de fréquence de retour.

Type de mobilité	Échelles	Retour	Fréquence	Durée à l'arrivée
Mobilité résidentielle	Villes /Région / Pays	non	Faible	De quelques semaines à une vie entière
Migration	Région / Pays	non / parfois	Faible	De quelques semaines à une vie entière
Voyage	Région / Pays	oui	Variable	Quelques jours à quelques mois
Mobilité quotidienne	Zone urbaine	oui	Quotidien, hebdomadaire	Quelques heures

Tableau 1 Différents types de mobilités, selon l'échelle, la fréquence de visite, la possibilité d'un retour et la durée à l'arrivée.

2.1.1 Déplacements sur de longues distances

Les notions de mobilités résidentielles et de migrations nous paraissent assez proches. Nous ne comprenons pas pourquoi Kaufmann limite la mobilité résidentielle à un déménagement dans un même bassin de vie (ou région) (Kaufmann *et al.,* 2004) alors qu'intuitivement l'échelle géographique n'aurait que peu d'importance. Néanmoins cette approche peut permettre de mieux étudier les phénomènes de gentrifications et de déclassement. Si nous considérons la mobilité résidentielle comme un changement de domicile à n'importe quelle échelle géographique (changement de quartier, de ville, de région ou de pays), la migration peut être alors considérée comme un sous-ensemble de la mobilité résidentielle, s'effectuant à des échelles géographiques plus petites (région, pays) et inclut également un changement culturel plus drastique (*e.g.* changement de région ou de pays, éclatement des liens familiaux et/ou sociaux, etc. (Brunet *et al.,* 1992)).

Le voyage s'effectue sur des distances variables, allant de la ville voisine au pays le plus éloigné et le retour au domicile s'effectue après une durée de quelques jours à quelques

semaines/mois, avec au moins un lieu visité. Il peut être lié au travail, ou aux loisirs (tourisme), aux pèlerinages religieux, etc. mais les différents cas de figure peuvent s'entremêler. Les causes et raisons poussant des personnes à migrer ou à voyager mobilisent des champs de recherche extrêmement larges et variés, que nous n'aborderons pas ici[34]. Nous nous intéresserons uniquement aux conséquences de tels déplacements dans la propagation de la dengue.

2.1.2 Les mobilités quotidiennes et urbaines

Les mobilités quotidiennes, sous leur forme la plus simple, sont associées au mouvement pendulaire des trajets domicile/travail. Elles sont évidements plus complexes, d'une part parce que beaucoup de personnes sont sans activités salariées (chômeurs, enfants, étudiants, femmes au foyer, retraités, etc.), et d'autre part parce que le travail ne constitue pas le seul motif de déplacement d'une personne notamment dans des sociétés où les loisirs et la consommation forment des pierres angulaires[35].

Une vision plus juste des mobilités quotidiennes serait donc se référer simplement aux déplacements de la vie quotidienne, à la mobilité du quotidien (Kaufmann *et al.*, 2004). Il peut s'agir effectivement d'aller à son travail, mais aussi d'aller chercher du pain, ou faire des courses sur le trajet du retour, ou encore aller parfois au cinéma ou au restaurant le soir, etc.

Nous considérons ici la mobilité urbaine comme tout déplacement effectué par un individu dans la zone urbaine dans laquelle il est domicilié, sans prise en compte des limites administratives. Ceci permet de prendre en compte des déplacements intra-urbains motivés par des événements rares, ne se produisant que quelques fois au cours d'une année ou d'une vie (par exemple aller voir un concert dans une grande salle de spectacle, assister à un mariage, ou visiter un lieu touristique, etc.). Car si certains de ces déplacements ne sont pas fréquents chez un individu, certains de ces lieux visités très occasionnellement à titre individuel peuvent néanmoins drainer un grand nombre de personnes (stade, salle de concert, lieux de célébrations de victoire de l'équipe nationale d'un sport quelconque, etc.). Dans le cadre d'une étude sur la propagation d'une maladie infectieuse en zone urbaine, il convient donc d'inclure ces déplacements occasionnels, car l'agrégation de ces deniers peut refléter une dynamique globale.

Nous ferons une distinction basée sur les fréquences de visite, séparant ainsi les mobilités quotidiennes routinières des déplacements plus exceptionnels. S'il n'existe pas de consensus pour définir une routine de mobilité (Meissonnier et Richer, 2015), nous considérons ici qu'elle est simplement constituée des différents lieux visités assez fréquemment où un individu exerce une activité.

34. Pour une étude détaillée sur les migrations en Inde du Sud, abordé sous un angle économiste, voir la thèse de Sebastien Michiels. Pour une geo anthropologie des voyageurs de longue durée au Maroc et en Thaïlande, voir la thèse de Brenda Le Bigot.

35. Sans parler de la flexibilité du travail lui même.

Kaufmann distingue les déplacements pendulaires des mobilités quotidiennes, car il définit ces derniers comme des déplacements de type navette domicile travail entre deux communes différentes (Kaufmann, 2012a). Nous n'effectuerons pas cette séparation qui n'a pas de sens lorsque l'on ne prend pas en compte les frontières administratives. Car tout dépend de la définition de ville et d'aires urbaines, qui sont parfois définies en prenant en compte les personnes d'autres régions qui y commutent (Cottineau *et al.,* 2018). De plus les travaux de Kaufmann sont surtout basés sur des cas d'études en Suisse, pays de 8,3 millions d'habitants (FSO, 2016), soit un peu moins que Bangkok et trois fois moins qu'à Delhi, et une distinction entre ces types de mobilité ne parait pas nécessaire dans des systèmes urbains déjà très complexes.

Mais un système urbain n'est pas composé uniquement de personnes y résidant à l'année. Ainsi, les populations de touristes fréquentent des types de lieux assez précis et participent à la dynamique urbaine pendant un laps de temps assez court (Cebeillac et Le Bigot, 2018). Nous pourrions les considérer ici comme un sous-ensemble de la population urbaine, appartenant temporairement au système urbain et n'effectuant que des déplacements exceptionnels, sur une durée relativement faible, avec une population variable au cours du temps, au gré des périodes touristiques.

Nous avons rapidement distingué différents types de mobilités, en fonction de l'emprise spatiale du déplacement (longues distances *versus* zones urbaines) de leur fréquence (routiniers vs occasionnels) et de leur retour (ou non) au point de départ (comme après un voyage touristique ou lors de déplacements pendulaires). Nous allons maintenant évaluer le rôle de ces différents types de mobilité dans la propagation de la dengue.

2.2 Les mobilités comme facteur de l'extension géographique de la dengue

Avant de prendre en compte les mobilités humaines dans les processus de diffusion des épidémies, il convient de rappeler que l'expansion des arboviroses est toujours précédée par la propagation globale de leur vecteur (Charrel *et al.,* 2014)Ainsi, l'extension géographique de la dengue s'opère si deux conditions *sine qua non* sont réunies :

- Le vecteur se propage et se développe dans une zone nouvelle où il est adapté aux conditions climatiques et où des hôtes sont présents pour assurer leur repas sanguin. En l'absence de concurrence ou de prédateurs, il peut alors se comporter comme une espèce invasive et coloniser un secteur donné.

- Le virus est ensuite importé dans cette zone par des vecteurs et/ou des hôtes infectés.

2.2.1 Propagation du vecteur

Dissémination « naturelle » du vecteur

La dissémination du moustique dans une zone où il est déjà implanté se fait naturellement, par capillarité, colonisant progressivement une région où les conditions environnementales sont propices à son développement.

Ae. aegypti est le principal vecteur de la dengue en milieu urbain. Sa propension à se déplacer et à s'éloigner de son gîte de naissance est un paramètre crucial dans sa colonisation de l'espace et *de facto* dans la propagation de la dengue. Si toutes les études s'accordent pour dire qu'*Ae. aegypti* ne se déplace pas au-delà d'un kilomètre, les ordres de grandeur diffèrent. Certaines études montrent qu'il ne se disperserait en moyenne qu'entre 20 et 50 mètres (González *et al.,* 2001 ; Maneerat et Daudé, 2016 ; Morlan et Hayes, 1958), d'autres estiment que les femelles peuvent s'éloigner d'entre 50 et 500 mètres de leur lieu de naissance (Muir et Kay, 1998 ; Reiter *et al.,* 1995 ; Trpis and Hausermann, 1986). Une équipe a montré en utilisant un marqueur radioactif, que les femelles *Aedes* pouvaient pondre à plus de 800 mètres de leur lieu d'origine (Honório *et al.,* 2003). Ces écarts assez importants dans l'estimation du potentiel de dispersion des moustiques peuvent s'expliquer par son comportement (recherche d'un repas sanguin ou d'un lieu de ponte), par les températures (dépense plus d'énergie s'il fait trop chaud) mais aussi par la structure urbaine locale, plus ou moins dense (Maneerat et Daudé, 2016). Ainsi, sans intervention de l'homme dans le transport du vecteur, la colonisation de nouvelles zones par le moustique serait relativement lente.

Transport du vecteur par l'Homme

Les moustiques du genre *Aedes* sont synanthropes[36], et peuvent être embarqués comme « passager clandestin », généralement sous forme d'œuf, étape de leur cycle durant laquelle ils sont les plus résistants, dans les nombreux modes de transports développés par l'Homme au cours de l'histoire (Gatrell, 2011 ; Kuno, 1995). Ils peuvent donc voyager sur de très grandes distances et essaimer dans des zones où les conditions environnementales permettent son développement (Gatrell, 2011 ; Gubler, 2006 ; Kuno, 1995 ; Wilder-Smith, 2014).

Aedes aegypti est selon toute vraisemblance originaire d'Afrique (Gubler, 2014). Il a été très certainement introduit en Asie et en Amérique par voie maritime (Kuno, 1995), lors notamment des grandes expéditions maritimes qui ont débuté à la fin du 15éme siècle pour l'Europe (Soper, 1967)La traite négrière aurait également pu favoriser l'implantation d'*Aedes aegypti* dans les régions tropicales du nouveau monde (Brown *et al.,* 2014). Sa propagation vers l'est s'est peut-être opérée plus tôt, avec l'ouverture de la route maritime de la soie entre la

36. du grec “syn”, qui signifie “avec”, et “anthropos”, “Homme”.

péninsule arabe et l'Asie il y a 2000 ans, ou encore avec les expéditions successives de l'Amiral Zeng He entre la Chine et la corne de l'Afrique au début du 15e siècle. En tout cas, à partir de la seconde moitié du 19e siècle, et surtout après les années 1930, on trouve un grand nombre d'*Aedes aegypti* tout le long des villes portuaires en Amérique (Soper, 1967), mais aussi en Asie du sud-est à partir des années 1960 (Gubler, 2014 ; Pant, 1974).

Le commerce de pneus usés par cargo entre les États-Unis et l'Amérique Centrale aurait propagé et entraîné une infestation d'*Adess aegypti* au Salvador en 1965 (Soper, 1970). Le même commerce entraîna l'importation d'*Aedes albopictus* originaire d'Asie (Gubler, 2014), vers les Amériques et le sud de l'Europe à la fin des années 1970 (Reiter, 1998 ; Reiter and Sprenger, 1987), associant de manière pérenne les gîtes larvaires que sont les pneus usés, à la pullulation des moustiques du genre *Aedes*. Nous pouvons aussi supposer que les moustiques peuvent également être transportés par le train ou les camions, même si peu d'études existent à ce sujet (Kuno, 1995). Le transport par voiture est probablement un élément important de la dissémination inter-urbaine du moustique (Eritja *et al.,* 2017 ; Kuno, 1995).

L'augmentation des liaisons aériennes longues distances depuis la seconde moitié du XXe siècle est aussi un facteur de propagation potentiel de moustiques (Gubler, 2014 ; Kuno, 1995 ; Wilder-Smith, 2014). Il a été montré très tôt, qu'*Aedes aegypti* pouvait survivre plus de 3 jours dans un avion effectuant de longues distances (Griffitts, 1933), et qu'ils survivaient à des vols non pressurisés entre l'Indonésie et le Japon (Misao et Ishihara, 1945). Cela dit, la politique de démoustication à bord des avions au décollage préconisé par l'OMS en 1985 ne s'avérera pas très efficace pour limiter la propagation des vecteurs (Kuno, 1995). Les nouvelles introductions de moustiques *Aedes* se feraient beaucoup plus par les voies maritimes[37] et terrestres qu'aériennes[38] (Kuno, 1995).

Une fois que le moustique est implanté durablement dans une zone géographique donnée, des épidémies sont susceptibles de se produire si le virus est à son tour introduit dans la zone.

2.2.2 Importation de la dengue

Auparavant, il est important de distinguer deux situations lorsqu'un cas de dengue est déclaré. Si la personne est contaminée localement, on parle de cas autochtone. Si une personne attrape la maladie dans une zone où le virus circule puis rentre dans une région où le virus est absent et est ensuite diagnostiquée, on parle de cas de dengue importé (Septfons *et al.,* 2015).

Lorsque des cas autochtones apparaissent dans un secteur donné, cela signifie que la maladie fut importée, très probablement par un humain (cas index), puisqu'elle a été transmise

37. 80 % du volume des marchandises exporté transite par la voie maritime (United Nations, 2015).

38. Nous pouvons également noter que de manière anecdotique, une personne ayant ramené des plantes de Martinique en Allemagne a vu se développer des *Aedes aegypti* chez lui, mais négatif à la dengue ou au Zika (Kampen *et al.*, 2016).

localement à des moustiques, qui la transmettent ensuite à d'autres personnes, entraînant des contaminations dites secondaires. Si les bonnes mesures de précaution ne sont pas prises, tous les facteurs sont potentiellement réunis vecteurs, virus et hôtes pour qu'une épidémie puisse débuter. Lorsque nous parlerons d'importation de la dengue dans une région géographique, nous sous-entendrons que des infections secondaires sont survenues, déclenchant des épidémies, tandis que les cas de dengue importés se réfèreront seulement aux personnes contaminées lors d'un court séjour dans une zone ou le virus circulait.

Les différentes souches de la dengue peuvent malgré tout être importées dans une nouvelle zone par des moustiques eux-mêmes déjà contaminés, mais les études sont assez rares à ce sujet. Cela dit, des moustiques infectés embarqués dans des bateaux peuvent contaminer l'équipage et les passagers (Vassal et Brochet, 1908), et la dengue aurait été introduite de cette manière dans le nouveau monde (Christie, 1881). Il apparaît aujourd'hui qu'à l'échelle de la planète, le déplacement des populations ou des personnes, notamment par avion, soit le facteur principal de l'importation de la dengue (ou de nouvelles souches) dans des zones jusque-là épargnées (Kuno, 1995 ; Wilder-Smith et Gubler, 2008). Ainsi, les sérotypes DENV1, DENV2, DENV4 auraient été introduits d'Asie vers les Amériques par des voyageurs, entre 1977 et 1981, et en 1994 pour DENV3 (Gubler, 2014).

L'apparition d'épidémies en contexte insulaire, comme ce fut le cas 14 fois en Polynésie française depuis le milieu des années 1940, avec généralement une alternance dans le type de souche de virus (Aubry *et al.,* 2017), illustre bien le rôle des déplacements humains. Par exemple, l'épidémie de 1988-1989 s'est propagée de manière chronologique d'îles en îles, et cette séquence de propagation est très liée avec le nombre de vols entre les îles (Chungue *et al.,* 1992). De même, à l'échelle nationale, les déplacements des populations entre différentes régions d'un pays semblent être le facteur déterminant de la propagation de la dengue à l'échelle nationale (Prevots, 1991 ; Teurlai *et al.,* 2012 ; Wesolowski *et al.,* 2015).

Cas de dengue importé en Europe

En avril 2013, le réseau de surveillance Geosentinel et l'Institut d'hygiène et de médecine Tropicale de Lisbonne ont enregistré 29 cas de dengue importés en Afrique du Sud, au Canada, en France, en Allemagne, au Portugal et en Israël, sans pour autant déclencher de cas secondaires. Toutes les personnes avaient comme point commun de s'être rendues à Luanda, en Angola, où sévissait une épidémie de dengue entre mars et mai 2013 (Schwartz et al., 2013). De la même manière, des voyageurs malades rentrant d'Afrique de l'Ouest ont permis de détecter l'expansion de DENV3 depuis 2009 dans cette région du monde (Franco *et al.,* 2010).

En France, le nombre de cas de dengue importé faisant l'objet d'une déclaration obligatoire pour 2010, 2013, 2014 et 2015 était respectivement de 596, 271, 201 et 167 et provenaient pour principalement d'Asie du Sud-Est (55 % en 2014 et 53 % en 2015), et

notamment de la Thaïlande (30 % en 2014) (Balestier *et al.,* 2016 ; Septfons *et al.,* 2015). En Europe, parmi les 242 personnes diagnostiquées positives à la dengue dans 12 sites du réseau Tropnet entre 2012 et 2014, 125 provenaient d'Asie, dont 56 de Thaïlande, 32 d'Indonésie et 27 d'Inde. Les 4 souches du virus furent détectées, avec une prédominance pour Denv1 (Neumayr *et al.,* 2017)

Cas de Dengue autochtone en Europe

La première grande épidémie de dengue survenue en Europe s'est déroulée en Grèce entre 1926 et 1928 (Wilder-Smith *et al.,* 2013). Le vecteur incriminé était alors *Aedes Aegypti*[39], et l'épidémie aurait fait environ 1 million de victimes (*ibid.)*. Il faudra attendre les années 2010 pour que de nouveaux cas autochtones soient à nouveau détectés en Europe. Entre octobre 2012 et février 2013, survient la première épidémie de dengue sur l'île de Madère où plus de 2000 résidents ont contracté la maladie (Tomasello et Schlagenhauf, 2013). Après une étude rétrospective passant par l'analyse des souches virales en circulation et des vols internationaux, le virus aurait très probablement importé du Venezuela (Wilder-Smith *et al.,* 2014). 78 cas ont ensuite été enregistrés sur le continent européen sans entraîner de cas secondaires (ECDC, 2013).

Dengue autochtone en France métropolitaine

Les premiers cas de dengue autochtones en France métropolitaine sont apparus en septembre 2010, où deux personnes (voisines) ont contracté la maladie à Nice sans avoir quitté récemment le pays (La Ruche et al., 2010). Quelques semaines plus tard, un autre cas est détecté en Croatie (Gjenero-Margan *et al.,* 2011), puis 17 autres après des recherches (relationnelles et sérologiques) approfondies (Tomasello et Schlagenhauf, 2013). Dans ces deux cas, le vecteur impliqué était *Aedes Albopictus.*

Un autre cas de dengue autochtone est enregistré dans les Bouches-du-Rhône en octobre 2013 (Marchand *et al.,* 2013). Un an plus tard, entre septembre et octobre 2014, deux cas d'infections secondaires sont détectés dans ce même département, et deux autres dans le Var (Giron *et al.,* 2015). Les premiers cas sont voisins, issus d'une même souche virale et proche dans le temps, tandis que les seconds ne sont pas liés, car issus de souches virales différentes et les infections sont non consécutives dans le temps[40] (Giron *et al.,* 2015)). Les derniers cas de dengue autochtones (à ce jour) sont survenus à Nîmes durant l'été 2015, où une personne infectée revenant de Polynésie a entraîné l'apparition de 7 cas secondaires dans son quartier (Succo *et al.,* 2016). Pour l'instant, les infections secondaires en France métropolitaine sont

39. Alors présent dans la région, puis éradiqué par la suite.

40. Quelques semaines plus tard, en octobre 2014, les premiers cas de transmission secondaires de chikungunya dans le département de l'Hérault sont enregistrés. La maladie a pu se propager dans le quartier de la personne qui avait contracté la maladie au Cameroun (Delisle *et al.,* 2015).

peu nombreuses et localisées autour du cas index, et le vecteur impliqué était très probablement *Aedes albopictus,* moins performant qu'*Aedes aegypti* pour transmettre la maladie.

Le rôle des voyageurs contaminés est donc déterminant dans l'importation de la maladie, car ils peuvent déclencher des épidémies dans des zones où *Aedes* est déjà implanté (Wichmann et Jelinek, 2004 ; Wilder-Smith et Gubler, 2008). Il serait donc pertinent de les inclure dans l'analyse et la modélisation des systèmes urbains ou la dengue est endémique. Ces derniers peuvent aussi permettre de révéler l'existence d'épidémies non officiellement déclarées lors de leur retour de pays où les systèmes de santé et de surveillance sont moins développés (Wilder-Smith, 2014). Les mobilités humaines sont grandement responsables de l'expansion géographique de la dengue à l'échelle mondiale et nationale via l'importation de nouvelles souches, et la réémergence de la dengue serait donc plus liée à des facteurs démographiques et écologiques, plutôt qu'a une adaptation du virus (Favier et al., 2005 ; Weaver et Vasilakis, 2009).

Cependant, qu'en est-il de la diffusion locale de la maladie, notamment en contexte urbain ? Les contributions des différents facteurs de propagation sont-elles les mêmes à résolution plus fine ?

2.2.3 Propagation de la dengue en milieu urbain

Alors que la capacité de dispersion naturelle d'*Aedes* réduit quasiment *de facto* son rôle dans la propagation du virus à l'échelle mondiale, cet aspect pourrait être plus important dans un contexte urbain. Cela dit, bien que pouvant théoriquement se déplacer sur une centaine de mètres, la présence de gîtes larvaires et les fortes densités de populations locales nécessaires aux repas sanguins limitent en général la dispersion d'*Aedes aegypti* à quelques dizaines de mètres (Maneerat et Daudé, 2016).

Le rôle des mobilités humaines et des axes de communications dans la propagation locale des épidémies de dengue a été observé depuis une quarantaine d'années, avec des exemples de diffusion radiale du centre-ville vers la périphérie (Kalra *et al.,* 1976 ; Wellmer, 1983). Ces observations sont confirmées par des travaux empiriques (Stoddard *et al.,* 2013, 2009), ou par modélisations mathématiques (Barmak *et al.,* 2011, 2016 ; Enduri et Jolad, 2014 ; Falcón-Lezama *et al.,* 2017 ; Perkins et al., 2014). De plus, la structure des relations sociales (familles, amis), et la fréquence de leur visite influencent la propagation de l'épidémie (Reiner *et al.,* 2014), tout comme les lieux habituellement fréquentés par les personnes contaminées (Perkins *et al.,* 2014).

Comme pour la propagation de la dengue à l'échelle mondiale ou nationale, la contamination locale en zone urbaine dépend fortement de la présence de moustiques infectés dans le quartier et la propagation de la dengue dans différentes zones de la ville serait

principalement due aux déplacements des populations (Stoddard *et al.*, 2013 ; Vazquez-Prokopec *et al.*, 2010).

Le risque de dissémination dépend donc des flux humains entre les différentes zones, qui peuvent d'ailleurs être d'échelles très variées (quartier, arrondissement, village, ville, région, pays, etc.) et du risque vectoriel dans chacune de ces zones. Si les flux inter-états sont en constante augmentation depuis les différentes révolutions des transports, ils sont surtout constitués de travailleurs, voyageurs et de migrants (Wilder-Smith, 2014; Wilder-Smith et Gubler, 2008). Ils sont donc proportionnellement nettement moins importants et fréquents (Bassand et Brulhardt, 1983) et plus facilement quantifiables (du moins pour les flux légaux) que les flux de mobilités intra-urbains surtout dans les mégapoles où des millions de navetteurs commutent quotidiennement.

Afin d'appréhender le rôle des mobilités humaines dans la propagation de la dengue en milieu urbain, il convient d'adopter un cadre théorique adapté. Les paradigmes employés influencent les angles d'attaques de toute recherche, et nous présentons ici celui qui nous semble être le plus adéquat pour considérer la diffusion de la dengue, à savoir la complexité.

3 La dengue en milieu urbain, sous l'angle de la complexité

3.1 Systèmes complexes et théorie de la complexité

La Science classique repose sur 3 piliers : « l'ordre », selon une vision mécaniste et déterministe du monde, la « séparabilité », soit la résolution des phénomènes ou des problèmes après leur décomposition en éléments simple et la « logique inductive-déductive-identitaire », qui cherche d'une part à créer des lois à partir d'observations discrètes, à obtenir une conclusion particulière à partir des lois générales et le rejet de la contradiction (Morin et Lemoigne, 1999). Ces piliers ont été ébranlés par l'apparition de trois nouvelles théories à partir des années 1940 : la théorie de l'information dans le traitement de l'incertitude, la cybernétique et l'idée de rétro-actions entre machines autonomes et la théorie des systèmes qui organise et hiérarchise les sous-systèmes, en sachant que « le tout est plus que la somme des parties » (Morin et Lemoigne, 1999). S'ajoutent ensuite les phénomènes d'auto-organisation et l'apparition (ou non) de points d'équilibres. Ainsi les phénomènes complexes sont régis par des interactions non-linéaires entre différentes entités à différents niveaux avec une part de stochasticité. De ce constat, Edgar Morin a formulé la théorie de la complexité[41], même s'il n'existe pas de définition universelle.

Alain Barrat[42] propose la définition suivante : *« C'est un système composé d'un grand nombre d'éléments interagissant sans coordination centrale, sans plan établi par un architecte,*

41. De *complexe*, du latin *cum plexus* signifiant tressé ensemble.
42. Centre de Physique Théorique de Marseille, dans le magasine "La Recherche".

et menant spontanément à l'émergence de "structures complexes", c'est-à-dire des structures stables avec des motifs présentant plusieurs échelles spatiales et temporelles » (Pajot, 2018).

En géographie, un système est défini comme « entité autonome par rapport à un environnement, organisée en structure stable (repérable dans la durée) et constituée d'éléments interdépendants, dont les interactions contribuent à maintenir la structure du système et à la faire évoluer » [43]. Un système complexe peut alors se définir par opposition à un système simple où son état à l'instant t ne peut être prédit, notamment à partir d'équations et de relations linéaires simples.

La complexité en géographie et en épidémiologie est donc basée sur l'analyse et l'étude des interactions entre les différentes entités d'un système (Eliot et Daudé, 2006) et « le fonctionnement d'un système complexe à une échelle d'observation donnée peut s'expliquer à partir de mécanismes simples interagissant à une échelle plus fine » (Perrier, 2014). En somme, un système complexe est organisé en plusieurs entités, réparties à différentes échelles (ou niveaux) et ayant des relations spécifiques de différentes natures (Eliot et Daudé, 2006).

Un des principes d'un système complexe largement accepté est l'émergence, soit l'apparition de phénomènes imprévisibles à partir de lois et d'interactions simples entre les objets à différents niveaux (Iltanen, 2012). Ainsi, une forme d'organisation peut paraître inattendue à l'échelle macro car résultant d'un grand nombre de processus se déroulant à une échelle plus fine (Eliot et Daudé, 2006).

3.2 La dengue, une maladie complexe

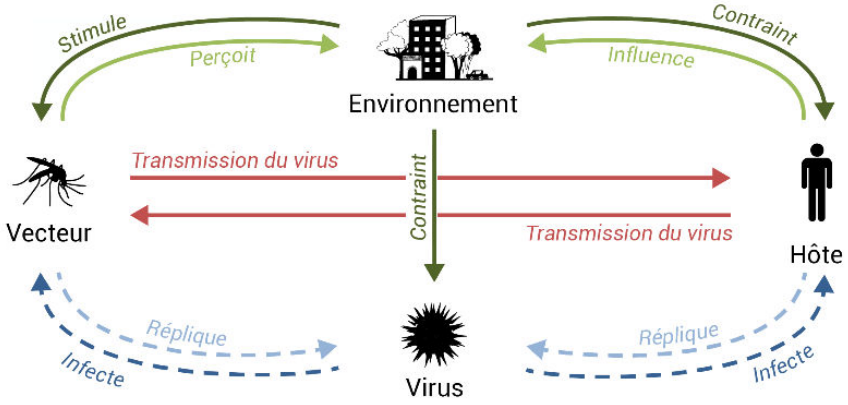


FIGURE 10 Le système de la dengue, ses composants et leurs interactions. Adapté de Daudé *et al.* (2015) par Renaud Misslin (2017).

Les quatre entités principales qui composent le système denguien, à savoir les hôtes, l'environnement, le vecteur et le virus, interagissent à différentes échelles selon des relations et

43. Denise Pumain, http://www.hypergeo.eu/sp p.php?art c e5

des rétro-actions complexes, non linéaires et de différentes natures [44] (figure 10), caractéristiques d'un système complexe (Daudé *et al.,* 2015).

Les hôtes agissent sur ce système à différents niveaux. Tout d'abord, ils influencent l'environnement en structurant l'espace et l'occupation et l'utilisation du sol via des planifications et des politiques urbaines ou des actions individuelles locales. Ils contribuent ou non à la création de gîtes larvaires, et la densité de l'habitat influence la température via les îlots de chaleurs. Cet environnement modifié va influencer l'Homme dans ses déplacements et dans sa pratique de l'espace et peut aussi contribuer à favoriser ou non le développement d'*Aedes.* L'Homme interagit également directement avec le moustique, que cela soit par l'intermédiaire d'une piqûre qui peut transmettre la dengue vers l'un des deux protagonistes, soit par la volonté d'éradication du nuisible par la destruction des gîtes vectorielles et la fumigation. La diversité des hôtes fait qu'ils réagissent de manière différente à la dengue, que cela soit d'un point de vue virologique où la génétique et l'âge expliqueraient que certaines personnes soient asymptomatiques, ou de la culture du risque qui pousse certaines personnes à prendre plus de précautions que d'autres.

Malgré ce schéma conceptuel et l'identification de nombreux facteurs, la méconnaissance et la difficulté de quantifier les différentes interactions dans des contextes locaux extrêmement variés entraîne la progression des épidémies de dengue (Gubler, 2011). La dengue est ainsi une maladie complexe, et chaque épidémie résulte d'un assemblage particulier de facteurs selon des niveaux d'interactions variables entre les différentes entités du système à une échelle et temporalité donnée (Daudé *et al.,* 2015 ; Telle, 2011). Il convient donc de traiter cette maladie avec humilité et les outils théoriques fournis par la théorie de la complexité. Chacune des interactions liant les entités du système mérite une exploration approfondie, et sa tentative de modélisation permettrait de mieux le comprendre (Daudé *et al.,* 2015), notamment en prenant en mieux en compte la mobilité urbaine des hôtes.

3.3 Le système des mobilités urbaines

Car les citadins évoluent dans des systèmes urbains avec discontinuités, qui sont aussi considérés comme des systèmes complexes (Batty, 2009a). En effet, un système urbain est composé d'un nombre impressionnant d'entités de natures et de degrés d'interactions extrêmement variés. De manière non exhaustive, nous pouvons citer l'organisation des transports en commun ou des quartiers, les politiques de la ville et de la planification urbaine, les pratiques individuelles et collectives, ou encore les différences de capital économique et culturel des personnes comme étant autant d'entités imbriquées de manière si intriquée que l'approche par la complexité paraît adaptée à l'analyse statique et temporelle de tels systèmes. Prenons

44. Pour une description détaillée de ces relations, voir (Daudé *et al.,* 2015).

un exemple particulier d'un quartier mal connecté au réseau de transport et où sa population se déplace moins que dans d'autres endroits de la ville. Ce constat (population plus sédentaire) est-il dû à la mauvaise connexion au réseau, ou est-ce que la connexion est moins bonne parce que la population du quartier y est plus sédentaire ? De même, quels sont les autres facteurs explicatifs de cette non-propension à changer de quartier ? Est-ce lié à un niveau de services et un bassin d'emploi suffisant dans le voisinage, ou *a contrario* à une absence de pouvoir d'achat et une forme de ségrégation sociale ? Ou est-ce dû à d'autres facteurs ? Une question simple soulève donc d'emblée de multiples d'hypothèses partielles et parfois antagonistes.

Dans un registre complètement différent, l'aspect complexe et d'auto-organisation des mobilités apparaît aussi dans les rues des villes de pays en voie de développement dont les politiques urbaines [45] autorisent les individus à prendre plus de liberté vis-à-vis du code de la route (s'il existe). Ceci entraîne une organisation du trafic routier suivant des règles tacites, comme lorsque chaque individu adapte sa conduite en fonction des personnes qui l'entourent et où le niveau de priorité semble proportionnel à la taille ou à la valeur du véhicule. Ces mêmes aspects peuvent s'observer de manière plus générale en cas d'incident sur une voie de communication ou suite à un accident de grande ampleur en milieu urbain. On voit alors que le système global précédemment stable devient chaotique puis se stabilise à nouveau sous une autre forme avant de reprendre en certains cas son état initial (Czura *et al.*, 2015).

Plus généralement, dans ce système complexe qu'est la ville, les mobilités quotidiennes d'une personne, décrites comme les endroits qu'elle fréquente régulièrement, forment une signature spatio-temporelle quasiment unique (de Montjoye *et al.*, 2013), et l'organisation des mobilités urbaines globales pourrait donc être perçue comme la somme de ces signatures individuelles. Les mobilités urbaines sont en général difficilement prévisibles sans données adaptées (Calabrese *et al.*, 2010 ; Song *et al.*, 2010b), et dépendent d'un grand nombre de facteurs individuels, parfois purement qualitatifs (Kaufmann et Jemelin, 2004), et le tout influencé par la structure urbaine et les pressions sociales.

(Brulhardt et Bassand, 1981) écrivaient d'ailleurs en 1981 dans leur article sur la mobilité spatiale en tant que système que : « *les divers flux de mobilité ne sont pas isolés les uns des autres, mais entretiennent entre eux des rapports de causalité, de complémentarité, de subsidiarité, de substitution, d'incompatibilité, etc. Ces divers mouvements sont articulés synchroniquement et diachroniquement, au point que la modification de l'un d'entre eux entraîne des changements dans les autres. Ces divers flux forment eux-mêmes un système* ». Au regard des avancées théoriques exposées précédemment, le système qu'ils décrivent est complexe.

45. Que certains pourraient qualifier de laxistes, d'autres de libertaires.

Synthèse

- La dengue est une maladie réémergente qui touche des millions de personnes chaque année, surtout les grandes villes des zones intertropicales. Elle est transmise par des moustiques du genre *Aedes*, notamment *Aedes aegypti,* particulièrement adaptés au monde urbain.

- Si les flux humains contribuent à la dissémination de la maladie à différentes échelles, leur étude en milieu urbain est tout à fait cruciale.

- Mais les mobilités urbaines forment un système complexe, au sein même du système complexe de la dengue.

Il convient donc d'essayer de prendre en compte les spécificités locales des zones urbaines où la dengue est endémique. Le chapitre suivant présentera les villes de Delhi (Inde) et Bangkok (Thaïlande), d'un point de vue descriptif, tant pour les différents types de quartiers qui les composent, les éventuelles inégalités sociales présentes, que sur la situation de la dengue.

Chapitre II: L'épidémie et le territoire : Le fardeau de la dengue à Delhi et Bangkok

Pour rappel, cette thèse s'inscrit dans le cadre de différents projets (ANR AEDESS & FP7 DENFREE) et dans la lignée d'autres travaux (Maneerat, 2016 ; Misslin, 2017 ; Telle, 2011), qui ont notamment pour zone d'étude Delhi et Bangkok, mégapoles où la dengue est endémique depuis des années. Des données sur les cas de dengues recensés dans chacune de ces villes ont de plus été fournies par des instituts locaux officiels, autorisant diverses analyses.

Ces mégapoles sont déstabilisantes pour le nouvel arrivant et/ou les non-initiés. Elles sont toutes deux très hétérogènes et discontinues en termes de densité de population, de niveaux socio-économiques et des inégalités associées, d'organisation des transports, des types de quartiers, etc. Des points communs existent entre Delhi et Bangkok. Par exemple un dynamisme qui se traduit par une forte attractivité sur les régions voisines, des quartiers très animés, la présence de bidonvilles dans des secteurs assez centraux, de grandes zones rurales aux périphéries des limites administratives, ou encore qu'il s'agisse de capitale de pays considérés comme émergents. Néanmoins nous estimons qu'aucun cadre théorique pertinent ne permet une comparaison rigoureuse de ces deux mégapoles fondamentalement différentes. Ce présent chapitre vise donc plus une présentation générale des composantes démographiques au regard de la dengue, qu'une étude comparée de ces deux villes si différentes.

En nous basant sur les quelques connaissances du terrain et des données issues de la littérature et des institutions officielles, nous décrirons la répartition des populations dans les différents types de quartiers d'habitations, leurs implications socio-économiques et leurs conséquences potentielles sur les mobilités quotidiennes des citadins. Nous mettrons ensuite ces éléments au regard des épidémies de dengue enregistrées ces dernières années. Ce travail sera d'abord effectué pour Delhi, puis pour Bangkok.

1 Delhi, une mégapole faite de ruptures. . .

Delhi, capitale de l'Inde, est une mégapole qui comptait 16,7 millions d'habitants d'après le recensement officiel de 2011, pour une superficie de 1 400 km². D'après des projections réalisées par les Nations-Unis, sa population serait estimée en 2016 à plus de 26,4 millions d'habitants, ce qui la positionne au second rang des métropoles les plus peuplées, derrière Tokyo et devant Shanghai [46].

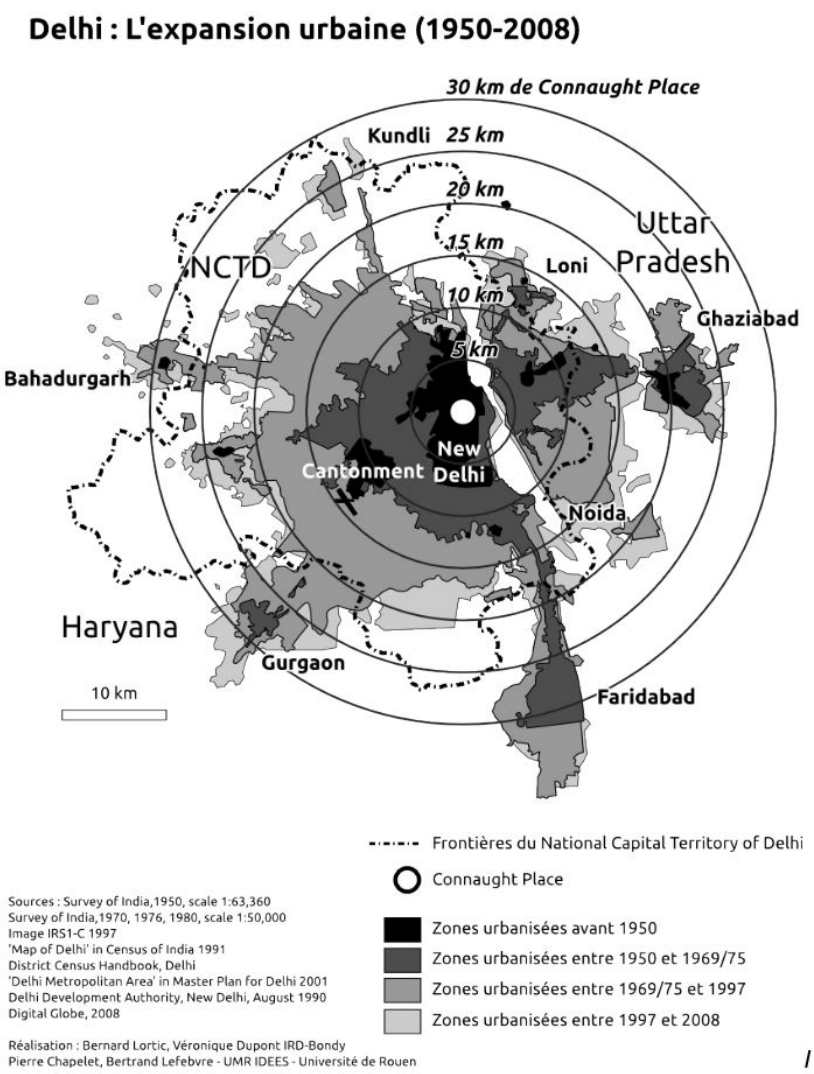


FIGURE 11 Évolution de la population à Delhi depuis 1950. Tiré de Lefevbre (2011)

Delhi ne comptait que trois secteurs urbanisés avant 1950, à savoir New et Old Delhi dans le centre, Delhi Cantonment au sud-ouest et une partie à l'est, sur la rive gauche de la

46. http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the worlds cities in 2016 data booklet.pdf

Yamuna (figure 11). Mais la ville a vu sa population augmenter de manière considérable depuis la partition de l'Inde de 1947, où des dizaines de milliers de déplacés originaires du Pakistan Occidental s'y sont installés. Depuis, la ville est restée très attractive puisqu'entre 1951 et 2001, le taux de croissance par décennie de la ville oscillait entre 47 et 53 %, ce qui est bien supérieur à un taux d'accroissement naturel, passant de 1,7 million d'habitants en 1951 à 13,8 millions en 2001 (Census 2011).

Ce taux baisse à 21 % entre 2001 et 2011, à la faveur du développement des villes satellites de Gurgaon, Noida, Faridabad, Ghaziabad, etc. qui captent l'essentiel des nouveaux arrivants (Economic survey of Delhi, 2016-2017). Néanmoins, une population qui augmente avec une telle tendance pendant 50 ans nécessite des plans d'urbanisme adaptés pour éviter les problèmes de logement. Mais malgré les différents MPD (Master Plan Developpement) pour les horizons 1962 et 2001, seulement 23,7 % de la population vivait dans des quartiers planifiés en 2000 et 46,6 % dans des bidonvilles (DUEIIP-2021). Ceci implique donc de grandes disparités économiques et sociales, tant sur l'accès aux logements et aux infrastructures de bases, que sur la qualité de ces derniers.

1.1 Une répartition inégale des populations

La figure 12 est le résultat d'une cartographie dasymétrique réalisée à partir du croisement entre des données de recensement de 2011 à l'échelle du ward (arrondissement) et une carte des densités de bâti dans des mailles de 250 m. Cette dernière fut obtenue à partir d'une estimation de l'occupation du sol réalisée sur des images SPOT 5, d'une résolution spatiale de 5 m. Les zones bâties ont ensuite été séparées des zones végétalisées, en eau et des sols nus, puis nettoyées avec une couche du réseau routier issu d'*OpenStreetMap*[47], et finalement agrégées dans un carroyage de cellules de 250 m de côté. La population des *wards* a ensuite été ventilée dans les mailles, en posant l'hypothèse que le nombre d'habitants est proportionnel à la densité des bâtiments d'un même arrondissement. Nous obtenons ainsi une répartition de la population nettement plus précise qu'à l'échelle du ward[48], ce qui nous permet de mieux apprécier les disparités locales de peuplement.

Nous pouvons noter quelques poches de très forte densité dans l'ouest et le nord-ouest, mais de manière générale, les districts de l'est et du nord-est et le quartier d'Old Delhi dans le centre affichent les plus grandes concentrations d'habitants. Il s'agit des quartiers historiques de la ville, à opposer à New Delhi, district planifié et créé par les colons anglais au début du XXe siècle et où la densité de population est la plus faible. Ceci s'explique par la politique hygiéniste des planificateurs britanniques et par le fait que ce quartier héberge la plupart des sièges des administrations et des ambassades, et où la population est principalement composée

47. www.*OpenStreetMap*.com – Une plateforme de cartographie collaborative (voir chapitres 8 et 11).

48. Nous pouvons souligner que nous ne faisons pas de distinction en fonction de l'usage du sol, ce qui sous entend que des populations peuvent être affectées à des zones qui ne sont pas dédiées aux habitations.

par des élites Delhiites qui vivent dans de grandes demeures entourées de jardins. Le sud est globalement peu densément peuplé, mais présente cependant des gradients très marqués, où des zones résidentielles peu denses côtoient des grands quartiers extrêmement peuplés [49].

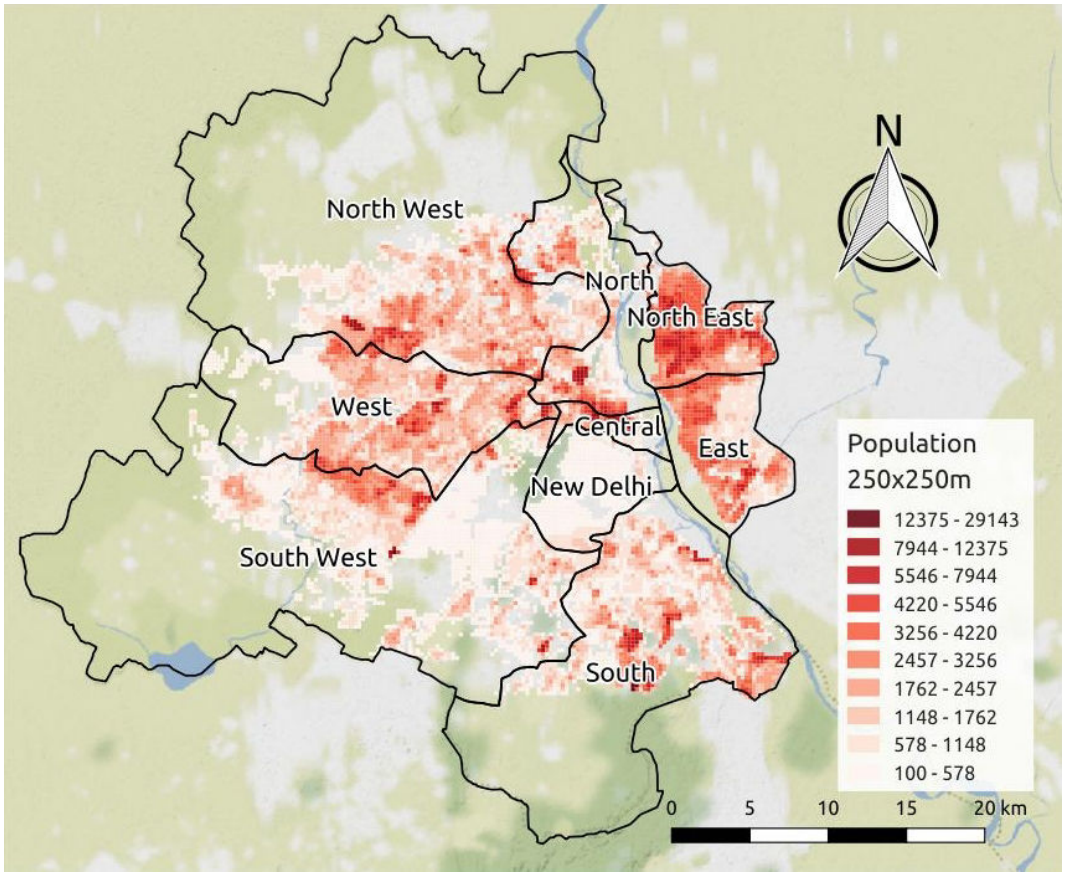


FIGURE 12 Répartition de la population à Delhi, selon un carroyage de maille de 250m de côté. Réalisé à partir d'images SPOT 5 et des données du recensement de 2011 à l'échelle de l'arrondissement (ward). Fond de carte *Google Terrain*.

1.2 Une grande variété de types de quartiers d'habitation

Ces grands contrastes dans la répartition des populations à Delhi sont liés aux sept types de quartiers, ou colonies, qui la compose, définis selon des oppositions entre planifié / non planifié, légal / illégal, formel / informel [50]. Pour une description précise de ces quartiers, nous vous renvoyons au papier de (Bhan, 2013), dont les prochains paragraphes ne sont qu'un court résumé [51] :

49. Comme entre le quartier aisé de Greater Kailash et le quartier densément peuplé de Govindpuri, ou encore entre Malviya Nagar / Saket et Mandagir.

50. Nous n'avons pas pu nous procurer de données ou de cartes fiables faisant ressortir toutes ces catégories.

51. Et pour comprendre les enjeux des politiques publiques en Inde, voir de manière plus générale les travaux du Centre for Policy Research, http://www.cpr nd a.org/.

- (1) *Les colonies planifiées* sont des quartiers légaux, construites suite à divers plans de développement (1962, 2001 et 2021) qui imposent des normes et des infrastructures de base telles que des routes assez larges, un approvisionnement en eau ou encore la présence de parcs, d'écoles et d'éclairage public *(MPD-2021).* Ces quartiers, notamment lorsqu'ils sont habités par des personnes relativement aisées, sont parfois fermés par des grilles et gardés, on parle alors de « gated community ». Les bâtiments construits selon le plan de 1962 sont limités à 4 niveaux, tandis que ceux construits plus tard, tel Dwarka au sud-est ou Rohini au nord-est sont plus denses et peuvent former des ensembles dépassant la dizaine d'étages.

- (2) *Les colonies non-autorisées* sont définies comme étant « construites sur des terrains non inclus dans les zones de développement du plan ou dans des zones du plan de développement non considérées comme des zones à usage résidentiel" [52] (Bhan, 2013).

- (3) Les *colonies régularisées.* Si après certains aménagements, les infrastructures de la colonie non autorisée se rapprochent des normes et spécifications des plans de développement, un processus de régularisation peut alors s'engager, qui ferait passer la colonie non-autorisée au statut de colonie régularisée et rendrait les titres de propriété des logements reconnus par la loi (Bhan, 2013 ; Sheikh et Banda, 2014).

- (4) *Les villages urbains* sont d'anciens villages qui ont été incorporés dans le plan de développement. On en retrouve un peu partout dans Delhi [53], leurs délimitations sont fixées, mais ils sont exemptés de normes sur le bâti et l'usage du sol, ce qui peut expliquer les fortes densités de population, les hauteurs parfois élevées des bâtiments ou encore l'étroitesse des rues (Bhan, 2013).

On retrouve également à Delhi des bidonvilles, répartis dans trois catégories :

- (5) *Les slums*, sont qualifiés par le Slum Areas Act de 1956 comme toute zone dont les bâtiments *« sont à tous égards impropres à l'habitation ou sont, en raison de la vétusté, du surpeuplement, d'arrangements et de conception défectueux, de l'étroitesse ou de la mauvaise disposition des rues, du manque de ventilation, de lumière ou d'installations sanitaires ou de toute combinaison de ces facteurs préjudiciables à la sécurité, la santé ou la moralité"* [54]*.* Ces zones d'habitations sont à la fois légales et non planifiées et leurs habitants sont protégés des expulsions.

52. *« unauthorised colony are built on land not included in the development area in the plan or one built on land within the developmental area but not yet zoned for residential use" (Bhan, 2013).*

53. Notamment Hauz Khas Village ou Hauz Rani.

54. *(a) are in any respect* unfit for human habitation ; or (b) are by reason of dilapidation, overcrowding, faulty arrangement and design of such buildings, narrowness or faulty arrangement of streets, lack of ventilation, light or sanitation facilities, or any combination of these factors, are detriment al to safety, health or morals – http ://lawmin.nic.in/ld/P ACT/1956/A1956 96.pdf

- (6) *Les jhuggi-jhopdi Cluster, ou JJ Cluste*r, sont également des zones relativement délabrées, similaires aux slums, mais qui n'en ont pas obtenu le statut et *a priori* ne l'obtiendront jamais et sont donc sans aucune protection. La distinction entre JJ cluster et colonie non-autorisées se fait principalement par le fait que l'on estime que les habitants de ces colonies non-autorisées sont propriétaires de leur terrain.

- (7) Les *resettlement colony,* sont les lieux d'accueil des habitants des JJ Clusters qui ont été expulsés de leur campement et qui ont été relogé ou du moins qui se sont vu attribuer un terrain dans des zones définies par le plan développement de la ville, et dont les titres de propriété ne sont pas transmissibles (Bhan, 2013).

À première vue tous ces différents types de quartiers peuvent impliquer des niveaux de ressources économiques très disparates, allant des personnes les plus pauvres et vulnérables dans ce qui s'apparente à des bidonvilles (Slums, JJ clusters et ressettlement camp) aux personnes ayant accès aux infrastructures de bases dans les colonies planifiées ou régularisées. Mais aussi une forme d'entre deux pour les habitants des colonies non-autorisées, car comme le souligne Bhan, (2013), les spectres socio-économiques dans ces colonies sont très larges, allant des travailleurs pauvres aux élites éduquées. Les villages urbains ne sont pas non plus systématiquement synonymes de pauvreté, car bien que la plupart d'entre eux soient peuplés par des personnes aux revenus modestes, certains profitent de l'absence de réglementation, notamment sur l'utilisation du sol, pour développer des activités économiques lucratives centrées par exemple sur les sorties ou la culture on pense ici à Hauz Khas Village ou à Shapur Jat, dans le sud de la ville. Ainsi, la catégorie du quartier de résidence n'est pas nécessairement un indicateur idéal de la classe sociale, du moins pour les villages urbains ou les colonies non-autorisées. Et parmi les colonies planifiées, certaines sont plus prisées que d'autres par les classes supérieures. Il convient donc de rechercher un autre indicateur pour rendre compte des disparités socio-économiques à une résolution spatiale fine.

1.3 De grandes disparités socio-économiques entre les quartiers

La taxe foncière, une synthèse d'indicateurs économiques

La taxe foncière dont s'acquittent les propriétaires à Delhi varie en fonction de la colonie. Elle est calculée en fonction de différents critères dont le prix locatif, l'âge et le type de la colonie (planifiée, régularisée, village urbain, etc.), mais aussi en fonction des différents services auxquels le quartier a accès, comme la qualité des infrastructures (route, réseau d'eau ou électrique), ou encore la proximité de marchés (Lefebvre, 2011 ; Telle, 2011). Chaque colonie ou quartier, se voit attribuer une note par des experts, qui varie entre A et H, pour des niveaux d'imposition

respectivement très élevés à très faible (voire inexistant)[55].

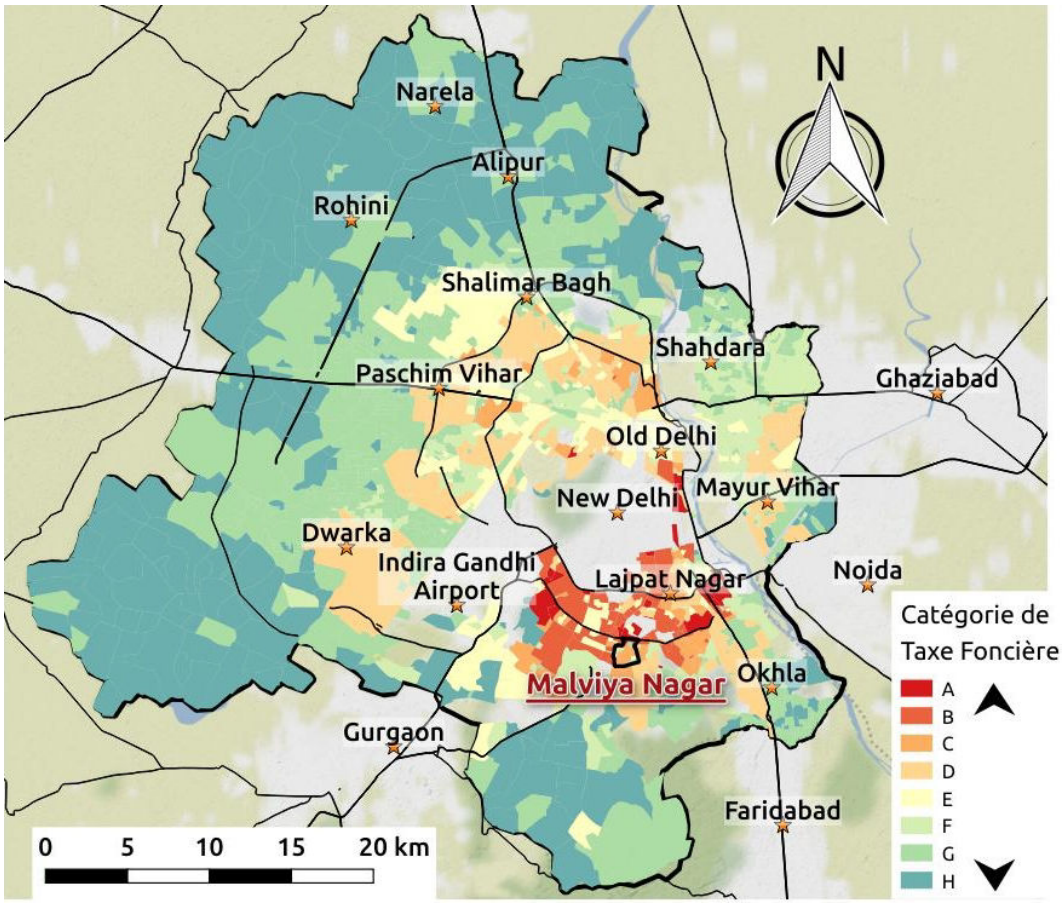


FIGURE 13 Répartition des colonies selon leur catégorie de taxe foncière à Delhi. Adapté de données fournies par Bertrand Lefebvre et Olivier Telle.

La figure 13 ci-dessus montre la répartition des différentes catégories de la taxe foncière dans la ville. Il en ressort que la plupart des quartiers où cette dernière est élevée (proche de A) se situent dans le sud. Le nord-est de la ville (zones de Shahdara, Yamuna Vihar et Seelampur) ne présente pas de quartier où les taux d'imposition sont élevés (entre H et E), tout comme les zones de la première couronne au nord de New Delhi mais dont le spectre des catégories est plus important (entre H et C). Les zones périphériques, principalement des villages ruraux en zone urbaine obtiennent la note la plus basse.

55. Ce système de notation entre A et H rappel vaguement la carte de la pauvreté à Londres, réalisée par Charles Booth entre 1886 et 1903, dans le cadre de ces travaux intitulés *Inquiry into Life and Labour in London*, où chaque rue était catégorisée en fonction des revenus et du niveau de précarité face à l'emploi de ces habitants (Topalov, 1991). – https ://booth.lse.ac.uk/map/14/ 0.1200/51.5000/100/0

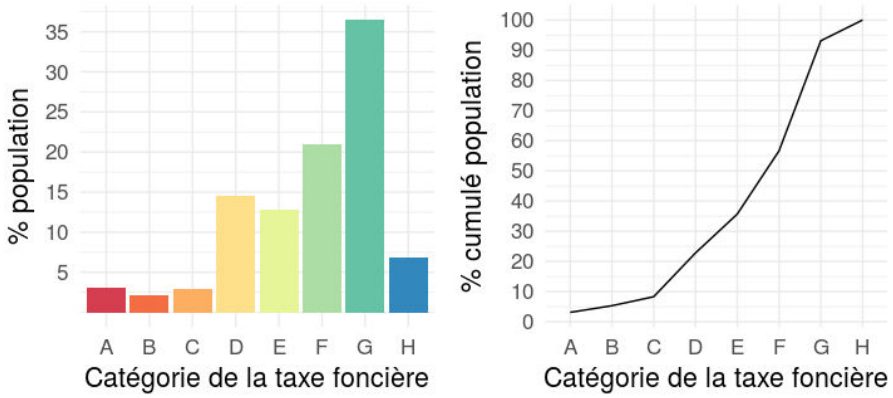


FIGURE 14 Répartition de la population en fonction de la taxe foncière. La figure de gauche représente les pourcentages de la population par catégorie de colonie. La figure de droite les pourcentages cumulés. Les quartiers de Delhi Cantonment et New Delhi, qui suivent un autre système de taxe foncière ont été considérés ici comme appartenant à la catégorie A.

Les frontières des différentes colonies ne recoupent pas les délimitations des sous-districts une colonie peut être à cheval sur plusieurs sous-districts, ou un sous-district peut contenir plusieurs colonies. Pour estimer la part de la population par catégories de taxe foncière (figure 14), nous avons agrégé les populations du carroyage par colonie et par catégories de taxe foncière. Il en ressort de fortes disparités, où la plupart de la population vit dans des zones où la taxe foncière est très faible, et moins de 10 % de la population dans des colonies où la taxe est supérieure à la catégorie D (A, B et C). Si nous partons du principe que la taxe foncière reflète globalement les niveaux de ressources économiques moyens des habitants des quartiers, cela suggère de très fortes inégalités spatiales dans la répartition des différentes classes sociales. Cela dit, les associations d'habitants de quartiers[56], notamment les plus aisés, ont tendance à faire du lobbying actif pour faire baisser la catégorie de la taxe foncière de leur quartier (Telle, 2011). Ceci peut expliquer par exemple pourquoi *Jangpura Extension*, une colonie de haut standing, aux prix locatifs prohibitifs, n'est classée qu'en catégorie D. Outre ces quelques limites, si nous revenons à la figure 13, le sud de Delhi présente une plus grande diversité dans les niveaux de taxe foncière. Cette zone parait donc appropriée pour approcher les différentiels de mobilités en fonction du capital économique, notamment par une enquête de terrain pilote que nous présenterons dans le chapitre 7.

Il convient maintenant d'étudier si ces grands contrastes dans les densités de population, les types de quartiers et les niveaux de richesses impliquent des potentiels de mobilités différents selon les zones géographiques.

56. « welfare association »

1.4 Des potentiels de mobilité différents en fonction des districts et très genrés

Si le recensement de 2011 n'a pas pour objet principal l'étude des mobilités individuelles, il contient tout de même des informations par district et par genre sur la distance parcourue pour se rendre à son lieu de travail (sauf secteur primaire). Le travail est ici considéré comme une activité, rémunérée ou non, productrice de richesse économique[57] (figure 15, ci-dessous). Les potentiels de mobilités sont très différents selon les secteurs de la ville et selon le genre. Il nous parait important de préciser que l'accès au travail est inégalement réparti entre les sexes, puisque seules 15,2 % des femmes entre 15 et 59 ans exercent une telle activité contre 75,7 % des hommes (census 2011, tables B-1). Mais ces chiffres sont très probablement sous-estimés puisque le travail informel est extrêmement présent en Inde, et notamment à Delhi[58]. Nous pouvons aussi souligner que 99,1 % des personnes dont l'activité principale est de s'occuper du foyer[59] sont des femmes (Census 2011, tables B-13) et que l'on compte dans la ville 867 femmes pour 1000 hommes.

La figure 15 montre d'abord que les femmes ont une plus grande propension à se déplacer sur de plus courtes distances (inférieure à 5 km) que les hommes pour exercer leur travail, et ce peu importe la zone de la ville. On peut ensuite observer que les différences sont assez marquées entre les districts, avec par exemple une relative parité à New Delhi, opposée à de forts niveaux de sédentarité chez les femmes de North et North-East Delhi. Nous pouvons noter que les niveaux de mobilités du South sont les plus proches de la tendance moyenne des districts de Delhi (NCT of Delhi).

Ces statistiques globales pointent une dichotomie des mobilités en fonction du genre. Les mobilités des femmes sont plutôt restreintes du fait de pressions et de constructions sociales très inégalitaires et tenaces bien qu'en voie d'atténuation chez les classes moyennes (Belliappa, 2013). À ces écarts dans les distances parcourues, s'ajoutent de grandes différences dans le choix des modes de transports (figure 16). Les femmes ont plus souvent tendance à se déplacer à pied, à prendre des transports collectifs (bus et train), ou privés (taxi, rickshaw, tempo), alors que les hommes prennent plus facilement des deux roues (motorisés ou non). Ceci est délicat à interpréter, mais cela suggère que les hommes sont plus libres dans leurs déplacements puisqu'ils conduisent leur propre véhicule[60]. Cela dit, nous pouvons noter que les femmes utilisent en proportion plus souvent la voiture que les hommes pour se rendre sur leur lieu de

57. « Work is defined as participation in any economically productive activity with or without compensation, wages or profit. Such participation may be physical and/or mental in nature. Work involves not only actual work but also includes effective supervision and direction of work. It even includes part time help or unpaid work on farm, family enterprise or in any other economic activity » (Indian census 2011). http://www.census nd a.gov. n/2011census/HLO/Metadata Census 2011.pdf.

58. Un département est d'ailleurs entièrement dédié à ces questions à l'Université Jawaharlal Nehru, le Centre for Informal Sector and Labour Studies – https://www.jnu.ac. n/sss/c s s

59. Ce qui n'est pas considéré comme un travail par le recensement indien.

60. De plus, le Sari n'est pas particulièrement adapté pour conduire un deux roues.

travail, peut-être parce que dans une ville perçue comme dangereuse pour les femmes[61], ce mode de transport privé procure un sentiment de sécurité[62] ?

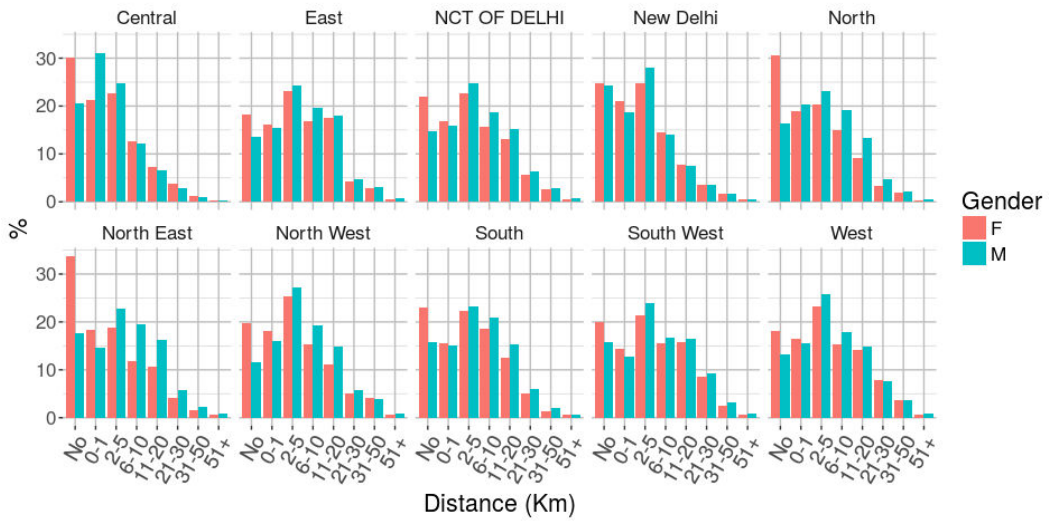


FIGURE 15 Distances parcourues à Delhi pour se rendre à son lieu de travail par district et par genre. D'après le recensement 2011, tables B-28.

Ce ressenti de l'espace, qui entraîne des catégorisations implicites de lieux en fonction de leur niveau de sécurité perçu, et les modes de transport utilisés par les femmes ont des répercussions sur les lieux fréquentés par ces dernières. Comme l'a montré (Borker, 2017) la représentation d'un trajet à parcourir entre le domicile et l'université influence le choix du lieu d'étude chez les jeunes indiennes de Delhi, qui ont tendance à favoriser le sentiment de sécurité au détriment du niveau de l'Université, ce qui tend à réduire leur mobilité sociale[63]. La dichotomie dans les potentiels de déplacements entre les hommes et les femmes est extrêmement marquée à Delhi. Ceci pourrait avoir un impact sur la dissémination du virus de la dengue, où les personnes les plus sédentaires contribueraient plus à une propagation locale (dans le voisinage) que les personnes les plus mobiles. De plus, se passer de la question du genre dans l'étude des mobilités urbaines dans la capitale indienne reviendrait à occulter des faits sociaux majeurs.

61. Nous ne ferons pas d'énumération des cas de violences faites aux femmes dans l'espace public à Delhi, mais nous rappellerons cependant le viol et la mort quelques jours plus tard d'une jeune femme en décembre 2012 et avait entraîné une indignation générale et largement marqué les esprits.http://www.theh ndu.com/news/nat ona /de h gangrape v ct m narrates the ta e of horror/art c e4230038.ece

62. Ce sentiment de sécurité pour les femmes en voiture a d'ailleurs été mis en avant en novembre 2017 par les opposants de la circulation alternée, alors que la ville connaissait des pics de pollutions démesurés, ce qui a poussé le gouvernement à abandonner la mesure – http://www.h ndustant mes.com/de h news/de h govt ca s off odd even scheme after ngt asks t to remove exempt ons/story 5KD3a4vrk6UAN0mqmUgooJ.htm

63. Et (Belliappa, 2013 ; Chatterjee, 1989) critiquent ces « politiques nationalistes d'éducation des femmes qui visent à faire d'elles des compagnes plus intelligentes et des mères responsables ».

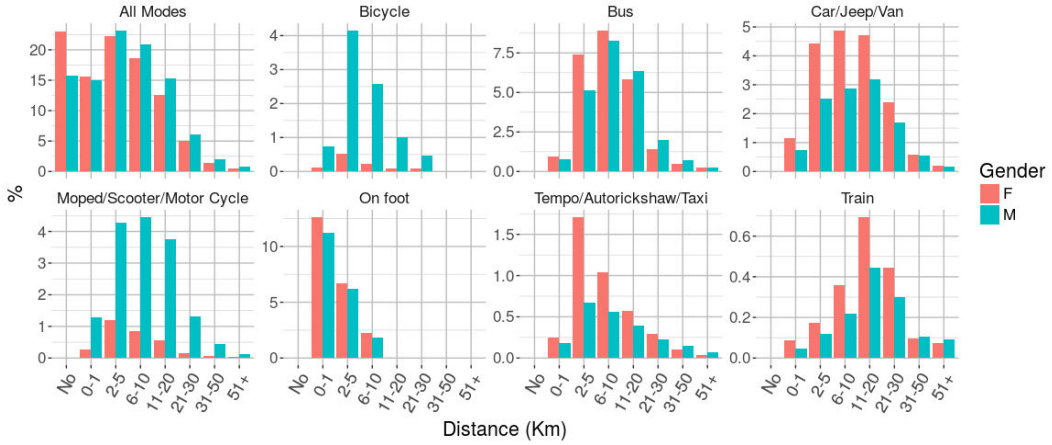


FIGURE 16 Distances parcourues pour se rendre au travail en fonction du genre et du type de transport, dans le sud de Delhi. D'après le recensement 2011, tables B-28.

1.5 Situation de la dengue à Delhi

Delhi est faite d'inégalités spatiales, qu'elles soient socio-économiques ou démographiques, et nous évaluerons ici si nous pouvons ajouter la dengue à cette liste. Cette dernière est considérée comme hyper-endémique à Delhi à partir de 2006 (Telle, 2015), c'est-à-dire que plusieurs sérotypes y circulent chaque année.

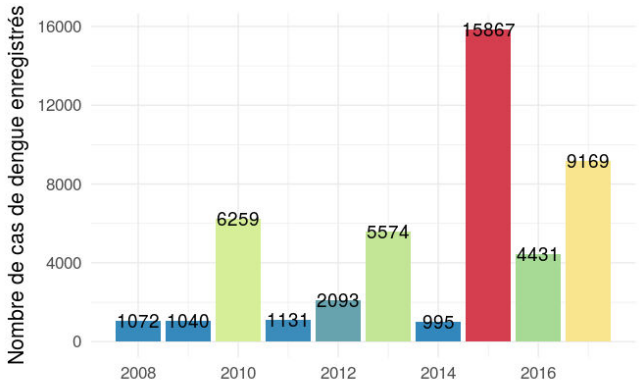


FIGURE 17 Nombre de cas de dengue enregistrés entre 2008 et 2017 à Delhi. source : National Vector Borne Disease Control Program http://nvbdcp.gov.in/den-cd.html

La figure 17 montre le nombre de cas enregistrés dans la capitale indienne depuis 2008. Il en ressort de grandes variations inter-annuelles, avec des années où la dengue est peu présente, avec moins de 3000 cas recensés (2008, 2009, 2011, 2014) et des années où l'épidémie est particulièrement active (2010, 2013, 2016, 2017 et surtout 2015 avec plus de 15 000 cas enregistrés). Mais il faut tout de même nuancer ces données officielles, car elles sous-estiment fortement les épidémies. En effet, seuls sont enregistrés les cas notifiés dans les hôpitaux sentinelles publics, et une grande partie de la population se fait dépister par son médecin

ou dans des cliniques ou hôpitaux privés (Daudé et Vaguet, 2015).

Néanmoins, si nous regardons la répartition spatiale des cas de dengue dans la ville au cours de trois années (2008, 2009, 2010, figure 18) nous observons des disparités dans les zones touchées (Telle, 2015). Un nombre de personnes équivalent a été contaminé par la dengue en 2008 et 2009 (~1000). Alors que le centre et l'est sont touchés ces deux années, nous pouvons noter que la dengue est très présente à l'ouest et quasiment absente du sud (zone plutôt aisée) la première année, alors qu'on observe l'inverse l'année suivante. En 2010, toutes les zones sont touchées, avec des concentrations plus importantes à l'est et au nord-ouest. Or, ces derniers secteurs sont les plus densément peuplés de la ville et furent systématiquement atteints lors de ces 3 années, d'où l'hypothèse d'un lien entre concentration de population et importance annuelle d'une épidémie.

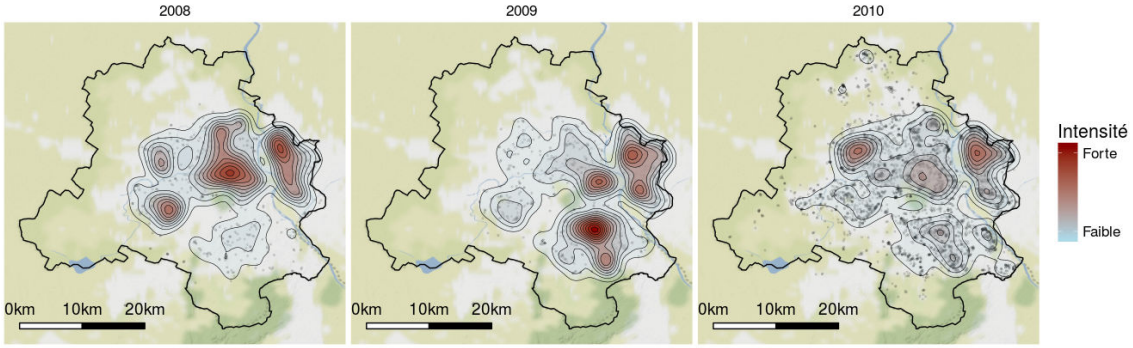


FIGURE 18 Intensité de la dengue à Delhi en 2008, 2009 et 2010, d'après des données du NIMR récupérées par Olivier Telle.

Si nous divisons notre carroyage de la population (figure 12) en décile avec des effectifs égaux le premier décile concentrant 10 % des mailles les moins peuplées, et le dernier décile 10 % des mailles les plus peuplées nous observons une loi de puissance positive (figure 19, gauche) ce qui confirme les inégalités dans les densités de peuplement, avec 10 % des zones peuplées qui concentrent plus de 40 % de la population. Si nous mettons maintenant en relation le nombre de cas de dengue dans chacun de ces déciles (figure 19), nous observons que le nombre de personnes contaminées augmente avec la densité de population, selon des lois de puissance.

La valeur de l'exposant de x, α, permet de décrire le niveau de concentration des cas de dengue selon la densité du quartier. Plus α est important, plus les taux d'incidence enregistrés ont été élevés dans les mailles les plus peuplées. Un α plus faible suggère une répartition un peu plus homogène des cas de dengue selon les densités de population. En 2008, la dengue touche surtout les quartiers les plus peuplés (et pauvres), et présente un α de 4,51. En 2009, l'épidémie atteint aussi le sud de la ville, zone à l'habitat relativement plus clairsemé, et l'α calculé est le plus faible (2,93). En 2010, alors que toute la ville est touchée, le coefficient de

x est entre celui de 2008 et 2009 (3,59). Cette approche permet donc de décrire de manière extrêmement synthétique le lien entre les concentrations de population et l'incidence de la dengue selon les années, en fournissant un indicateur global sur la répartition de l'épidémie selon des considérations de densités de peuplement.

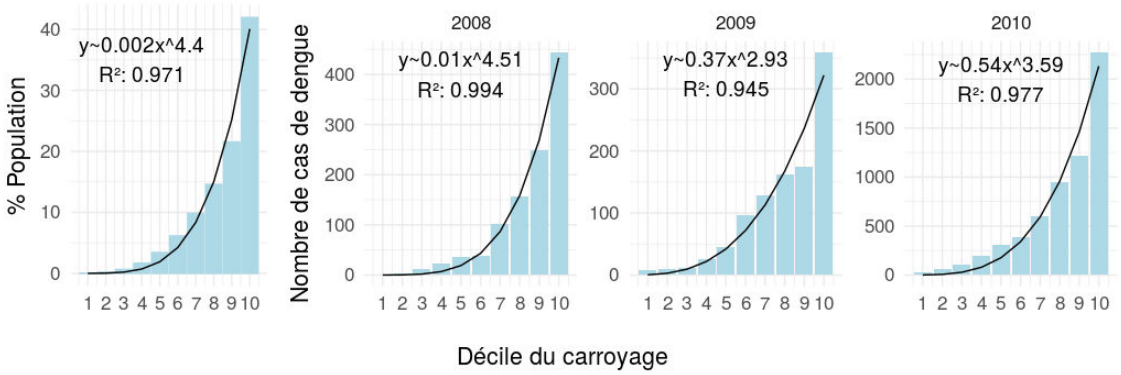


FIGURE 19 Population par décile de population (gauche) et nombre de cas de dengue par décile de population du carroyage (entre 2008 et 2010). Les courbes et coefficients sont obtenus en utilisant la fonction NLS (Non-linear Least Square) native dans R avec l'algorithme plinear.

Plus les zones sont densément peuplées, plus les taux d'incidences sont proportionnellement plus importants, ce qui va dans le sens des observations de (Telle, 2015), où les effets de proximité dans des secteurs densément peuplés induisent un plus grand risque de contamination locale. Par ailleurs, les mobilités humaines joueraient également un rôle majeur dans la dissémination, notamment dans les quartiers plus aisés et centraux de la ville (Telle *et al.*, 2016). Pour une lecture plus complète de la situation de la dengue à Delhi, nous vous renvoyons aux travaux de Somsakun (Maneerat, 2016 ; Telle, 2011 ; Telle *et al.*, 2016)et d'Olivier Telle (2011, 2015, 2016).

Nous avons présenté ici les différents types de quartiers de Delhi, ainsi que les tendances de déplacements selon les secteurs de la ville d'après des données du recensement. Les mobilités semblent bien jouer rôle dans la propagation de la dengue à Delhi, sans pour autant être pour l'instant précisément déterminé. Nous reprendrons dans la prochaine section la même approche descriptive que nous appliquerons à notre autre zone d'étude : Bangkok.

2 Les discontinuités de Bangkok

2.1 Généralités sur Bangkok

Bangkok, ou *Khrung Thep* en thaï [64], s'étend en zone intertropicale sur des longitudes comprises entre 100.2 et 100.9° et des latitudes entre 13.5 et 13.95°. C'est la capitale du royaume de Thaïlande, qualifié de « démocratie à 99,99 % » par le premier ministre Prayut Chan-o-Cha [65], arrivé au pouvoir en 2014 après un *n*-ieme coup d'état. D'un petit comptoir commercial établi sur les rives de la Chao Phraya au 15e siècle, la ville s'est développée progressivement pour devenir la capitale du roi Taksin en 1768. La métropole, bâtie autour de canaux (ou *khlongs)* connectés au fleuve a subi des transformations majeures après des politiques de modernisation et d'aménagement entamées dès la fin du 19e siècle, pour développer un caractère terrestre au détriment de son côté fluvial (Pichard-Bertaux, 2011 ; Shinawatra, 2012).

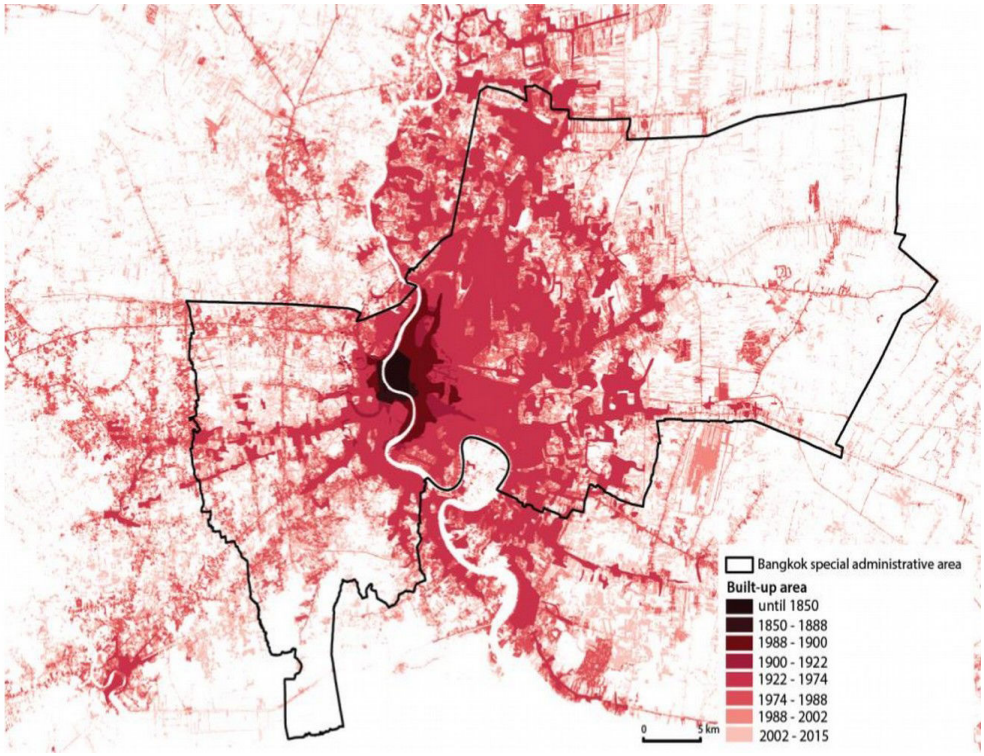


FIGURE 20 Évolution de la surface bâtie à Bangkok depuis 1850 (Heeckt *et al.*, 2017), d'après des données du Lincoln Institute.

La ville s'est lentement agrandie à partir de son centre historique, surtout sur la rive

64. de son nom complet : « Krungthepmahanakhon Amonrattanakosin Mahintharayutthaya Mahadilokphop Noppharatratchathaniburirom Udomratchaniwetmahasathan Amonphimanawatansathit Sakkathattiyawitsanukamprasit », qui signifie « *Ville des anges, grande ville, résidence du Bouddha d'émeraude, ville imprenable du dieu Indra, grande capitale du monde ciselée de neuf pierres précieuses, ville heureuse, généreuse dans l'énorme Palais Royal pareil à la demeure céleste, règne du dieu réincarné, ville dédiée à Indra et construite par Vishnukarn ». (Source : wikipedia).*

65. http://www. emonde.fr/as e pac fique/art c e/2015/03/30/en tha ande prayuth chan ocha homme qu n a me pas es journa stes 4606047 3216.htm

gauche du fleuve jusque dans les années 1940 (Heeckt *et al.,* 2017 et figure 20). La population de la ville est en effet passée d'un peu plus de 400 000 habitants en 1919, à environ 1,2 million en 1947 (figure 21, NSO). Après la seconde guerre mondiale, Bangkok observe un afflux de personnes provenant d'autres provinces et de Chine, qui entraîna « *Des zones d'habitat spontané et précaire (qui) se multiplièrent tandis que les villages englobés par la ville devenaient vite surpeuplés* » (Vongrattanatoh, 2011). Elle devient aussi simultanément une ville industrielle clé du pays, où se mélangent pauvres, classes moyennes et touristes (Askew, 2002).

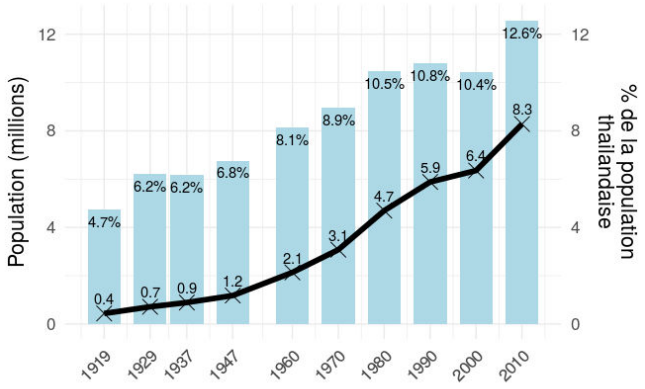


FIGURE 21 Évolution de la population à Bangkok entre 1919 et 2012. La courbe noire représente la part de la population de la capitale par rapport à la population nationale tandis que l'histogramme correspond à la population enregistrée par le recensement (NSO).

Bangkok comptait 3,1 millions d'habitants en 1970 puis 5,9 millions en 1990, ce qui représentait alors presque 11 % de la population nationale. La ville a aussi profité du boom économique en Asie du Sud-Est dans les années 80-90. S'y sont alors implantés de nombreux de sièges de multinationales (Heeckt *et al.,* 2017), et une très forte migration intra-nationale s'est opérée sur la période, du fait de la grande attractivité de la ville, malgré une politique d'encouragement de développement des villes secondaires (Browder *et al.,* 1995). Le grand nombre de travailleurs pauvres et la pression foncière peuvent en partie expliquer que 16,2 % de la population vivait alors dans l'un des 1744 bidonvilles[66] que comptait la ville en 1992 (Askew, 2002).

La ville s'est agrandie par ces franges, surtout par l'action de promoteurs privés et leurs projets de très grandes ampleurs (Browder *et al.,* 1995), aboutissant à la destruction et au déplacement des *slums* se trouvant sur les parcelles concernées. Mais la croissance de Bangkok s'est nettement ralentie au tournant des années 2000, après le *crash* économique de 1997 dont la Thaïlande a énormément souffert. La crise absorbée en quelques années, la croissance démographique de la ville a ensuite repris à nouveau pour atteindre 8.3 millions d'habitants en 2010 (National Statistical Office, 2010), dont 5,7 millions enregistrés auprès de

66. qualifiés de « congested area » par les autorités (Choiejit et Teungfung, 2005).

la municipalité (Bangkok GIS, 2018) et plus de 12 % de la population Thaï y était concentrée. En 2000 1,486,700 personnes habitaient toujours dans les bidonvilles [67] (United Nations Human Settlements Programme, 2008).

Bangkok s'étend aujourd'hui sur 1 568 km² soit environ 15 fois Paris, selon une limite administrative définie en 1972. Elle est divisée en 50 districts, les *Khets*, et en 169 sous-districts, les *Khwaengs* l'équivalent des arrondissements français.

Si l'on regarde Bangkok dans son ensemble, il s'agit d'une ville essentiellement monocentrique, où les zones les plus peuplées sont situées au centre de la ville, ainsi qu'à l'est de l'hyper-centre, et le long des grands axes de communication pour les zones périphériques (figure 22). Mais ce constat ne tient plus à une échelle plus fine, car on note l'existence de plusieurs centres qui se sont développés lors de différentes étapes de croissance de la ville (Clément-Charpentier, 2011).

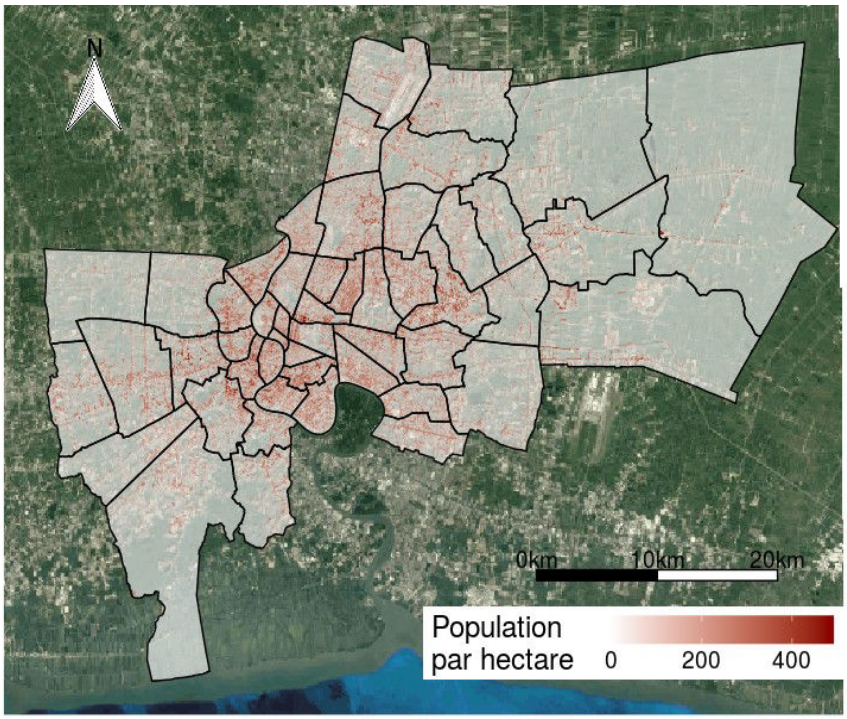


FIGURE 22 Répartition de la population à Bangkok, d'après Misslin et Daudé, (2016), agrégée sur une grille de maille de 100 m de côté.

Des quartiers d'habitation très contrastés

Si Bangkok a maintenant perdu une grande partie de son caractère fluvial, il n'en demeure pas moins qu'une grande partie de la ville, surtout en périphérie, est toujours parcourue de canaux même si une grande partie sert plus à l'évacuation des eaux usées ou à l'irrigation

67. Nous n'avons pas trouvé de chiffres plus récents.

qu'à la navigation. Au bord de ces canaux se trouvent généralement de petites maisons individuelles[68], dont le standing peut grandement varier dans un périmètre relativement réduit. C'est le cas par exemple dans le quartier de Bang Wa, à l'ouest de la ville où certaines rives sont constituées de bicoques en tôle, tandis que l'on peut trouver à quelques centaines de mètres des maisons plus cossues, au centre de jardins fleuris (figure 23).

FIGURE 23 Exemples de maisons au bord des *khlongs* dans le secteur de Bang Wa.

Nous pouvons aussi observer, toujours dans le secteur de Bang Wa, la présence de quartiers de densité moyenne, allant de petites résidences relativement huppées (figure 24.d) à des immeubles collectifs de 2 à 6 étages (figures 24.a, b, c, e). Il n'est pas rare de trouver des grilles au rez-de-chaussée des immeubles ou des maisons. S'il s'agit dans bien des cas de l'entrée du bâtiment, cet espace peut aussi être reconverti en un local dédié à une activité économique, allant du petit garage de réparation de moto (figure 24.e) au petit commerce d'alimentation (figure 24.b) ou encore à des ateliers d'imprimeries (figure 24.a). On parlera alors de "Shophouse".

68. Élément constitutif de la ville au 19e siècle, les maisons flottantes sont maintenant très rares dans les *khlongs* de Bangkok (Shinawatra, 2012).

FIGURE 24 Exemples d'habitations, entre Bang Wa et Wutthakat.

Mais Bangkok c'est aussi, contrairement à Delhi, un nombre très important d'immeubles de grande hauteur, dépassant allègrement les 30 étages et surtout localisés dans le centre-ville et le long des grands axes de communication. La figure 24.a est prise depuis le haut d'un immeuble de l'université de Chulalongkorn et est orientée vers le nord. Au premier plan nous pouvons distinguer le stade national, puis le *mall* MBK un peu sur la droite. Le second plan est rempli de gratte-ciels à vocation probablement commerciale. Les figures 25.b, c, d, e et f ont été prises depuis des stations de métro de la ligne verte qui dessert l'est et le sud-est de la ville. Si un grand nombre d'immeubles très élevés est visible entre Nana (figure 25.b) et après Thong Lo (figure 25.c), ce qui représente tout de même plus de 5 km de linéaire urbain, leur densité se réduit progressivement à mesure que l'on se rapproche du sud (xx.d, e). Dans la zone de Bearing, à la limite sud de Bangkok, nous pouvons noter la présence de petits immeubles assez

récents, ne dépassant pas les 7-8 étages (figure 25.f). Orientée vers le nord-ouest, la figure 25.f montre la proéminence des grands immeubles dans le centre-ville.

FIGURE 25 Quelques ensembles d'immeubles à Bangkok.

Néanmoins, si certains quartiers semblent être dominés par de grands ensembles architecturaux, il s'agit assez souvent d'un effet en trompe-l'œil, car derrière ou aux pieds des grands immeubles se trouvent très fréquemment des bâtiments assez bas, entre 1 et 4 étages, comme c'est le cas notamment dans le quartier de Sathorn (figure 26). Alors que l'artère principale sur laquelle passe également le métro aérien est jalonnée de grands immeubles d'habitation ou de bureaux (26.a), il suffit de prendre les petites rues parallèles (ou *soi*) pour arriver sur ce qui ressemble davantage à un petit village urbain, aux maisonnettes d'un ou deux étages, aux ruelles étroites, et où de nombreux rez-de-chaussée font souvent office de petits commerces (26.b, 26.c). Comme le résume assez bien le très bel article de *Pichard-Bertaux,*

(2011)[69] *: « Les perspectives verticales sont brisées par des rangées d'immeubles bas ; les éléments anciens côtoient les tours modernes ; les soi tortueux se jettent dans des avenues démesurées ».*

FIGURE 26 Bangkok et ses quartiers très contrastés. Exemple de Sathorn, où l'artère principale est jonchée de grandes tours, tandis que l'intérieur du quartier est composé d'un dédale de ruelles bordées de petites maisons mitoyennes.

La figure 27 ci-dessous reprend les informations du recensement sur les parts de chacun des types d'habitation évoqués précédemment. Tout d'abord, nous pouvons noter que les immeubles d'habitation et appartements sont surtout localisés au nord et à l'est du centre historique, ce dernier étant largement composé de *shophouse*. Les maisons individuelles se trouvent quant à elle principalement dans la grande périphérie, au cœur de zones agricoles. Les quartiers à forte densité de *shophouse*, qui mélangent résidentiel et commercial sont des lieux où les brassages des populations sont potentiellement plus importants. S'il peut s'agir selon les zones d'une clientèle plutôt locale, certains quartiers ont un rayonnement plus important, comme le quartier chinois de Yaowarat.

Une part non négligeable de la population vit aussi dans des quartiers d'habitat spontané, qui se situent généralement le long des canaux, des voies rapides, des chemins de fer et en cœur d'îlot (Gerbeaud, 2011), information qui n'apparaît pas sur la figure 22. L'un des plus peuplés

69. Qui d'ailleurs décrit bien mieux Bangkok que nous pourrions le faire.

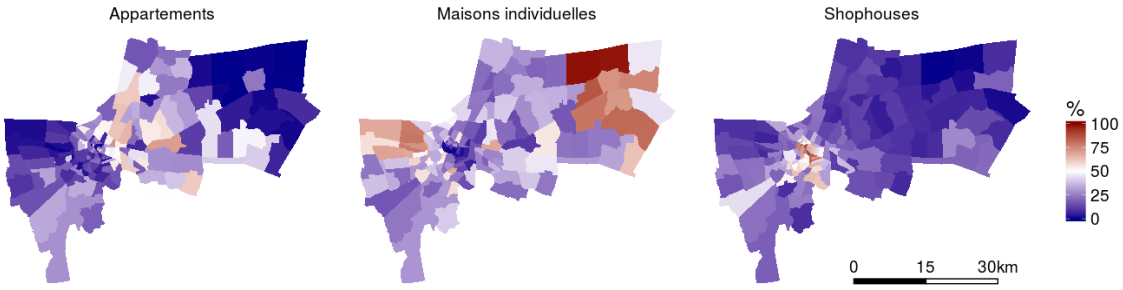


FIGURE 27 Répartition des types de bâti dans Bangkok - Source : NSO 2010.

d'entre eux serait celui situé à Khlong Toey[70], soit dans le grand centre de Bangkok (à 2 km au sud-est de Sathorn).

Ces grandes différences dans les types de bâtiment d'habitation, allant d'immeubles récents avec piscine dans les quartiers centraux, aux petites bicoques le long des *khlongs* périphériques sont susceptibles de traduire des inégalités sociales assez marquées.

Les inégalités sociales à Bangkok

Contrairement à Delhi, nous n'avons pu trouver d'indicateurs (*e.g.* la taxe foncière) pour estimer le niveau socio-économique moyen par arrondissement. Néanmoins, avec un PIB par habitant estimé[71] à 28 000 $ en 2010, la capitale Thaï est globalement plus riche que son homologue indienne seulement 7 000 $ par habitant (McKinsey & Company, 2018). Le revenu mensuel moyen par foyer est quasiment deux fois supérieur au revenu national, avec 45 572 baht à Bangkok contre 26 915[72] en Thaïlande pour l'année 2015 (Office of the National Economic and Social Development Board, 2017a). La contribution de Bangkok au PIB national est restée stable sur 20 ans, avec une part d'environ 30 % entre 1995 et 2011 (Office of the National Economic and Social Development Board, 2011).

Le niveau de pauvreté est d'ailleurs étonnamment bas dans la capitale, puisqu'il fluctue entre 1 et 2 % entre 2012 et 2016, contre 7,2 et 12,64 % pour l'ensemble du pays (Office of the National Economic and Social Development Board, 2017b). L'indice de Gini qui permet d'apprécier les niveaux d'inégalités est plus faible à Bangkok que dans l'ensemble du pays, avec des valeurs de 0.45 et 0.39 en 2013[73] et 2015, contre 0.46 et 0.44 dans le royaume Thaï aux mêmes périodes (Office of the National Economic and Social Development Board, 2017c).

70. http://www.borgenmagaz ne.com/bangkoks k ong toey s um/ (mais cet article ne cite pas suffisamment ses sources).
71. Par une entreprise privée qui ne détaille pas sa méthodologie.
72. Soit 1 182 € pour Bangkok et 698 pour le reste du pays, avec un taux de change à 1 euro pour 38,5 baht.
73. Revenant alors au niveau de 1988, où l'indice était de 0,388 (Falkus, 1999).

Une ville sans pauvres ? Vraiment ?

Si ces statistiques dressent un portrait plutôt flatteur de Bangkok (une forte contribution au PIB, de hauts salaires, une pauvreté quasiment inexistante et des inégalités moins marquées qu'à l'échelle nationale) il faut tout de même remettre ces chiffres dans un contexte plus large, où une grande partie de l'économie Thaï se fait de manière informelle. Certains auteurs estiment que la « shadow economy », ou économie de l'ombre, c'est-à-dire les transactions non déclarées constituent 39 % de l'économie du pays (Schneider, 2017). Si nous nous référons aux observations du terrain, il est vrai que Bangkok compte un nombre très important de vendeurs ambulants (Hazan, 2017), et les petits commerces familiaux, aux rez-de-chaussée des maisons, sont relativement généralisés dans certains secteurs de la ville. De plus, un grand nombre de migrants peu qualifiés provenant des provinces voisines ou des états voisins (Laos, Cambodge et Birmanie) travaillent également à Bangkok, et ne sont pas nécessairement pris en compte dans les statistiques (Heeckt *et al.,* 2017).

Si toutes les personnes vivant dans des quartiers informels d'habitats spontanés ne sont pas nécessairement pauvres, nous pouvons toutefois rappeler qu'elles étaient presque 1,5 million en l'an 2000 (United Nations Human Settlements Programme, 2008). Un taux de pauvreté qui fluctue entre 1 et 2 % représente donc entre 85 000 et 170 000 personnes en 2010. En admettant que la population dans les *slums* soit restée stable malgré la croissance de la ville ou des politiques de destruction / réhabilitation de ces quartiers, et que tous les pauvres vivent dans ces derniers, cela signifierait que seulement 5 à 10 % des habitants des *slums* vivent sous le seuil de pauvreté. Si de nos courts séjours à Bangkok (3 mois) et à Delhi (2,5 ans) ressort l'impression d'une pauvreté ambiante nettement moins perceptible dans la capitale thaïe, ces chiffres officiels de 1 à 2 % de personnes sous le seuil de pauvreté nous laissent perplexes. Mais nous pouvons noter que ce seuil de pauvreté est très bas, car il concerne les personnes dont les revenus sont inférieurs à 83 $/mois (Ratanawaraha et Chalermpong, 2016), alors que le revenu mensuel moyen à Bangkok est de 1 182 €. Ainsi, bien que le coût de la vie est toutefois beaucoup plus élevé en ville qu'à la campagne, très peu de personnes sont au-dessous de ce seuil à Bangkok, alors qu'un grand nombre de travailleurs perçoivent pourtant un salaire très faible (*ibid)*[74].

Comme le décrit Clément-Charpentier, (2011), Bangkok est *« une ville hétérogène, où les gratte-ciels des quartiers d'affaires côtoient des quartiers résidentiels ou des bidonvilles, avec des classes sociales et des modes de vie très diversifiés ».* Il peut donc en résulter des potentiels de mobilités différents selon les quartiers et les classes sociales.

74. Et même si ce papier repose sur études conduites dans des quartiers considérés comme les plus pauvres de la ville, ils ne remettent en cause ni ce seuil, ni la méthodologie.

2.2 Les tendances de mobilité à Bangkok, d'après la littérature

2.2.1 Tendances globales

Un bon indicateur pour estimer les tendances de mobilité globales et les navettes domicile-travail revient à cartographier les densités de population au domicile (le soir) et le jour (figure 28). La population de la ville est surtout concentrée dans le grand centre, et le long des axes de communication pour les zones les plus périphériques, avec quelques zones satellites localement plus densément peuplées (figure 28.a). La population en journée (figure 28.b) est encore plus concentrée dans le centre de la ville.

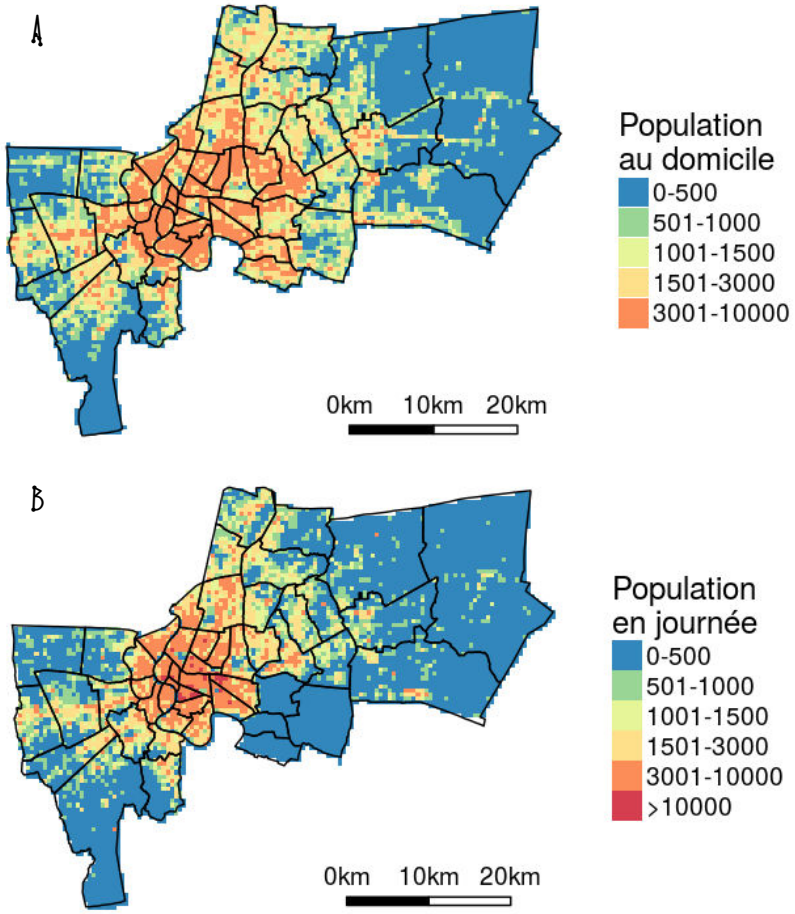


FIGURE 28 Répartition de la population à Bangkok au domicile (a) et en journée (b) selon une grille de 500 m. La population au domicile est issue d'une agrégation des résultats obtenus par Misslin et Daudé,(2016), d'après une cartographie dasymétrique réalisée à partir des données du recensement 2010 et d'une estimation des zones de bâti d'après des images Landsat 8. La population en journée, fournie par Bertrand Lefebvre, a été estimée par l'*Asian Institute of Technology*, en 2009, d'après des études de terrain, selon une méthodologie qui nous est inconnue.

Si cela suggère déjà des mouvements pendulaires de type « centre-périphérie » assez marqués, il convient d'apporter quelques nuances. Car le grand centre de Bangkok reçoit effectivement des travailleurs venant de toute la ville, tandis que les quelques zones de densité de population de rang inférieur situées en périphérie accueillent aussi des travailleurs, mais principalement originaires des zones voisines (Vichiensan, 2009). Il s'agit là d'un résultat assez attendu au regard de la figure 28, et du fait que Vichiensan utilise un modèle gravitaire dans son étude. Si un grand nombre de personnes travaillent dans le district de leur domicile (Choiejit et Teungfung, 2005), cette structure de la ville, au polycentrisme embryonnaire (Choiejit et Teungfung, 2005 ; Vichiensan, 2007) implique néanmoins un nombre important de flux entre le centre et la périphérie, qui vont impacter la circulation dans la ville.

N'importe qui de passage à Bangkok peut se rendre compte que la ville est congestionnée aux heures de pointe. Les bouchons sont récurrents entre 6 h et 9 h et entre 16 h et 19 h (Supatn, 2011), et pour plusieurs raisons. Tout d'abord, plus de 4,24 millions de voitures, 3,5 millions de motos et 1,3 million de vans ou pick-up y sont enregistrés, soit plus de 9 millions de véhicules privés (DLT, 2018), pour environ autant d'habitants. Le récent réseau de métro permet de traverser la ville bien plus rapidement qu'en voiture. Mais il n'est probablement pas assez développé et optimisé et les transports publics sont utilisés quotidiennement par moins d'un million de personnes (Chalermpong et Ratanawaraha, 2015). Dès lors, cette situation peut se traduire par des routes congestionnées et des temps de transports relativement longs pour effectuer une activité donnée.

2.2.2 Temps de transport selon les genres et les âges

Si le recensement indien fourni des indications sur les distances parcourues par genre et par zone de Delhi, le *time use survey* de Bangkok permet d'avoir des informations générales et agrégées sur le temps consacré en moyenne et quotidiennement à une activité (NSO, 2009). Y figure notamment le temps de trajet aller-retour pour effectuer chacune des 15 activités recensées. Compte tenu de la situation des transports dans la ville, raisonner en termes de durée parait plus pertinent qu'en termes de distance. Parmi les 15 catégories d'activités assez larges que recense le *time use survey*, nous en présenterons ici 8 :

- Le travail dans le secteur formel
- Travail pour le foyer, soit les personnes qui participent à l'économie du foyer sans nécessairement être enregistrées ou rémunérées
- S'occuper des enfants
- Étudier

- Fournir des services à sa communauté
- Participer à des événements culturels ou sportifs
- S'adonner à des *hobbies*, définis comme des passe-temps
- Pratiquer une activité sportive

La figure 29 présente deux types d'informations. Tout d'abord les effectifs par genre se déplaçant pour effectuer une des 8 activités (29.a) ainsi que les temps de transport moyen par tranche d'âge par activité (29.b). Elle révèle déjà des éléments intéressants en termes de répartition des activités et de genre. Par exemple, le nombre de personnes travaillant est équivalent selon les genres, et les femmes auraient même légèrement plus accès à l'éducation que les hommes. Néanmoins, les activités consistant à accompagner des enfants ou à fournir des services à sa communauté sont largement plus effectuées par des femmes. On compte en revanche plus d'hommes qui s'adonnent à des *hobbies*, pratiquent une activité sportive et se rendent à des événements culturels ou sportifs.

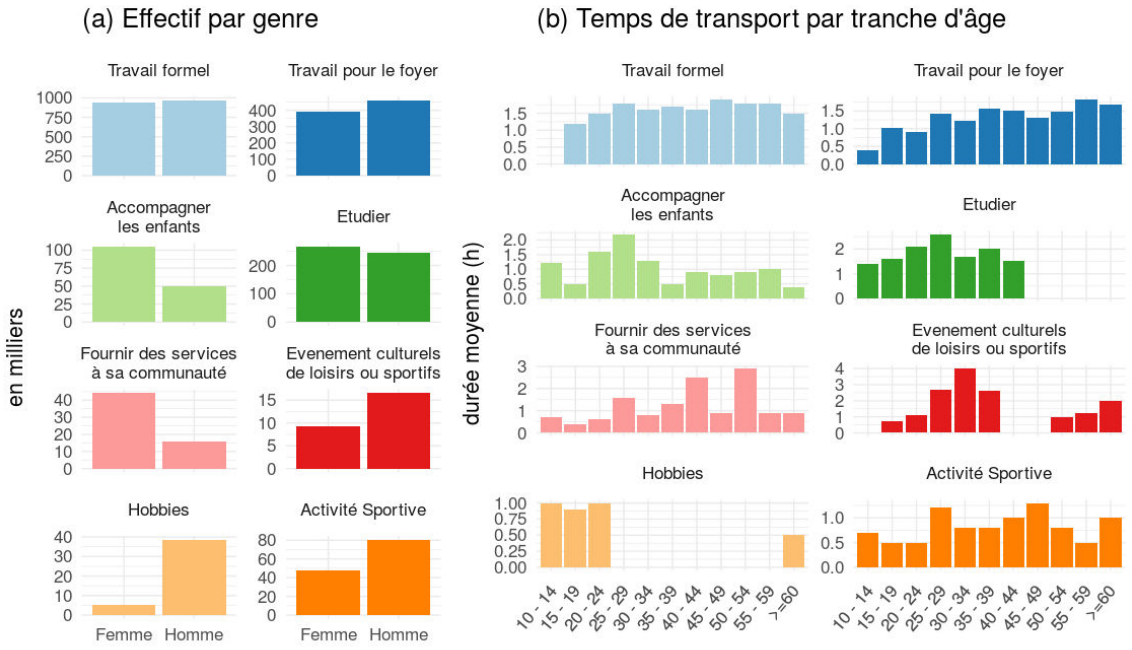


FIGURE 29 Tendances de déplacements à Bangkok selon les activités à réaliser (NSO, 2009). Le nombre de personnes qui se déplace pour réaliser une activité selon le genre (a), et les temps moyens de transport quotidien pour réaliser une activité selon les tranches d'âges (b).

La figure 29.b présente les temps moyens de transport quotidien pour se rendre à une activité par tranche d'âge. Le temps passé pour aller dans un lieu d'éducation double entre 10

et 30 ans, passant de 1h12 à 2h24 quotidiennes (aller et retour). Ce fait intéressant traduit la raréfaction des lieux d'éducations selon les niveaux scolaires (et la taille des établissements). Ceci implique leur éloignement progressif des domiciles, où le maillage des écoles est plus dense que celui des collèges, lui-même plus important que celui des lycées, etc.

Les temps de déplacement pour effectuer un travail formel oscillent entre 1h12 (pour les plus 15-19) ans et plus de 1h54 (45-49 ans), avec une moyenne à 1h42. Pour ce qui est de la participation à l'activité économique du foyer, les plus jeunes se déplacent moins longtemps, avec 25 minutes pour les 10-14 ans, alors que les plus de 15 ans mettent entre 1h et 1h50 par jour, avec une moyenne à 1h26. Les activités sportives sont généralement effectuées à proximité du domicile, avec des temps de déplacement qui varient de 30 minutes à une heure et demie. Les 30-34 ans seraient prêts à passer 4 heures dans les transports pour assister à des événements sportifs ou visiter des lieux culturels. Les fortes variations entre les tranches d'âges sont difficilement explicables pour les autres activités suggèrent que les échantillons n'étaient peut-être pas assez grands.

Ce *time use survey* fourni donc quelques éléments intéressants. Car si les hommes et les femmes ont le même accès à l'emploi et à l'éducation, les autres activités sont très genrées. Le temps passé dans les transports est globalement très long, et a aussi tendance à augmenter avec l'âge pour certaines activités (étudier, travailler). Mais ces informations ne sont que des moyennes, et nous ne connaissons pas les effectifs de chacun de ces sous-groupes. Par exemple combien de personnes constituent les 20-24 ans qui prennent en moyenne 2 h par jour pour se rendre dans un lieu d'éducation ? Et il serait aussi plus intéressant d'avoir pour chaque catégorie les distributions des temps passés à se rendre dans chacune de ces activités. Cette enquête sur l'utilisation du temps ne prend pas non plus en compte les niveaux de richesse économique, susceptibles d'influencer les déplacements.

Mobilité des classes les plus défavorisées

Avec des temps de transport très longs dans une ville souvent congestionnée, la notion de « valeur temps » s'est progressivement développée, surtout chez les classes moyennes et les hommes d'affaires de la ville, qui préféreront alors prendre le métro, plus rapide que les transports par la route (Richardson et Jensen, 2008). Si les mêmes auteurs font le constat que les personnes les plus défavorisées n'ont pas les moyens de prendre le métro à Bangkok, un système de bus et train gratuit a été mis en place par la municipalité à partir de 2008 permettant à environ 470 000 bénéficiaires de voyager quotidiennement (Ratanawaraha et Chalermpong, 2016).

Dans leur étude reposant sur des enquêtes qualitatives auprès d'environ 500 personnes aux revenus les plus modestes, Punpuing et Ross, (2001) ont montré que la plus grande partie des

personnes interrogées travaillaient près de chez elles (57 %). Parmi les personnes qui commutent sur des distances plus importantes, 45 % se déplaçaient de la périphérie vers le centre, 18 % restaient dans le centre et 28 % dans les zones périphériques et 8 % allaient du centre vers la périphérie (Punpuing, 1993). Seule une minorité de ces voyageurs aurait une perception de temps de transport aberrants (Punpuing et Ross, 2001), mais l'étude se base sur des données collectées dans les années 1990.

Une autre étude s'est basée sur l'interview de 463 personnes résidant dans 15 des quartiers les plus pauvres de la ville, situés à la fois dans le centre, le péricentre et la périphérie (Ratanawaraha et Chalermpong, 2016). S'il existe quelques variations au sein des groupes et notamment des catégories professionnelles, le temps moyen par voyage était d'environ 30 minutes, avec un écart-type entre 10 et 20 minutes, ce qui suggère encore que finalement, les personnes les plus défavorisées auraient tendance à travailler relativement à proximité de leur domicile. Les hommes utiliseraient plus la moto, tandis que les femmes seraient plus dépendantes des transports publics (Ratanawaraha et Chalermpong, 2016).

Bangkok est donc une ville très hétérogène, composée d'un grand nombre de types de quartiers. La population est principalement concentrée dans un grand centre, avec toutefois en périphérie des zones satellites relativement peuplées. Les navettes domicile-travail sont principalement de type périphérie / centre, même si un grand nombre de personnes commute localement. Les durées consacrées quotidiennement au transport sont assez importantes, et fluctuent selon l'âge des individus, le genre le niveau social et les activités visées (travail, loisirs, etc.).

Ces considérations permettent de dresser un rapide portrait des mobilités dans la ville. La section suivante évaluera si ces informations générales peuvent déjà servir à améliorer la lecture du déroulement des épidémies de dengue ces dernières années dans la capitale thaïe.

2.3 Dengue à Bangkok

La localisation de Bangkok en zone intertropicale est propice au développement des moustiques du genre *Aedes.* La dengue y est d'ailleurs endémique, et chaque année des milliers de personnes sont contaminées, surtout des jeunes (75 % des cas suspects étaient des personnes de moins de 30 ans figure 30). La dengue y est considérée comme une maladie à déclaration obligatoire, c'est-à-dire que tout médecin diagnostiquant la maladie doit faire remonter l'information aux hôpitaux sentinelles de la BMA. Néanmoins, il a été estimé que seuls un peu plus de 10 % des cas de dengue symptomatiques[75] étaient reportés en Thaïlande (Undurraga *et al.,* 2013), ce qui implique des biais difficilement quantifiables.

75. Sans compter les cas asymptomatiques.

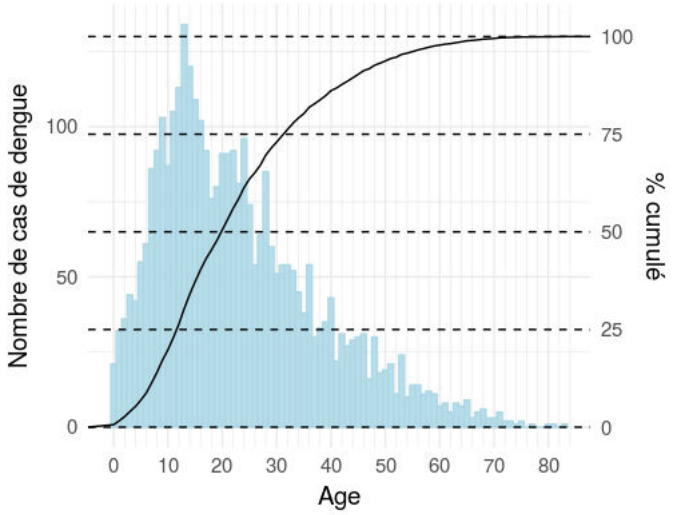


FIGURE 30 Nombre de cas de dengue à Bangkok en 2013 par tranche d'âge. Source : BMA).

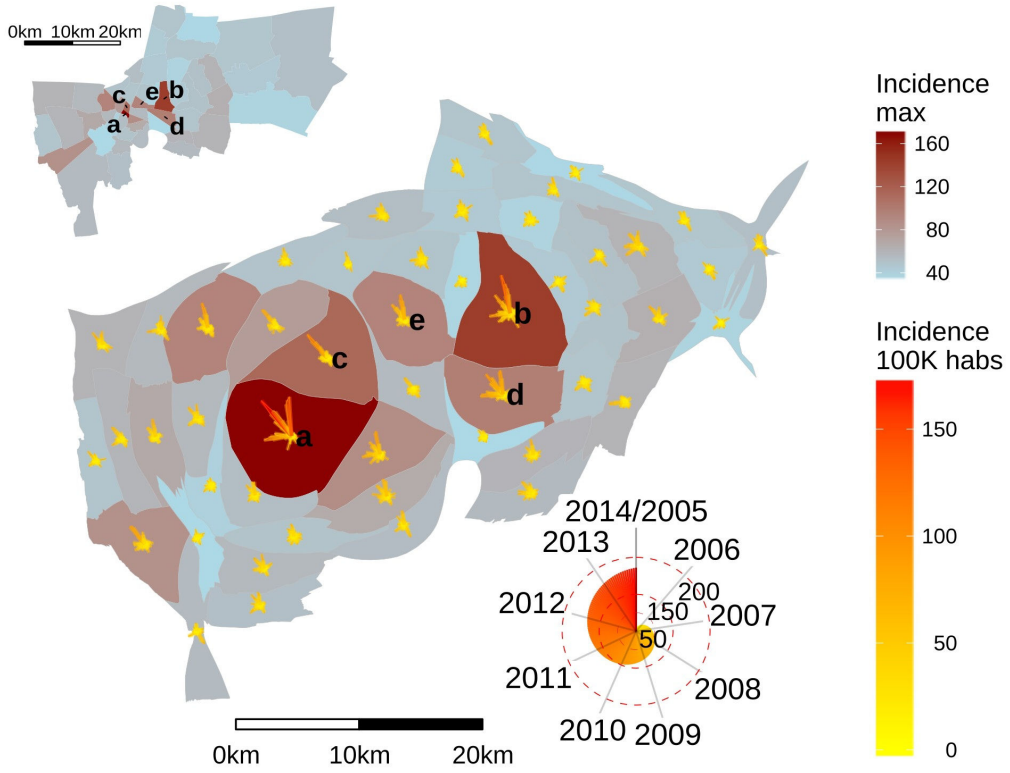


FIGURE 31 Carte par anamorphose de la répartition des cas de dengue par khet entre 2005 et 2013. La taille du khet est ici proportionnelle à l'incidence maximale enregistrée sur la période. L'angle des traits correspond à la période enregistrée, tandis que le gradient jaune/orange et la longueur de traits indiquent l'ampleur de l'incidence des cas de dengue pour 100 000 habitants. Source : BMA.

La figure 31 est une carte par anamorphose qui présente l'incidence de la dengue à l'échelle des districts (*khets)*, soit le nombre de personnes atteintes une année donnée divisée par la population du secteur. La taille des *khets* est proportionnelle à l'incidence maximale enregistrée entre 2005 et 2014, l'angle des traits correspond à un mois de la période, avec un écart de 40° entre chaque année (3,3° entre chaque mois). La longueur des traits indique l'incidence enregistrée un mois donné. Les labels de "a" à "e" correspondent aux 5 *khets* les plus touchés sur l'ensemble de la période et figurent également sur la carte en encart en haut à gauche ce qui permet de mieux les situer dans l'espace (sans anamorphose).

Nous pouvons noter tout d'abord que c'est dans les districts de l'hyper-centre que sont enregistrées les incidences les plus importantes, avec néanmoins de grandes variations saisonnières et inter-annuelles. Alors que le *Khet* "a" présente trois grandes épidémies en 2011, 2012 et 2013, le *Khet* "c", pourtant voisin, n'a observé une incidence très élevée qu'en 2012, et l'incidence maximale enregistrée pour le *Khet* "b" l'a été en 2013. À l'est de la ville, les incidences les plus importantes ont surtout été observées avant 2011, alors que les épidémies les plus marquées se sont plutôt déroulées après 2012 pour les districts à l'ouest.

2.3.1 Aspects climatiques

Les températures jouent un rôle déterminant dans le cycle de vie des moustiques (voir chapitre 1), mais celles-ci varient peu à Bangkok et sont quasiment toujours favorables au développement du vecteur de la dengue, du fait du climat intertropical de la ville (Misslin, 2017). Si l'îlot de chaleur urbain induit par la densité de population de l'hyper-centre influence la durée et la précocité des épidémies de dengue (Misslin *et al.,* 2018, 2017), le développement des moustiques et le caractère saisonnier très marqué des épidémies à Bangkok (figure 32) seraient plus dues au régime des précipitations qu'aux températures (Misslin, 2017), contrairement à ce qui est observé à l'échelle mondiale, notamment par (Kraemer *et al.,* 2015b). La figure 32 montre le régime des précipitations et les cas de dengue enregistrés mensuellement entre 2005 et 2013. Les données de précipitation sont issues de 5 stations météorologiques situées à Bangkok (RCCES, 2017), puis moyennées. Les cas de dengue nous ont été fournis par la BMA. Nous pouvons noter que les précipitations suivent un régime globalement bi-modal, avec de fortes pluies enregistrées entre février et mars, puis entre août et septembre. Concernant la dengue, le nombre de cas enregistrés est relativement bas en début d'année, entre janvier et mars, et atteint un maximum soit vers avril/mai (2005, 2006), parfois juin ou juillet (2011, 2008), soit vers septembre/octobre (2007, 2009, 2010, 2012). Ceci suggère déjà la présence d'une relation entre les précipitations et le nombre de cas de dengue enregistrés, car une flambée épidémique apparaît généralement après les grosses pluies de début ou de fin d'année.

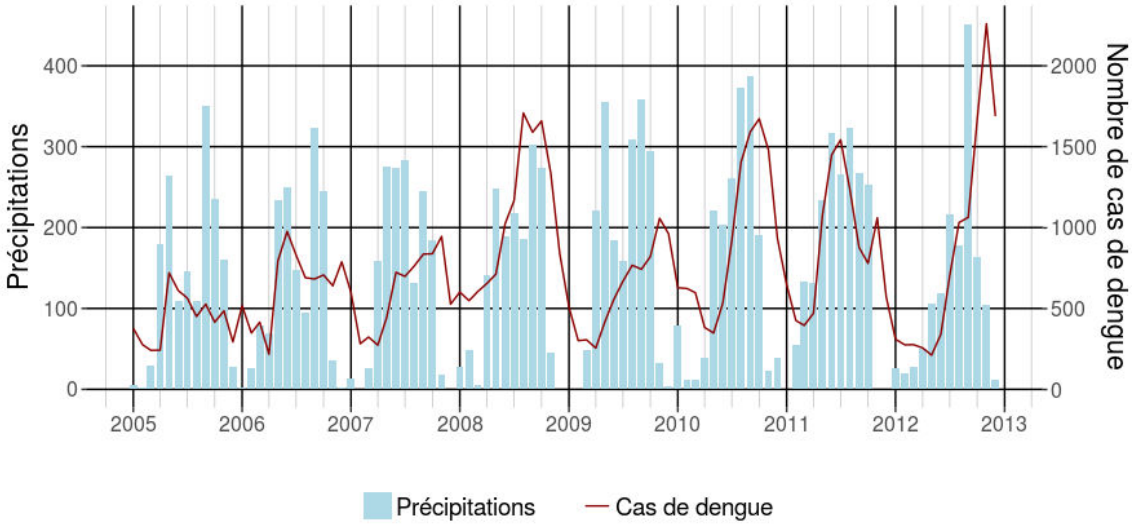


FIGURE 32 Lien entre précipitation et cas de dengue enregistrés à Bangkok entre 2005 et 2012.

Pour étudier plus en détail les caractères périodiques de ces deux séries temporelles et essayer d'apprécier l'influence des précipitations sur l'ampleur de l'épidémie, nous allons décomposer ces dernières selon l'algorithme STL (Seasonal and Trend decomposition using Loess) introduit par Cleveland *et al.* (1990). Cette méthode à la fois flexible et robuste se base sur les principes de la théorie des signaux présentée au XIXe siècle par Joseph Fourier. Dans le cas d'une STL, le signal D(t) est décomposé selon une redondance sur une période donnée, ici 1 an, définissant la saisonnalité S(t), soit les valeurs mensuelles moyennées, à laquelle s'ajoutent (2) la tendance de l'évolution générale sur T(t) sur l'ensemble de la période étudiée[76] ainsi que (3) des valeurs non expliquées par les données internes de la série temporelle, R(t), avec D(t) = S(t)+T(t)+R(t).

La figure 34 montre la décomposition des cas de dengue par l'algorithme STL selon une période de 12 mois. Elle montre respectivement de haut en bas les données observées, la périodicité (ou saisonnalité), la tendance générale ainsi que les écarts non expliqués statistiquement. La courbe de tendance présente les plus grandes valeurs absolues en termes de cas de dengue, suggérant le caractère endémique et évolutif de l'épidémie à Bangkok, indépendamment des cycles annuels. Le côté saisonnier est ici mis en évidence, avec la présence de deux pics annuels, aux troisième et quatrième trimestres, mais avec une contribution relativement faible (entre -200 et 250). Finalement, les résidus représentent une part non négligeable du signal initial (entre -400 et +1000), ne présentent pas de caractère cyclique et forment en quelque sorte les discontinuités temporelles de l'épidémie.

76. Calculée selon une régression locale (ou loess) sur les données originales auxquelles sont soustraites les valeurs de saisonnalités calculées précédemment.

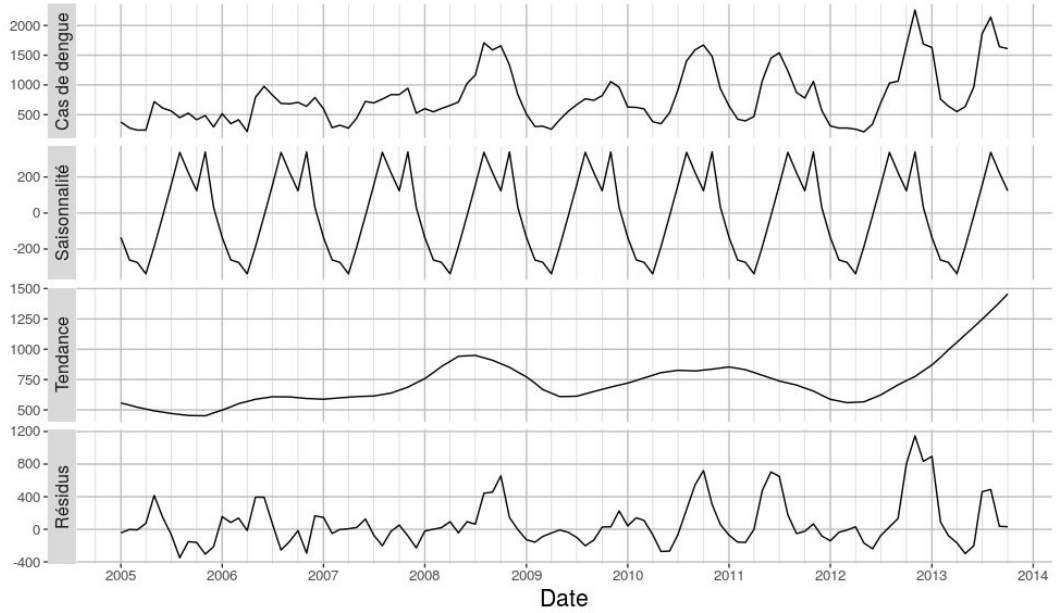


FIGURE 33 Décomposition temporelle des cas de dengue à Bangkok selon la méthode STL.

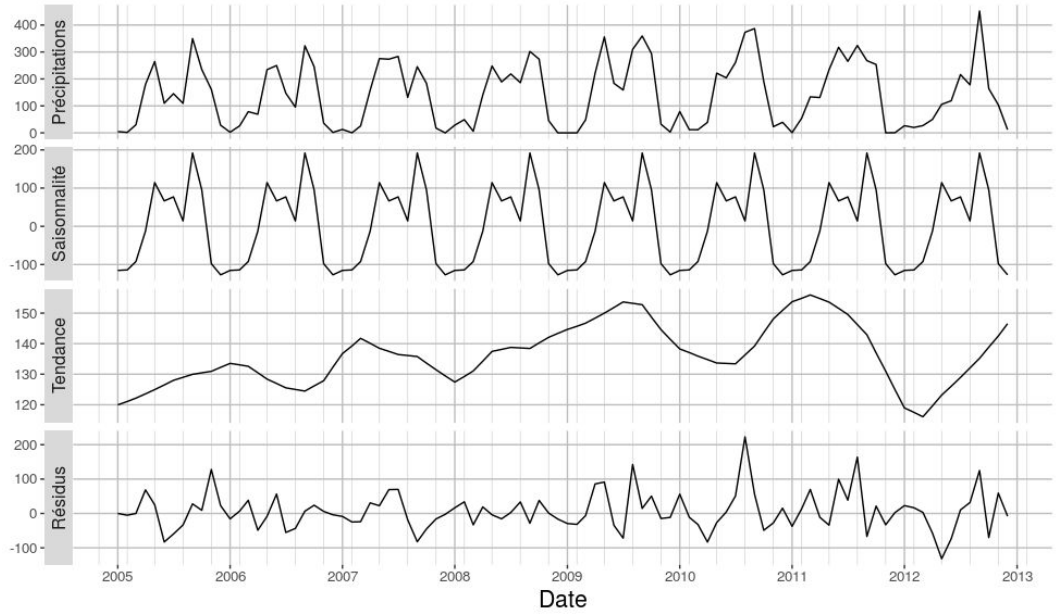


FIGURE 34 Décomposition temporelle des précipitations à Bangkok selon la méthode STL.

Pour ce qui est de la décomposition temporelle des précipitations à Bangkok (figure 34), nous pouvons noter que la part de l'aspect saisonnier est plus importante que pour les cas de dengue, ce qui suggère une plus grande régularité. Nous observons également une courbe de tendance croissante, suggérant des saisons des pluies qui se sont intensifiées entre 2005 et 2013.

Nous allons dans un premier temps supprimer dans chacune des séries la part non

expliquée (les résidus) afin d'obtenir des séries temporelles "type". Une première étude des corrélations croisées entre ces deux séries (figure 35), montre que leur niveau de corrélation est très important dès lors que la phase est différente de -1,-2 et 6 mois (modulo 12) avec un maximum à 2 mois (modulo 12). Ceci nous permet de poser l'hypothèse que les pics de dengue surviennent environ 2 mois après des pics de précipitation (figures 36 et 37), constat déjà observé au Bangladesh (Salje *et al.*, 2016).

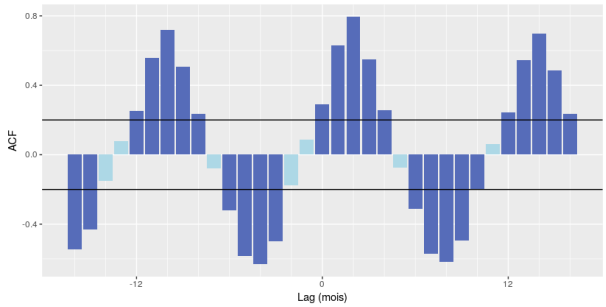


FIGURE 35 Analyse des corrélations croisées temporelles entre les données non bruitées des précipitations et des cas de dengue. Les barres en bleu foncé montrent des corrélations croisées significatives.

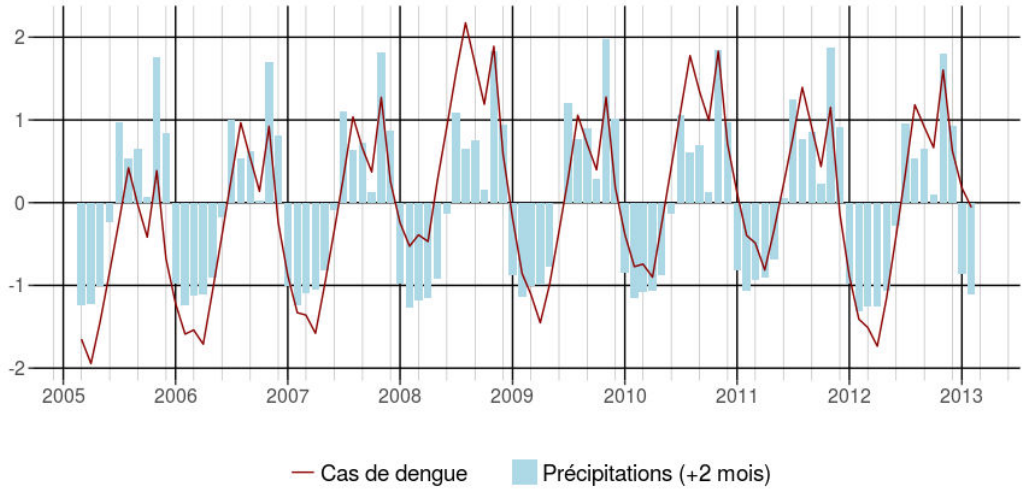


FIGURE 36 Superposition des données débruitées et centrées-réduites pour les cas de dengue et les précipitations, avec l'ajout d'une phase de 2 mois pour ces dernières.

Les figures 36 et 37 permettent de confirmer visuellement l'hypothèse d'un décalage global dans l'arrivée des pics de dengue à la suite de précipitations importantes. En effet, le premier pic de cas de dengue apparaît 3 mois après le premier pic de précipitation (en général moins marqué), tandis que le second pic de dengue apparaît 2 mois après le 2e pic des précipitations. Il y a donc un temps de latence de quelques semaines, probablement lié au temps de développement des moustiques et à la propagation de l'épidémie sur l'ensemble du

territoire.

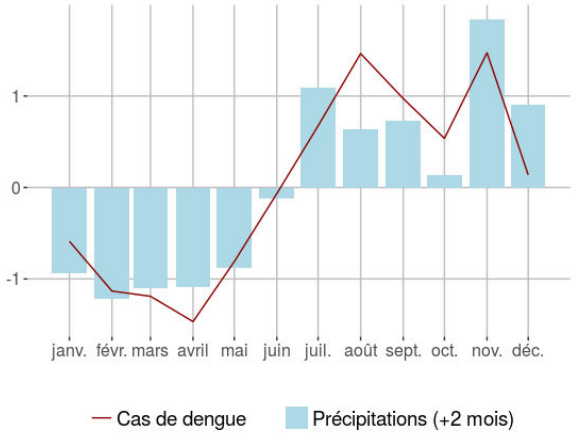


FIGURE 37 Superposition des séries temporelles de la saisonnalité, centrées-réduites pour les cas de dengue et les précipitations, avec l'ajout d'une phase de 2 mois pour ces dernières.

Ainsi, contrairement à ce que Brady *et al.* (2014) ou Kraemer *et al.* (2015b) ont pu montrer à l'échelle globale, il semblerait que cycles des précipitations jouent un rôle très important sur les épidémies de dengue à Bangkok, confirmant les observations de Misslin (2017). L'importance des résidus dans la décomposition temporelle des cas de dengue pourrait quant à elle s'expliquer par d'autres facteurs, notamment des défaillances dans le système de surveillance, des différences d'immunité en fonction des souches de dengue en circulation, ou encore par les mobilités humaines qui peuvent contribuer à diffuser plus ou moins largement l'épidémie dans la ville.

Des discontinuités spatio-temporelles des cas de dengue

L'objet de cette section est de comparer les séries temporelles des incidences de dengue par *Khet* avec les séries correspondantes des nombres de cas de dengue enregistrés à l'échelle de la ville. Pour les *Khets*, nous avons choisi de travailler sur les incidences plutôt que sur le nombre de cas pour prendre en compte la densité de population des sous-districts. Pour les données à l'échelle de la ville, incidence ou effectif ne font pas de différence (identique à un coefficient près), ce qui nous permet de comparer les deux jeux de données.

Pour apprécier si un *Khet* voit la temporalité de ces épidémies évoluer conformément à la tendance de la métropole thaïe, nous calculons pour chaque district les coefficients R^2 ajustés entre les séries temporelles des taux d'incidences et le nombre de cas déclarés à la même période dans l'ensemble de la ville sauf dans le district en question afin de limiter le risque d'auto-corrélation. Les calculs de ces R^2 portent sur les données brutes et les données décomposées (saisonnalité, tendance et résidus) et sont visibles dans la figure 38. Ils montrent donc dans quelle mesure l'évolution temporelle de l'incidence de la dengue par *Khet* corrèle

avec le nombre de nouveaux cas enregistrés dans l'ensemble de la ville.

Pour ce qui est des données brutes, les coefficients de corrélation varient entre 0.28 et 0.78, avec des valeurs importantes au nord-est de la ville. Les plus faibles valeurs sont enregistrées à l'est de la ville, mais également dans certains districts de l'hyper-centre. Les coefficients de corrélation les plus élevés sont observés lorsque nous comparons les aspects saisonniers des séries, soit la correspondance entre les pics mensuels moyens sur la période, avec des valeurs comprises entre 0.4 et 0.95. Ces R^2 très importants peuvent en partie s'expliquer par le fait que les coefficients de corrélation ne sont calculés que sur 12 dates. Mais beaucoup de districts présentent tout de même une saisonnalité très en phase avec l'épidémie globale (notamment dans le centre-est de la ville), mais parfois voisins de *Khet* dont les pics de dengue sont beaucoup moins synchrones avec ceux de la ville.

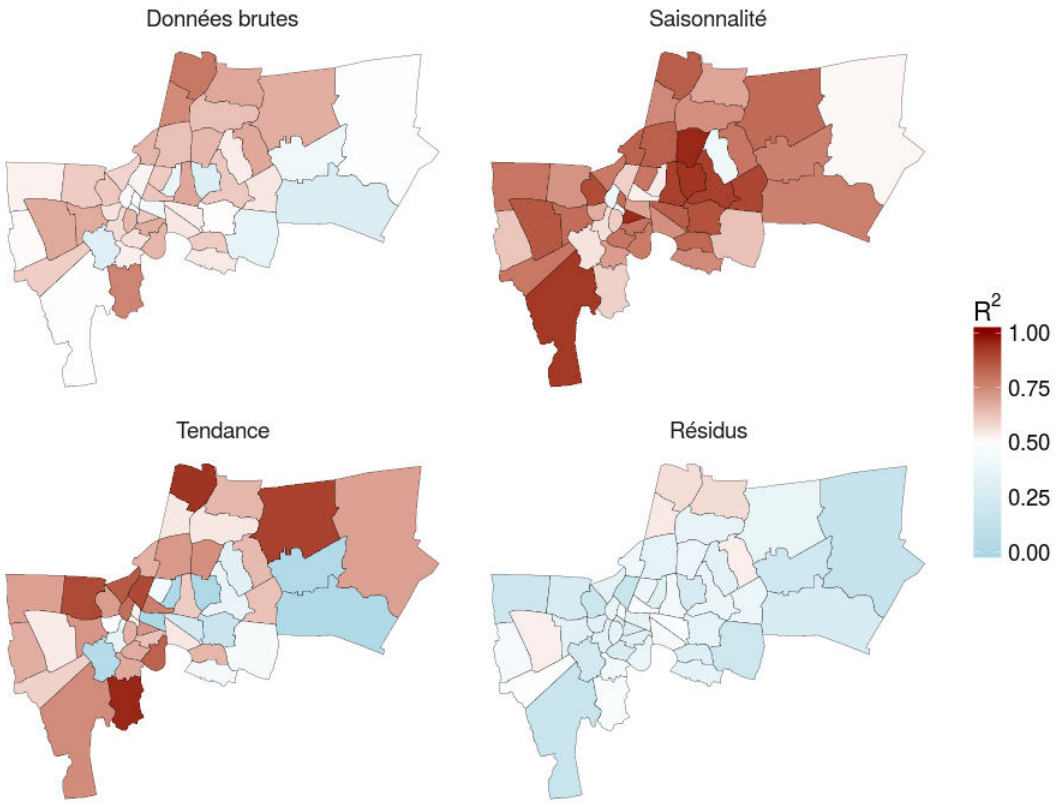


FIGURE 38 Spatialisation des coefficients R^2 issus d'une régression linéaire entre les séries temporelles des cas de dengue par *Khet* et l'ensemble de Bangkok.

La série temporelle des tendances (~correspondance dans l'amplitude) est celle qui présente les plus grands écarts de R^2. Alors que certains districts du centre et du nord ont des niveaux de corrélation importants avec la tendance globale, des districts pourtant voisins présentent des R^2 très faibles. Finalement, le niveau de correspondances entre les résidus est

relativement faible, ce qui peut tout d'abord s'expliquer par leur caractère non périodique. Néanmoins, certains *Khets* du nord, et dans une moindre mesure à l'ouest ou au sud ont des séries temporelles de résidus assez proches de l'ensemble de la métropole. À noter également que les districts dont les pics mensuels moyens correspondent relativement à ceux de l'épidémie globale ne présentent pas nécessairement des amplitudes (tendances annuelles) en phase avec le déroulement de l'épidémie à l'échelle de la ville.

Cette rapide étude permet de montrer l'importance des discontinuités spatio-temporelles dans le déroulement de la maladie, mettant en avant que certaines zones voient leurs cas de dengues évoluer en phase avec l'épidémie globale, alors que d'autres zones sont plus décorrélées, sans qu'aucune organisation spatiale ne ressorte clairement.

2.3.2 Dengue et population

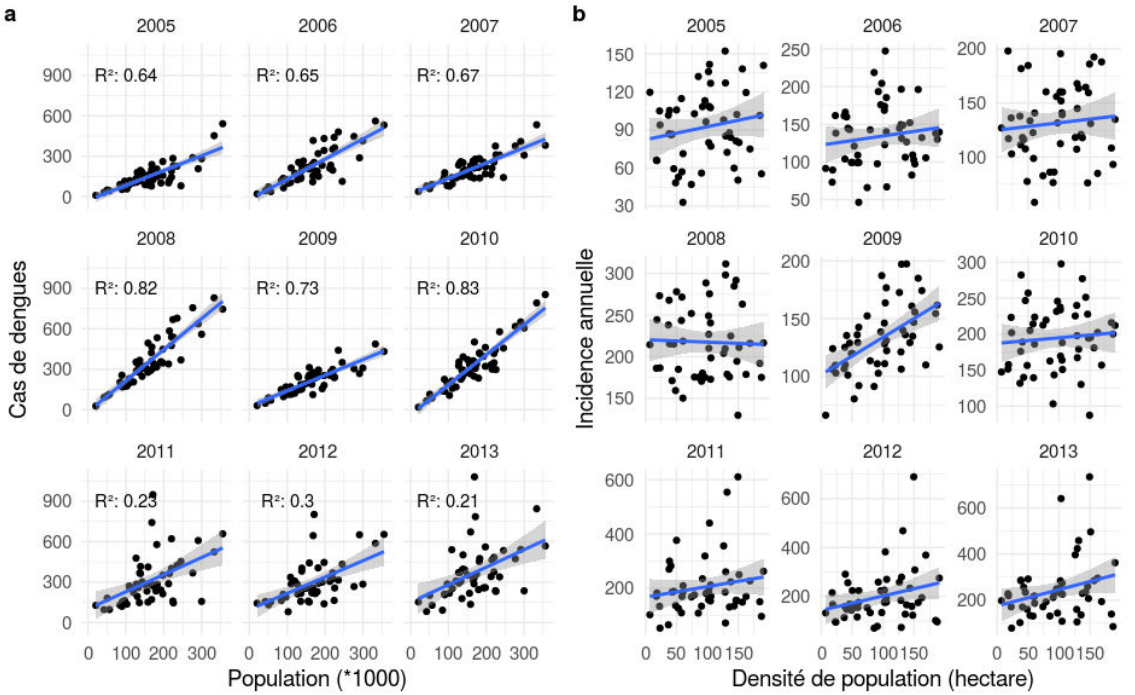


FIGURE 39 Lien entre le nombre de cas de dengue enregistrés et la population par *Khet* (a) et lien entre l'incidence annuelle et la densité de population (b).

Nous allons maintenant observer les liens entre le nombre de cas de dengue enregistrés une année donnée dans un *Khet* et la population associée (figure 39.a) puis l'impact de la densité de population sur l'incidence de la dengue (figure 39.b). N'ayant en notre possession que les incidences et non les effectifs, nous calculons le nombre de personnes infectées en multipliant ces incidences par la population de 2010, ce qui peut poser un premier biais. De même, la densité de population est définie à partir des mêmes données du recensement. Alors qu'une relation linéaire relativement forte existe entre le nombre de cas enregistrés et la population

par district entre 2005 et 2010 (R^2 entre 0,64 et 0,83), cette relation disparaît après 2011 (R^2 entre 0,21 et 0,3), où quelques *Khets* de taille intermédiaire (entre 150 000 et 200 000 habitants) sont extrêmement touchés, notamment dans le centre-ville historique. En revanche, aucune relation, linéaire ou non, n'apparaît lorsque nous confrontons l'incidence et la densité de population, contrairement à ce que nous avons montré à Delhi, ou comme l'ont noté à Bangkok (Salje *et al.,* 2017) en créant des chaînes de contaminations à partir de génotypes et de sérotypes géolocalisés et datés. La même étude a d'ailleurs mis en évidence l'importance des contaminations locales, qui peuvent expliquer les flambées dans certains quartiers du centre-ville.

À noter que dans nos données, les cas de dengue sont enregistrés au domicile et ces derniers ne sont pas nécessairement les lieux de contamination des individus du fait du caractère diurne du moustique incriminé dans la propagation de la maladie. Ceci renforce l'importance de l'étude des mobilités individuelles et collectives entre différents secteurs, notamment lorsqu'une zone très attractive subit une flambée importante.

Synthèse

- Delhi et Bangkok sont des mégapoles très hétérogènes, faites de discontinuités et de ruptures entre les différents quartiers. Les inégalités sociales sont perceptibles dans les deux capitales, mais nettement plus marquées et visibles à Delhi. Les potentiels et tendances de mobilités individuelles sont différents selon le genre, l'âge, le capital économique et les zones des villes.

- La dengue est endémique dans ces deux villes. Si un lien apparaît parfois entre l'ampleur des épidémies de dengue une année donnée et les conditions climatiques ou la densité de population, ces dernières n'expliquent pas totalement les discontinuités des incidences observées plus localement dans la ville. Les mobilités humaines semblent donc bien jouer un rôle dans leur propagation.

Néanmoins, le rôle de ces mobilités reste à être déterminé plus précisément. Les différentes analyses proposées sont plus descriptives qu'explicatives. La taille de ces villes et leurs spécificités locales font qu'il est nécessaire de trouver d'autres méthodes pour cerner les mobilités urbaines et leur impact dans la propagation des maladies infectieuses. Une approche est de passer par la modélisation de ces divers phénomènes et systèmes, afin de pouvoir mieux les comprendre.

Chapitre III: Prendre en compte les mobilités dans la modélisation des arboviroses : La possibilité de modèles

La modélisation est un ensemble de méthodes permettant d'obtenir une représentation une représentation simplifiée de la réalité en se basant sur des simplifications des relations entre les facteurs influençant le phénomène observé (Eliot et Daudé, 2006 ; Varenne, 2008). Elle est très utile dans la compréhension d'un comportement d'un système complexe lorsque différentes entités interagissent de manières non-linéaires (Batty, 2009b ; Manson, 2001). Elle permet également de faire émerger des propriétés non prévues ou des processus non triviaux (Iltanen, 2012 ; Parrott, 2002). La modélisation peut avoir différents objectifs, comme expliquer de manière rétrospective le déroulement d'un phénomène, essayer de le prédire, ou quantifier les impacts de différents facteurs (Varenne, 2008).

Ce chapitre s'intéresse aux différentes méthodes de modélisation appliquées aux arboviroses et notamment à la dengue. Il traitera d'abord de l'évolution des techniques de modélisation en épidémiologie, puis s'intéressa à l'apport des mobilités humaines dans ces modèles. Nous présenterons finalement les concepts qui semblent être les mieux adaptés pour associer les mobilités humaines à la modélisation dans le contexte complexe de la dengue en zone urbaine.

1 Des débuts des modèles compartimentaux à la prise en compte du vecteur de la dengue

La première modélisation en épidémiologie a été effectuée par Bernoulli, où il présente en 1760 une approche mathématique de l'évolution de la petite vérole d'abord sans immunité, puis avec une inoculation d'une forme bénigne à la naissance[77] (Bacaër, 2011 ; Bernoulli, 1760 ; Smith *et al.,* 2012). Son approche, ensuite discutée par D'Alembert est très simplificatrice de la réalité, mais pose les bases de l'approche par la formulation mathématique des épidémies permettant une meilleure compréhension de leur déroulement Bacaër, 2011).

1.1 Ross et les premiers modèles mathématiques en épidémiologie

Au début du 20e siècle, les chercheurs avaient déjà réparti les épidémies infectieuses dans trois grandes classes, en fonction de leur cycle d'apparition constaté et de leur évolution au cours du temps (Ross, 1916). Les épidémies qui fluctuent peu à travers le temps, comme la lèpre et la tuberculose d'un côté ; celles qui sont endémiques, mais qui se déclenchent soudainement à des intervalles fréquents, telle la malaria ; et finalement celles qui disparaissent complètement après des épidémies aiguës, comme la peste ou le choléra (Ross, 1916). Pour comprendre l'origine de ces différences et les phénomènes et relations sous-jacents, Ronald Ross[78] s'intéressa à la structure des courbes épidémiques et notamment à leur formulation mathématique (Bacaër, 2011 ; Smith *et al.,* 2012). Il proposa un premier modèle sur la transmission de la malaria (Ross, 1908), puis un système à deux équations différentielles prenant en compte les hôtes humains et les vecteurs (Ross, 1911a, 1911b). Il montra notamment qu'en dessous d'une population seuil de moustique, l'épidémie de malaria devrait disparaître. Reformulé par (Lokta, 1923), cela revient à dire qu'avec une population de moustique relativement faible, le nombre moyen d'infections secondaires R_0 devient inférieur à 1 et l'épidémie ne se propage plus (Bacaër, 2011 ; Smith *et al.,* 2012).

Waite commenta ensuite les travaux de Ross, en insistant sur les facteurs qui influencent le taux de personnes infectées par la malaria : « *In general, the rate is continually changing owing to (a) new infections, (b) recoveries, (c) emigration and immigration, (d) the birth and death rates, and (e) the extent to which cases are isolated, as well as owing to changes in the mosquito population* » (Smith *et al.,* 2012 ; Waite, 1910). Tout comme Ross, Waite souligna la nécessité d'avoir des données et des statistiques, et notamment sur les mobilités et les migrations[79].

77. Sa préconisation d'inoculer une forme bénigne de la petite variole ne fut pas suivie de fait. Le roi Louis XV mourra de cette maladie en 1774.

78. D'ailleurs le premier Britannique prix Nobel de médecine et de physiologie en 1902 pour ces travaux sur la transmission de la malaria.

79. "As emigration and immigration vary considerably in different localities, and in the same locality at different times, their influence on the malaria rate cannot be satisfactorily dealt with except in particular cases where the necessary statistics are available ; neither would results in general terms be of much practical

Via ces formulations mathématiques, Ross va notamment préconiser l'éradication des moustiques par le contrôle des gîtes larvaires (Ross, 1911a, 1902) dont le coût est relativement faible par rapport aux conséquences sanitaires des épidémies de malaria (Smith *et al.,* 2012). Il insista ensuite sur la grande sensibilité et l'interdépendance des différents paramètres qui régissent ces équations, car une faible variation de ces derniers entraîne des changements drastiques dans les distributions théoriques (Ross et Hudson, 1917). Mais Ross manquait de données et d'informations pour calibrer ces modèles, et ces travaux vont aussi orienter les collectes de données en entomologie et épidémiologie, en orientant les recherches sur les paramètres les plus importants, comme la force d'infection, la densité de moustiques ou encore le taux de guérison (Smith *et al.,* 2012).

1.2 Les premiers modèles compartimentaux

Dans la lignée des travaux de Ross, McKendrick va diviser une population donnée en trois catégories (ou compartiments) selon une séquence d'états vis-à-vis d'une maladie au cours du temps (McKendrick, 1926). Le premier état concerne la sous-population non-immunisée ou susceptible (S) d'être infectée. Seule une portion de cette catégorie pourra être contaminée à l'itération suivante et rejoindra alors le groupe des Infectés (I). Après un laps de temps, les personnes infectées sont guéries et développent une immunité face à la maladie considérée et sont donc immunisées (R, pour Recovered). Les individus sont décrits ici sous forme de stock en fonction de leur état vis-à-vis d'une maladie.

$$\mathrm{S} \rightarrow \mathrm{I} \rightarrow \mathrm{R}$$

Cette approche peut être déterministe, c'est-à-dire qu'elle considère les changements d'état comme régis par des paramètres posés et fixés. Par exemple, le premier modèle déterministe de ce type formulé mathématiquement (Kermack et McKendrick, 1927) se compose de 3 équations différentielles régissant l'évolution de chacun des états S, I et R à travers le temps et en fonction de la force d'infection β et d'un taux de guérison γ :

$$\frac{dS}{dt} = -i\beta SI \tag{1}$$

$$\frac{dI}{dt} = i\beta SI - i\gamma I \tag{2}$$

$$\frac{dR}{dt} = i\gamma I \tag{3}$$

use."(Waite,1910)

À chaque pas de temps, le nombre de personnes nouvellement infectées est décrit comme étant le produit de la capacité β (ou force d'infection) d'une personne infectée *I* à transmettre une maladie à une personne susceptible *S*. Cependant, les personnes infectées finissent par guérir (*R)* selon un taux de guérison γ, inversement proportionnel à la durée de l'infection, et acquièrent une immunité. Ainsi, au pas de temps suivant, le nombre de personnes susceptibles *S* est réduit de βIS, soit le nombre de personnes nouvellement infectées (équation 1). Le nombre de personnes infectées *I* augmente de βIS auquel on soustrait le nombre de personnes ayant guéri γI (équation 2). Ces personnes guéries sont finalement ajoutées au compartiment R (équation 3). Dans ce cas précis, à chaque pas de temps, la somme des *S*, *I* et *R* est constante et est égale à la population totale initiale.

Aux états S, *I*, *R* peut s'ajouter l'état exposé *E, c'est-à-dire* que les personnes sont contaminées par la maladie, mais ne sont pas tout de suite infectieuses. Le lien entre les compartiments du modèle suit la relation :

$$\mathrm{S} \rightarrow \mathrm{E} \rightarrow \mathrm{I} \rightarrow \mathrm{R}$$

Nous pouvons également considérer que l'immunité diminue avec le temps et que les personnes guéries redeviennent susceptibles, ce qui revient à ajouter une relation entre les compartiments *R* et *S*, donnant ainsi un modèle *SEIRS* :

$$\mathrm{S} \rightarrow \mathrm{E} \rightarrow \mathrm{I} \rightarrow \mathrm{R} \rightarrow \mathrm{S}$$

La formulation sous forme d'équations différentielles de ces modèles *SEIR*, et SEIRS, suivent la même logique que celle montrée précédemment, avec l'ajout du compartiment représentant les personnes exposées.

Le choix des compartiments dépend de l'épidémie étudiée. Par exemple, les modèles de type *SI* ne prennent en compte que les personnes susceptibles et infectées, suggérant que les individus ne peuvent guérir de la maladie. Ils sont généralement appliqués à des maladies telles que le VIH. Les modèles de type *SIS* suggèrent quant à eux qu'une fois guéris, les individus n'acquièrent pas d'immunité et redeviennent susceptibles.

Ces modèles peuvent se complexifier car d'autres facteurs peuvent être ajoutés. Si on considère que la maladie étudiée est potentiellement mortelle, il convient d'ajouter un taux de mortalité aux personnes infectées. Si la durée de l'épidémie est relativement longue, il faut également prendre en compte les naissances, ce qui revient à augmenter le nombre de personnes susceptibles à chaque pas de temps selon un taux de natalité (sauf si le nouveau-né acquiert une immunité *in utero*). Dans ces deux cas, on parle de modèles ouverts, car la population n'est plus constante au cours du temps.

Ces systèmes d'équations différentielles sont résolus mathématiquement, et sont extrêmement sensibles aux paramètres initiaux, notamment la force et la période d'infection. Une force d'infection élevée (dans le cas d'une maladie très contagieuse) va entraîner une rapide contamination de l'ensemble de la population, tandis qu'une période d'infection longue implique que plus de personnes sont malades en même temps.

1.3 La prise en compte du vecteur

Les modèles déterministes classiques utilisés dans le contexte de maladies vectorielles reprennent les bases des modèles *SIR* ou *SEIR* et ajoutent de nouveaux compartiments décrivant les différents états du moustique, leur croissance et surtout leurs interactions avec l'Homme (Andraud *et al.*, 2012). Le premier modèle spécifique à la dengue a été développé par (Newton et Paul Reiter, 1992). Il ne prend en compte qu'un seul sérotype et est inspiré du modèle de (Bailey, 1975) initialement développé dans le contexte du paludisme (Andraud *et al.*, 2012). L'approche de Bailey est un modèle SIR classique pour les hôtes auquel s'ajoute des compartiments S et I pour les vecteurs :

Hôtes :

$$\frac{dS_h}{dt} = i\mu_h N_h - \frac{i\beta_h b}{N_h} S_h I_v - i\mu_h S_h \tag{4}$$

$$\frac{dI_h}{dt} = \frac{\beta_h b}{N_h} S_h I_v - (i\gamma_h + i\mu_h) I_h \tag{5}$$

$$\frac{dR_h}{dt} = i\gamma_h I_h - i\mu_h Rh \tag{6}$$

Vecteurs :

$$\frac{dS_v}{dt} = A - \frac{i\beta_v b}{N_h} I_h S_v - i\mu_v S_v \tag{7}$$

$$\frac{dI_v}{dt} = \frac{i\beta_v b}{N_h} I_h S_h - i\mu_v I_v \tag{8}$$

Les $_h$ et $_v$ sont relatifs aux hôtes et aux vecteurs, N étant la population. μ correspond au taux de mortalité, tandis que γ est l'inverse de la période d'infection (la durée où le malade est contagieux) et b le taux de piqûres par unité de temps. β_h et β_v sont les probabilités d'infection du vecteur vers l'hôte et de l'hôte vers le vecteur, respectivement. A est le nombre de moustiques naissant durant chaque intervalle de temps. Il s'agit d'une extension d'un modèle SIR classique, où la force d'infection est ici décrite comme $i\beta_h b I_v / N_h$, soit dépendante des paramètres du moustique. Elle est donc modulée au cours du temps en fonction du nombre de moustiques infectés et de paramètres fixes (β_h, b et N_h).

Dans le cadre des modèles déterministes appliqués à la dengue, les différents états des hôtes sont la plupart du temps décrit selon une séquence *SIR* ou *SEIR*. La plupart des études se focalisent sur une formalisation plus précise des caractéristiques du vecteur, ce qui influence la force d'infection qui conditionne le nombre de personnes nouvellement infectées. Certains gardent la structure *SI* des moustiques et ajoutent des probabilités de contact entre les hommes et les vecteurs et modulent la probabilité de piqûre en fonction de l'état du moustique (Derouich *et al.,* 2003) tandis que d'autres rajoutent un état exposé au moustique. Des travaux plus poussés ont largement augmenté le nombre de compartiments chez le vecteur en prenant en compte les stades aquatiques et conditionnant le développement par la température et les précipitations, et ajoutent de l'aléa (Lourenço et Recker, 2014). Certaines études prennent aussi en compte plusieurs souches de virus, et d'autres les niveaux de susceptibilité différents entre différents groupes d'une population. Par exemple des études ciblant les touristes en Thaïlande ont été effectuées selon un modèle SIR en deux groupes, séparant les touristes, plus naïfs face au virus, de la population locale (Polwiang, 2016 ; Pongsumpun et al., 2004). Mais ce dernier modèle ne prend pas en compte les mobilités et interactions entre les provinces.

Ainsi, malgré le fait que les mobilités et migrations soient des paramètres importants dans le déroulement et la simulation d'épidémies vectorielles et connus et discutés depuis plus d'un siècle (Waite, 1910), il existe étrangement peu d'études qui les prennent en compte dans le contexte de la dengue.

2 Modélisation d'épidémies et mobilité humaine

Le déplacement des hôtes est un paramètre primordial dans des modèles spatialement explicites de maladies infectieuses (Buckee *et al.,* 2013 ; Daudé et Eliot, 2005 ; Eliot et Daudé, 2006 ; Riley, 2007 ; Tizzoni *et al.,* 2014). Ces déplacements peuvent être traités de deux manières : sous forme de flux de stock, ou sous la forme de déplacements individuels. Les flux de stock sont généralement représentés par des matrices origines/destinations, où chaque variable correspond au nombre de personnes du lieu d'origine i se déplaçant vers le lieu de destination j. Une variante peut être une matrice de probabilité de déplacement, où un taux correspondant à la probabilité d'aller de i à j remplace la variable de stock.

La deuxième approche est individu-centrée et chaque individu de synthèse suit ainsi un itinéraire plus ou moins prédéfini en fonction des hypothèses et données de départ. Cet angle d'attaque nécessite donc de poser des hypothèses adaptées et mobilise diverses disciplines, allant de la sociologie des mobilités aux mathématiques appliquées. Une telle approche permet de suivre les individus, mais également de raisonner en termes de stock lorsque l'on agrège ces mobilités individuelles.

Dans le cadre de modèles SIR "classiques", il est possible à tout moment de la simulation d'injecter des individus infectés dans le modèle pour signifier l'arrivée de personnes malades. Nous pouvons citer le cas de l'épidémie de dengue survenue à Madère en 2012 qui a été simulée en utilisant le trafic aérien vers l'île comme proxy pour l'importation de l'épidémie (Lourenço et Recker, 2014). Une autre étude a adopté une approche en deux temps : une épidémie est d'abord simulée dans une zone donnée, ce qui permet d'obtenir le nombre de personnes infectées et injectent des personnes infectées dans d'autres zones en fonction des probabilités d'interactions entre les différents secteurs (Wesolowski *et al.,* 2015a). Cette étude présente également la particularité d'utiliser des données réelles, issues des statistiques d'appels au Pakistan, permettant d'établir des probabilités de déplacements entre les différents districts du pays et de calibrer un modèle gravitaire. Nous reviendrons sur ce type de données dans le chapitre 5.

Au-delà de ces petites "astuces", il existe d'autres types de modèles où les mobilités sont (ou peuvent être) prises en compte dans leur formulation. Nous pouvons citer notamment les modèles métapopulations, basés sur des matrices origines destinations, les modèles en réseaux et les modèles à base d'agents individu-centrés (Riley *et al.,* 2015).

2.1 Modèles compartimentaux métapopulations

Les modèles métapopulations ont d'abord été introduits en écologie, afin d'appréhender les différences démographiques dans des populations situées dans des zones différentes (Ball *et al.,* 2015 ; Levins, 1969). Dans le cadre des épidémies, il s'agit de modèles plus ou moins complexes (SIR, SEIR, etc.) qui permettent de prendre en compte des sous-groupes aux caractéristiques sérologiques différentes (Balcan *et al.,* 2009) ainsi que les déplacements des individus entre différentes zones, ou *patchs* (Arino et van den Driessche, 2003 ; Sattenspiel et Dietz, 1995 ; Wang et Zhao, 2004). Il n'est cependant pas possible de pouvoir suivre individuellement l'évolution des états sérologiques des individus (Riley *et al.,* 2015). Les déplacements des individus selon leur état sérologique (*e.g.* susceptibles, exposés, infectés ou guéris) peuvent être matérialisés par une matrice de flux origine-destination (Arino, 2005). Le travail d'Arino, initialement un modèle SEIR avec des taux de mortalité et de naissance (Arino et Van den Driessche, 2006) est simplifié ici en un modèle SIR :

$$\frac{dS_i}{dt} = -i\beta SI + \sum_{j=1}^{p} m_{ij} S_j - \sum_{j=1}^{p} m_{ji} S_i \tag{9}$$

$$\frac{dI_i}{dt} = i\beta SI - i\gamma I + \sum_{j=1}^{p} m_{ij} I_j - \sum_{j=1}^{p} m_{ji} I_i \tag{10}$$

$$\frac{dR_i}{dt} = i\gamma I + \sum_{j=1}^{p} m_{ij} R_j - \sum_{j=1}^{p} m_{ji} R_i \tag{11}$$

Où p est le nombre de zones (ou patch), et pour chaque zone i :

$\sum_{j=1}^{p} m_{ji}$ la somme des taux de déplacement entrant en i

$\sum_{j=1}^{p} m_{ij}$ la somme des taux de déplacement sortant de i (<1)

En sommant les équations, on obtient :

$$\frac{dN}{dt} = \sum_{i=1}^{p} m_{ij} N_j - \sum_{j=1}^{p} m_{ji} N_i \tag{12}$$

Ce qui signifie que la population totale est stable au cours du temps, mais que si la somme des flux entrants est différente des flux sortants, la population des différentes zones varie au cours du temps. Les zones ayant un solde positif verront leur population augmenter jusqu'à ce que les zones ayant un solde négatif soient totalement vidées de leur population. Il est également possible de moduler les déplacements en fonction de l'état des individus, par exemple réduire les interactions si dans le cadre d'une certaine épidémie les personnes infectées sont moins aptes à se déplacer.

Le premier modèle métapopulation déterministe spécifique à la dengue a été développé relativement récemment. Il part du principe que le nombre de moustiques infectés croit selon le nombre d'hôtes infectés résidant ou visitant cette zone et seuls les hôtes ne se déplaçant pas sont susceptibles d'être contaminés localement (Torre, 2009). Les mobilités ont par la suite été estimées à travers un modèle gravitaire et appliqué à trois *patchs* au Pérou (Sarzynska *et al.*, 2013). Ce modèle présente l'avantage d'impliquer une conservation de la population des différentes zones au cours du temps, mais sous-entend que les personnes susceptibles ne peuvent pas être contaminées lors d'un déplacement.

Ces modèles métapopulations sont relativement simples à mettre en place et fournissent en général des simulations assez convaincantes à de petites échelles (Arino, 2005 ; Ball *et al.*, 2015 ; Riley *et al.*, 2015 ; Sattenspiel et Dietz, 1995). Mais ils requièrent d'avoir des bonnes estimations des matrices de flux entre les différentes zones (Sattenspiel et Dietz, 1995), et dépendent donc de données mobilités de bonne qualité (Tizzoni *et al.*, 2015, 2014). À l'échelle locale, la trop grande simplification des processus et des dynamiques des populations résulte d'une moins bonne performance de cette approche (Ball *et al.*, 2015), due à l'homogénéité des sous-groupes (Frias-Martinez *et al.*, 2011). Ils n'échappent pas aux limites inhérentes des modèles compartimentaux, à savoir la dépendance à un grand nombre de paramètres difficilement estimables. Ils ne permettent pas non plus de tracer des individus isolés et tendent

à surestimer la propagation et les pics des épidémies (Ajelli *et al.,* 2010 ; Keeling *et al.,* 2010). Aussi, cette approche entraîne généralement un modèle ouvert où les populations dans les différentes zones changent au cours du temps. Or, les déplacements quotidiens en contexte urbain sont la plupart du temps pendulaires, c'est-à-dire qu'une personne part de chez elle le matin pour aller quelque part et finit par rentrer le soir. Nous proposons et discutons d'un modèle métapopulation fermé avec maintien des populations dans chacun des différents patchs (Annexe A).

2.2 Modèle épidémiologique en réseaux

L'approche par les modèles en réseaux permet de pallier certaines limites des modèles méta-populations et se base sur les théories des graphes (Pellis *et al.,* 2015). Ces systèmes sont composés de nœuds (ou vertex) reliés entre eux par des liens (edges). Cette approche relativement flexible, permet de représenter des systèmes complexes de différentes natures et composés d'éléments en interaction (Boccaletti *et al.,* 2006). D'un point de vue épidémiologique, les nœuds peuvent correspondre à des personnes prises individuellement, ou en groupe (foyers, villes, régions), et les liens au type de relation entre ces nœuds par exemple une probabilité de transmission, d'interaction ou de déplacements d'individus, variable au cours des simulations (Andersson, 1999 ; Danon *et al.,* 2011 ; Pellis *et al.,* 2015). Mais ces modèles n'ont été appliqués que très rarement à des maladies vectorielles. Un modèle en réseau individu-centré prenant en compte une population dont les liens sont hétérogènes a par exemple permis d'obtenir de meilleures estimations de l'épidémie de dengue survenue sur l'île de Pâques en 2002 (Favier *et al.,* 2005) que si ces contacts avaient été homogènes.

Nous pouvons aussi noter le modèle meta-population déterministe en réseaux développé par (Xue *et al.,* 2012) sur la Rift Valley Fever (RVF)[80]. L'aspect métapopulation n'intervient que pour décrire l'évolution sérologique des sous-populations dans les nœuds du réseau. La température et les précipitations affectent le développement des moustiques, et le vent leur dissémination. Les liens entre les nœuds du réseau correspondent aux interactions, soit les déplacements des populations d'hôtes et de vecteurs. Ce modèle qui requiert 21**n* équations, où *n* est le nombre de zones considérées, a pu reproduire la tendance globale de l'épidémie de RVF en 2010 en Afrique du Sud (Xue *et al.,* 2012). Cette approche semble pertinente dans la mesure où le rôle des mobilités humaines est bien inférieur à celle du bétail dans la propagation de cette maladie (Chevalier *et al.,* 2005 ; Xue, 2013), mais demeure difficilement transposable et suit approche uniquement mathématique.

Ces modèles qui suivent la théorie des graphes restent toutefois délicats à mettre en œuvre, car ils requièrent une connaissance relativement exhaustive des règles qui régissent

80. Une maladie vectorielle émergente transmise par des moustiques du genre *Aedes* et *Culex*, et dont les hôtes sont des mammifères, ce qui fait que cette maladie est au moins aussi complexe que la dengue.

chacun des nœuds et liens du réseau (Pellis *et al.,* 2015). Cela dit, comme nous le verrons dans la partie dédiée aux données (chapitre 5), les avancées en termes de connaissances, notamment sur les mobilités et les interactions entre individus, devraient tendre à lever ces verrous (Tizzoni *et al.,* 2015).

2.3 Modélisation à base d'agents

Avec l'augmentation des vitesses de calcul, il est maintenant possible de développer des modèles individu-centrés qui autorisent une meilleure description des sociétés et de l'environnement (Ajelli *et al.,* 2010 ; Ball *et al.,* 2015 ; Frias-Martinez *et al.,* 2011 ; Maneerat et Daudé, 2016 ; Pellis *et al.,* 2015). L'entité dont on cherche à évaluer le comportement *e.g.* des humains, des moustiques, ou des types d'environnements prend la forme d'un agent. Il s'agit d'objets informatiques qui partagent des processus et attributs communs (comme piquer les hôtes pour se nourrir pour les moustiques ou effectuer une activité pour les Hommes) mais dotés de caractéristiques individuelles spécifiques, définies et évoluant en fonction de différents critères et paramètres stochastiques ou déterministes (Bretagnolle *et al.,* 2006). Lorsqu'une seule classe d'agent est présente, on parle de modèle à base d'agents (ou Agent Based Model, ABM), et de modèles ou système multi-agent (SMA) lorsqu'il existe plusieurs classes d'agents. Ainsi, les humains peuvent être vu comme des agents différenciés, dotés de capacités de déplacements intrinsèques et variables par exemple selon leur zone de résidence, leur âge, leur statut socio-économique, l'heure de la journée ou l'activité qu'ils exercent.

Cette approche permet également d'apprécier les interactions entre différents types d'agents, notamment dans le cadre des SMA. Par exemple, un agent moustique va chercher un agent gîte afin de pouvoir y pondre ses œufs, et cette action de ponte ne pourra s'effectuer que si la capacité de mouvement du moustique est suffisante et s'il trouve un gîte mis en eau par l'action de l'Homme ou par les précipitations (Maneerat et Daudé, 2016 ; Misslin et Daudé, 2016). Ainsi, dans un contexte épidémique, les ABM permettent de capturer, à partir d'un grand nombre de simulations, le comportement individuel des agents. Les SMA sont de bons outils pour modéliser et simuler les dynamiques dans des milieux où les composantes individuelles sont très hétérogènes, et permettent d'ajouter des règles qualitatives et quantitatives à différents niveaux hiérarchiques (Bretagnolle *et al.,* 2006, Sanders, 2006). Ils sont donc *a priori* plus performants que les modèles métapopulations pour rendre compte de l'évolution d'une épidémie, de par une prise en compte des comportements individuels et des différentes interactions complexes non linéaires entre différents types d'agents (Ajelli *et al.,* 2010).

Application à la dengue

En se basant sur un modèle spatial et stochastique pour les dynamiques vectorielles (Otero *et al.,* 2008), (Barmak *et al.,* 2016) ont créé une petite ville artificielle sous forme de 20×20

blocs et composée de 100 individus plus ou moins mobiles, selon les scénarios. Ils partent du principe que les mobilités des personnes suivent deux règles : elles doivent être prédictibles (Song *et al.,* 2010b), et suivre une loi de distribution de Levy tronqué (González *et al.,* 2008). Leur étude montre que les mobilités de leurs agents influent sur la vitesse de propagation, sur les pics et la morphologie de l'épidémie simulée. Une étude a simulé les épidémies de dengue à partir d'un système multi-agent prenant en compte la mobilité, dans la ville de Cairns en Australie (Karl *et al.,* 2014). La ville a été divisée selon une grille de cellule de 30 m². Les mobilités routinières ont été définies de manière pseudo-aléatoire pour les populations en âge de travailler, tandis que les enfants ont été affectés à l'école la plus proche de leur domicile. Il en ressort que dans le cas de l'épidémie de dengue de 2008/2009, les mobilités humaines auraient joué un rôle relativement faible dans la flambée épidémique comparativement à l'impact de l'introduction d'une nouvelle souche de dengue (Karl *et al.,* 2014). L'influence des mobilités quotidiennes dans l'émergence de la dengue a également été abordée dans la ville de Santa-Cruz au Pérou (Garcia Lopez, 2010). À partir d'agents humains dont les déplacements ont été calibrés à partir d'enquêtes de terrains, cette étude a notamment montré l'importance des contaminations extra domiciliaires.

Notre rapide revue bibliographique tend donc à montrer que les SMA sont les outils de modélisation les plus à même de fournir les résultats les plus réalistes et permettent de mieux décrire l'évolution des épidémies, surtout lorsque les mobilités humaines sont prises en compte de manière individuelle ((Bian, 2004). Ils sont également particulièrement adaptés pour l'analyse numérique des résultats (Otero *et al.,* 2011), et aux tests d'hypothèses (Barmak *et al.,* 2016 ; Karl *et al.,* 2014). Ils sont aussi bien adaptés à l'analyse de systèmes complexes (Banos, 2013), particulièrement dans le contexte denguien (Daudé *et al.,* 2015 ; Karl *et al.,* 2014 ; Maneerat and Daudé, 2016). Mais pour définir les différents attributs qui peuvent constituer des agents mobiles, il convient tout d'abord de reprendre quelques concepts liés aux mobilités urbaines et d'étudier dans quelles mesures ils sont transférables aux SMA.

3 Ontologie d'un modèle de mobilité urbaine

3.1 Quels concepts utiliser pour analyser et modéliser les mobilités humaines ?

3.1.1 La motilité, ou les différents aspects sociaux à l'origine des mobilités

Le concept de *motilité*[81] *est relatif aux motivations et au potentiel de déplacement de chaque individu* (Kaufmann et Jemelin, 2004). Il peut être défini comme « l'ensemble des

81. Terme utilisé en médecine et en biologie pour définir la capacité d'une cellule ou d'un animal à se mouvoir. Nous pouvons noter que Kaufmann utilise aussi le terme de "réversibilité", pour définir le caractère définitif ou non d'une mobilité, emprunté lui aussi à la physique (Kaufmann *et al.*, 2004). Faut il interpréter cela comme une recherche de légitimité par rapport aux sciences considérées comme plus "dure" ?

caractéristiques personnelles qui permettent de se déplacer, c'est-à-dire les capacités physiques, le revenu, les aspirations à la sédentarité ou à la mobilité, les conditions sociales d'accès aux systèmes techniques de transport et de télécommunication existants, les connaissances acquises, comme la formation, le permis de conduire, l'anglais international pour voyager, etc. » (Kaufmann, 2012b).

Cette approche part donc du principe qu'il existe une forme de capital de mobilité variable en fonction des individus. Très discuté par (Borja *et al.,* 2015), nous préférons le terme de potentiel de déplacement que de capital de mobilité. En effet, en fonction de l'âge, du sexe, de la zone géographique et des attributs socio-économiques et culturels, une personne n'a pas la même possibilité ou propension à se déplacer. Un enfant n'effectuera bien souvent que de courts trajets, souvent en compagnie de ses parents. Des étudiants fréquenteront des lieux qui leurs sont relativement spécifiques (universités, quartiers étudiants, etc.). Une femme indienne sera vraisemblablement plus sédentaire que son mari (chapitres 2 et 7), ce qui n'est probablement pas aussi évident en France ou en Thaïlande. Des personnes aux revenus modestes habitant dans les zones périphériques feront probablement de plus longs trajets que des personnes vivant dans un centre-ville aisé et ne fréquenteront probablement pas le même genre d'endroit compte tenu des différentiels dans l'accès à certains types de loisirs, conditionnés bien souvent par le capital économique et culturel. Il existe en somme des personnes moins contraintes dans leurs déplacements que d'autres, que cela soit pour des raisons économiques, sociales, culturelles, personnelles ou d'âge.

La figure 40 présente quelques relations entre certains facteurs susceptibles d'influencer les mobilités urbaines d'un individu, à savoir :

- Des considérations socio-économiques, démographiques et culturelles.
    - Un capital économique (niveau de richesse) et culturel (niveau et type d'étude, opinions politiques, références culturelles, etc.).
    - Les membres du réseau social.
    - L'âge de l'individu et son état de santé physique.
- Des localisations dans la ville.
    - Le domicile.
    - L'activité principale (lieu de travail, école, etc.).
    - La répartition des activités commerciales et récréatives.

- Les domiciles des membres du réseau social.

- Les modes de transport susceptibles d'être utilisés pour se rendre dans chacune des localisations.

Schématiquement, la localisation du domicile d'un individu peut être influencée entre autres par celle de l'activité principale (plus ou moins loin), à laquelle s'ajoutent des considérations culturelles (choix du type de quartier), sociales (proximité physique de son réseau social) en fonction des possibilités économiques et du mode de transport (les zones rurales sont peu adaptées aux personnes sans permis de conduire par exemple). L'âge d'un individu influe sur son capital culturel, les réseaux d'amis, les modes de transports qu'il peut ou préfère emprunter, tout comme l'activité principale et la santé physique. Le réseau social peut se former dans le cadre d'une activité principale (collègues, camarades de classe) ou encore lors de la pratique d'activités récréatives et/ou de consommation (rencontrer des gens sur le marché, discuter sur Internet, etc.). De même, il peut permettre de trouver une activité principale (recommandations, cooptation). Les lieux fréquentés par la famille ou les amis pendant les heures non ouvrées influencent aussi le choix de fréquentation de tels lieux cela peut aller du choix d'un restaurant en fonction de la localisation et du capital économique des convives, à des sessions de shopping entre amis dans un centre commercial donné, en fonction des besoins et envies du moment, etc.

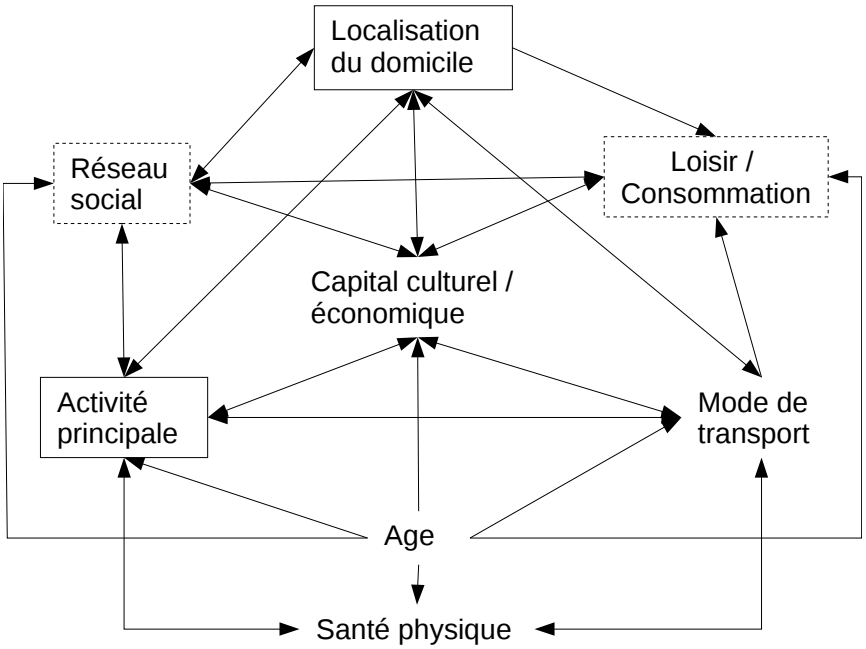


FIGURE 40 Représentation schématique de quelques éléments susceptibles d'influencer les mobilités individuelles en zone urbaine.

Mais cette figure est loin d'être exhaustive, car si on tente de répertorier tout ce qui

est susceptible d'influencer les comportements de mobilités individuels, il faudrait prendre en compte la personnalité des individus et les différents modes de vie qu'ils ont adoptés (plus ou moins consumériste, partisan de la marche, etc.). Sans oublier de considérer le genre de la personne, qui est fortement susceptible d'influencer les possibilités de déplacements (Hanson, 2010 ; Law, 1999), que cela soit par la pression sociale (les femmes ont un rôle plus ou moins sédentaire en fonction des familles et des cultures), ou la perception du danger (un quartier ou un mode de déplacement peut être perçu comme plus à risque, notamment vis-à-vis des agressions sexuelles en fonction des heures de la journée (Borker, 2017)). Sans compter l'accès au marché du travail et les écarts de salaire à diplôme et travail égal qui peuvent influencer leur capital économique. Il faudrait de plus envelopper le tout du contexte social de la région d'étude (la caste a-t-elle un impact sur les mobilités urbaines en Inde ?) ou encore les infrastructures de transport qui organisent la circulation dans les villes et rend des secteurs plus ou moins accessibles.

Cette logique de vouloir inclure le plus grand nombre de facteurs susceptibles d'influencer les mobilités individuelles peut paraître intéressante dans le cadre d'études sociologiques hautement qualitatives, car elle permettrait d'apprécier divers niveaux de déterminismes des mobilités. Mais elle parait délicate à mettre en œuvre dans un modèle à base d'agent, car il faudrait pour cela arriver à quantifier l'impact de chacun des éléments sur les autres, et pour chaque individu, en prenant en compte l'organisation et la structure des villes. Et en admettant que cela soit possible, cela compliquerait d'autant plus le système de la dengue, déjà plutôt complexe et dont les mobilités font partie.

Une autre approche pourrait donc être de mobiliser des éléments de l'esquisse de théorie de la "simplexité", proposé par Alain Berthoz (2009). Ce dernier part du constat « qu'entre cette complexité qui nous écrase et cette sur-simplification qui ne résout pas non plus les problèmes, il y a peut-être une troisième voie »[82]. La simplexité passerait ainsi par « des principes simplificateurs qui permettent de traiter des informations ou des situations, en tenant compte de l'expérience passée et en anticipant l'avenir. Ce ne sont ni des caricatures, ni des raccourcis ou des résumés. Ce sont de nouvelles façons de poser les problèmes, parfois au prix de quelques détours, pour arriver à des actions plus rapides, plus élégantes, plus efficaces » (Berthoz, 2009).

L'auteur prend notamment l'exemple de l'Amour, qui serait une solution simplexe apparue pour assurer la stabilité des relations, et maintenue car présentant alors un avantage sélectif. Il prend aussi l'exemple d'un joueur de tennis qui doit renvoyer une balle arrivant à toute allure. Le calcul de la trajectoire et le renvoi de la balle passent par des processus cérébraux et physiques extrêmement complexes, accélérés par l'expérience de l'individu et se traduisent par des actions observables relativement simples. Une interprétation de ce concept encore naissant pourrait

82. https://www.youtube.com/watch?v=2898sXZmEPQ

être de dire que finalement, de la même manière qu'un tennisman qui renvoie une balle est le résultat des calculs et ajustements complexes de son corps, les lieux qu'une personne fréquente ne seraient-ils pas l'expression observable de son potentiel de mobilité ? Sans toutefois être vraiment certain qu'il s'agisse là d'une forme de "simplexification" ou non[83], ceci permettrait de songer à une approche centrée sur les lieux fréquentés et les activités, utilisable dans le cadre d'une simulation à base d'agents.

3.1.2 L'espace d'activité et la time-geography

Car bien avant les théories de la complexité, de la simplexité ou de la motilité, fut développé à la fin des années 60 le concept d'espace d'activité (Hägerstrand, 1970), par le mouvement de la *time geography* de l'école de Lund[84], qui propose d'excellents outils pour décrire les mobilités quotidiennes. Il s'agit d'un concept individu-centré, basé sur les séquences d'activités effectuées dans des lieux à des horaires donnés. L'espace d'activité d'une personne est une portion de l'espace urbain total, défini comme la somme des lieux fréquentés plus ou moins régulièrement par cette personne (Horton et Reynolds, 1971). Dans chacun de ces lieux peut s'effectuer une activité donnée. Il peut s'agir du lieu de travail, d'un lieu de loisir, du domicile, etc. Il est aussi possible de représenter la localisation d'un individu dans un espace en trois dimensions (X, Y pour le plan géographique, et Z pour l'instant considéré, figure 41).

83. Est ce que l'utilisation d'un proxy relève de la simplexité ? Selon certains partisans de la simplexité, il s'agirait là encore d'une approche complexe car « Confrontés à la complexité du monde réel, les chercheurs se reconnaissant du domaine des systèmes complexes cherchent à comprendre les mécanismes expliquant le comportement des systèmes complexes environnementaux, biologiques ou sociaux. Ils se positionnent comme des observateurs percevant le système de façon *a priori* objective (par exemple via des capteurs artificiels de données, pendants des capteurs sensoriels naturels)" (Perrier, 2014).

84. http://www.hypergeo.eu/sp p.php?art c e540

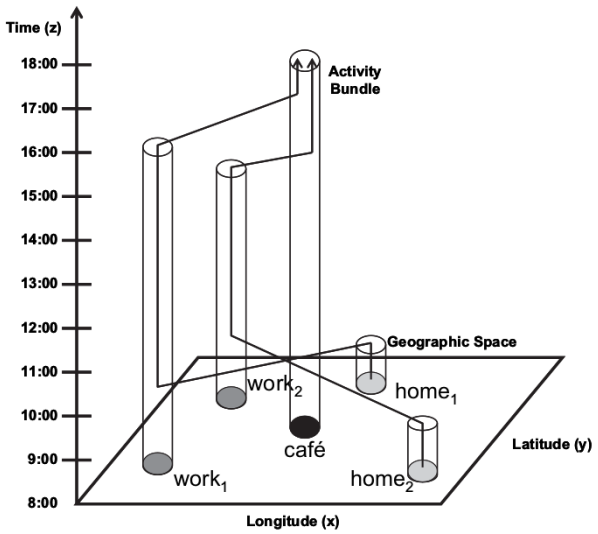


FIGURE 41 Exemple d'espace d'activité pour deux individus. Si la plupart du temps ces deux individus sont dans des lieux différents, ils fréquentent un même café au même moment. D'après Rainham *et al.* (2010).

Chacune des activités exercées par une personne est plus ou moins flexible dans le temps et dans l'espace (Hägerstrand, 1970).

- Les activités dites « figées » se déroulent à un moment donné et/ou dans un lieu précis.
  - Ainsi les activités « dormir » et « travailler » sont généralement fixes dans l'espace (domicile et lieu de travail) et suivent des plages horaires régulières (durant la nuit, ou durant les heures ouvrées). Une personne peut aussi par exemple suivre des cours de guitare dans la même école de musique un jour donné de la semaine.
  - En revanche, certaines activités peuvent se dérouler toujours aux mêmes plages horaires, mais être effectuées dans différents lieux. Par exemple si certaines personnes mangent toujours au même endroit (domicile, restaurant d'entreprise, etc.), d'autres peuvent fréquenter différents restaurants.
  - De même, une personne peut réaliser une activité toujours dans le même endroit, mais à des horaires différents. Une personne peut faire ces courses dans le même supermarché, mais pas forcément dans les mêmes tranches horaires.
- A contrario, les activités dîtes « flexibles » sont moins contraintes dans le temps et l'espace, moins planifiées. Il peut s'agir d'aller boire un verre en terrasse parce que le temps s'y prête ou bien de rencontrer un ami de passage en ville, dans un endroit choisi de manière opportuniste.

L'espace d'activité a généralement plusieurs centres (Axhausen *et al.,* 2002), l'un étant le domicile, l'autre le lieu où s'effectue l'activité principale en journée, avec des activités autour et entre chacun de ces lieux (Golledge et Stimson, 1996). Ce concept permet également de prendre en compte les interactions sociales, lorsque deux personnes se retrouvent au même endroit, au même moment (figure 41). On parle alors de convergence des trajectoires (Rainham *et al.,* 2010).

Comme le souligne Sonia Chardonnel (2007), l'espace d'activité permet d'avoir une approche plutôt holistique, en accord avec la complexité, et qui prend en compte le contexte du déplacement. Il est alors possible de développer des analyses longitudinales qui permettent de suivre ce qui régule l'utilisation du temps et de l'espace, tout en proposant des représentations graphiques synthétiques. Cette approche est également un bon outil utilisé en géographie de la santé pour mesurer certains risques liés à des expositions dans différents environnements (Perchoux *et al.,* 2013).

L'espace d'activité englobe donc les notions d'espaces, de temps, de fréquences, d'activités, avec différents niveaux de flexibilités. Ce concept semble parfaitement adapté pour traiter des mobilités quotidiennes. D'un point de vue qualitatif, l'enquêteur peut reconstruire l'espace d'activité des personnes interrogées sous forme de carnet de bord, et l'analyser selon des considérations relatives à ses déplacements. Cette approche permet de synthétiser des informations d'un individu où chaque activité est associée à une localisation et à une heure de la journée, autorisant alors l'emploi de méthodes plus quantitatives. Le concept d'espace d'activité paraît donc bien adapté à la modélisation à base d'agents (Banos, 2013 ; Banos *et al.,* 2005).

3.2 Vers une formulation abstraite d'un modèle de mobilité

3.2.1 Quels attributs pour un agent mobile dans le système de la dengue ?

Associé au concept d'espace d'activité, chaque agent du système peut se voir attribuer une série d'activités qu'il exerce dans différents lieux à différents horaires, avec un niveau de flexibilité variable et défini par le modélisateur. Par exemple, Karl *et al.,* (2014) ont utilisé cette approche en définissant des activités simples que peuvent réaliser leurs agents à différents endroits et à différentes plages horaires. Chaque individu s'est vu attribuer un lieu de domicile et un lieu de travail, où ils sont respectivement présents en semaine entre 15 h et 9 h et entre 9 h et 15 h. Ils ont également ajouté un peu de stochasticité en attribuant des lieux pris au hasard pour simuler d'autres types d'activités (Karl *et al.,* 2014).

Cette approche revient à créer pour chaque individu un agenda simple, avec d'une part des activités assez figées (être à son domicile ou à son travail), et d'autres par des activités flexibles, définies aléatoirement. Dans ce cas, l'espace d'activité créé repose sur des hypothèses

assez fortes, mais non vérifiables et la définition de la localisation des activités de chacun n'est pas réaliste (à part peut être pour les enfants dont le lieu de travail est l'école la plus proche du domicile).

En nous inspirant de cette approche, nous allons maintenant discuter des différents attributs qui nous paraissent importants à implémenter dans un modèle de mobilité urbaine centré sur l'espace d'activité des individus, et en lien avec la transmission des épidémies. Nous verrons dans les chapitres suivants quels éléments sont effectivement modélisables.

Définir des catégories d'activités raisonnables

Nous considérons que le fait qu'un agent soit à son domicile est une activité à part entière. Un agent devra aussi avoir une activité principale qui peut aussi se dérouler à son domicile. Il s'agira d'un lieu de travail, ou d'un lieu d'éducation pour les plus jeunes.

Les zones qui brassent un grand nombre d'individus jouent un rôle clé dans la propagation des maladies infectieuses. Ces lieux, construits à différents moments de la journée par la présence d'individus de groupes sociaux plus ou moins hétérogènes peuvent être de différentes natures. Il peut s'agir de lieux liés aux transports, comme les gares et les stations de métro, ou des lieux de consommation, comme les marchés, les centres commerciaux et différents quartiers tournés vers les sorties et les commerces. Les lieux de cultes peuvent rassembler un grand nombre de croyants pratiquants, certains d'entre eux pouvant attirer aussi des touristes. Les parcs urbains sont aussi à prendre en compte, car ils peuvent être plus ou moins fréquentés et forment aussi un habitat pour les moustiques. Enfin, notre étude se déroulant en contexte épidémique, il convient également de prendre en compte les hôpitaux et les différentes cliniques.

Le réseau social

Le réseau social d'un individu est susceptible d'influencer les lieux qu'il fréquente, que cela soit pour se rendre dans un espace public, marchant ou non, ou privé (Cho *et al.,* 2011). Il conviendrait donc d'incorporer ce paramètre dans notre étude car la proximité entre les personnes influence la propagation des épidémies (Andersson, 1999 ; Frias-Martinez *et al.,* 2011 ; Perkins *et al.,* 2014 ; Stoddard *et al.,* 2013). L'approche des mobilités quotidiennes par l'espace d'activité permet une simplification partielle des relations sociales entre les agents d'un modèle. Il n'est en effet pas nécessaire de définir des liens sociaux entre les individus d'un autre foyer, ces derniers étant décrits de manière implicite, par exemple lorsque deux agents exercent la même activité au même endroit et aux mêmes plages horaires. Néanmoins, il pourrait être intéressant d'ajouter des contraintes relationnelles entre les agents.

D'un point de vue formalisation, chaque agent aurait donc des liens avec d'autres, qu'il s'agisse de membre de la famille, d'amis ou de collègue. Ils partageraient alors plus ou moins

d'activités en commun activité principale, domicile, ou des sorties par exemple et à des fréquences temporelles variables. Dans le cadre de l'espace d'activité, cela reviendrait pour un individu à ajouter à une activité donnée, une liste d'agents susceptibles d'être au même endroit au même moment.

Prise en compte du mode de transport

L'espace d'activité d'un individu peut être vu comme une sorte d'archipel dont les îles (activités dans un lieu) sont reliées entre elles par des déplacements. Le risque vis-à-vis de la dengue n'est pas le même lorsqu'on marche dans la rue, que l'on prend un bus ouvert ou lorsque l'on est dans sa voiture. Le mode de transport utilisé peut aussi contraindre les lieux qu'une personne peut fréquenter et influencer la séquence des activités quotidienne d'une personne. Les larges avancées dans la modélisation des transports urbains (Ortúzar S. et Willumsen, 2011), associées à des informations sur la fréquentation des transports en commun et l'usage des voitures peuvent permettre de définir un mode de déplacement à chaque agent, conditionné éventuellement par l'activité qu'il compte exercer, même si cela reste loin d'être évident.

Dans le cadre d'une simulation à base d'agent, si le pas de temps entre deux itérations est suffisamment élevé, par exemple supérieur à 1h, nous pouvons considérer que l'agent va directement d'un lieu où s'effectue une activité A, à une activité B. Si le pas de temps est inférieur à une heure, dans ce cas l'individu se trouvera quelque part entre les lieux A et B, selon une méthode qu'il conviendra de définir.

Une approche SEIR, appliquées aux 4 souches du virus

Chaque agent se doit d'avoir un état vis-à-vis de chacune des 4 souches du virus de la dengue. Il peut être susceptible, exposé (piqué, mais pas encore contagieux), infecté puis guéri ou décédé. Certaines études montrent qu'une personne ayant déjà subi une primo infection a plus de chance de développer des cas sévères de la dengue en cas de nouvelle contamination par d'autres souches (Halstead *et al.,* 1969). Ce paramètre serait intéressant à prendre en compte.

Prise en compte des cas asymptomatiques

En cas d'infection, il convient aussi de distinguer les personnes chez qui la maladie s'exprime de celles chez qui elle s'exprime moins ou pas du tout (ten Bosch *et al.,* 2018). En effet, un individu infecté mais asymptomatique ne devrait pas changer ses habitudes de mobilités, contrairement à une personne malade, qui serait alors soit chez elle et soignée par ses proches, soit à l'hôpital ou dans une clinique.

Importance de l'âge

Le risque de contracter la dengue n'est pas le même en fonction de l'âge, car les plus jeunes, en général plus naïfs vis-à-vis de la maladie, ont plus de chance d'être contaminés, que cela soit en Inde (Angel et al., 2017 ; Cecilia, 2014 ; Mishra *et al.,* 2016) ou en Thaïlande (Limkittikul *et al.,* 2014). Il convient donc d'associer un âge à chacun des agents, ce qui permet de conditionner les probabilités d'avoir déjà été infecté, en se basant notamment sur la littérature. Les personnes âgées sont aussi plus susceptibles de développer des complications en cas d'infection.

Prendre en compte le genre ?

Ajouter un attribut de type "genre" à un agent peut être pertinent au regard de ce que nous avons montré dans les parties précédentes sur les contraintes de mobilité. Tout dépendra des données mobilités que nous arriverons à recueillir, car nous ne nous risquerons pas à faire des hypothèses hasardeuses.

Un agent touriste ?

Étant donné que Bangkok est une ville très attractive pour les touristes, et qu'environ un tiers des cas de dengues importés en France proviennent d'Asie du Sud-Est, créer un « agent touriste », plus naïf vis-à-vis de la maladie et visitant des lieux spécifiques parait aussi être une piste intéressante.

3.2.2 Attributs d'un agent mobile idéal

À partir de ces considérations, nous allons poser les attributs qu'il conviendrait d'associer à un agent mobile en contexte épidémique.

- ***Paramètres socio-démographiques et épidémiologiques de l'agent :***
  - Sérologie vis-à-vis des quatre souches de dengue
    - Susceptible, Exposé, Infecté, ou Guéri
    - Si infecté, est-il asymptomatique ?
  - L'âge, qui module :
    - L'activité principale (école, université, travail, retraité, autre)
    - Le niveau d'immunité vis-à-vis des différentes souches de dengue
  - Un réseau social
    - Personnes partageant le même foyer (~famille)
    - Amis ou collègues
      - Localisation des lieux partagés en commun (domicile ou autre)
      - Fréquence et durée de rencontre
  - Le genre, qui peut influencer la composition de l'espace d'activité
  - Résident ou touriste ?

        - Influence l'espace d'activité (les lieux fréquentés et la localisation du domicile, ou "hôtel" pour un touriste)
        - L'agent sera plus ou moins naïf vis-à-vis de la dengue
- ***Espace d'activité de l'agent :***
    - Définis par différentes activités localisées :
        - Domicile, figé dans l'espace
        - Activité principale très figée dans le temps et dans l'espace
        - D'autres activités, parfois plus facultatives ou épisodiques comme les lieux de consommation et de restauration, de sorties, de loisirs, les lieux de cultes, les parcs ou encore le domicile de membres du réseau social
    - Un nombre de lieux correspondant à chacune de ces activités
    - Des distributions de fréquence et de durée de visite
    - Le transport utilisé pour se rendre à une activité

3.2.3 Processus de modélisation :

À partir de ces caractéristiques, nous pouvons proposer un plan très schématique de génération d'individus mobiles :

1. Créer des foyers, composés d'agents :
    - Affecter un domicile, en prenant en compte la répartition de la population totale.
    - Définir les caractéristiques des agents du foyer (nombre, âge, sexe, situation professionnelle) associées à la zone de domicile, d'après des caractéristiques de recensement (Chapuis *et al.,* 2018).
    - Définir le niveau d'immunité de l'agent selon l'âge, à partir de la littérature.
2. Définir les relations sociales de tous les agents.
    - Nombre d'amis.
    - Nature des relations.
3. Générer un agenda (ou séquences temporelles d'activités) pour chaque agent. En admettant que nous ayons pu collecter suffisamment de données sur les espaces d'activité d'un échantillon assez représentatif, il devrait être possible de définir des fréquences de distribution et de tirer plus ou moins aléatoirement pour un agent :
    - Un nombre d'activités et de lieux fréquentés.
    - Des types d'activités effectuées.
    - Des heures, durées et fréquences de réalisation de ces activités.
    - Du niveau de flexibilité temporel de ces activités.
4. Contraindre les horaires de réalisation de certaines activités en prenant en compte le réseau social :

- Ajouter une notion de synchronicité entre des agents "amis" dans la réalisation de certaines activités, qui est
- Modulée selon la proximité sociale des agents (plus ou moins amis).

5. Attribuer une localisation à chacun de ces lieux, qui peut passer par
   - Des modèles géographiques (gravitaires ou radiatifs ou autre).
   - Des données de fréquentation et d'attractivité des différentes zones de la ville.
   - Des tendances générales de déplacements
6. Faire se déplacer les agents
   - Selon leur agenda spatialisé
   - Selon leur sérologie

Bien entendu, il ne s'agit là encore que d'une formalisation assez naïve et idéaliste, car un grand nombre de contraintes restent à dépasser. Il convient en effet de trouver des données adaptées qui permettent à la fois d'observer les mobilités individuelles, collectives et de définir des espaces activités d'un grand échantillon d'individus. Arriver à prendre en compte et créer un réseau social pour un agent serait évidemment un plus, mais comme nous l'avons vu, cet aspect peut être implicite au regard du concept d'espace d'activité. Nous discuterons de ces aspects dans le reste de la thèse, et il est fort probable que nous soyons obligés de simplifier notre schéma conceptuel au gré des données collectées dans les différentes zones d'études.

Synthèse

Nous sommes revenus dans ce chapitre sur les principes de la modélisation en épidémiologie, en insistant sur les méthodes qui peuvent prendre en compte les mobilités urbaines. Il ressort que les SMA, de par leur grande flexibilité et leur formalisation nécessairement réductrice mais néanmoins concrète, semblent être très appropriées dans le cadre du système complexe qu'est la dengue.

La mobilisation du concept individu-centré de l'espace d'activité permet de synthétiser de manière efficace un grand nombre de notions relatives aux mobilités quotidiennes individuelles. D'un point de vue qualitatif, il permet de résumer simplement des différences de potentiels de déplacements entre individus, tandis qu'une approche plus quantitative autorise son implémentation dans un système à base d'agent.

Nous avons proposé une ontologie de modèle, mais qui passe par la collecte des données adaptées, objet de la prochaine partie de ce travail.

Partie B:

Les traces numériques : de «nouvelles» données pour aborder les mobilités

Jusqu'à récemment, il n'y avait pas énormément de cas de figure lorsqu'il s'agissait d'utiliser des données pour étudier des mobilités urbaines. On pouvait se baser sur des études institutionnelles[85] réalisées sur un grand échantillon et prenant en compte des mobilités, pour peu que ces dernières soient récentes et disponibles dans la zone d'étude et que notre approche soit plutôt quantitative. On pense ici typiquement à des enquêtes domicile-travail, ménage-déplacement[86] (Banos, 2013 ; Commenges, 2013 ; Guvry *et al.,* 2008) ou encore sur l'utilisation du temps[87]. Ces études sont en général jugées plutôt fiables de par une méthodologie d'échantillonnage éprouvée, contiennent des informations socio-économiques et démographiques et servent souvent de bases pour qualifier la qualité d'un autre jeu de données (*e.g.* Calabrese *et al.,* 2011a ; Lenormand *et al.,* 2014).

En cas d'absence de ce type d'étude, ou choix délibéré d'effectuer une enquête plus qualitative ou hybride, il est toujours possible d'aller sur le terrain et de récolter des données en procédant à des interviews individuelles, ou effectuer des comptages dans des lieux stratégiques. Beaucoup d'études se basent sur l'analyse des traces issues de GPS fournis à des enquêtés consentant (*e.g.* Chevalier, 2018 ; Drevon *et al.,* 2014 ; Nguyen-Luong, 2012 ; Vazquez-Prokopec *et al.,* 2009 ; ou encore Shen et Stopher, 2014, pour une revue). Il s'agit notamment de combler les lacunes des entretiens classiques, par exemple l'oubli de mentionner la fréquentation d'un lieu donné, en étudiant généralement les comportements de mobilités d'une catégorie d'individus (enfants (Depeau *et al.,* 2017 ; Loebach et Gilliland, 2016), personnes âgées (Hirsch *et al.,* 2014 ; Shoval *et al.,* 2010)). Si ce type d'études permet de recueillir des informations géographiques très précises, elles ne peuvent cependant pas être réalisées sur un grand

85. Financée au moins en partie par des institutions publiques.

86. voir aussi http://www.stat st ques.deve oppement durab e.gouv.fr/repondre enquetes/enquete mob te personnes 2018 2019.htm

87. Time use survey, ou Time Budget survey

échantillon essentiellement pour des raisons de coût ni sur de trop longues périodes car assez intrusives et contraignantes pour l'enquêté. Les approches par enquêtes qualitatives (ou quantitatives sur un petit échantillon) n'ont pas vocation à être extrapolées, mais elles permettent toutefois d'acquérir une bonne connaissance du terrain et de l'échantillon, et surtout de contextualiser des résultats plus quantitatifs.

Nous qualifierons ici de « classique » ces méthodes de collectes de données qui passent par des enquêtes *in situ*. Et aussi pertinentes soient-elles, elles restent relativement coûteuses en temps et/ou mains d'œuvres, et souvent restreintes à une petite zone géographique ou à des échelles spatiales et/ou temporelles inappropriées selon le niveau de précision requis par l'étude.

Le développement technologique dans les domaines des communications et de l'informatique depuis les années 1990 et surtout après les années 2000, a vu la généralisation de certaines pratiques, comme l'usage des téléphones portables et depuis peu des réseaux sociaux, au gré d'un meilleur accès à des réseaux de communication (meilleurs débits, baisse des prix, bond technologique des téléphones portables, etc.). Ces comportements en ligne et en itinérance impliquent une production de données ou traces numériques parfois géolocalisées et donc potentiellement utilisables dans l'analyse des mobilités individuelles ou collectives. Mais contrairement aux données plus « classiques », qui présentent la caractéristique que leur création et collecte ont pour objectif explicite l'analyse des mobilités, ce n'est pas le cas de la plupart des traces numériques, dont les conditions et objectifs de création sont *a priori* déconnectés de leur usage et de leur analyse. Ainsi, si ces grands volumes de données permettent d'effectuer des analyses spatiales jusqu'alors inédites ou encore de calibrer des modèles de déplacements, elles sont intrinsèquement associées à des considérations éthiques, du fait du consentement probablement trop tacite des utilisateurs qui pourrait se rapporter à une sorte de pacte faustien inconscient entre ces derniers et les plateformes Internet, et des enjeux sur la vie privée qui en découlent.

Le premier chapitre de cette partie abordera les mécanismes de création et de collecte de ces traces numériques, surtout celles qui sont associées à des informations sur la géolocalisation. Si ces données collectées par divers acteurs émanent d'une action plus ou moins consciente d'un utilisateur d'un service de communication *lambda,* les intentions initiales de ce dernier sont quelque peu détournées. Il s'agira en effet ici d'expliquer comment le simple fait de passer un coup de fil ou d'envoyer un message géolocalisé pour par exemple signaler sa présence dans un lieu donné aux membres de son réseau social peut aussi permettre à des tiers d'obtenir des informations sur des comportements de mobilités. Ces données à forts potentiels économique et technique sont ainsi sujettes à de possibles dérives. Nous aborderons donc dans cette section certains aspects d'ordre éthique et juridique.

Le chapitre 5 est quant à lui un état de l'art sur l'utilisation de ces nouvelles données dans les études sur les mobilités, et se focalise principalement sur les mobilités urbaines et sur ce qu'elles apportent à l'analyse spatiale et à la modélisation.

Les données individuelles issues de *Twitter* que nous avons enregistrées à Delhi et Bangkok et les données agrégées provenant de *Facebook* à Bangkok seront présentées dans le chapitre 6. Nous y détaillerons les méthodes de collecte et les prétraitements réalisés, prérequis avant leur application dans nos zones d'études (parties C & D).

Chapitre IV: L'abondance des traces numériques géolocalisées

Nous allons présenter ici les principaux modes de création et de collecte de données en ligne. Nous nous focaliserons surtout sur celles qui permettent une géolocalisation, mais nous aborderons aussi le sujet de manière plus globale, car les aspects techniques et éthiques qui en découlent touchent un large pan de notre société actuelle et incluent des considérations d'ordre géographiques, notamment de traçage d'individus. Il convient au préalable de définir quelques termes.

Quelques définitions

- ***Trace numérique***

Une trace se définit comme la « marque physique, matérielle laissée par quelqu'un ou quelque chose sur, en quelqu'un ou quelque chose » [88]. Une trace numérique est par extension une donnée de nature binaire, enregistrée par une machine, notamment lors de protocoles de connexion (Merzeau, 2009). Ces données enregistrées permettent, en fonction des éléments qui les constituent de tirer des conclusions variables. Comme le suggère Gérard Berry dans un de ses cours au Collège de France, il est important de distinguer « donnée », « information » et « connaissance ». Une donnée est simplement une valeur, par exemple 42, tandis qu'une information est la nature ou le type de la donnée, par exemple une température en Celsius. La connaissance est ce qui permet d'interpréter cette information, dans ce cas nous savons qu'une température de 42 °C n'est pas confortable pour l'être humain. Une trace numérique géolocalisée est une information contenant des coordonnées géographiques.

- ***Données personnelles***

L'article 2 de la loi 78-17 informatique et liberté du 6 janvier 1978 modifiée qualifie de donnée à caractère personnel « toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne » (Droit Français, 1978). Ce que la CNIL résume comme « Toute information identifiant directement ou indirectement une personne physique » (Commission Nationale Informatique et Liberté, 2004).

88. http://www.cnrt .fr/defin t on/trace

Le Correspondant Informatique et Liberté (CIL) du CNRS précise également que « Les technologies de l'information et de la communication génèrent de nombreuses données personnelles (un appel passé par un téléphone portable, une connexion à Internet) et aussi des "traces informatiques" facilement exploitables grâce aux progrès des logiciels, notamment les moteurs de recherche » (Conseiller Informatique et Liberté - CNRS, 2016). Une donnée personnelle est donc une donnée que l'on peut rattacher à un identifiant qui désigne de manière plus ou moins équivoque une personne physique, notamment s'il est possible d'effectuer des recoupements entre des bases de données.

Ainsi, le numéro de sécurité sociale est relié à une personne physique et permet aux autorités compétentes d'accéder à une partie de son état civil (âge, sexe, adresse) et à son dossier de santé, ce qui en fait une donnée personnelle par excellence. Tout comme le numéro fiscal qui relie l'état civil d'une personne à ses revenus déclarés. Ces bases de données personnelles sont inhérentes au fonctionnement d'un état pourvu d'un système de sécurité sociale et pratiquant un système de redistribution de richesse par l'intermédiaire de la levée d'un impôt.

Certains acteurs majeurs du web peuvent offrir à la même personne des services de messageries, de recherche en ligne, de réseaux sociaux ou encore d'aide au déplacement. Ces derniers peuvent donc techniquement connaître le nom de cette personne, ses contacts, ses centres d'intérêt, ses habitudes de mobilités et les horaires de fréquentation de certains lieux, et autres données intimes[89]. À partir de ces données personnelles, il est alors possible d'effectuer des profilages et des classifications d'individus, pour l'amélioration des services, ou pour leur valeur marchande. Même si l'usage des données personnelles par des tiers est encadré par la loi française, notamment par la Commission Nationale Informatique et Liberté (CNIL) (section 4).

L'utilisation de ces données individuelles par des tiers revêt alors un caractère éthique, défini par le TLFI[90] comme la « science qui traite des principes régulateurs de l'action et de la conduite morale ». Laurence Devillers propose une définition inspirée des travaux de Paul Ricoeur, où « l'éthique est une discipline philosophique pratique et normative, visant à indiquer comment les êtres humains doivent se comporter, agir et être, entre eux et envers ceux qui les entourent. L'éthique propose souvent des compromis afin de concilier règles morales, désirs et capacités » (Devillers, 2017). La conception de l'éthique vis-à-vis de la collecte et des traitements des données peut donc varier selon les individus et cultures. Nous y reviendrons plus tard, mais les débats autour de l'éthique des données personnelles en France et en Europe tournent surtout autour des questions de consentement de la personne concernée, de leur sécurisation et du respect de la vie privée.

Quatre types de données personnelles peuvent être distinguées : celles liées à la technique et aux protocoles de connexion, celles déclaratives, ou encore les données navigationnelles et comportementales (Ertzscheid, 2013). Les données liées à la technique sont toutes les traces

89. https://www.courr er nternat ona .com/art c e/vous etes prets vo c tout ce que facebook et goog e savent sur vous

90. Trésors de la Langue Française Informatisé http://at f.at f.fr/t f.htm

enregistrées lors de l'utilisation des services, pour des raisons protocolaires. Il peut s'agir des heures auxquelles un téléphone sans fil s'est connecté à une antenne relais ou de l'adresse IP d'une personne qui visite un site Internet . Les données personnelles dîtes « déclaratives » sont fournies directement par un utilisateur. Il peut s'agir de données liées à la création d'un compte en général d'un identifiant, d'une date de naissance, d'un sexe, d'un nom, d'un prénom ou encore d'un courriel ou d'un numéro de téléphone ou d'informations (messages, photos) postées sur un réseau social. Les données navigationnelles sont les données collectées par les moteurs de recherches et les données comportementales s'appuient sur l'historique de navigation (Ertzscheid, 2013). Bien qu'une telle classification est pertinente, nous présenterons dans les sections suivantes les modes de création des données lors des usages de tous les jours, que cela soit en naviguant sur Internet, en téléphonant ou en prenant le métro. Nous distinguerons les données « passives », liées aux fonctionnements techniques et protocolaires et qui sont transparentes pour l'utilisateur, des données « actives », dont la création dépend de l'activité *a priori* consciente de l'Internaute, comme lors de la publication d'un commentaire sur un réseau social.

1 Des données protocolaires...

1.1 ...Générées lors de l'utilisation d'un téléphone mobile...

« Un cellulaire allumé, la limousine est repérée », Claude M'Barali (1998)

Les technologies téléphoniques ont bien évolué depuis l'époque où les redirections des appels (commutation) étaient effectuées manuellement par des opérateurs. Entre temps un croque-mort lassé de voir sa clientèle potentielle systématiquement redirigée vers son concurrent via l'intermède de sa femme, opératrice, breveta le premier commutateur automatique en 1889. Ce fut les débuts balbutiant des systèmes de routages analogiques et l'émergence du réseau téléphonique commuté public (RTCP), surclassé maintenant par le tout numérique et le multiplexage[91]. Les technologies utilisées de nos jours en télécommunications sont extrêmement nombreuses et relativement complexes, nous nous contenterons ici d'expliquer très simplement les méthodes d'accès au réseau des technologies les plus courantes, communément appelées 2, 3 et 4G[92]. Même si ces technologies sont très différentes en termes de bandes passantes, de fréquences utilisées et d'organisations internes, les principes généraux et les architectures schématiques restent relativement similaires.

Une communication téléphonique mobile obéit à quelques règles de base. Tout d'abord il faut qu'elle soit au moins bidirectionnelle, c'est-à-dire que l'appareil puisse à la fois envoyer et recevoir des données. Le mobile doit donc être localisable, même inactif (pour pouvoir recevoir des appels) et la communication maintenue lors de déplacement (Servin, 2006). Schématiquement, un appel téléphonique consiste en deux individus reconnus par des identifiants qui communiquent entre eux via un réseau téléphonique, composé notamment d'antennes relais. L'identification d'un appareil mobile (ou Mobile Station, MS) se fait via la carte SIM (Subscriber Identity Module) qui contient le numéro IMSI (International Mobile Subscriber Idendity). Cet identifiant unique se compose du code du pays d'origine, de l'identifiant de l'opérateur téléphonique, du numéro de l'abonné (MSIN pour Mobile Subscriber Identification Number) répertorié dans la base de données centrale (HLR, Home Location Register).

Les points d'accès au réseau téléphonique sont les antennes relais (BTS pour Base Tranceiver Station dans un réseau GSM[93]) qui gère le trafic radio avec le mobile. Elles sont réparties de manière très inégale sur les territoires car elles ne peuvent assurer qu'un nombre limité de communications simultanées, ce qui signifie que pour assurer un service de qualité, leur densité doit être relativement proportionnelle à la population. Ainsi, en zone urbaine elles sont assez rapprochées, de l'ordre de la centaine de mètres, tandis qu'elles sont plus dispersées

91. Le multiplexage est une technique permettant de faire passer plusieurs communications sur un même canal de transmission (Servin, 2006).

92. Respectivement le réseau GSM (Groupe Spécial Mobile, devenu Global System for Mobile Communications), UMTS (Universal Mobile Telecommunications System) et LTE (Long Term Evolution).

93. L'équivalent du BTS pour le réseau 3G est l'antenne node B, et enodeB pour le réseau 4G

en zone rurale (de quelques kilomètres à dizaines de kilomètres). La couverture spatiale d'une antenne est appelée cellule, représentée par un polygone de Voronoi (figure 42). Les BTS sont regroupés en BSS (base station subsystem) formant des réseaux de zones de plusieurs cellules (d'une dizaine à une centaine, suivant les besoins locaux), les « Location Area Network », caractérisées par un numéro LAC (Location Area Code).

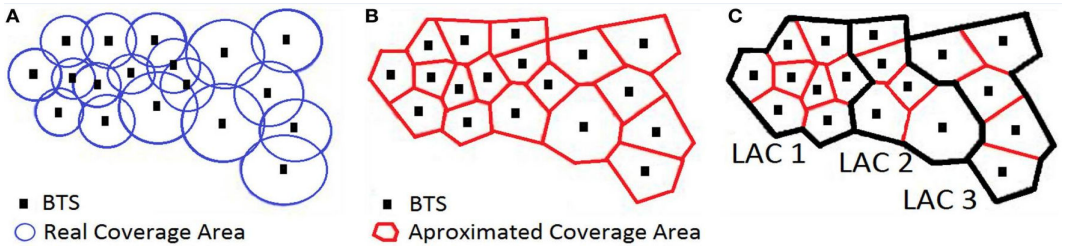


FIGURE 42 Organisation schématique du réseau d'antennes téléphonique. La zone couverte par chaque antenne forme dans des conditions optimales un cercle de portée variable (A). L'emprise spatiale des antennes peut être approximée par l'utilisation de polygones de Voronoi (B). Les antennes sont ensuite regroupées en LAC, ou sous-réseaux (C). D'après (Oliver *et al.*, 2015).

Les BTS émettent en permanence des informations générales sur leurs cellules par voie hertzienne sur un canal de signalisation. Un mobile recherche alors ces signaux et se connecte au BTS dont le niveau de réception est le plus élevé et s'y inscrit. Le mobile se voit attribuer un canal de communication aléatoire dans cette cellule. Les données de l'utilisateur contenu dans la HLR sont ensuite recopiées dans la base de données locale des visiteurs de la cellule (VLR, Visitor Location Register), la localisation est enregistrée et un nouvel identifiant éphémère est attribué, le TMSI (Temporary Mobile Station Identity) qui sera échangé sur le réseau. Lorsqu'un appel est émis, le réseau cherche la localisation (cellule) de la personne à joindre via le numéro de téléphone (MSIN) distant enregistré dans le HLR et la connexion s'établit.

Au final, lors d'un appel ou de l'envoi d'un SMS, un BTS enregistre à minima dans les statistiques d'appels (Call Duration Report ou CDR) l'identifiant de l'émetteur, celui du récepteur, l'horodatage de l'appel, le type d'appel (voix, SMS) et ceci pour des raisons de facturation. Si l'utilisateur se déplace, comme la localisation des BTS est connue, il est possible de le suivre de cellule en cellule lorsque celui-ci appelle ou est appelé. De manière plus passive, le téléphone de l'utilisateur communique plusieurs fois par seconde avec le réseau pour vérifier la bonne inscription à la bonne antenne (Michael et Clarke, 2013). Ainsi, les informations de changement de zone (ou de LAC) sont également enregistrées, même si la personne n'a pas téléphoné, ceci pour assurer la continuité du service[94]. Ces données sont gérées par les opérateurs téléphoniques. Les informations du téléphone et de la carte SIM relatives aux antennes relais détectées (nom, heure, et puissance du signal) ne sont *a priori* accessibles à de tierces personnes que si l'utilisateur installe des applications externes et leur autorise l'accès.

94. Ou « handover »

Les CDR, enregistrés à la base pour des raisons pratiques (facturation et continuité du service) ont vu leur utilisation détournée à des fins de recherche partir du milieu des années 2000, notamment dans le cadre d'études sur flux de mobilités humaines (voir chapitre 5). La figure 43 ci-dessous montre un exemple théorique d'une personne ayant passé quatre appels dans une journée. Comme cette personne se déplace, elle a été localisée près de quatre antennes. Il est alors possible de reconstituer ces déplacements, avec une précision qui varie en fonction de l'agrégation temporelle (plage de temps) et spatiale (regroupement de BTS). D'un point de vue collectif, l'agrégation du nombre de personnes localisées à un BTS dans le temps permet d'observer la pulsation urbaine. Nous reviendrons sur ces aspects mobilité individuelle et dynamiques urbaines dans le chapitre suivant.

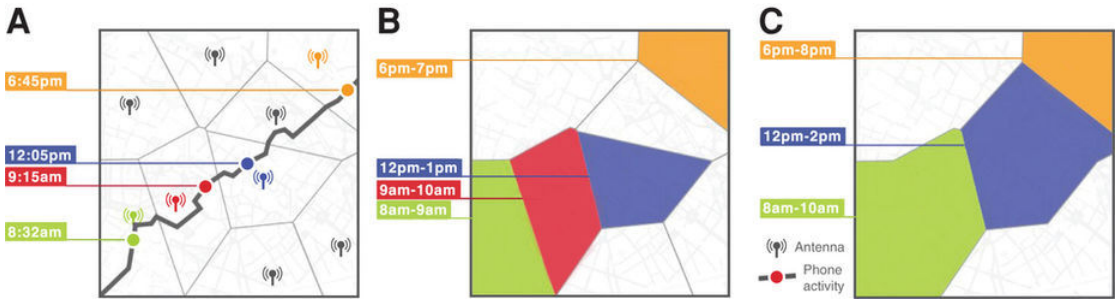


FIGURE 43 Localisation d'un téléphone auprès d'antennes relais (de Montjoye *et al.*, 2013). (A) représente la localisation exacte du téléphone, et (B) et (C) les polygones de Voronoi où se trouve le téléphone selon des critères d'agrégation temporels.

1.2 ...Lors de l'usage d'Internet...

Internet est un immense réseau de réseaux, constitué de serveurs interconnectés qui suivent des protocoles de connexions, de transferts de données et de routage relativement sophistiqués. Il existe grossièrement deux manières d'accéder à Internet, soit par une connexion d'un appareil sur le réseau téléphonique, si ce dernier supporte la technologie (3G ou +), soit par des réseaux domestiques. Dans ce dernier cas, la connexion passe systématiquement par un routeur et peut se faire via une connexion filaire (Ethernet) ou sans fil (Wifi). Le routeur est relié à un point d'accès Internet (PAI) géré par le fournisseur d'accès Internet (FAI) qui sert de passerelle en routant les informations sur le réseau Internet.

Une connexion à Internet classique passe par la suite des Protocoles Internet, soit un ensemble de contraintes permettant d'établir la connexion. En attendant la finalisation de la migration vers IP (Internet Protocol) v6, les protocoles utilisés sont de type IPv4. Une fois que la machine est connectée à un routeur, elle se voit attribuer une adresse IP unique sur le réseau, composée de 4 séries de chiffres variant entre 0 et 255 et séparée par des points, par exemple 192.168.1.2 pour une machine située dans un réseau local[95]. Pour accéder à Internet, il faut

95. Un réseau domestique pouvant accéder à Internet s'il dispose d'une passerelle (routeur lui même connecté à Internet).

que le routeur soit doté d'une adresse IP dite publique reconnue sur le réseau mondial. Pour cela, des blocs d'adresses IP sont accordés en cascade par des organismes internationaux de régulation[96], où chaque plage dépend de la zone géographique et du FAI. Par exemple, toute adresse IP comprise entre les plages 92.88.0.0 et 92.88.5.255 désigne des personnes habitants à Paris et sous contrat avec SFR. De même qu'une adresse IP comprise entre 61.19.146.0 et 61.19.151.255 correspond à une partie des utilisateurs de la ville de Bangkok.

La limite de la précision de localisation est celle du PAI sur lequel se connecte le routeur du client. La personne se trouve alors dans un cercle ayant pour centre le point d'accès, et de rayon maximum la distance entre ce point d'accès et le client le plus éloigné. Une distance élevée au point d'accès (plus de 4 km), entraîne une forte atténuation du signal et donc une baisse de débit entrant pour le client. Les points d'accès sont donc relativement rapprochés, laissant un flou géographique allant de quelques centaines de mètres à quelques kilomètres. Dans le cas d'une connexion à Internet depuis un téléphone portable via un réseau UMTS ou 3G, l'adresse IP attribuée à la machine par le fournisseur d'accès ne fournit pas d'information sur la localisation de l'appareil, car les plages IP utilisées ne renseignent que le siège social ou technique de l'opérateur. *A priori*, seul l'opérateur est à même de connaître l'adresse physique de son client, même s'il peut être amené à communiquer des informations à diverses autorités dûment mandatées qui recherchent une personne associée à une adresse IP[97]. Cela dit, il est extrêmement simple de modifier son adresse IP, en utilisant notamment un réseau privé virtuel (VPN). Cette méthode est notamment utilisée pour contourner les restrictions géographiques d'accès à certains sites, pour des raisons de licence d'utilisation ou de censure étatique[98]. À cette adresse IP attribuée, s'ajoute l'adresse physique des cartes réseaux des machines, ou adresse MAC, permettant les connexions sans fils (Wifi, bluetooth) filaires (Ethernet). Il s'agit d'un identifiant à priori unique[99] codé en hexadécimal sur 6 bits enregistré par le routeur sur lequel se connecte la machine, et n'est pas diffusée sur le réseau.

Dans le cadre d'une connexion à un site Internet, la requête est acheminée (ou routée)

96. Comme l'ICANN (Internet Corporation for Assigned Names and Numbers) en Amérique du Nord ou encore le RIPE (Réseaux IP Européen) en Europe.

97. suivant le cadre légal défini par article L34 1 du code des postes et des communications électroniques du 20 décembre 2013 https://www. eg france.gouv.fr/affichCodeArt c e.do?c dTexte= LEGITEXT000006070987& dArt c e=LEGIARTI000006465770

98. Par exemple en Turquie, avec le blocage temporaire de *Twitter* en avril 2015. www.lemonde.fr/pixels/article/2015/04/07/contourner la censure un jeu d enfant pour les internautes turcs 4610829 4408996.html. Mais l'usage de VPN est par exemple interdit aux Emirats Arabe Unis ou il constitue dorénavant un crime – www.lemonde.fr/pixels/article/2016/08/01/aux emirats arabes unis les utilisateurs de vpn risquent desormais des peines de prison 4977057 4408996.html. L'usage de VPN est également de plus en plus surveillé en Chine http://www.courrierinternational.com/article/chine Internet les reseaux prives virtuels dans le viseur du pouvoir et en Russie http://www. emonde.fr/p xe s/ art c e/2017/07/31/en ch ne et en russ e es vpn sont nterd ts par a censure 5167006 4408996.htm . La Société *Apple* a supprimé en juillet 2017 les applications fournissant des VPN de son catalogue Chinois https://techcrunch.com/2017/07/30/app e ssues statement regard ng remova of un censed vpn apps n ch na/?nc d=rss

99. Même s'il est extrêmement simple de modifier les adresses MAC. Mais certaines législations peuvent considérer cela comme de l'usurpation d'identité.

entre l'adresse d'origine et de destination [100], grâce à la structure du datagramme IP qui renseigne à la fois l'adresse de l'émetteur et celui du destinataire. Ceci implique que n'importe quelle machine distante (ou site web visité) connaît la source émettrice de la requête [101]. Les serveurs du site enregistrent ces informations en mode séquentiel, qui fournissent les « *logs* » de connexion. En plus de l'adresse IP du routeur domestique qui visite le site, sont également enregistrées la date, l'action effectuée par le site (un chargement d'image, de vidéo, etc.), les actions effectuées par la machine *e.g.* un clic sur une vidéo, ou sur un autre lien dans le site. Le site enregistre également le système d'exploitation de la machine et le navigateur utilisé, ceci à des fins d'optimisation de l'affichage un téléphone portable n'aura pas le même rendu qu'un ordinateur. Ces données permettent au gestionnaire du site d'avoir des informations sur le nombre de visites par jour, sur les contenus regardés, et permettent d'enregistrer les éventuels bugs. L'adresse IP qui renseigne sur l'origine géographique peut également servir à afficher des publicités sur des produits se trouvant dans la région de la personne qui visite le site, ou à détecter des connexions provenant de lieux inhabituels lorsque le site requiert une authentification [102].

Certains sites Internet ont recours à l'utilisation de « cookies », des petits fichiers textes non exécutables installés sur le terminal de l'utilisateur et qui stockent des données relatives à l'utilisation du site web (préférences d'affichages, achats effectués, etc.). La plupart des sites présentent du contenu provenant d'autres sites (par exemple des boutons « j'aime » de *Facebook*), ce qui implique que certains cookies sont communs à plusieurs noms de domaine (cookie tierce partie). Ce dernier aspect permet ainsi de « tracer » les comportements des utilisateurs lorsqu'ils fréquentent d'autres sites Internet et d'enrichir ainsi les bases de données les concernant [103].

La connexion à un site Internet donne des informations à l'échelle macro sur le lieu et l'heure de connexion d'une personne. Si la personne accède à un site qui utilise des cookies ou qui requiert une authentification par identifiant et mots de passe, le gestionnaire du site a techniquement la possibilité de suivre les déplacements de cette personne. Plus localement, si un utilisateur se connecte par exemple au réseau wifi d'un grand centre commercial ou d'une université, il est possible de suivre ses déplacements au gré des connexions aux différentes bornes du réseau (Sapiezynski *et al.*, 2015) et la personne est alors identifiée par son adresse MAC.

100. L'adresse de destination, par exemple www.dsi.cnrs.fr est convertie en adresse IP (ici 194.57.136.114) par un serveur de nom de domaine, le DNS (Domain Name Server).
101. En général, l'adresse IP du routeur domestique.
102. Comme c'est le cas avec les sites facebook.com et hotmail.com. Ces sites demandent alors de vérifier le compte, ce qui peut passer par donner des informations sur sa vie personnelle ou sur ces relations sociales.
103. Il existe cependant divers plugins qui permettent de supprimer ces cookies du navigateur. Nous pouvons citer ici ghostery et privacy badger pour les utilisateurs de Firefox.

1.3 ...Ou à d'autres moments

Les données personnelles enregistrées et conservées suite à des protocoles de connexions ne se cantonnent pas uniquement à la téléphonie mobile ou à l'accès à Internet. Nous présentons ici deux cas de figure de la vie quotidienne. D'autres exemples sont présentés en annexe B et C.

1.3.1 Données bancaires

Tout retrait bancaire effectué auprès d'un distributeur automatique de billets (DAB) entraîne la création de données, tel que le numéro de la carte bancaire et la banque associée, le numéro du DAB et la banque propriétaire, l'heure et le jour de la transaction, ainsi que le montant. Ces données seront accessibles à la banque propriétaire de l'automate ainsi qu'à la banque de l'utilisateur afin de produire un relevé de retrait bancaire et de faire des recherches en cas d'erreur (par exemple un débit dans le compte en banque sans obtention des billets pour le client, comme cela arrive régulièrement en Inde). Une banque sait donc où ses clients ont effectué des retraits via le numéro du DAB qui est géolocalisé, et à quelle heure, via les *logs* de connexions.

Toute entreprise ou commerce qui détient un terminal de paiement électronique (TPE) pour permettre à ses clients de payer par carte doit au préalable enregistrer la domiciliation de son établissement et déclarer sa raison sociale auprès d'une banque. Ainsi, lorsqu'un client paie par carte, les données enregistrées sont la date, le montant de la transaction, ainsi que le nom de l'établissement. La localisation du commerce est connue, tout comme le type d'activité ou de services qu'il propose, ce qui permet techniquement à une banque de savoir les tendances de déplacements et les habitudes de consommation de ces clients, mais l'utilisation de ces informations est théoriquement régulé par la Commission Nationale Informatique et Liberté (CNIL, voir section suivante). Le développement des cartes de paiement dîtes « sans contact », fonctionnant grâce à des puces RFID[104], permet de simplifier les transactions en se passant de code et pourrait encore inciter les gens à préférer les paiements par cartes plutôt que par espèce, créant encore plus de traces numériques.

1.3.2 Données transports publics

Au-delà du simple fait d'être en règle en validant son titre de transport lors d'un trajet, ce geste répété par des milliers d'utilisateurs crée des bases de données indispensables aux planificateurs urbains, car elles permettent de connaître le nombre de passagers au cours du temps et donc d'adapter et optimiser les offres du réseau. En fonction du mode de validation du

104. Radio Frequency Identifier. Il s'agit de puces passives qui s'activent lorsqu'elles reçoivent un signal et transmettent alors des informations à l'émetteur – par exemple des données bancaires ou des informations sur les abonnements dans le cadre d'une carte de transport.

système de transport, ces données générées entraînent des méthodes d'exploitation différenciées.

Dans le cas de la validation d'un billet à tarif unique, comme c'est le cas dans la plupart des villes françaises, les régies de transports connaissent le nombre de personnes par ligne, par arrêt et par intervalles de temps, ce qui donne une information sur le nombre de passagers entrant à l'instant *t*, sans pour autant connaître les flux sortants. Cet aspect disparaît lorsque les tarifs des tickets dépendent de la distance à parcourir, car il y a en général une obligation de valider pour pouvoir sortir de la station, comme c'est le cas à Delhi à Bangkok, ou sur les lignes du réseau RER à Paris. Dans ces deux cas de figure, les informations collectées sont des données agrégées et il n'est possible d'apprécier que les déplacements collectifs nombre de personnes entrant et sortant à chaque station.

Les utilisateurs réguliers ont souvent recours à des abonnements, ce qui, en plus d'un tarif préférentiel leur donne accès à une carte de transport contenant des puces électroniques[105] dans lesquelles sont enregistrées des informations les concernant identifiant, type d'abonnement et/ou solde, etc. Il est alors possible de connaître les lieux de montée de ces utilisateurs, ce qui permet en théorie de reconstruire des trajectoires individuelles. Cela dit, pour le cas du passe Navigo, seuls les trois derniers déplacements sont enregistrés[106], afin de protéger les informations sur les utilisateurs. Étant donné qu'une grande partie des déplacements sont routiniers, il y a de grandes chances que le domicile et le lieu de travail fassent partie de ces 3 déplacements enregistrés l'heure de pointage permettant de distinguer l'un de l'autre. À noter toutefois que dans la lignée de l'ouverture des données publiques, la RATP met à disposition une partie de ces données agrégées[107].

Après cet aperçu technique sur la création de traces liées aux protocoles de connexions et au fonctionnement des sites web, nous abordons maintenant les traces laissées plus activement par les utilisateurs d'Internet.

2 Création et collecte de données sur les réseaux sociaux

« Elle m'a pas followback quand je l'ai follow », Stanilas Dina Pinto (2017).

Un réseau social peut se définir comme un graphe ou les nœuds sont les personnes que nous connaissons et les liens les degrés d'interactions avec ces personnes, qui peuvent être des gens de notre famille, des amis, des collègues de travail ou des amis d'amis, etc. Par extension, un réseau social sur Internet est une plateforme permettant d'accéder à une page personnelle

105. Par exemple le pass' Navigo en île de France, la Oyster card à Londres, la Smart Card à Delhi ou le smartpass et la Rabbit card à Bangkok.
106. Voir http://www.st f.org/IMG/pdf/Nav go 2.pdf. Il existe également des passes navigo « anonymes », ou moyennant 5€ de plus, le nom du possesseur de la carte n'est inscrit dans aucun fichier. http://www.nav go.fr/wp content/up oads/2016/06/20120912 cgu carte nav go decouverte.pdf
107. https://data.ratp.fr/exp ore/?sort=mod fied

dans laquelle figurent nos contacts.

La définition des protocoles IRC (Internet Relay Chat, pour discussion relayée par Internet) à la fin des années 1990 a probablement posé les fondements théoriques des actuels réseaux sociaux sur Internet. Ce protocole permet entre autres de discuter directement avec d'autres personnes sur des canaux publics sur lesquels tout le monde peut se connecter ou sur des canaux privés, géré par un administrateur qui autorise ou non l'accès à la discussion. Au gré d'évolutions mineures, il fut ensuite possible d'enregistrer une liste de contacts permettant de voir lorsque ces derniers sont connectés ou non sur la plateforme de discussion. Suivant une logique assez similaire, certains sites Internet commencèrent à créer des forums de discussion (publics ou privés), où des gens pouvaient par exemple poser une question et d'autres y répondre. En parallèle, se sont développé des « blogs » soit des pages personnelles ou un individu partage des informations de nature très diverse sur une page, accessible à tous ou non.

Les réseaux sociaux sur Internet d'aujourd'hui sont donc basiquement un mélange de ces approches : une partie est dédiée à l'expression (publique ou privée), une autre à une page personnelle contenant les informations que nous souhaitons partager (publiquement, ou uniquement auprès de certaines personnes).

Par exemple *Facebook*, qui avec plus de 2 milliards d'utilisateurs[108], est le réseau social en ligne le plus utilisé, peut basiquement se résumer à une sorte de blog (page personnelle) couplé à un système proche d'IRC (messagerie). Avec cependant pour challenge technique de hiérarchiser l'affichage de contenus créés et partagés (messages, photos, liens vers d'autres sites) par les personnes qui font partie du même réseau (ou « amis »)[109]. La vocation de *Facebook* est plutôt de permettre aux utilisateurs d'échanger des informations de manière privée, c'est-à-dire visible par les membres de leur réseau social même s'ils peuvent choisir de rendre certains contenus visibles à tous. L'ensemble des informations, publiques comme privées est bien entendu détenu et analysé par *Facebook*[110]. Les réseaux sociaux ont de fait une vision quasiment panoptique sur les pratiques et comportements de leurs utilisateurs.

Parmi les contenus que les personnes peuvent partager sur les réseaux sociaux, ne nous intéressent ici que ceux qui sont associés à une géolocalisation (figure 44). Une localisation peut être partagée directement par l'utilisateur, de manière purement déclarative sans que la plateforme d'expression n'ait accès aux données de géolocalisation de l'appareil, en publiant par exemple un message « Je suis bien arrivé à Rouen ». Dans le cadre d'un usage sur un smartphone, l'utilisateur peut aussi autoriser une application ou un site Internet à avoir accès aux données permettant une géolocalisation[111], à savoir les informations du GPS de l'appareil,

108. http://www. emonde.fr/p xe s/art c e/2017/06/27/facebook passe a barre des 2 m ards d ut sateurs 5152063 4408996.htm

109. Dont les algorithmes tendraient à créer un « effet bulle », où chaque personne ne verrait que du contenu confortant son opinion (Pariser, 2011).

110. https://www.facebook.com/fu data usea po cy

111. Pour *Facebook* voir : https://fr fr.facebook.com/he p/275925085769221?he pref=faq content pour *Twitter*, voir plus bas, pour Foursquare : https://support.foursquare.com/hc/en us/art c es/201065420

ainsi que des informations sur le niveau de signal reçu des antennes relais (dont la localisation est souvent connue) par l'appareil, ce qui permet de détecter par trilatération [112] la position du téléphone portable [113]. Si l'utilisateur peut selon les plateformes partager cette information à ses contacts (figure 44), cette dernière est aussi détenue par l'application, qui peut, selon les conditions d'utilisations, la transmettre à d'autres entités.

FIGURE 44 Illustration d'un réseau social avec Géolocalisation. Tiré de *« Ultimex en Enfer, tome 1 : Ni Dieu, ni Maîtres, ni glaçons dans le Whisky »* de Gad, édition Lapin Vraoum (2017).

Nous y reviendrons plus longuement dans la section 3, mais ces informations peuvent être transmises directement à des sites partenaires (Zang *et al.*, 2015) ou accessibles à des tiers par l'intermède d'*API* (Application Program Interface) qui consistent en une série de protocoles développés par une entreprise et qui permettent à un utilisateur distant d'accéder à certaines de ces bases de données, en fonction de ces accréditations [114]. Pour faire simple, il est en général possible d'accéder gratuitement à un échantillon des données possédées par certains réseaux sociaux. Ces derniers définissent des limites de volumes (quantité d'informations), de requêtes (nombre de fois qu'il est possible d'interroger la base par heure ou par jours) ou d'exhaustivité (accès a certaines informations uniquement), qu'il est possible de dépasser en payant un certain montant. Les prochains paragraphes ne décriront que les principales plateformes de réseaux sociaux en termes d'utilisateurs et dont des données publiques et géolocalisées sont accessibles gratuitement via des *API*.

Locat on Sett ngs ou encore pour *FlickR* : https://uk.he p.yahoo.com/kb/fl ckr w ndows up oadr/enab e ocat on features mob e dev ce s n24002.htm

112. La trilatération est une méthode permettant de calculer une localisation à partir des distances entre à minima 3 points dont la localisation est connue – et non des angles comme c'est le cas pour la triangulation.

113. Un téléphone portable n'est enregistré que sur une seule antenne relais, celle dont le signal est le plus important. Mais le téléphone a une liste d'antennes dont il perçoit le signal, avec leur niveau d'atténuation. Cette information, disponible que sur le téléphone permet d'estimer la distance entre chaque antenne, et donc par trilatération de connaître la position approximative du téléphone.

114. le principe de fonctionnement des *API* est assez bien expliqué ici : http://www. efigaro.fr/secteur/h gh tech/2018/03/30/32001 20180330ARTFIG00015 au fa t c est quo une ap .php

2.1 Twitter ou le micro-blogging

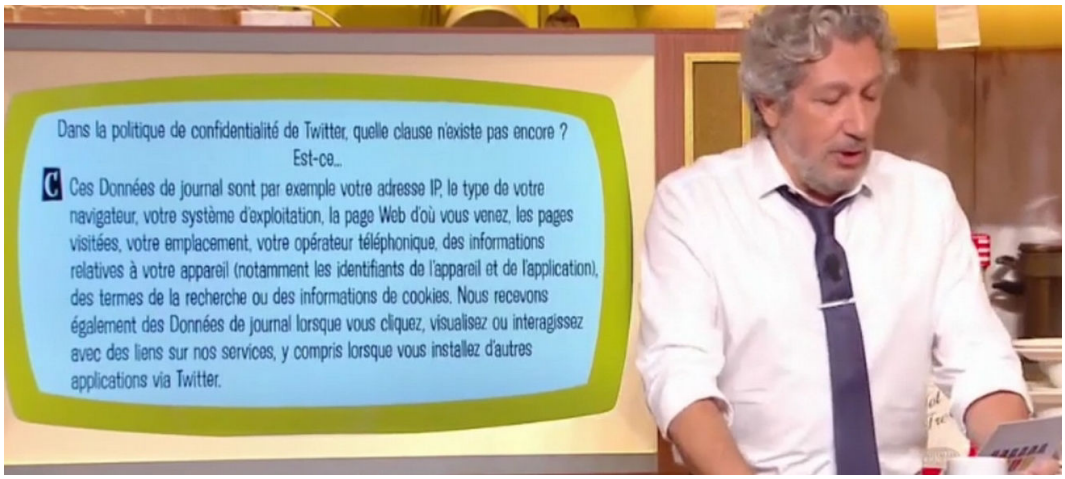


FIGURE 45 Extrait de la politique de confidentialité de *Twitter* concernant la collecte des données utilisateurs. Dans l'émission "Burger Quiz" du 20/06/2018, diffusée sur TMC.

Twitter est une plateforme de « micro-blogging » créé en 2006 et orienté vers l'envoi et le partage public de courts messages de 280 caractères[115]. Utilisé par plus de 300 millions de personnes, le service a activé en 2009 une fonction permettant l'ajout d'une localisation à un message. Cette option permet d'ajouter soit la localisation précise de l'appareil, soit un lieu parmi une liste. Conformément à la politique de confidentialité du service (figure 45), une partie de ces données est accessible à des tiers en temps réel par une *API*, permettant de récupérer des données longitudinales, où chaque message est associé à un identifiant, une heure de réception, et parfois à une localisation. Nous nous étalerons plus longuement sur le sujet dans les chapitre 5 et 6.

2.2 Sites de « check-in »

Gowalla, Foursquare et Swarm

La plateforme *Gowalla*, disponible sur smartphone, fut ouverte en 2007, et bien que fermée en 2012, elle inaugure le concept de « *check-in* », ou enregistrement. Le principe est simple : il permet à une personne qui fréquente un lieu présent dans la base de l'indiquer aux membres de son réseau social présents sur le site. En contrepartie l'utilisateur reçoit des récompenses, allant d'une sorte de petit badge numérique à des bons de réductions. Le même principe fut repris par *Foursquare*[116], lancé en 2009[117]. Ce dernier compte environ 50 millions d'utilisateurs

115. 140 caractères jusqu'en novembre 2017.
116. Www.foursquare.com
117. Toujours active, bien que l'entreprise ne fasse toujours pas de profit. https://www.entrepreneur.com/article/290543. Son « business model » est basé principalement sur des levées de fonds, soit des injections ponctuelles d'argent par des investisseurs. L'entreprise a reçu au total 166 millions de dollars depuis sa

mensuels et recense plus de 93 millions de « lieux », principalement des commerces, dans lesquels sont enregistrés plus de 9 millions de *check-in* quotidiens [118]. La plateforme permet également de noter les établissements fréquentés. L'entreprise s'est divisée en deux entités [119] en 2014, « *Foursquare City-guide* » et « *Swarm* ». La première entité se focalise maintenant sur la recommandation de lieux de sorties basée sur l'historique des lieux fréquentés par l'utilisateur et les notes données par les membres de son réseau social. La seconde, « *Swarm* », disponible sur smartphone, se veut plus ludique et reprend les mêmes fonctionnalités de *check-in* tout en permettant de voir si un membre du réseau social est situé à proximité. Chaque *check-in* est récompensé par un petit badge numérique et la personne qui a accumulé le plus grand nombre de badges sur une période de 30 jours devient, aux yeux de ces contacts sur la plateforme, le « maire » de la ville (figure 46). Dans les deux cas, les localisations partagées sur *Foursquare* ou *Swarm*, peuvent également être partagées sur *Twitter* ou *Facebook*, si la personne le souhaite et possède un compte sur ces plateformes.

FIGURE 46 Illustration d'un « *check-in* » sur Swarm. D'aprés *Dans l'ombre de la peur : le big data et nous* de Michael Keller et Josh Neufeld, éditions ça et là (2017).

Les données sur le nombre de visites et d'utilisateurs unique pour chaque lieu de la base de *Foursquare* sont accessibles gratuitement et en temps réel grâce à une *API* dédiée [120]. Les données publiques concernant l'historique des lieux visités par un utilisateur sont également accessibles, pour peu que l'on connaisse son pseudonyme [121]. Nous y reviendrons plus tard, mais ces principes de récompense lorsqu'un utilisateur partage sa localisation lorsqu'il se rend dans un lieu et de suggestions de lieux situés à proximité et susceptibles de lui plaire peut influencer son déplacement et les traces numériques qu'il laisse et biaiser les analyses de mobilités basées

création, dont 45 millions en janvier 2016. https://www.crunchbase.com/organ zat on/foursquare/fund ng rounds

118. https://foursquare.com/about – visité le 19 juillet 2017.
119. 3 si on ajoute foursquare location intelligence, la partie développement pour entreprise.
120. https://deve oper.foursquare.com/docs/venues/search
121. https://deve oper.foursquare.com/docs/users/venueh story

sur ce type de réseau social.

Facebook Places

Inspiré sans complexe des *check-in* de *Foursquare*, *Facebook* propose depuis 2010[122] la même fonctionnalité, sans pour autant offrir de contreparties (bons points ou coupons de réductions) à ces utilisateurs. Peut-être influencée localement par les établissements que les personnes fréquentent (typiquement si vous faites un *check-in* vous on vous offre le café), l'utilisation de cette option permet également de montrer son espace de sociabilité, et peut être ainsi de se démarquer socialement. *Facebook*, qui est un réseau social plutôt privé, c'est-à-dire que le contenu produit est orienté pour être vu par le réseau social (le groupe "d'amis"). Il n'en demeure pas moins que les lieux fréquentés sont considérés comme publics et qu'il est possible de faire des requêtes et des recherches par lieux, notamment avec l'*API Place search* qui permet de « rechercher des millions de lieux dans le monde entier et récupérer des détails propres à un lieu, comme le nombre de visites, les avis ou l'adresse, le tout en une seule demande »[123]. Il n'est pas possible de savoir qui a déclaré être présent dans tel lieu (au-delà des contacts de l'utilisateur). Mais bien que ces données soient non longitudinales, il est très aisé de récupérer des informations sur le nombre de *check-in* dans tous les lieux. Nous détaillerons cela dans le chapitre 6, section 2.2.

2.3 Partage de photos

« Merci Arnaud d'être venu t'prendre en photo pour alimenter tes réseaux sociaux »,
Aurélien Cotentin (2017).

FlickR

L'une des premières plateformes à permettre le partage de contenus géolocalisés est le site d'hébergement et de partage de photo et de vidéo *Flickr*[124]. Fondé en 2004[125], il permet à des utilisateurs de mettre en ligne des photos et s'ils le désirent, de les associer à des coordonnées géographiques. Ces dernières peuvent être publiques, ou alors d'accès restreint (réservées à un groupe). Le site hébergerait ainsi en 2017 13 milliards de photos et 2 millions de groupes d'utilisateurs[126]. En récupérant par une *API*[127] les photos publiques associées à l'identifiant de la personne les ayant postées, il est donc possible de voir où la personne est allée et d'apprécier les lieux les plus photographiés si l'on raisonne de manière agrégée[128] (Girardin *et*

122. http://ed t on.cnn.com/2010/TECH/soc a .med a/08/18/facebook. ocat on/ ndex.htm
123. https://deve opers.facebook.com/docs/p aces/web/search
124. www.flickr.com
125. Puis racheté en 2007 par Yahoo! Et propriété de Verizon depuis juin 2017.
126. www.flickr.com
127. https://www.fl ckr.com/serv ces/ap /
128. Voir par exemple https://www.fl ckr.com/photos/wa k ngsf/sets/72157623971287575/

al., 2008, 2007). Dans la même veine, on trouve également le site *Panoramio*, mais de moindres importances et fermé depuis 2016.

Instagram

Instagram est un service qui propose des fonctionnalités similaires à *FlickR*, mais plus orienté sur le partage et les commentaires. Racheté par *Facebook* en 2012[129], *Instagram* a notamment popularisé l'utilisation de filtres permettant de « styliser » les photos. Cette plateforme très simple d'utilisation, notamment sur des smartphones compte désormais plus de 700 millions d'utilisateurs mensuels[130] qui postent quotidiennement 95 millions de photos ou vidéos[131].

Ces données publiques présentes sur la plateforme peuvent être récupérées par une *API*[132], notamment lorsqu'elles sont associées à un lieu[133]. Il existe deux niveaux de permissions pour accéder aux données : le mode test (ou *sandbox*[134]) et le mode opérationnel. Le mode test permet de récupérer des données par lieu, mais uniquement sur les utilisateurs qui se sont enregistrés dans l'application du développeur (maximum 10 personnes), ce qui paraît extrêmement limité. Pour passer au niveau suivant, il faut envoyer une demande (avec démonstration vidéo) qui est étudiée par l'entreprise et seules sont acceptées celles qui permettent à un utilisateur de partager son contenu, ou qui sont relatives à l'analyse ou au ciblage d'une audience pour des publicitaires ou des entreprises, ou encore à des diffuseurs ou éditeurs de contenus[135]. L'utilisation de l'*API* semble donc assez délicate à des fins de recherches, même si certains y ont eu accès ((Giridhar *et al.*, 2017 ; Giridhar et Abdelzaher, 2017). Cela dit, il est tout à fait possible de faire de simples requêtes web pour extraire par exemple le nombre d'images associées à une ville ou un quartier, et d'apprécier l'évolution temporelle de la quantité de contenus associés à ce lieu[136].

2.4 Des données géographiques volontaires ?

Toutes ces données postées sur les réseaux sociaux posent la question de leur caractère volontaire, et de leur réutilisation, notamment à des fins de recherche. Dans un article paru en 2007, Goodchild propose le terme de « VGI » (pour Volonteer Geographic Information) afin de décrire le phénomène de création d'information géographique par des néophytes (Goodchild, 2007). En prenant comme exemple les sites de cartographies participatives tels

129. Pour un montant d'un milliard de Dollars
130. http://b og. nstagram.com/post/160011713372/170426 700m on
131. https://www.b ogdumoderateur.com/ch ffres nstagram/
132. https://www. nstagram.com/deve oper/
133. https://www. nstagram.com/deve oper/endpo nts/ ocat ons/
134. https://www. nstagram.com/deve oper/sandbox/
135. https://www. nstagram.com/deve oper/rev ew/
136. Par exemple https://www. nstagram.com/exp ore/tags/bangkok/ renvoyait 16089576 documents le 26/07/2017 à 17h et 53 de plus 5 minutes plus tard.

que *Wikimapia*[137] et *OpenStreetMap*[138] (*OSM*) ou des services cartographiques privés tels que le globe virtuel *Google Earth*[139] qui incitent les personnes à ajouter des « lieux », il insiste sur le rôle bénéfique des « amateurs » dans l'observation géographique. Il va même un peu plus loin en disant que les êtres humains peuvent dans certains cas s'apparenter à des capteurs intelligents qui interprètent les informations locales. La qualité des données créées est certes à prendre avec des pincettes, mais leur quantité, leur coût et leur disponibilité font des VGI une source de donnée intéressante, prometteuse et de grande valeur pour les géographes (Goodchild, 2007).

Mais au moment où cet article était écrit (2007), les smartphones n'existaient pas (Quesnot, 2016), et il paraissait clair que les personnes qui produisaient volontairement des données, par exemple en cartographiant les routes de leur quartier sur Openstreetmap savaient que leur travail non rémunéré allait être utile à d'autres personnes et que l'usage de ces données ne serait pas dévoyé, compte tenu de la structure du site Internet. Entre temps, le développement d'outils de partage de position géographique dans les réseaux sociaux a incité les utilisateurs à divulguer leur localisation, ce qui implique un approfondissement du concept de *VGI*.

Le GDR Magis propose la définition suivante : « L'information géographique volontaire (aussi appelée participative, collaborative), (...) est l'information géographique créée par la participation répandue (au travers d'Internet) d'un grand nombre de citoyens (contributeurs), ayant ou non des compétences en géographie, topologie ou géomatique »[140]. Ce qui peut inclure les *check-in* de *Facebook* et les *tweets* géolocalisés dans les VGI. Harvey propose en 2013 de distinguer d'une part les informations géographiques volontaires (VGI) des contributions d'informations géographiques (CGI, contributed geographical Information) (Harvey, 2013). L'opposition se fait en fonction du niveau de clarté et de contrôle de la collecte et de l'utilisation de ces données. Très clair chez Openstreetmap, qu'il qualifie de « VGI», moins clair pour les données issues des téléphones portables, réseaux sociaux et puces RFID utilisées dans les transports, qu'il classe sous le terme de « CGI ». Cette distinction implique des données de qualité différentes, mais également des enjeux éthiques bien particuliers. Quesnot (2016) met en avant que le terme « *Volunteered* » employé par Goodchild serait plus à prendre au sens britannique, soit « contributif », ce qui écarte le caractère conscient du concept de VGI.

Hildebrandt (2013) propose de sortir du paradigme données personnelles / non-personnelles et de réfléchir plutôt en termes de données volontaires, observées ou inférées. Elle définit les données volontaires comme « créées et partagées de manière explicite par les individus, par exemple les profils des réseaux sociaux »[141]. Les données observées sont les données enregistrées permettant « l'amélioration de l'expérience de l'utilisateur » lorsqu'il visite

137. www.wikimapia.org/
138. http://openstreetmap.fr/
139. https://www.*Google*.com/earth/
140. http://apvg mag s. gn.fr/
141. « created and explicitly shared by individuals, *e.g.* social network profiles »

un site Internet, l'optimisation d'un site web ou encore la profitabilité d'un système économique. La création d'un profil de l'utilisateur par le croisement de bases de données constitue des données inférées. Cette distinction entre données volontaires et observées ne se base pas sur les notions de consentement ou du caractère personnel de la donnée et Hildebrandt (2013) suggère pour chaque type de donnée une protection légale différenciée.

Le débat peut être très long. Nous observons simplement que d'un point de vue descriptif et sémantique, le fait de poster un *tweet* géolocalisé ou d'effectuer un *check-in* est une action volontaire qui entraîne la création de données géolocalisées. Mais la distinction entre un contributeur d'OSM et un *tweetos* (utilisateur de *Twitter*) ne se fait non pas par la plateforme utilisée, ni le type de donnée créée, mais plutôt dans l'intention derrière la création de cette donnée. Nous considérons ici les traces numériques provenant de réseaux sociaux comme des données géographiques produites volontairement, mais dont l'usage est détourné. L'utilisation de ces données est alors opportuniste, éloignée du contexte initial de leur création et sans que l'utilisateur-producteur (consommateur ?) en soit vraiment conscient. Elles revêtent cependant un énorme potentiel dans l'analyse des mobilités (voir chapitre 5, état de l'art) et sont déjà largement utilisées à des fins commerciales (section suivante). Leur utilisation dans la recherche ne nous parait donc pas aberrante [142], même si cela implique de prendre des précautions d'ordre technique (sécurisation des données) et éthique (maintient autant que possible du plus grand niveau d'anonymisation). Car comme nous allons rapidement le voir dans la section suivante, ce système de création/collecte de données est pourvoyeur de dérives de degrés divers.

3 Collecte massive et identité numérique

« L'habit ne fait pas l'moine dans la ruée vers l'or
Dès lors les techniques se perfectionnent
La carte à puce remplace le Remington », Claude M'Barali (1994).

L'héritage de la philosophie des « pionniers de l'informatique et du net » [143], considérés comme des « hackers [144] libertaires » a fait que la plupart des services disponibles sur Internet sont gratuits et accessibles au plus grand nombre (Biagini, 2012 ; Sadin, 2015). La contrepartie étant principalement la présence de publicités que beaucoup semblent accepter compte tenu de la qualité de certains services et de la richesse des contenus. Le système économique des entreprises d'Internet d'aujourd'hui est généralement composé d'un versant extrêmement

142. Il faut cependant garder à l'esprit que l'accès à ces bases de données par des chercheurs qui vont publier leurs résultats peut donner des idées d'utilisations des différents services et peut faire office de recherche et développement ou de publicité gratuite pour ces entreprises.

143. Nous pouvons citer Richard Stallman, initiateur des logiciels libres avec la création de la licence publique générale GNU en 1983, base du système Linux, ou encore Tim Berners Lee, considéré comme l'inventeur du « World Wide Web ».

144. Le terme de « Hacker » peut se rapporter à « bricoleur ».

déficitaire (le fonctionnement et le maintien d'un service gratuit[145]) qui doit être compensé notamment par la vente d'espace publicitaire. Par exemple sur les 24.7 milliards de dollars de revenus de l'entreprise Alphabet (maison mère de *Google*) au 1er trimestre 2017, 21.4 provenait de la publicité[146].

L'importance de ce segment entraîne la nécessité de l'optimisation de la présentation de publicités aux consommateurs potentiels, qui passe notamment par une meilleure connaissance de leur profil personnel (Kessous et Rey, 2009). Cette logique impose donc d'engranger le plus de données possible concernant les individus, que des traitements algorithmiques plus ou moins pointus peuvent ensuite mettre dans des « cases » (Douplitzky, 2009). Les données sont collectées par différents intermèdes. Les sites web utilisés par l'Internaute, notamment les réseaux sociaux n'ont simplement qu'à enregistrer les *logs* de connexions de leurs utilisateurs ainsi que toutes les interactions qu'ils effectuent sur les pages. Un rapide portrait peut dès lors être créé selon les réactions à divers messages ou sur les contenus partagés. L'utilisation de cookies (section 1.2) permet également de récolter des informations sur d'autres pages visitées, complétant encore le profil de l'internaute (centres d'intérêt, etc.).

Les différents services en ligne où certaines applications installées sur des téléphones dotés de systèmes d'exploitation de type *Android* ou *iOS* échangent directement des informations de l'utilisateur avec des entités tierces, les "*third party*". Par exemple, l'entreprise de paiement en ligne *PayPal* est susceptible d'échanger des informations auprès de plus de 400 partenaires[147]. Ces transferts de données sont parfois effectivement indispensables au fonctionnement du service. Dans le cas de *Paypal*, il s'agit par exemple de valider les transactions de banques. Cependant, dans bien des cas, la nature des données échangées ou vendues dépasse largement les besoins de fonctionnement et n'améliore en rien l'expérience utilisateur.

L'étude de Zang et al. (2015) sur la question dresse une partie de la cartographie de ces échanges d'informations. Si 73 % des applications qu'ils ont testés sur Android échangeaient des données personnelles des utilisateurs, comme l'adresse mail, avec d'autres entreprises, 47 % des applications sous *iOs* transféraient des informations relatives à la géolocalisation à d'autres entreprises. Si cet aspect peut être préjudiciable à la vie privée des utilisateurs de ces applications, la collecte de données géographiques associées à un individu est la base du géo-marketing moderne, qui peut être amené à traiter des profils comportementaux en prenant en compte les lieux visités par les personnes. D'un point de vue consommateur, cela peut entraîner dans le meilleur des cas une différenciation des publicités en fonction des lieux dans lesquels nous nous trouvons et du contexte de la période (vacances, travail, jours de repos, etc.), les incitant à faire des achats dans tel ou tel magasin, au gré de coupons de réductions opportuns.

Certaines entreprises profitent aussi de la complexité des réglages des paramètres des

145. Comme le rappel le slogan du réseau social *Facebook* « It's free and always will be »
146. https://abc.xyz/ nvestor/news/earn ngs/2017/Q1 a phabet earn ngs/
147. https://www.paypa .com/ e/webapps/mpp/ua/th rd part es st

applications qu'elles proposent. Par exemple *Google* enregistrerait la localisation de toute personne accédant à certains de ces services comme *Maps* (via *Android* ou *iOs*), même si l'option « historique des positions » est désactivée. Pour qu'aucune localisation ne soit enregistrée, il faudrait aussi désactiver l'option «Activité sur le Web et les applications».[148]

Les différentes données collectées, soit directement par l'entreprise, soit par échanges directs (probablement monnayés), soit via des *API*, ou simplement par l'achat auprès de revendeurs, les « *databrokers* », permettent d'établir un profil partiel, plus ou moins précis d'un identifiant associé à une personne (Douplitzky, 2009). Ceci entraîne la possibilité d'établir une identité numérique qui peut être définie comme « la collection des traces (...) que nous laissons derrière nous, consciemment ou inconsciemment, au fil de nos navigations sur le réseau » (Ertzscheid, 2013). La figure 47 ci-dessous exprime en quelque sorte cette idée de reconstitution de profil, permettant d'avoir une idée assez précise de l'objet initial. Malgré une dispersion des éléments constitutifs, une fois ces derniers grossièrement rassemblés, il est possible de reconnaître l'entité originelle, et les initiés peuvent en deviner la marque et conclure sur ces caractéristiques.

FIGURE 47 « Laissez votre empreinte », une allégorie de l'identité numérique. Extrait d'une publicité.

De nombreuses personnes considèrent donc les données personnelles comme le « nouveau pétrole du XXIe siècle ». En effet, comme pour cet hydrocarbure, les données sont à l'origine brutes, c'est-à-dire non traitées et inexploitable en tant que tel. Dans ce cas leur valeur est moindre. Mais elles peuvent être « raffinées », c'est-à-dire analysées et enrichies en contexte, ce qui démultiplie leur intérêt marchand et scientifique. Cela dit, Eveny Morozov souligne dans

148. https://www.apnews.com/828aefab64d4411bac257a07c1af0ecb/AP Exclusive: Google tracks your movements, ket or not. (visité le 13/08/2018)

un article pour « Le monde diplomatique »[149] le côté encore plus capitaliste[150] des données personnelles :

« Le pétrole ne devient pas meilleur ou plus précieux parce que vous en stockez davantage dans votre entrepôt. En revanche, les données, oui : plus vous en collectez, plus vous gagnez en finesse d'analyse et plus importantes seront les économies, que vous pouvez répercuter sur les citoyens ».

Le début de l'année 2018 a été très riche en termes de révélation sur les dérives dans les collectes de données personnelles. Par exemple dans l'affaire de *Cambridge Analytica*, des données concernant plusieurs dizaines de millions d'utilisateurs de *Facebook* ont été récoltées légalement par un tiers et furent ensuite revendues *a priori* illégalement à une entreprise spécialisée dans le profilage et le ciblage dans l'optique d'optimiser l'envoi de messages à certains électeurs indécis, notamment lors de l'élection américaine de 2016[151]. Selon Eyvgeny Morozov, il ne s'agit là que de la partie émergée de l'iceberg[152] : « *pokes and likes is simply how our data comes to the surface much like energy firms drill deep into the oil wells : profits first, social and individual consequences later* »[153].

Une autre affaire nous a particulièrement intéressée, celle de l'application de rencontre *Grindr,* basée sur la géolocalisation. Un utilisateur de l'application peut voir les profils d'autres utilisateurs dans un rayon donné et peut décider de rentrer ou non en contact avec l'un d'eux. Le principe est exactement le même que celui de *Tindr*, mis à part que *Grindr* cible ouvertement une population homosexuelle. Développée par des Canadiens, puis vendue à un consortium chinois en 2016[154], l'application revendique plus de 3 millions d'utilisateurs. Sintef, une organisation norvégienne indépendante a montré en utilisant le même procédé que Zang et al. (2015), que *Grindr* transmettait énormément d'informations personnelles de ces utilisateurs à des tiers, dont certaines n'étaient même pas cryptées[155]. Si les données de géolocalisations sont récupérées de manière plus sécurisée par deux entreprises, *Apptimize* et *Localytics*, ces dernières ont également accès aux positions sexuelles préférées et surtout aux informations relatives au statut HIV, renseigné par l'utilisateur[156].

Des centaines d'autres d'exemples sur le potentiel de l'utilisation de ces données

149. https://b og.monded p o.net/2016 12 15 Pour un popu sme numer que de gauche
150. Dans le sens ou plus le capital est important, plus la création de valeur ajoutée est importante, selon une sorte d'effet boule de neige.
151. lire par exemple http://www. emonde.fr/affa re cambr dge ana yt ca/
152. http://www.abc.net.au/rad o/programs/pm/cambr dge ana yt ca just the t p of the b g data ceberg/9573270
153. https://www.theguard an.com/techno ogy/2018/mar/31/b g data e exposed s mp y b am ng facebook wont fix rec a m pr vate nformat on
154. http://www. emonde.fr/as e pac fique/art c e/2016/01/12/ app cat on de rencontres gay gr ndr passe sous pav on ch no s 4846153 3216.htm
155. https://g thub.com/SINTEF 9012/gr ndr pr vacy eaks
156. http://www. emonde.fr/p xe s/art c e/2018/04/03/donnees pr vees e s te de rencontres gr ndr m s en cause 5279794 4408996.htm

personnelles, qu'elles soient collectées sur Internet ou ailleurs pourraient être ici développés, allant de la prédiction de la grossesse en regardant les changements de comportement de consommation [157], de l'utilisation des réseaux sociaux pour estimer la crédibilité d'une personne demandant un crédit [158], ou encore simplement pour connaître les habitudes de ces clients [159]. Nous citerons ici l'exemple d'une personne de notre entourage qui a reçu une offre de son assureur garantissant la couverture des frais d'obsèques jusqu'à un certain montant moyennant une cotisation mensuelle. Leur calcul coût/bénéfice implique une estimation de la durée de vie de 11 ans et 3 mois [160], après quoi ils perdent de l'argent. Nous n'avons aucune idée des données et des algorithmes employés qui leur ont obtenir ce chiffre.

Si la plus grande partie des données personnelles utilisées à des fins commerciales est récoltée et analysée sans que la personne concernée par ces traitements ne soit au courant, il arrive qu'une exhibition en ligne sur des sites accessibles à tous puisse suffire à reconstruire une identité numérique. Nous pensons par exemple au portrait de Marc L* (Meltz, 2008) démontrant qu'il est très simple de trouver la date de naissance, le numéro de téléphone, de reconstituer les réseaux relationnels et professionnels d'une personne si cette dernière fait preuve de négligence sur ce qu'elle publie en ligne. Il convient donc d'être vigilant, car il est extrêmement simple pour un acteur quelconque d'agir comme une sorte d'éboueur du web et de ramasser des informations personnelles, potentiellement intimes (figure 48). Il existe cependant en France et en Europe des appareils législatifs indépendants qui encadrent l'utilisation des données personnelles.

FIGURE 48 La métaphore des éboueurs. Tiré de *« Trashed »*, de Derf Backderf, Edition « ça et là » (2015).

157. Comme ce fut le cas pour la chaîne de magasin *Target* qui a deviné qu'une adolescente était enceinte avant même que son père ne soit au courant https://www.forbes.com/s tes/kashm rh /2012/02/16/how target figured out a teen g r was pregnant before her father d d/#7426ab896668 ou encore http://www.nyt mes.com/2012/02/19/magaz ne/shopp ng hab ts.htm

158. https://www.amer canbanker.com/news/banks to use soc a med a data for oans and pr c ng

159. http://www. efigaro.fr/secteur/h gh tech/2017/08/02/32001 20170802ARTFIG00264 e bhv asp re es donnees de ses c ents ma s est o n d etre e seu .php

160. Soit à peu près 6 ans de moins que l'espérance de vie moyenne.

4 L'appareil législatif : la CNIL

Si les finalités et motivations sont différentes d'une entreprise de services numérique, ces traces numériques et données personnelles peuvent aussi intéresser les chercheurs, notamment en sciences humaines. Lorsque ces données sont récoltées de manière *ex-situ,* ces derniers vont tenter d'inférer des critères à minima démographiques, géographiques et socio-économiques pour analyser leur niveau de représentativité et en retirer le plus d'information exploitable pour mener à bien leur recherche. Cela dit, une grande différence réside dans l'intérêt porté à la personne dont on utilise des données qu'elle a pu créer. Alors que les compagnies d'assurances, de marketing, de publicité, etc. ont tout intérêt à connaître l'identité (numérique ou réelle) de la personne qu'elles tracent [161], les chercheurs sur les mobilités vont plutôt considérer les traces numériques d'une personne comme étant un élément d'un échantillon dont l'analyse individuelle ne présente pas d'intérêt scientifique aux regards des comportements de déplacement généraux et des distributions statistiques de l'ensemble de l'échantillon (distance parcourue, nombre de lieux fréquentés, dispersion autour du domicile, etc.). Mais tous travaux, indépendamment de leur finalité (marketing ou recherche), doivent être encadré par la Commission Nationale Informatique et Liberté (CNIL) dès lors qu'il touche à des données personnelles afin de préserver la vie privée et le relatif anonymat des personnes concernées.

4.1 Contexte de création

• • • LE MONDE — 21 mars 1974 — Page 9

JUSTICE

Tandis que le ministère de l'intérieur développe la centralisation de ses renseignements

Une division de l'informatique est créée à la chancellerie

En ordre dispersé, les départements ministériels tentent de développer à leur profit, à leur seul usage, l'informatique et son outil, l'ordinateur. Ce n'est pas tout à fait un hasard si, à l'époque où le Journal officiel *va publier un arrêté créant une « division de l'informatique » au ministère de la justice, celui de l'intérieur met la dernière main à la mise en route d'un ordinateur puissant destiné à rassembler la masse énorme des renseignements grapillés sur tout le territoire; pas un hasard non plus si le projet SAFARI (Système automatisé pour les fichiers administratifs et le répertoire des individus) destiné à définir chaque Français par un « identifiant », qui ne définisse que lui, maintenant terminé, est l'objet de convoitises ardentes; le ministère de l'intérieur y souhaite jouer le premier rôle. En effet, une telle banque de données, soubassement opérationnel de toute autre collecte de renseignements, donnera à qui la possédera, une puissance sans égale.*

Ainsi se trouve d'évidence posé un problème fondamental, même s'il est rebattu : celui des rapports des libertés publiques et de l'informatique. Son importance exigerait qu'il en fût, au Parlement, publiquement débattu. Tel ne paraît pas être, pourtant, la solution envisagée par le premier ministre dans les directives qu'il vient d'adresser au ministère de la justice, intéressé au premier chef si l'on s'en rapporte à la Constitution qui dans son article 66 fait de l'autorité judiciaire le gardien des libertés individuelles.

« Safari » ou la chasse aux Français

FIGURE 49 Coupure de presse du monde, 21 mars 1974. Source : http://bugbrother.blog.lemonde.fr/2010/12/23/safari-et-la-nouvelle-chasse-aux-francais/

En 1974, Philippe Boucher révèle dans Le Monde l'existence du projet SAFARI (Système Automatisé pour les Fichiers Administratifs et le Répertoire des Individus) mis en place par le ministère de l'Intérieur [162, 163](figure 49). Comme l'explique le site du Sénat, « Ce système prévoyait de créer une base de données centralisée de la population, en utilisant le fichier de

161. En plus des comportements collectifs.

162. http://www.mag securs.com/news/art c etype/art c ev ew/art c e d/23700/de safar a edv ge 35 annees d8217une h sto re oub ee ma gre a creat on de a cn .aspx

163. http://bugbrother.b og. emonde.fr/2010/12/23/safar et a nouve e chasse aux franca s/

sécurité sociale comme identifiant commun à tous les fichiers administratifs »[164].

Un reportage de Bernard Rapp sur Antenne 2 en 1976 précise la nature du projet : « Il s'agit de centraliser au ministère de l'Intérieur, grâce à un puissant ordinateur, près de 100 millions de fiches réparties dans les quelque 400 fichiers des services de police. Pour certains, c'est la porte ouverte à la mise en fiche des citoyens. En effet, en rassemblant les données enregistrées dans les mémoires et les fichiers des divers services publics ou parapublics telles que les services de police, ministère de la justice, ministère des armées, sécurité sociale, banques, etc., en rassemblant ces données, on peut, d'une seule pression sur un seul bouton, tout savoir sur un individu »,[165]. Au final, devant le tollé suscité par la mise en lumière de ce projet, ce dernier fut retiré et une commission Informatique et Liberté fut créée afin de proposer une réglementation sur l'utilisation de l'informatique et des données. C'est ainsi que la loi 78-17 « Informatique et liberté » fut votée le 6 janvier 1978, actant la création d'une autorité indépendante : la Commission Nationale Informatique et Liberté ou CNIL. La France rejoint alors quelques états ayant adoptés des initiatives similaires, comme la Suède (1973), les États-Unis (1974), la république fédérale d'Allemagne et la Belgique (1976)[11]. Un groupe de travail rassemblant 29 CNIL (ou équivalents) européennes fut créé en 1995[166] et une unification des règlements des CNIL à l'échelle européenne à vu le jour en mai 2018[167].

Dans la même lignée que *SAFARI*, l'Inde a mis en place à partir de 2009 un système d'identification biométrique de l'ensemble de sa population, le programme Aadhaar[168]. Chaque citoyen est associé à un identifiant unique et doit fournir une photo d'identité, ses empreintes digitales, et deux scans de ses iris, le tout stocké dans une base de données centralisée. En janvier 2018, des journalistes ont montré qu'il était possible d'acheter un compte administrateur pour seulement 500 roupies (~6 €), ce qui permet l'accès à l'ensemble de la base de données[169].

4.2 Rôle de la CNIL

Si nous revenons à la France et à la CNIL, cette dernière a quatre missions principales vis-à-vis des particuliers et professionnels[170] :

- Informer et protéger, avec notamment la mise à disposition d'outils pratiques et

164. https://www.senat.fr/evenement/arch ves/D45/context.htm
165. http://www. na.fr/v deo/CAB7600764601
166. https://www.cn .fr/fr/ e g29 groupe des cn europeennes
167. https://www.cn .fr/fr/comprendre e reg ement europeen et http://eur ex.europa.eu/ ega content/FR/TXT/?ur =CELEX%3A32016R0679. Pour suivre les péripéties du projet de loi européen, voir le documentaire « Democracy ; La ruée vers les datas » de David Bernet (2015) http:// nfo.arte.tv/fr/ a ruee vers es datas, disponible également ici : https://www.youtube.com/watch?v=Ob0uRM IT38.
168. https://u da .gov. n/
169. https://t mesofind a. nd at mes.com/ nd a/fir fi ed n aadhaar data eak case a nfo safe says u da /art c eshow/62373114.cms
http://abonnes. emonde.fr/p anete/art c e/2018/03/02/ es rates de dent ficat on b ometr que en nde 5264646 3244.htm
170. Pour des informations plus exhaustives, voir https://www.cn .fr/fr/ es m ss ons

pédagogiques[171] ainsi que la promotion de l'utilisation des technologies de chiffrement de données[172].

- Accompagner et conseiller, c'est-à-dire aider les organismes a être en conformité avec la loi, en enregistrant dans un registre tout traitement de données à caractère personnel[173].

- Contrôler que l'usage de données personnelles par des tiers est conforme à loi informatique et liberté, et sanctionner si des infractions sont avérées[174].

- Anticiper en faisant une veille sur les technologies employées qui peuvent potentiellement porter atteinte à la vie privée, et conduire une réflexion sur les problèmes d'éthique.

Ce dernier point est primordial compte tenu de l'évolution rapide des technologies (dont le niveau « d'utilité sociale » est très variable) utilisant des données personnelles (Bahu-Leyser, 2009).

4.3 CNIL et données personnelles

4.3.1 Liste des données personnelles selon la CNIL

Les données à caractère personnelles telles que définies par la CNIL sont de natures très variées[175] et leur collecte doit faire l'objet d'une demande d'inscription au registre de

171. Mais avec seulement 77500 abonnés sur *Twitter* et 26000 sur *Facebook* en juin 2017, la visibilité de la CNIL sur Internet paraît très limitée.

172. Nous pouvons noter ici une certaine contradiction avec la loi sur le renseignement de juillet 2016, notamment l'article L852 1 du code de la sécurité intérieure relative à l'interception des correspondances électronique : https://www. eg france.gouv.fr/affichCodeArt c e.do;jsess on d= 9353749EB84796BF8A446D452B80D978.tpd a21v 2?c dTexte=LEGITEXT000025503132& dArt c e= LEGIARTI000032925430&dateTexte=20170619&categor eL en= d#LEGIARTI000032925430. Ce dernier article remet également en cause la l'article 12 de la déclaration universelle des droits de l'Homme des Nations unies, qui bien que n'ayant aucune portée juridique précise : « Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes". http://www.un.org/fr/un versa dec arat on human r ghts/.

173. « Constitue un traitement de données à caractère personnel toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. » Article 2 de la loi Informatique et liberté – https://www.cn .fr/fr/ o 78 17 du 6 janv er 1978 mod fiee#CHAPITRE1.

174. Pour une liste des sanctions prononcées par la CNIL : https://www.cn .fr/fr/ es sanct ons prononcees par a cn ; Nous pouvons noter par exemple que la société *Google*, maintenant Alphabet, valorisé à 555 milliards de dollars en février 2016 (http://www.lemonde.fr/entreprises/article/2016/02/02/alphabet *Google* devient la premiere capitalisation mondiale 4857552 1656994.html), a essuyé une amende « record » de 100 000€ pour non respect des droits d'opposition et de suppression des données.

175. Voir http://communaute un vers ta re.un v rouen.fr/ es donnees a caractere personne 477012.kjsp?RH=1435224458605&RF=1436260972935 pour a ste comp ète des données concernées.

la CNIL. En fonction du type de données, différentes déclarations sont à envisager[176]. Par exemple les données de localisation font l'objet d'une déclaration normale par le Correspondant Informatique et Liberté (CIL)[177] de l'Université. Tandis que les données sur les origines raciales ou ethniques, biométriques, génétiques, les infractions et condamnations font l'objet d'une demande d'autorisation auprès de la CNIL. Les demandes d'avis se font lorsque les finalités sont relatives à la sûreté/sécurité publique, à des mesures pénales ou de sûreté, à l'utilisation du numéro de sécurité sociale, à l'utilisation de données biométriques pour le contrôle de l'identité de personnes, ou encore lors de recensement de la population. Il faut également demander une autorisation en cas de recherche médicale ou d'évaluation de pratique de soin. L'objectif de ces démarches est d'assurer le cadre légal de l'utilisation de ces données personnelles, dès qu'elles concernent ou sont réalisées par une personne morale française.

4.3.2 « Principes clés de la protection des données personnelles »

La loi Informatique et liberté a établi 5 grands principes que toute personne ou organisme privé ou public s'apprêtant à collecter des données à caractère personnel doit respecter. Résumé de manière très succincte, cela revient à dire que toute collecte de données doit suivre des objectifs (ou finalités) bien définis dès l'origine, et que seules les informations nécessaires à la réalisation de ces objectifs doivent être récoltées auprès de personnes consentantes. Ces données doivent également être stockées pendant une durée limitée, avec des mesures de sécurités maximales pour éviter d'éventuelles fuites[178].

Par exemple, les banques sont tenues de respecter la délibération n°80-22 du 8 juillet 1980 et ne peuvent donc pas effectuer de traitements automatisés sur les données de leurs clients pour en déduire, les lieux fréquentés et d'en déduire le type d'activité qu'ils ont pu exercer en ce lieu donné en fonction du commerce fréquenté[179]. Cependant, nous pensons que les habitudes d'achat sont analysées individuellement, notamment lors de la demande de prêt bancaire.

La loi oblige également les sites Internets qui utilisent des cookies à afficher un message explicite afin de prévenir l'utilisateur : *« les internautes doivent être informés et donner leur consentement préalablement à l'insertion de traceurs. Ils doivent disposer d'une possibilité de choisir de ne pas être tracés lorsqu'ils visitent un site ou utilisent une application. Les éditeurs ont donc l'obligation de solliciter au préalable le consentement des utilisateurs. Ce consentement est valable 13 mois maximum. Certains traceurs sont cependant dispensés du recueil de ce consentement »*[180].

176. https://www.dec arat on.cn .fr/dec arat ons/dec arat on/accue .act on
177. Il s'agit d'un référent faisant l'interface entre l'université (ou tout autre organisme) et la CNIL. Pour plus d'informations voir : https://www.cn .fr/fr/ e ro e du c et ses benefices.
178. https://www.cn .fr/fr/comprendre vos ob gat ons/ es pr nc pes c es
179. https://www. eg france.gouv.fr/affichCn .do? d=CNILTEXT000017654312
180. https://www.cn .fr/fr/cook es traceurs que d t a o

Concernant le consentement des personnes, l'alinéa 3 de l'article 32 de la loi 78-17 précise cependant : *« Lorsque les données à caractère personnel n'ont pas été recueillies auprès de la personne concernée, le responsable du traitement ou son représentant doit fournir à cette dernière les informations énumérées au I dès l'enregistrement des données ou, si une communication des données à des tiers est envisagée, au plus tard lors de la première communication des données. (...) Ces dispositions ne s'appliquent pas non plus lorsque la personne concernée est déjà informée ou quand son information se révèle impossible ou exige des efforts disproportionnés par rapport à l'intérêt de la démarche»* [181].

Cet alinéa semble pouvoir s'appliquer au cas où une personne aurait accepté des conditions d'utilisation d'un service qui prévoit une utilisation et\ou un partage vers d'autres organismes des données fournies volontairement ou non par ladite personne. Ainsi, si des tiers récupèrent des données personnelles provenant d'un autre service (par exemple *Twitter*), et qu'il n'est pas possible de prévenir les personnes concernées (notamment lorsque l'envoi de message direct n'est pas possible), ces données peuvent tout de même être utilisées. Cela dit, la plupart des personnes ne lisent pas les conditions d'utilisation [182] (dont le caractère contractuel est discutable) et n'ont probablement pas conscience de la nature des différents traitements réalisés sur leurs données personnelles [183].

Créées à l'origine pour protéger les citoyens contre une volonté de surveillance étatique trop intrusive, la CNIL et la loi Informatique et Liberté sont censées être les « garde-fous » qui protègent les citoyens Français face à d'éventuelles dérives sur l'utilisation de leurs données personnelles par des organismes, qu'ils soient publics ou privés. Mais nous pouvons noter un décalage significatif entre l'ampleur colossale des missions de la commission et les moyens qui lui sont attribués, avec seulement 192 agents pour 16 millions d'euros de budget annuel [184] en 2017. De même, le montant maximum des sanctions économiques que peut attribuer la CNIL en cas de manquement grave à la loi n'est que de 150 000 euros [185], ce qui paraît bien dérisoire et peu dissuasif au regard des capitalisations boursières et fonds propres de certaines compagnies dont le système économique est en grande partie basé sur la quantité, le traitement ou encore la revente

181. https://www.cn .fr/fr/ o 78 17 du 6 janv er 1978 mod fiee#Art c e32

182. Pour une satire sur le fait d'accepter les conditions d'utilisation sans les lire, voir l'épisode 1 de la saison 15 de South Park http://southpark.cc.com/fu ep sodes/s15e01 humancent pad. Pour un documentaire plus sérieux sur le sujet, voir « Terms and condition may apply » http://tacma.net/.

183. Voir par exemple le documentaire « Nothing to Hide » https://noth ngtoh dedoc.wordpress.com/ ou le « serious game » data dealer sur le commerce des données https://datadea er.com/about .

184. https://www.cn .fr/fr/ e fonct onnement

185. Comme ce fut le cas avec *Facebook*, pour 6 chefs d'inculpations dont l'absence d'information sur l'utilisation des données : *« Les sociétés ne délivrent aucune information immédiate aux internautes sur leurs droits et sur l'utilisation qui sera faite de leurs données notamment sur le formulaire d'inscription au service »* et l'analyse des données sans consentement *: « ils [les utilisateurs] ne consentent pas à la combinaison massive de leurs données et ne peuvent s'y opposer, que ce soit lors de la création de leur compte ou a posteriori. Ils sont donc dépourvus de tout contrôle sur cette combinaison. »* https: //www.cn .fr/fr/facebook sanct onne pour de nombreux manquements a o nformat que et bertes

des données personnelles qu'elles possèdent[186]. Mais depuis mai 2018 est entré en vigueur le règlement européen UE 2016/679 du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractères personnels, ou « Règlement européen sur la protection des données ». De nombreuses clauses y figures, concernant notamment le consentement des Internautes ou doit figurer de clairement l'usage des données, leur durée de stockage, leur communication à des tiers ainsi que la possibilité de leur suppression ou de leur transfert vers d'autres plateformes[187]. D'un point de vue sanctions économique, l'article 83, alinéa 5, prévoit des « amendes administratives pouvant s'élever jusqu'à 20 000 000 EUR ou, dans le cas d'une entreprise, jusqu'à 4 % du chiffre d'affaires annuel mondial total de l'exercice précédent »[188]. Cette sanction peut s'élever potentiellement à plusieurs dizaines, voire centaines de millions d'euros. Elle devrait pousser les entreprises intéressées à se conformer à la législation en vigueur dans l'Union européenne si elles souhaitent se maintenir dans cette zone ou les données forment un marché majeur estimé à 60 milliards d'euros en 2016[189].

186. Ainsi, chez *Twitter*, la revente des données représentait 85 des 574 millions de recettes (15 %) au second trimestre 2017 https:// nvestor.tw tter nc.com/resu ts.cfm

187. Des résumés assez simple et clair sont disponibles ici : https://www.us nenouve e.com/art c e/rgpd e reg ement qu change tout.N674994 et https://www.numerama.com/tag/rgpd/ mais surtout : https:// sc nfo ex.com/2018/07/18/donnees personne es et recherche sc ent fique que e art cu at on dans e rgpd/

188. http://eur ex.europa.eu/ ega content/FR/TXT/?ur =CELEX%3A32016R0679

189. Et 300 milliards si l'on considère les retombées indirectes, http://www. emonde.fr/econom e/art c e/ 2017/05/28/des donnees personne es tres convo tees 5135092 3234.htm (article payant).

Discussion

« With the development of television, and the technical advance which made it possible to receive and transmit simultaneously on the same instrument, private life came to an end[190] *»*
George Orwell, 1949

Un nombre sans précédent de données personnelles est créé quotidiennement et accessible à des tiers. Dans la plupart des cas, la personne concernée n'a probablement pas donné son consentement. Ces données, qui alimentent un marché juteux de ciblage et profilage sont surtout utilisées par des entreprises à des fins commerciales, dont ne bénéficie pas l'intérêt général.

Les données issues d'un Internet mondialisé posent le problème de la souveraineté des états et des citoyens à leur égard. L'appareil législatif européen tend progressivement à encadrer leur usage, parfois sous l'impulsion d'une société civile très active et concernée. Nous pensons ici à la *Quadrature du net*[191], aux *Moutons numériques*[192], ou encore à la *Ligue des droits de l'Homme*[193].

Sur un plan plus technique, certaines entreprises ou associations mettent à disposition des outils censés garantir le maintien de la vie privée de leur utilisateur. Ainsi, le moteur de recherche français *Qwant*[194] s'oppose clairement à *Google* et déclare ne pas collecter d'informations individuelles sur ces utilisateurs et calibre ces algorithmes uniquement à partir de données agrégées. L'association Framasoft[195] développe des logiciels libres et open-sources, et propose une alternative a à peu près tous les services numériques disponibles, avec notamment *Framasphère*, proche de *Facebook* et calqué sur *Diaspora**[196] ou encore *Framapiaf*, un succédané de *Twitter*. Ces initiatives restent néanmoins marginales en termes d'utilisateur, et leurs possibilités d'agrandir leurs infrastructures sont plutôt limitées du fait de capitaux assez faibles.

Accéder à Internet est un droit constitutionnel, et son usage fait partie intégrante du quotidien de la plupart des personnes. Peut-être faudrait-il envisager de créer une entité publique et indépendante chargée de développer des plateformes de communication, à l'image de l'audiovisuel publique, où chaque citoyen contribuerait par l'intermédiaire du prélèvement d'un impôt. La collecte des données pourrait alors se limiter aux données protocolaires, tandis que l'utilisation des données publiées consciemment serait strictement encadrée par la législation. Mais cette dernière peut évoluer en France et en Europe, au gré de différentes élections. Aussi, la série d'attentats survenus en France depuis une dizaine d'années a par exemple entraîné le

190. *« Avec le développement de la télévision et le perfectionnement technique qui rendit possible, sur le même instrument, la réception et la transmission simultanées, ce fut la fin de la vie privée ».*
191. https://www. aquadrature.net/
192. https://mouton numer que.org/
193. https://www. dh france.org/
194. www.qwant.com
195. https://framasoft.com/
196. https://d aspora fr.org/

vote de la loi sur le renseignement, jugée par beaucoup comme plutôt préoccupante sur le plan des libertés individuelles[197]. Mais un tel système pourrait aussi conduire à la segmentation géographique des différentes plateformes et réseaux sociaux, selon les régions du monde, ce qui n'est évidemment pas souhaitable.

Cette approche ne pourrait fonctionner non plus dans des États dont la conception des droits individuels est extrêmement différente de la philosophie européenne issue de la Déclaration des droits de l'Homme, au risque de créer un État totalitaire au sens Orwellien du terme. C'est par exemple ce qui est en train d'arriver en Chine, ou l'état s'immisce dans la vie en ligne de ses citoyens, au gré d'une proximité avec les géants du net chinois (comme *Tencent* qui commercialise *WeChat*) et d'une législation aux contours parfois assez flous[198].

Nous pouvons aussi penser à la création de fondations fonctionnant sur le don, un peu à la manière de *Wikipedia*, ou simplement mieux doter les associations comme *Framasoft*. L'idée étant de sortir du cynisme et du mépris d'un système économique ou chaque individu qui utilise un service en ligne n'est qu'un consommateur en puissance, dont le profil, défini par l'accumulation des traces qu'il laisse en ligne est commercialisable.

En attendant que surviennent une évolution radicale dans la collecte et le traitement des données d'*Homo Numericus,* nous présenterons dans le chapitre suivant un état de l'art des travaux traitant des traces numériques géolocalisées, et de leur potentiel dans les études sur les mobilités. Selon les situations, ces données pourraient s'avérer être d'intérêt général (Shah, 2018), notamment en contexte épidémiologique.

197. http://abonnes.lemonde.fr/societe/article/2015/09/21/renseignement-jean-marie-delarue-se-dit-preoccupe-par-le-controle-du-renseignement_4765449_3224.html

198. http://www.slate.fr/story/159364/wechat-application-totalitaire-revee-gouvernement-chinois

Chapitre V: « Nouvelles données » et mobilités urbaines : état de l'art

Ce chapitre présente un état de l'art sur l'utilisation de données, principalement issues des CDR (Call Detail Record, soit les statistiques d'appels enregistrées aux antennes relais) ou des réseaux sociaux en ligne, dans l'analyse des mobilités humaines, notamment en contexte épidémique. Cette compilation non systématique et non exhaustive regroupe 126 articles, selon des recherches par mots clés [199] sur des sites comme *Google* scholar [200], researchgate [201], science direct [202], arxiv [203]. Quelques revues de littérature existent déjà, que cela soit sur l'utilisation des CDR en général (Blondel *et al.,* 2015), ou appliqué aux trajectoires individuelles (Chen *et al.,* 2016), ou à l'analyse des dynamiques urbaines (Calabrese *et al.,* 2014). Steiger *et al.,* (2015a) ont effectué une première revue sur l'utilisation de la géolocalisation de *Twitter*, et (Weiler *et al.,* 2015) sur la détection d'événements à partir du même réseau social. Nous utiliserons également une partie des références de ces revues qui nous paraissent pertinentes, l'objectif étant de réaliser une synthèse [204] de l'utilisation des CDR ou des données des issues des réseaux sociaux dans le cadre d'études sur les mobilités, principalement urbaine en vue d'établir un modèle de mobilité à base d'agent. À noter qu'une excellente revue de littérature sur la modélisation des mobilités vient de paraître au moment où nous finissions ce chapitre (Barbosa *et al.,* 2018), et nous ne pouvons que vous conseiller sa lecture qui pourrait permettre d'éclaircir certains aspects ou travaux cités, peut être mal expliqués dans ce chapitre.

La figure 50 montre l'évolution du nombre de papiers publiés traitant des mobilités humaines en fonction des données utilisées qui seront repris dans cette section. Les CDR et les données issues des réseaux sociaux sont très employées, notamment depuis 2010. Parmi les réseaux sociaux, la surreprésentation des études sur *Twitter* peut s'expliquer par la simplicité de la collecte des données, la diversité des thématiques notamment lorsque l'on couple des analyses de contenus à la géolocalisation, ainsi qu'une focalisation de notre part, car l'utilisation de ces données servira de base aux chapitres 6 et 8, 9 10 et 11. Avant la mise en place de la

199. Comme « Mobility », « Mobile Phone », « Location Based Social Network », « *Twitter* », « Foursquare », « epidemic ».
200. https://scholar.*Google*.fr/
201. https://www.researchgate.net/
202. www.sciencedirect.com
203. https://arx v.org/
204. Même s'il est très possible que des papiers pertinents soient passés au travers de notre radar.

géolocalisation dans les *tweets* à partir de 2009, les études portant sur les mobilités à partir des réseaux sociaux se cantonnaient surtout à l'étude de *FlickR* ou de *Brightkite*. Nous pouvons noter que la mise à disposition des données issues des GPS des taxis, comme à New York depuis 2009 [205], autorise plus d'études sur ce sujet à partir de 2011.

FIGURE 50 Répartition par année des types de données utilisées dans les études sur les mobilités, dans la revue bibliographique

Nous n'avons pas fait de statistiques sur les zones d'études, mais la plupart d'entre elles se situent aux États-Unis ou en Europe. Quelques études utilisant des CDR sont faites en Afrique, notamment dans les contextes de la Malaria, Ebola ou de la fièvre Jaune (Bengtsson *et al.*, 2015 ; Buckee *et al.*, 2017, 2013 ; Kraemer *et al.*, 2017 ; Wesolowski *et al.*, 2014, 2013, 2012), ou dans le cadre de comparaison géographique (Amini *et al.*, 2014 ; Li *et al.*, 2017 ; Yan *et al.*, 2014). Les études utilisant *Weibo*, le « *Twitter* Chinois », sont bien évidemment réalisées en Chine (Liu *et al.*, 2014 ; Y. Liu *et al.*, 2015). Les travaux se basant sur *Twitter* ou *Foursquare* sont principalement réalisés aux États-Unis, en Europe et en Australie, même si on peut noter pour *Twitter* des études en Inde (Cebeillac et Rault, 2017), en Thaïlande (Cebeillac *et al.*,

205. http://www.nyc.gov/htm /t c/htm /about/tr p record data.shtm

2018 ; Cebeillac et Le Bigot, 2018), ou aux Philippines (Coberly *et al.*, 2014).

Ces études sont classées en fonction des thématiques abordées, selon qu'elles cherchent à définir le niveau de représentativité de l'échantillon concerné, établir des lois de mobilités, des modèles d'interactions spatiales ou de déplacements, la saisie des dynamiques urbaines et la détection d'événements, le rôle des relations sociales dans les déplacements, le couplage avec une utilisation du sol pour une approche centrée sur les activités, et leur utilisation en contexte d'épidémies de maladies infectieuses.

1 Représentativité des différents jeux de données

1.1 Part de la population concernée par le jeu de données utilisé

La figure 51 ci-dessous montre que la plupart des pays sont bien équipés en téléphones portables avec cependant des taux de pénétration extrêmement variables en fonction des régions. Ainsi, parmi les 252 pays ou territoires considérés par les données de la banque mondiale, seuls 17 ont un taux de pénétration des téléphones portables inférieurs à 50 %, principalement localisés en Afrique subsaharienne, et 88 entre 50 et 100 %. Près de 60 % des pays ou territoires ont un taux de pénétration supérieur à 100 %, suggérant plus de téléphones que d'habitants.

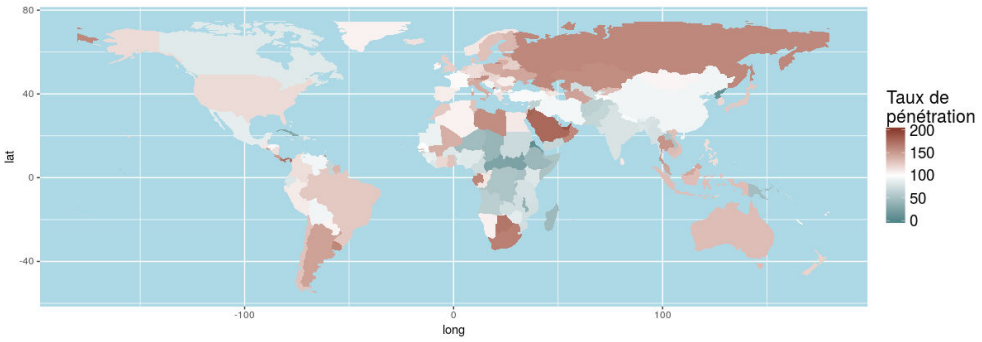


FIGURE 51 Répartition des taux de pénétration de téléphones portables dans le monde en 2015. Source : Banque mondiale - http://databank.worldbank.org/data/reports.aspx?Id=5494af8e&Report_Name=Mobile-penetration-

La figure 52 montre quant à elle le taux de pénétration d'Internet par pays. Nous pouvons observer un fossé bien plus important que sur la figure précédente, qui peut s'expliquer par le fait que les réseaux de téléphonies mobiles sont plus anciens et plus simples à déployer qu'un réseau Internet filaire classique. Cela dit, la mise en place progressive de la 3, 4 et 5G permettra aux opérateurs de proposer des connexions à Internet sur des téléphones portables (munis d'une carte réseau) dans plus de zones du monde et à moindre coût. Il est en effet probablement moins coûteux de quadriller le territoire avec des antennes relais que de dérouler des câbles jusqu'aux

domiciles de toute une population.

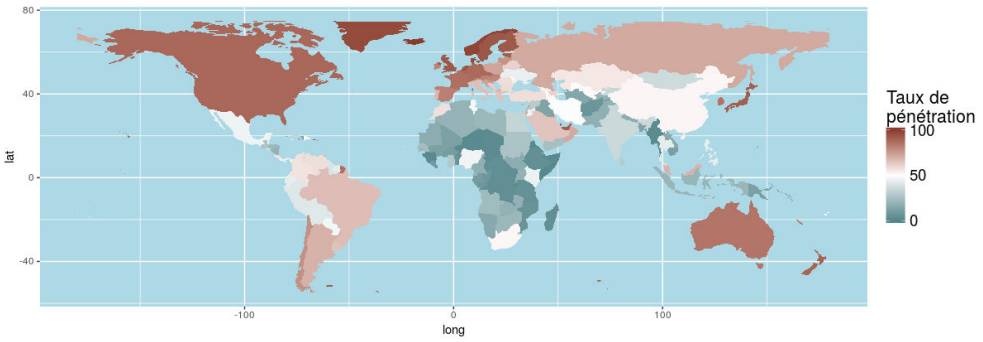


FIGURE 52 Taux de pénétration d'Internet en 2016 par pays. Source : Internet Live Stats http://www.Internetlivestats.com/Internet-users-by-country/

Ces deux figures mettent en avant quelques points importants sur le niveau de représentativité des jeux de données. En effet, l'utilisation des CDR (figure 51) dans des pays où le taux de pénétration des téléphones est supérieur à la population risque d'entraîner des doublons. Tandis que dans les pays moins développés, les personnes ne possédant pas de téléphones, ou téléphonant moins pour des raisons économiques ne seront pas présentes dans l'échantillon. Il faut aussi prendre en compte quelques pratiques régionales, comme l'utilisation de plusieurs cartes SIM d'opérateurs différents dans des régions où les fournisseurs de réseaux téléphoniques ne couvrent pas tout le pays, comme c'est le cas par exemple en Haïti (Bengtsson *et al.,* 2015). Toujours dans des pays relativement pauvres, certaines personnes peuvent utiliser le même téléphone portable (Wesolowski *et al.,* 2013), ou l'emprunter pour aller au village voisin par exemple.

Les fortes variations dans le niveau de connectivité à Internet impliquent que les quantités de données produites sur les réseaux sociaux seront très différentes en fonction des pays. De plus, dans des pays où l'accès à Internet en itinérance est plutôt réservé à une relative élite urbaine connectée, les traces laissées ne devraient pas être représentatives de la population. En plus d'un accès à Internet différenciés entre les pays riches et pauvres et très probablement entre les villes et les campagnes, il convient d'apprécier l'utilisation des différents réseaux sociaux en fonction des régions du monde (figure 53).

En guise d'indicateur du niveau d'utilisation des différents réseaux sociaux en ligne dans le monde, nous avons récupéré des données des 50 sites les plus visités par pays d'après le site www.alexa.com [206], un service *d'Amazon* qui mesure la popularité et le trafic sur des sites Internet, grâce notamment à des utilisateurs ayant installé des « add-ons » dans leur moteur de recherche.

206. www.alexa.com/topsites

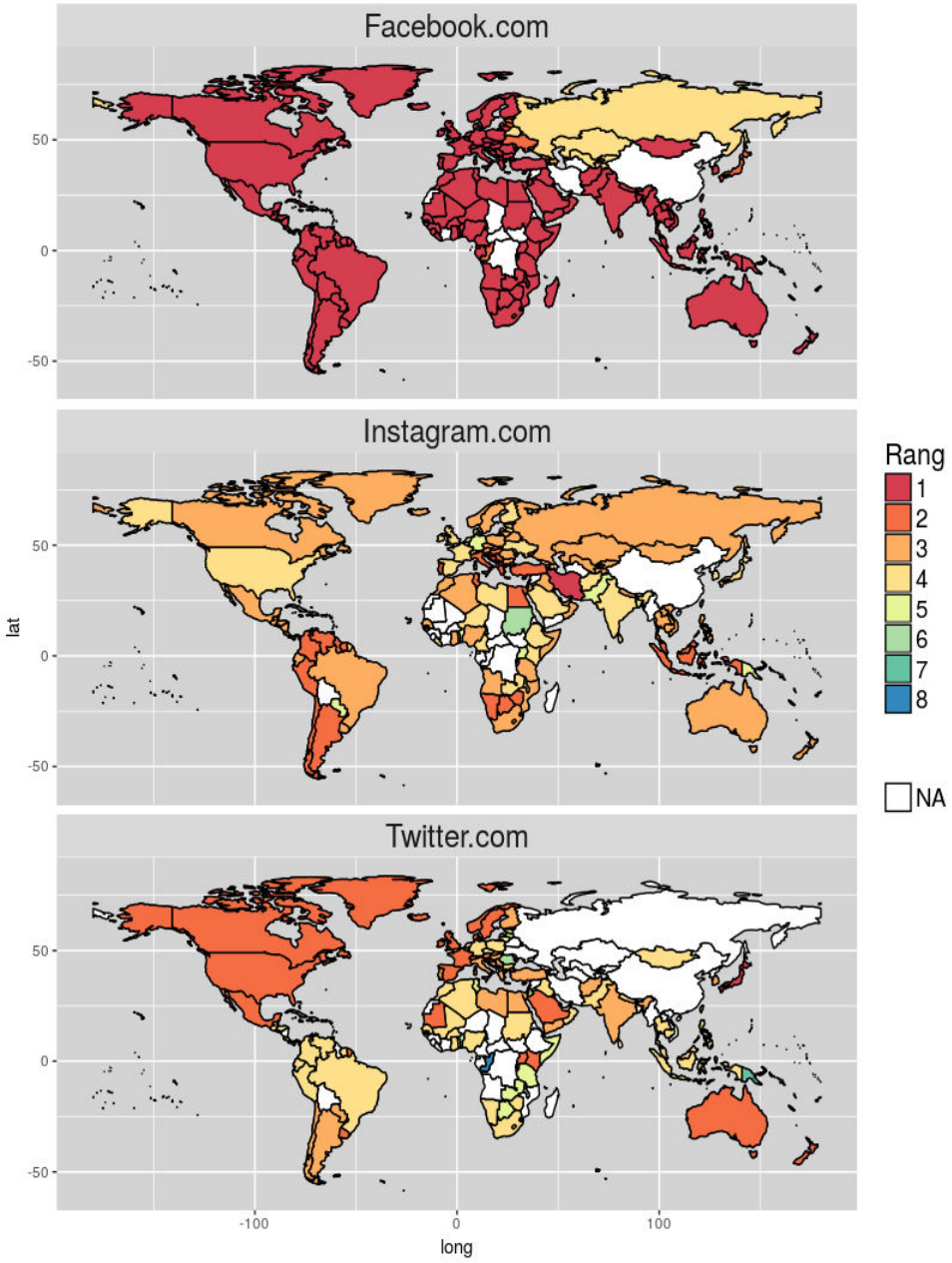


FIGURE 53 Rang par pays pour *Facebook*, *Twitter* et Instagram. À noter que malgré le fait que *Twitter* ait moins d'utilisateurs qu'Instagram, il demeure plus utilisé dans plus de pays. Source : alexa.com

Parmi ces sites visités, nous avons classé par pays les sites considérés comme des réseaux sociaux et la figure 53 montre le rang de trois d'entre eux : *Facebook*, *Instagram* et *Twitter*. *Facebook* est de loin le réseau social le plus utilisé, avec plus de 2 milliards d'utilisateurs en juillet 2017, même s'il est absent en Chine et en Iran, où il est remplacé par un service équivalent. Alors qu'*Instagram* compte environ 700 millions de membres, soit 2 fois plus que *Twitter*, il semble être détrôné dans beaucoup de pays par ce dernier, notamment en Europe et en Amérique du

Nord. Bien que *Foursquare*, fort de 50 millions d'utilisateurs sert de base à beaucoup d'études, il ne figure dans aucun des tops 50 des sites les plus influents pour alexa.com. Ainsi, lorsqu'il s'agit de parler de la représentativité de l'échantillon, il semble pertinent d'apprécier dans un premier temps la part d'utilisateurs du service dont les données sont utilisées pour étudier les mobilités.

1.2 Représentativité et connaissance de l'échantillon

Plus important encore que la taille de l'échantillon, le niveau de représentativité des données recueillies de manière *ex situ* doit pouvoir être quantifié. Cette vérification n'est pourtant pas systématique, que cela soit dans des études portant sur des données issues de la téléphonie mobile ou des réseaux sociaux.

1.2.1 Valider avec des données de recensement

La première approche serait de comparer les données récoltées aux recensements de population, pour apprécier les résidus dans les distributions spatiales des populations. Typiquement, cela revient à estimer la population d'utilisateurs vivant dans une entité géographique et de confronter les résultats à la population officielle. Jurdak *et al.* (2015a) et J. Liu *et al.* (2015) ont fait des études à partir des données *Twitter* en Australie et ont considéré le domicile des utilisateurs comme étant l'unité administrative où l'utilisateur avait le plus *tweeté*. Il en ressort un coefficient de Pearson global à 0.816 (R^2 de 0.665), mais les résultats sont meilleurs à l'échelle nationale que métropolitaine, suggérant une différence d'usage en fonction des villes. La même équipe a par la suite travaillé à l'échelle urbaine et considéré que la population d'un quartier était le nombre d'utilisateurs unique ayant visité ce quartier (Khan *et al.,* 2017), ce qui peut expliquer le faible niveau de représentativité de leurs résultats.

La méthode qui nous semble la plus appropriée est dans un premier temps d'estimer la localisation du domicile d'un utilisateur et de comparer ensuite avec des données de recensement. Ce travail a été fait à partir de données téléphonies en France et au Portugal, où le domicile a été estimé comme étant le lieu où la personne a le plus souvent utilisé son téléphone entre 20 h et 07 h (Deville *et al.,* 2014). Il en ressort un coefficient de Pearson de 0.89, ce qui est un peu moins bon que leur estimation à partir d'images satellites, mais qui permet néanmoins d'apprécier les dynamiques des populations au cours du temps. Une autre étude a estimé la population de la même manière, en se basant sur des jeux de données issues de la téléphonie mobile et du réseau social *Twitter* à Barcelone et à Madrid (Lenormand *et al.,* 2014). Les deux jeux de données sont bien corrélés avec le recensement, même si les données téléphonies offrent de meilleurs résultats et que le niveau de représentativité dépend la taille de la cellule de base.

Mais bien que ces résultats soient valables pour les deux plus grandes villes d'Espagne, ce n'est absolument pas le cas à Delhi, où l'usage de *Twitter* n'est absolument pas représentatif de la population (chapitre 6), et serait plutôt l'apanage des tranches sociales moyennes supérieures (Cebeillac et Rault, 2016) Certaines études estiment la localisation du domicile en fonction de l'activité nocturne sur *Twitter*, élaborent des modèles, mais ne comparent pas les domiciles estimés aux recensements (notamment Kurkcu *et al.,* 2016), ce qui nous laisse perplexes.

Si l'approche la plus classique pour l'estimation du domicile se base sur les horaires d'envois de messages, certaines études utilisent une analyse sémantique pour définir le lieu de domicile. Ainsi, Steiger et al. (2015b) ont repris des mots clés, et bien qu'ils arrivent à peu près à corréler les lieux de travail avec les recensements (R^2 de 0.75), leur méthode ne fonctionne pas pour les domiciles (R^2 à 0.08).

Les données issues de la téléphonie sont plus représentatives de la population que des données issues des réseaux sociaux comme *Twitter* ou *Foursquare*, ne serait-ce que part la plus grande utilisation de téléphone portable par rapport aux réseaux sociaux. La grande quantité de données d'utilisateurs est souvent mise en avant pour justifier de la crédibilité de l'étude, sans pour autant en estimer les couches sociales correspondantes. Il est pourtant admis que les mobilités individuelles sont très influencées par des critères socio-économiques et cognitifs (Kaufmann et Jemelin, 2004). Il convient dès lors d'arriver à estimer les catégories socio-professionnelles des personnes présentes dans l'échantillon, notamment dans le cas de l'utilisation de données issues des réseaux sociaux, où finalement peu d'informations sont disponibles sur les statuts des utilisateurs, car souvent déclaratifs et parcimonieux.

1.2.2 Caractérisation socio-économique et démographique de l'échantillon

Le recoupement entre des données téléphonies et des comptes bancaires (e.g. Lenormand *et al.,* 2015 ; Luo *et al.,* 2017) ou l'utilisation de données de *check-in* provenant d'un nombre important de comptes certifiés d'un réseau social (*e.g.* Zhong *et al.,* 2015) sont sans conteste de bonnes approches pour estimer les catégories socio-économiques des individus présents dans l'échantillon. L'étude de transactions bancaires géolocalisées dans deux villes espagnoles a montré des tendances de déplacements différents en fonction de l'âge, du sexe et du travail (Lenormand *et al.,* 2015a). Une autre étude basée sur l'analyse d'environ 160 000 comptes certifiés sur le réseau social chinois *Weibo*[207], contenant des millions de *check-in* associés à des profils assez complets[208], a permis de mettre en avant des mobilités différenciées à Beijing et Shanghai (Zhong *et al.,* 2015).

Mais ce genre de données est très difficilement accessible et pas forcément disponible dans

207. L'équivalent de *Twitter* et *Facebook* en Chine.

208. Date de naissance, genre, statut marital, niveau d'éducation, etc., le tout vérifié par l'envoi d'un scan de la pièce d'identité et d'un certificat d'études.

toutes les régions du monde. En leur absence, l'âge, le sexe, et les catégories socio-économiques doivent donc être déduits en utilisant d'autres approches. Tout d'abord, pour ce qui est des données issues de *Twitter*, une première méthode est d'analyser à la fois le profil déclaré des personnes, qui consiste en un pseudonyme, une photo et quelques lignes descriptives, ainsi que le contenu de messages. Sloan *et al.* (2015) ont ainsi inféré les âges et les occupations des personnes au Royaume-Uni, et ont remarqué la jeunesse relative des utilisateurs de *Twitter* par rapport à la courbe des âges issue des recensements. En utilisant une approche bayésienne selon le lieu de résidence estimé et du nom de l'utilisateur, Luo *et al.* (2016) ont développé une méthode permettant de détecter des communautés « ethniques » à Chicago. Pour connaître le genre, ils ont utilisé d'une part le prénom de la personne, et ont cherché d'autre part les utilisateurs dans la base de données de *Facebook*, où le genre est plus souvent renseigné que sur *Twitter*.

D'autres études ont utilisé les services de *Google*, notamment le *DoubleClick Ad Planner*[209], une régie publicitaire qui installe des *trackers* et crée des profils individuels en fonction des comportements en ligne. En associant des comptes de *Foursquare* à Austin et à Chicago à ce type d'information, il fut alors possible de reconstruire la structure socio-économique et démographique de l'échantillon (Jin *et al.,* 2014 ; Yang *et al.,* 2015).

Il est également possible, en ayant détecté le domicile des personnes et connaissant (1) la valeur de la taxe foncière du quartier et (2) le type de lieux fréquentés, d'estimer la catégorie socio-économique des individus, notamment des plus aisés (Cebeillac et Rault, 2016). Cette approche par les lieux fréquentés permet également de détecter des étudiants (qui visitent régulièrement des universités, voir chapitre 11) ou des touristes. Ces derniers, bien que souvent écartés des études car considérés comme non-résidents, ont de par leur comportement de mobilité potentiellement caractérisé par un grand nombre de lieux fréquentés à différentes échelles spatiales (Cebeillac et Le Bigot, 2018). Ils ont également un système immunitaire naïf et un rôle prépondérant dans la diffusion d'épidémies de dengue à l'échelle nationale et mondiale (voir chapitre 1).

Les premières études sur les lieux fréquentés par les touristes furent effectuées à partir de données issues de partage de photographie sur des plateformes comme *FlickR* (Girardin *et al.,* 2008, 2007 ; Y. Sun *et al.,* 2013) ou *Panoramio* (Schlieder et Matyas, 2009) et on fait ressortir les dynamiques temporelles des lieux touristiques. Une étude de large envergure a montré que les déplacements à l'échelle de la planète, enregistrés sur le réseau social *Twitter* étaient très fidèles aux données des flux de passagers aériens et le nombre de touristes à destination, et reproduisent assez fidèlement les périodes de vacances des pays d'origine (Hawelka *et al.,* 2014). Ceci suggère que *Twitter* peut être un bon proxy pour capter les déplacements des touristes

209. Remplacé aujourd'hui par Display Planner https://support.goog e.com/goog e ads/answer/2475441?h =fr

tout comme les hot-spots touristiques à l'échelle d'une ville comme Barcelone (Manca *et al.,* 2017) ou de la planète (Bassolas *et al.,* 2016).

1.2.3 Hybridation avec des enquêtes de terrains

Malgré la longue tradition d'enquêtes de terrain pour la collecte de données sur les mobilités, la plupart des études se cantonnent à comparer des bases de données qu'ils n'ont pas créées eux-mêmes[210]. Nous pouvons noter l'étude de Wesolovsky et al, (2015b) sur les comparaisons des mobilités entre des enquêtes de terrain effectuées dans le cadre d'études sur les déplacements dans le contexte de la propagation de la Malaria et des données téléphoniques au Kenya (Wesolowski et al., 2015b). Ce travail conclut sur la nécessité d'une hybridation des données notamment en contexte rural dans un pays en développement. Nous sommes les seuls à notre connaissance à avoir effectué des études comparatives entre des données issues des réseaux sociaux et des enquêtes de terrain, que cela soit pour les espaces d'activités des plus aisés à Delhi (Cebeillac et Rault, (2016) et partie C), ou sur des pratiques de mobilités dans un quartier de Delhi (chapitres 7 & 8), ou des touristes à Bangkok (Cebeillac et Le Bigot, 2018).

Améliorer la connaissance de l'échantillon et arriver à apprécier les biais de représentativité dans les données est à nos yeux la première chose à faire lorsque l'on travaille sur des traces numériques géolocalisées dépourvues de méta-données. Cela dit, tout dépend des objectifs des études. Par exemple, compte tenu des volumes des données récoltées dans les zones urbaines, notamment les *CDR*, il est possible d'essayer de tirer des lois de mobilités individuelles générales, et réutilisables à des fins de modélisation.

2 À la recherche de lois sur les mobilités individuelles

D'après Song *et al.,* (2010a), jusqu'au début des années 2000, à défaut d'avoir des informations suffisantes, les modélisateurs considéraient sans vraiment y croire que les activités et comportements humains individuels étaient relativement aléatoires dans le temps, et ces derniers étaient alors souvent estimés par des lois de Poisson (Barabasi, 2005 ; Haight, 1967). Cette approche a été remise en cause, d'abord par analogie à des études qui portaient sur les déplacements d'animaux (Albatros et fourmis par exemple) qui ont montré que leurs comportements pouvaient être estimés et approximés par une loi de type Lévy Flight (Edwards *et al.,* 2007 ; Shlesinger et Klafter, 1986 ; Viswanathan *et al.,* 1996). Un Lévy Flight est un cas particulier de la « random walk », ou « marche aléatoire ». Schématiquement, une marche aléatoire suit une approche itérative, où un individu se déplace à chaque pas de temps dans une direction aléatoire (avec une portée plus ou moins fixée). Les déplacements à l'instant t étant

210. Tels que des recensements, ou des enquêtes publiques de ménage déplacement.

indépendant des déplacements précédents, le système n'a donc pas de mémoire. Le Lévy Flight est un cas particulier de la marche aléatoire, où la distance parcourue entre deux itérations suit une loi de distribution proche d'une distribution exponentielle ou de puissance, mais non bornée[211] (Brockmann *et al.,* 2006a ; Shlesinger et Klafter, 1986). Ceci implique que la plupart des déplacements sont proches dans l'espace et qu'une frange infime est très espacée.

Une des premières analyses des mobilités et déplacements humains sur de longues distances s'inspirant de ces résultats a été effectuée à partir du suivi d'un demi-million de billets de banque aux États-Unis d'Amérique. Cette étude fut basée sur des données issues du site web participatif https://www.wheresgeorge.com/ qui permet à des internautes d'enregistrer le numéro de série d'un billet et de l'associer à un lieu. Ils peuvent également ajouter un commentaire sur le billet avant de le remettre en circulation. Lorsqu'une personne rentre sur le site Internet un numéro de série d'un billet qu'elle a trouvé et que ce numéro est reconnu, la personne renseigne alors sa localisation, et la distance parcourue par le billet est ainsi connue. À partir de l'étude de la propagation d'un demi-million de billets de banque, Brockmann *et al.* (2006) ont ainsi montré que la distribution des distances de déplacement entre 10 et 3 200 km décroît selon une loi puissance et que ces trajectoires peuvent être modélisées par une « continuous time random walk »[212] (Brockmann *et al.,* 2006b). Dans la même veine, à partir de l'analyse de traces GPS de 101 volontaires Rhee *et al.,* (2011) ont montré que les déplacements individuels en contexte urbain suivent une loi de Levy Flight tronquée. En d'autres termes, les individus fréquentent plus souvent des lieux proches entre eux et plus occasionnellement des lieux éloignés, ce qui parait assez intuitif.

Mais c'est avec l'utilisation de données issues des communications téléphoniques que les avancées dans la compréhension des mobilités humaines furent les plus importantes (Blondel *et al.,* 2015). Ceci s'explique par leur caractère ubiquiste, correspondant à un échantillon important de la population[213] (entre 10 et 60 % selon les opérateurs téléphoniques et les pays), avec une bonne résolution temporelle et une résolution spatiale correcte pour analyser les mobilités urbaines, de l'ordre de quelques centaines de mètres en centre-ville (Blondel *et al.,* 2015).

Ces données ont en effet permis de révéler le haut niveau de régularité spatial et temporel des mobilités humaines qui s'explique par le fait que les personnes passent le plus clair de leur temps dans quelques lieux (González et al., 2008). Ces résultats corroborent avec une étude portant sur des données téléphoniques d'un échantillon de 100 000 utilisateurs au Portugal qui montre que le nombre moyen de lieux fréquemment visités est de 2.14 et que 95 % des personnes fréquentent régulièrement moins de 4 lieux différents (Csáji *et al.,* 2013). Une étude portant sur les connexions d'appareils mobiles sur différents réseaux sans fil publics (généralement

211. Heavy tailed distribution, où loi de distribution à queue lourde.
212. « Marche aléatoire continue dans le temps », soit une autre variante de la « random walk », où la distance parcourue et la durée entre deux itérations dépendent d'une « heavy tailed distribution ».
213. Même si les jeunes (< 10 15 ans) s'avèrent évidemment sous représentés.

universitaire) a également montré que la durée entre deux connexions au même réseau suit une loi de puissance tronquée, ce qui sous-entend une assez faible diversité dans les lieux fréquentés (Chaintreau *et al.,* 2007). Ce même genre de données a montré que les déplacements des étudiants dans le campus de Dartmouth suivaient une loi log-normale, selon des directions calquées sur routes et les chemins (Kim *et al.,* 2006).

González *et al.,* (2008) ont remarqué que comme dans l'étude de Brockmann *et al.,* (2006b), les distances des trajets individuels agrégés forment également une loi de puissance tronquée, résultats des caractéristiques de déplacement individuels. La même conclusion fut tirée à partir de l'analyse des données *Foursquare* (Noulas *et al.,* 2012). Dans la lignée des travaux de l'équipe de González, un modèle de déplacement a été proposé où lorsqu'un utilisateur décide de changer de lieu, ce dernier peut soit visiter un nouveau lieu avec une probabilité qui décroît avec le nombre de lieux déjà visités, soit revenir dans un lieu déjà visité modèle EPR, pour Exploration and Preferential Return (Song *et al.,* 2010a). Cette formalisation relativement simple a permis d'expliquer à la fois l'augmentation du nombre de lieux fréquentés au cours du temps ainsi que la distribution de la probabilité de présence en chaque lieu. La même équipe a également étudié le caractère entropique[214] des mobilités individuelles de 50 000 utilisateurs téléphonant suffisamment régulièrement pour semer des traces relativement continues dans le temps. Ils en ont conclu que 93 % des déplacements pouvaient être prédits, et que ce taux était constant sur l'ensemble de la population, indépendamment des distances généralement parcourues (Song *et al.,* 2010b). Néanmoins, comme le souligne Pablo Jensen dans son ouvrage « pourquoi la société ne se laisse pas mettre en équations ? » (Jensen, 2018), il convient de relativiser ce résultat. En effet, nous passons généralement la plupart de notre temps entre notre domicile et notre lieu de travail, à des horaires relativement réguliers. Il y a donc de fortes chances qu'a l'heure suivante, un individu soit encore présent dans ce type de lieu, surtout la nuit ou lors des heures ouvrées.

D'autres auteurs ont aussi utilisé un échantillon issu de données téléphoniques (un million de personnes sur 4 mois), pour créer un algorithme prédictif pour estimer la future localisation des utilisateurs de leur échantillon (Calabrese *et al.,* 2010). En se basant sur les lieux visités précédemment par une personne et calibré par les lieux visités par l'ensemble des individus, leur algorithme a pu estimer correctement 60 % des prochains lieux visités par une personne. L'utilisation de chaînes markoviennes a permis à (Li et Chen, 2009) de prédire le prochain lieu visité pour 49 % de leur échantillon d'utilisateurs du réseau social *Brightkite*[215].

Selon une approche centrée non pas sur les distances parcourues mais sur les interactions entre les lieux par individus (par exemple A $\Rightarrow$ B $\Rightarrow$ C $\Rightarrow$ A, ou A $\leftrightarrow$ B $\Rightarrow$ C), Schneider *et al.* (2013) ont montré qu'il n'existait que 17 de ces combinaisons, ou « motifs », quelles que

214. L'entropie d'un système caractérise son côté chaotique ou prévisible voir chapitre 9.
215. Une plateforme permettant le partage de localisation, de texte et de photos https://br ghtk te.com/.

soient les villes qu'ils ont étudiées (Stuttgart, Chicago ou Paris). Ceci permettrait d'améliorer les simulations, notamment lors de la génération d'individus synthétiques.

Partant de ces régularités spatio-temporelles, de Montjoye *et al.* (2013) ont montré qu'à partir de quatre traces d'appels téléphoniques prises aléatoirement dans le temps et dans l'espace, il était possible d'identifier 95 % des personnes de leur échantillon de manière unique. Ceci sous-entend l'existence d'une signature spatio-temporelle propre à chaque individu et soulève donc des questionnements éthiques sur la vie privée et l'usage de ce type de données [216]. À ce propos, des chercheurs ont montré qu'il était possible de dés-anonymiser un jeu de donnée téléphonique avec un taux de réussite de l'ordre de 45 % (Gambs *et al.,* 2014).

Si les données massives, notamment téléphoniques, ont permis d'établir des lois de déplacement prétendues générales, il s'avère que les mobilités individuelles varient grandement en fonction des pays, comme l'a montré une étude comparative entre le Portugal et la Côte d'Ivoire (Amini *et al.,* 2014). De plus, une étude basée sur les cartes de métro à Londres a montré que la distance des déplacements suivait une loi binomiale décroissante (Roth *et al.,* 2011), alors qu'elle suit une distribution gamma lorsque l'on se base sur les GPS des taxis à Lisbonne (Veloso *et al.,* 2011), ou encore une loi exponentielle avec l'analyse des trajectoires de voitures privées à Florence (Bazzani *et al.,* 2010). Ces hétérogénéités dans les distributions des déplacements, tant sur leurs distances que sur les fonctions qui les caractérisent sont résumés dans (Alessandretti *et al.,* 2017). Il en ressort qu'environ deux tiers des études trouvent des lois de puissances, environ un quart des lois exponentielles, et environ 15 % des lois log-normales, polynomiales, ou des négatives binomiales. Une étude suppose que la nature des fonctions représentant la distribution des distances de déplacement serait due à l'organisation et la répartition des populations dans les zones étudiées, généralement très dense au centre et décroissante vers la périphérie ce qui implique qu'un grand nombre de personnes se déplace sur des distances relativement faibles, et très peu sur des distances plus importantes (Liang *et al.,* 2013).

Ainsi, il n'y a pas vraiment de consensus autour des lois caractérisant les mobilités individuelles, même s'il y a un certain accord autour d'une loi de puissance pour décrire les distributions des distances de déplacements de l'ensemble de la population (Schneider *et al.,* 2013). Mais les lois résultantes de l'analyse de traces numériques semblent tout de même dépendre du type de données, de la zone d'étude et de l'échelle de l'analyse, ce qui rend délicat leur transposition dans d'autres régions du monde où les comportements de déplacement sont probablement différents et spécifiques. Il convient donc d'acquérir des données adaptées à la zone d'études et de s'assurer du niveau de représentativité plutôt que d'appliquer des lois de déplacements fonctionnant par exemple à l'échelle des districts aux États-Unis à des situations

216. Ainsi, remplacer le nom d'une personne par un identifiant peut s'avérer insuffisant pour maintenir son anonymat.

géographiques très différentes, comme à Delhi ou Bangkok [217].

Au-delà des différentes lois qui sont censées régir les déplacements humains, les études de (González *et al.*, 2008 ; Song *et al.*, 2010a) mettent en évidence que trois paramètres sont primordiaux pour décrire les mobilités :

(1) La distribution des distances des trajets, soit la probabilité qu'une personne parcoure une certaine distance entre deux lieux distincts.

(2) Le rayon de giration, qui est un indicateur de dispersion. Utilisé à l'origine pour décrire le comportement d'un objet autour d'un axe, le rayon de giration a été appliqué à l'analyse des portées et des trajectoires de mobilités par (González *et al.*, 2008). Il s'agit d'un indice r_g qui mesure le niveau de dispersion pour un individu, à partir d'un lieu central et l'ensemble des lieux de son espace d'activité. Il se calcule comme suit :

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\bar{a}_i - \bar{a}_{cm}|^2} \tag{13}$$

Où n est le nombre de lieux fréquentés par une personne, a la localisation de chacun des lieux i et a_{cm} la localisation du lieu de référence (domicile, barycentre, etc.). Une faible valeur du rayon de giration indique une faible portée des déplacements, tandis qu'une valeur élevée indique que l'utilisateur se déplace sur de plus longues distances.

(3) Le nombre de lieux fréquenté au cours du temps, avec notamment la fréquence de retour de retour en un lieu.

3 Analyse des interactions spatiales

Outre l'obtention de lois globales régissant les déplacements, ces nouvelles données (CDR et réseaux sociaux) permettent d'analyser les interactions spatiales entre différents lieux à différentes échelles. Les paragraphes suivants traiteront dans un premier temps de l'utilisation de ces données pour créer des matrices Origine Destination (OD), premier pas vers l'utilisation et la calibration de modèles d'interactions spatiales.

3.1 Création de matrices origine-destination

Les matrices OD sont un outil essentiel dans la compréhension des mobilités. D'échelles spatiales et temporelles très variées en fonction des problématiques, elles consistent en une estimation du nombre de personnes vivant dans une zone i qui se déplace vers une zone j. À

217. Même si les aspects routiniers et centrés sur quelques lieux des mobilités quotidiennes peuvent s'apparenter à des généralités transposables dans la plupart des villes du globe.

l'échelle d'un pays ou de la planète, elles permettent de connaître les flux migratoires entre différentes zones. À l'échelle de la ville, elles donnent un aperçu du nombre de navetteurs (personnes commutant quotidiennement) d'un quartier vers un autre, généralement pour des raisons professionnelles. De telles matrices sont primordiales, notamment pour l'optimisation des transports publics et la politique de la ville (Calabrese *et al.,* 2011a).

L'estimation de ces flux se fait classiquement par de longues enquêtes de type « Domicile Travail » ou « ménage-déplacement » (Commenges, 2013). Bien que coûteuses en temps, en main d'œuvre, et mise à jour peu fréquemment (en général tous les 5-10 ans) (Calabrese *et al.,* 2011a ; Yang *et al.,* 2015), l'échantillon concerné est censé être représentatif de la population. La réalisation d'enquêtes par GPS pour tracer les déplacements d'un échantillon consentant (Wolf *et al.,* 2001) permet d'avoir des résultats plus précis sur les pratiques des mobilités, mais ces dernières prennent plus de temps et coûtent très cher, notamment lorsque la taille de l'échantillon doit être importante pour que les résultats soient significatifs. D'autres méthodes, basées notamment sur la mise en place de points de comptages et le dénombrement du nombre de voitures sortant et entrant dans une zone permet d'en estimer le solde des déplacements quotidiens (Bell, 1983).

Mais l'utilisation de nouvelles sources de données, issues des communications téléphoniques ou des réseaux sociaux orientés sur la géolocalisation, permettrait de créer ces matrices à moindre coût et de les actualiser plus fréquemment. Ainsi l'utilisation des statistiques d'appel a permis d'établir ces matrices, notamment à Boston, et ces dernières, en plus d'être corrélées aux données issues d'enquêtes de recensement, faisaient également ressortir les déplacements en semaine, en week-end ainsi que leur caractère saisonnier (Calabrese *et al.,* 2011a). En se basant sur les fréquences et les horaires d'appels, il est également possible de reconstruire les flux domiciles / travail et les flux vers d'autres activités, ce qui suggère que ces données sont adaptées dans le cadre de modélisations des mobilités centrées sur les activités (Alexander *et al.,* 2015). Appliqué à 31 villes espagnoles, et en se focalisant sur les flux domicile travail, ces données téléphonies ont permis de mettre en évidence la prépondérance des flux entre zones mixtes (*i.e.* centre d'activité couplé à du résidentiel), par rapport aux autres types de zones (Louail *et al.,* 2015b). Ainsi, l'utilisation des données de téléphonies permet de bien estimer les commutations à l'échelle urbaine (Frias-Martinez et al., 2012).

Cela dit, ces données bien que présentant des avantages indiscutables en termes de taille de l'échantillon sont difficilement accessibles, car elles requièrent la signature de contrats et de clauses de confidentialités auprès des opérateurs téléphoniques, pas toujours enclins à partager ce type de données sensibles.

L'utilisation de données issues de réseaux sociaux dont la vocation est le partage public de localisation semble être une alternative prometteuse. Ainsi, comme expliqué précédemment,

Foursquare permet aux utilisateurs du service de faire des « *check ins* », c'est-à-dire de spécifier qu'ils ont fréquenté un lieu donné. À partir de données contenant la localisation des lieux et le nombre de personnes qui s'y sont enregistrées, des études ont pu dériver des matrices origine-destination assez corrélées avec des enquêtes domicile travail pour les villes de Chicago et Austin (Jin *et al.,* 2014 ; Yang *et al.,* 2015). De même, à partir de *tweets* géolocalisés à Los Angeles, Gao *et al.,* (2014) ont pu recréer des matrices origine-destination en accord avec le recensement local. Pour cela, ils ont agrégé les tweets selon la même unité spatiale que les données *in situ*, et ont trié les *tweets* par utilisateurs et par date afin de repérer les changements de zones de chacun des utilisateurs. Une fois créées, ces matrices OD permettent d'étudier les flux entre diverses zones, permettant de qualifier et quantifier les interactions spatiales.

3.2 Modèle gravitaire

À la fin du XIXe siècle, Ravenstein (1885) met en évidence à partir d'observations empiriques sur les migrations que les flux entre deux zones décroissent avec la distance[218]. 85 ans plus tard, Tobler énonce son principe : « *everything is related to everything else, but near things are more related than distant things* » (Tobler, 1970). Ainsi, depuis plus d'un siècle on cherche à modéliser à partir d'équations relativement simples les paramètres déterminant ces flux de mobilités. La paternité des modèles d'interactions revient souvent à William J. Reilly, qui propose en 1931 une équation mathématique permettant d'estimer les aires de marchés entre deux villes, soit la ligne de partage de leur influence géographique (Reilly, 1931) :

$$d_{ip} = \frac{d_{ij}}{1 + \sqrt{(M_j/N_i)}} \tag{14}$$

Où d_{ij} est la distance entre la ville *i* et *j*, M et N les populations respectives en j et i, et d_{ip} la distance entre la ville i et le point d'équilibre. Cela dit, il s'agit plus d'un modèle de position que d'un modèle d'interaction puisqu'il permet de décrire les lieux plus que leur relation[219].

En reprenant les observations de Ravenstein (1885) et l'approche mathématique de (Reilly, 1931), (Stewart, 1942) propose un premier modèle de potentiel gravitaire. Ce dernier sera repris et développé, notamment par Dodd (1950) et Zipf (1946) (Commenges, 2016). Même si certains estiment que la première formulation rigoureuse du modèle est celle de Casey (1955) (Ortúzar S. et Willumsen, 2011). Cette dernière, aussi appelée *modèle gravitaire à deux paramètres*[220], reprend les équations de Newton et se formule comme suit :

218. Claude Grasland, Interaction spatiale, accessible : http://www.hypergeo.eu/sp p.php?art c e2.
219. http://gras and.scr pt.un v par s d derot.fr/go303/ch5/doc ch5.htm
220. Car ne dépendant que de la constante C et du coefficient de fiction γ.

$$F_{ij} = C\frac{m_i n_j}{d_{ij}^{\gamma}} \tag{15}$$

Où F_{ij} est le flux de *i* vers j, *m* et *n*, les populations en zone *i* et *j*, d_{ij} la distance entre *i* et *j*, C une constante, et γ un paramètre à calibrer qui représente la rugosité de l'espace entre *i* et *j*. Cette équation montre que les flux d'une zone vers une autre dépendent de la population de ces zones et sont pondérés par une variable de friction, ici la distance[221] mise à une puissance. Ceci implique que les flux sont plus importants vers les lieux les plus peuplés, et que le nombre de flux décroît rapidement avec la distance entre les zones. Un peu plus tard, Alonso (1976) introduira un modèle gravitaire à quatre paramètres, définis par :

$$F_{ij} = C\frac{m_i^{\alpha} n_j^{\beta}}{d_{ij}^{\gamma}} \tag{16}$$

Où α et β modulent l'importance des flux des zones *i* et *j*. Dans ces deux approches génériques du modèle gravitaire, connaissant la distance et les populations, les flux F_{ij} peuvent être estimés en posant des hypothèses sur les valeurs des différents paramètres. Mais un intérêt de ces modèles est surtout la caractérisation des interactions entre les lieux d'un territoire. Si les flux F_{ij} sont connus, il est alors possible de déduire les différents paramètres en résolvant l'équation. Une première étape est d'appliquer une fonction logarithmique aux différents arguments, afin de supprimer les exposants. L'application de régressions (multi) linéaires permet ensuite d'estimer ces coefficients γ, α *et* β. Ainsi, plus le paramètre γ est élevé, plus la friction est importante, c'est-à-dire que les personnes auront tendance à se déplacer sur de plus courtes distances, favorisant la proximité à la densité. Les paramètres α et β permettent quant à eux de rendre de compte de l'importance des déplacements en fonction des zones étudiées, soit la propension des habitants d'un lieu à se déplacer. Ces deux modèles sont présentés ici sous leur forme la plus simple, et il est possible de les complexifier en ajoutant d'autres paramètres explicatifs. Pour une explication détaillée des modèles gravitaires, voir l'ouvrage (Ortúzar S. et Willumsen, 2011). La publication de Chen,(2015) contient également une revue bibliographique sur les différentes formalisations de la force de friction.

À partir de données issues d'un réseau social chinois de partage de localisation, Liu *et al.* (2014) ont pu calibrer un modèle gravitaire décrivant les flux entre les villes chinoises. L'utilisation des photos géolocalisées et disponibles publiquement sur le site *Flickr* a également permis de construire un modèle gravitaire permettant d'étudier les voyages internationaux et d'investiguer sur l'influence de la popularité de certains lieux dans le choix des voyages (Beiro *et al.,* 2016 ; Yuan et Medel, 2016). Des données issues du réseau social *Twitter* ont également permis de calibrer des modèles gravitaires (2 et 4 paramètres) en Australie, à différentes échelles (Pays, états et métropoles) (Jurdak *et al.,* 2015a ; Khan et al., 2017 ; J. Liu *et al.,* 2015). De

221. Ou n'importe quelle autre variable de coût, comme la durée du trajet par exemple.

manière générale, la popularité du modèle gravitaire fait qu'il s'agit souvent du premier modèle auquel on confronte de nouvelles sources de données de mobilité, notamment lorsque l'on estime des flux entre différentes zones (Commenges, 2016). Mais d'autres modèles géographiques, plus ou moins inspirés de modèles ou de propriété physiques sont également utilisés.

3.3 Modèle d'opportunité et modèle radiatif

L'utilisation des données téléphoniques a permis à Simini *et al.* (2012) de proposer et de valider un nouveau type de modèle de commutation entre districts, le modèle radiatif, censé dépasser les limites des modèles gravitaires conventionnels, notamment par l'absence de paramètres de calibration. Ce modèle de déplacement, inspiré des lois de la physique des particules, est très proche du modèle d'opportunité de (Stouffer, 1940) dont le principe veut que « le nombre d'individus se déplaçant à une distance donnée est proportionnel au nombre d'opportunités offertes à cette distance et inversement proportionnel au nombre d'opportunités interposées » [222]. Ce dernier est formulé comme :

$$T_{ij} = k_i m_i (e^{-\alpha x_{ji}} - e^{-\alpha x_j}) \tag{17}$$

Où *m* est la population de la zone *i*, x_j les opportunités en j, x_{ji} les opportunités entre *i* et *j*, alpha une constante et k_i un paramètre d'ajustement. En reprenant le principe des opportunités, ou d'absorptions potentielles Simini *et al.,* (2012) formulèrent le modèle radiatif tel que :

$$T_{ij} = T_i \frac{m_i n_j}{(m_i + n_j)(m_i + n_j + S_{ij})} \tag{18}$$

Ce qui signifie que le nombre de personnes vivant en zone i, de population m et allant en zone *j*, de population *n*, ne dépend que du nombre de personnes *Ti* quittant la zone *i*, de la population *i* et *j*, et de l'ensemble de la population dans un cercle de centre *i* et de rayon *ij* (S_{ij}). En somme, plus la population entre *i* et *j* est importante, moins il y aura de personnes allant de *i* vers *j,* car ils seront « absorbés » en chemin, ayant trouvé d'autres opportunités.

Ce modèle, présenté comme universel, permettrait alors de s'affranchir des données issues de la téléphonie (ou autres) pour créer facilement des matrices d'interactions entre différents quartiers, villes ou régions. Mais il dépend malgré tout d'une bonne connaissance de la distribution des populations et notamment du nombre de personnes qui commutent, et ces données ne sont pas forcément connues dans toutes les zones du monde (Blondel *et al.,* 2015). De plus, si ce modèle estime assez bien les flux lointains vers des zones peu peuplées, il serait moins bon que le modèle gravitaire pour estimer les flux proches vers des zones très peuplées (Commenges, 2016). Une étude comparative de ces deux modèles, effectuée à diverses échelles

222. Cité par (Commenges, 2016).

(macro au Pays de Galles et en Angleterre et urbaines à Londres), avec différentes données (flux de transports estimés à partir du nombre des fréquences des navettes, recensement sur les navetteurs et de la population) a montré que le modèle gravitaire donnait cependant de meilleurs résultats dans l'estimation des flux que le modèle radiatif (Masucci et al., 2013). Cette conclusion est partagée par d'autres études comparatives, appliquées aux données d'enquêtes ménage-déplacement à diverses dates en Île-de-France (Commenges, 2016), dans divers pays d'Europe et d'Amérique du Nord (Lenormand *et al.,* 2016a) et à des données issues du réseau social *Twitter* et à diverses échelles (pays, région et ville) en Australie (Jurdak *et al.,* 2015a ; Khan et al., 2017 ; J. Liu *et al.,* 2015). En comparant les tendances de mobilités urbaines au modèle radiatif à Pékin, Londres, Chicago et Los Angeles[223], une étude balaye tout simplement l'usage de ce modèle en contexte urbain (Liang *et al.,* 2013). Le modèle radiatif fut cependant utilisé à Dhaka (Bangladesh) dans le contexte de la propagation du Choléra (Perez-Saez *et al.,* 2016) et conjointement avec un modèle gravitaire pour apprécier le rôle des mobilités interrégionales entre l'Angola et la République « Démocratique » du Congo dans la propagation de la Fièvre Jaune (Kraemer *et al.,* 2017).

En alternative au modèle radiatif, Yan *et al.* (2014) ont proposé un modèle d'opportunité pondéré par la population (Population-weighted opportunities, PWO), explicité comme :

$$T_{ij} = T_i \frac{m_j(1/S_{ji} - 1/M)}{\sum_{k \neq i}^{K} m_k(1/S_{ki} - 1/M)} \tag{19}$$

Le nominateur correspond au nombre de personnes en *j*, multiplié par l'inverse de la population dans le cercle de centre *j* et de rayon *ji* auquel on soustrait l'inverse population totale *M.* Le tout est ensuite pondéré par la somme des populations en chaque lieu, divisé par les populations dans tous les cercles de centre les destinations *k* et de rayon *ki* moins l'inverse de la population totale. Appliqué à 4 villes avec des jeux de données différents[224], ce modèle a donné de meilleurs résultats que le modèle radiatif pour estimer les distances globales parcourues (Yan *et al.,* 2014). Nous pouvons cependant noter qu'il requiert énormément de calculs de population S_{ki}, ainsi que le nombre de personnes T_i sortant d'une zone *i*. Et même si l'écriture de cette équation est plutôt simple (peu de paramètre) elle nous parait peu intuitive.

Ces différents modèles d'interactions spatiales permettent d'estimer les flux globaux entre deux zones. D'un point de vue individuel, un individu résidant dans une zone donnée aura une probabilité d'aller dans un autre secteur selon les flux estimés entre ces deux zones. Ils peuvent donc servir dans des modèles individu-centrés, notamment à base d'agents. Néanmoins, les CDR ou les données longitudinales issues des réseaux sociaux permettent aussi de connaître

223. Avec respectivement les enregistrements GPS de 10 000 taxis, 5 % des trajets sur des cartes de transports et des recensements pour les villes américaines).
224. Des enregistrements de Taxi pour Pékin et Shenzen, des donnes issues de téléphones portables pour Abidjan, et des enquêtes déplacements pour Chicago.

les caractéristiques individuelles de déplacements, ce qui devrait rendre d'autant plus aisée la modélisation à base d'agents.

4 Modélisation à base d'agents

Comme vu précédemment, les récentes observations faîtes sur ces jeux de données, notamment le fait que les mobilités individuelles sont plutôt prédictibles (González *et al.*, 2008 ; Song *et al.*, 2010a, 2010b), que la plupart des personnes fréquentent souvent les mêmes lieux (Csáji *et al.*, 2013 ; González et al., 2008), ou les descriptions mathématiques des comportements de déplacement collectifs sont des éléments de connaissance pertinents pour établir des modèles de mobilités urbaines individu-centrés.

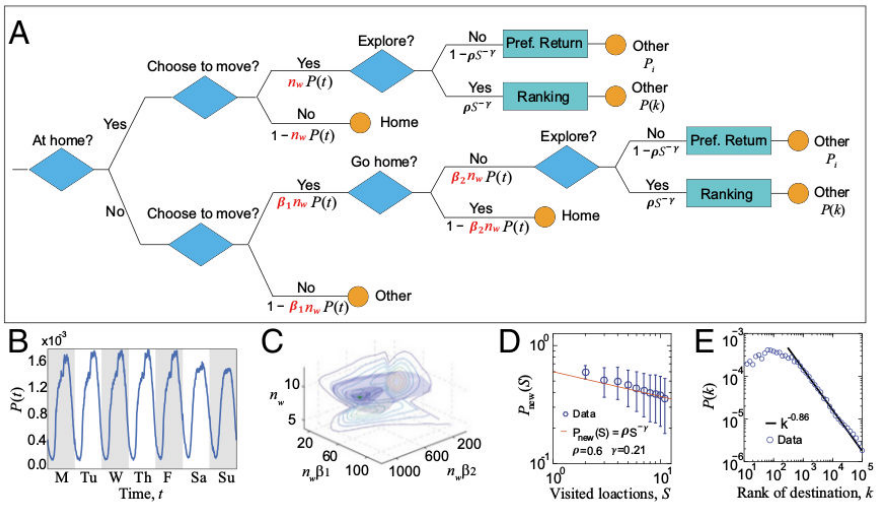


FIGURE 54 Le modèle r-EPR proposé par Jiang *et al.*, (2016).(A) représente les choix possibles en fonction du lieu et du temps, par étape, contrôlé par des paramètres individuels. (B) montre la probabilité de se déplacer par tranche horaire, issue des CDR, et pour les personnes qui ne commutent pas vers leur lieu de travail. (C) met en relation les différents paramètres utilisés en (A).(D) est la loi de probabilité de visiter un nouveau lieu et (E) la probabilité de choisir un lieu de rang k.

À partir de leurs observations sur les CDR, Song *et al.* (2010a) ont introduit le modèle *EPR*, pour « Exploration and Preferential Return ». Dans un premier temps, chaque agent situé dans un lieu a une probabilité de retourner dans un des lieux qu'il a déjà fréquenté, et une probabilité complémentaire de visiter d'autres lieux, ou d'explorer. Lorsqu'un agent retourne vers un lieu qu'il a déjà fréquenté, la probabilité de choisir un de ces lieux est alors proportionnelle au nombre de fois qu'il a déjà visité chaque lieu, conformément à la régularité observée par (González *et al.*, 2008). Lorsqu'un agent est en phase d'exploration, il va visiter de nouveaux lieux en suivant un random-walk dont la distribution des sauts (ou distances) suit une loi puissance tronquée.

Les deux parties qui composent ce modèle ont par la suite été améliorées. Ainsi, pour le choix du lieu dans le cadre d'un retour, Barbosa *et al.,* (2015) ont modulé la probabilité de sélectionner un lieu par le niveau d'ancienneté, en plus de la fréquence de visite. Jiang et al. (2016) ont modulé le choix des lieux préférentiels en fonction du fait qu'un individu se déplace à partir de son domicile ou non. Ils ont également ajouté lors de la phase d'exploration, une notion de rang, basée sur la distance des lieux de destination potentiels en fonction du lieu d'origine ce qui signifie que parmi tous les lieux que l'individu peut explorer, il devrait choisir les plus proches (figure 54).

D'autres auteurs (Pappalardo *et al.,* 2016a ; Pappalardo et Simini, 2017a), se sont basés sur des CDR et sur des données de GPS de voitures et ont pondéré le choix des lieux à explorer en fonction de l'attractivité de ces derniers basés sur le nombre de visites de l'ensemble des agents. Plus un lieu est fréquenté, plus un agent aura une probabilité élevée de s'y rendre lors de ces phases d'explorations. Ils proposent également un modèle où un agenda de mobilité est généré pour chaque agent (soit une liste de lieux, sans prise en compte de l'activité qui s'y exerce), à partir duquel sont déduites des trajectoires de déplacement.

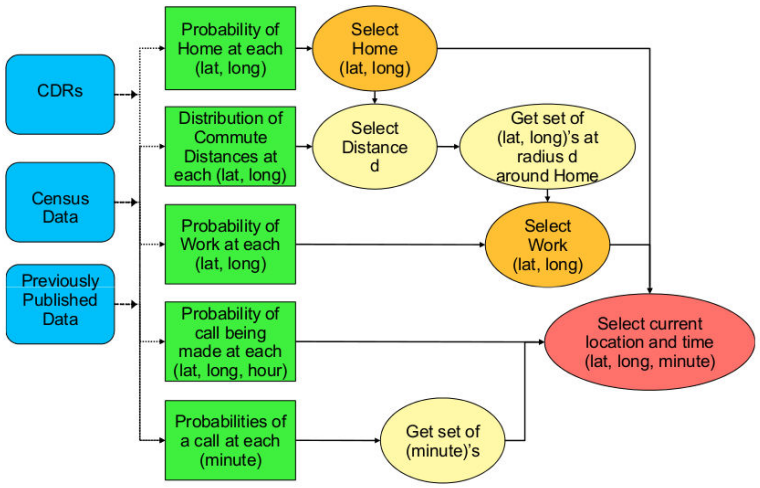


FIGURE 55 Représentation schématique du modèle WHERE, mis en place par Isaacman *et al.* (2012)

Isaacman *et al.* (2012) ont proposé le modèle *WHERE*, qui bien que calibré à partir de CDR peut exploiter diverses sources de données, dès lors qu'elles permettent d'obtenir des distributions de distance lors des changements de lieux (figure 55). Ce modèle produit des distributions réalistes des densités temporelles des agents, notamment à New York et Los-Angeles, mais ne prend pas en compte pour l'instant d'autres activités que le travail.

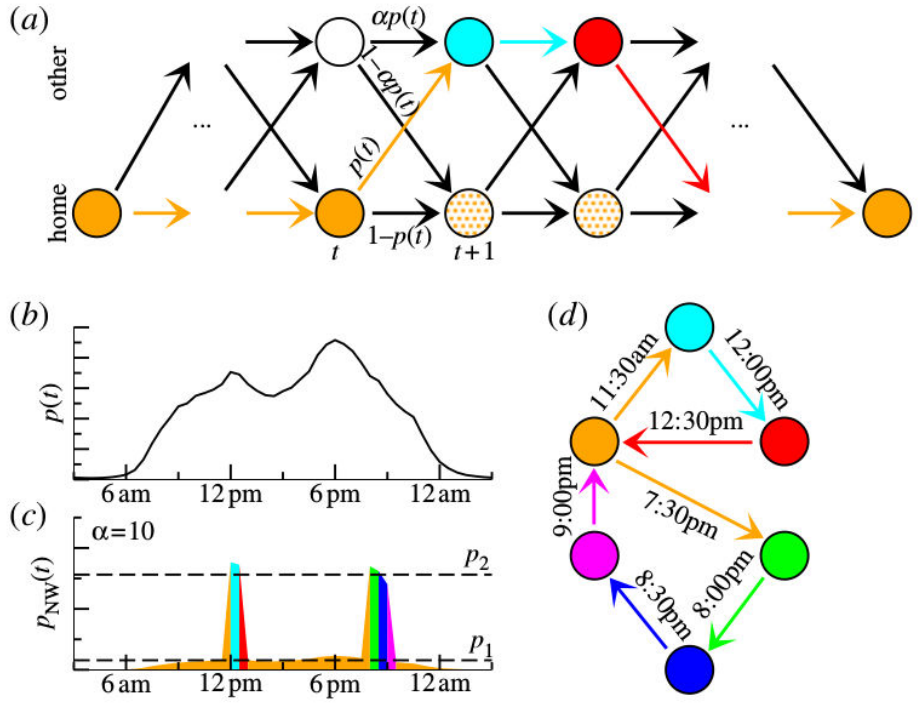


FIGURE 56 Présentation du modèle de Schneider *et al.* (2013) pour un agent ne travaillant pas. (a) montre les trajectoires possibles qu'un agent peut suivre, en commençant et finissant par son domicile. (b) montre la distribution des probabilités de déplacement issue des CDR. (c) est un exemple d'affectation de localisation en fonction de l'heure, p1 étant la probabilité d'être au domicile et p2 la probabilité d'être ailleurs. Le changement de lieux par tranche horaire pour cet agent est finalement résumé en (d).

Après avoir montré d'après des CDR et des données de recensements que les déplacements quotidiens individuels pouvaient être répartis selon 17 motifs, soit des petits réseaux de lieux visités (par exemple A → B, ou A → B → C → A, etc) Schneider *et al.,* (2013) ont créé un modèle à base d'agents qui peuvent se trouver dans trois types de lieux (Domicile, travail et autres) (figure 56). Alors que les activités de type « travail » et « rester au domicile » sont assez figées dans le temps et l'espace, ce n'est pas le cas pour les autres activités. La probabilité de réalisation de chacune de ces trois activités par tranche horaire suit une loi de distribution spécifique déduite des comportements globaux. Une fois qu'une activité est exercée par un agent, la probabilité d'effectuer une autre activité augmente de manière significative, ce qui permet aux activités flexibles moins fréquentes de pouvoir s'enchaîner, tout en ayant une durée plus courte. Ainsi, effectuer une activité autre que le travail ou de rester au domicile intervient comme une perturbation du cycle routinier, et ces activités flexibles ont plus de chance de se dérouler successivement (figure 56). Cela permet par exemple de simuler une sortie un samedi après-midi où beaucoup de lieux différents sont visités. Leur modèle, bien que calibré pour une seule journée a pu reproduire les différentes fréquences de visites et les motifs des jeux de données (CDR et recensements). Cela dit, les intervalles de temps entre les différentes

activités simulées ne reproduisent pas parfaitement les distributions observées.

Des études basées sur des réseaux sociaux de *check-in* ont également cherché à valider les hypothèses résultantes de l'analyse de leurs jeux de données en utilisant des modèles à base d'agents pour comparer les simulations aux données observées. C'est par exemple ce qu'on fait (Noulas *et al.*, 2012) qui ont utilisé des données issues du réseau social *Foursquare* pour tracer les déplacements d'environ 900 000 utilisateurs dans 5 millions de lieux dans 34 villes réparties dans 11 pays. Contrairement à d'autres études qui ont dû se contenter de données agrégées aux lieux (Yang *et al.*, 2015), ils pouvaient tracer les utilisateurs de manière individuelle, comme cela peut se faire avec des données issues de *Twitter* par exemple. Inspirés par les modèles d'opportunités et partant du principe que la densité de services proposés en un lieu influence les déplacements d'une personne entre ce lieu et les autres, ils proposèrent un modèle rang-distance alliant à la fois offre et distance. Pour ce faire, ils ont calculé pour chaque couple de lieux {u,v} un $rang_u(v)$ défini comme étant le nombre de lieux w se trouvant à une distance inférieure à la distance entre u et v (figure 57).

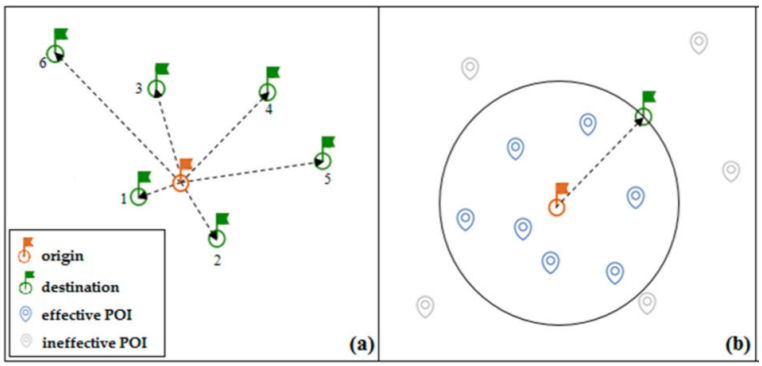


FIGURE 57 Principe du modèle rang / distance de Noulas *et al.* (2012) (a) et définition du rang entre le lieu d'origine u (en orange) et un lieu de destination v (en vert) (b). Ici, le lieu u compte 7 lieux plus proches que v, d'où un $rang_u(v)$ égal à 7. D'après (Abbasi *et al.*, 2017).

Noulas *et al.* (2012) posent ensuite la probabilité P_{uv} d'aller de la zone u à la zone v comme étant proportionnelle à l'inverse du $rang_u(v)$ à une puissance alpha :

$$P_{uv} \propto \frac{1}{rang_u(v)^{\alpha}} \tag{20}$$

avec $rang_u(v) = |\{w : d_{uw} < d_{uv}\}|$

Ils ont ensuite créé un modèle à base d'agents, où ces derniers se déplacent :

- Selon le modèle rang-distance avec une probabilité d'aller de u en v définie comme :

$$P_{uv} = \frac{rang_u(v)^{-\alpha}}{\sum_{w=1}^{W} rang_u(w)^{-\alpha}} \quad (21)$$

Où *W* correspond à l'ensemble des lieux, P_{uv} représentant le nombre de lieux entre *u* et *v* divisé par la somme des lieux entre *u* et tous les autres lieux.

- Selon un modèle gravitaire, où les masses correspondent au nombre de lieux dans un rayon donné.

Il en résulte que leur modèle gravitaire surestime les trajets de courtes distances, tandis qu'un exposant alpha commun a été trouvé pour les 34 villes de l'étude dans le cas du modèle rang-distance. De plus, les paramètres des modèles gravitaires sont beaucoup plus variables en fonction des lieux, tout comme le rayon permettant d'en définir la masse. Ils concluent sur le fait que la différence dans les déplacements enregistrés entre les villes serait plus due à des différences dans les distributions spatiales des lieux plutôt qu'a des pratiques socio-culturelles ou cognitives spécifiques (Noulas *et al.,* 2012). Cependant, compte tenu de leur échantillon, que cela soit sur le nombre d'individus (900 000 utilisateurs dans 34 villes de population comprise environ entre 1 et 20 millions d'habitants), du nombre de lieux par ville (en moyenne 11 692 lieux, ce qui paraît faible) ou de la nature même de la source (un réseau social pas si utilisé, et probablement non représentatif de la population, et dont les *check-in* sont récompensés), nous pouvons arguer le fait que même si leur approche est pertinente, innovante et très prometteuse, leur modèle ne décrit que les déplacements des utilisateurs de *Foursquare* dans les lieux de la base de données de *Foursquare*[225], et doit faire l'objet d'une confrontation avec d'autres bases de données. De plus, le formalisme de ce modèle décrit surtout l'absorption en cours de route plutôt que le caractère attractif de certains pôles urbains, notamment les zones qui regroupent beaucoup de lieux de services[226]. Ce modèle a été utilisé par la suite, notamment à partir de données issues du réseau social *Whrrl*[227], (Chen *et al.,* 2017). Une autre étude pondère ce modèle de rang par le nombre de *check-in* (issu de *Foursquare*) dans chacun des lieux, ce qui améliore le caractère prédictif du modèle (Abbasi *et al.,* 2017). Ce modèle a également été simplifié sous la forme :

225. Dennis Crowley, président du conseil de Foursquare, à propos de Swarm, l'application qui remplace maintenant l'aspect « *checkin* » de Foursquare : "Swarm was designed to help you become aware of your routines, and then nudge you to do more — encouraging you to try a new place or explore a different neighborhood. This idea of 'software that encourages you to do/see/experience things you normally wouldn't' is something that has been core to us since our early days, and I'm excited to see it live on and thriving in Swarm." dans http://geoawesomeness.com/why you need to check out the new swarm app at east once/.

226. Ce qui pourrait être remédié par exemple par l'agrégation du nombre de lieux selon un buffer pour chaque destination, ou en pondérant par une fonction de puissance décroissante en fonction de la distance entre i et j.

227. Une plateforme maintenant fermée, ou les utilisateurs pouvaient explorer, noter et partager des points d'intérêt par l'intermédiaire d'un « *check in* ».

$$T_{ij} \approx T_i \frac{P_j}{rang_i(j)\ln(M)} \quad (22)$$

Avec P_j la population en destination, et M la population totale (Liang *et al.*, 2015). Cette simplification permettrait d'obtenir des résultats convaincants, tout en s'affranchissant d'un paramètre (alpha) et de réduire la complexité des calculs.

Dans une toute autre approche et à partir de 15 millions de *check-in* répartis dans 97 000 lieux pour environ 250 000 utilisateurs du réseau social Chinois Weibo à Shanghai, Wu *et al.* (2014) ont proposé un modèle basé sur les séquences d'activités réalisées, contraintes par la localisation. Ils ont tout d'abord regroupé les activités en 6 catégories : domicile, travail, transport, restaurant, loisirs et autres. Ils ont ensuite extrait les trajets des individus, soit les différentes séquences de *check-in* réalisés avec un intervalle de temps maximum de 12 h entre deux *check-in*. Ils ont ensuite analysé les transitions entre les activités en prenant en compte l'heure de la journée par exemple travailler l'après-midi puis aller au restaurant le soir. Ceci leur a permis de construire une matrice de probabilité de transition par activité et par intervalles de temps. Ils ont aussi considéré les types d'activités en fonction de leur niveau de flexibilité spatiale, un peu comme le suggère Hägerstrand (1970), séparant les activités contraintes dans l'espace typiquement le domicile et le travail des activités plus libres. À la fin de chaque trajet, chaque agent choisit une autre activité dans un autre lieu. Cette opération est conditionnée par la matrice de transition (qui dépend de l'activité et du niveau de flexibilité de cette dernière) et la distance au lieu précédent selon une loi de puissance. Ils ont réalisé des milliers d'essais permettant de calibrer au mieux la sélection des lieux où se déroulent les activités avec les tendances de déplacements globales (distance entre deux activités). Les différents profils de fréquentation horaire des activités sont relativement bien reconstruits d'après leurs simulations, du moins pour les principaux pics et les formes globales, avec néanmoins une surreprésentation des activités nocturnes.

S'il est possible de tester et de calibrer des modèles de mobilités à base d'agents à partir de ces traces numériques géolocalisées, ces données permettent aussi de révéler des informations sur les dynamiques spatio-temporelles des villes étudiées. Néanmoins, si les sites de *check-in* comme *Foursquare* indiquent directement l'activité réalisée, ce n'est pas le cas pour les *tweets* géolocalisés et les CDR.

5 Des approches centrées sur les temporalités des activités

Dans le cadre d'approches orientées sur les lieux fréquentés, reprenant notamment le concept d'espace d'activités, il est indispensable de pouvoir croiser les lieux visités avec une activité qui s'y déroule typiquement est-ce Un lieu de domicile ? Un lieu de travail ? Un

lieu commercial ? Une école ? etc. L'objectif est alors relativement simple, c'est-à-dire arriver à associer des déplacements avec une activité à réaliser. En fonction des pays, il est possible d'utiliser des données institutionnelles, qui répertorient le nombre et le type de commerces par unité géographique, comme c'est le cas de la base de données commerces de Paris (Fleury *et al.*, 2012). Jiang *et al.*, (2012) ont par exemple exploité l'enquête sur l'utilisation du temps et des activités de Chicago, où pour chacune des 30 000 personnes de l'échantillon, les lieux et horaires des activités sont connus, pour faire ressortir les dynamiques urbaines et individuelles, et ont montré que les personnes peuvent être regroupées en 7 ou 8 groupes, en fonction de leur agenda individuel.

Cela dit, dans bien des pays ce type de données n'existe pas ou est difficilement accessible. La plupart des services de cartographies en ligne proposent maintenant en plus des routes des points d'intérêts (*POI* pour Point of Interest), c'est-à-dire des points géographiques localisés et associés à une catégorie, par exemple un type d'établissement commercial ou un service. C'est typiquement ce qui apparaît lorsque l'on cherche un cinéma ou un restaurant dans une ville sur *Google maps*, *Bing maps*, *Yahoo maps*, ou *Foursquare*. Outre le fait que ces catégories sont souvent multiples (un bar peut aussi être considéré comme un restaurant) parfois ambiguë (une école de kung-fu peut être considérée comme un lieu d'éducation) et dans certains cas créées de manière participative par des quidams, sans contrôles *a posteriori* (ce qui peut ajouter des biais), ces *POI* n'en demeurent pas moins de nouvelles sources d'informations spatiales qu'il est possible d'utiliser pour caractériser l'utilisation du sol.

Dans cette section, nous traiterons d'abord des études qui ont mis en relation des données d'utilisation du sol à des données de mobilités de type CDR ou de réseaux sociaux. Nous regarderons ensuite les quelques études qui ont cherché à catégoriser l'utilisation à partir de ces *POI* ou de données institutionnelles désagrégées, puis nous présenterons quelques travaux qui ont cherché à définir l'utilisation du sol à partir des activités enregistrées sur les antennes relais ou sur les réseaux sociaux.

5.1 Mise en relation de l'utilisation du sol et des mobilités

Une première approche peut être d'utiliser des données issues de réseaux sociaux de *check-in* comme *Foursquare* qui répertorie à la fois le nombre de personnes ayant fréquenté un établissement, et la nature de ce dernier. Nous pouvons citer Hong (2015) qui a cartographié les différences de fréquentation par quartier et par type d'établissements (restaurants, commerces, et services) à Séoul, mais sans prendre en compte les variations temporelles. De manière plus convaincante, Noulas *et al.*, (2011) ont montré les dynamiques temporelles dans différents types de lieux de *Foursquare*, en étudiant également les transitions d'un lieu de type A vers un lieu de type B. Wu *et al.* (2014) ont effectué une étude dont une partie est assez similaire à celle

de Noulas et al, sur les *check-in* issus de *Weibo* à Shanghai. L'étude des trajectoires des taxis à New-York, couplée à des données issues de *Foursquare* et du recensement a par exemple permis d'établir des profils de déplacements associés à des finalités (*e.g.* aller au restaurant dans un quartier branché) (Espín Noboa *et al.*, 2016).

D'autres auteurs (Gong *et al.*, 2015 ; X. Liu *et al.*, 2015) ont travaillé sur des données de trajets de taxi à Shanghai et ont essayé d'estimer quel endroit une personne va fréquenter à sa sortie du taxi, sachant qu'il y a plusieurs lieux à proximité. En se basant sur une base de données de 30 000 *POI* de la municipalité, ils ont mis en œuvre une méthode bayésienne permettant d'estimer la probabilité que la personne visite tel *POI*, sachant la distance au point de sortie du taxi et des hypothèses sur les horaires de fréquentation de lieux. Leurs résultats sont assez proches des études faites par la municipalité de Shanghai.

Dans une démarche plus orientée sur l'attractivité de zones de la ville, Sun *et al.* (2016) ont cherché à savoir s'il était possible d'estimer le nombre de personnes descendant à une station de métro en fonction du niveau d'attractivité du voisinage. Pour ce faire, ils se sont basés sur des données de transport en commun de la ville de Pékin, ainsi que sur 100 000 *POI* de la ville. Ils ont d'abord réparti ces derniers en 5 catégories, les lieux d'éducations, de travail, de domicile, de commerces, et de loisirs. Ils ont ensuite défini autour de chaque station une emprise géographique en fonction d'un temps de trajet en voiture ou à pied. Dans chacune de ces zones est noté le nombre de *POI* et ils ont ensuite défini un indice d'attractivité basé sur le niveau de mixité de l'utilisation du sol, ainsi que la densité des surfaces commerciales. Ils ont ensuite pu corréler le nombre de passagers par station avec leur indice d'attractivité de la station en appliquant des régressions linéaires et quadratiques.

Alhazzani *et al.* (2016) proposent un angle d'attaque intéressant essayant de comprendre quels types de lieux attirent différents types de flux. Pour cela, ils ont procédé en deux étapes. Ils ont tout d'abord utilisé des données issues de la téléphonie pour créer des matrices origine-destination à Riyad. Ils ont ensuite classé les lieux en fonction du nombre de visiteurs, de la distance parcourue par ces derniers et du niveau de dispersion spatial autour des lieux. Ils répartissent ainsi la ville en trois grandes zones, l'une correspondant au centre très commercial, une autre aux lieux résidentiels et enfin les autres zones importantes. Ils associent ensuite à chacune de ces zones les 12 000 *POI* officiels de la ville et effectuent pour chaque catégorie des tests de Fischer pour estimer la surreprésentation de certaines catégories d'utilisation du sol parmi les trois grandes zones qu'ils ont définies préalablement. Ils sont ainsi en mesure d'expliquer en partie les différents flux urbains en fonction des caractéristiques de l'utilisation du sol.

À partir de trajectoires individuelles issues de CDR dans la ville de Boston couplée à l'utilisation du sol de la ville, Jiang *et al.* (2013) ont proposé une méthode pour estimer l'activité

d'une personne, permettant d'aller au-delà des problèmes d'échelles. En effet, la localisation d'une personne se fait à une antenne relais, alors que les activités potentielles sont à une granularité plus fine. Pour cela, ils se basent sur les différentes utilisations du sol dans le rayon d'une antenne et posent des hypothèses quant à l'activité la plus plausible en sachant l'heure de la journée, la durée de l'activité et les autres activités effectuées.

5.2 Caractériser l'utilisation du sol à partir de POI

Lorsqu'un utilisateur effectue un *checkin* sur *Foursquare*, ce denier peut le partager sur *Twitter*. Dans ce cas, en plus de la géolocalisation du lieu où le *check-in* a été effectué, un lien Internet raccourci apparaît dans le message qui redirige vers le lieu en question. À partir de ce constat, Zhan et al. (2014) ont récupéré les Tweets renvoyant à des *checkin* de *Foursquare* à New-York et ont pu croiser les informations spatiales (*Twitter*) et thématiques (*Foursquare*). Ils ont ensuite réduit les 365 catégories de lieux disponibles sur *Foursquare* à 9 : Domicile, travail, restaurant, divertissement, loisirs, shopping, services sociaux, lieux d'éducations et lieux liés aux transports et déplacements. Le domicile regroupe les catégories de type « Private » et « Appartment/Condo », tandis que les lieux de travail contiennent « Office », « Co-working Space », « Tech Startup », ou encore « Design Studio ». Chaque lieu et son type sont ensuite agrégés dans des mailles de 200 m de côtés. Chaque cellule contient donc un nombre d'activités auxquelles ils ont appliqué un algorithme de classification par nuée dynamique, le *kmeans,* en 5 classes. Ils ont ainsi établi une utilisation du sol comprenant 2 types de lieux de résidences, les zones commerciales, les lieux ouverts et de divertissement ainsi que les lieux de transports. Ils ont également appliqué une classification supervisée de type Random Forest. Dans les deux cas, les résultats obtenus coïncident assez bien avec les données de la ville de New York sur l'utilisation du sol.

À partir de *POI* issus de Yahoo ![228], dont la création se fait de manière volontaire par les utilisateurs [229] Phithakkitnukoon *et al.,* (2010) ont défini 4 classes d'activité : Manger, faire du shopping, les lieux de divertissement et de loisirs. Ils ont ensuite agrégé ces *POI* dans une maille de 500 m et appliqué une kmeans en 4 classes. Chaque cellule est ensuite affectée à une de ces 4 classes selon une approche bayésienne et leur niveau de vraisemblance. Ils ont ensuite pu à partir de données téléphoniques comparer les activités à différentes heures de la journée et ont remarqué que les personnes qui travaillent dans la même zone ont des profils d'activités quotidiennes similaires. Rodrigues *et al.,* (2012) ont cherché à harmoniser les types de lieux renseignés dans les *POI* de Yahoo avec les catégories de l'utilisation du sol officielle. Ils ont pu apprécier un bon niveau d'équivalence du nombre de commerces par catégorie par blocs de recensement entre les *POI* et les données institutionnelles. La même équipe conclue plus tard

228. https://deve oper.yahoo.com/geo/p acefinder/gu de/examp es.htm
229. Le type de lieu renseigné et libre et non contraint par un cahier des charges par Yahoo.

qu'il est possible de désagréger les données sur le travail à l'échelle des *POI*, ce qui permet à échelle fine d'estimer où les personnes sont susceptibles d'aller travailler (Jiang *et al.*, 2015).

Hu *et al.* (2016) ont proposé une méthode intéressante pour cartographier l'utilisation du sol à Pékin à partir d'images satellites (Landsat 8) auxquelles ils ont intégré des *POI* d'OpenStreetMap. Ils ont appliqué aux 10 catégories de *POI* retenues d'OSM une KDE (kernel density estimation) de 500 m de portée, permettant de définir 10 nouveaux canaux auxquels ils ont incorporé des indices de végétation (NDVI) et de bâtiment (NDBI) calculés à partir des images satellites. Une classification supervisée leur a ensuite permis d'obtenir leur couche d'utilisation du sol, relativement en accord avec les données officielles.

5.3 Définir l'utilisation du sol à partir des données de mobilités

Par corollaire, il est envisageable, connaissant les horaires de fréquentation d'un type de lieux, par exemple des lieux de sorties ou des domiciles, d'en définir une signature spatio-temporelle et de qualifier ainsi les lieux en fonction de l'activité enregistrée à des antennes relais ou sur les réseaux sociaux. C'est pas exemple ce que Soto et Frías-Martínez (2011) ou Toole *et al.* (2012) ont fait à partir de données téléphoniques et de données institutionnelles sur l'utilisation du sol, où ils ont défini des signatures spatio-temporelles sur différents types de lieux sur une partie de leur base SIG et ont appliqué ensuite des algorithmes pour inférer l'utilisation du sol à partir des enregistrements téléphoniques sur l'autre partie de la base. Frias-Martinez et Frias-Martinez, (2014) ont utilisé une approche similaire à Madrid, Londres et New-York en utilisant des données *Twitter*, en montrant ainsi que l'activité sur ce réseau social permet de caractériser les types quartiers. À partir des signatures temporelles de chaque unité spatiale obtenues à partir de CDR, il est aussi possible de définir une distance entre chaque zone selon la proximité des profils et d'appliquer ensuite des algorithmes de partitionnement (*e.g.* Louvain, voir chapitre 10). Cette approche a permis de définir l'utilisation du sol dans 5 villes espagnoles (Lenormand *et al.*, 2015c).

À partir de données et de trajectoires individuelles et parfois sans connaissance de l'utilisation du sol, l'activité d'une personne en un lieu peut toujours être estimée et approximée en posant des hypothèses sur l'heure et la durée de fréquentation d'un lieu donné (Huang *et al.*, 2010 ; Xie *et al.*, 2009).

Après avoir construit des réseaux de mobilités individuels à partir d'environ 150 000 traces de déplacements en voiture, Rinzivillo *et al.* (2014) ont utilisé des trajets dont les objectifs de déplacements ont été annotés (*e.g.* aller au travail, ou aller faire des courses) comme données d'entraînement à leur méthode de classification des activités en cascade (Activity-Based Cascading). Leur algorithme permettrait d'obtenir une meilleure précision dans l'estimation de

l'activité que d'autres algorithmes utilisés précédemment (par exemple Random Forest).

Ainsi, les traces numériques permettent, associées ou non à des données sur l'utilisation du sol, de décrire les structures et les temporalités des activités dans les villes. Il existe encore d'autres applications possibles, que nous allons rapidement aborder dans la section suivante.

6 Autres études en contexte urbain

6.1 Détection d'événements

L'analyse de l'activité à l'instant t, tant sur les réseaux sociaux que des CDR permet également de détecter des événements de diverse nature. Le principe est simple : il suffit d'avoir une base de données, contenant par exemple le nombre d'appels émis par tranche horaire, sur une période suffisamment longue, ce qui permet d'isoler le bruit de fond et d'obtenir une activité moyenne. Ainsi lorsqu'à l'instant t l'activité enregistrée présente des caractéristiques éloignées de la normale, il est très probable qu'un événement se produise. Candia *et al.* (2008) se sont par exemple intéressés aux activités moyennes enregistrées par des antennes relais et ont appliqué une analyse de variance pour détecter les activités anormales.

D'autres études similaires ont été appliquées à des *CDR*, mais la possibilité de relier une activité anormale avec une analyse de messages, comme pour le cas de *Twitter* a suscité un assez grand intérêt dans la littérature. Ainsi, de nombreuses études ont utilisé cette approche pour détecter des événements (Abel *et al.,* 2012 ; Becker *et al.,* 2011 ; Boettcher et Lee, 2012 ; Chae *et al.,* 2012 ; Yardi et Boyd, 2010), notamment des épidémies (Lampos et Cristianini, 2010 ; Sofean, 2012), ou des catastrophes naturelles (Murthy et Longwell, 2013 ; Sakaki *et al.,* 2010). Nous vous renvoyons à (Steiger *et al.,* 2015a ; Weiler *et al.,* 2015) pour une revue bibliographique très détaillée.

Nous pouvons noter les travaux de Coberly *et al.* (2014) qui proposèrent une étude pilote dont l'objectif était de savoir si l'augmentation de mots clés associés à la dengue pouvait permettre à *Twitter* d'être un proxy dans la détection des épidémies de dengue aux Philippines. Il en ressort beaucoup de limites, que cela soit sur le nombre d'utilisateurs insuffisant aux Philippines, les problèmes de langues, la variation des mots clés en fonction des zones. Mais cette approche simple et peu coûteuse pourrait être envisagée comme un appoint au système de surveillance traditionnel, au même titre que les activités enregistrées sur le moteur de recherche de *Google*[230].

Certains chercheurs ont également combiné des données issues de *Twitter* et d'*Instagram*,

230. https://www.*Google*.org/flutrends/about/ pour des modèles de prédictions spécifiques à la dengue et à la grippe. Ou plus simplement : https://trends.*Google*.fr/trends/ pour une recherche par mots clés spécifique.

ce qui, toujours appliqué à la détection d'événements, a permis de réduire le nombre de faux négatifs et de faux positifs (Giridhar *et al.*, 2017 ; Giridhar et Abdelzaher, 2017).

Si nous revenons maintenant à des aspects plus individuels, les moyens de communication servent à connecter les membres d'un même réseau (amis, familles, travail). Les données issues des réseaux sociaux en lignes ou des CDR peuvent donc permettre d'analyser l'impact de tiers sur les déplacements d'un individu.

6.2 Rôle du réseau social dans les mobilités

« Elle a son réseau, j'ai mon réseau, mon beau frère a un réseau, on a tous un réseau, mais on ne mélange pas. » Pierre Mondy, dans 'la 7e compagnie au clair de Lune'.

Car au-delà des navettes entre le domicile et le travail, et les activités très routinières comme faire des courses, certains des autres déplacements en zones urbaines sont très probablement influencés par le réseau social d'une personne, au sens relations humaines dans la vie réelle. Par exemple rendre visite à ses amis, aller boire un verre dans un bar, ou au cinéma, etc. sont autant d'activités réalisées par plusieurs membres d'un même groupe socio-affectif. La localisation et le moment où ses activités seront effectuées devraient donc dépendre d'un compromis trouvé entre ces différents individus. Et l'utilisation de traces numériques géolocalisées a permis à quelques études de rendre compte de ces phénomènes.

Les jeux de données issues de la téléphonie mobile contiennent souvent, selon les accords avec les opérateurs, à la fois la localisation d'un individu, mais également des informations sur les personnes contactées. Ainsi, selon une étude, la plupart des personnes qui s'appellent se sont rencontrées physiquement au moins une fois dans l'année, et la fréquence des rencontres est très corrélée avec celle des appels et la distance qui les sépare (Calabrese *et al.*, 2011b). Cette étude montre également que deux personnes qui s'appellent alors qu'elles sont dans la même cellule d'un BTS ont de fortes chances de se rencontrer physiquement [231].

À partir de jeux de données variés (téléphonie et réseaux sociaux tels que *Gowalla* et *Brightkite*) il a également été montré que les relations sociales influencent les déplacements sur de longues distances (Cho *et al.*, 2011). Ces observations, assez évidentes (si nos amis habitent loin de chez nous, leur rendre visite influence nos déplacements sur de longues distances) permettent cependant, grâce aux données massives disponibles de réaliser des modèles de déplacement (prédictif ou non) et contraint par ces liens sociaux (Frias-Martinez *et al.*, 2011 ; Toole *et al.*, 2015 ; Wang *et al.*, 2011).

Les réseaux sociaux ne sont pas en reste, puisqu'un grand nombre de comptes partage publiquement la liste des personnes qu'ils suivent et qui les suivent, et l'utilisation de ces

231. Typique du « Allô ? T'es où ?» .

informations permet par exemple d'améliorer des algorithmes de localisation qui n'utilisent que des informations contenues dans les messages (Cheng *et al.,* 2010). Une des rares études publiées par des chercheurs de *Facebook* montre que la prise en compte des « amis » des personnes permet de prédire à une assez bonne résolution (plus précise que celle des adresses IP) la localisation de centaines de millions de personnes (Backstrom *et al.,* 2010).

Ces données issues de plateformes en ligne ou de la téléphonie permettent d'analyser et de modéliser des tendances de déplacement, de détecter des événements, ou encore d'analyser l'influence du réseau social d'un individu sur sa mobilité. Elles ne sont néanmoins pas exemptées de critiques. Au-delà de points de vue purement scientifique et technique, elles impliquent des enjeux éthiques, du fait que leur utilisation est déconnectée des motifs de leur création[232] et des retombées sur la vie privée des personnes si elles sont mal anonymisées et protégées. Cela dit, elles permettent de combler un vide dans des régions où peu de données institutionnelles sont disponibles, et ces dernières régions, en générales plus défavorisées sont souvent le théâtre de nombreuses épidémies où les mobilités humaines jouent un rôle prépondérant dans leur propagation (Buckee *et al.,* 2013 ; Oliver *et al.,* 2015 ; Wesolowski *et al.,* 2015a)

7 Utilisation en contexte d'épidémies – maladies infectieuses

Privilégiant une approche pragmatique plutôt que dogmatique, certains chercheurs se sont réunis autour d'une association, *Flowminder*[233], et utilisent des nombreux jeux de données, notamment des données téléphonies anonymisées pour étudier le rôle des déplacements humains dans la propagation des épidémies. Leurs cas d'études sont très concrets et nombreux, par exemple dans le cas de l'épidémie d'Ebola de 2014 (Wesolowski *et al.,* 2014) où leurs travaux soulignent qu'un accès rapide et coordonné aux données de téléphonies permettrait d'estimer la propagation de l'épidémie en temps réel. Et l'utilité de telles données ne se limite pas à Ebola. Ces données sur les dynamiques des populations sont également primordiales dans le cas de la malaria à l'échelle régionale (Wesolowski *et al.,* 2012 ; zu Erbach-Schoenberg *et al.,* 2016), et comme pour la dengue, les déplacements humains ont une contribution plus importante dans la propagation de l'épidémie que le moustique lui-même. Mais si les lacunes dans les données épidémiologiques rendent difficile la modélisation de la propagation de l'épidémie (Buckee *et al.,* 2013), elles permettent d'estimer les principaux flux responsables de l'importation de la maladie et les zones à haut risque, ce qui permettrait d'améliorer le système de surveillance (Buckee *et al.,* 2017).

Ces données téléphoniques ont également permis d'établir rétrospectivement un modèle

232. Dans un but de maintien de services et de facturation pour les données téléphonies et d'expressions publiques pour les données issues des réseaux sociaux.
233. http://www.flowm nder.org/about

de déplacement à l'échelle de Haïti qui permet de mieux estimer les zones les plus à risque dans le cas de l'épidémie de choléra de 2010, qu'à partir des modèles gravitaires (Bengtsson *et al.,* 2015). Cette étude insiste également sur l'importance d'une très grande réactivité, que cela soit dans la déclaration des cas que dans le partage des données téléphoniques, car les premiers jours sont cruciaux dans l'évolution de l'épidémie.

Évidemment, le domaine de recherche sur les mobilités en contexte épidémique à partir de données téléphoniques n'est pas uniquement *trusté* par les membres de *Flowminder*. Des groupes de recherches, notamment les équipes de Tizzoni et de Frias-Martinez ont pu obtenir par eux même ce genre de données. Nous pouvons également noter l'initiative « Data for Development » [234], ou *D4D*, lancée en 2012 par l'opérateur de télécommunication Orange qui a permis à des chercheurs d'avoir accès à des CDR en Côte d'Ivoire et au Sénégal. L'objectif d'une telle mise à disposition est d'apprécier dans quelles mesures ces données peuvent « contribuer au développement et au bien-être des populations » dans les domaines de la santé, de l'agriculture, du transport de l'énergie et des statistiques nationales [235]. Grâce à ces données, de nombreuses études, que nous détaillerons ci-après, ont pu être menées au Sénégal et en Côte d'Ivoire (de Montjoye *et al.,* 2014 ; Finger *et al.,* 2016 ; Kafsi *et al.,* 2013 ; Lima *et al.,* 2015, 2013). Par exemple, à partir de ces CDR au Sénégal, Finger *et al.* (2016) ont pu créer des matrices OD variables dans le temps et ont appliqué un modèle métapopulation (voire chapitre 3) qui met en avant l'importance des rassemblements et notamment des pèlerinages dans la propagation de l'épidémie de choléra de 2005.

Alors que les précédentes études reliant mobilité et épidémies ont raisonné de manière analytique, les paragraphes suivants détaillent des travaux de modélisation de mobilité et de propagation d'épidémies faites à partir de CDR ou de données issues des réseaux sociaux.

Tizzoni *et al.* (2014) ont comparé des CDR récoltés dans 3 pays européens (France, Espagne et Portugal), avec des données de recensement et un modèle radiatif calibré à partir de ces enquêtes. Ils ont montré que leur matrice OD déduite à partir des CDR capture 87 % des flux enregistrés par les recensements, même si les flux sont surestimés avec les CDR. Ce dernier aspect a des répercussions lorsqu'ils appliquent leur modèle métapopulation SIR (voir chapitre 3) orienté dans un contexte de grippe, où l'épidémie simulée se propage plus vite lorsqu'ils utilisent les CDR. Toujours en prenant les études de mobilités institutionnelles comme référence, ils montrent que le modèle radiatif donne de meilleurs résultats que les CDR lorsque le point de départ de l'épidémie est au centre du réseau de mobilité, et notent l'inverse lors un début de propagation dans les zones périphériques. La même équipe a utilisé le même type de modèle métapopulation SIR, mais cette fois-ci en mobilisant des *tweets* géolocalisés en Europe et aux États-Unis (Tizzoni *et al.,* 2015). Le lieu de domicile d'un utilisateur est ici considéré comme étant la zone, ou bassin, cette personne a été le plus active sur le réseau, pendant

234. http://www.d4d.orange.com/en/Accue
235. http://www.d4d.orange.com/en/presentat on/the object ves of the cha enge

plus de 50 % du temps. À partir des différentes zones fréquentées autres que les domiciles, ils ont pu établir des degrés de connectivité entre chacun des bassins de population. Ils ont ensuite défini et calibré le paramètre de contact entre individus, soit le niveau d'interaction entre personnes d'une même zone suivant la méthode employée par (Schlapfer *et al.,* 2014). L'intérêt de cette étude est surtout méthodologique, car ils concluent que le niveau de contact et l'hétérogénéité des flux influencent la propagation de l'épidémie, ce qui est intrinsèque aux modèles paramétriques déterministes. Un peu dans la même veine, la même équipe a d'abord créé à partir de CDR collectés en France un réseau d'interaction défini par une matrice origine-destination, où le domicile et le lieu de travail d'une personne sont définis comme étant à proximité de l'antenne relais où l'activité sur le réseau téléphonique est respectivement la plus importante et la deuxième plus importante (Panigutti *et al.,* 2017). Ils ont ensuite appliqué un modèle métapopulation de diffusion d'épidémie suivant ce réseau agrégé dans 329 unités spatiales. Ils ont défini une force d'interaction et un nombre de personnes susceptibles variable en fonction du lieu de travail et du lieu de domicile. À partir de cela, ils ont simulé 658 000 épidémies et ont noté que les principaux points communs dans les propagations de ces dernières étaient dus aux niveaux de connectivité, de trafic et de la taille de la population dans chaque unité spatiale.

Ces études, bien qu'associant modélisation en milieu épidémique et mobilité humaine à partir de nouvelles sources de données présentent pour nous deux inconvénients : tout d'abord leur échelle globale n'est pas appropriée aux études de propagation en milieu urbain, et leur modélisation, basée sur des modèles métapopulations présente beaucoup trop d'incertitudes, notamment vis-à-vis des paramètres et de l'échelle employée.

Wesolowski *et al.,* (2015b) ont travaillé sur les interactions entre les districts du Pakistan et la propagation de la dengue à partir de données environnementales et de 40 millions de CDR. Ils ont d'abord calibré un modèle SEIR à partir des cas reportés dans le sud du Pakistan (ou la dengue est endémique) et dans les villes où la dengue est saisonnière. Leur modèle, directement inspiré de Lourenço et Recker (2014) prend en compte l'évolution des comportements des moustiques lors de leurs différents stades en fonction de la température et de l'humidité. Les CDR servent à établir les probabilités d'interaction entre chaque district et donnent des résultats différents lorsqu'ils sont comparés à un modèle gravitaire classique. Ils ont estimé les dates d'importation de la dengue de Karachi vers les autres districts en récupérant le nombre de personnes infectées à l'instant t résultant du modèle à Karachi et en combinant avec la probabilité d'interaction entre Karachi et chacune des autres zones à partir des CDR. En comparant les résultats en fonction des données utilisées, les CDR entraînent une importation de la dengue plus précoce dans le nord du pays et plus tardive dans le sud. Leur étude souffre néanmoins d'un grand nombre de limites et de biais, inhérents au contexte denguien, notamment liés aux grandes incertitudes quant aux taux de transmission de la maladie, aux variables climatiques, aux nombres de cas reportés, au nombre important de personnes

asymptomatiques, de données sur le statut immunitaire de la population, etc. Ils ont cependant pu produire des cartes dynamiques du risque denguien, qui permettent de se préparer en cas de début d'épidémie.

Lima *et al.* (2015, 2013) ont travaillé sur des CDR en Côte d'Ivoire, d'après des données issues de la compagnie Orange et l'initiative *D4D* et ont fait ressortir les interactions entre différentes zones et ont appliqué un modèle métapopulation. Ils ont ensuite fait des tests sur différentes mesures de santé, comme la quarantaine (réduction des flux entre différentes zones) et la prévention, en partant du principe que l'information du danger se propage comme la maladie, mais augmente le nombre de personnes immunisées (pour la simplicité du modèle). Il ressort que la deuxième méthode permet de réduire la propagation de l'épidémie théorique, tandis que la première, même si elle réduit la propagation dans certaines zones n'entraîne pas de décalage temporel du pic épidémique dans le reste du pays.

À partir du même jeu de donnée Kafsi *et al.* (2013) ont dans un premier temps analysé les appels entre régions, et ont détecté 30 communautés distinctes, regroupées en *cluster* dans le pays. Ils ont ensuite créé un modèle de déplacement, en divisant une journée en 3 grandes tranches horaires et ont séparé les jours de semaine du week-end. Ils ont défini le lieu de domicile à l'antenne qui recueille le plus grand nombre d'enregistrements et posèrent que la probabilité de fréquenter un lieu à un moment donné dépend de l'heure de la journée, du moment de la semaine et de la localisation du domicile. Chaque personne a une probabilité de visiter un lieu selon une distribution multinomiale dépendant du nombre d'antennes d'où il a passé un appel. Ils appliquent ensuite un modèle métapopulation de type SIR et testèrent trois types de comportements ne sortez pas des frontières de votre communauté, restez dans votre cercle social et restez chez vous sur leur modèle. Ils montrèrent que certaines de ces recommandations, facilement communicable par SMS peuvent réduire drastiquement l'évolution d'une épidémie.

À partir de CDR, Frias-Martinez *et al.* (2011) ont construit un modèle de mobilité à base d'agents appliqué à la grippe H1N1 en prenant en compte à la fois les déplacements, ainsi que le réseau social. Ce dernier aspect est déterminé à partir du nombre d'appels entre deux personnes, notamment lorsqu'elles sont localisées autour d'une même antenne relais, en général le symptôme d'une future rencontre physique (Calabrese *et al.*, 2011b). Frias-Martinez *et al.* (2011) ont donc contraint leur modèle par la prise en compte d'une plus forte probabilité de contact entre une personne infectée et une personne saine si ces dernières se sont déjà appelées auparavant et se trouvent dans la même zone. Ils ont ensuite appliqué leur modèle aux cas de H1N1 enregistrés en 2009 dans une ville au Mexique et ont estimé que les mesures de contraintes de déplacements imposées par le gouvernement auraient réduit de 10 % le pic de personnes contaminées. Il s'agit, à notre connaissance de la seule étude de simulation à base d'agent des mobilités quotidiennes à partir de CDR n'ayant pas recourt à des modèles métapopulations de type SIR et appliqué en contexte épidémique.

Synthèse de l'état de l'art

Publication	Orienté sur les activités	Prise en compte du réseau social	Contexte épidémique	Échelle	Données mobilité	Données Activités
(Pappalardo and Simini, 2017)	Potentielle-ment	Non	Non	Urbaine	CDR + taxi	non
(Isaacman et al., 2012)	Domicile / travail	Non	Non	Urbaine	CDR + autres	non
(Jiang et al., 2016)	Domicile / autre	Non	Non	Urbaine	CDR	non
(Schneider et al., 2013)	Domicile / travail / autre	Non	Non	Urbaine	CDR	non
(Frias-Martinez et al., 2011)	Non	Oui	Oui	Urbaine	CDR	Non
(Noulas et al., 2012)	Oui	Non	Non	Globale / urbaine	Checkin Foursquare	Foursquare
(Wu et al., 2014)	Oui	Non	Non	Urbaine	Checkin Weibo	Weibo
(Silveira et al., 2016)	Non	Oui	Non	Urbaine / Nationale	CDR + Twitter	non
(Karl et al., 2014)	Domicile / travail / autre	Non	Oui	Urbaine	Census	Census
(Rinzivillo et al., 2014)	Oui	Non	Non	Urbaine	GPS voitures	Annotation volontaire
(Jiang et al., 2012)	Oui	Non	Non	Urbaine	Census	Census
(Kafsi et al., 2013)	Non	Oui	Oui	Nationale	CDR	hypothèses

Tableau 2 Synthèse des travaux sur les modèles de mobilité humaine à base d'agents. Ne figurent que les études qui prennent en compte au moins un critère parmi une approche orientée sur les activités, l'utilisation du réseau social et l'application en contexte épidémique.

Ces traces numériques géolocalisées (CDR ou données en lignes provenant de réseaux sociaux), relativement récentes, permettent de quantifier les mobilités et d'établir des lois et des modèles de déplacement. De nombreuses études ont mis en avant certains caractères des déplacements humains à différentes échelles, qu'il s'agisse de la distribution des distances entre chaque trajet, ou de différents paramètres de dispersions des individus. Ces informations servent de base dans l'établissement de modèles de déplacements.

Les données issues des réseaux sociaux sont de plus en plus utilisées, et elles présentent les avantages d'un accès relativement simple, par des *API*, tout en conférant une meilleure

précision spatiale (GPS du téléphone) que les *CDR*. Il faut cependant prendre en compte le taux de pénétration du réseau social dans la zone étudiée, ainsi que le niveau de représentativité des données. Néanmoins, ces données permettent d'analyser les pulsations urbaines et peuvent apporter d'énorme bénéfice dans la compréhension et la gestion des mobilités urbaines (Lenormand et Ramasco, 2016). Mais finalement, comme le montre le tableau 2, peu d'études ont développé des modèles à base d'agents à partir de traces numériques, où les déplacements sont motivés par la réalisation d'une activité, et en contexte épidémique.

Discussion

Les CDR et les données issues des réseaux sociaux offrent un large éventail de possibilité dans l'analyse, le traitement et la modélisation des mobilités urbaines. Elles permettent de combler et de compléter certaines lacunes des données mobilités plus classiques telles que les enquêtes de type ménage-déplacement, notamment lorsque ces dernières sont inexistantes dans la zone d'étude ou pas assez actualisée pour refléter les dynamiques actuelles.

Mais la production de ces données et leur utilisation entraînent des implications éthiques du fait du caractère personnelle des données et de l'absence de consentement explicite des personnes concernées. Pour protéger les utilisateurs, les opérateurs téléphoniques qui fournissent les CDR aux chercheurs passent d'abord par une phase d'anonymisation, dont l'objectif est de ne plus pouvoir relier un identifiant à un utilisateur[236]. Or, comme le souligne de Montjoye *et al.* (2014), il n'y a pas de technique parfaite pour anonymiser des jeux de données, notamment des CDR, sauf si elles sont agrégées ou grandement modifiées et dans ce cas elles perdent de l'intérêt.

Les données issues des réseaux sociaux présentent vis-à-vis des CDR l'avantage que l'on peut considérer l'utilisateur comme informé du caractère public de son message, notamment lorsqu'il utilise *Twitter*, où ce service à la vocation explicite d'être orienté sur un partage public des contenus. Certains *tweets* sont d'ailleurs repris tels quels par les médias, ou le pseudonyme de l'auteur est souvent cité. Ainsi, contrairement aux CDR, l'utilisateur est très probablement conscient que ces informations publiques vont être lues par d'autres personnes ce qui réduit les implications éthiques[237]. Mais comme pour les CDR, il ne se doute pas des traitements qu'il est possible d'effectuer à partir de ses données.

Ces données impliquent également une réécriture des rapports de puissance, que cela soit entre l'entreprise qui les possède et l'État qui est censé protéger ses citoyens, ou encore entre le chercheur et les sujets de l'étude (Taylor, 2016). Ces risques de ré-identification et l'application de traitements dans le dos de l'utilisateur sont cependant à mettre au regard des possibilités et de leurs perspectives d'utilisation dans l'intérêt général, et notamment en cas d'épidémies potentiellement mortelles (de Montjoye et al., 2014). C'est par exemple ce que vantait le projet « *D4D* » ou data for développement, ou même la fondation « Flowminder » qui parle de « data for good » (Decuyper, 2016). Turktelekom, un opérateur téléphonique Turque a

236. Changer le nom ne suffit pas. Un des exemples notoire d'échec d'anonymisation fut le cas du partage de la base de données d'AOL en 2006. Outre le fait que la requête la plus tapée sur leur moteur de recherche était « google », cette base partagée à des fins de recherches, voyait les utilisateurs identifiés par un numéro aléatoire, mais il fut très simple de ré identifier certaines personnes en regardant leur recherches.

237. Nous osons ici d'une métaphore, ou l'utilisation de CDR peut s'apparenter à une forme d'espionnage, tandis que l'utilisation des données de *Twitter* est plus une forme de voyeurismes vis à vis d'exhibitionnistes.

lancé en décembre 2017 le projet "Data For Refugees" [238]. Si le projet se veut philanthrope, il va sans dire que l'application de méthodes déjà testées dans *D4D*, comme la détection de communautés et l'analyse des déplacements prend une toute autre tournure éthique, du fait d'une possible récupération par un régime qualifiable d'autoritaire et de la relation ambiguë avec l'Union Européenne en ce qui concerne « la crise des migrants ». De plus, ces données pourraient aussi plus servir à contrôler, traquer ou expulser des personnes plutôt que de les protéger [239].

Au-delà des aspects primordiaux du respect des personnes, du droit à l'anonymat, de ne pas être tracé, et de garder un droit de regard sur les données personnelles, d'autres critiques peuvent être émises sur ce genre de données. Ainsi leur volume colossal et les nécessités de connaissances techniques et informatique pour pouvoir les traiter font que ces données sont plus facilement analysées par des informaticiens et des physiciens que par des géographes et des sociologues (Louail, 2016), ce qui entraîne l'apparition d'un fossé entre ceux qui analysent les données et ceux qui peuvent les comprendre (Taylor, 2016). Par exemple, sur les 150 équipes de recherches qui ont travaillé avec les données de *D4D* en Côte d'Ivoire, seule une est allée sur le terrain. Il paraît donc pertinent de croiser des approches pour éviter le « tout quantitatif », qui bien que faisant ressortir énormément de choses peut laisser passer des notions plus « humaines » des mobilités, appréciable par des enquêtes qualitatives de terrain. Il faut également être conscient du niveau de représentativité de la donnée utilisée, et ne pas hésiter à croiser différentes sources, allant d'enquêtes de terrains, à des données sur l'utilisation du temps ou des enquêtes ménage-déplacement, ou même différents réseaux sociaux.

238. http://d4r.turkte ekom.com.tr/presentat on/object ves
239. http://www.w red.co.uk/art c e/europe mm grat on refugees smartphone metadata deportat ons

Chapitre VI: Traces numériques à Delhi et Bangkok

Une quantité sans précédent de traces numériques est produite quotidiennement (chapitre 4), et au-delà des aspects éthiques de consentement et de vie privée, ces dernières peuvent être très utiles dans l'analyse et la modélisation des mobilités, notamment quotidienne (chapitre 5). Aussi, toutes ces données n'offrent pas le même potentiel en termes de représentativité, de précision spatiale et temporelle, et d'accessibilité. Si les statistiques d'appels sont relativement continues dans le temps, leur résolution spatiale dépend du maillage des antennes relais et leur accès est conditionné à juste titre par des accords avec les fournisseurs d'accès. Si la collecte de données issues des réseaux sociaux, qu'il s'agisse de *Twitter* ou *Facebook*, est nettement plus aisée, leur potentiel varie selon leur nature (donnée longitudinale ou agrégée) et leur volume, ce qui implique des avantages et des limites, que nous détaillerons dans le présent chapitre.

Nous allons tout d'abord présenter une étude comparative des données issues de *Twitter* à Delhi et à Bangkok, en nous focalisant sur les prétraitements et l'estimation du niveau de représentativité, afin d'évaluer leur potentiel d'utilisation. Nous présenterons ensuite les données de *check-in* collectées sur plusieurs semaines à Bangkok depuis *Facebook Places*, en insistant sur les différents filtres nécessaires pour limiter leur biais.

1 Données *Twitter* à Delhi et Bangkok

1.1 Présentation du service

Twitter est un réseau social fondé en 2006 qui compte actuellement 328 millions d'utilisateurs mensuels[240], ce qui en fait l'un des médias sociaux les plus populaires mais qui n'a jamais fait de bénéfices depuis son lancement[241]. Il permet à ces membres d'envoyer des messages, ou « *tweets* » à destination de personnes abonnées, ou 'follower'. Ces messages contenaient au maximum 140 caractères, reprenant ainsi le format SMS de 160 caractères dont vingt sont réservés au nom de l'utilisateur[242]. En novembre 2017, le nombre maximum de caractères est doublé, passant à 280[243]. Ces messages sont publics par défaut, c'est-à-dire visible par tous[244] et intégrés dans le moteur de recherche de *Twitter*[245]. Cet aspect rappelle le principe des flux RSS, mis en place en 1999 et qui permet d'accéder aux contenus produits en temps réel par des pages web[246]. Cela dit, les messages peuvent aussi être « protégés », c'est-à-dire visible uniquement par les personnes abonnées au compte et dont la demande a été validée par l'utilisateur[247]. Un utilisateur peut également « retweeter » un message, c'est-à-dire faire suivre un *tweet* envoyé par une autre personne à l'ensemble de son réseau. Un *tweet* public peut également être adressé à une ou plusieurs personnes, par l'intermédiaire de l'ajout d'un « @ » avant le pseudonyme de l'utilisateur ciblé par le message, créant ainsi une sorte de courte lettre de 280 caractères ouverte destinée @USER. L'ajout de « # » avant un mot-clé permet de créer une page temporaire, qui regroupera tous les messages contenant le même mot-clé[248]. Par exemple tous les messages comportant #dengue seront regroupés sur la page https://twitter.com/hashtag/dengue.

Depuis 2015, *Twitter* autorise l'envoi de messages directs, c'est-à-dire qu'une personne peut envoyer des messages visibles uniquement par une partie de ses abonnés. Chaque personne de ce groupe peut ensuite autoriser certains de ces abonnés à participer à une discussion, la taille

240. Au premier trimestre 2017, dans la *Letter to Shareholders :* http://fi es.shareho der.com/down oads/AMDA 2F526X/4838254700x0x939175/D7BAFE57 DCBD 42E9 9909 7F587047FCED/Q117 Shareho der Letter.pdf.

241. Elle perd entre 1 et 2 millions de dollars par jour depuis quelques années et ne survit que par des levées de capitaux https:// nvestor.tw tter nc.com/resu ts.cfm& https://www.crunchbase.com/organization/twitter.

242. *Qu'est ce que Twitter ?* https://support.tw tter.com/art c es/247973

243. https ://techcrunch.com/2017/11/07/twitter officially expands its character count to 280 starting today/

244. "What you say on the *Twitter* Services may be viewed all around the world instantly. You are what you Tweet !" *Twitter* terms of use https://tw tter.com/tos? ang=en#bas cterms

245. Et depuis aout 2015 directement dans *Google* http://www. emonde.fr/p xe s/art c e/2015/ 08/22/tw tter conforte sa presence dans e moteur de recherche de goog e 4733192 4408996.htm & http://fi es.shareho der.com/down oads/AMDA 2F526X/2695012492x0x874448/7F88ED74 4727 4B42 8568 60B0C0DA92C7/Q4 15 Shareho der Letter.pdf.

246. Il suffit d'ailleurs de remplacer « USER » par un nom d'utilisateur de *Twitter* dans le lien suivant pour accéder à ces *tweets* sur un agrégateur de flux RSS https://tw trss.me/tw tter user to rss/?user=USER.

247. *À propos des Tweets publics et des Tweets protégés :* https://support.tw tter.com/art c es/115718}.

248. https://support.tw tter.com/art c es/49309#

maximale du groupe étant de 50 personnes [249]. Pour envoyer un message privé, il faut cependant que la personne suive le compte, sauf lorsque la personne a activé une option permettant de recevoir des messages privés de tout le monde.

Mis à part les « @ » et « # », *Twitter* n'apporte rien de très nouveau d'un point de vue conceptuel, puisqu'il s'agit en quelque sorte d'une plate-forme d'envoi de SMS qui suit une philosophie calquée sur celle des flux RSS, auquel s'est ajouté un système de messagerie instantanée rudimentaire. Cependant, *Twitter* doit faire face à d'énormes challenges techniques dans la gestion et l'acheminement d'une quantité phénoménale d'information, puisqu'entre 300 et 500 millions de messages sont envoyés quotidiennement [250], produisant environ 7 Téra Octets de données [251].

1.1.1 Utilisation du réseau social dans le monde

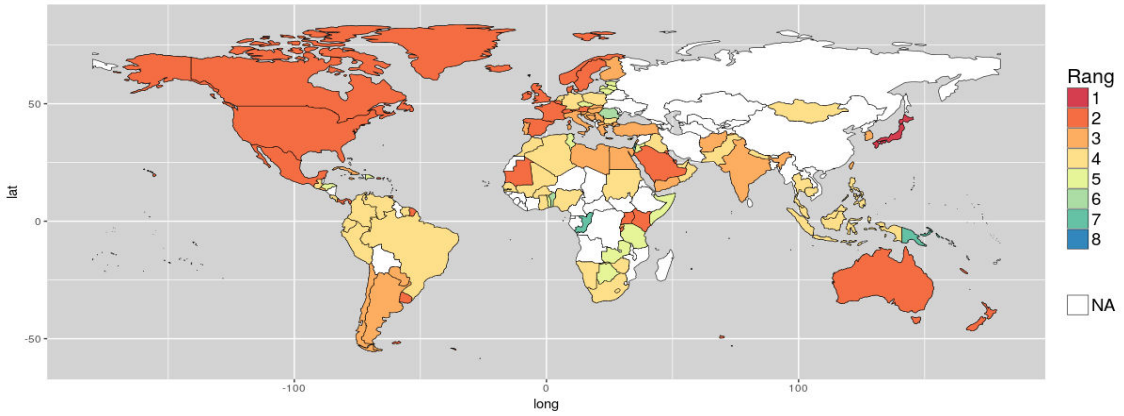


FIGURE 58 Rang de *Twitter* parmi les autres réseaux sociaux par pays. D'après un classement issu d'Alexa.com en juillet 2017. Source : http://www.alexa.com/topsites/

Si pour des raisons de transparence liées à sa présence en bourse [252], *Twitter* communique facilement sur le nombre d'utilisateurs aux États-Unis et dans le monde, l'entreprise ne donne pas plus d'information sur l'utilisation du service par pays. Afin de donner un ordre de grandeur de l'utilisation de *Twitter* dans le monde, nous avons utilisé les données du service *alexa* d'*Amazon*

249. *À propos des messages privés :* https://support.tw tter.com/art c es/223526.

250. 500 millions de *tweets* étaient envoyés quotidiennement en novembre 2013 alors que le nombre d'utilisateurs actifs était de 215 millions. Depuis la compagnie ne communique plus sur ce genre d'informations. Certains estiment aujourd'hui qu'environ 300 millions de messages sont envoyés quotidiennement sur la plateforme http://www.businessinsider.fr/uk/*tweets* on twitter is in serious decline 2016 2/.

251. Ce qui, comme le souligne Jean Gabriel Ganascia, professeur à l'UPMC, correspondrait environ à la moitié du volume des ouvrages de la Bibliothèque François Mitterrand si ces deniers étaient codés en langage binaire – https://www.francecu ture.fr/em ss ons/ a methode sc ent fique/ e numer que fa t de nous des numeros.

252. http://www.nasdaq.com/fr/symbo /twtr

et avons calculé son rang parmi les réseaux sociaux (figure 58).

Le site figure parmi les réseaux sociaux les plus populaires, notamment en Amérique du Nord et en Europe de l'Ouest et du Nord, ainsi qu'en Australie où il est second après *Facebook*. *Twitter* serait numéro un au Japon, et dans le top 4 en Asie du Sud-Est. À noter l'absence de *Twitter* dans le top 50 des sites les plus visités en Chine et en Russie, où il est détrôné par d'autres plateformes similaires qui s'y sont largement développées en parallèle d'une volonté de souveraineté de ces États comme *Sina Weibo*, *QQ* et *Wechat* en Chine et *VK* en Russie.

Il est également important de nuancer les données d'alexa.com car beaucoup de personnes accèdent aux réseaux sociaux directement via des applications sur leur téléphone portable et ne sont donc pas prises en compte par le service. Par exemple, alors que *Twitter* compte moins d'utilisateurs qu'*Instagram* (environ 328 millions contre plus 700 millions en 2018), alexa.com classait en juillet 2017 *Twitter* au 11e rang mondial en termes de trafic, contre 18 pour *Instagram* (et 3e pour *Facebook*).

1.1.2 Twitter et la géolocalisation

À partir de 2009 l'entreprise a ajouté une option de géolocalisation des messages (désactivée par défaut) qui permet à un utilisateur de relier sa localisation à son tweet, ajoutant ainsi un contexte géographique. Il existe deux types de géolocalisations : Une géolocalisation générale, et une géolocalisation précise.

- La géolocalisation générale consiste à ajouter une adresse connue, fournie par l'entreprise *Foursquare* et située en général à proximité de la localisation réelle [253]. Il s'agira par exemple d'une ville, d'un quartier, d'un monument ou d'un commerce. Mais un utilisateur peut localiser son tweet dans n'importe quel lieu répertorié par *Foursquare*.

- La géolocalisation précise récupère en revanche la localisation quasi-exacte de l'appareil mobile [254], si ce téléphone portable est doté d'un système d'exploitation de type Android ou iOs [255].

Afin de pouvoir utiliser les informations de localisation de l'appareil, l'utilisateur doit au préalable autoriser l'application de *Twitter* à accéder aux données de géolocalisation du téléphone, c'est-à-dire les données du GPS du téléphone, ainsi que les informations sur les signaux des BTS captés, et éventuellement une liste de réseaux Wifi publics détectés à

253. https://support.tw tter.com/art c es/231371 "*Twitter* peut utiliser divers signaux pour déterminer la localisation précise de votre appareil, y compris le GPS, le signal d'une antenne relais et les données des points d'accès sans fil voisins".
254. Utiliser le service de localisation sur les appareils mobiles : https://support.tw tter.com/art c es/20170767.
255. Comme ce fut le cas pour 98.9 % des smartphones vendus au 1er trimestre 2016. http://www.gartner.com/newsroom/ d/3323017

proximité[256]. La précision de la localisation dépend ensuite du GPS de l'appareil et des conditions (si l'appareil se trouve à l'intérieur ou à l'extérieur, la précision varie entre 30 et 100 m (Zandbergen et Barbeau, 2011) et du nombre d'autres antennes captées. *Twitter* réalise ensuite une trilatération, à partir des informations GPS, des données BTS et Wifi en prenant en compte l'atténuation du signal comme indicateur de distance à une antenne relais ou au point d'accès wifi le plus proche. Avant mai 2015, lorsque la géolocalisation des tweets était activée pour un message, elle l'était également pour tous les messages suivants, jusqu'à ce que l'utilisateur désactive l'option[257]. Cela dit, une faible part des *tweets* est géolocalisée entre 1.5 et 3 % en 2011 (Murdock, 2011).

1.1.3 Décorticage d'un Tweet

La structure complète d'un Tweet contient bien plus de 280 caractères (tableaux 3, 4, 5, 6). Les informations relatives au message et transmises en même temps que ce dernier (tableaux 3 et 4), bien que très nombreuses, ne paraissent pas superflues d'un point de vue technique[258]. En revanche, nous ne comprenons pas vraiment pourquoi autant d'informations sur l'utilisateur sont transmises avec le message (tableaux 5 et 6), alors que l'identifiant et le nom pourraient suffire car toutes ces informations sont accessibles sur la page du profil de l'utilisateur[259] et que ces données redondantes ont un coût de stockage[260].

D'après expérience, lorsqu'un utilisateur active sa localisation précise mais géolocalise son message selon un lieu dans la liste fournie par Foursquare, la géolocalisation du téléphone est enregistrée dans *geo_point_coordinates*, mais les autres informations géographiques (*bounding_box, place_id country_code, etc.*) font référence à la localisation sélectionnée par l'utilisateur. Il est également possible d'envoyer des messages en forçant la géolocalisation précise[261]. Dans ce cas, « geo_point_coordinates » renvoie à la géolocalisation usurpée, mais « source » ne devrait pas faire référence à un système d'exploitation classique[262]. Nous y

256. « If you have location services enabled, *Twitter* may use a variety of signals to determine the precise location of your device, such as GPS, cell tower signal and data about nearby wireless access points. Whether or not you have location services enabled, if you own a wireless access point (for example, if you have set up a wireless network), *Twitter* may use certain publicly broadcast information from that access point, such as its name/SSID, MAC address, frequency, and signal strength, to power *Twitter*'s location services » https://support.tw tter.com/art c es/118492#

257. https://support.tw tter.com/art c es/231371

258. L'information sur le système d'exploitation permet d'optimiser l'affichage, les données sur le nombre de retweet permettent de voir la popularité du message, etc.

259. https://twitter.com/USER pour l'adresse du profil de l'utilisateur USER. Adresse à laquelle on ajoute /following pour avoir la liste des utilisateurs que la personne suit, /followers pour la liste des personnes qui la suivent, /likes pour les messages qu'elle a "aimés" et /medias pour les photos ou vidéos qu'elle a pu envoyer.

260. Il est tentant de poser l'hypothèse d'une erreur de conception, mais il doit y avoir une explication plus rationnelle, comme l'optimisation des tris et des filtres pour accéder au message ou pour la performance de l'affichage.

261. En passant par une *API* et notamment via des scripts python et la librairie *tweepy*. Par exemple : api.update status(status=tweet,lat=13.8,lon=100.4), ou tweet est le message.

262. Mais au nom de l'application créée par la personne et enregistrée sur *Twitter* pour pouvoir accéder

reviendrons plus tard dans ce chapitre.

Balise html	Signification
id	Identifiant du message
created at	Date de création du message
text	Contenu du message (280 caractères)
in reply to status id	En réponse à un statut ou non
in reply to user id	En réponse à un utilisateur via son identifiant ou non
in reply to screen name	En réponse à un utilisateur via son nom qui apparaît ou non
source	Type d'appareil utilisé lors de l'envoi du message - Android, Iphone, etc.
retweeted	Si le message a été retweeté
retweet count	Le nombre de fois que ce message a été retweeté
favorited	Si le message a été mis en favoris
favorite count	Le nombre de fois que ce message a été mis en favoris
hashtags	Si le message contient un #
url	L'url éventuelle raccourcie contenu dans un message
expanded url	L'url complète

Tableau 3 Informations contenues dans un *tweet* concernant le message

à l'*API*.

Balise html	Signification
geo enabled	Si la localisation a été activée
Geo Point coordinates	Les coordonnées du point où le message a été géolocalisé. Uniquement lorsque la géolocalisation précise a été activée
bounding box	Les coordonnées d'un polygone de taille variable dans lequel fut envoyé le message géolocalisé. Apparaît dès que la géolocalisation (précise ou non) est activé.
Place id	L'identifiant du lieu renseigné
place type	Le type de lieu (ville, pays, etc.)
name	Le nom de ce lieu
full name	Le nom complet de ce lieu
url	Lien vers tous les tweets étant envoyés de ce lieu
country code	Le code du pays, par exemple TH pour Thaïlande, FR pour France
country	Le nom complet du pays

Tableau 4 Informations contenues dans un *tweet* concernant la géolocalisation du message

Balise html	Signification
user id	Le numéro Twitter identifiant l'utilisateur
name	Le nom renseigné par l'utilisateur
location	La ville de domiciliation renseignée
screen name	Le nom apparaissant à l'écran
url	Un lien éventuel vers un autre site blog, page Facebook, etc.
description	Une rapide description faite par l'utilisateur
protected	Si le compte est protégé
verified	Si la personne a fait vérifier son compte
followers count	Le nombre de personnes qui suivent ce compte
friends count	Le nombre de comptes suivit par cette personne
favourites count	Le nombre de messages mis en favoris
statuses count	Le nombre de mises à jour de statut
created at	La date de création du compte
utc offset	Le décalage horaire (par rapport à l'UTC)
time zone	Le nom de la zone pour l'heure
lang	La langue parlée par défaut

Tableau 5 Informations contenues dans un *tweet* concernant l'utilisateur

Balise html	Signification
profile background color	La couleur de fond
profile use background image	Si l'utilisateur utilise une image de fond
profile background image url	Lien vers l'image de fond
profile image url	Lien vers l'image de profil
profile banner url	Lien vers la bannière de profil
default profile	S'agit-il d'un profil personnalisé ?
default profile image	S'agit-il de l'image de profil par défaut ?

Tableau 6 Informations contenues dans un *tweet* concernant la page de l'utilisateur

1.1.4 Récupération des données

À partir du moment où des données sont publiques, elles sont par essence potentiellement accessibles à tous. Plutôt que de voir ces données récupérées par des tierces-personnes inconnues, l'entreprise a fait le choix de mettre à disposition des *API* qui permettent aux entités extérieures d'obtenir ces données de manière plus aisée, tout en pouvant contrôler leurs usages et en valorisant leur monétisation. Les bénéfices pour l'entreprise sont considérables en termes d'image et de visibilité, notamment lors des campagnes présidentielles, où l'ouverture de la base permet aux journalistes de relayer des commentaires postés sur le réseau ou d'analyser les « tendances » au travers des #, tout en citant le service. *Twitter* permet l'accès aux données publiques de deux manières :

- Avec l'*API Rest*[263] pour les données historiques, stockées sur le site, telles que les messages anciens ou les profils d'utilisateurs.

- l'*API Stream*[264] pour les flux de message en temps réel.

Dans les deux cas, la personne désirant accéder à ces données doit d'abord créer un compte, puis créer une sorte de projet dans la partie dédiée aux développeurs[265]. Elle obtiendra ensuite des identifiants et des mots de passe permettant d'accéder à ces services. Les différentes *API* permettent de rechercher des identifiants, des profils d'utilisateurs, des mots clés et ce qui nous intéresse ici, des *tweets* émis depuis certaines zones géographiques. Utilisées gratuitement, ces *API* présentent des limites de nature différentes. Pour l'accès via Rest, les limites sont en termes du nombre de requête par tranche de 15 minutes. Par exemple il n'est possible d'avoir accès qu'a 900 pages (ou *timelines*) par quart d'heure[266]. Pour ce qui est de l'accès au flux de messages en temps réel et gratuitement, les limites se présentent comme un pourcentage de la masse globale de messages envoyés au même moment, autour de 1 %. Il s'agit de l'échantillon

263. https://dev.tw tter.com/rest/pub c
264. https://dev.tw tter.com/stream ng/overv ew
265. https://dev.tw tter.com/
266. https://dev.tw tter.com/rest/pub c/rate m ts

« Spritzer » (Morstatter et al., 2013 ; Wang *et al.,* 2015). Mais lorsque l'on effectue une requête, dont les critères de recherches (des noms, des #, des zones géographiques) renvoient moins de 1 % de l'activité, dans ce cas l'ensemble des messages est récupéré[267]. Ce dernier point semble être confirmé par une petite analyse réalisée à Bangkok où nous avons configuré une *API* pour envoyer des messages automatiquement dans la zone d'étude et ces derniers ont tous été récupérés via l'*API Stream*. Il faut cependant préciser que le nombre de messages envoyés par nos soins est relativement faible (900) et qu'il est possible que les messages envoyés via des *API* soient tous récupérés par défaut par l'*API Stream*. Il est possible d'augmenter la taille de l'échantillon collecté en temps réel en payant des sommes assez importantes pour accéder à l'échantillon « Gardenhouse », fourni par Gnip (une filiale de *Twitter*), qui permet de récupérer jusqu'à 10 % des flux totaux en temps réel (Wang *et al.,* 2015).

Ainsi, l'utilisation de ces *API* (*Rest* & *Stream*) peut s'avérer similaire au concept de « freemium », dans le sens où un échantillon est accessible gratuitement, et qu'il est possible de payer pour accéder à des fonctionnalités (ici des volumes de données) plus exhaustives. Ainsi, le flux *Gardenhouse* est susceptible d'intéresser des médias, ou des entreprises de communication cherchant à cibler une clientèle pour promouvoir des produits. Comme montré précédemment, ce ne fut pas nécessaire dans notre cas, car notre recherche est basée sur des *tweets* géolocalisés (seulement entre 1 et 3 % des Tweets) dans des zones géographiques réduites, ce qui induit un volume suffisamment faible de messages pour ne pas dépasser le seuil de 1 % du flux total.

1.1.5 Twitter et les bots

Mais *Twitter* ne fournit pas que des *API* permettant de récupérer des messages, elle offre également des outils permettant d'en envoyer. Gérés par un opérateur ou par des algorithmes dont la finalité n'est pas toujours bien définie en général de marketing, de promotions ou d'influence d'opinion ces comptes ne sont plus associés à des individus physiques, mais principalement à des entreprises ou des personnalités célèbres et peuvent être comparés à des robots, ou « bots ». *Twitter* a annoncé en 2014 que près de 8.5% de ses utilisateurs actifs utilisaient des services tiers[268], une part d'entre eux étant donc potentiellement des bots (Ferrara *et al.,* 2014). Ce pourcentage reste stable en décembre 2015[269], et représenterait donc entre 26 et 32 millions de comptes en 2017.

Parmi ces bots, certains utilisent la géolocalisation, car la possibilité d'envoyer des messages contenant des informations géographiques qui ne sont pas la position actuelle du

267. Comfirmé par certains employés de *Twitter* et de Gnip (filiale de *Twitter*) sur des sites dédiés à la communauté des programmeurs https://tw ttercommun ty.com/t/d ffence between samp e and fi ter stream ng ap /15094 : « the keywords you are tracking account for less than 1% of the firehose, you will receive all the matching Tweets, otherwise you will be capped »)

268. http://qz.com/248063/tw tter adm ts that as many as 23 m on of ts act ve users are actua y bots/

269. https://www.sec.gov/Arch ves/edgar/data/1418091/000156459016021918/twtr 10q 20160630.htm

propriétaire du compte *Twitter* revêt des avantages certains dans différents domaines. Cela permet par exemple de signaler un événement comme un accident sur la rocade de Bangkok [270] ou l'arrivée d'un dictateur à l'aéroport de Genève [271]. L'objectif peut également être juste promotionnel, tel un *show case* d'artistes quelconques de *Kai-Pop* [272], des soldes dans un magasin de cosmétique de Bangkok [273], un appartement à louer dans une *gated-community* de Delhi ou de Mumbai [274], ou encore des offres d'emplois dans le secteur des services aux Philippines [275]. Il existe des bots exploitant la géolocalisation dont la finalité demeure plus mystérieuse tel *@googuns_lulz* qui envoie des séries hexadécimales de 140 caractères toutes les minutes à des positions géographiques qui paraissent aléatoires [276].

Certains programmeurs ont également conçu des algorithmes assez sophistiqués pour envoyer et répondre à des messages de manière quasi autonome, définis comme étant des « socials bots » (robots sociaux) (Ferrara *et al.*, 2014). Ils peuvent ainsi interagir et envoyer des messages ciblés à des personnes qui les suivent, suivant des desseins généralement commerciaux, promotionnels ou de vitrine en termes d'intelligence artificielle [277]. De nombreuses études se concentrent également sur le fonctionnement et les impacts de ces robots sociaux, notamment lors de l'élection américaine de 2016 (Shao *et al.*, 2017). Nous avons nous aussi configuré une application capable d'envoyer des messages à « localisation précise » artificielle, afin de vérifier le bon fonctionnement de nos programmes et pour estimer la part de l'échantillon enregistré via l'*API Stream* (voir plus haut). Tous ces tweets envoyés furent récupérés, ce qui peut donc fausser l'échantillon, car ces derniers sont difficilement différenciables de messages émis par des utilisateurs "réels" si l'on ne récupère uniquement les coordonnés, l'heure et l'identifiant du message. Un des enjeux est donc de distinguer les utilisateurs *lambda* des autres comptes professionnels, commerciaux, ou bots à peu près autonomes.

270. https://tw tter.com/ ongdotraffic à Bangkok, qui a envoyé environ 300 000 *tweets* en 7 ans.
271. https://tw tter.com/gva watcher
272. https://tw tter.com/Koreaboo
273. Par exemple https://tw tter.com/EVEANDBOY
274. Avec par exemple en inde le compte https://tw tter.com/propert es nd a qui a envoyé plus de 1.3 millions de *tweets* depuis sa création en novembre 2009
275. https://tw tter.com/tmj ph jobs
276. Une partie de l'explication se trouverait ici : http://v ctorz.ca/bots/googuns.
277. On se rappellera de l'expérience de l'entreprise Microsoft qui avait lancé un bot dédié à la conversation, nommé "Tay". Il était censé interagir et apprendre de ces interactions avec les internautes. 8 heures et 96 000 tweets plus tard, le bot avait tellement appris de ces interlocuteurs qu'il réfutait l'holocauste et estimait qu'Hitler ferait un meilleur travail que les hommes politiques d'aujourd'hui. http://www. emonde.fr/p xe s/art c e/2016/03/24/a pe ne ancee une nte gence art fic e e de m crosoft derape sur tw tter 4889661 4408996.htm et http://rue89.nouve obs.com/2016/03/ 25/am s es bots 263573. Deux autres chatbots, développé eux aussi par Microsoft sur le réseau social chinois QQ ont été interdits car ils avaient fini par tenir des propos non favorables au régime en place http://www. epar s en.fr/h gh tech/ a ch ne deconnecte deux chatbots ant commun stes 04 08 2017 7175022.php#xtor=RSS 1481423633.

1.2 Données brutes à Delhi et Bangkok

1.2.1 Collecte des données

Twitter met donc à disposition des « développeurs » des *API* qui permettent de récupérer les *tweets* publics envoyés sur sa plateforme, notamment STREAM, qui donne accès aux messages en temps réels tant que le résultat de la requête ne dépasse un certain seuil de l'échantillon global. Un *package* permettant d'accéder à ces *API* pour récupérer les *tweets* est disponible sous R (TwittR[278]). Nos compétences rudimentaires dans ce langage au moment de nos essais (juin 2014), non compensées par des discussions sur le sujet sur des forums Internet spécialisés (*e.g. stackoverflow*) ont fait que nos tentatives sont restées infructueuses. Malgré aucune notion, nous avons testé Python, et la librairie Tweepy[279] qui permet d'accéder de manière similaire à l'*API*, avec l'avantage d'être à l'époque bien documentée[280] et agrémentée de tutoriels en ligne[281]. Plusieurs heures de programmation, notamment en reprenant et adaptant des bouts de codes trouvés en ligne, ont permis d'aboutir à un petit programme fonctionnel, d'une centaine de lignes, capable d'accéder et d'enregistrer les flux de *tweets* publics et géolocalisés à Bangkok et dans l'ensemble de l'Inde. Nous avons par la suite étendu la zone d'étude de Bangkok à l'ensemble de la Thaïlande, dans l'optique d'étudier les mobilités inter-urbaines non présenté dans ce présent travail.

Notre programme fut lancé sur un serveur distant[282] ce qui permet d'accéder aux *tweets* de manière continue via l'*API Stream* tout en profitant d'une grande capacité de stockage. À cette période, nous ne récupérions que les identifiants des utilisateurs, la géolocalisation du message, ainsi que la date d'envoi. D'autres informations, comme la plateforme d'envoi du message, dont les contenus ne furent récupérés que plus tard, à partir de novembre 2016 et ne figureront pas dans ce travail exploratoire. Conformément aux recommandations de la CNIL vis-à-vis des données personnelles, les données brutes, qui comportent le pseudo de l'utilisateur sont des données personnelles et doivent être sécurisées. Elles sont stockées en deux endroits : sur une partition chiffrée[283] d'un serveur hébergé par Huma-num[284], et sur une partition chiffrée[285] sur un disque dur externe pour permettre des traitements en local. La description des traitements soumis à la CNIL est disponible en annexe D.

1.2.2 Descriptions des données

Volumes

278. https ://cran.r project.org/web/packages/twitteR/twitteR.pdf
279. http://www.tweepy.org/
280. Avec quelques exemples : https://g thub.com/tweepy/tweepy/tree/master/examp es
281. Par exemple https://www.youtube.com/watch?v=pUUxmvv 2FE
282. Tout d'abord un serveur OVH mis à disposition par le Centre de Science Humaines de Delhi, puis par le service Huma num.
283. en utilisant ecryptfs https://doc.ubuntu fr.org/ecryptfs
284. http://www.huma num.fr/.
285. En utilisant veracrypt https://doc.ubuntu fr.org/veracrypt

La période d'enregistrement des *tweets* s'étend du 26 juin 2014 au 4 décembre 2015[286], soit 528 jours, et concerne l'ensemble du sous-continent indien et la région de Bangkok[287]. Nous avons ensuite élargi la zone d'enregistrement de la capitale thaï à l'ensemble de la péninsule thaïlandaise à partir du 5 janvier 2015. Plus de 35,8 millions de tweets ont été collectés en Thaïlande contre 13,6 en Inde, malgré un nombre équivalent d'utilisateurs respectivement 543 857 et 507 482 ce qui suggère que les usagers de *Twitter* utilisant la géolocalisation sont plus actifs et nettement plus nombreux en Thaïlande qu'en Inde proportionnellement à la population et à la taille du territoire.

Nous n'avons ensuite conservé que les *tweets* des utilisateurs ayant envoyé au moins un message depuis Delhi[288] ou Bangkok. Nous avons ainsi enregistré 29 499 225 tweets en Thaïlande émanant de 307 428 utilisateurs, et 70 % de ces messages étaient envoyés depuis la capitale Thaï (20 711 488). De même, 63 496 utilisateurs à Delhi ont émis 3 167 005 tweets en Inde dont 1 617 530 (51 %) provenait de la capitale indienne. Ces volumes bien plus importants à Bangkok qu'à Delhi sont visibles sur la figure 59 qui montre le nombre de *tweets* publics et précisément géolocalisés enregistrés quotidiennement dans les deux mégapoles. La ligne bleue verticale sur les graphiques pour Bangkok correspond à une extension de la zone d'étude à l'ensemble de la Thaïlande (5 janvier 2015). La présence de "trous" dans la collecte est liée à des problèmes techniques entre les serveurs, notamment des désynchronisations entraînant des pertes de connexions. Ainsi, nous n'avons des informations que pour 498 jours.

Nous pouvons aussi observer une chute drastique du nombre de *tweets* récoltés à partir du 28 avril 2015. Après quelques recherches, il s'avère que cette baisse est très probablement liée à une mise à jour de l'application sur les systèmes d'exploitation Android et iOs[289]. Auparavant, lorsqu'un utilisateur activait la géolocalisation précise pour un message, cette dernière option restait activée et donc appliquée aux messages suivants, ce qui n'est plus le cas à partir de cette date pour les utilisateurs ayant des versions à jour de l'application sur leurs téléphones portables. Alors que cette mise à jour entraîne une baisse d'environ 66 % des messages enregistrés à Bangkok et 61 % à Delhi, le nombre d'utilisateurs quotidiens n'est pas réduit de la même manière puisqu'il décroît de 32 % à Bangkok contre 57 % à Delhi (figure 59). À noter que le nombre de messages envoyés quotidiennement par utilisateur est divisé par deux à Bangkok, mais reste stable à Delhi, ce qui pourrait signifier que les utilisateurs les plus actifs sortirent des radars dans la première zone d'étude, mais pas dans la seconde.

286. Date à laquelle nous avons compilé et traité les données après un énième crash informatique.
287. Longitude comprise entre 100.1° et 100.9° et latitude entre 13.5 et 14.2°.
288. Longitude comprise entre 76.6° et 77.7° et latitude entre 28.28 et 28.9°.
289. https ://support.twitter.com/articles/231371# la version 5.56.0 pour Andro d est sort e e 28 avr 2015 (https://www.apk4fun.com/h story/2699/), et a vers on 6.26 pour Os e 14 avr 2015 (https: //a mychanges.com/p/ os/tw tter/)
.

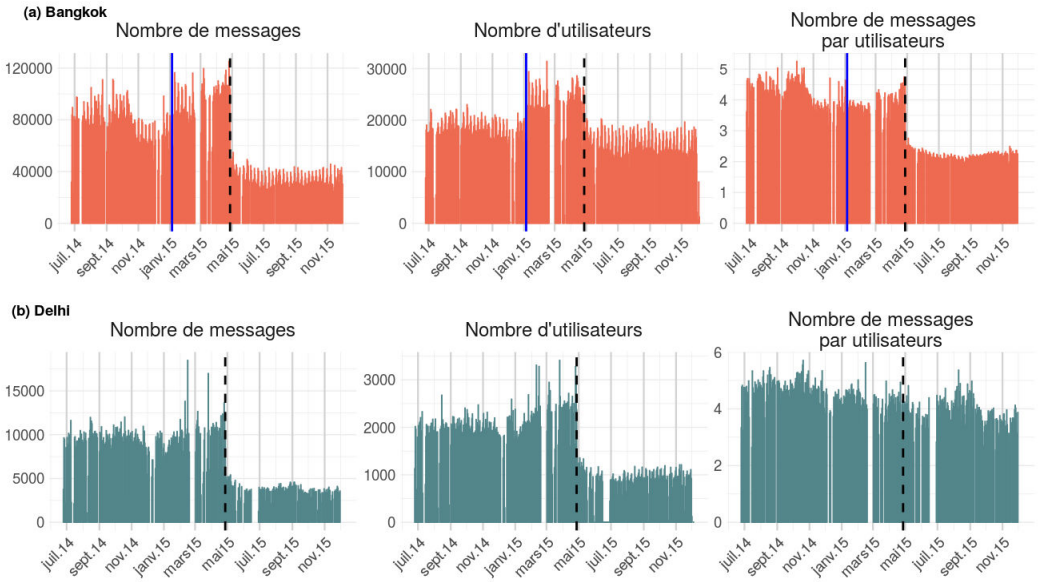


FIGURE 59 Comparaison des volumes de données *Twitter* collectés à Bangkok et à Delhi. Avec respectivement le nombre de messages enregistrés quotidiennement (gauche), le nombre d'utilisateurs associés (centre) et le nombre de messages envoyés par utilisateurs (droite). La ligne bleue correspond à la date de l'extension de la zone d'étude à l'ensemble de la Thaïlande pour le jeu de données sur Bangkok, tandis que la ligne en pointillé correspond à une chute globale du nombre de *tweets* collectés.

Activité sur le réseau

La figure 60.a ci-dessous présente la part d'utilisateurs ayant *tweeté* au moins une fois à Bangkok et Delhi et le nombre de messages géolocalisés qu'ils ont envoyés sur la période. La figure 60.b montre la même information, mais en cumulant ces pourcentages. Dans les deux zones, la distribution de l'activité des utilisateurs suit une loi puissance à longue queue, ce qui signifie que la plus grande partie des comptes envoi peu de messages et qu'un faible nombre en envoi énormément, avec de grandes amplitudes au sein de ce dernier groupe (longue queue, non bornée). Dans un graphique en log/log, cette distribution est relativement linéaire sur un certain intervalle (jusqu'à environ 500 messages), puis se disperse après une zone d'inflexion (figure 60.a). À Delhi, une plus grande part de compte a envoyé moins de 10 messages qu'a Bangkok (65 % contre 47 %). Si 80 % des utilisateurs ont émis moins de 100 messages à Bangkok, cette part monte a plus de 90 % à Delhi (figure 60.b). Ainsi, les comptes *Twitter* à Delhi ont tendance à envoyer moins de messages géolocalisés qu'a Bangkok. Néanmoins, dans les deux villes, certains utilisateurs émettent énormément de *tweets*, avec plusieurs dizaines de milliers de messages envoyés, et peuvent déjà être qualifiés de bots.

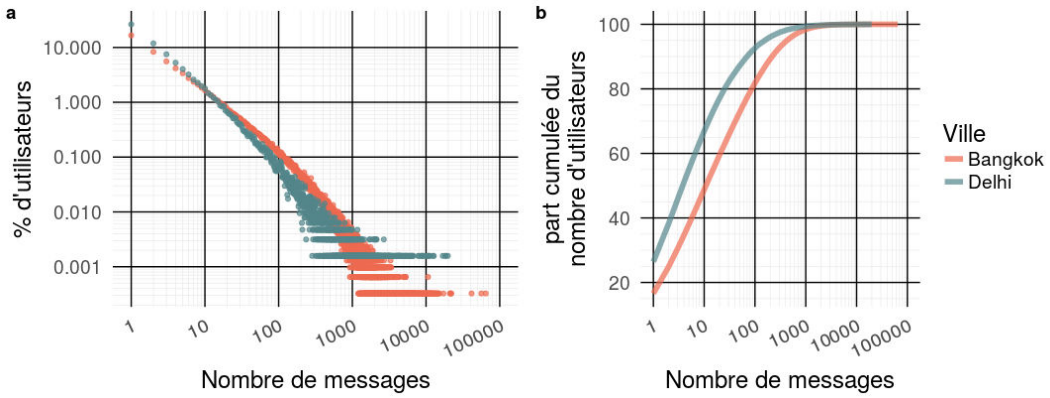


FIGURE 60 Nombre de *tweets* par utilisateurs à Delhi et Bangkok. (A) Part d'utilisateur et (B) part cumulée du nombre d'utilisateurs selon le nombre de messages envoyés.

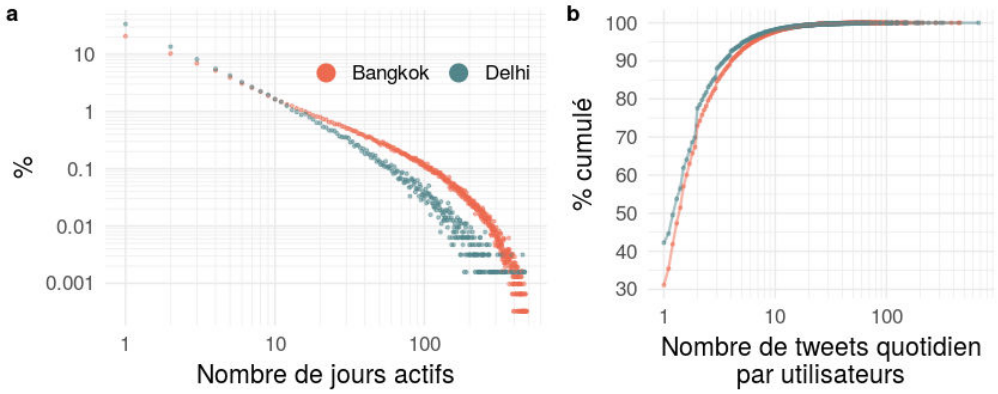


FIGURE 61 Nombre de jours actifs sur la période et nombre moyen de *tweets* envoyé par jours par utilisateur (activité).

Concernant l'activité des utilisateurs sur le réseau, nous pouvons observer que la distribution de la part d'utilisateurs en fonction du nombre de jours où ces derniers ont été actifs (figure 61.a), suit une loi puissance composée de deux fonctions linéaires décroissantes l'une sur un intervalle entre 1 et 200-300 jours, l'autre après cette zone d'inflexion, avec un coefficient directeur plus important (pente plus forte). Les utilisateurs de *Twitter* à Bangkok ont été plus actifs pendant plus de jours qu'à Delhi, et une plus grande proportion de comptes de la capitale indienne a émis des messages pendant moins de dix jours (figure 61.a). Concernant l'activité quotidienne moyenne sur le réseau (figure 61.b), 97 % des utilisateurs envoient moins de dix par jours en moyenne, et 67 % et 77 % moins de deux, respectivement à Bangkok et Delhi.

Si nous nous intéressons à la distance entre deux messages successifs pour un même utilisateur (figure 62), nous pouvons encore noter que la distribution suit une loi de puissance linéaire décroissante, jusqu'à environ 20 kilomètres, puis décroît plus fortement après cette distance. Plus de 40 % des messages sont envoyés à moins de 50 m du message précédent, et

entre 7 et 10 % à environ 100 m. À noter la présence de pics aux alentours de 500-700 km pour Bangkok et à diverses distances pour Delhi (200, 300, 900, 1100, 1200 km). Cette augmentation correspond sans doute à des sauts entre deux villes, par exemple lorsqu'une personne envoie un *tweet* avant et après avoir pris l'avion. Chiang Mai et certaines îles thaïlandaises prisées des touristes se situent en effet entre 500 et 600 km de Bangkok, tout comme Kolkata ou Mumbai sont situés à plus de 1 000 km de Delhi.

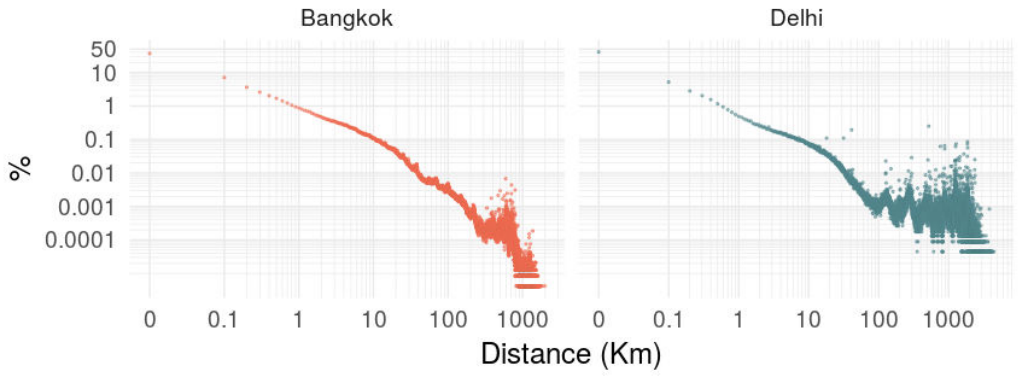


FIGURE 62 Part des distances entre deux messages successifs pour chaque utilisateur à Delhi et Bangkok.

Cette forte propension à envoyer des messages sur de courtes distances a pour corollaire une part très importante de *tweet* envoyé dans un pas de temps très court (figure 63 a et b). Ceci est très cohérent, notamment vis-à-vis de caractère "instantané" de *Twitter*, où l'envoi d'un message peut susciter des réponses immédiates et des réactions en chaîne, de la part d'un ou plusieurs utilisateurs. À Bangkok et Delhi, respectivement 32 et 48 % des *tweets* géolocalisés sont envoyés moins de dix minutes après un message précédent et moins d'un jour pour 80-85% d'entre eux. Nous pouvons observer que la courbe de la figure 63.a ondule après un intervalle d'un jour, dans le sens où l'on note une série de pics et de creux et qui se rapprochent du fait de l'utilisation d'une échelle logarithmique.

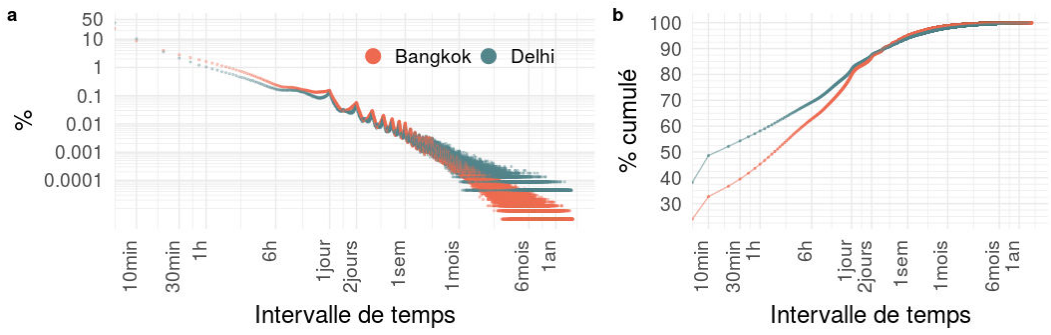


FIGURE 63 Proportion (a) et pourcentage cumulé de *tweets* selon l'intervalle de temps entre deux messages successifs (envoyés par chaque utilisateur) à Bangkok et Delhi.

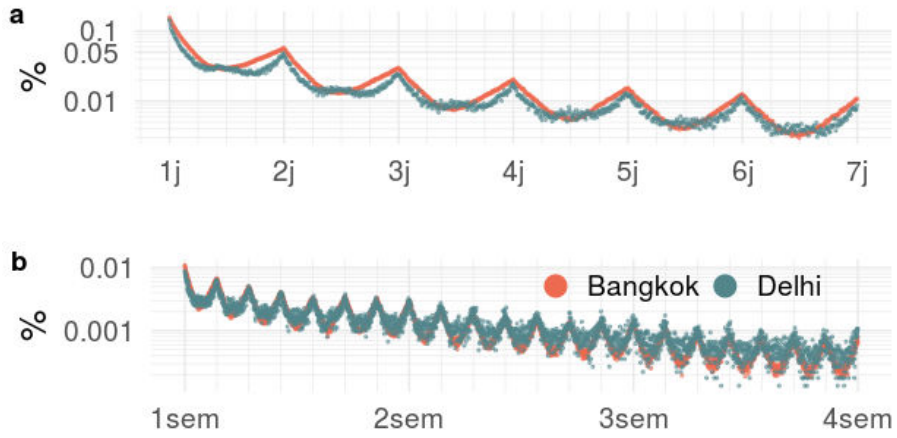


FIGURE 64 Proportion de *tweets* selon l'intervalle de temps entre deux messages successifs envoyés par chaque utilisateur. Zoom sur une semaine (a) et sur un mois (b).

Le fait d'effectuer un zoom sur les intervalles de temps situés entre un jour et une semaine (figure 64.a) ou entre une semaine et un mois (figure 64.b), permet de rendre compte du caractère périodique de l'utilisation du service, avec des pics toutes les 24 h et un léger pic toutes les 12 h pour Delhi. Cela peut signifier que les personnes qui tweetent peu fréquemment ont plus de chance de le faire aux mêmes moments de la journée. Par exemple une personne peut envoyer un message en sortant de son travail un jour donné, et le message suivant quelques jours plus tard, mais approximativement à la même heure, lorsqu'elle est disponible pour émettre un message ou pour signaler par exemple la fin de cette activité. Collectivement, cela peut signifier la présence de pics horaires quotidiens très marqués. À noter que la distribution issue des données provenant de Bangkok est beaucoup plus nette et moins bruitée que celle de Delhi, ce qui s'explique par une plus grande régularité.

Connaissant la distance et l'intervalle de temps entre deux messages, il est possible de calculer les vitesses de déplacement des utilisateurs entre deux *tweets* (figure 65). La figure 65.a montre la part de messages envoyés selon la vitesse de déplacement. Nous observons encore une loi de puissance décroissante à longue queue, ou un nombre non négligeable d'activités sur *Twitter* suggère des vitesses de déplacement aberrante, supérieure à celle d'un avion (~600 km/h). Cette part est d'autant plus importante à Delhi qu'à Bangkok (figure 65.b), où plus de 30 % des messages impliquent des vitesses supérieures à 1 000 km/h parmi les utilisateurs dans la capitale indienne contre moins de 1 % pour Bangkok. Si environ 92 % des tweets envoyés dans la capitale thaï supposent des vitesses de déplacement inférieure à 10 km/h ce qui est tout à fait crédible cette part n'est que 75 % à Delhi. Ainsi, les pourcentages de *tweets* émanant de bots devrait être nettement plus important à Delhi qu'à Bangkok. La vitesse moyenne de chaque compte *Twitter* (figure 65.c et 65.d) peut être un autre indicateur révélateur de la présence de bot. En effet, en contexte urbain, nous pouvons supposer qu'un

individu ne peut se déplacer plus vite qu'une voiture ou qu'un train. Or, entre 2 et 4 % des comptes *Twitter* présentent des vitesses moyennes dépassant les 50 km/h, seuil où l'on observe une légère inflexion de la courbe (figure 65.d). Néanmoins, 85 et 90 % des utilisateurs ont une vitesse moyenne inférieure à 10 km/h à Delhi et Bangkok, respectivement. Ces pourcentages passent à 50 et 55 %, pour les utilisateurs arborant une vitesse moyenne inférieure à 1 km/h, du fait d'une faible fréquence temporelle dans l'envoi des messages et/ou de petites distances parcourues.

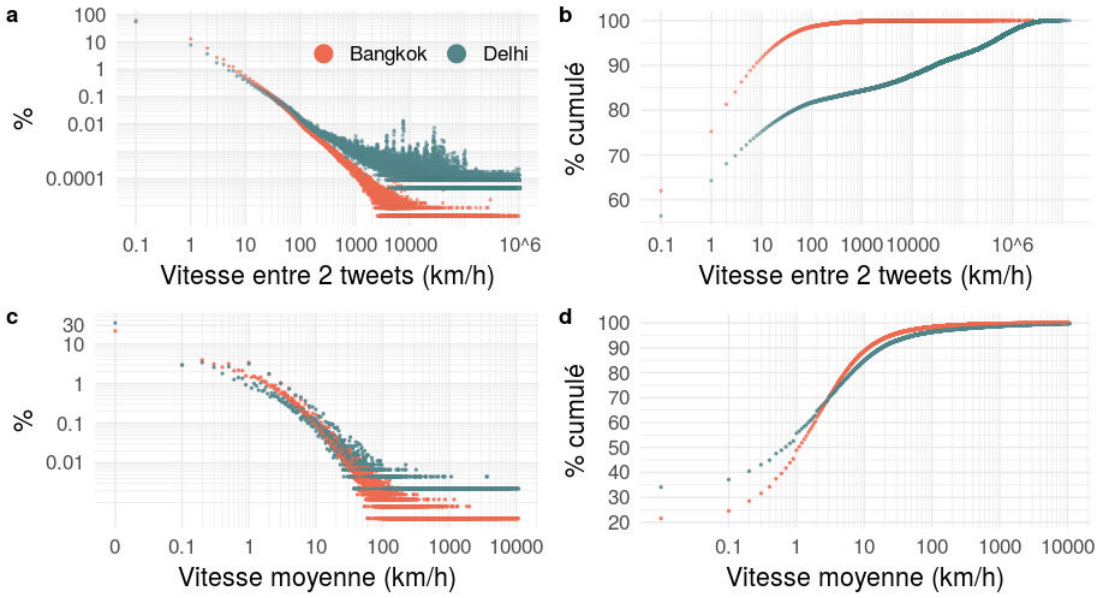


FIGURE 65 % de *tweets* envoyés à une vitesse donnée (a) et pourcentages cumulés (b). % du nombre utilisateurs en fonction de leur vitesse moyenne (c) et cumulée (d).

Ces analyses descriptives des données permettent déjà d'éclairer quelques tendances de l'échantillon à notre disposition. Les distributions du nombre de messages envoyés (figure 59.) et de l'activité (figure 61) des utilisateurs, ou des distances parcourues (figure 62), de l'intervalle de temps (figures 63 & 64) et des vitesses (figure 65) entre deux messages présentent toutes la caractéristique de suivre des lois de puissances. Ceci peut paraître fascinant pour des mathématiciens ou des physiciens, et même si les coefficients de ces courbes sont parfois différents localement, il n'est pas étonnant qu'un bon nombre d'études se soient focalisées sur leur caractère prédictif de tels jeux de données, notamment des déplacements (voir chapitre précédent). Néanmoins, ces relations, proche des lois rang/taille impliquent d'un côté qu'une grande proportion d'utilisateurs *tweet* très peu, sur quelques jours seulement, et qu'une infime part a une activité démesurée et des vitesses de déplacement tout simplement aberrantes. Les premiers peuvent être des utilisateurs très occasionnels, dont la faible quantité de traces numériques laissées sur le réseau social ne peut en aucun cas être représentative de l'ensemble des lieux qu'ils ont fréquentés, et les seconds des bots qu'il convient de supprimer.

1.3 Filtres et prétraitements

Comme présenté précédemment, nous avons à notre disposition de grands volumes de données, concernant énormément de comptes *Twitter*. Mais nous n'avons aucune idée de la représentativité de notre échantillon, et une part non négligeable de messages est très certainement émise par des bots. De plus, si dans un contexte d'analyse simple des niveaux d'attractivité de certaines zones de Bangkok ou Delhi, le fait de dissocier les bots des utilisateurs réels peut être suffisant, mais notre démarche s'ancre dans l'étude des mobilités quotidiennes individuelles et la création d'espace d'activité (chapitre 6). Nous faisons ici le postulat que plus le nombre de traces laissées par un individu est important, plus l'écart entre l'espace d'activité virtuel (l'ensemble des traces *Twitter* géolocalisées et datées) et réel (l'ensemble des lieux réellement fréquentés) tend à se réduire. Ceci induit qu'il convient également de supprimer de notre base les utilisateurs dont la quantité de *tweets* n'est pas assez importante pour arriver à reconstruire un espace d'activité individuel le plus proche possible de la réalité.

1.3.1 Suppression des utilisateurs occasionnels et des bots

Le type de média utilisé dans l'envoi d'un tweet aurait pu être utilisé pour filtrer et séparer les utilisateurs qui envoient leurs messages depuis l'une des plate-formes de *Twitter* (via Android, iOs, etc.) des autres. Mais nous n'avons à notre disposition que le nom de l'utilisateur, la date et la localisation d'un message. Néanmoins, les observations précédentes permettent déjà de poser que certains comportements sur le service (activité, vitesse entre les messages)sont purement artificiels, et l'application de seuils, même assez arbitraire, devrait permettre de supprimer ces comptes.

Tout d'abord, nous posons qu'un utilisateur réel ne doit pas avoir une activité trop importante et décidons de supprimer ceux qui envoient en moyenne plus de 20 messages par jours. Si cela ne concerne qu'environ 0,6 % des comptes, ces derniers contribuent à 11 % et 29,6 % des *tweets* envoyés à Bangkok et Delhi respectivement (tableau 7).

Comme nous l'avons vu, l'intervalle de temps et la distance entre deux *tweets* consécutifs impliquent parfois des vitesses de déplacement irréalistes, parfois égales ou largement supérieures à celle d'un avion. C'est l'approche choisie par (Hawelka *et al.,* 2014 ; Jurdak *et al.,* 2015b ; Lamanna *et al.,* 2018), en supprimant les utilisateurs ayant une vitesse respectivement supérieure à 1000 et 864 km/h (240 m/s) et 100 km/h. Mais il est possible qu'un utilisateur prenne réellement l'avion, comme le suggèrent les pics sur les distances parcourues qui correspondent aux distances entre nos zones d'études et d'autres grandes zones urbaines du pays. De plus, nous n'avons que peu d'information sur le fonctionnement de l'*API Stream*, et certains messages géolocalisés à différents endroits mais à différents moments peuvent être reçus en "rafale" par l'*API*, ce qui entraîne *de facto* de très grandes vitesses de déplacement, sans pour autant

qu'il s'agisse de messages envoyés par un bot. De plus, parmi les comptes présents dans notre échantillon, nous avons repéré un utilisateur connu par des membres notre entourage. Si nous appliquons l'un des seuils proposés précédemment, ce dernier serait considéré comme un bot. Nous décidons donc de laisser une part de doute, en supprimant les utilisateurs dont plus de 2 % de leurs messages envoyés à des vitesses supérieures à 600 km/h. Si ces seuils sont difficiles à justifier (voir annexe H), ils permettent de prendre en compte les incertitudes évoquées ci-dessus, et de supprimer 17,4 % des messages à Delhi et 2,17 % à Bangkok pour respectivement 2,41 et 1,6 % des utilisateurs.

Le pendant de ces vitesses ponctuellement irréalistes est une vitesse moyenne de déplacement extravagante. L'acte de tweeter n'est pas continu dans le temps, et les vitesses moyennes des utilisateurs calculées à partir des données *Twitter* sont donc forcément inférieures à la vitesse moyenne de déplacement réel. Néanmoins une personne peut envoyer des messages lorsqu'elle se trouve dans un moyen de transport, et nous décidons donc de supprimer tous les utilisateurs présentant des vitesses moyennes supérieures à 50km/h. Ce seuil assez élevé, mais crédible en contexte urbain est conforté par la présence d'une inflexion dans la figure 65.d. 2,25 % des utilisateurs sont ainsi concernés à Bangkok et 3,46 % à Delhi, représentant respectivement 3,41 et 22,2 % des messages envoyés (tableau 7). Ces deux critères de vitesses (instantanée et moyenne) permettent donc de supprimer des bots potentiels, aux déplacements irréalistes, sans écarter le compte *Twitter* associé à une personne réelle qui nous sert ici d'étalonnage.

Quelques comptes *Twitter* ont également pour pseudonyme une succession de 12 caractères hexadécimaux, ce qui peut s'apparenter à des adresses MAC (chapitre 4.1.2) et sont donc vraisemblablement des bots. Ils représentent 1275 comptes à Bangkok et 1643 à Delhi.

À ces utilisateurs que nous considérons comme des bots et que nous supprimons de notre échantillon, s'ajoutent ceux dont l'activité sur le réseau social nous paraît trop faible pour pouvoir tirer des conclusions ultérieures. Il peut également s'agir de personnes de passage, comme des touristes visitant l'Inde ou la Thaïlande. Si leur présence participe à la création de certains espaces, et si une sorte de routine apparaît lorsque l'on raisonne en termes de groupe et non d'individu (Cebeillac et Le Bigot, à paraître), nous préférons dans un premier temps exclure ces personnes de notre échantillon. Pour cela, nous ne considérons pas les comptes *Twitter* ayant été actifs pendant moins de 15 jours différents, ce qui concerne tout de même près de 200 000 comptes à Bangkok et plus de 50 000 à Delhi, avec des volumes de *tweets* relativement faibles respectivement 5,86 et 12 % à Bangkok et Delhi (tableau 7).

Le tableau 7 consigne tous les volumes de données supprimés. Certains comptes peuvent être évincés dans plusieurs catégories. Par exemple des utilisateurs qui présentent plus de 2 % de leurs *tweets* envoyés à plus de 600 km/h peuvent aussi avoir une moyenne de déplacement

supérieure à 50 km/h. Comme l'avaient suggéré les analyses des données brutes effectuées précédemment, une part nettement plus importante de comptes et d'utilisateurs sont écartés à Delhi qu'à Bangkok.

	Bangkok		Delhi	
	Nombre de Tweets	Nombre d'utilisateurs	Nombre de Tweets	Nombre d'utilisateurs
Données initiales	**29 492 225**	**307 428**	**3 167 004**	**63 495**
plus de 20 tweets par jours	3 240 272 (10,98 %)	1803 (0,58 %)	925 005 (29,20 %)	389 (0,61 %)
plus de 2 % des tweets envoyés à plus de 600 km/h	639 564 (2,17 %)	5104 (1,66 %)	551 883 (17,42 %)	1535 (2,41 %)
Vitesse moyenne supérieure à 50 km/h	1 007 616 (3,41 %)	6925 (2,25 %)	705 288 (22,27 %)	2200 (3,46 %)
moins de 15 jours de Tweets	1 730 153 (5,86 %)	195 646 (63,63 %)	381 236 (12,03 %)	51 832 (81,63 %)
pseudo hexadécimal	22 796 (0,07 %)	1275 (0,41 %)	15 021 (0,47 %)	1643 (2,58 %)
Total supprimé	5 613 875 (19,03 %)	200 341 (65,1 %)	1 503 501 (47,47 %)	53 019 (83,5 %)
Total gardé	**23 878 350 (80,96 %)**	**107 087 (34,8 %)**	**1 663 503 (52,52 %)**	**10 476 (16,5 %)**
plus de 50 % des jours de *tweet* depuis Delhi ou Bangkok	**17 838 798 (60.5 %)**	**76 547 (24.9 %)**	**915 918 (28.9 %)**	**5831 (9.18 %)**

Tableau 7 Évolution du nombre de *tweets* et d'utilisateurs selon les critères de sélections.

Sur les 107 087 et 10 476 comptes de notre échantillon que nous considérons comme exploitable à Bangkok et Delhi, nous ne gardons que ceux habitants potentiellement dans les zones d'études. Nous ne conservons ainsi que les utilisateurs ayant émis des *tweets* depuis ces zones lors de plus de la moitié des jours où ils ont été actifs sur le réseau social. Cette opération conserve 76 547 utilisateurs (71,5 % des 107 087) pour 17,84 millions de *tweets* (74,7 % des 23,87 millions) à Bangkok, et 5831 comptes pour 915 918 *tweets* à Delhi. Ces 76 547 comptes *Twitter* conservés à Bangkok et ses alentours représentent 0,5 % des 15,1 millions d'habitants de la zone (census 2010), pour près de 18 millions de traces numériques, ce qui en fait un échantillon assez colossal. La population de Delhi et ces villes satellites était d'environ 22 millions en 2011[290] (census 2011), et les 5831 comptes ne représentent alors qu'environ 0,026 % des habitants de la région. Ainsi, le résultat de ces filtres suggère que l'utilisation de *tweets* géolocalisés sera probablement plus pertinente à Bangkok qu'a Delhi.

290. Avec en millions : Delhi : 16,78 ; Gurgaon : 0,87 ; Faridabad 1,4 ; Ghaziabad : 2,35 et Noida : 0,64.

1.3.2 Des points GPS partagés par plusieurs utilisateurs...

Bien après avoir filtré ces données, nous nous sommes rendu compte en étudiant les données de *Facebook Places* (section suivante) qu'un nombre non négligeable d'utilisateurs fréquentaient des localisations avec exactement les mêmes coordonnées. Par exemple à Bangkok, sur les 23,8 millions de *tweets*, nous n'avons enregistré que 6,42 millions de localisations différentes.

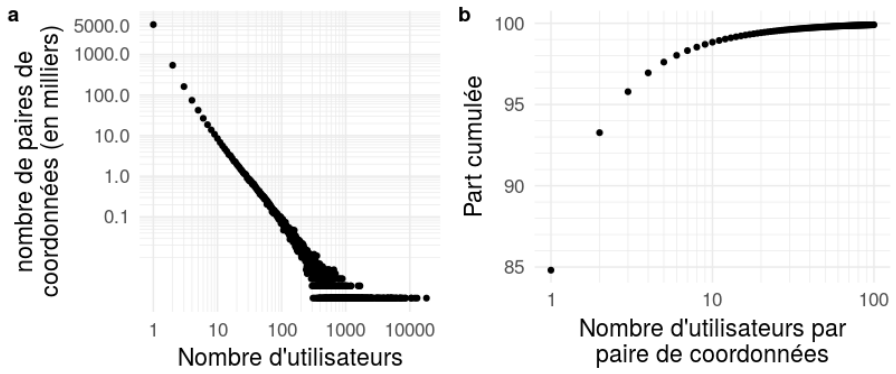


FIGURE 66 Des utilisateurs qui *tweet* exactement dans la même localisation à Bangkok. Nombre de paires de coordonnées en commun à plusieurs utilisateurs (a). Et part cumulée du nombre d'utilisateurs ayant des *tweets* aux coordonnées identiques.

Sur la figure 66, nous pouvons voir que 85 % des lieux à Bangkok ont des coordonnées propres à un seul utilisateur, c'est-à-dire qu'il est le seul a avoir envoyé un *tweet* depuis ce lieu unique. Si 15 % des lieux sont communs à plusieurs utilisateurs, moins de 1 % des lieux sont partagés par plus de 10 personnes (figure 66.b). Nous pouvons poser plusieurs hypothèses explicatives. Tout d'abord, il est possible que les coordonnées renvoyées par l'*API STREAM* soient parfois tronquées à deux ou trois décimales, ce qui peut déjà associer deux localisations très proches. Il est également envisageable que *Twitter* adapte la géolocalisation du message en triangulant avec d'autres signaux, comme des antennes relais ou des Wi-fi publics[291]. Si ces hypothèses peuvent valoir pour des lieux partagés par quelques personnes, la récupération de localisations provenant de bases de données externes paraît être une piste plus sérieuse pour les coordonnées partagées par un grand nombre de personnes. Ainsi, des personnes qui utilisent Swarm, Foursquare ou encore Instagram peuvent également partager leur géolocalisation sur *Twitter*. Dans ce cas, ce n'est plus la géolocalisation de l'appareil qui est envoyée, mais les coordonnées précises du lieu que la personne dit visiter. Ceci implique des biais, de nature différente en fonction des plateformes utilisées. Dans le cas de *Swarm* ou *Foursquare*, l'utilisateur doit être à proximité d'un lieu de leur base pour pouvoir faire un *check-in* (ou enregistrement). En revanche, dans le cas d'*Instagram*, la personne choisit un lieu dans une liste sans qu'il y ait

291. https://support.tw tter.com/art c es/231371 "*Twitter* peut utiliser divers signaux pour déterminer la localisation précise de votre appareil, y compris le GPS, le signal d'une antenne relais et les données des points d'accès sans fil voisins".

de réelles contraintes spatiales même si les lieux les plus proches sont proposés en priorité. Si les biais relatifs à la précision de la géolocalisation réelle sont donc potentiellement plus grands pour les messages émanant de la plateforme *Instagram*, nous pouvons supposer que le partage cette information est effectué assez rapidement sur *Twitter* ce qui ne devrait pas trop impacter la précision temporelle. Néanmoins, l'envoi de messages contenant une géolocalisation et provenant d'autres plateformes (*Foursquare* ou *Instagram*) peut aussi être à l'origine de vitesses de déplacement très élevées (voir supra), notamment si un individu envoie un message contenant la géolocalisation précise de son téléphone sur *Twitter*, puis dans la foulée un message provenant d'*Instagram* et associé à un lieu très éloigné.

Coordonnées	Nombre d'utilisateurs / Nombre de *tweets*	Nom	Base de données
100.5346,13.74602	18 182 / 76 086	Siam Paragon	*Foursquare*
100.5394,13.74586	13 004 / 35 001	Central World	
100.561,13.81628	11 979 / 38 616	Central Plaza	*Foursquare*
100.5328,13.74593	10 639 / 27 466	Siam Center	
100.7512,13.69378	10 442 / 24 091	Suvarnabhumi Airport	
100.551,13.79955	8575 / 15 975	Chatuchack	*Foursquare*
100.5344,13.74623	8471 / 17 532	Siam Paragon	*Facebook*
100.6007,13.91428	7855 / 16 804	DMK Airport	
100.5605,13.73728	7481 / 19 535	Terminal21 Asok Mall	
100.494,13.7522	7049 / 17 414	Bangkok	*Facebook*

Tableau 8 Liste des 10 lieux les plus fréquentés à Bangkok et le nom du lieu associé, obtenu en cherchant les coordonnées via *Facebook / Instagram* et *Foursquare*. La colonne 'Base de données' est renseignée uniquement si les coordonnées de nos tweets sont exactement les mêmes que celles des points issus de *Foursquare* ou *Facebook/Instagram*.

Le tableau 8 présente les 10 principaux lieux dont les coordonnées sont communes au plus grand nombre d'utilisateurs et le nom correspondant recherché dans les bases de données d'*Instagram* (fournie par *Facebook*) et *Foursquare* (voir annexe G pour plus de précision). Seule la moitié de ces lieux a exactement les mêmes coordonnées géographiques dans une de ces bases de données. S'il est néanmoins très aisé de nommer les autres lieux compte tenu de leur localisation, ils doivent probablement provenir d'autres bases de données. Nous pouvons noter que le *mall* de

Siam Paragon apparaît à fois dans la base de *Foursquare* (18 182 utilisateurs uniques) et celle d'*Instagram* (8471 utilisateurs). La ville de Bangkok figure également dans ce classement (7049 utilisateurs pour 17 414 tweets), et il s'agit de personnes ayant partagé sur *Twitter* un message provenant d'*Instagram*.

Si les *malls*, les aéroports ou le marché de Chatuchack se réfèrent à des lieux concrets, la ville de Bangkok ne peut à notre échelle se résumer à des coordonnées géographiques ponctuelles. Nous extrayons de la base de données *Facebook* (voir section suivante) les 837 lieux de catégorie 'city', 'region' ou 'country' et trouvons 428 lieux présents dans nos *tweets* avec les mêmes coordonnées. Ils correspondent à 26 588 *tweets*, pour 10 989 utilisateurs ce qui reste très marginal au regard du volume total de nos données et nous les supprimons. Nous effectuons le même travail pour Delhi.

1.4 Temporalité à Bangkok et Delhi

À partir de ces données filtrées, il est maintenant possible de faire ressortir l'activité moyenne par tranche horaire sur une semaine type à Delhi et à Bangkok. La figure 67 illustre cela, en comptant pour chaque utilisateur le nombre de *tweets* unique émis par tranche horaire depuis un lieu de son espace d'activité.

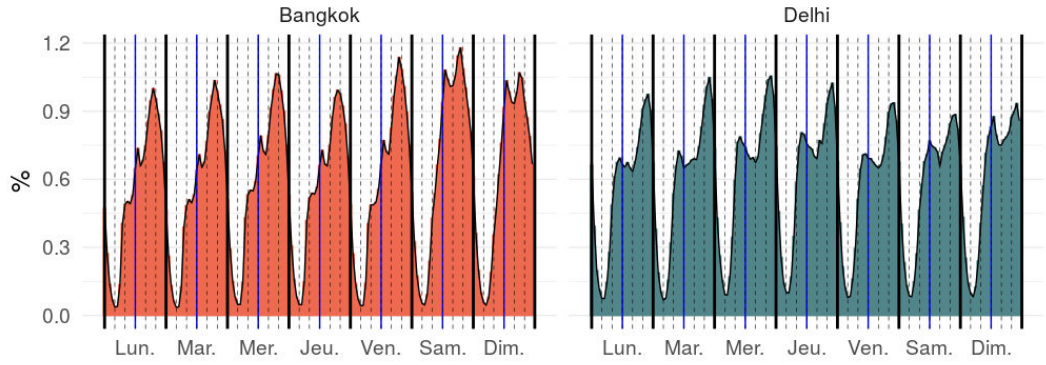


FIGURE 67 Nombre de messages envoyés par tranche horaire sur une semaine à Delhi et Bangkok, après les filtres. Les lignes en pointillé symbolisent un pas de temps de 4 h. La ligne bleue représente midi tandis que la ligne noire minuit.

Nous pouvons tout d'abord noter que si les profils dans les deux villes sont très différents, une distinction nette est visible entre l'activité en semaine et durant le week-end. Les profils des pourcentages des *tweets* envoyés par tranche horaire à Bangkok sont plutôt très cohérents, car nous pouvons noter les jours de semaines un premier pic d'activité vers 8h du matin soit le départ du domicile ou de l'arrivée au travail puis un léger plateau jusqu'à un second pic vers 12-13h correspondant à la pause déjeuné. L'activité augmente rapidement à partir de 16 h, pour

atteindre un maximum aux alentours de 20 h, plus important le vendredi que les autres jours, soit après le travail et pendant les heures de sorties. Le pourcentage minimum de tweet est observé vers 4 h du matin. Le pic de 8 h est absent le week-end ce qui est tout à fait cohérent. Plus de *tweets* sont enregistrés les après-midi, avec un écart d'amplitude entre les pics de midi et du soir nettement moins marqué qu'en semaine. À noter que le pic de soirée survient plus tôt le dimanche, vers 18h, et que l'activité y est globalement moins importante que le samedi.

Pour ce qui est de l'activité horaire enregistrée à Delhi, nous pouvons noter la présence de deux pics. Le premier, moins marqué survient les jours de semaine entre 9 h et 11 h et plus vers midi le week-end. Le pic du soir est plus tardif qu'à Bangkok puisqu'il survient aux alentours de 22 h, tous les jours de la semaine. Ces différences majeures entre Bangkok et Delhi peuvent peut-être s'expliquer par les différences de temporalités dans les deux villes, mais surtout par la taille et la composition de l'échantillon. En effet, à Delhi nous n'avons que 5831 utilisateurs ce qui n'est probablement pas suffisant pour faire ressortir des tendances fiables. En revanche à Bangkok, la taille de l'échantillon est assez importante et les profils temporels de l'activité sur *Twitter* aisément justifiable par les diverses étapes d'une semaine type.

1.5 Création des espaces d'activité individuels

1.5.1 Formalisation de la méthode

Ces données filtrées, et conscient des biais évoqués précédemment, nous allons maintenant reconstruire les espaces d'activités de nos utilisateurs en créant des groupes de *tweets* selon leur proximité spatiale. Car les traces numériques des utilisateurs de *Twitter* ne sont pas uniformément réparties dans l'espace et le nombre de *tweets* localisés dans une zone donnée dépend d'une part de la fréquence de visite de ce lieu par l'utilisateur et d'autre part de la volonté de ce dernier de le signaler sur le réseau social. Dès lors, compte tenu de la durée de collecte des données (environ un an et demi) il est très probable qu'un utilisateur ait pu tweeter à plusieurs moments dans une même zone géographique relativement restreinte. Regrouper les *tweets* provenant d'une même zone permet alors d'avoir un aperçu des jours et fréquences de visites de ces dernières, et l'ensemble de ces groupes de *tweets* pourrait alors s'apparenter à l'espace d'activité de l'utilisateur.

Nous allons donc dans un premier temps regrouper les *tweets* de chaque utilisateur en fonction de leur proximité, de façon à former des groupes, ou *cluster*. Nous considérerons ensuite chacun de ces *clusters* comme un lieu de l'espace d'activité de l'utilisateur. Nous posons ainsi l'hypothèse que plus une personne laisse de messages géolocalisés, plus l'écart entre l'espace d'activité réel et virtuel rétrécit. Il existe différentes méthodes pour effectuer cela et nous avons eu recours à l'algorithme dont l'utilisation est la plus répandue dans la bibliographie, à savoir dbscan (Ester *et al.*, 1996). Son principe est relativement simple pour les personnes

habituées aux SIG car il consiste tout simplement à effectuer un *buffer* d'une distance donnée sur l'ensemble des points géolocalisés d'un individu. Les couches spatiales qui s'intersectent sont ensuite fusionnées dans un même *cluster* (figure 68.a,b et c). Nous choisissons ici une distance de 50 m, qui sera la distance maximale entre deux points d'un même cluster.

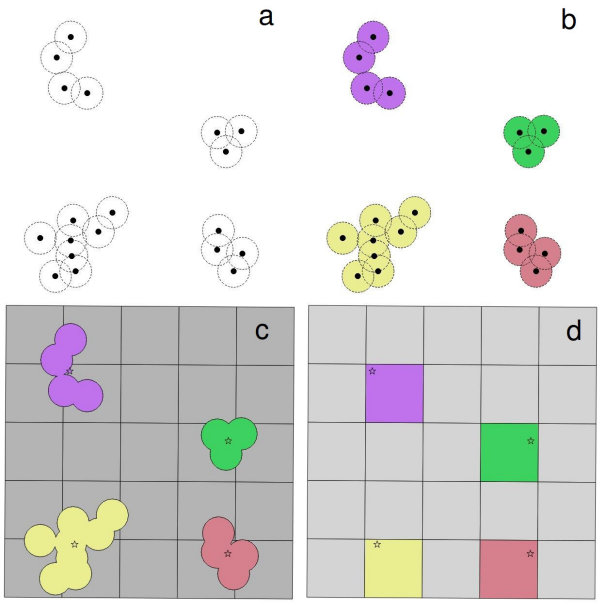


FIGURE 68 Principe de fonctionnement de l'algorithme *db-scan* et agrégation des *tweets* dans des "lieux" . Les points correspondent ici à des *tweets* auxquels nous appliquons un *buffer* (a). Les polygones qui s'intersectent sont regroupés entre-eux (b). Nous prenons ensuite le barycentre (c) que nous affectons à une maille de la grille (d). Ici, 20 points sont ainsi associés à 4 lieux, formant l'espace d'activité.

Nous affectons ensuite à chaque cluster les coordonnées moyennes des points qui le composent (figure 68.c). Ces barycentres sont ensuite agrégés à une grille de maille 180 m (68.d) pour diverses raisons. Tout d'abord, comme vu précédemment, ce travail s'ancre dans un projet de recherche plus large, et se doit d'être compatible avec les travaux de Maneerat (2016) et Misslin (2017) dont l'unité de base est une grille de 30 m, du fait de la résolution des images de Landsat 8 et des potentiels de déplacement des moustiques 180 m en est donc un multiple. Aussi, cette unité correspond environ à un pâté de maison, ce qui permet d'ajouter un peu de flou spatial sur la localisation exacte de la personne. Finalement, l'utilisation d'une grille nous a permis d'établir une typologie des fonctionnalités commerciales de la ville de Bangkok à partir des données issues de *Google Map* (Cebeillac *et al.*, 2017).

Nous avons ainsi défini 306 466 clusters à Delhi et 4 110 741 à Bangkok. Ces derniers forment ainsi les différents lieux des espaces d'activités des utilisateurs de notre échantillon. Nous pouvons noter une grande différence dans le nombre de lieux fréquentés selon les utilisateurs, mais également selon les zones d'études (figure 69). Si la plupart des individus fréquentent moins

de 50 lieux sur l'ensemble de la période de collecte, nous observons un mode aux alentours de 10 à Delhi, et de 17 à Bangkok, avec une part non négligeable d'individus ayant fréquenté plus de 100 lieux dans la capitale thaï.

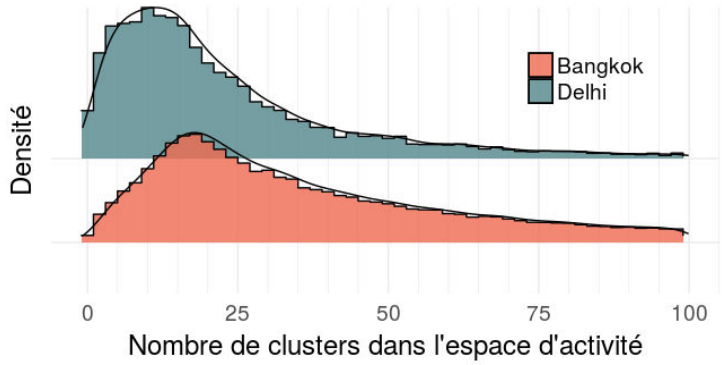


FIGURE 69 Densité du nombre d'utilisateurs en fonction du nombre de *clusters* (lieux) qui composent leur espace d'activité.

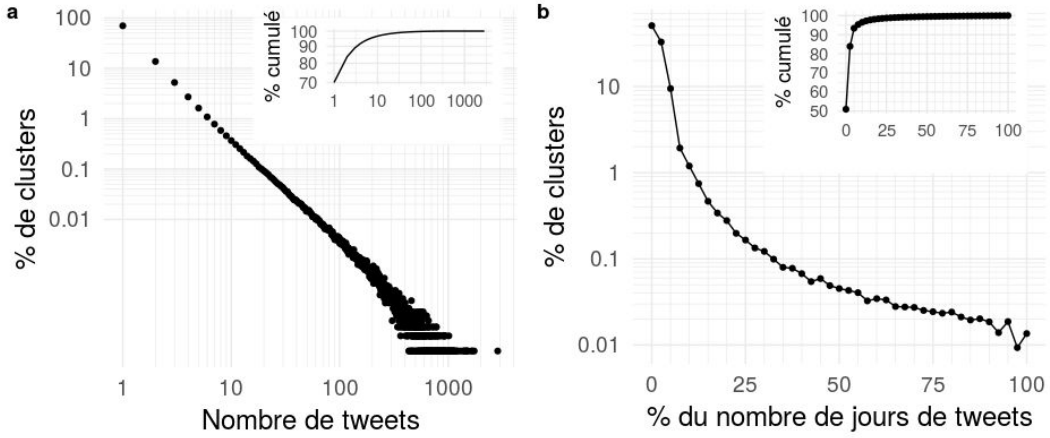


FIGURE 70 Pourcentage de clusters en fonction du nombre de *tweets* qui les composent (a) et pourcentage du nombre de jours où un utilisateur a *tweeté* depuis ce lieu (b), à Bangkok.

Chacun de ces lieux est aussi caractérisé par un nombre de *tweets* et les jours et heures d'envois de ces derniers. La figure 70.a montre que 70 % des clusters individuels créés à Bangkok ne sont composés que d'un seul tweet et 90 % d'entre eux en ont plus de 5 les mêmes observations étant également faites à Delhi. Cela signifie qu'un grand nombre de lieux est caractérisé par un nombre très faible de traces numériques, avec une fréquence de visite (visible sur la figure 70.b par le pourcentage de jours où un utilisateur a émis des messages) relativement basse. Ainsi, si nous reprenons le concept d'espace d'activité, cela peut se traduire par un grand nombre de lieux fréquentés de manière plutôt exceptionnelle, et un faible nombre de lieux plus routiniers.

Parmi tous ces lieux, un premier enjeu est d'arriver à estimer lequel est le domicile probable de l'utilisateur de *Twitter*. Il s'agit en effet d'une information primordiale dès lors qu'il s'agit de faire des recoupements avec d'autres bases de données, provenant notamment des recensements, afin d'évaluer le niveau de représentativité spatiale de l'échantillon.

1.5.2 Détection du domicile

Méthode

Le lieu du domicile est par essence très fréquenté, surtout le soir et se situe en zone résidentielle. Nous mobilisons dans un premier temps la base de données libre *OpenStreetMap* qui renseigne un type d'utilisation du sol (voir chapitre 8 pour une description plus détaillée). Nous partons du principe que le lieu de domicile ne se situe pas dans une zone commerciale, un parc, une rivière, un aéroport, une route ou encore une gare. Nous écartons les clusters définis précédemment qui intersectent cette base de données.

Nous partons du principe que le lieu de domicile est caractérisé par une activité globale importante, avec notamment un pourcentage de jours actifs sur *Twitter* et un nombre de *tweets* émis plutôt élevés, mais aussi par une activité entre le soir et le matin relativement prépondérante sur les autres lieux. La figure 71 présente les heures d'envois de messages pour deux utilisateurs pris au hasard, dans les différents lieux qui constituent leur espace d'activité. Si l'utilisateur 1 a envoyé des messages en soirée dans ses clusters 3, 6 et 9, l'activité de cette personne est bien plus importante dans le cluster 6, qui est un bon candidat pour être son lieu de domicile. Nous pouvons faire le même constat avec l'utilisateur 2, où le cluster 8 correspond probablement à son lieu de résidence, sous réserve qu'il soit situé en zone résidentielle.

De manière similaire à Luo *et al.* (2016), nous définissons les plages horaires nocturnes comme se déroulant entre 20 h et 08 h. Nous comptons ainsi pour chaque cluster le nombre de *tweets* uniques envoyé par heure et par jour dans cette tranche horaire, mais sans prendre en compte les *tweets* ayant des coordonnées géographiques partagées avec d'autres utilisateurs (voir supra). Ces derniers étant très probablement des lieux présents dans des bases de données, ils ne peuvent être un lieu de domicile. Néanmoins, si quelques *tweets* au sein d'un cluster émanent de ce type d'endroit, nous n'écartons pas le cluster pour autant, car le lieu de domicile peut se trouver dans le voisinage de cafés ou restaurants présent dans les bases de données de *Foursquare* ou d'*Instagram*. À partir de ces informations, ne reste plus qu'a définir des seuils limites de nombre de *tweets*, que nous calibrerons à partir des données de Bangkok, plus prolixes qu'a Delhi. Nous avons ainsi posé qu'un utilisateur doit avoir *tweeté* depuis son lieu (ou cluster) de domicile au moins 10 % des jours où il a été actif, avec au moins 6 jours d'activité et que les *tweets* émis entre 20 h et 8 h doivent représenter plus de 10 % des *tweets* uniques envoyés par heure. Au final, 38 696 utilisateurs (13 000 933 tweets dans 2 247 339 lieux) habitants dans la région de Bangkok ont des lieux qui répondent à tous ces critères et 25 127 utilisateurs

résideraient à Bangkok même. À Delhi, ces effectifs sont moins importants puisque seuls 4089 utilisateurs (et 758 552 tweets) habitent dans la région de la capitale Indienne, dont 2573 à Delhi même.

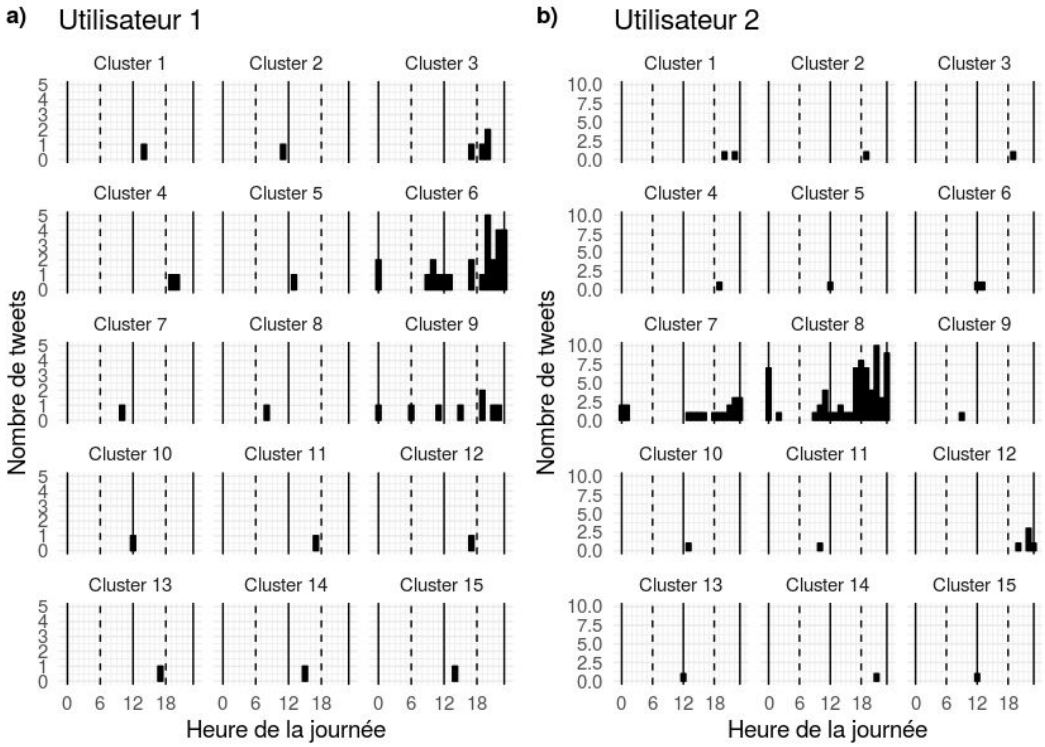


FIGURE 71 Nombre de *tweets* par tranche horaire par lieu fréquenté pour deux utilisateurs.

Représentativité spatiale

Après avoir estimé les localisations des domiciles des utilisateurs de notre échantillon, nous allons maintenant estimer le niveau de représentativité spatiale en confrontant le nombre d'utilisateurs de *Twitter* domiciliés dans un sous-district (*Khwaeng* à Bangkok et *Ward* à Delhi) et la population répertoriée par les différents recensements (2010 à Bangkok et 2011 à Delhi).

Bangkok

Le niveau de proportionnalité entre le nombre d'utilisateurs de *Twitter* par sous-district et la population enregistrée est très élevé à Bangkok, avec un R^2 de 0.81 (figure 72.a), ce qui suggère une très bonne représentativité spatiale. Ce R^2 passe à 0,79 si nous prenons en compte l'ensemble des sous-districts de Bangkok et sa région (annexe H). Cette baisse peut s'expliquer par la présence de beaucoup de sous-districts périphériques avec peu d'utilisateurs. À noter que dans des papiers précédents nous obtenions un R^2 plus faible, de 0,67 (Cebeillac *et al.*,

2017 ; Cebeillac et Le Bigot, à paraitre), car nous n'excluons pas les lieux présents dans d'autres bases de données dans l'estimation du domicile. Loin de palinodier ces précédents travaux, ces meilleurs résultats témoignent d'abord de l'importance d'une bonne connaissance des données à traiter, et confirment d'autant plus la bonne répartition des utilisateurs de *Twitter* dans Bangkok. Ce dernier aspect est d'ailleurs mis en avant par la figure 72.c qui de par le calcul du *K-ripley* montre que la localisation des domiciles des utilisateurs ne suit pas une loi de poisson et sont plutôt regroupés. De manière intéressante, le nombre de domiciles d'utilisateurs de *Twitter* est d'autant plus corrélé à la population au sous-district lorsque nous passons en logarithme. Ceci signifie que plus un sous-district est densément peuplé, plus le nombre d'usagers de *Twitter* y est important, suggérant un effet de masse.

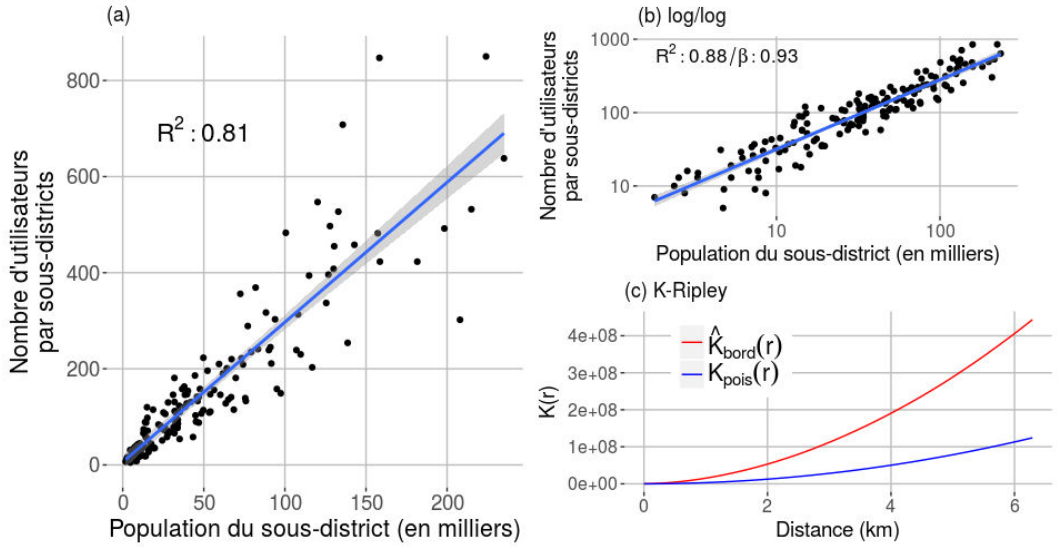


FIGURE 72 Représentativité de l'échantillon à Bangkok. Lien entre la population enregistrée au sous-district par le recensement de 2010 et le nombre d'utilisateurs de *Twitter* habitant dans les mêmes *Khwaeng* (a). En logarithme (b). Et écart à un modèle Poissonien d'après le K de Ripley (c).

La figure 73, cartographie les résidus sous forme de taux de variation. Un écart positif signifie une surreprésentation des utilisateurs de *Twitter* par rapport à la tendance générale. La plupart des sous-districts présentent des écarts à la prédiction compris entre -25 et 25 %, ce qui signifie un bon niveau de représentativité. Les sous-districts où la surreprésentation de l'utilisation du service est la plus grande sont surtout situés dans le centre-ville, zones relativement riches, avec une forte proportion d'expatriés (chapitre 2). Les secteurs où l'utilisation de *Twitter* est inférieure à la tendance générale se situent en périphérie, notamment dans le sud-ouest, zone assez peuplée mais néanmoins plutôt rurale.

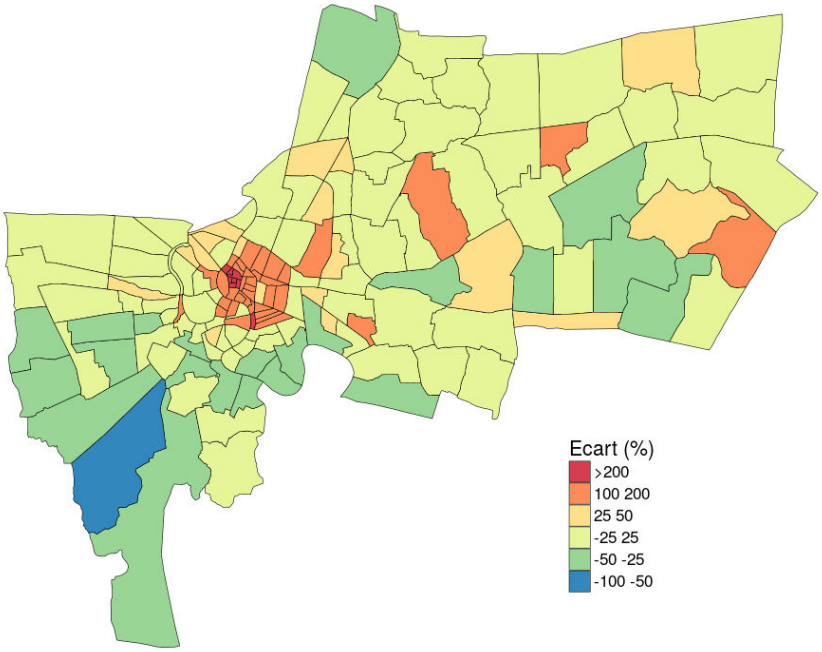


FIGURE 73 Cartographie des résidus entre la population enregistrée par le recensement et le nombre de domiciles estimés pour les utilisateurs de *Twitter*. Des valeurs élevées (qui tendent vers le rouge) signifient une surreprésentation des utilisateurs de Twitter par rapport à la population du sous-district, tandis que des valeurs négatives (bleu) indiquent l'inverse.

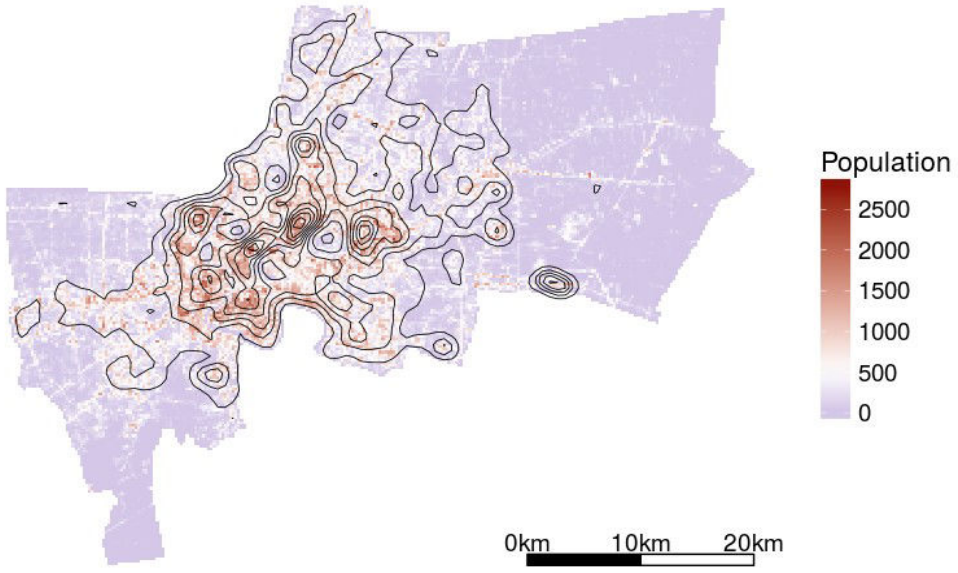


FIGURE 74 Densité des domiciles des utilisateurs de *Twitter* à Bangkok (isolignes) au regard de la population dans des mailles de 250m (adapté de Misslin et Daudé (2016)).

Le bon niveau de représentativité spatiale est aussi observable à une résolution plus fine (figure 74). Les plages de couleurs correspondent à la population estimée d'après une cartographie dasymétrique effectuée à partir des données de recensement et d'une cartographie de l'occupation du sol à partir d'image Landsat 8 (Misslin and Daudé, 2016) agrégée dans des mailles de 250 m de côté. Les isolignes correspondent à la densité du nombre de domiciles des usagers de *Twitter* (par tranche de 10 %), calculée selon une distance de 2,5 km. Les zones les plus densément peuplées sont globalement celles où nous enregistrons le plus grand nombre de domiciles des utilisateurs du réseau social.

Delhi

À Delhi en revanche, aucune relation n'existe entre la répartition du nombre de domiciles estimés et la population enregistrée au sous-district (figure 75). Les plages de couleurs de la correspondent à la population estimée d'après une cartographie dasymétrique effectuée à partir des données de recensement et d'une cartographie de l'occupation du sol à partir d'image Spot 5, agrégée dans des mailles de 250 m de côté (voir chapitre 2).

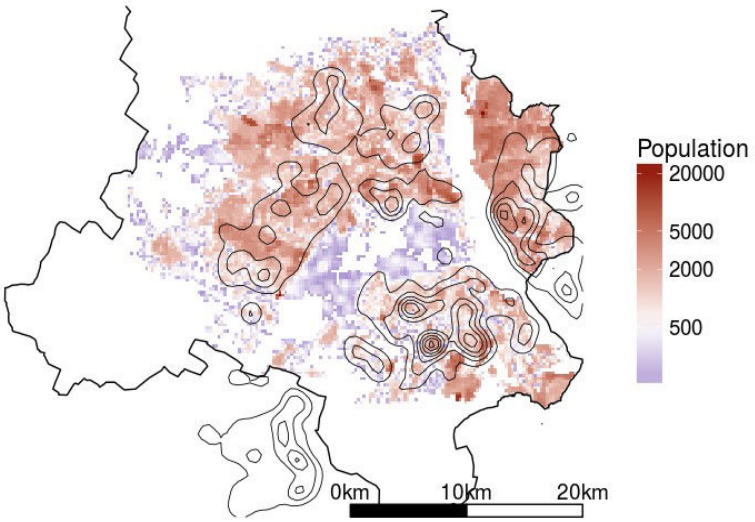


FIGURE 75 Densité des domiciles des utilisateurs de *Twitter* à Delhi au regard de la population dans des mailles de 250m.

Compte tenu de la grande disparité dans les densités de population, nous avons opté pour une transformation logarithmique afin de faire ressortir les nuances dans les quartiers moyennement peuplés. Les isolignes correspondent à la densité du nombre de domiciles des usagers de *Twitter* (par tranche de 10 %), calculée selon une distance de 2,5 km. Les utilisateurs de *Twitter* sont surtout concentrés dans les quartiers aisés du sud de la ville, ainsi qu'à l'est

et dans les villes satellites de Noida et Gurgaon. Quasiment aucun utilisateur de *Twitter* n'est domicilié dans le nord-est de la ville, pourtant l'une des zones les plus peuplées de la ville. Il n'est donc pas étonnant qu'aucune relation n'apparaisse entre la population enregistrée au *ward* par le recensement en 2011 et le nombre d'utilisateurs de *Twitter* habitant dans ces mêmes arrondissements (figure 76). À noter que la zone du domicile de la personne dont les traces numériques nous ont servi à étalonner les filtres précédents a été correctement estimée - alors qu'elle habitait à Delhi et que les seuils choisis l'ont été d'après l'observation des données à Bangkok (annexe H).

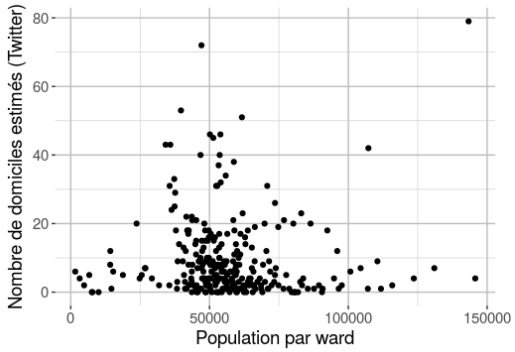


FIGURE 76 Lien entre la population enregistrée au sous-district par le recensement de 2011 à Delhi et le nombre d'utilisateurs de *Twitter* habitant dans les mêmes *ward*.

1.6 Synthèse

Les volumes de données enregistrés à Delhi et Bangkok sont très différents. Si *Twitter* est relativement peu utilisé dans la capitale indienne, nous y avons détecté une plus grande proportion de *bots*. Le niveau de représentativité spatiale est également très différent. Alors que les utilisateurs de *Twitter* à Bangkok sont plutôt bien répartis dans la ville, avec un bon niveau de représentativité au regard de la population enregistré au sous-district, ce n'est pas le cas à Delhi. Ainsi, ces données ne pourront être utilisées de la même manière dans nos deux zones d'études, et leur potentiel dans l'analyse des mobilités quotidiennes devrait être nettement meilleur dans la capitale thaï.

Les données *Twitter* que nous avons collectées sont des données individuelles. Elles permettent de reconstituer des espaces d'activités de chacun des usagers du service. Mais il est délicat d'évaluer le niveau de correspondance entre les traces numériques laissées dans certains secteurs de la ville et les temporalités de visite réelle (heures et fréquences) de ces lieux. Néanmoins, nous pouvons considérer que plus une personne émet des tweets géolocalisés, plus son espace d'activité réel est révélé. Ce dernier aspect soulève des questions éthiques vis-à-vis du respect de la vie privée et du maintien de l'anonymat. Nous utilisons ici une approche qui passe par l'ajout de divers niveau de flou. Tout d'abord un flou spatial, en regroupant les

différents lieux des utilisateurs dans des mailles de 180 m². Ensuite un flou temporel, en ne prenant pas en compte les jours et heures exacts de visites. Si une personne fréquente un lieu donné le mardi 7 avril 2015 à 9h10, nous ne prendrons en compte que le fait qu'il s'agissait d'un mardi à 9 h. Nous changeons également le pseudonyme de l'utilisateur, en leur attribuant un nouveau code, de manière aléatoire. Il est délicat d'obtenir un niveau d'anonymat élevé, car la plupart des individus ont des signatures spatio-temporelles (soit l'ensemble des lieux qu'ils fréquentent à des horaires donnés) uniques. Si par exemple plusieurs personnes habitent dans le même secteur et travaillent dans la même zone de la ville, il est fort probable que les lieux qu'ils fréquentent de manières moins routinières soient différents, ce qui devrait permettre de les différencier. Mais le fait d'avoir juste un espace d'activité propre à un individu (dont le pseudonyme a été changé) ne suffit pas à retrouver directement son identité. Il faudrait pour cela faire par exemple des enquêtes de terrains ciblées.

Les données *Twitter* que nous avons collectées permettent de raisonner à deux échelles : individuelle, où il est possible de reconstruire les espaces d'activités de chaque utilisateur ; ou agrégées, ce qui permet de déduire des niveaux d'attractivités en fonction des tranches horaires et des secteurs de la ville. Mais il est difficile d'apprécier précisément le niveau de véracité de ces données, à savoir si les espaces d'activités créés à partir des données *Twitter* correspondent aux lieux réellement fréquentés. En revanche, pour ce qui est des tendances globales de fréquentation des différentes zones de la ville, d'autres données sont mobilisables afin d'établir des points de comparaison, comme les *check-in* de *Facebook*.

2 *Check-in Facebook* à Bangkok

2.1 Présentation

Fondé en 2004, *Facebook* est le réseau social en ligne le plus utilisé au monde, avec 1,4 milliard d'utilisateurs quotidiens et 2,13 milliards d'utilisateurs mensuels en décembre 2017 [292]. Il est notamment leader en Inde et en Thaïlande (figure 53). Comme nous l'avons rapidement vu dans le chapitre 4, l'entreprise a activé une nouvelle fonctionnalité en 2010 qui permet aux personnes d'effectuer un "*check-in*", [293] c'est-à-dire de signaler leur présence dans un lieu donné. Contrairement à des sites comme *Foursquare* ou *Swarm*, l'utilisateur n'a pas de récompenses lorsqu'il effectue un *check-in*. Mais certaines études ont montré que plus une personne partage de données sur sa page personnelle, plus elle aurait tendance à effectuer de *check-in*, et que ce comportement relèverait plus d'une forme d'exhibitionnisme que de narcissisme (Wang et Stefanone, 2013).

Les données relatives à chaque lieu de la base de données, dont le nombre de *check-*

292. https://newsroom.fb.com/company nfo/
293. Appelé "Je suis là" en France.

in répertorié au moment de la requête est public et accessible facilement depuis l'*API Places Search*[294]. À partir de ces *check-in* et des informations sur les utilisateurs (données longitudinales), des chercheurs de *Facebook* ont montré dès 2011 qu'il était possible de prédire la localisation du prochain *check-in*, tout en montrant que des contacts sur *Facebook* ont plus de chance de fréquenter les mêmes lieux (Chang et Sun, 2011). Mais à notre niveau, nous ne pouvons récupérer que des données agrégées (nombre de *check-in* dans chaque lieu), ce qui peut expliquer que peu d'études ont eu recours à ces informations, qui représentent pourtant des volumes de données considérables (voir ci-après).

Création des données

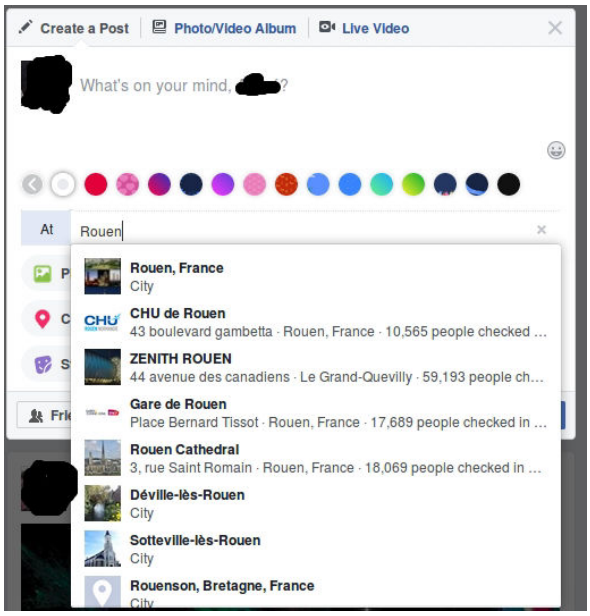


FIGURE 77 Création d'un *check-in* sur *Facebook*.

Lorsqu'une personne désire créer du contenu sur *Facebook*, elle peut choisir de signaler sa présence en cliquant sur l'onglet *"check-in"*. Elle doit néanmoins activer le service de géolocalisation[295] si elle utilise un téléphone fonctionnant sous *iOs* ou *Android*. Elle peut alors sélectionner un lieu dans une liste (figure 77), *Facebook* proposant en priorité les lieux situés à proximité de l'appareil.

Une personne qui utilise un téléphone fonctionnant sous *iOs* ou *Android* peut également créer un lieu. Dans ce cas une carte apparaît avec un signet représentant la localisation du téléphone que la personne peut déplacer, pour ajuster la position du lieu. L'internaute peut ensuite nommer ce lieu, ajouter diverses informations (horaires d'ouverture, site Internet, adresse exacte, etc.) mais elle doit surtout choisir le type d'activité correspondant, parmi

294. https://deve opers.facebook.com/docs/p aces/web/search
295. https://www.facebook.com/he p/andro d app/174846215904356?he pref=p atform sw tcher&ref=p atform sw tcher

plusieurs milliers et réparti dans 8 grandes classes : *Nourriture et boisson*, *Art et divertissement*, *Éducation*, *Fitness* et *loisirs*, *Hôtel et hébergement*, *Santé/médecine*, *Commerces*, ou encore *transports et voyages*. À noter qu'il est possible de renseigner jusqu'à 3 catégories.

Lorsqu'une personne effectue un *check-in*, elle peut également ajouter des personnes présentes dans ces contacts, qui seront associées à cet endroit et donc comptabilisées. Si par exemple un utilisateur de *Facebook* crée un message contenant un *check-in* avec 8 de ces "amis" dans un lieu donné, le nombre de *check-in* dans le lieu sera incrémenté de 9. Si la personne à l'initiative du message efface ce dernier, le nombre de *check-in* baissera de 9. Si une personne citée supprime son pseudonyme du message, le nombre de *check-in* ne baissera que de 1.

Validation des données

Étant donné qu'il est extrêmement simple de créer un lieu et que les utilisateurs de *Facebook* sont assez libres dans la manière de renseigner des informations associées (localisation, type, nom, etc.) *Facebook* a mis en place un système de validation qui repose sur la contribution des utilisateurs. Il suffit de cliquer sur l'onglet "Suggérer des modifications", présent sur la page d'accueil pour ouvrir une nouvelle fenêtre. Le site va alors présenter un lieu, et poser une question à l'internaute, comme s'il s'agit de la bonne adresse, de la bonne ville ou encore de la bonne catégorie.

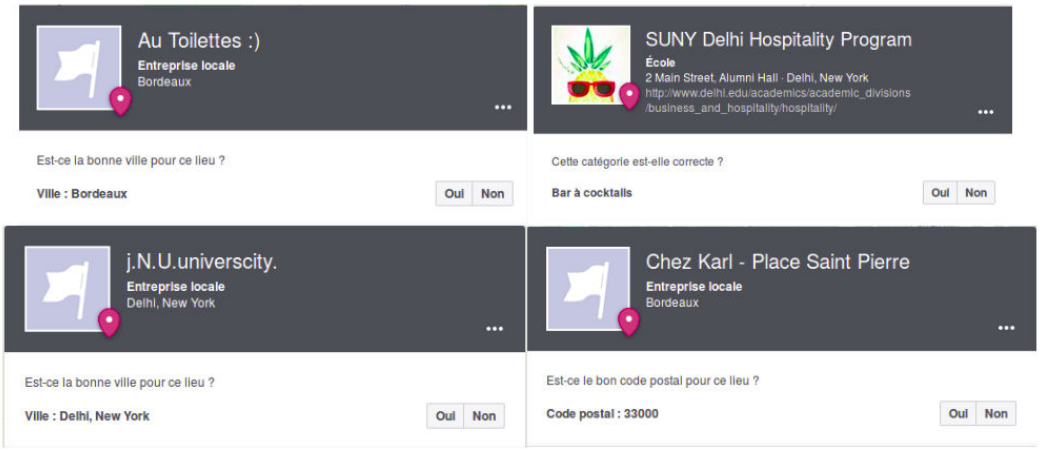


FIGURE 78 Quelques exemples de lieux mal référencés dans la base de *Facebook*.

La figure 78 illustre quelques questions relatives à 4 lieux créés. Par exemple une personne a créé un lieu nommé "Au Toilettes :)", considéré comme une entreprise locale et *Facebook* demande s'il s'agit de la bonne ville. De même, *Facebook* s'interroge si le "SUNNY Delhi Hospitality Program" est bel et bien un bar à cocktail, si l'Université JNU de Delhi, en Inde, se trouve bien à Delhi dans l'État de New-York aux états Unis, ou encore si le restaurant "Chez Karl - Place Saint Pierre", qui se situe en réalité place du Parlement à Bordeaux, a le bon code

postal. Ces exemples, volontairement sélectionnés pour leur manque d'exactitude et/ou leur incohérence permette de montrer l'existence de biais assez important, au moins pour certains lieux.

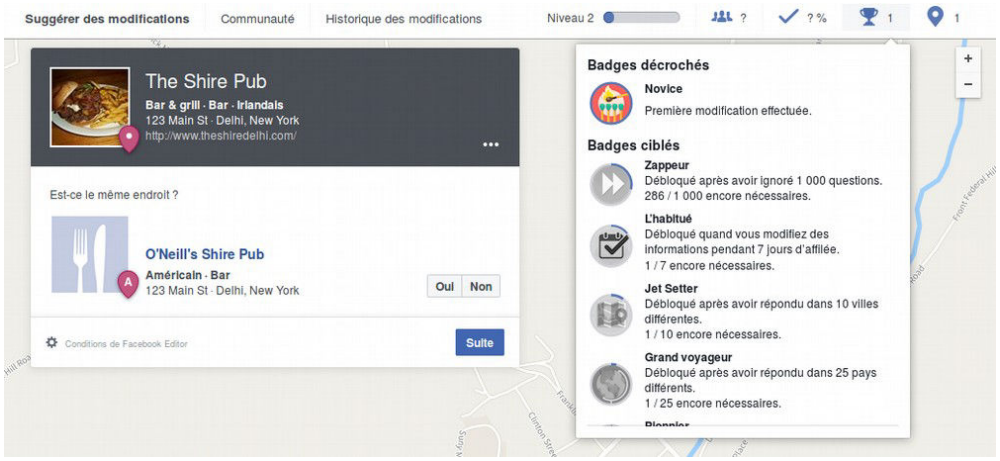


FIGURE 79 Exemple de badges et autres récompenses qu'une personne éditant des contenus sur *Facebook* peut recevoir.

Si l'utilisateur effectue des modifications, il se voit récompensé par l'attribution de badge en fonction des "quêtes" qu'il a pu compléter (par exemple répondre à des questions dans 10 villes différentes) et voit alors sa barre de niveau monter légèrement (figure 79), ce qui confère un caractère plutôt ludique. L'utilisateur peut néanmoins donner des réponses erronées[296], qui sont tout de même comptabilisées par *Facebook*. L'entreprise valide ou non l'information, en croisant probablement les modifications fournies par d'autres utilisateurs. Si *Facebook* estime que l'information est fausse, l'utilisateur perd des points. Chaque contributeur se voit alors définir un niveau de confiance, que *Facebook* prend probablement en compte dans le processus de validation de l'information. Un onglet « Communauté » permet aussi de voir les niveaux atteints et le nombre de points chez ces contacts, ce qui peut peut-être susciter chez certaines personnes enclines à la compétition une motivation supplémentaire. Les contributeurs les plus actifs par région géographique voient par défaut leur pseudonyme et leur score apparaître publiquement.

En somme, le système de *Facebook Places* a tous les attributs des Informations Géographiques Volontaires (*VGI*, chapitre 4), puisqu'un utilisateur réalise un *check-in apriori* en toute conscience, qu'il peut créer un lieu, et participer gratuitement à la correction de la base de données[297] en étant récompensé par des badges, ou par la satisfaction d'avoir plus contribué que les membres de sa communauté. Néanmoins, le signalement de présence dans

296. Fournir volontairement des informations erronées est interdis par les conditions d'utilisations de *Facebook*. https://www.facebook.com/ ega /terms
297. D'après les conditions d'utilisation : "L'utilisation de *Facebook Editor* et de tout groupe connexe est purement volontaire", "vous reconnaissez le faire pour votre propre loisir et divertissement".

un lieu, ou la création et/ou la correction de la base de données se fait gratuitement au profit d'une entreprise qui malgré une « mission » affichée teintée de philanthropie[298], a un modèle économique extrêmement lucratif basé sur le profilage, la vente d'espaces publicitaires ou d'informations, notamment des données personnelles et comportementales collectées auprès de ces utilisateurs. Si quelques chercheurs ont montré que les contributeurs de la base de données géographique libre et gratuite *OpenStreetMap* sont principalement motivés par le désir de participer à la création d'un bien commun (Duféal et Noucher, 2017), il serait intéressant de mener des études similaires sur les personnes qui contribuent largement à l'amélioration d'une base de données géographique d'une entreprise privée, qui s'enrichit et monétise leur travail bénévole.

Comme pour toutes *VGI*, il n'est pas évident de définir le niveau de véracité des *check-in* de *Facebook*, car une personne peut renseigner sa présence dans un lieu, parfois mal localisé ou catégorisé, sans y être nécessairement au moment où elle partage cette information avec ces contacts. Néanmoins, nous pouvons poser l'hypothèse que la plupart des personnes qui effectuent des *check-in* sont réellement allées dans le lieu, que l'instantanéité qui caractérise le partage de l'information sur les réseaux sociaux implique un décalage temporel relativement faible, et qu'un grand volume de données agrégées peut tendre vers un profil de fréquentation horaire quotidien moyen relativement proche de la réalité. La section suivante détaillera la collecte et le volume des données, ainsi qu'une rapide méthode permettant de filtrer les données, notamment en supprimant des pics de fréquentation irréalistes.

2.2 Collecte des données

La collecte des informations sur les différents lieux de la base de données de *Facebook* passe par l'*API Places Search*[299]. Après avoir créé un compte et obtenu une clé[300], il suffit alors de renseigner des coordonnées géographiques et un rayon de recherche pour obtenir certaines informations sur au maximum 100 lieux dans le périmètre géographique défini. La figure 80 est un exemple de résultat de requête réalisée aux alentours de l'Institut National des Sciences Appliquées (INSA) de Rouen en novembre 2017. En plus des informations générales sur le type d'activité qui s'exerce dans ce lieu, l'adresse et les coordonnées GPS, 3429 *check-in* (*checkins*) et 5174 et *likes* (*fan_count*) y était comptabilisés au moment de la requête, depuis la création de la page.

298. https://fr.newsroom.fb.com/company nfo/ , *notre mission* : "*Facebook* vise à favoriser le partage entre les gens et à rendre le monde plus ouvert et connecté. Les utilisateurs de *Facebook* souhaitent rester en contact avec leurs amis et leur famille, savoir ce qui se passe dans le monde, partager et exprimer ce qui leur tient à cœur".
299. https://deve opers.facebook.com/docs/p aces/web/search
300. https://deve opers.facebook.com/docs/facebook og n/access tokens? oca e=fr FR

checkins: 3429
category_list:
0:
id: "2602"
name: "College & University"
1:
id: "2601"
name: "School"
2:
id: "108051929285833"
name: "College & University"
location:
city: "Saint-Étienne-du-Rouvray"
country: "France"
latitude: 49.385049297931
longitude: 1.0680389754214
street: "Avenue de l'Université"
zip: "76800"
name: "INSA Rouen Normandie"
fan_count: 5174
id: "185751271450722"

FIGURE 80 Exemple de résultat d'une requête sur *Facebook Places Search*.

Connaissant le nombre de *check-in* à l'instant *t*, il suffit de relancer une requête sur le même lieu à l'instant *t+1* puis d'effectuer une soustraction pour connaître le nombre de personnes ayant fait un *check-in* durant la période. Plus l'intervalle de temps est court, plus précis sera le profil de fréquentation temporel du lieu.

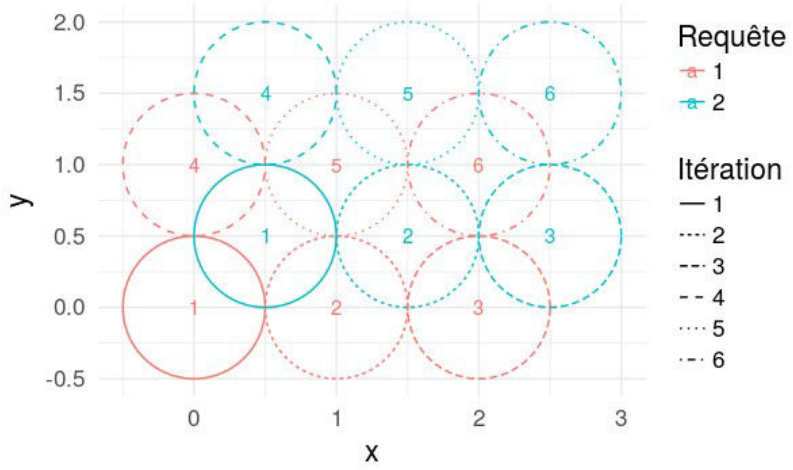


FIGURE 81 Principe de fonctionnement de la fenêtre mobile permettant de collecter les données *Facebook*.

Nous avons écrit un code en python qui permet de faire des requêtes selon une fenêtre mobile, afin de couvrir la zone d'étude de Bangkok longitudes comprises entre 100,1° et 100,9° et latitudes entre 13,45° et 14°. Si nous avons également couvert la zone de Delhi, seuls les résultats de Bangkok figureront dans ce travail, pour des raisons de temps impartis. L'objectif étant de passer plusieurs fois sur la même zone avec l'intervalle de temps le plus faible possible, nous avons lancé simultanément deux requêtes de rayon de recherche d'un kilomètre, l'une ayant un point de départ décalé de 500 m en longitude et en latitude. Une fois que

l'ensemble des lieux ont été récoltés dans une zone donnée, la fenêtre se déplace d'un kilomètre vers l'est - puis d'un kilomètre vers le nord lorsque la ligne est complétée. La figure 81 illustre le principe de fonctionnement de notre fenêtre mobile, où nos deux requêtes simultanées et décalées spatialement permettent de couvrir toute la zone, et donc d'enregistrer la totalité des lieux de la base. L'*API* n'est pas restrictive quant au nombre de requêtes, ce qui permet de lancer l'application en boucle. Néanmoins, certaines coupures de connexions sont survenues, interrompant ponctuellement la collecte des données.

Les requêtes renvoient au maximum 100 résultats par pages, mais notre programme permet d'accéder aux résultats de toutes les pages. Enregistrer l'ensemble des données sur la zone d'étude mettait environ 1h10. Nous avons appliqué cet algorithme pour les mois de juillet et août 2017. Cette première méthode entraîne l'enregistrement de doublons, car des lieux peuvent être communs à plusieurs fenêtres de recherches, ce qui augmente la durée de balayage total tout en enregistrant un nombre inutilement important d'informations sur les lieux. Nous avons donc changé de méthode et utilisé tous les identifiants enregistrés dans la zone d'étude ("id"), et fait une recherche sur ces derniers. Ceci permet de limiter le nombre de requêtes tout en accélérant le temps de retour sur un même lieu (~15-20min).

Environ 150 Go de données ont ainsi été récupérés entre juillet et novembre 2017[301]. Nous avons géré les doublons (des nombres de *check-in* différents par identifiant et par tranche horaire) en prenant pour chaque identifiant le nombre maximum de *check-in* par tranche horaire, puis nous avons soustrait au nombre de *check-in* enregistrés à *t* celui enregistré à *t+1*. Au final, prés de 50 millions de *check-in* (47 874 025) ont été enregistré dans 153 771 lieux, au cours de 108 jours d'enregistrement. Alors que 2 millions de *check-in* étaient mondialement créés quotidiennement en 2010 (Chang et Sun, 2011), le nombre de *check-in* quotidiens moyen était ainsi de 443 278 à Bangkok en 2017, ce qui suggère une généralisation de l'utilisation de l'outil par les utilisateurs de *Facebook*.

La figure 82.a présente le nombre de lieux selon le nombre de *check-in* enregistrés. Nous retrouvons encore une fois une loi de puissance à longue queue, où l'écrasante majorité des lieux n'a enregistré que très peu de *check-in*, tandis qu'un nombre relativement restreint en enregistre énormément. La figure 82.b montre le nombre de jours où des lieux ont enregistré des *check-in*, ce qui est indicateur de régularité et peut être vu comme le corollaire de la figure précédente, car la plupart des lieux n'enregistrent des traces numériques synonymes de visites que très rarement.

301. Les compiler fut un défi technique pour notre ordinateur personnel classique dotée de 12 giga de RAM.

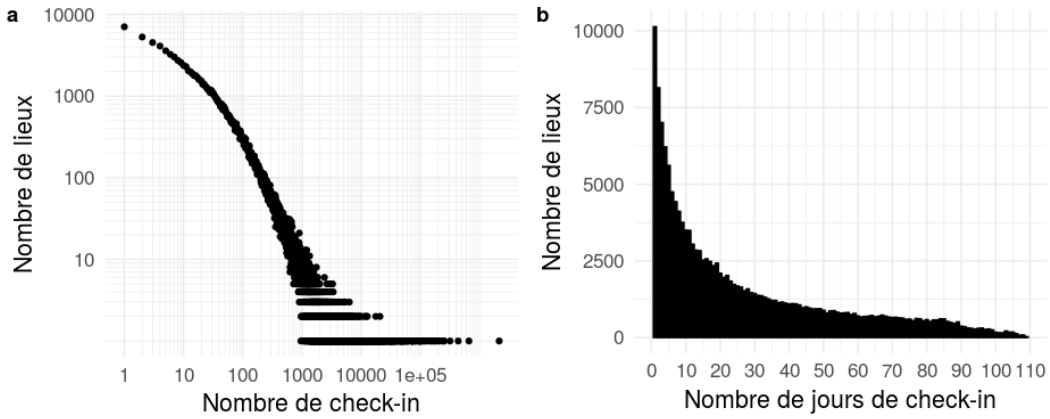


FIGURE 82 Distribution du nombre de *check-in* (a) et du nombre de jours différents de check-in (b) par lieu.

Le nombre de *check-in* envoyés par tranche horaire sur l'ensemble de la période d'enregistrement (figure 83) montre d'abord le caractère circadien de l'activité sur *Facebook*, soit trivialement que l'activité nocturne est moindre que l'activité diurne. On observe aussi généralement un plus grand nombre de *check-in* émis les jours de week-end - mis en évidence par la moyenne mobile (ligne orange). Nous pouvons aussi noter une tendance décroissante dans le nombre de *check-in* enregistrés. Plusieurs hypothèses explicatives peuvent être posées, tout d'abord l'aspect saisonnier où plus de personnes, notamment des touristes, peuvent se rendre à Bangkok pendant la période estivale et faire savoir à leurs contacts les lieux qu'ils fréquentent par l'intermédiaire de *check-in*, ce qui peut faire office de carnet de voyage. Un autre aspect, plus d'ordre technique, serait lié au changement de notre méthode de captation des *check-in* opérée à partir du mois de septembre. En effet, à partir de cette période, nous ne nous focalisons que sur les identifiants de lieux déjà répertoriés, ce qui implique que nous n'enregistrons plus les *check-in* dans des lieux nouvellement créés.

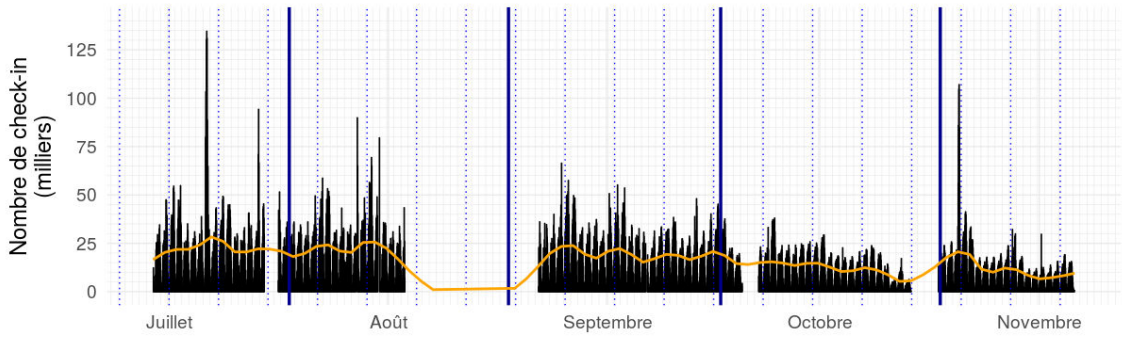


FIGURE 83 Nombre de *check-in* enregistrés par tranche horaire. Les lignes verticales continues bleues indiquent le début de chaque mois, celles en pointillé les vendredis à minuit. La ligne orange est une moyenne mobile calculée selon la méthode de LOESS.

Il est également possible que certains lieux soient supprimés par *Facebook*, voire que certains identifiants soient fusionnés, et leurs *check-in* agrégés (figure 84). Ce dernier aspect pourrait expliquer les quelques pics proéminents que nous observons les 20 et 27 juillet, les 10 et 13 août, et le 3 novembre. Car si certains événements ponctuels peuvent entraîner localement une plus grande fréquentation, cet argument ne peut suffire à justifier les 640 000 *check-in* enregistrés au palais des congrès de Bangkok le 20 juillet 2017.

FIGURE 84 *Facebook* et la gestion des lieux doublons. Lorsque l'entreprise estime que deux lieux sont identiques, ils sont alors fusionnés.

Ces pics pourraient également avoir pour origine des achats de *check-in*. Prosaïquement, cela signifierait que les gérants de pages *Facebook* désireux d'augmenter artificiellement le rayonnement de leur activité et leur réputation en ligne auraient recours à des compagnies externes, qui part divers intermèdes incitent des comptes *Facebook* (dont certains peuvent s'apparenter à des bots) à faire des *check-in* dans le lieu en question. Ces pics pourraient donc aussi provenir de l'impact de « fermes à clic », phénomène assez répandu en Thaïlande[302]. Tous ces éléments nous poussent donc à effectuer des filtres.

2.3 Filtres

2.3.1 suppressions de lieux

Avant d'effectuer des traitements sur les pics enregistrés, il convient tout d'abord de sélectionner les lieux selon leur pertinence pour notre étude. En effet, certains lieux ne font pas nécessairement référence à des localisations précise et ponctuelle. Nous pensons par exemple aux catégories de type « Quartier », « Village », « Ville », « District », « Province » ou « Pays », qui représentent des entités géographiques plus vastes que de commerces ou des universités, et qui ne sont pas contextualisable en termes d'activités effectuées. Nous supprimons donc de notre base tous les lieux ayant au moins un de ces types renseignés dans une des trois catégories qui définissent les lieux de *Facebook*. Certaines pages *Facebook* qualifiées de lieux sont associées à des sites web ou des chaînes de télé et enregistrent un grand nombre de *check-in*. Ceci parait irréaliste, car ces lieux ne peuvent accueillir un grand nombre de personnes, et il est possible qu'il y ait une confusion de la part de l'utilisateur de *Facebook*, et le *check-in* peut être aussi vu

302. https://www.francetv nfo.fr/econom e/emp o /recherche d emp o /emp o 2 0 et reseaux soc aux/tha ande fermeture dune ferme a c cs denvergure 2235997.htm

comme une autre forme d'adhésion comparable au *like*. Il conviendrait d'étudier plus en détail chacune des catégories renseignées et les profils de fréquentations des lieux correspondants. Nous présenterons peut-être ce travail ultérieurement, mais nous décidons dans un premier temps de ne pas prendre en compte les lieux catégorisés comme "website", "tv channel" et "just for fun".

Ce premier filtre sur les activités entraîne la suppression de 3645 lieux correspondant à 5 440 034 *check-in*. Les lieux ayant au moins une catégorie renseignée comme « Ville » représentent presque 3 millions de *check-in* (2 998 208) et ceux contenant « Region » plus d'un million (1 034 089). Nous décidons également de supprimer les lieux les moins fréquentés, pouvant être créés à des fins ludiques (voir plus haut) et qui ont enregistré moins de 5 *check-in* sur moins de deux jours sur la période, soit 21 949 lieux pour 55 829 *check-in*. Il nous reste au final 128 467 lieux cumulant 42 378 802 *check-in*.

2.3.2 suppressions des pics

Ponctuellement, certains lieux présentent des pics de fréquentations relativement inhabituels. Si nous ne pouvons poser des hypothèses d'ordre technique sur l'apparition de ces derniers fusion de certains lieux, achat de *check-in*, ou plus localement une personne qui fait un *check-in* en mentionnant l'ensemble de ces contacts ils peuvent également avoir pour origine des événements assez exceptionnels. Quoi qu'il en soit, un de nos objectifs est plus de travailler sur les temporalités de fréquentation les plus routinières, en mettant de côté les éventuels aléas.

Nous allons donc, dans un premier temps, selon une logique d'optimisation des temps de traitements, focaliser notre travail sur les lieux qui présentent de grandes variations dans leur nombre de *check-in* enregistrés par tranche horaire. Nous calculons pour chaque lieu le coefficient de variation, soit le rapport entre l'écart type et le nombre de *check-in* moyen enregistré, ainsi que le rapport entre le nombre de *check-in* maximum et le nombre moyen de *check-in*. Nous sélectionnons les lieux ayant respectivement un coefficient de variation supérieur à 1 et une valeur maximale supérieure à 8 fois la valeur moyenne. 17 751 lieux (13,8 %) pour 25,8 millions de check-in (61 %) remplissent ces critères. Nous utilisons ensuite la fonction "*despike*" du *package* "*oce*" (Kelley et Richards, 2017), développé à la base pour débruiter les données de relevées océaniques. L'algorithme fonctionne en calculant pour chaque lieu une série temporelle de référence, selon l'algorithme "smooth", qui définit automatiquement une fenêtre de lissage (Tukey, 1977). Si la valeur à l'instant t est supérieure à 4 fois la valeur de référence, la valeur en question se voit attribuer la valeur de référence. Nous recalculons ensuite l'écart type de la nouvelle série. Si ce dernier est supérieur à 50, nous appliquons à nouveau l'algorithme, en définissant la série temporelle de référence par interpolation linéaire entre les valeurs comprises entre 0 et 5 fois l'écart type (voir Kelley et Richards (2017) pour

une explication détaillée de l'algorithme *despike*, avec les fonctions "smooth" et "trim"). Ces différents seuils ont été choisis de manière empirique, après un grand nombre d'essais. Il est en effet assez délicat de trouver les paramètres optimaux compte tenu des fortes différences entre les lieux en termes de fréquentation.

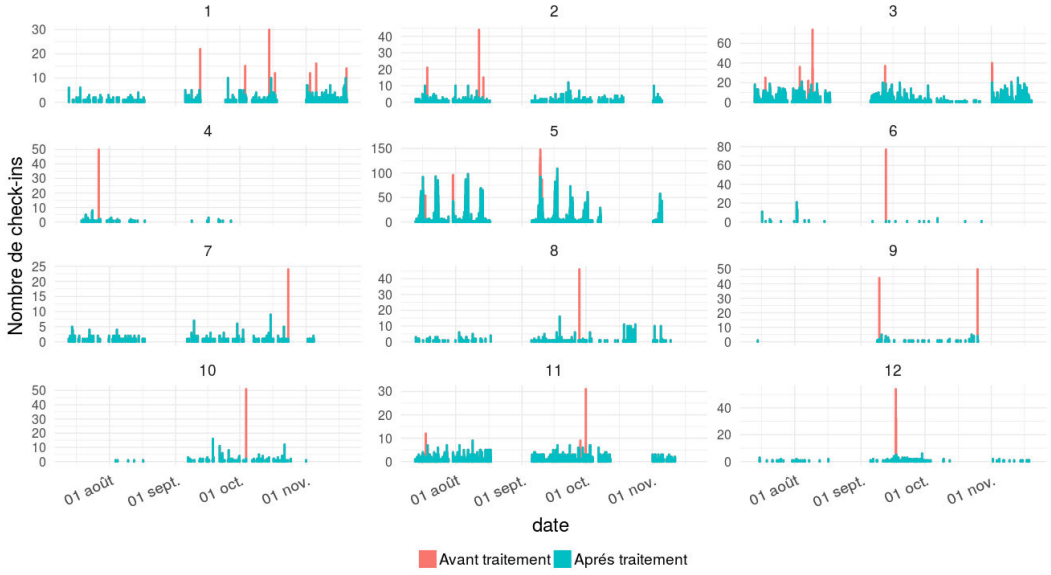


FIGURE 85 Résultat de notre algorithme de suppression de pics. Exemple dans 12 lieux où apparait le nombre de *check-in* par tranche horaire avant (en rose) et après traitement (en cyan).

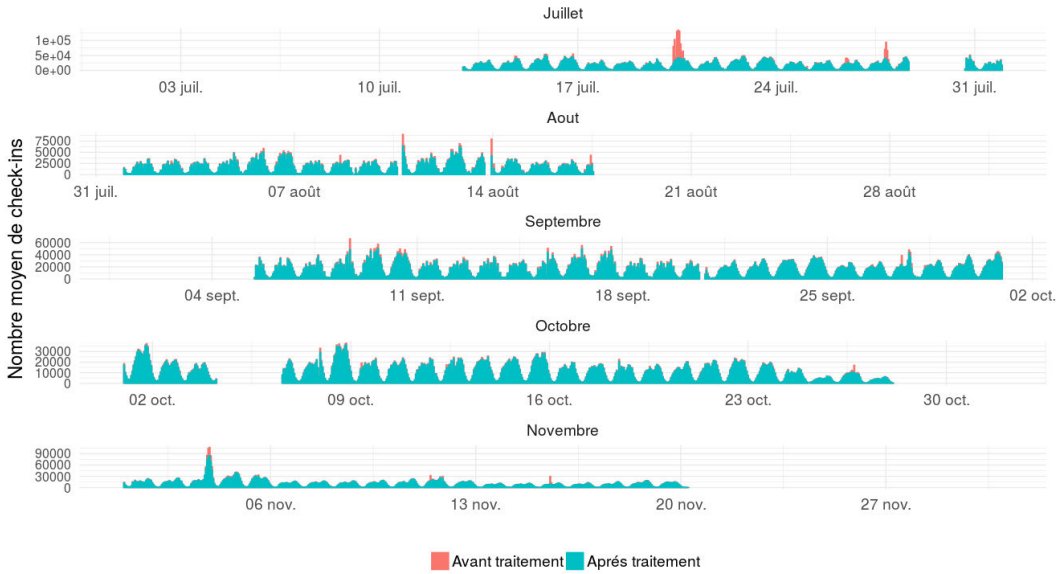


FIGURE 86 Résultat de notre algorithme de suppression de pics sur l'ensemble des données. Apparait le nombre de *check-in* par tranche horaire avant (en rose) et après traitement (en cyan).

Si nous regardons le résultat de notre algorithme dans 12 lieux (figure 85), nous pouvons noter que la plupart des pics sont bel et bien supprimés. Si globalement ces seuils semblent adaptés, il arrive néanmoins, comme pour le lieu 5, que certains pics de fréquentations soient rabotés inutilement, sans pour autant changer la tendance. Nous observons que notre méthode a d'un point de vue agrégé (figure 86) entraîné un lissage des pics aberrants même si le pic du 3 novembre n'est pas aussi abrasé que celui du 20 juillet. Nous conservons au final un peu moins de 40 millions de *check-in* (39 208 042), toujours dans 128 467 lieux.

2.4 premiers résultats

Ces données filtrées sont géolocalisées et datées. Nous présentons ici quelques résultats sur leurs potentiels d'utilisation davantage de traitements seront effectués dans la partie D.

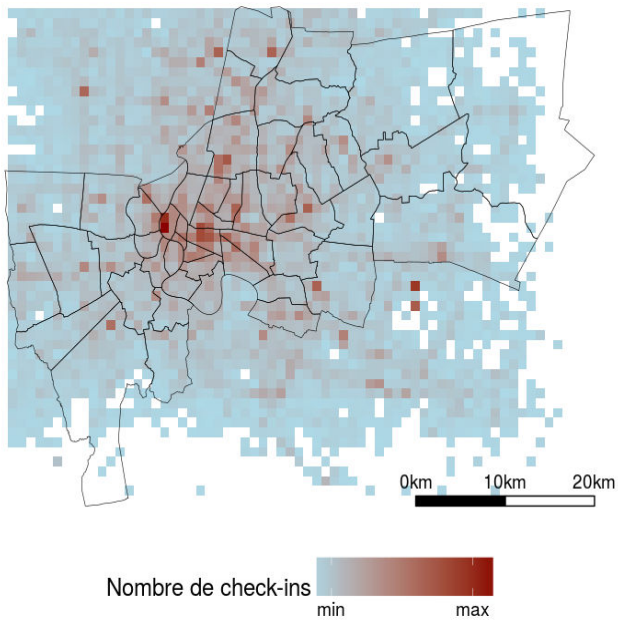


FIGURE 87 Répartition des *check-in* à Bangkok - agrégés dans des mailles d'un kilomètre.

Tout d'abord, nous pouvons voir globalement quels secteurs de la ville ont enregistré le plus de *check-in* (figure 87). Cette répartition n'est pas homogène dans l'espace car plus concentrée dans le centre-ville. Nous pouvons noter la présence de clusters de taille secondaire éparpillés dans les zones périphériques de la ville. Ceci suggère une forte attractivité du centre-ville, avec néanmoins des petits pôles locaux, susceptibles d'attirer un grand nombre de personnes.

La figure 88 présente le nombre moyen de *check-in* enregistrés par tranche horaire à Bangkok avant et après traitement. Nous pouvons noter que les niveaux d'activités sont

similaires du lundi au vendredi, avec cependant un plus grand nombre d'enregistrements le vendredi soir. L'activité sur le service atteint un premier palier vers 9 h, puis apparaît un pic en début d'après-midi, et un second aux alentours de 20 h. Plus de *check-in* sont enregistrés les jours de week-end, notamment les après-midis. Si les pics ne sont pas marqués de la même manière que pour les données de *Twitter* (figure 67), les tendances générales sont relativement similaires.

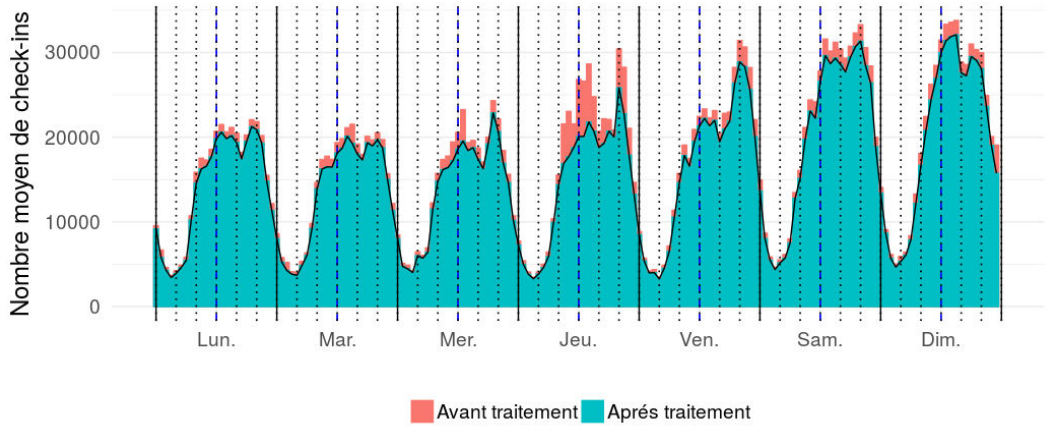


FIGURE 88 Nombre moyen de *check-in* effectué par tranche horaire sur une semaine. Les lignes en pointillé symbolisent un pas de temps de 4 h. La ligne bleue représente midi tandis que la ligne noire minuit.

Les données de *Facebook* sont aussi associées à un type de lieu et revêtent alors un intérêt non négligeable surtout lorsque l'on utilise le concept d'espace d'activité. 1297 catégories étaient renseignées à Bangkok. Nous avons regroupé les occurrences les plus fréquentes dans de nouvelles catégories, plus restreintes sémantiquement. Par exemple tous les lieux contenant « Restaurant » (Thai Restaurant, Chinese Restaurant, etc.) furent renommés « Restaurant ». Tous les petits commerces (Mobile phone shop, clothings, etc.) sont maintenant labellisés « Shopping and retail » (voir annexe F pour plus de détails).

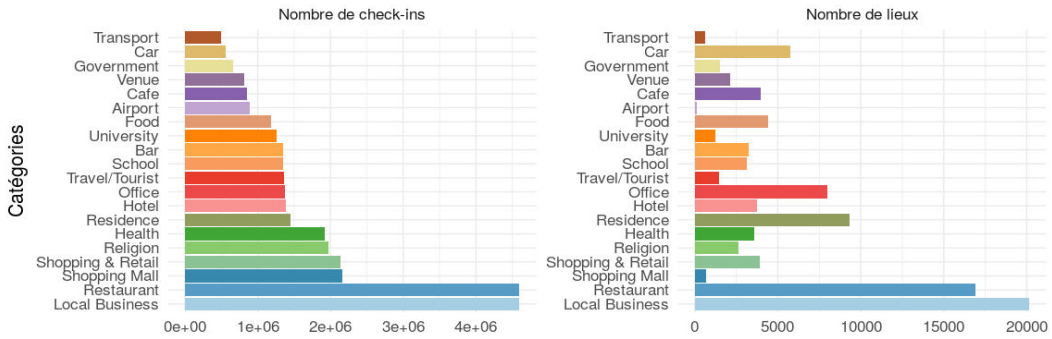


FIGURE 89 Nombre de lieux et de *check-in* en fonction des catégories.

La figure 89 montre d'une part le nombre de *check-in* enregistrés selon les catégories et d'autre part le nombre de lieux associés. Nous pouvons noter que les lieux de types *malls*, Tourist/Travel (qui regroupent les lieux touristiques et des agences de voyages), *University,* ou encore *Religion* (qui regroupent les temples et les lieux de cultes) enregistrent énormément de *check-in* par rapport aux nombres de lieux qu'ils représentent. D'autres catégories, comme *local business*, *residence*, ou *office* sont associées à un grand nombre de lieux, mais aussi à un nombre global de *check-in* assez important, ce qui suppose qu'individuellement, peu de *check-in* sont enregistrés dans chacun de ces lieux.

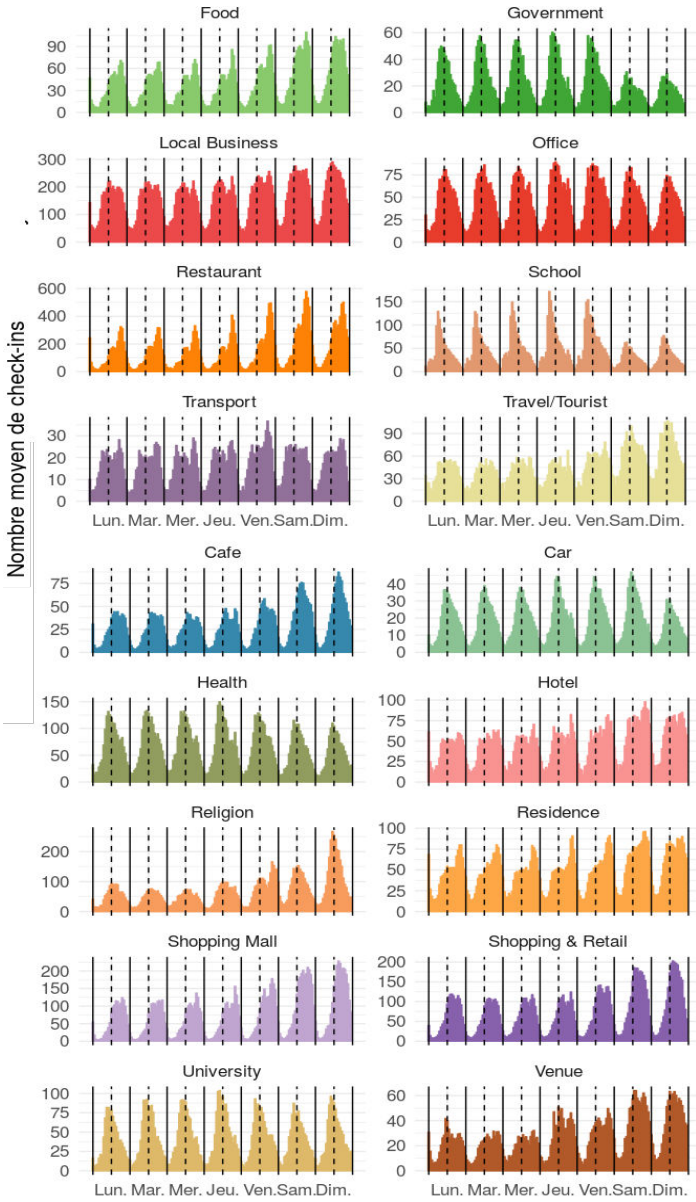


FIGURE 90 Profils temporels sur une semaine des 20 catégories les plus visitées sur *Facebook*.

Il est également possible de faire ressortir les profils de fréquentation horaire de chacune des catégories (figure 90). Ainsi, les bars sont surtout fréquentés le soir, et principalement le vendredi et le samedi, un peu comme les restaurants ou les lieux de type *food* (lieux associés à des commerces de nourritures, non-labellisés contrairement à *restaurant*). Les cafés sont surtout fréquentés l'après-midi et d'autant plus le week-end. Les lieux de type résidences présentent un pic les soirs de semaines. Les écoles ou les administrations (*governement*) présentent un pic marqué le matin des jours de semaine, absent le week-end. Les *malls* ou les commerces (*shopping & retail*) sont très fréquentés, mais surtout le week-end. Les transports (gares, autoroutes, arrêts de métro ou de bus) présentent deux pics de fréquentations correspondants aux heures de pointe, etc. En somme, les données issues de *Facebook* confirment des intuitions quant aux horaires de fréquentation des activités, mais en quantifiant ces dernières. Malgré les biais liés au fait qu'une personne n'est pas obligée de se trouver dans un lieu pour effectuer un *check-in*, ces profils temporels montrent que cet aspect n'est probablement pas une limite à l'analyse des activités dès lors que les données sont agrégées.

Synthèse

Ce chapitre a présenté les données *Twitter* et *Facebook* que nous utiliserons par la suite. Nous avons ici insisté sur les protocoles de collecte et les prétraitements qui nous paraissent indispensables pour réduire les biais lors d'analyses futures.

Les données *Twitter* sont individuelles et permettent la création d'espaces d'activités et il est possible d'estimer la localisation des domiciles des utilisateurs. Nous avons tenté, autant que faire se peut d'avoir une approche qui préserve au mieux l'anonymat des personnes de notre échantillon.

Les données que nous avons récoltées sur *Facebook* sont quant à elles agrégées, ce qui empêche le traçage des utilisateurs. Elles donnent aussi directement des informations sur les horaires de fréquentation des lieux. Ces sources d'informations sont donc très complémentaires, et nous présenterons dans les prochains chapitres des méthodes permettant d'estimer les activités réalisées dans les espaces d'activités de l'échantillon issus de *Twitter*.

Si les données *Twitter* à Delhi ne semblent pas autoriser d'études potentiellement concluantes permettant la compréhension des déterminants des mobilités dans la capitale indienne, la partie C mobilisera des données de terrain ce qui permettra d'évaluer l'apport et les limites de chacune des approches. Nous reviendrons sur l'hybridation des données de *Twitter* et de *Facebook* dans la partie D, dédiée à Bangkok.

Partie C:

Approche hybride des mobilités à Delhi

Dans le grand nombre de travaux cherchant à analyser et/ou modéliser les mobilités quotidiennes à l'aide de traces numériques (voir chapitre 5), la plupart des processus de validation passent par l'utilisation d'enquêtes ménages déplacements ou de données de recensements (*e.g.* Calabrese *et al.,* 2011 ; Gao *et al.,* 2014 ; Tizzoni *et al.,* 2014), ou bien par d'autres sources de traces numériques (*e.g.* (Lenormand *et al.,* 2014)). Loin de remettre en cause la pertinence de ces approches, nous pouvons tout de même en souligner leur caractère « 100 % *ex situ* », dans le sens où aucun terrain n'est effectué par les membres de l'équipe[303]. Les durées et coûts importants des enquêtes institutionnelles, pour avoir simplement une vision « instantanée » des mobilités sont très souvent perçus comme des limites[304], que les traces numériques sont censées dépasser. Il s'agit bien souvent de montrer que des données quantitatives *ex situ* peuvent se substituer aux données quantitatives *in situ* pour analyser des flux de mobilités et des tendances de déplacements.

Si le caractère ubiquiste des données téléphoniques est peu remis en cause[305], se pose la question de l'échelle géographique qui dépend de la densité du réseau d'antenne relais (de l'ordre de la centaine de mètres en zone urbaine, voir chapitre 3). Cette dernière, bien que très dense en milieu urbain et *a priori* adapté à une analyse à l'échelle d'une métropole ne permet pas d'avoir une résolution spatiale suffisamment fine pour apprécier les mobilités des habitants dans un périmètre réduit à un quartier (Perkins *et al.,* 2014). Or, la contamination locale joue un rôle important dans la propagation des épidémies, notamment de dengue (Stoddard *et al.,* 2013,

303. Les données recensements et les enquêtes ménage déplacement sont effectuées par des institutions différentes.

304. "The process involved in the calculation of an OD matrix, from the initial data gathering to the exploitation of the first results, is lengthy and may take years to only get a snapshot of the travel demand ; the collected data has shortcomings both in terms of spatial and temporal scale" (Calabrese *et al.,* 2011a)

305. Tout du moins dans les pays développés, et même si les données utilisées sont en général celles d'un seul opérateur et qu'il peut y avoir des différences sociologiques plus ou moins marquées entre les clients des différents opérateurs, notamment en fonction des tarifs proposés.

2009 ; Telle *et al.,* 2016). À titre de comparaison, les traces numériques géolocalisées issues des réseaux sociaux, notamment *Twitter*, offrent un niveau de précision spatial de l'ordre de la dizaine de mètres, avec cependant de grandes inconnues sur la représentativité de l'échantillon et du niveau de correspondance entre les traces numériques laissées par les utilisateurs et leur espace d'activité réel. *Twitter* n'est en effet pas utilisé de la même manière en fonction des tranches d'âge (Longley *et al.,* 2015 ; Sloan *et al.,* 2015a), des groupes sociaux (Luo *et al.,* 2016) et des régions, et l'intention derrière une trace numérique géolocalisée dans un lieu donné est propre à chaque individu et médium (Lenormand *et al.,* 2016b).

De plus, ces approches purement quantitatives ex-situ, bien qu'ayant largement révolutionné l'analyse des mobilités quotidiennes (voir chapitre 5), sont relativement déconnectées de certaines réalités et pratiques sociales, dans le sens où un individu est simplement défini par une succession de positions inscrites dans le temps et l'espace, avec parfois la notion d'activité réalisée, mais sans plus d'informations sur la personne[306], sur ces motivations ou sur sa perception de la ville et de son environnement.

Certaines études allient traces numériques et entretiens individuels, notamment en fournissant des GPS aux quelques personnes échantillonnées (e.g.Chevalier, 2018a, 2018b ; Drevon *et al.,* 2014 ; Feildel, 2014 ; Vazquez-Prokopec *et al.,* 2009). Ceci permet de récolter des informations sur les mobilités, mais aussi de contextualiser les déplacements et d'évaluer les biais entre les lieux déclarés être fréquentés et les lieux réellement visités. Mais peu d'études ont à notre connaissance essayé de comparer des tendances et profils de mobilités obtenus à partir des traces numériques issues des réseaux sociaux en lignes à des données collectées *in situ* via des enquêtes de terrain qualitatives.

Analyser les mobilités avec des données issues de *Twitter*, dans une ville où ce service est peu utilisé (chapitre 6) ne semble pas à première vue des plus pertinents. Mais compléter ces données avec une enquête de terrain sur un petit échantillon d'habitants d'un quartier peut procurer des éléments de contextualisation, notamment à l'échelle d'un quartier. Ainsi, le fait de combiner ces deux sources de données selon une méthode mixte[307] (ou hybride) confère des avantages majeurs, comme la comparaison des résultats et la validation / infirmation des hypothèses de départ ou encore l'élaboration de nouvelles hypothèses et la préconisation de nouveaux protocoles de collecte (Rossman et Wilson, 1985). Combiner des données quantitatives (*Twitter*) et qualitatives (entretien de terrain) collectées indépendamment, devrait théoriquement permettre d'interpréter et de mieux comprendre une situation (Condomines et Hennequin, 2013)

306. Nous pouvons noter néanmoins l'étude de (Lenormand *et al.,* 2015b) qui utilise des données bancaires comme proxy pour les mobilités en connaissant le capital économique de chacun.
307. "Mixed methods research is empirical research that involves the collection and analysis of both qualitative and quantitative data" (Punch, 2014).

Delhi est une mégapole qui, comme beaucoup de grandes villes de pays en développement, affiche des niveaux d'inégalité socio-économiques très élevés et très marqués spatialement. En effet, des bidonvilles ont tendance à se développer dans les zones non construites à proximité de quartiers ou « colonies » plus ou moins planifiées, ce qui implique des niveaux de ségrégation socio-économiques très forts, mais pas forcément de barrière dans les déplacements. Nous mobiliserons dans ce chapitre le concept de motilité (Kaufmann et Jemelin, 2004), et nous essayerons de voir dans quelles mesures le capital économique influence les potentiels de mobilités. Combiner les données *Twitter* à des enquêtes de terrains devrait aussi permettre d'un point de vue méthodologique de montrer les limites et avantages de chacun des corpus[308], d'un point de vue conceptuel de valider ou nuancer l'hypothèse d'une mobilité influencée par le capital économique.

Nous présenterons dans le chapitre 7 notre enquête de terrain réalisée dans un quartier du sud de Delhi. L'échantillon et la zone d'étude seront largement décrits, et nous évaluerons notamment si le quartier du domicile est un bon indicateur des caractéristiques socio-économique des habitants. Nous analyserons ensuite différentes métriques de mobilités ou des types de lieux fréquentés en fonction des couches sociales. Nous mettrons en avant des lieux où les niveaux de coprésences et de mixité sociales sont assez importants, ce qui peut avoir des répercussions sur la propagation de maladie vectorielles (Daudé et Eliot, 2005).

Les questionnaires de mobilités seront ensuite transcrits sous forme d'agendas spatialisés, où chaque personne échantillonnée se verra attribuer une séquence d'activité qu'elle effectue probablement à un moment de la journée dans un lieu donné. Cette méthode permet une implémentation relativement aisée dans un système à base d'agents et permet de formuler des préconisations méthodologiques, notamment sur les informations à collecter ou à estimer.

Après avoir évalué le niveau de représentativité sociale des utilisateurs de *Twitter* à Delhi, nous définirons une activité probablement réalisée dans chaque lieu de l'espace d'activité de l'échantillon. Pour ce faire, nous créerons tout d'abord une couche d'utilisation du sol en mobilisant des données issues de différents services cartographiques. Nous proposerons ensuite une méthode permettant de passer d'espaces d'activités discrets dans le temps à des agendas continus temporellement. Finalement, nous testerons et discuterons d'une première méthode permettant la génération d'agendas pour des individus de synthèses, en nous basant sur les agendas reconstitués issus de l'échantillon *Twitter* et des données de terrain.

308. Comme souligné par (Winter, 1983) à propos d'enquêtes qualitatives et quantitatives : « au delà de l'opposition factice entre qualitatif et quantitatif [...], il s'agit de promouvoir des systèmes d'investigation dans lesquels chaque mode d'approche, chaque type d'investigation, garde sa spécificité, mais valide l'autre »

Chapitre VII: Des enquêtes de terrain pour appréhender les mobilités à Delhi et poser les bases d'un modèle à base d'agents

Nous présenterons dans ce chapitre les résultats d'une enquête de terrain réalisée dans le sud de la ville. Compte tenu de la taille et de la population de la ville, la réalisation d'une enquête de terrain pilote qui vise à apprécier qualitativement les potentiels de mobilités du plus large spectre socio-économique possible, convient d'être effectuée dans une zone où les niveaux de cohabitations semblent très élevés du fait de grande variété dans les types de quartiers et dans les niveaux de taxes foncières, tout en présentant des profils de mobilités relativement proches des tendances moyennes. C'est ainsi que notre choix s'est posé sur la zone de Malviya Nargar, dans le sud de Delhi, qui répond à tous ces critères (chapitre 2). Il s'agit également du secteur où a été expérimenté le modèle de déplacement du moustique *MOMA* (Maneerat et Daude, 2017), autre versant du projet MO³ dans lequel cette thèse s'inscrit.

1 Malviya Nagar et ses alentours, un bon laboratoire pour aborder les mobilités à Delhi

1.1 De grandes hétérogénéités socio-économiques et spatiales...

Le secteur de Malviya Nagar se trouve donc au sud de Delhi, dans un environnement relativement aisé par rapport au nord ou l'est de la ville. Néanmoins, les discontinuités socio-économiques y sont très perceptibles, notamment via des ruptures très marquées dans les unités architecturales. Il suffit de traverser une route pour passer d'un quartier aisé et calme, fermé par des grilles et jalonné d'immeubles de haut standing dans un environnement verdoyant bien entretenu (Panshchila Park ou Shivalik), et arriver dans des bidonvilles aux fragiles habitations parfois colorées, tels le *slum* de Begumpur, le JJ Cluster de Lal Gumbad ou encore le camp de relocalisation Valmiki camp. Entre ces deux extrêmes se trouve la colonie planifiée de Malviya Nagar[309], puis à l'ouest Khirki Extension, une colonie qui s'est établie illégalement, sans respect des politiques d'aménagement. Au sud de la zone se trouve Hauz Rani, un village urbain très dense (figure 91). Nous décrirons plus en détail chacun de ces quartiers dans les sections suivantes, mais nous pouvons déjà noter que nous trouvons dans un périmètre restreint tous les types de quartiers de Delhi (chapitre 2) (sauf les colonies régularisées).

309. La partition de 1947 48 a vu l'arrivé à Delhi de centaines de milliers de migrants venant du Pakistan Occidental, qui ont été repartis dans des campements en zones périphériques rurales, dont faisait partie Malviya Nagar (Jha, 1988). La colonie de Malviya Nagar a ensuite été réhabilitée et certains de ces migrants ont vu leur situation être régularisée et sont devenu propriétaire du terrain qu'ils occupaient.

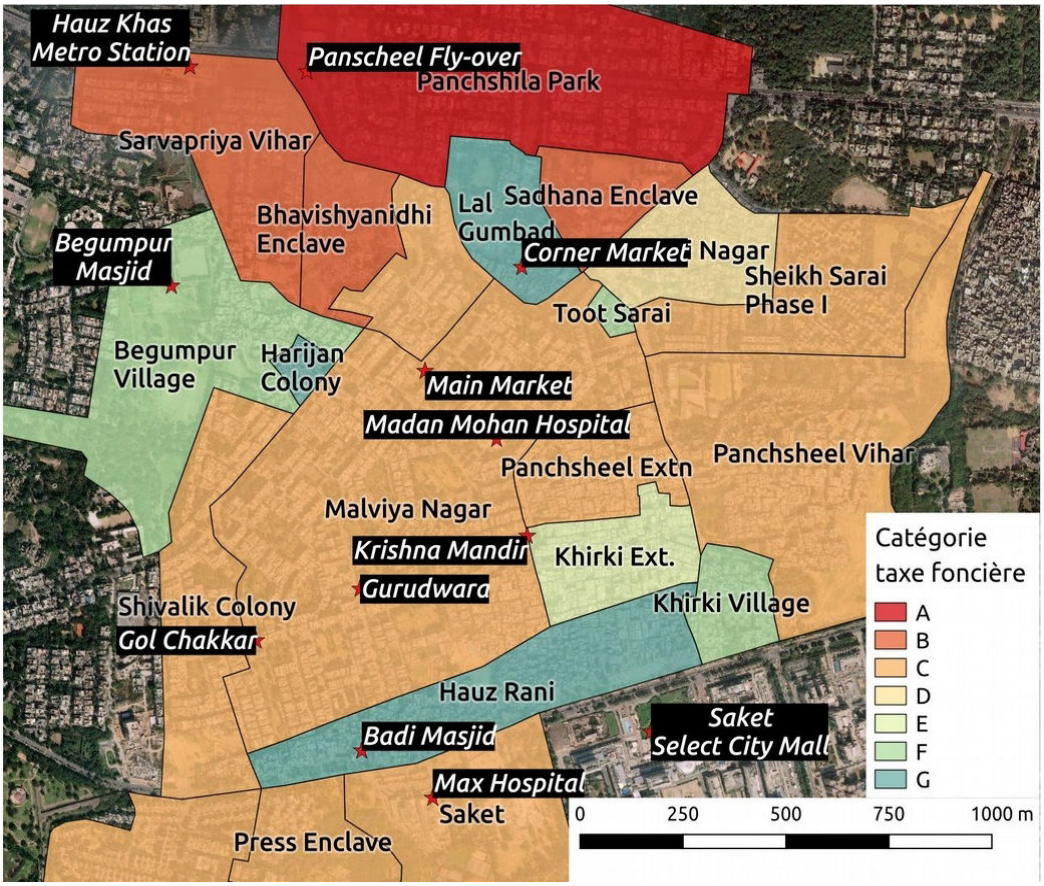


FIGURE 91 Catégorie de taxe foncière des colonies du secteur de Malviya Nagar.

La figure 91 ci-dessus reprend les valeurs de la taxe foncière par colonie, et nous pouvons noter que toutes les catégories de taxes foncières se retrouvent dans une zone de 2Km², allant des colonies où la taxe est maximale (A, Panschila Park) au *slum* de Harijan Colonie (G), en passant par les villages de Begumpur (F), Khirki Extension (E) ou encore Malviya Nagar (C). Cette hétérogénéité dans les types de quartiers et dans les niveaux d'imposition fait de la zone de Malviya Nagar un terrain d'étude *a priori* idéal pour interroger des personnes de toutes classes sociales, vivant dans un périmètre assez restreint[310], afin d'apprécier dans quelles mesures les mobilités des individus sont contraintes par leur capital économique et où et comment les différentes strates de la société sont susceptibles de se croiser.

1.2 ...matérialisé par des quartiers très différents

Nous avons choisi de découper la zone de Malviya Nagar en 8 quartiers (figure 92), sectionnés en fonction des catégories des habitations et du niveau de la taxe foncière. Cette zone d'étude couvre environ 1,5 km² et hébergerait environ 53 000 personnes d'après nos

310. Toute proportion gardée, si nous nous référons à la carte de la pauvreté à Londres de Booth, cela reviendrait à choisir un quartier tel que Nothing Hill https ://booth.lse.ac.uk/map/16/0.2100/51.5110/100/0

données de population carroyées (chapitre 2).

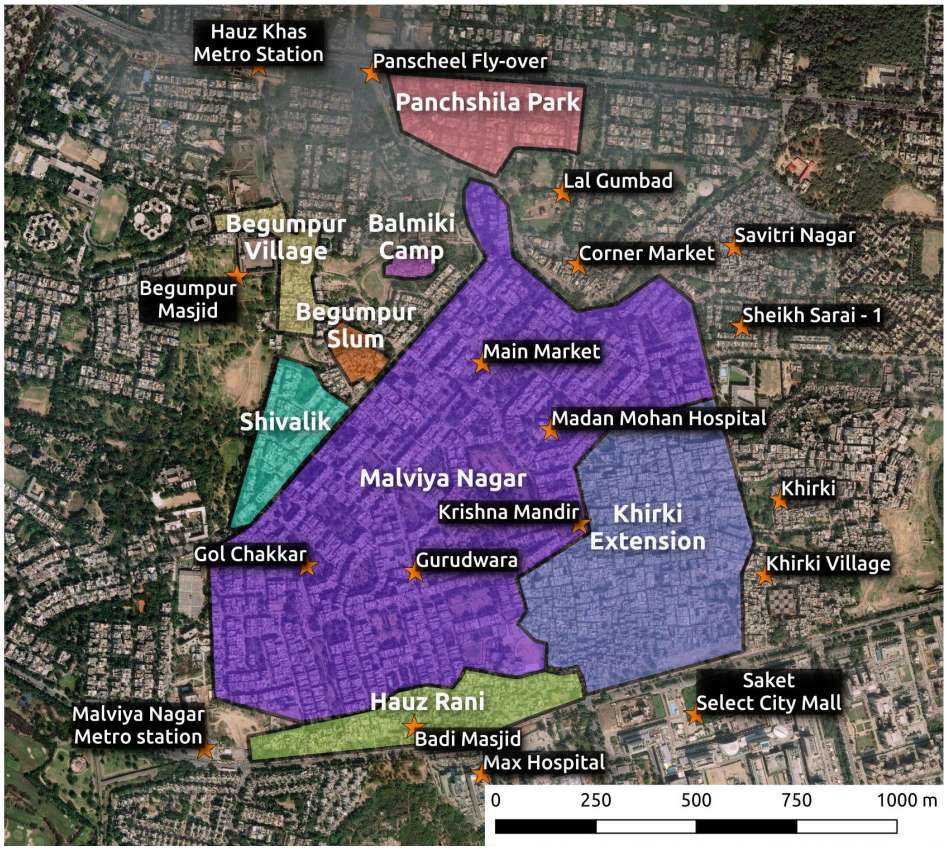


FIGURE 92 Localisation des différents quartiers de la zone d'étude.

Malviya Nagar

Malviya Nagar est le plus grand quartier de la zone. Il s'agit d'une colonie planifiée, où la taxe foncière est évaluée à C. Les rues y sont assez larges pour faire passer des voitures, et jalonnées de nombreux parcs plus ou moins bien entretenus (figure 93). Les commerces sont plutôt répartis le long des rues principales, avec une très forte concentration au niveau du marché principal, aussi nommé *Main Market*. On y trouve également de nombreux temples hindous (Laxmi & Khrishna Mandir), et une gurudwara (temple Sikh) de taille assez importante du fait d'une grande communauté Sikh, installée dans la colonie depuis la partition de l'Inde.

Au nord de Malviya Nagar se trouve le quartier très aisé de Panschilla Park. Il s'agit d'une *gated-community* principalement résidentielle, où sont présents de nombreux parcs bien entretenus et où les chauffeurs patientent à proximité des voitures de leurs employeurs.

FIGURE 93 Quelques photos de Malviya Nagar (Maneerat 2014).

Hauz Rani

Hauz Rani est un village urbain, à majorité musulmane. Les rues y sont très étroites, et les bâtiments, d'une hauteur de 3 à 5 étages présentent souvent des commerces aux rez-de-chaussée (figure 94). Nous pouvons noter une artère un peu plus large qui relie la grande avenue du sud à Malviya Nagar et qui fait office de zone commerçante.

Khirki

Au nord-est de Hauz Rani se trouve Khirki Extension 95, un quartier dont l'organisation générale est assez similaire à celle d'Hauz Rani, avec cependant des ruelles généralement un peu plus larges et rectilignes. Ce quartier est une colonie non autorisée, construite en dehors du plan de développement de la ville, sur des terrains inoccupés à l'est de Malviya Nagar. Hauz Rani et Khirki Extension ont un accès à l'eau plus ou moins régulier en fonction des maisons, et les coupures d'électricités sont généralement plus fréquentes qu'à Malviya Nagar.

FIGURE 94 Hauz Rani (Maneerat 2014).

FIGURE 95 Khirki (Maneerat 2014).

Begampur Village, Begampur slum et Valmiki Camp

FIGURE 96 Begampur Village. Photo : Eloise Layan, 2017.

Situé entre un grand parc à l'ouest et Malviya Nagar à l'est, le village de Begumpur 96 s'est construit à côté d'une ancienne mosquée (ou *masjid*) du 14e siècle dédiée à l'une des femmes du *vizir* de l'époque. L'ambiance des rues y est relativement similaire à celles d'Hauz Rani, avec des ruelles à peine plus larges. Le *slum* de Begumpur est inséré entre le village et Malviya Nagar, dont il est séparé par une grande route. Il se compose de petites habitations de briques peintes, de un à deux niveaux, dont certaines sont équipées de latrines. Au nord-est du *slum* de Begumpur, derrière un autre parc, se trouve le petit camp de relocalisation de Valmiki (figure 97). Un assez grand nombre de personnes y vit, entouré d'animaux d'élevages (cochons, et chèvres) dans des conditions d'hygiènes déplorables[311]. Les habitations sont faites de briques et de toits en tôle, l'accès à l'eau se fait via des pompes, et les latrines sont communes.

311. Un incendie s'y est d'ailleurs déclaré le 30 avril 2017, qui ôta la vie à trois personnes http ://www.gettyimages.fr/event/fire at Valmiki camp begumpur 700042475.

FIGURE 97 Extérieur du Valmiki Camp, Begumpur. Photo : Eloise Layan, 2017.

1.3 Une bonne accessibilité et des zones commerçantes attractives

Outre ces grandes différences dans les types de quartiers, cette zone est aussi relativement bien accessible, bordée au nord par l'Outer Ring-Road (la rocade extérieure) et desservie par deux arrêts de métro de la ligne jaune au nord avec la station Hauz Khas, et au sud-ouest par la station Malviya Nagar et un grand nombre de lignes de bus. Ainsi, nous pouvons postuler que si certaines personnes ne sont pas ou peu mobile, ce serait plus dû a des contraintes sociales et/ou économiques qu'à un enclavement géographique structurel.

Pour ce qui est de l'accès aux commerces, outre les nombreuses petites échoppes disséminées un peu partout, l'une des artères principales de Malviya Nagar fait office de marché (que l'on nommera *"Main Market"* par la suite) (figure 98) et regroupe un grand nombre d'enseignes de types, allant du vendeur ambulant à des chaînes de *fast-food*, en passant par des

magasins d'électroménagers et de vêtements. Même si ce grand marché ne concentre pas autant de commerces que Lajpat Nagar et très peu, voire aucune boutique de luxe (contrairement à Khan Market ou South Extension) et n'est pas spécialisé dans un domaine particulier (comme le marché de Sarojini Nagar dans le textile), il n'en demeure pas moins l'axe attractif majeur de la zone, où le trafic est très souvent congestionné le soir. Plus au sud, derrière une large avenue, se trouve un immense complexe commercial composé de deux *malls* adjacents, le « Select City » d'un côté et le « DLF Place » de l'autre (figure 99). Inaugurés en 2007, ces *malls* accueillent notamment deux cinémas ainsi que de nombreuses franchises de magasins et restaurants allochtones[312], dont les prix sont plus élevés que sur le *Main Market* et dont l'ambiance suggère un idéal de consommation plus mondialisé (Rault *et al.*, 2018)

FIGURE 98 Le *Main Market* de Malviya Nagar. (source : Somsakun Maneerat, 2014)

FIGURE 99 Le mall Select Citywalk de Saket. (source : http://www.selectcitywalk.com/)

Le *main marke*t de Malviya Nagar et les *malls* de Saket représentent deux formes de zones commerciales bien distinctes. La première est un mélange de commerces formels et informels, où des vendeurs ambulants côtoient des magasins construits et dont l'ambiance se rapproche parfois d'une sorte de chaos organisé, où il est parfois délicat de se frayer un chemin le soir. L'autre au contraire, est l'archétype du projet planifié, proposant une manière bien rangée et structurée de pratiquer la consommation dans un espace fermé et climatisé. Il paraît donc intéressant d'interroger les pratiques et niveaux de fréquentation de ces lieux commerciaux

312. Pour une liste des 175 commerces et services disponibles à Select City : http://www.se ectc tywa k.com/shopp ng category.php#

antagonistes, mais accessibles par les habitants de Malviya Nagar et des alentours.

1.4 Une zone où la dengue sévit

Concernant la dengue, nous pouvons également noter que le premier cas de 2009 fut enregistré à Malviya Nagar [313], mais seuls 63 cas ont été répertoriés dans la zone entre 2008 et 2010 (source NMIR & Olivier Telle). Si nous relions ces cas de manière spatiale et temporelle pour former des clusters regroupant des cas situés à une distance inférieure à 300 m et à des pas de temps de moins de 21 jours, comme effectué par (Telle, 2015), mais en utilisant l'algorithme st-dbscan [314], qui est une version tri-dimentionnelle de *dbscan* (voir chapitre 6), nous obtenons la figure 100. S'il est délicat de définir sans investigation poussée l'origine de la contamination (locale ou importée) du premier cas d'un cluster ou cas index les autres cas du même cluster ont de fortes chances d'être des contaminations locales, autour du lieu de domicile.

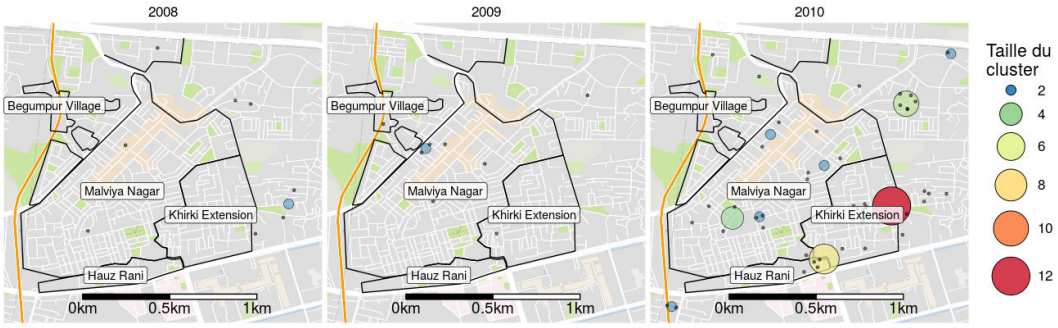


FIGURE 100 Les cas de dengue à Malviya Nagar. Répartition des cas de dengue (points noirs) et taille des différents clusters spatiotemporels enregistrés en 2008, 2009 et 2010.

Très peu de cas furent enregistrés en 2008 et 2009 (figure 100), avec seulement un cluster de 2 cas pour chacune de ces années, respectivement à Khirki Village et à Malviya Nagar, en face de Begumpur. En 2010 la dengue fut nettement plus active à Delhi, et on note la présence de gros cluster à Khirki (12 personnes), Chirag Delhi (10 personnes) et Hauz Rani (8 personnes). Malviya Nagar compte quant à elle 4 clusters, de 2 à 4 personnes, et aucun cas n'est enregistré dans Begumpur, tant au village que dans le *slum*, ni à Valmiki Camp.

Ainsi, malgré la présence du Madan Mohan Hospital, un hôpital sentinelle chargé de faire remonter les cas de dengue qui y sont enregistrés, très peu de cas ont officiellement été déclarés dans la zone, ce qui pourrait être le reflet d'un système de surveillance imparfait, où la plupart des malades se font soigner dans des cliniques ou hôpitaux privés, et les plus pauvres ne se font hospitaliser qu'en cas d'urgence absolue (Daudé *et al.*, 2017 ; Daudé et Mazumdar, 2016). La figure 101 montre une campagne de fumigation réalisée en 2015 dans le quartier de Khirki, qui

313. http://arch ve. nd anexpress.com/news/c ty s first dengue case reported from ma v ya nagar/489441/
314. Dont le code pour une utilisation sous R est disponible ici : https://g thub.com/fitrahmun r/Web C uster ng ST DBSCAN

suggère que les moustiques vecteurs de la dengue y étaient largement présents.

FIGURE 101 Fumigation à Khirki Extension. Photo : Sambit Dattachaudhuri, 2015

Avec une grande diversité socio-économique et de types de quartiers, des zones marchandes attractives, a *priori* pas d'entraves structurelles dans les potentiels de mobilités, le secteur de Malviya Nagar parait être la zone idéale pour effectuer une enquête de terrain.

2 Présentation de l'enquête de terrain

2.1 Protocole

Notre enquête de terrain a été effectuée en binôme, sur 20 jours entre le 1er avril et le 5 mai 2014 dans les différents quartiers de la zone décrite précédemment. Notre collègue, Shankare Gowda, docteur en science politique et chercheur indépendant, très expérimenté dans les enquêtes de terrain [315] fut plus que notre traducteur puisqu'il posait les questions et participait activement à la contextualisation des réponses fournies par les interviewés. Nous pouvons déjà préciser que les températures à midi étaient supérieures à 30 °C, et elles dépassaient les 37 °C à partir de la mi-avril. Ceci implique que beaucoup de gens étaient fatigués lorsqu'on les interrogeait en milieu de journée, c'est pourquoi nous avons privilégié les horaires matinaux et de soirée.

En plus de l'état civil (âge, sexe profession), le questionnaire de l'enquête se compose de 3 parties : une partie sur la connaissance et les pratiques (KAP pour *Knowledge Attitudes and Practices*), une partie sur les aspects socio-économiques du foyer, et une partie sur les pratiques de mobilités [316].

315. Il a effectué plusieurs dizaines d'études avec des membres du Centre de Science Humaines de Delhi, ou avec le Centre for Policy Research.
316. Voir annexe I pour l'ensemble du questionnaire

Les deux premières parties furent grandement inspirées par des questionnaires existants, développés par le NIMR et Olivier Telle. Le KAP visait à savoir si les personnes ou leurs connaissances avaient déjà eu la dengue, s'ils étaient au courant des modes de transmissions de la maladie et quelles méthodes utilisaient-ils pour s'affranchir de la nuisance des moustiques, que cela soit par l'utilisation de produit répulsif ou du contrôle des gîtes larvaires. Des informations générales telles que le nombre de personnes dans le foyer, le nombre de pièces à disposition, le fait d'être propriétaire ou locataire ainsi que le type de bâtiment furent notés afin de contextualiser le mode de vie des personnes. Différents éléments pouvant jouer un rôle plus ou moins favorable dans la création de gîtes larvaires furent également demandés, comme le système d'adduction d'eau, la présence de latrines, de *cooler*[317] ou de climatiseurs (voir chapitre 1). D'autres informations sur la possession de moyen de transport furent renseignées, telles que le nombre de deux roues motorisés et de voitures par foyer, ce qui permet d'apprécier leur potentiel de mobilité et leur niveau socio-économique. Nous avons également décidé de ne pas prendre en compte les castes et la religion, sujets et concepts beaucoup trop complexes pour un géographe non « Indianiste »[318].

Avant de commencer notre enquête sur les mobilités, trois jours ont été nécessaires pour tester différentes méthodes afin d'optimiser la collecte de l'information géographique et des déroulements temporels des différentes activités.

Collecter les lieux fréquentés

Pour ce qui est de la collecte des lieux fréquentés par les enquêtés, nous avons d'abord demandé aux gens de pointer sur une carte les lieux qu'ils visitaient régulièrement. Mais manifestement, la lecture des cartes n'était pas évidente chez la plupart des personnes interrogées, car bien que connaissant le nom des localités et/ou les différents « *landmark* » (points de repère) situés à proximité, peu de personnes arrivaient à localiser précisément et sans peine les lieux en question même ceux situés dans le quartier[319]. Nous avons donc décidé de changer de méthode et de répertorier simplement les lieux cités, en notant le plus d'éléments descriptifs des lieux qui nous étaient inconnus, et de faire des recherches ultérieures des localisations en passant par différents services de cartographies en ligne. Chaque lieu cité est de plus associé à un niveau de précision géographique « précis » lorsque le lieu a été bien retrouvé, « colonie » ou « district » si nous n'avons pas plus d'informations nous ont pu être donnée. Retranscrit de manière schématique, cela pouvait donner :

317. Système composé d'un bac rempli d'eau qui permet de rafraîchir l'air pulsé par un ventilateur.
318. Même si la religion transparaît soit par les attributs vestimentaires, soit dans les lieux fréquentés (temple, mosquée ou gurudwara).
319. Et si la réalisation de cartes mentales est une méthode pertinente pour apprécier les perceptions et représentations des territoires, cette approche s'éloigne de l'objectif de notre enquête, soit la collecte d'information sur les espaces d'activités des personnes du quartier

- *"I visit my relatives in East Delhi"*

- *"Where in East Delhi ?"*

- *"in Seelampur"*

- *"where in Seelampur ?"*

- *"In Seelampur only"*.

Ainsi, dans cet exemple, nous n'arriverons pas à obtenir d'information plus précise qu'une localisation quelque part dans un quartier hyper-peuplé de l'est de Delhi, ce qui ajoute encore des biais.

Collecter les fréquences de visites

Pour ce qui est des aspects temporels, nous aurions pu procéder comme pour les enquêtes de mobilité quantitatives, telles les enquêtes ménages déplacement ou d'utilisation du temps, c'est-à-dire demander aux personnes les lieux qu'ils ont fréquentés un jour donné, ou la veille, et ce, heure par heure. Cependant ces méthodes reposent sur un échantillonnage multi-niveaux [320] et ne sont fiables et pertinentes que si un nombre suffisant et représentatif d'individus est interviewé. Nos moyens étant très modeste et nos objectifs pas aussi ambitieux d'un point de vue de la représentativité, nous avons simplement demandé dans un premier temps aux gens interviewés (voir section suivante pour le choix de l'échantillon) de nous décrire une journée type, avec une forte précision horaire. Nous avons néanmoins écarté rapidement cette approche. En effet, notre questionnaire sur les mobilités arrivait après les questionnaires purement relatifs à la dengue et aux aspects socio-économiques, qui prenaient environ 5 à 10 minutes, durée suffisante pour voir apparaître un effet de lassitude, accentué par un effort assez important de mémorisation. Ensuite certaines personnes n'avaient pas forcément de journée très normée, et une réponse assez récurrente était *"it's depend"*. Compte tenu des écarts de précision entre les interviewés, certains pouvant dire à 10 minutes près ce qu'ils ont fait, d'autres n'ayant qu'une idée très vague de leur journée type, ceci impliquerait une grande difficulté dans la retranscription des questionnaires et dans leur traitement ultérieur, compte tenu des fortes hétérogénéités des niveaux de précision temporels. Nous avons donc revu nos exigences à la baisse et décidé de simplifier notre questionnaire. Pour ce faire, nous avons pré-établi une liste d'activités que les personnes pouvaient potentiellement exercer, à savoir aller au travail, se promener dans un parc, faire du sport, visiter des lieux religieux, visiter des proches, aller au restaurant, au marché, au

320. Il s'agit typiquement de sélectionner de manière plus ou moins aléatoire des foyers dans un quartier, puis de choisir une personne qui répondra au questionnaire pour le cas de Bangkok ou toute personne âgée de plus de 5 ans dans le cadre des EMD en France http://www.terr to res v e.cerema.fr/emd edvm et edgt methodes et gu des a679.htm .

cinéma ou dans les *malls*[321]. Nous avons également ajouté une catégorie de type « autre », permettant d'ouvrir le questionnaire. Ainsi, nous n'avions plus qu'à lister chacune des activités, et à demander à la fois la localisation de cette dernière, des tranches horaires (ou à défaut un moment de la journée), si possible une durée, ainsi qu'une fréquence hebdomadaire. Cette méthode semi-directive nous a permis de réduire drastiquement la durée des entretiens, tout en captant les informations qui nous paraissent indispensables : localisation, type d'activité, plage horaire et/ou durée, ainsi que la fréquence de réalisation.

Il nous a également semblé qu'une telle approche qui implique un certain flou géographique et temporel entraînait moins de réticences de la part des interviewés que lorsque l'on demandait précisément les lieux et horaires de fréquentation d'une journée type. En effet, un espace d'activité est quelque chose de très personnel, et insister pour qu'une personne dévoile heure par heure les activités qu'elle réalise peut être perçu comme une forme d'interrogatoire qu'elle désire arrêter avant terme. *A contrario*, une approche moins injonctive, qui passe simplement en revue des types d'activités auxquelles on demande d'associer une localisation et une fréquence de visite semblait moins déranger les personnes interrogées, tout en permettant une collecte des informations nécessaires à notre recherche.

Biais et limites

Comme toute enquête de terrain, outre les biais sur l'échantillonnage (voir section suivante), de nombreuses limites sont déjà à noter. Tout d'abord, les personnes interviewées peuvent oublier d'énumérer des lieux qu'elles fréquentent pourtant. Elles peuvent également avoir une perception erronée de leur fréquence de visite de certains lieux. À leur décharge, la plupart des activités dîtes flexibles ne sont pas nécessairement associées à une forte régularité. Par exemple il peut y avoir des moments où une personne va régulièrement au cinéma car les films lui plaisent, et de longue période où les films projetés lui sont moins attractifs. Définir une fréquence pour une telle activité est délicat dans ce contexte. Si certaines activités sont associées à des émotions plutôt positives, par exemple pour un jeune enfant faire un pique-nique avec ses parents un soir à l'India Gate, il y a de fortes chances qu'elle soit retranscrite dans nos interviews. À ce moment, la notion de fréquence (ici très faible) permet de relativiser ces activités plutôt anecdotiques.

Dans tous les cas, il ne faut pas omettre la notion de désirabilité sociale (Parry and Crossley, 1950), où les réponses des personnes peuvent être motivées de manière plus ou moins consciente par une volonté de renvoyer une image positive, en accord avec ce qu'elle pense être les pratiques jugées comme acceptables par la société (ou l'interviewer). Typiquement, si la personne estime qu'aller dans un bar ou acheter de l'alcool est perçu négativement, il est possible

321. Et si certaines personnes peuvent voir des proches pour aller au restaurant puis faire du shopping dans un *mall*, nous avons considéré cela comme aller dans un *mall*.

qu'elle ne mentionne pas ces activités alors qu'elle les effectue plus ou moins régulièrement. De même, une personne croyante pourrait exagérer la déclaration de ses visites de lieux de cultes pour passer pour plus dévote qu'elle ne l'est réellement. À ces remarques générales s'ajoute également la représentation qu'ont les personnes interrogées du binôme qui effectue l'enquête, à savoir un Indien originaire du sud de l'Inde et un jeune Français blond qui ne parle pas bien le hindi. De ce fait, certaines personnes vont peut-être surestimer leurs fréquences de visite de *malls* ou de restaurants pour coller à un imaginaire des pratiques des sorties des Européens. *A contrario*, certaines personnes relativement pauvres vont peut-être réduire inconsciemment ou non leur espace d'activité pour se forger une image de personne plus précaire qu'elle ne l'est réellement, afin de susciter de l'empathie.

2.2 Présentation de l'échantillon

2.2.1 Déroulement de l'enquête

FIGURE 102 Une ambiance aux abords d'un marché en Inde. Par Arka Alam (2013), artiste Bengali ayant vécu à Delhi - https://www.instagram.com/tumbleslumber/.

Un des objectifs de notre terrain est de constituer, non pas un échantillon représentatif de la population, mais un panel de la plus grande diversité démographique (âge, sexe) et socio-économique possible, afin d'apprécier notamment les potentiels de mobilité de chacun des différents groupes et des divers lieux de coprésences. Ces interviews se sont déroulés principalement en matinée et en début de soirée, auprès des personnes disponibles rencontrées au gré des déambulations (nous avons parfois interpellé quelques personnes qui étaient dans la cour de leur maison). Au fur et à mesure que les personnes furent interrogées, les nouveaux interviewés furent ciblés de manière à réduire la surreprésentation de certaines catégories.

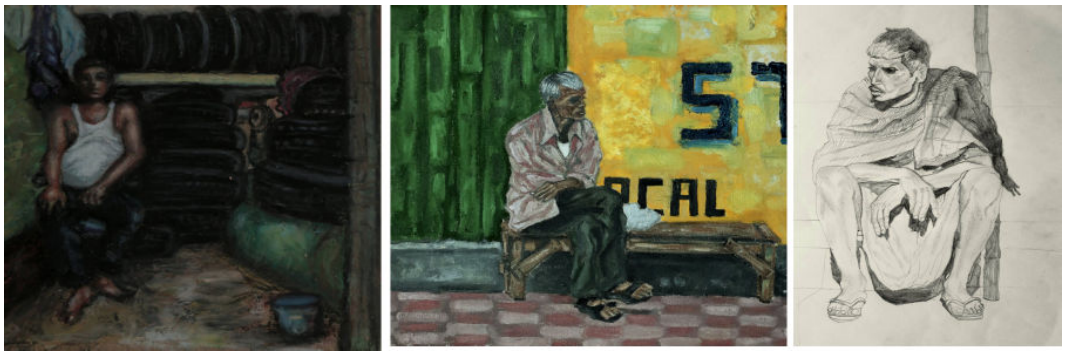


FIGURE 103 Des Hommes qui attendent. Illustré par "*Man in a tyre shop*", "*Smoke break*" et "*Man sitting beside a bamboo pole*". Arka Alam (2017,2017,2016).

Par exemple, en journée, la population des abords du marché de Malviya Nagar est constituée d'un grand nombre d'hommes entre 30 et 50 ans qui attendent (figure 103). Certains sont dans leurs commerces, prêts à recevoir leur clientèle, d'autres sont assis dehors, l'air songeur. Ces personnes sont souvent d'accord pour répondre à nos questions, ayant manifestement du temps disponible. Nous en avons interviewé 9, allant du vendeur de glaces, au vendeur d'appareils électroniques, au gardien de distributeurs de billets ou au ramasseur d'ordures en pause à un stand de chai. Cela dit, malgré la relative accessibilité des personnes dans cette zone, nous avons choisi ensuite de cibler préférentiellement des personnes du quartier d'autres catégories socio-démographiques, à savoir des personnes plus aisées, des femmes et des jeunes, afin d'équilibrer notre échantillonnage dans le quartier de Malviya Nagar. Il va de soi que lorsque nous interrogions des personnes à Valmiki Camp ou dans le *slum* de Begumpur, nous recherchions plus un équilibre démographique qu'un large spectre de niveau de richesse la population de ces quartiers étant visiblement uniformément (très) pauvre.

FIGURE 104 Portraits de femmes. *Lady In A Wildflower Field*, - *untitled* - *a girl with a flower. Arka Alam (2014, 2015, 2016).*

Interviewer des femmes dans des lieux très fréquentés était assez délicat et parmi la

faible part de femmes présentes, la plupart d'entre elles refusaient de prendre part à notre questionnaire. Si nous mettons de côté le fait qu'elles n'avaient simplement pas envie ou d'autres choses à faire, nous ne pouvons écarter que ces réticences sont peut-être aussi le fruit de constructions mentales d'une société indienne très genrée, qui expliquerait de manière très simpliste que les femmes sont globalement moins enclines à répondre à des questions posées par deux hommes dont un étranger dans un espace public assez fréquenté. Loin de faire des généralisations, nous pouvons souligner que dans des zones moins passantes, plus proches des habitations, elles devenaient plus accessibles et nous avons pu interviewer 20 femmes, entre 10 et 70 ans.

Nous n'avons interviewé qu'une personne par foyer, sauf vers la fin de l'enquête où afin de recueillir des informations sur les plus jeunes, nous avons interrogé plusieurs enfants d'un même foyer (3 foyers pour 7 enfants entre 8 et 15 ans).

Nous avons également pu interviewer quelques personnes assez riches, des propriétaires rentiers, une femme de journaliste, ou encore un avocat à la haute cour de New Delhi. Finalement, la tâche la plus délicate fut d'interviewer des personnes très riches. Bien que nous focalisant quelques jours dans la colonie aisée et fermée de Panschilla Park, les quelques personnes qui s'y trouvaient étaient principalement les domestiques et des femmes au foyer qui ne souhaitaient pas répondre à notre questionnaire. Les seules personnes du quartier ayant accepté de participer à notre enquête étaient un jeune avocat travaillant dans le quartier, et un étudiant qui aime bien se promener dans un des parcs, mais aucun d'eux ne résidait dans la colonie. Nous avons néanmoins pu interviewer dans le grand parc de Begumpur un jeune issu d'une famille très aisée du quartier riche de Shivalik.

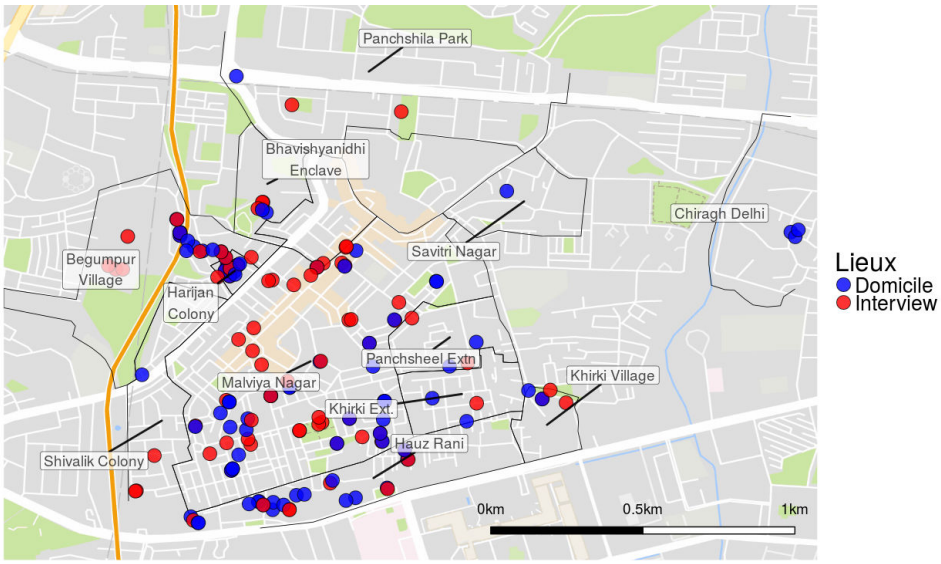


FIGURE 105 Localisation des lieux d'interview et des domiciles des personnes interrogées.

Nous avons au final interviewé 99 personnes dans notre zone d'étude, dont 80 habitaient

dans l'un des 8 quartiers présentés plus haut (figure 105). Parmi les 99 personnes interrogées, 4 d'entre elles ont déjà eu la dengue (en 2009, 2010 et 2 en 2013), et ont fait leurs analyses dans des cliniques ou hôpitaux privés, tandis que 20 interviewés ont des connaissances qui ont déjà contracté la maladie. Ces premières informations sur un très faible échantillon permettent déjà de relativiser les 63 cas enregistrés entre 2008 et 2010, soutenant l'hypothèse d'un système de surveillance qui tend à sous-estimer les différentes épidémies de dengue.

2.2.2 *Structure démographique et socio-économique*

La figure 106 présente la structure démographique de notre échantillon, constitué majoritairement d'hommes 79 contre 20 femmes. Cet écart est essentiellement dû aux difficultés d'interviewer les femmes, comme expliqué précédemment, et rendra délicat l'analyse des différentiels de mobilités entre les genres. Bien qu'ayant fait un effort pour interviewer toutes les classes d'âges, les 20-45 ans sont bien plus représentés.

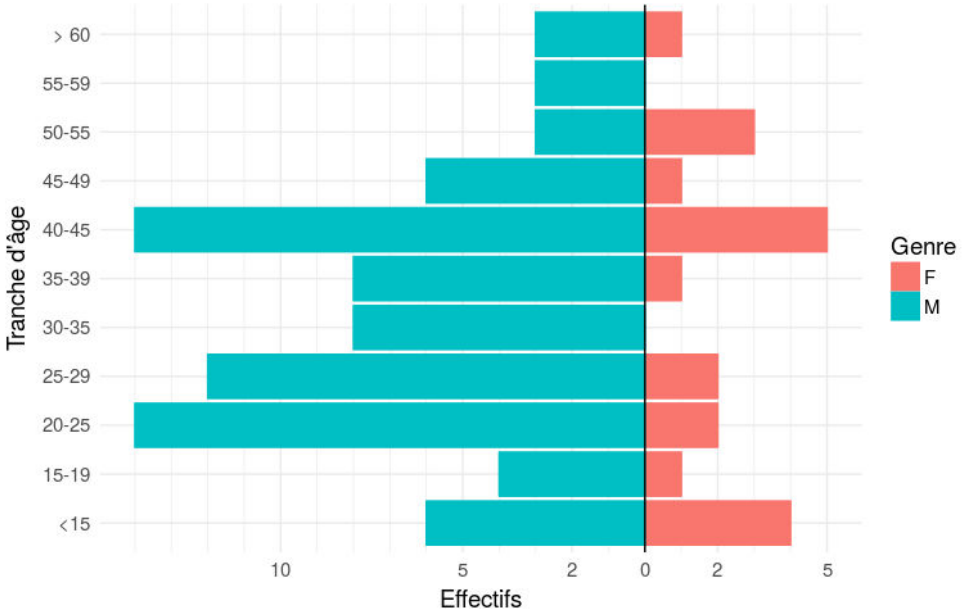


FIGURE 106 Structure démographique de l'échantillon.

Les 12 activités principales, les plus représentées dans notre échantillon sont regroupées dans la figure 107. Les étudiants et écoliers constituent le groupe le plus important, avec les commerçants de rue [322] et les commerçants [323]. 8 personnes se considéraient comme « Business man », concept assez large et équivoque [324]. Nous avons également interviewé des femmes sans

322. considérés comme propriétaires de leur commerce, mais ce dernier étant informel et dans la rue.
323. considérés comme propriétaires de leur commerce, mais ce dernier étant installé dans un bâtiment en dur.
324. Certaines entrevues pouvaient alors ressembler à cela :
What is your job ?
Business

activité professionnelle, et certaines d'entre elles se présentant plus spontanément que d'autres comme femme au foyer. Les femmes de ménage et les chauffeurs furent sans surprises, rencontrés soit sur le lieu de travail de leur employeur, soit dans les quartiers les plus pauvres (Begampur et Hauz Rani). Les chauffeurs auront *de facto* une mobilité très importante car liée à celle de la famille qui les emploie.

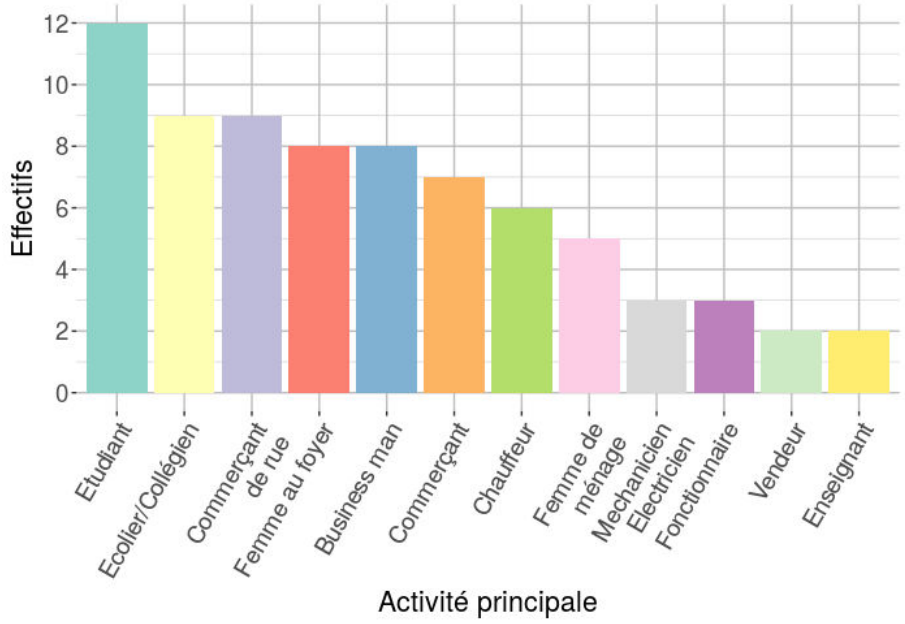


FIGURE 107 Effectifs des personnes interrogées par type d'activité principale.

À ces catégories principales s'ajoutent divers métiers dont un photographe, un DJ de mariage et autres cérémonies religieuses, un développeur web, un avocat, un ramasseur d'ordures, un autre avocat, mais à la haute cour de Delhi, un traducteur au *Max Hospital*, une jardinière qui parcourt 30 km pour s'occuper du parc de Baghan Singh, un peintre, un gérant d'une entreprise de voyage qui passe sa journée à jouer au cricket, un décorateur d'intérieur, un « Guru » local qui se considère comme consultant, un garde sécurité dont le pouce de la main droite est dédoublé, une personne qui travaille dans le cinéma, mais au guichet, ou encore un gérant d'hôtel.

Déterminer les classes sociales à partir de la taxe foncière

Les revenus des personnes ont été demandés mais ne furent pas systématiquement communiqués. Nous allons ici mettre en relation ces niveaux de ressources déclarés à la catégorie de la taxe foncière afin d'en évaluer le niveau de crédibilité. Cette dernière, de par son élaboration et malgré les quelques limites exposées dans le chapitre 2, semble être *a priori* un bon marqueur

What kind of business ?
You know, Business".

du niveau de richesse. Un simple croisement entre le lieu de domicile et la colonie devrait permettre d'avoir une première indication sur la classe sociale de la personne, à mettre au regard des revenus déclarés. Cela dit, nous avons interviewé trois personnes habitant dans le Valmiki Camp qui relevaient de la taxe foncière de Bhavishyanidhi Enclave (B), ce qui ne reflète pas leur réalité. Nous les avons donc mis en catégorie G, comme pour le *slum* de Begumpur. Nous avons également interviewé un ramasseur d'ordures, qui dort sous le *fly-over*[325] de Panscheel, et une intersection de son lieu de « domicile » avec la carte de la taxe foncière ferait de lui quelqu'un de très riche. Nous l'avons donc également mis en catégorie G. La figure 108 montre la localisation des domiciles des personnes interrogées, tandis que la figure 109 montre les effectifs par colonie en fonction de la taxe foncière. Il ressort que toutes les personnes interrogées habitent dans le sud de Delhi, avec une majorité de personnes en catégorie C, F et G.

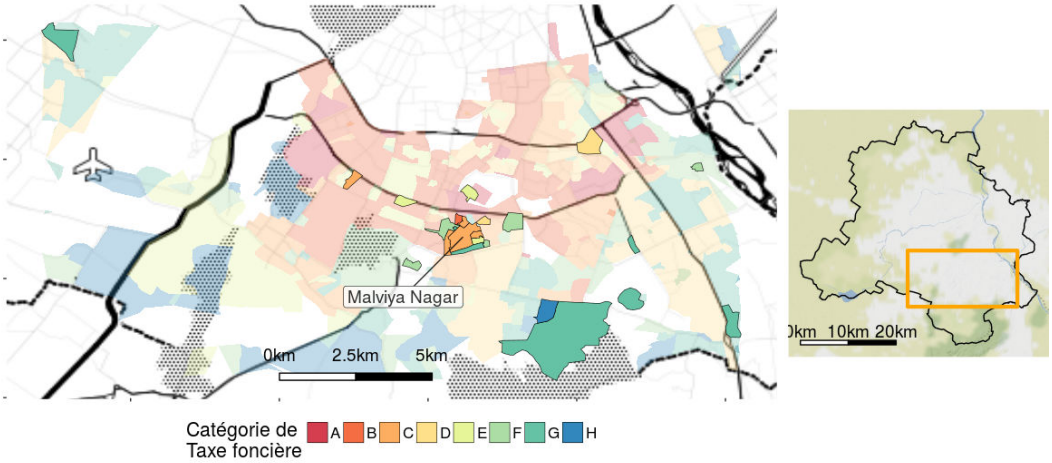


FIGURE 108 Localisation des domiciles des personnes interrogées. Les colonies sans transparence accueillent au moins une personne de l'échantillon.

La figure 110 présente une répartition de la population par catégorie d'imposition d'après nos données carroyées à l'échelle de la ville (gauche), dans la zone d'étude (centre), et dans notre échantillon (droite). Comme vu précédemment, la zone de Malviya Nagar regroupe toutes les catégories (sauf les villages ruraux, H), avec une plus grande part de personnes vivant dans des colonies aisées (A, B, et C) par rapport à l'ensemble de la ville. Notre enquête n'a pas pu cibler de personnes vivant dans des catégories A et B, mais une part à peu près équivalente aux données de la ville pour les personnes résidant en zone F et G, et aux données du quartier pour les catégories C, D et E.

325. zone où la rocade passe au dessus d'une autre route, formant un abri.

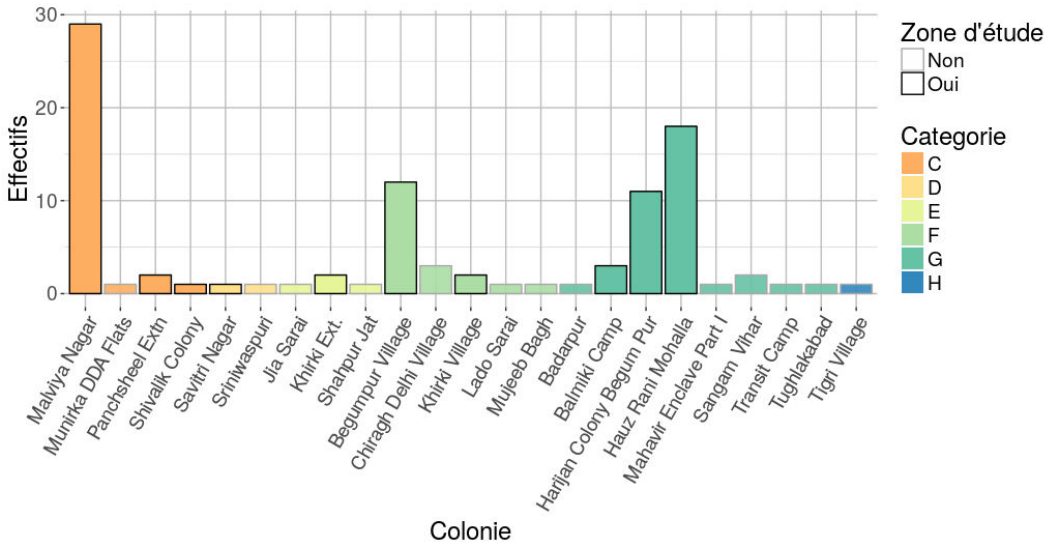


FIGURE 109 Répartition des effectifs par colonie et par catégorie de la taxe foncière.

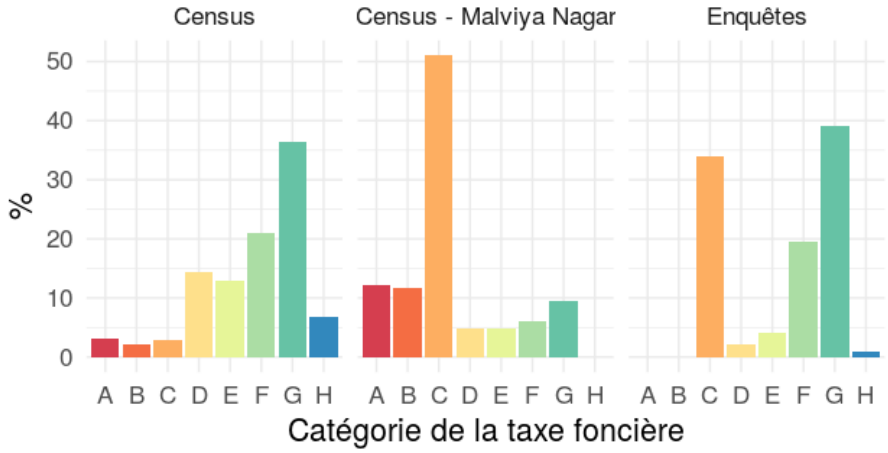


FIGURE 110 Répartition de la population par catégorie de la taxe foncière. Sur l'ensemble de Delhi (à gauche), dans la zone d'étude (centre), et pour notre échantillon (droite). Les zones de New Delhi et Delhi Cantonment sont considérées comme étant en catégorie A.

Le revenu mensuel moyen à Delhi est, à titre de référence, estimé à 25 000 roupies, mais environ 10 % de la population vit avec moins de 1134 roupies en 2017 [326]. La figure 111 montre la répartition des revenus individuels (gauche) ou du foyer (revenus du foyer divisé par le nombre de membres, à droite), en fonction de la catégorie de la taxe foncière du quartier de domicile.

Malgré un revenu moyen de 20 360 roupies (N=64) proche de la moyenne des statistiques officielles, nous pouvons noter que la plupart des personnes interrogées déclarent des revenus inférieurs à 15 000 roupies (˜200 €). Pour les personnes ayant déclaré les ressources de leur foyer (N=23), la moyenne par membre est proche des 10 000 (10 566 INR ˜150 €), même si là

326. http://www.delhi.gov.in/wps/wcm/connect/doit des/DES/Our+Services/Statistical+Hand+Book/

encore, la plupart des personnes avaient des revenus individuels très faibles (<6000 roupies). De plus, nous n'avons pas d'informations pour 19 personnes, soit parce qu'elles n'ont pas souhaité les partager, soit parce qu'elles ne le connaissaient pas. Comme le montre la figure 111 ci-dessus, les personnes qui résident dans des villages urbains ou dans des *slum* (catégorie F / G) tendent à avoir les revenus les plus faibles et la plupart des personnes ayant des revenus supérieurs à la moyenne habitent dans des zones de catégories E à C. Tout ceci tend à valider que la taxe foncière est un proxy relativement correct du niveau de ressource des résidents.

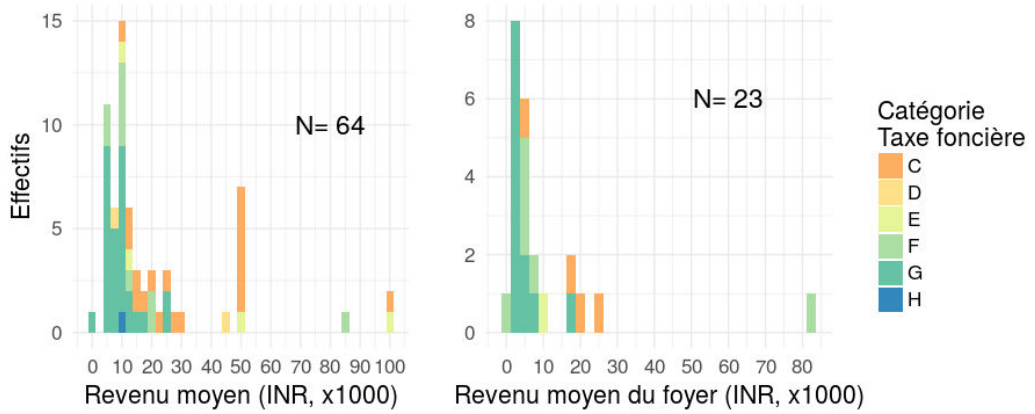


FIGURE 111 Répartition de l'échantillon selon le revenu moyen individuel (gauche) et le revenu moyen du foyer (droite).

Nous pouvons cependant noter quelques exceptions chez deux habitants de Begumpur Village qui présentent des revenus assez conséquents. L'un d'entre eux est étudiant mais fait partie d'une famille de propriétaire rentier. L'autre a un statut assez particulier car il s'agit d'un guru hindou qui, bien que semblant vivre dans le dénuement dans une petite bicoque d'une pièce faisant office de temple, à la limite de Begumpur *slum*, déclare gagner environ 85 000 roupies par mois de par ces activités de « conseiller ».

Néanmoins nous devons également souligner le fait qu'il y a parfois des incohérences, notamment quand certaines personnes déclarent payer un loyer quasiment égal à leurs revenus, ou lorsque le niveau de ressource annoncé parait en désaccord avec leur niveau de vie apparent (possession de plusieurs véhicules, climatiseurs dans toutes les pièces, etc.). Nous allons donc essayer de rechercher des critères autres que les revenus déclarés pour estimer les catégories socio-économiques de notre échantillon.

D'autres indicateurs de richesse / pauvreté

Lors de nos entretiens, nous demandions aux personnes si elles avaient des systèmes de refroidissement thermiques à leurs domiciles, car les moustiques sont très susceptibles de pondre dans les réservoirs des coolers, et beaucoup moins dans les climatiseurs. Nous demandions

aussi quels étaient les véhicules disponibles dans le foyer, afin d'apprécier les potentiels de mobilités, ainsi que le nombre de pièces de leur domicile et la taille du foyer, comme indicateur de promiscuité.

La réalisation d'une ACP (figure 112) à partir de ces informations[327] montre que le nombre de climatiseurs par pièce du domicile (pc_ac) est très lié à la possession de voitures (pc_car), opposée à la part de *cooler* (pc_cool). À noter que la part de deux roues motorisées (pc_scoot) augmente avec la place disponible par individus dans le foyer (pc_room). La partie droite de l'axe 1 décrit ainsi une plus grande propension à avoir des équipements relativement coûteux et de l'espace dans le domicile, qui peuvent être vus comme des indicateurs de richesse. La taille du foyer (nb_hh) a la plus grande contribution à la construction de l'axe 2 et un cos^2 très élevé, mais avec une valeur propre proche de 0 sur l'axe 1, cette variable semble plus opposer les "joints family", c'est-à-dire les familles agrandies où plusieurs générations vivent sous le même toit, aux foyers plus restreints.

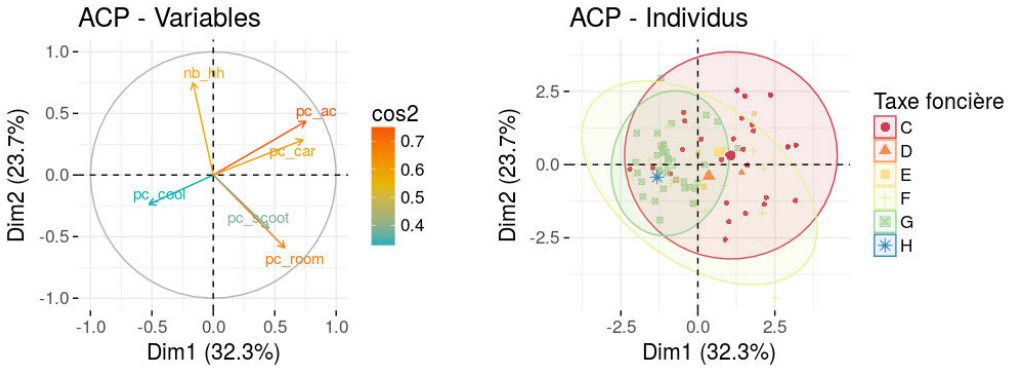


FIGURE 112 Résultat d'une ACP sur des indicateurs de richesse où nb_hh est la taille du foyer, pc_ac le nombre de climatiseurs par pièce, pc_cool, le nombre de *coolers* par pièce, pc_car, le nombre de voitures par personne du foyer, pc_scoot, le nombre de deux roues motorisés par foyer, et pc_room, le nombre de chambre par membre du foyer. Voir en annexe J pour la contribution des variables et individus pour les axes de l'ACP.

La visualisation des individus avec l'ajout de la taxe foncière comme variable qualitative supplémentaire ne fait pas apparaître de groupes bien séparés, car même si les individus vivant dans les bidonvilles ou à Hauz Rani (catégorie G) sont plutôt situés à gauche de l'axe 1, ce groupe est englobé par l'ellipse des habitants des catégories C. Ce dernier aspect tend à confirmer les observations de la figure 111 sur les revenus déclarés, à savoir que la plupart des personnes résidant dans les zones de catégories F et G ont en général moins de ressources, mais que vivre dans une zone de catégorie C ne confère pas pour autant un niveau de richesse forcément plus élevé.

327. Sur des individus âgés de plus de 17 ans, pour éviter d'avoir plusieurs personnes d'un même foyer.

Si la taxe foncière semble être un bon indicateur global des niveaux de richesses des personnes résidant dans les quartiers, plus de subtilités apparaissent logiquement lorsque nous raisonnons à l'échelle individuelle. Nous allons donc définir un nouvel indicateur socio-économique, basé non pas sur le lieu de résidence, mais sur les possessions matérielles des personnes[328]. Nous partons du principe que plus le foyer d'une personne a de voitures, plus ce dernier est riche. De même, un foyer équipé de climatiseurs est plus riche qu'un foyer équipé de *coolers*. Les critères sont répertoriés dans le tableau 9 ci-dessous et entraînent des classes relativement équilibrées en termes d'effectifs.

Critère	Niveau de richesse	Effectifs
climatiseur > 0 & Voiture > 1	Très élevé	9
climatiseur > 0 & 0 < Voiture <=1	Élevé	17
climatiseurs = 0 & voiture > 0 ou climatiseurs > 0 & voiture =0	Moyen	19
cooler > 0 ou motorbike > 0	Faible	29
Rien	Très Faible	18

Tableau 9 Répartition des effectifs selon le niveau de richesse, défini à partir de critères de possession.

Il ne s'agit cependant que d'une classification très relative pour plusieurs raisons. Tout d'abord les personnes considérées ici comme ayant un niveau de richesse très élevé est à relativiser par rapport aux véritables élites économiques de la ville, et les limites entre ces catégories sont assez arbitraires et finalement très poreuses, notamment pour les niveaux les plus faibles car dépendant de bien matériels parfois assez accessible à la propriété, comme un deux roues motorisé de piètre qualité. De plus certaines personnes assez aisées vivant dans le village de Begumpur n'ont pas de voitures car les ruelles sont trop étroites pour pouvoir circuler. Les effectifs de chacune de nos nouvelles classes en fonction du revenu déclaré et du lieu de domicile sont synthétisés dans la figure 113 ci-dessous.

Nous pouvons apporter quelques précisions à la figure 113, en ajoutant que les personnes considérées comme ayant un niveau de richesse élevé mais qui habitent en zone E, F et G sont respectivement avocat, agent immobilier / rentier et traducteur au max Hospital. La personne considérée comme niveau de richesse très faible à Malviya Nagar est un vendeur de tchai qui dort près de son commerce. Une personne du *slum* de Begumpur, chauffeur de métier, nous a également dit qu'il était propriétaire d'une voiture, ce qui le classe *de facto* comme une personne aux ressources élevées.

328. "On sait ce que tu es quand on voit ce que tu possèdes" (Geoffroy Mussard, 1997)

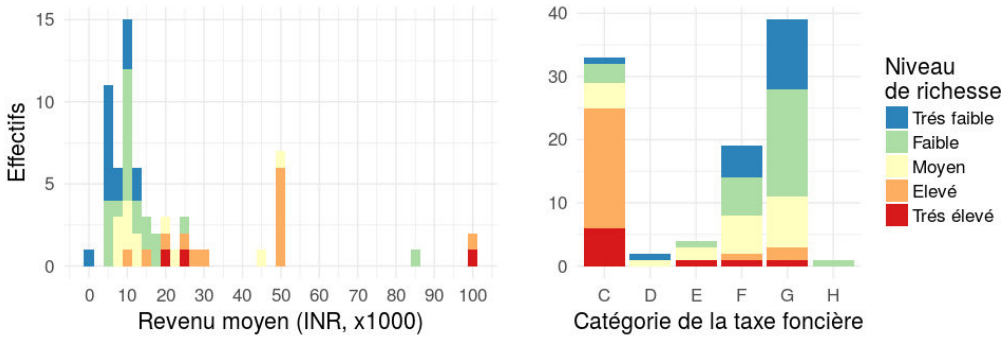


FIGURE 113 Lien entre le niveau de richesse estimé et la catégorie de la taxe foncière du lieu de domicile.

En somme, le revenu déclaré parait tout de même être une information relativement crédible, malgré les quelques incohérences évoquées plus haut. Néanmoins, vu que ces données sont incomplètes et que les différences sont assez importantes entre le revenu déclaré individuellement et le revenu du foyer, nous n'utiliserons pas cette information. De manière générale, la taxe foncière est un indicateur du niveau de richesse relativement correct, mais notre indice basé sur la possession semble plus en accord avec ce que nous avons observé, car basé sur des individus et non sur une population comme la taxe foncière. De plus il permet une segmentation en groupe assez équilibré, ce qui devrait faciliter l'interprétation des différents espaces d'activités et des potentiels de mobilités en fonction du niveau de ressources.

2.2.3 *Étude des espaces d'activités*

Comme vu dans la partie A, le concept d'espace d'activité nous semble être une bonne approche pour travailler sur les mobilités urbaines. En effet, ce concept simple et souple, classe et hiérarchise les activités effectuées par individu selon un niveau de flexibilité spatio-temporelle et de fréquence de visite, ce qui autorise, en plus d'apprécier de manière individuelle ou collective les portions de la ville fréquentées, permet aussi une approche analytique plus quantitative. Notre enquête qui suivait ce concept nous a permis d'associer chaque lieu fréquenté à une activité, une fréquence de réalisation et une plage horaire plus ou moins précise.

La figure 114 présente pour chaque individu (en abscisse) le nombre de lieux dans lesquels une personne effectue une activité, par exemple visiter un « *mall* », « Restaurant », « Marché », « Cinéma », ou « Gare / Aéroport ». Nous considérons ici l'activité principale comme l'activité qui prend le plus de temps à une personne et qui s'effectue hors du domicile[329]. L'activité secondaire est en lien avec l'activité principale. Typiquement si une personne a plusieurs emplois il s'agira de celui qu'il effectue le moins fréquemment et dans le cas d'un écolier, cela correspondra aux cours du soir dans un organisme privé. La catégorie « Parc / Loisir » regroupe les activités qui sont effectuées dans un parc, ce qui prend en compte le fait de discuter sur un

329. Nous ne considérons donc pas les personnes qui travaillent à domicile.

banc, les promenades matinales ou encore les matchs de cricket. L'activité « Amis/Famille » consiste en la visite d'un proche, et « Religion » la fréquentation d'un lieu de culte. Les lieux considérés comme « Autre » sont ceux où la personne a été interviewée, ou qui ne rentrent dans aucune catégorie, comme « faire un tour en moto ». À noter qu'aucune femme au foyer n'a cité « aller chercher les enfants à l'école » comme étant une activité.

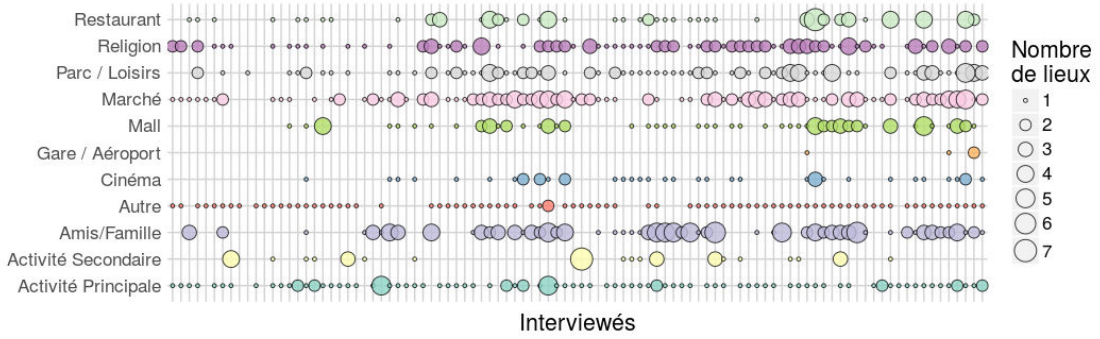


FIGURE 114 Répartition des activités en fonction des interviewés. Chaque ligne verticale correspond aux activités effectuées par une personne interviewée. La taille des cercles correspond au nombre de lieux différents où un individu effectue chaque activité.

La figure 115 montre d'un côté la distance au domicile par activité (en haut), et de l'autre la fréquence hebdomadaire de réalisation de ces dernières (en bas), pour l'ensemble de notre échantillon certaines personnes réalisant plusieurs fois la même activité. Sans surprise, l'activité principale est effectuée très régulièrement, et la moitié d'entre elles se trouvent à moins de 2 kilomètres du domicile (contre 32 % dans d'après le recensement 2011 pour le sud de Delhi). Les fréquentations des *malls*, cinémas, restaurants et les visites chez les proches ont relativement le même profil temporel, et sont plutôt occasionnelles pas toutes les semaines. *mall* et cinéma ont à peu près le même profil de distances, ce qui est relativement normal puisque les cinémas et *malls* les plus fréquentés sont situés à proximité (Saket). Néanmoins les *malls* semblent avoir un pouvoir attractif plus important avec la distance que les cinémas. La plupart des personnes vont dans des restaurants situés à moins de 5 kilomètres de chez eux. Les distances que les gens parcourent pour rendre visite à leurs proches sont propres à chacun, et sont parfois très importantes du fait de la localisation de leurs connaissances dans la ville et d'une probable relativisation des contraintes de déplacements pour s'adonner à une activité de socialisation.

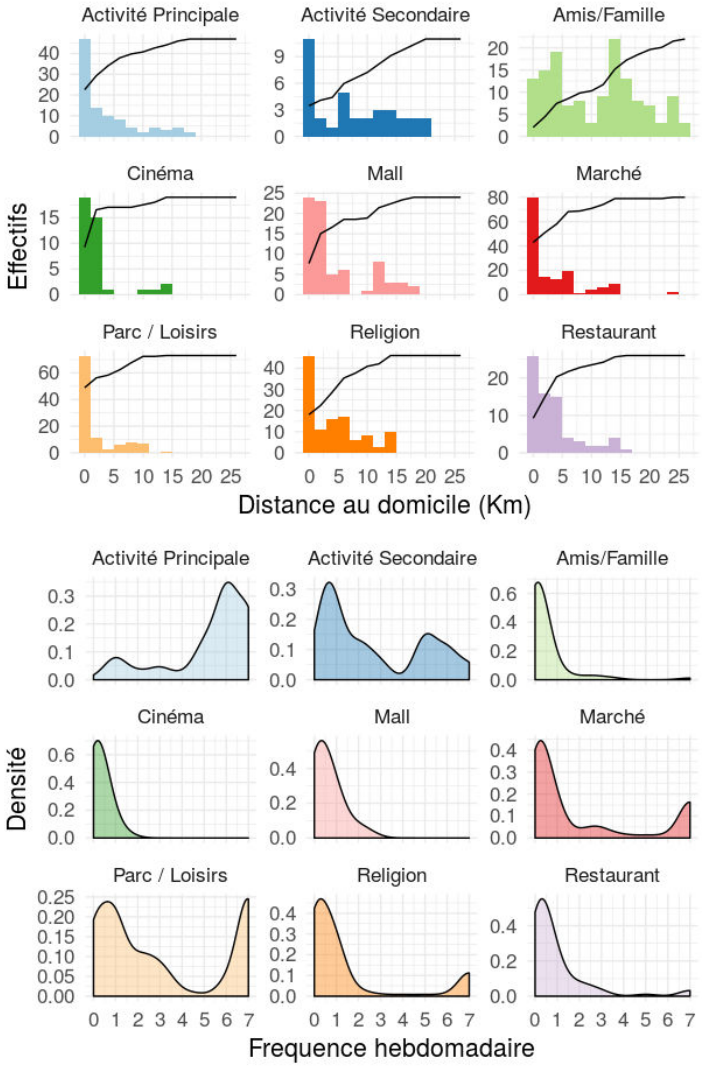


FIGURE 115 Distance des activités au domicile (haut) et fréquence de visite hebdomadaire (bas). La ligne noire (haut) montre les effectifs cumulés.

La figure 116, qui synthétise la figure 115, montre pour le cas des marchés 3 types de comportements : les personnes qui vont presque tous les jours au marché assez proche de chez eux, ceux qui y vont moins fréquemment, et ceux qui y vont de temps en temps dans un marché plus éloigné, et probablement plus grand ou spécialisé. Nous pouvons faire des observations similaires pour les fréquentations des lieux religieux, ou les lieux de cultes locaux sont fréquentés dans le cadre d'activités quotidiennes ou hebdomadaires, tandis que les temples (*e.g.* Kailkaji Mandir), mosquées (*e.g.* Jama Masjid) ou gurudwaras[330] (*e.g.* Bangla Sahib) plus importants sont visités dans le cadre d'offices de premier plan (*e.g.* Durga Puja, Maha Shivatri, Aïd Mubarak, Aïd el-kibir, Guru Nanak Gurpurab, etc.). Le niveau de fréquentation des parcs dépend des habitudes de chacun, mais on retrouve encore le même genre de relation entre la

330. Les temples Sikh.

distance, la fréquence de visite et la taille ou l'importance du lieu.

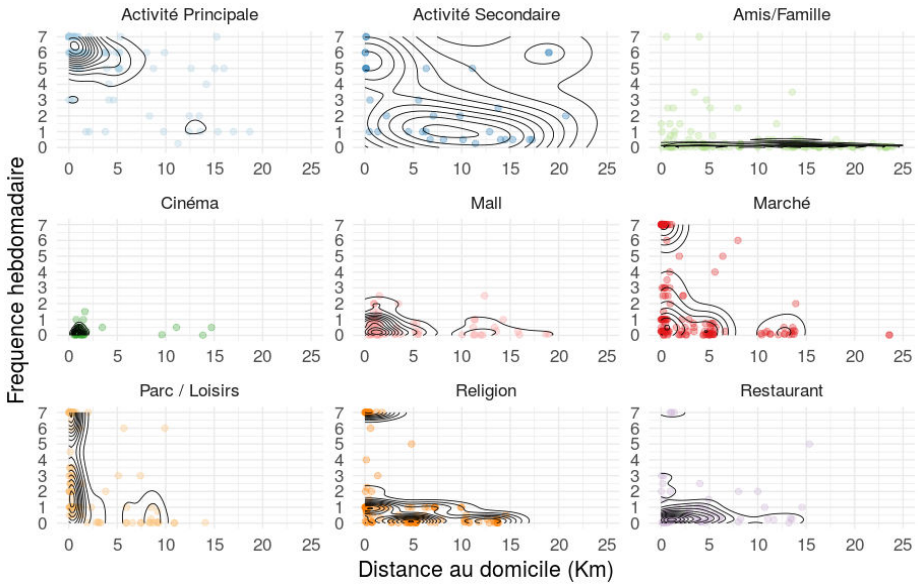


FIGURE 116 Fréquence de visite hebdomadaire d'une activité selon sa distance au domicile. Les isolignes représentent les clusters de densités

Ces observations sur l'ensemble de l'échantillon donnent des informations générales sur les fréquences de visites et sur les distances parcourues généralement pour effectuer telle ou telle activité. Nous allons maintenant mobiliser d'autres informations tels que le genre et notre estimation du niveau de richesse pour conduire une double analyse : l'étude des coprésences des différents groupes dans divers lieux, et l'analyse de leur niveau de mobilité à travers diverses métriques de déplacement.

2.2.4 *Étude des coprésences*

Selon (Levy et Lussault, 2004), la coprésence « se caractérise par le rassemblement et l'agrégation en un même lieu de réalités sociales distinctes ». Il s'agit d'un paramètre important à prendre en compte, que cela soit dans les processus de création d'un lieu que dans la propagation des épidémies (Daudé et Eliot, 2005). L'idée ici est de voir où, quand et dans quelles mesures les différentes classes sociales et genres se mélangent, et ce que cela peut impliquer dans une optique de modélisation des mobilités.

Coprésences Homme / Femme

La figure 117 montre les lieux fréquentés en fonction du genre dans le quartier de Malviya Nagar et dans l'ensemble de la ville. Comme vu précédemment, les écarts d'effectifs dans les échantillons ne nous permettent pas de tirer des conclusions définitives, mais il ne semble pas

avoir de différences très marquées dans les zones visitées dans le quartier. En revanche, à l'échelle de la ville, les hommes paraissent fréquenter plus de lieux éloignés de leur domicile, ce qui va dans le sens des données du recensement (chapitre 2).

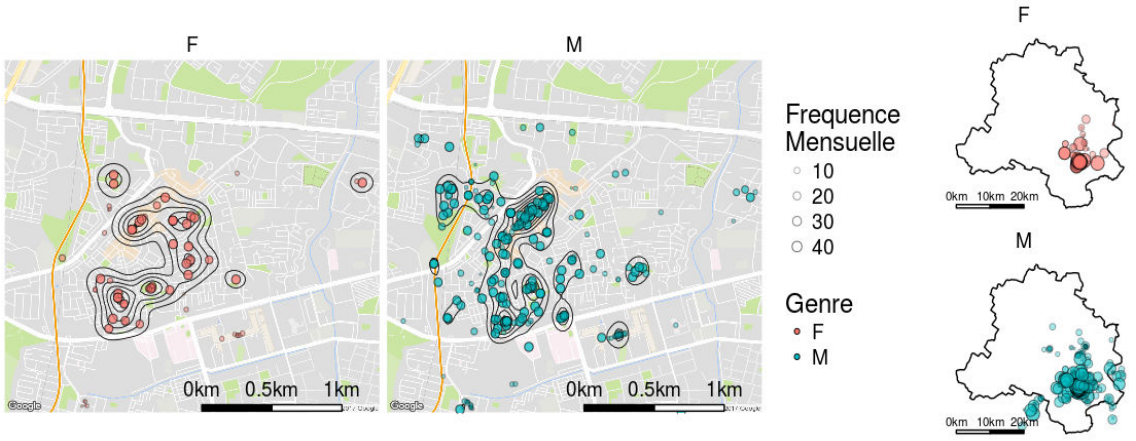


FIGURE 117 lieux déclarés comme fréquentés par les hommes et les femmes interrogés, outre le lieu de domicile. L'encart de gauche correspond à Malvia Nagar, celui de droite à l'ensemble de la ville. La taille des cercles et le niveau de transparence indiquent la fréquentation mensuelle, tandis que les isolignes les densités de fréquentation.

Coprésences des différentes classes sociales

La figure 118 présente les zones fréquentées à Malviya Nagar et à Delhi, en fonction des différents niveaux de richesses que nous avons établies précédemment, sans prendre en compte les lieux de domicile. Autant il est délicat de définir des zones propres à chaque groupe, autant il apparaît que certains lieux sont fréquentés par tous, tels que le *Main market* et les *malls*. À l'échelle de la ville, aucune distinction claire n'apparaît.

La figure 119 présente pour chaque groupe social défini selon notre indicateur de richesse, la part de fréquentation de certains lieux. Nous pouvons noter que certaines activités sont effectuées de manière similaire par l'ensemble de notre échantillon, à savoir aller au marché, dans les parcs ou dans les lieux de cultes[331]. Aussi, plus les personnes de notre échantillon sont *a priori* aisées, plus elles auraient tendance à fréquenter des restaurants, des *malls* ou des cinémas.

331. Néanmoins, toutes les personnes interrogées qui appartiennent à ce que nous avons défini comme étant les plus riches (N=9) ont déclaré fréquenter un lieu de culte, ce qui n'est pas le cas des autres classes. La comparaison à Tartuffe de Molière est tentante, mais loin de remettre en doute les réponses des interviewés qui sous le fait de désirabilité sociale aurait tendance a être plus dévot que la moyenne, un biais d'échantillonnage n'est pas à écarter.

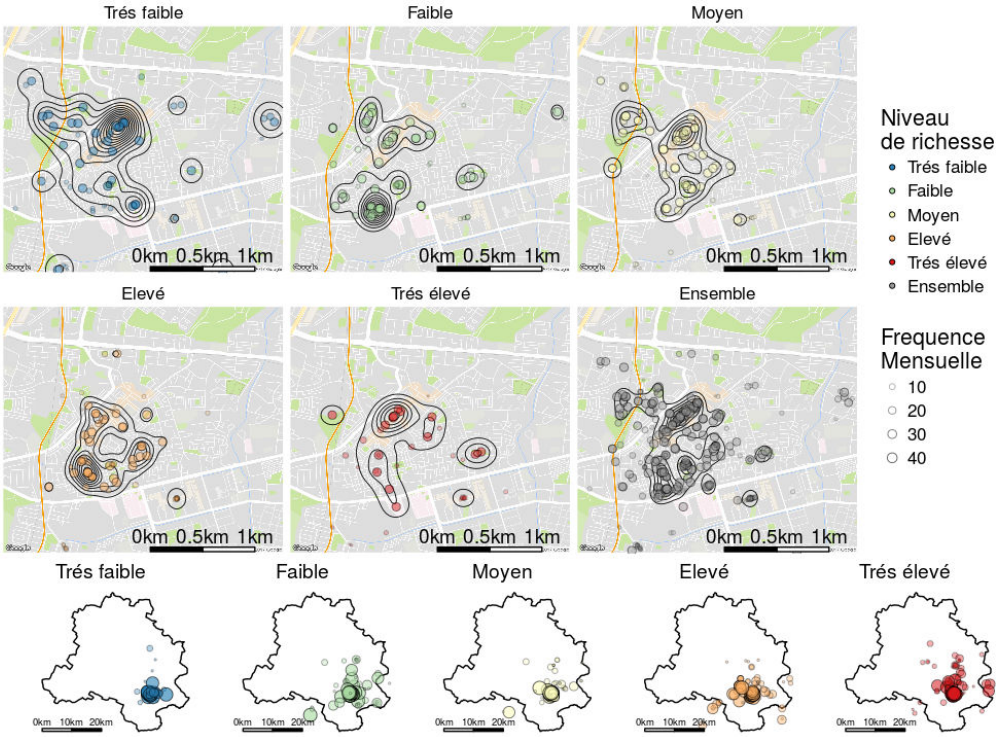


FIGURE 118 Lieux fréquentés (sauf le domicile) en fonction de notre estimation du niveau de richesse. L'encart du haut correspond à Malvia Nagar, celui de bas à l'ensemble de la ville. La taille des cercles et le niveau de transparence indiquent la fréquentation mensuelle, tandis que les isolignes les densités de fréquentation.

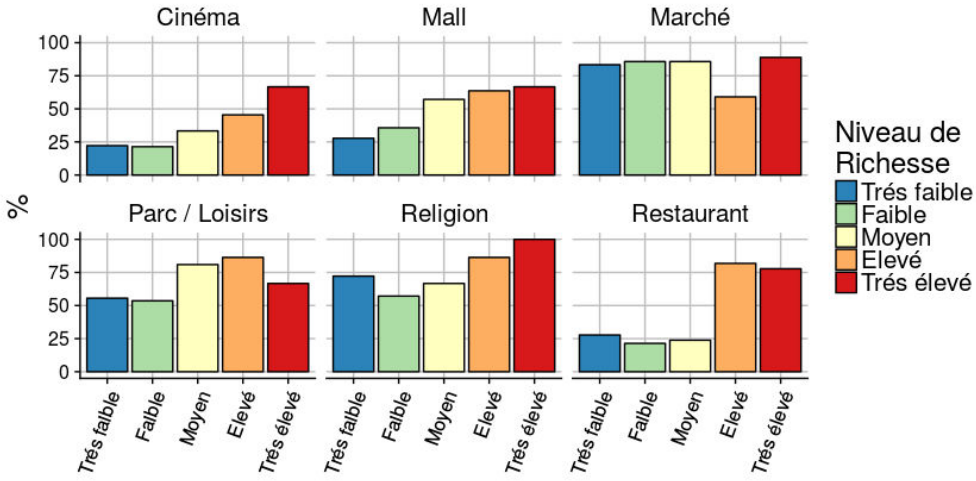


FIGURE 119 Part de chaque groupe socio-économique déclarant effectuer une activité donnée.

Lorsque l'on demandait aux personnes si elles mangeaient parfois au restaurant, et mis à part le sans-abri, la plupart répondaient globalement : *"I eat home food only, it's healthy"*.

Puis, en insistant un peu (*"But sometimes, you go to the restaurant ?"),* elles finissaient par compléter leur réponse. Il convient de préciser que les tarifs des restaurants suivent un éventail extrêmement large, allant de la *street-food* à quelques dizaines de roupies (20 à 30 roupies pour 12 *momos* végétariens) et qui est de fait très abordable, en passant par des chaînes de restauration qui produisent des ersatz de pizza ou de burger à moins de 200 roupies, à des restaurants plus onéreux, dont les plats dépassent facilement les 300/400 roupies et peuvent parfois atteindre les 1000 roupies pour le Panschilla Rendez-vous.

À Delhi, les *malls* ne sont pas non plus que des lieux de consommation, mais également des lieux de chalandise et de promenade (Cebeillac et Rault, 2016)(Rault *et al.,* 2018) Ainsi, ce n'est pas parce que les personnes n'ont pas les moyens de consommer qu'elles ne peuvent pas faire du lèche-vitrine, ou y retrouver leurs amis (Rault *et al.,* 2018). Cette pratique de ce type d'espace explique pourquoi ces lieux sont fréquentés par toutes nos catégories. Néanmoins, nous observons un gradient assez marqué, inversement proportionnel au niveau de pauvreté, car bien que le shopping actif n'y soit pas obligatoire, cela reste un centre commercial où les tarifs sont plus élevés que la moyenne.

La figure 120 montre pour chaque groupe les lieux fréquentés localement en fonction des activités. Si nous nous focalisons sur les lieux du domicile et de l'activité principale, il ressort clairement que localement, les personnes les moins aisées ont tendance à venir travailler dans des zones plus riches en l'occurrence Malviya Nagar l'inverse n'étant pas observé [332]. Ceci peut s'expliquer assez facilement du fait (1) de la présence du *main market*, marché attractif où de nombreux commerçants de rue interrogés exercent leur travail, et (2) qu'il s'agit d'une zone plus aisée, ou les personnes les plus pauvres peuvent fournir des prestations de services aux classes médianes supérieures, qu'il s'agisse des femmes de ménage ou des chauffeurs.

Tous les groupes se croisent dans le marché principal de Malviya Nagar, moins dans le marché plus local de Hauz Rani, fréquenté principalement par des personnes résidant à proximité. Les parcs de Baghan Singh et ses voisins, pourtant de proportions modestes par rapport au parc de Begumpur accueillent toutes nos catégories, probablement du fait de leur localisation, en plein centre de Malviya Nagar, et entourés de zones résidentielles.

Si nous revenons aux problématiques de la transmission de la dengue, de ces coprésences spatiales que nous retrouvons dans les parcs, sur le marché principal ou chez les personnes ayant recours à une main d'œuvre pour des tâches domestiques peut découler des potentielles contaminations et transmissions inter-quartier. En effet, les quartiers pauvres ou riches offrent

332. Comme le souligne Bruno Cousin à propos des élites délhiites dans l'ouvrage collaboratif « *Ce que les riches pensent des pauvres* » dont une section fut écrite par Jules Naudet, chercheur au CNRS, spécialiste des élites indiennes : « Le fait de vivre exclusivement dans les beaux quartiers et dans quelques autres lieux soigneusement sélectionnés afin de minimiser autant que possible les interactions avec la masse des pauvres est perçu comme un comportement naturel et allant de soi. » https ://www.alterechos.be/nous vivons dans des societes tellement inegalitaires que les differences entre les riches et les pauvres vont de soi/

tous des gîtes larvaires, de natures différentes certes (bidons, pneus, ou divers conteneurs d'un côté, et coupelles de pots de fleurs de l'autre), mais qui laissent présager par exemple qu'une femme de ménage puisse être indirectement contaminée par une femme au foyer (et vice-versa), ce qui peut entraîner une importation de la maladie dans le quartier de domicile. Ces mouvements de va-et-vient sont donc susceptibles de propager les épidémies de dengues entre quartiers riches et pauvres. À noter que l'intérieur des *malls* sont des lieux où les risques de contamination par la dengue sont probablement relativement modestes du fait d'une température en général trop fraîche pour les moustiques et de l'absence de gîtes larvaires.

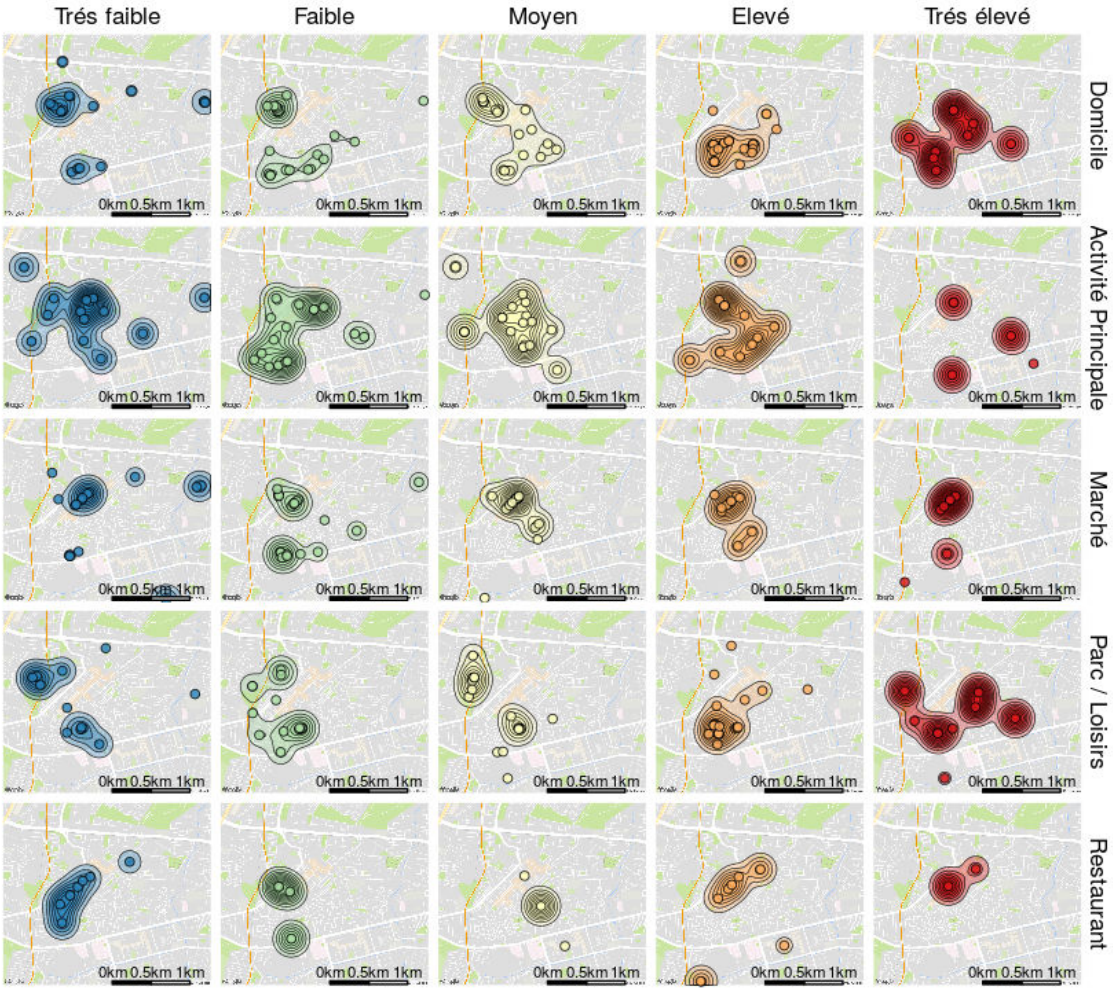


FIGURE 120 Localisation des activités effectuées à Malviya Nagar, selon le groupe socio-économique.

2.2.5 Analyse des mobilités

Après cette rapide analyse des co-présences, nous allons maintenant décrire les potentiels de mobilités, en mobilisant notamment diverses métriques : nombre de lieux fréquentés, le nombre d'activités effectuées, la distance au domicile, le rayon de giration et l'enveloppe

convexe. Le rayon de giration, comme vu dans le chapitre 5 est calculé ici en prenant le domicile comme point de référence et en pondérant les lieux en fonction des fréquences de visites plus un lieu est fréquenté régulièrement, plus il contribuera au rayon de giration. L'aire de l'enveloppe convexe, qui inclut l'ensemble des lieux fréquentés par un individu, est ici utilisée comme un autre indicateur de dispersion, et correspond à la superficie de l'emprise de l'espace d'activité.

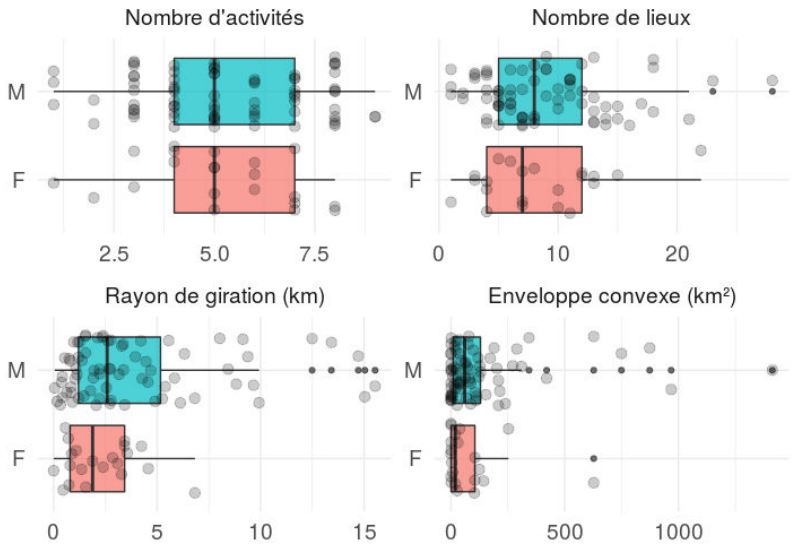


FIGURE 121 Différentes métriques de dispersions selon le genre. F pour les femmes, M pour les hommes.

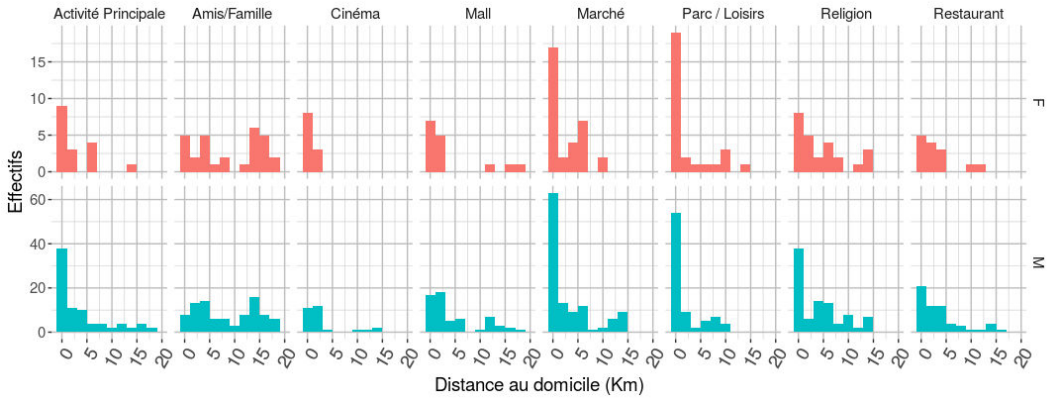
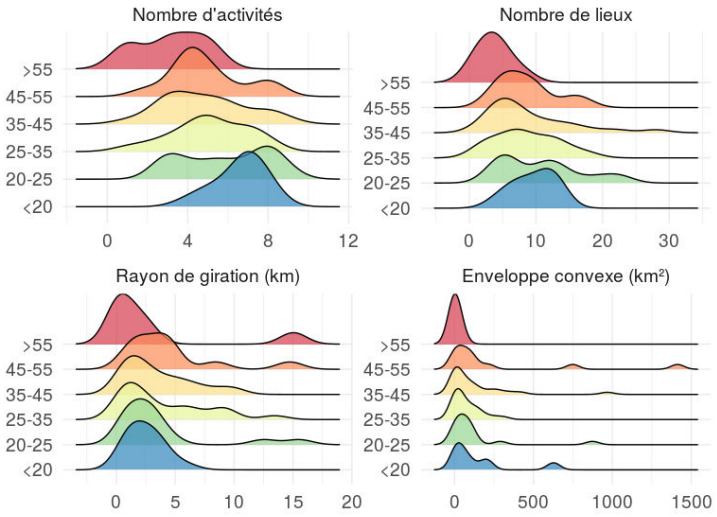


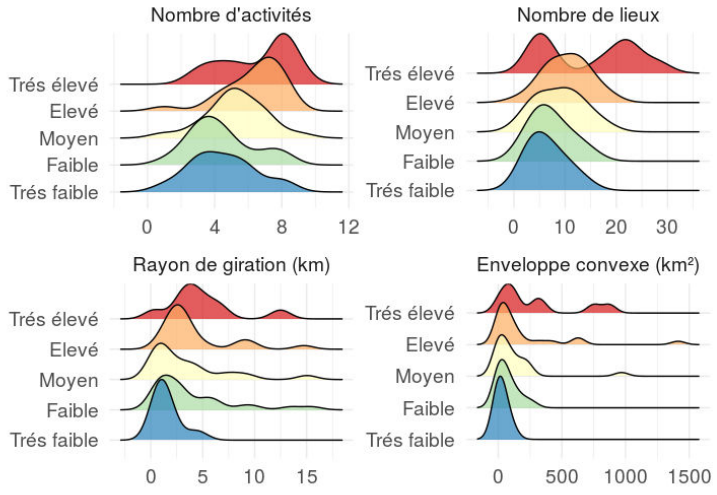
FIGURE 122 Distance parcourue pour effectuer une activité en fonction du genre.

La figure 121 montre par exemple ces différents critères appliqués aux femmes et aux hommes de notre échantillon. Il ressort que les différences entre ces deux groupes sont assez mineures, même si les femmes auraient tendance à avoir un potentiel de dispersion un peu moins important que les hommes (rayon de giration et enveloppes convexes globalement inférieures). Nous n'observons pas non plus de différences importantes dans les distances à parcourir pour

exercer une activité en fonction du genre (figure 122). Ceci peut s'expliquer soit par un biais d'échantillonnage, soit par le fait que la plupart des activités peuvent s'effectuer en couple ou en compagnie d'autres personnes. Nous n'observons pas non plus de différences très significatives entre les tranches d'âges (figure 123.a), si ce n'est que les personnes les plus âgées de notre échantillon ont en général moins d'activités, et sont moins mobiles pour la plupart (rayon de giration relativement faible).



(a)



(b)

FIGURE 123 Fonction de densité exprimant le nombre d'activités, de lieux, le rayon de giration et enveloppe convexe de l'échantillon en fonction (a) de l'âge des individus et (b) de leur niveau de richesse estimé.

Si nous regardons maintenant ces différentes métriques en fonction du niveau de richesse

estimé (figure 123.b), il apparaît que les personnes les plus aisées ont en général plus d'activités, un rayon de giration plus élevé et une enveloppe convexe plus importante que les personnes les plus pauvres. Nous pouvons néanmoins noter que les personnes les plus riches peuvent se scinder en deux groupes, l'un visitant un nombre restreint de lieux, l'autre un nombre très important d'endroits différents. Ceci défendant donc l'hypothèse d'un potentiel de mobilité plus important chez les classes sociales les plus aisées, même si certaines personnes considérées comme pauvres ou de niveau de richesse moyen doivent se déplacer sur de très longues distances, notamment pour se rendre à leur travail.

Pour aller au-delà de ces partitions selon le sexe, l'âge ou la classe sociale, et voir s'il n'y a pas des regroupements intrinsèques, nous avons réalisé une classification par nuées dynamiques (kmeans) en utilisant le nombre de lieux, le nombre d'activités et le rayon de giration comme variables en entrée. Ces valeurs sont centrées et réduites afin de pouvoir être comparées raisonnablement. Nous n'utilisons pas l'enveloppe convexe qui ne semble pas apporter plus d'information quant au potentiel de dispersion que le rayon de giration. Nous choisissons trois groupes, comme suggéré la méthode de la silhouette moyenne (annexe J).

La figure 124 présente les distributions des effectifs par groupe (de la classification) et par variable, selon une fonction de densité (intégrale égale à 1). Le groupe 2 concerne les personnes qui fréquentent peu de lieux et qui ont le plus faible potentiel de dispersion. Les membres du groupe 3 fréquentent globalement le plus de lieux et effectuent le plus d'activité, avec un rayon de giration un peu supérieur aux membres du groupe 2. Le groupe 1, finalement, concerne les personnes les plus mobiles (rayon de giration très élevé). Mais la distribution du nombre de lieux et d'activité est assez étalée, ce qui ne permet pas de dégager de tendances quant à ces critères de mobilités.

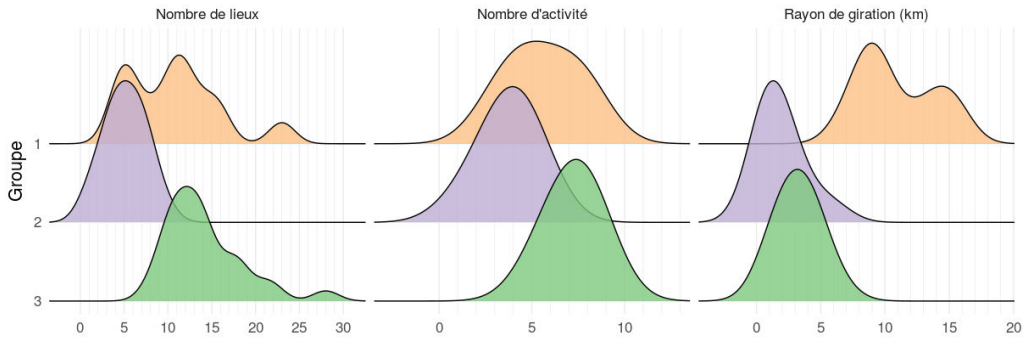


FIGURE 124 Fonction de densité représentant les tendances de déplacements des différents groupes de la classification.

Si nous regardons maintenant la répartition des tranches d'âges, des genres et des classes (niveaux de richesses) dans chacun de ces groupes (figure 125) répartis en pourcentage selon

les groupes de la classification nous pouvons faire différents constats. Tout d'abord, c'est dans le groupe 2, qui concerne les personnes les moins mobiles où la part de femme est la plus importante. Concernant les niveaux de richesses, nous observons une tendance décroissante, dans le sens où plus le niveau de richesses augmente (jusqu'à "élevé"), moins il y a de personnes qui appartiennent à ce groupe. C'est dans le groupe 3, défini par un grand nombre de lieux visités et d'activité réalisée, avec une dispersion plus importante que chez les membres du groupe 2, que l'on retrouve les parts les plus importantes de personnes de niveau de richesse *élevé* à *très élevé* (entre 40 et 50 % des effectifs de ces classes). Le groupe 1, qui regroupe les personnes qui se déplacent sur de plus grandes distances sans nécessairement effectuer un grand nombre d'activités concerne plus les hommes que les femmes, et plus les personnes âgées de plus de 20 ans que les plus jeunes.

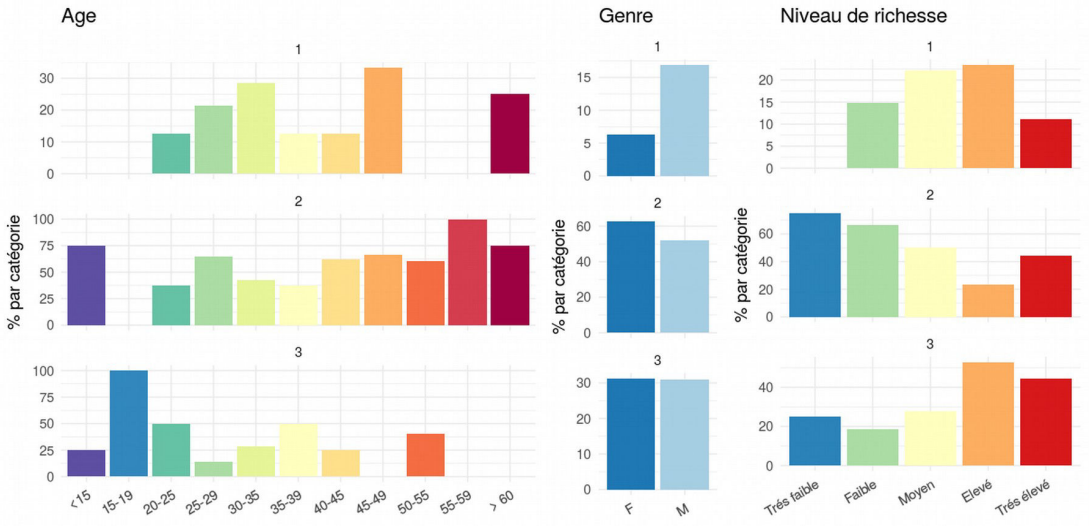


FIGURE 125 Caractéristiques socio-démographiques des différents groupes de la classification.

Finalement, cette classification complète la figure 123.b (profils de différentes métriques de déplacement en fonction niveau de richesse) et suggère que bien qu'il semble y avoir un lien entre potentiel de mobilité, genre, âge et classe sociale, ce lien n'est ni direct ni systématique et doit probablement dépendre des habitudes, des modes de vie, du réseau social, ainsi que des contraintes liées à la localisation du travail.

2.2.6 Synthèse de l'enquête

La nature des différents quartiers de Malviya Nagar implique de fortes ségrégations sociales. Néanmoins, l'analyse de nos questionnaires a montré qu'il existe différents types de lieux où l'on peut observer des coprésences (1) multidirectionnelles, où toutes les classes sociales peuvent se retrouver, tels que les parcs, les marchés ou les *malls*, mais aussi (2)

monodirectionnelles, comme les domiciles des personnes aisées qui emploient du personnel vivant dans des quartiers plus précaires.

Les lieux fréquentés et les capacités de dispersions des interviewés sont étroitement liés au capital économique d'un individu, mais un niveau de pauvreté élevé n'implique pas pour autant une mobilité réduite dans la ville. Pour les personnes les plus défavorisées, les déplacements sur de longues distances sont surtout contraints par la localisation de leur travail ou motivés par des raisons sociales (visiter les proches) ou religieuses. Les mêmes aspects peuvent s'observer chez des personnes plus aisées, avec en plus une facilitée de déplacement pour atteindre des zones de loisirs assez éloignées.

Nous sommes néanmoins conscients que notre échantillonnage n'est pas représentatif de la population, que les réponses fournies par les interviewés sont de qualités très différentes, plus ou moins parcellaires et biaisées à différents niveaux (spatial et temporel). De plus, les écarts de potentiels de mobilité entre homme et femme ne sont pas aussi nets dans nos données qu'ils semblent l'être en réalité. Mais notre étude pilote renforce néanmoins l'idée d'une très grande hétérogénéité dans les mobilités quotidiennes à Delhi, dont les aspects socio-économiques individuels forment un déterminant majeur, non unique. De plus, les lieux de coprésence socio-économiques sont un paramètre crucial dans la diffusion et la propagation de la dengue.

Un des objectifs de la présente thèse est d'évaluer le potentiel de différentes sources de données en vue d'une modélisation des mobilités individuelles selon une approche à base d'agents, orientée sur la localisation et la temporalité des activités effectuées. Si notre enquête nous a permis de montrer quelques aspects et tendances de mobilités dans un quartier de Delhi, nous savons pertinemment que notre échantillon ne peut servir de base à une génération d'individus de synthèse aux caractéristiques socio-économiques et aux mobilités statistiquement représentatives de l'ensemble de la population. Mais d'un point de vue purement méthodologique, il parait judicieux de voir dans quelles mesures ces données récoltées sur le terrain peuvent permettre de reconstruire des agendas individuels avec une part de stochasticité pour combler les différentes lacunes de notre corpus notamment le jour de réalisation et la durée d'une activité.

3 Des données terrain à des agendas individuels

Les informations que nous avons recueillies sur le terrain ne sont pas continues dans le temps. Nous n'avons en effet que des lieux associés à des activités et des fréquences de visites plus ou moins précises. Une première étape est donc d'arriver à reconstituer des agendas, où pour chaque tranche horaire, chaque individu effectue une activité dans un lieu donné. L'objectif est donc d'avoir des séquences temporelles d'activités localisées, un peu à la manière d'une enquête ménage déplacement sauf que dans notre cas le pas de temps sera plus important, mais la fenêtre temporelle plus grande car s'écoulant sur plusieurs jours.

3.1 Création des agendas

3.1.1 Harmonisation des plages horaires

La première limite de nos données recueillies provient du fait que toutes les personnes interrogées n'étaient pas en mesure d'être très précises sur les plages horaires durant lesquelles elles effectuent leurs activités. Autant certaines personnes pouvaient fournir ces informations heure par heure, jour par jour, autant la plupart d'entre elles se contentaient d'approximations : plutôt en matinée, en journée, en soirée, le matin et le soir, etc. Nous allons donc poser quelques hypothèses en définissant des plages horaires bornées en lien avec les informations fournies (tableau 10). Pour des raisons de simplicité, nous choisissons un pas de temps d'une heure.

Plages horaires	Horaire de début	Horaire de fin
Journée	8 h	18 h
Matin	8 h	12 h
Midi	12 h	14 h
Après Midi	14 h	18 h
Soir	18 h	22 h
Nuit	23 h	6 h
Dîner	18 h	22 h

Tableau 10 Approximation des moments de la journée en plage horaire. Par exemple, lorsqu'une personne déclare fréquenter un lieu le matin, nous poserons qu'il peut le fréquenter entre 8h et 12h.

Une autre limite est que bien que les personnes aient renseigné la fréquence de visite de chacun des lieux, les jours de visite ne sont pas systématiquement donnés. Nous posons donc l'hypothèse que les activités effectuées 5,6 et 7 fois par semaine se déroulent respectivement du lundi au vendredi, du lundi au samedi et tous les jours de la semaine.

Les temporalités journalières des activités de chaque individu sont dupliquées sur une semaine, et sont ensuite divisées par la fréquence de visite hebdomadaire. Par exemple, si une personne déclare aller au marché le soir 4 fois par semaine, les tranches horaires 18-22h se verront attribuer une probabilité de 4/7 d'être associée à l'activité marché. La somme des parts d'activité par tranche horaire n'est pas systématiquement égale à 1. Par exemple si la même personne ne déclare pas d'autres activités le soir, dans ce cas nous posons que cette personne est à son domicile, selon une probabilité égale à 1 4/7, soit 3/7. Si la personne déclare d'autres activités aux mêmes plages horaires, avec des fréquences hebdomadaires assez élevées, comme se balader dans un parc et rendre visite à des proches respectivement 2 et 3 fois par semaine, nous appliquerons alors un produit en croix de telle sorte que la somme des probabilités d'effectuer une activité à ces horaires donnés soit égale à 1.

Cette rapide correction de notre corpus nous permet d'obtenir des agendas spatialisés où chaque personne a une probabilité par tranche horaire d'effectuer une activité dans un lieu donné, illustré par les figures 126 et 127. Ces dernières se divisent en deux parties : la première, à gauche correspond à l'agenda sur une semaine type, tandis que la partie de droite montre les lieux fréquentés la taille des cercles correspond à la fréquence de visite dans le quartier et dans la ville (encadré). La figure 126 montre des agendas pour deux personnes ayant une activité professionnelle. La personne dont l'identifiant est 81 (en haut) habite dans le quartier aisé de Shivalik et fréquente 5 jours par semaine une université privée à Noida, dans l'est de Delhi. Elle se rend assez régulièrement dans divers *malls* et cinéma de la ville, et rend visite à ses connaissances, mais sans percevoir de réelle routine ("it depends"). Elle va également assez souvent au marché et au parc (environ 10 fois par mois), plutôt le soir pour le parc, et à des horaires très variables pour le marché. D'après cette représentation graphique, cet interviewé a environ 70 % de chance d'effectuer son activité principale dans les heures de journées les jours de semaines, et 20 % de chance de se trouver au marché.

La personne numéro 33 travaille tous les jours de la semaine et va au restaurant matin midi et soir et habite dans une zone qui dépend de la colonie de catégorie A Panschilla Park. Ces éléments sont à re-contextualiser car cet homme est ramasseur d'ordures, il habite sous le fly-over et mange dans des restaurants de rue parce qu'il n'a pas de domicile. Il n'a pas les moyens d'avoir des activités de loisirs payantes, et son espace d'activité est restreint à Malviya Nagar.

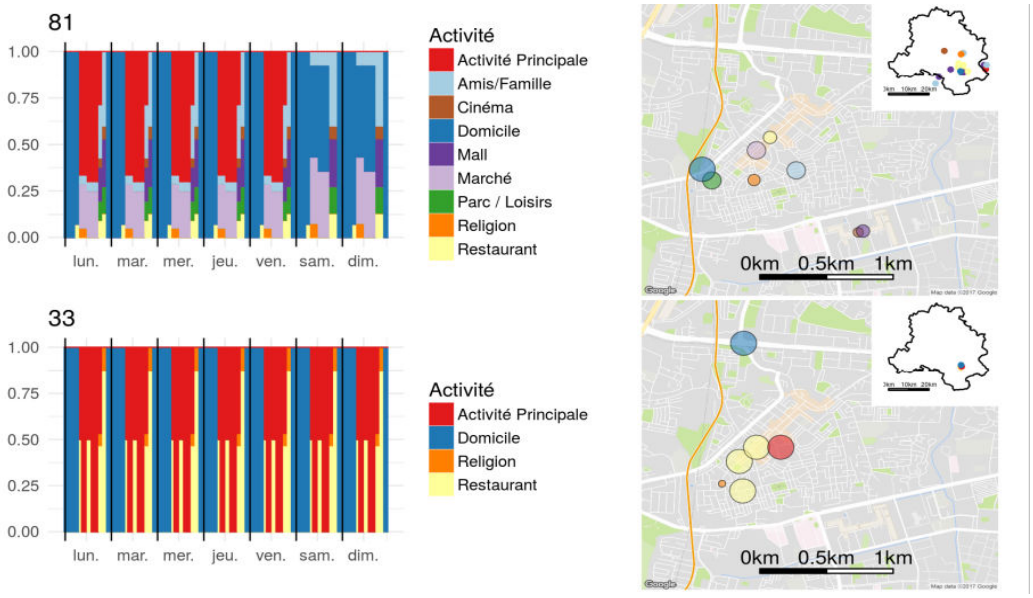


FIGURE 126 Probabilité d'effectuer une activité à une tranche horaire donnée (gauche), et localisation de l'espace d'activité (droite). Sont représentées ici deux personnes ayant une activité professionnelle.

La personne 86 (figure 127) est une femme au foyer de 18 ans qui habite dans le *slum* de Begumpur, et reste principalement chez elle, sauf le soir où elle va au marché et parfois flâner dans le *mall* de Saket ou se promener dans un parc. Son agenda (comme beaucoup d'autres) est caractérisé par une interchangeabilité des activités quotidiennes, dans le sens où tous les jours de la semaine se ressemblent.

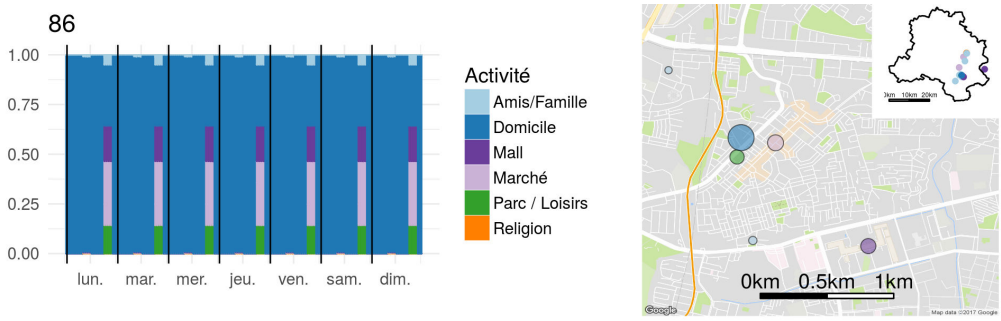


FIGURE 127 Probabilité d'effectuer une activité à une tranche horaire donnée (gauche) et localisation de l'espace d'activité (droite) d'une personne n'ayant pas d'activité rémunérée.

Ces agendas permettent d'envisager une reconstruction plausible des lieux fréquentés par tranche horaire, en fonction de la probabilité d'exercer une activité dans un lieu à un instant donné. Néanmoins, cette approche souffre de quelques lacunes, liées à la durée d'une activité qui est parfois peu réaliste (voir section suivante), et ne repose pas sur les aspects séquentiels du déroulement des activités (possibilité de va-et-vient entre 2 activités).

Une autre limite de l'utilisation des résultats des figures 126 et 127 dans une optique de modélisation des mobilités est que chaque visite d'un lieu est généralement équiprobable au cours de la semaine. Si une personne déclare aller au temple une fois par semaine, une approche par tirage aléatoire entraîne qu'elle aura une chance sur sept d'effectuer cette activité un jour donnée aux horaires définis. Ceci implique qu'elle peut faire cette activité plusieurs fois en une semaine ou aucune fois, selon les tirages, ce qui ne correspond pas vraiment au vécu de la personne. Et si une personne déclare aller une fois par mois au cinéma, il y a très peu de chance que cette activité soit effectivement réalisée un jour donné (1/28).

Il convient donc de définir une méthode qui permette d'affecter un jour où se réalise une activité flexible, et de définir de manière plus réaliste les plages horaires des activités.

3.1.2 Affectation d'un jour pour effectuer une activité

Nous décidons de raisonner sur un mois, selon une méthode relativement simple, en posant que si l'activité est régulière (fréquence hebdomadaire ≥ 1) elle se produira toutes les semaines. Nous utilisons un processus en plusieurs étapes pour les activités effectuées de manière plus anecdotique (fréquence hebdomadaire <1). La probabilité qu'un lieu soit visité la première semaine est égale à sa fréquence hebdomadaire. Si le lieu est choisi, la probabilité qu'il soit aussi fréquenté la semaine suivante décroît, et inversement s'il n'a pas été sélectionné selon la formule :

$p_i = p_{i-1} \pm (1 - p_{i-1})/(4 - i)$ avec p_i la probabilité de réaliser une activité une semaine i, avec i entre 2 et 4 (p_1 étant la fréquence de visite d'un lieu fournie par la personne interviewée).

Par exemple si une personne va dans un marché donné deux fois par mois, il a une chance sur deux d'y aller la première semaine. S'il y va, il n'aura plus qu'une chance sur trois d'y aller la semaine suivante. S'il n'y va pas en semaine 2, il aura une chance sur deux d'être sélectionné pour la semaine 3, et dans la négative il sera visité en semaine 4. Les quatre semaines sont ensuite mélangées de manières aléatoires, afin d'éviter que les dernières semaines concentrent les activités avec le moins de probabilité d'être effectuée.

Les semaines où les différents lieux sont visités étant définies, il nous reste à choisir le(s) jour(s). Pour cela, nous partons du principe que les activités autres que principales et secondaires ont plus de chance d'être effectuées en week-end. Nous posons ici que les jours de week-end ont deux fois plus de chance d'être tirés que les jours de semaines, ce qui revient à affecter une probabilité de réalisation de 1/9 pour les jours de semaines, contre 2/9 pour le samedi et le dimanche. De telles activités ont donc 5/9 d'être effectuée la semaine et 4/9 un week-end. Il est toutefois envisageable de reconsidérer par la suite ces affectations un peu abruptes, et d'augmenter par exemple les probabilités de réalisation un week-end. Dans tous les

cas, il conviendra de prendre en compte plus précisément les jours de la semaine où les activités sont réalisées lors de prochaines enquêtes de terrains.

On affecte ensuite un jour dans une semaine donnée pour les lieux visités occasionnellement (fréquence hebdomadaire <1), et un nombre de jours correspondant à la fréquence de visite hebdomadaire pour les lieux visités plus fréquemment (fréquence hebdomadaire ≥1).

Dans le cas de lieux aux fréquences de visites impaires (par exemple une personne qui va dans un parc 10 fois par mois, soit 2,5 fois par semaine) nous tirons aléatoirement pour chaque semaine l'arrondi supérieur ou inférieur. Dès qu'un jour de visite est affecté à un lieu, la probabilité qu'un lieu suivant choisisse le même jour baisse de 50 %, afin de répartir la fréquentation des lieux plus équitablement dans la semaine. Les probabilités sont ensuite réinitialisées au début de chaque semaine.

Cette section peut aussi se formuler sous forme de pseudo-code, avec A le type d'activité associée au lieu, F la fréquence hebdomadaire de visite du lieu ; S_i, la semaine de visite, avec i entre 1 et 4 ; Ps_i la probabilité que le lieu soit visité la semaine i, et J, le jour de la semaine avec Pd_j la probabilité que le lieu soit visité un jour j, avec j dans J {lundi, mardi, mercredi, jeudi, vendredi, samedi, samedi, dimanche, dimanche}.

- Si A = Activité principale
  - Semaines de réalisation : S_1,S_2,S_3,S_4
    - Jours de réalisation :
      - si F = 7 : {lundi, mardi, mercredi, jeudi, vendredi, samedi, dimanche}
      - si F = 6 : {lundi, mardi, mercredi, jeudi, vendredi, samedi
      - si F = 5 : {lundi, mardi, mercredi, jeudi, vendredi}
      - si F < 5 : On tire F parmi {lundi, mardi, mercredi, jeudi, vendredi, samedi, dimanche}
- si F ≥ 1 & A != Activité principale
  - Semaines de réalisation : S_1,S_2,S_3,S_4
    - Jours de réalisation D : on tire F fois *j* avec une probabilité de Pd_j
      - À chaque tirage : Si F >1 --> $Pj_j = Pj_j/2$

- si F < 1 & A != Activité principale
  - Semaine de réalisation :
    - pour i dans {1,2,3,4}, on tire {réalisation ; non réalisation} avec une probabilité respective de {Ps_i ; 1-Ps_i}
      - Si "réalisation" ==> S_i & Ps_{i+1}= Ps_i - (1-Ps_i)/(4-i)
      - Si "non-réalisation" ==> Ps_{i+1}= Ps_i + (1-Ps_i)/(4-i)
  - Jour de réalisation dans une semaine lorsque S_i = réalisation
    - on tire un jour D dans J selon Pj_j

3.1.3 Estimation des durées et des plages horaires des activités

Les activités comme faire le marché, aller voir des proches, ou se rendre sur les lieux de cultes, peuvent se dérouler sur des durées très variables. Pour remédier à cela, nous ajoutons des contraintes sur les durées des activités, en posant qu'aller au restaurant ne prend pas plus de 3 h, tout comme faire le marché. Les films indiens peuvent être assez longs, et nous définissons donc un intervalle de 2 à 5 h pour le cinéma, ce qui permet de prendre en compte les temps des activités connexes, comme discuter ou manger. Rendre visite à des proches peut également s'étaler sur un large spectre horaire, allant de la simple visite de courtoisie à une journée entière (tableau 11).

Activité	**Durée minimale**	**Durée maximale**
Restaurant	1h	3h
Marché	1h	3h
Cinéma	2h	5h
Religion	1h	4h
Visiter des amis	1h	12h
Parc / Loisirs	1h	4h
mall	1h	4h

Tableau 11 Durée minimale et maximale pour chaque activité.

De plus, aller faire des courses en semaine au marché d'à côté ne prend pas autant de

temps que d'aller dans un marché plus important et/ou spécialisé dans le week-end. De même, se rendre à un lieu de culte majeur est souvent lié à des cérémonies religieuses importantes, et de fait, la durée sur place peut se voir allongée par rapport à un office plus routinier. Il convient donc de trouver une méthode pour pondérer ces durées selon des critères plausibles. Nous pourrions prendre en compte la fréquence d'une activité et la distance au domicile, en partant du principe que plus cette dernière est rare et éloignée, plus la durée de cette dernière se rapprochera de la durée maximale évoquée dans le tableau 11, et inversement pour les activités proches du domicile. Pour des raisons de simplicité, nous ne prendrons en compte que la distance (x) et poserons que la durée (y) d'une activité suit une loi de probabilité concave sur $[0;+\infty[$ de type $y \sim 1/x + x^2$, bornée par la durée minimale et maximale de l'activité définie dans le tableau 11. Cette fonction, choisie de manière assez arbitraire, présente l'avantage d'être d'abord décroissante sur un court intervalle et devient fortement croissante ce qui permet d'appliquer une plus grande probabilité d'effectuer une action de courte durée pour les lieux les plus proches du domicile, et inversement pour les lieux les plus éloignés. La figure 128 ci-dessous illustre cette augmentation de la probabilité de tirer une durée plus longue en fonction de la distance, ce qui permet aussi de prendre en compte de manière indirecte les temps de transports non récoltés lors de nos interviews.

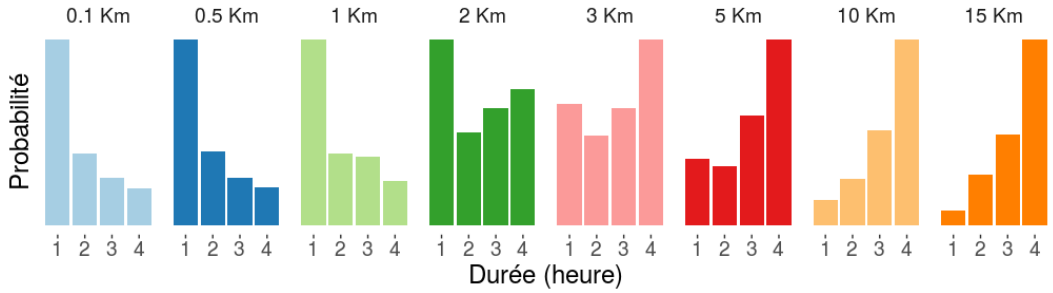


FIGURE 128 Exemple théorique montrant pour une activité de durée maximum de 4h, la probabilité de tirer une durée (entre 1 et 4h) en fonction de la distance x au domicile, suivant une fonction du type $y=1/x+x^2$, après 1000 itérations.

Un pseudo-code pourrait se formuler de la manière suivante, dur_{min} et dur_{max} les durées minimales et maximales d'une activité posée dans le tableau 11, et définie selon la distance *dist* au domicile :

- On définit une séquence x de longueur dur_{max} - dur_{min}, comprise entre 0 et *dist*, et triée par ordre croissant.
- On applique la fonction $1/x+x^2$ à cette séquence, permettant de définir une série de probabilités.
- On choisit une durée T, comprise entre dur_{min} et dur_{max}, selon les probabilités définies

précédemment[333].

Il convient ensuite d'affecter une heure de début à ces activités. À partir de ces estimations de durée, et connaissant la plage horaire, nous posons d'abord que si la durée tirée est supérieure à la largeur de plage horaire, l'activité se réalisera sur l'ensemble des heures attribuées par le tableau 11. Par exemple si une personne déclare visiter des amis, mais uniquement l'après-midi (14h-18h), alors que nous avons tiré une durée de 8 h, cette activité s'effectuera entre 14 h et 18 h, et non entre 14 h et 22 h. En revanche, si la durée de l'activité tirée est inférieure à la largeur de la plage horaire, l'heure du début de l'activité sera choisie aléatoirement parmi les heures comprises entre le début et la fin de la plage horaire, moins la durée de l'activité tirée. Par exemple, si une personne déclare se rendre au marché local en journée (entre 8 h et 18 h), et que nous tirons une durée de 2 h, l'heure du début de l'activité sera choisie de manière aléatoire entre 8 h et 16 h.

Nous considérons ici l'activité principale comme un cas particulier, dans le sens où cette activité est *a priori* fixe dans le temps et l'espace. Pour les personnes dont nous n'avons pas de tranches horaires de références, nous posons que cette activité se déroule entre 8 h et 20 h. Nous posons ensuite que la durée de l'activité est comprise entre les deux tiers de la durée maximale et la durée maximale (soit entre 8 et 12 h). Nous appliquons ensuite à ces bornes notre fonction de distance, comme précédemment, sauf que nous prenons la valeur moyenne sur 28 itérations (correspondant à 4 semaines). De même, nous estimons 28 heures de débuts d'activité et choisissons la valeur moyenne. Ainsi, lors de chaque simulation sur un mois, la durée et la plage horaire de l'activité principale reste fixée.

Finalement, les plages horaires où aucun lieu n'est défini sont associées au domicile de l'interviewé.

3.1.4 Gestion des redondances

Cet algorithme n'empêche cependant pas l'apparition de redondances, c'est-à-dire que des lieux distincts peuvent être visités au même moment. Nous pourrions tenter d'optimiser la chose, en faisant par exemple tourner des boucles jusqu'à ce qu'aucun doublon ne soit observé, mais cette méthode "dure" risque d'impliquer des boucles infinies, notamment si une personne a déclaré visiter beaucoup de lieux avec des fréquences de visites trop importantes,

333. Fonction qui peut s'écrire sous R :

prob=runif(min=0,max=d,n=c(dur_{max},dur_{min}+1)) %>% sort # on tire autant de valeur que la durée entre 0 et la distance d

prob=1/prob+prob^2 # on applique notre fonction y~1/x+x^2

sample(dur_{min} :dur_{max},1,prob=prob) #on sélectionne une valeur avec la probabilité définie précédemment, avec dur_{min}, dur_{max}, et d respectivement les durées minimales, maximales, et la distance du lieu ou s'exerce l'activité au domicile. Les librairies des fonctions les plus pertinentes seront mises à disposition dans les semaines / mois qui suivront, lorsque ces dernières seront rendues plus génériques.

ce qui peut entraîner des incohérences. Nous pourrions également faire quelques ajustements, en testant au préalable si une plage horaire est libre ou non avant de l'attribuer. Nous aurions également pu faire un modèle bayésien, en définissant une *prior* pour chaque lieu pour chaque individu. À noter que de très bons algorithmes génétiques permettent aussi de générer un ensemble de combinaisons possibles à partir de contraintes initiales (ici les fréquences de visites, les durées et jours des activités), comme la plateforme OpenMole[334] qui mériterait d'être testée ultérieurement. Il serait également envisageable d'utiliser un solveur prolog (Carlsson *et al.*, 1997) afin de vérifier la cohérence de l'agenda, en posant des règles strictes (pas de chevauchement d'activité, déroulement aux bonnes tranches horaires, etc.), et réduire éventuellement sa complexité (Banos *et al.*, 2006).

Nous avons choisi ici la simplicité, et posé des règles simples : si l'activité principale, dont les plages horaires ne sont pas toujours connues, est en "conflit" avec une autre activité, cette dernière est prioritaire. Ceci permet de reproduire quelques ruptures dans les activités, comme prendre une pause pour faire des courses au marché, ou sortir plus tôt de son travail. Si des activités autres que principales sont attribuées aux mêmes plages horaires, nous choisissons au hasard un des lieux et l'attribuons à l'ensemble des plages horaires successives où le conflit est observé.

3.1.5 Quelques exemples d'agendas de synthèses

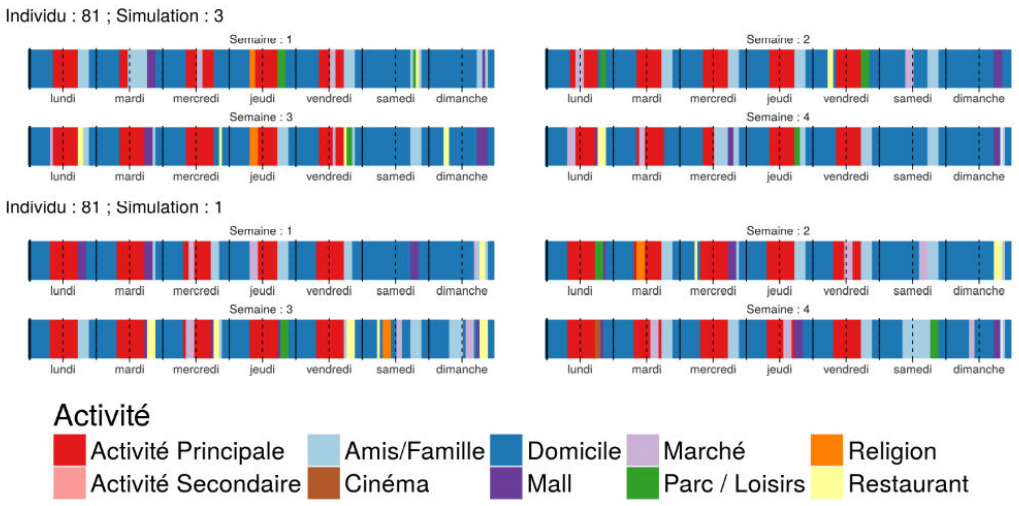


FIGURE 129 Exemple d'agenda reconstitué pour un interviewé (81).

Les figures 129, 130 et 131 montrent deux simulations d'agendas générés sur un mois pour trois personnes (81, 33 et 86). L'interviewé 81 déclarait exercer beaucoup d'activités dans différents lieux et ceci de manière assez régulière, ce qui transparaît sur les différentes simulations, avec le maintien d'une activité principale les jours de semaines même si certains

334. https ://www.openmole.org/

jours il ne travaillera que le matin.

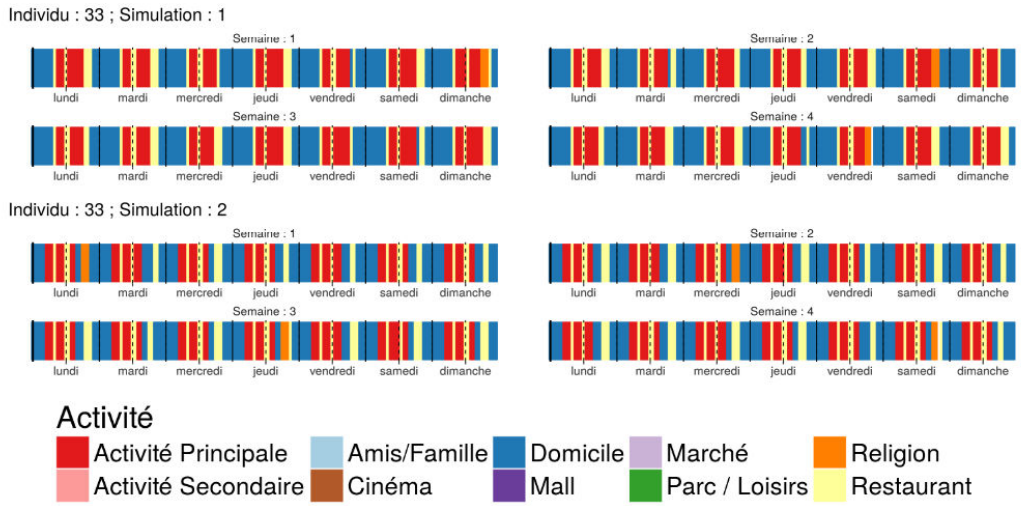


FIGURE 130 Exemple d'agenda reconstitué pour un interviewé (33).

La personne 33 effectue peu d'activité (travail, restaurant et visite de lieux de culte), son agenda est donc relativement stable, avec néanmoins quelques variantes selon les jours et les semaines. À noter que les durées au restaurant sont probablement exagérées, compte tenu du fait que la personne mange dans la rue. La figure 131 montre les différents agendas reconstitués pour une personne n'ayant pas déclaré exercer beaucoup d'activités différentes et restant la plupart du temps à son domicile. Elle va cependant au marché local 1,5 fois par semaine et 3 fois par mois au marché de Sarojini et au parc une fois par semaine. Ces caractéristiques ressortent bien dans les différents agendas générés.

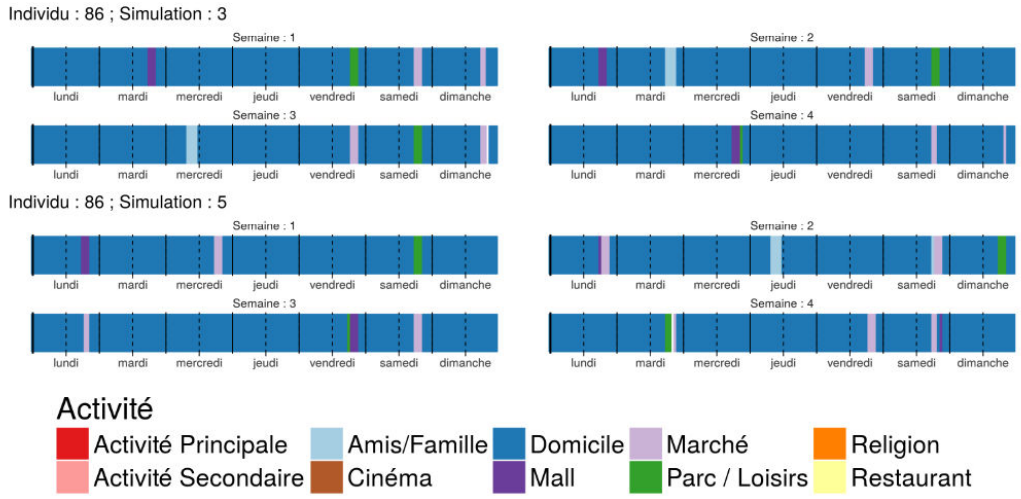


FIGURE 131 Exemple d'agenda reconstitué pour un interviewé (86).

En somme, à partir d'informations partielles sur les durées, fréquences et horaires de visite

d'un lieu donné, nous avons généré des agendas relativement crédibles en posant des hypothèses simples. L'approche stochastique employée permet de générer des séquences d'activités variables selon les jours et les semaines, permettant de rendre compte de la flexibilité dans la réalisation de certaines activités. Quelques détails peuvent être améliorés, notamment dans l'attribution de plages horaires uniques à un lieu pour éviter des redondances, notamment lorsqu'un individu déclare exercer un grand nombre d'activités, avec des fréquences hebdomadaires possiblement surestimées. Il conviendrait également de prendre en compte la distance entre les lieux fréquentés afin d'améliorer la robustesse théorique des agendas, en estimant le potentiel de réalisation d'une activité en prenant par exemple en compte la durée d'une activité et de la distance à l'activité précédente. Il est aussi envisageable d'apporter ultérieurement des améliorations aux différentes étapes de la reconstruction de ces agendas.

3.2 Analyse des agendas générés et perspectives pour la simulation à base d'agents

Ces agendas reconstruits, il est maintenant possible d'aller au-delà d'une simple analyse comparative des nombres de lieux fréquentés et de paramètres de dispersions en fonction de différents groupes, en étudiant notamment les séquences temporelles des activités. Un des objectifs est ici d'explorer différentes approches qui pourraient permettre une génération d'agendas de synthèse inspirés de ces agendas reconstruits.

Nous pouvons envisager par exemple d'effectuer diverses analyses et classifications en fonction des séquences de déroulement d'activités, afin de créer des sous-groupes d'où nous pourrions extraire des informations spécifiques que nous appliquerons à des agents de synthèses. Mais les méthodes classiques de classification n'ont pas été prévues pour des données séquentielles, et les analyses factorielles fournissent des résultats assez limités dans ce type de contexte (Lesnard et de Saint Pol, 2006). Nous allons étudier ici des méthodes d'appariement optimal qui permettent de comparer les degrés de similarité de séquences. Nous extrairons ensuite des taux de transitions qui nous permettront d'obtenir des chaînes markoviennes qui présentent l'intérêt de définir pour chaque individu et à chaque pas de temps une probabilité de changer d'activité. Ces dernières sont utilisées depuis longtemps dans les modélisations orientées sur les activités (Marble, 1964 ; Pappalardo et Simini, 2017a ; Perkins *et al.*, 2014).

3.2.1 Regrouper des individus selon les similarités de leur agenda

Utilisées d'abord en bio-informatique, notamment pour analyser les séquences d'ADN, les méthodes d'appariement optimal furent ensuite appliquées à la socio-démographie, notamment dans l'analyse de trajectoires professionnelles et autre parcours de vie (Lesnard, 2006). La première étape consiste à calculer un niveau de similitude entre les séquences, ce qui revient

pour chaque personne de l'échantillon à calculer le plus petit nombre d'activités par heure qu'il faudra changer pour que son agenda soit identique à celui d'un autre individu. Si nous prenons par exemple 2 séquences sur 5 heures, A, B et C étant des activités réalisées par les individus 1 et 2 (tableau 12) :

Heure	1	2	3	4	5
Individu 1	A	A	B	B	A
Individu 2	A	B	B	C	A

Tableau 12 Exemple de séquences d'activités (A, B et C) pour deux individus (1 et 2).

Plusieurs transformations sont possibles afin de rendre ces deux séquences égales. Nous pouvons par exemple remplacer les activités de l'utilisateur 1 réalisées aux heures 2 et 4 par B et C, ou celles de l'utilisateur 2 par A et B. Nous pouvons aussi transformer l'activité de l'utilisateur 1 à l'heure 2 en B, et celle de l'utilisateur 2 à l'heure 4 en C, etc. L'idée est de trouver pour chaque séquence temporelle d'une personne de l'échantillon le nombre minimum de changements permettant d'obtenir une séquence temporelle d'un autre individu. Ceci permet d'obtenir une distance entre deux agendas, aussi appelé dissimilarité. Nous avons choisi aléatoirement une simulation de l'ensemble de nos agendas à laquelle nous avons appliqué la métrique « Optimal Matching » choix par défaut de la librairie *TraMineR* de R (Gabadinho *et al.,* 2011). Les distances entre chaque agenda sont ainsi calculées, permettant d'obtenir une matrice de dissimilarité entre tous les agendas. Il existe bien entendu un grand nombre de méthodes et de paramètres permettant de définir ces distances, qui donnent un poids plus ou moins important aux changements de séquences, avec une tolérance temporelle plus ou moins flexible. Par exemple le fait de passer du domicile à l'activité principale lors d'une plage horaire (plus ou moins large) peut être vu comme plus discriminant que de partir d'un marché pour aller visiter des amis. Dès lors, deux agendas qui ont un couple domicile / activité principale assez synchrone peuvent se voir définir des distances assez proches, malgré des différences majeures dans la temporalité de réalisation d'autres activités. Pour une revue des différentes méthodes de calculs des dissimilarités, voir (Robette et Bry, 2012).

Une fois la matrice obtenue, il ne reste plus qu'à effectuer une classification pour obtenir des regroupements. Parmi les nombreuses méthodes qui existent, nous avons choisi la classification ascendante hiérarchique (CAH) couramment utilisée dans ce genre d'analyse, en nous basant sur la méthode de Ward (1963). La structure du dendrogramme et la répartition de son pourcentage d'inertie en fonction des classes nous incitent à choisir 5 clusters (figure 132).

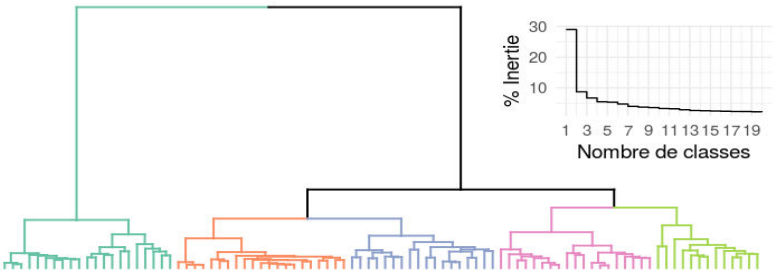


FIGURE 132 Résultat d'une CAH réalisée sur la matrice de dissimilarité entre agendas

La figure 133 montre pour chaque classe la proportion de personnes effectuant une activité par tranche horaire, complétée par la figure 134 qui montre les séquences temporelles de chaque individu (en ordonné). Le nombre de jours hebdomadaire où un individu exerce son activité principale semble être un paramètre très discriminant. En effet, les personnes qui travaillent jusqu'à sept jours par semaine sont regroupées dans la classe 2, et ceux qui exercent une activité principale entre 5 et 6 jours se retrouvent dans les classes 1 et 4. À noter que les agendas des personnes de la classe 1 sont plus redondants et présentent moins de variations de séquences que ceux de la classe 4. La classe 3 rassemble des individus qui n'ont pas ou qui ne passent que peu de temps dans leur activité principale. La classe 5 reprend des personnes qui ont beaucoup d'activités, et des agendas plutôt morcelés.

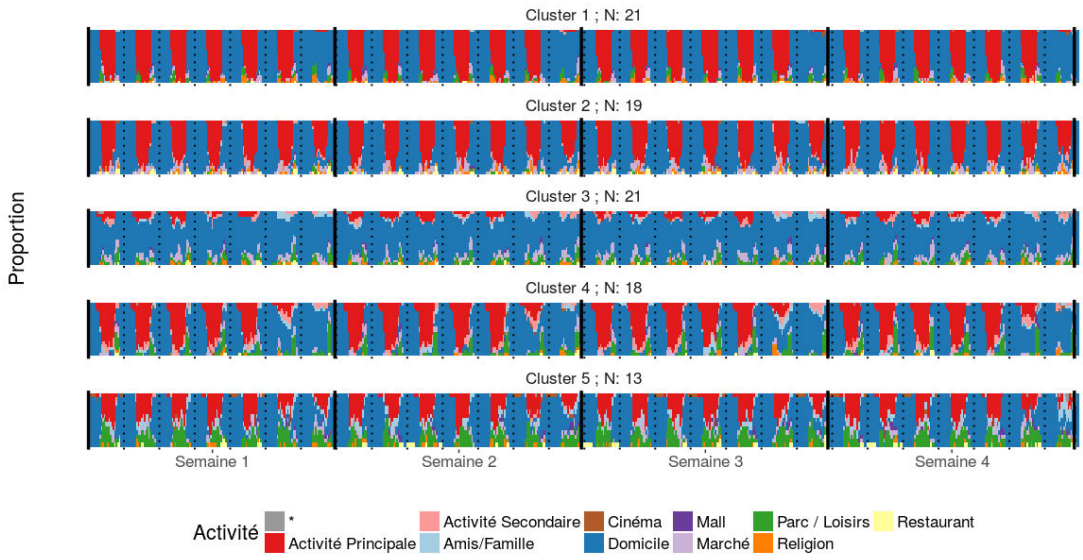


FIGURE 133 Séquences d'activités moyennes par classes.

La figure 134 met en avant les séquences d'activités individuelles pour chaque groupe ce qui permet d'observer plus en détail la variabilité des agendas qui composent chacun des groupes. Par exemple, les individus du cluster 1 ont quasiment tous les mêmes séquences du

lundi au samedi, avec quelques variantes le dimanche. Le groupe 2 est surtout caractérisé par une activité principale réalisée les 7 jours de la semaine, avec d'assez grandes différences dans les activités réalisées en journée, même si ces dernières sont relativement peu nombreuses. Les membres du groupe 4 travaillent entre 5 et 6 jours, mais ont tendance à aller dans les parcs les soirs de week-end. Le cluster 3 concerne essentiellement les personnes de l'échantillon qui n'ont pas d'activité principale, et qui restent surtout chez eux. Enfin, les agendas individuels des membres du groupe 5 sont très différents, et semblent être surtout caractérisés par beaucoup d'activités réalisées en journée.

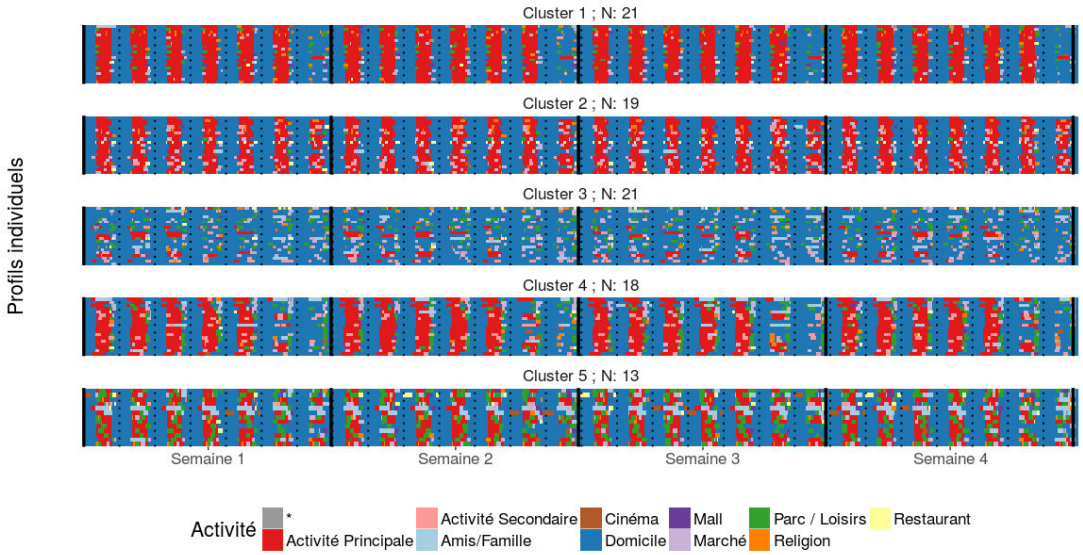


FIGURE 134 Séquence d'activités de chaque individu par classe.

FIGURE 135 Répartition des individus par classe en fonction de leur niveau de richesse estimé.

Il est aussi possible de mettre ces regroupements d'individus en fonction de leurs agendas au regard de leur niveau de richesse estimé précédemment (figure 135). Aucune association très

marquée n'apparaît, même si les personnes les moins aisées (niveau de richesse faible et très faible) se retrouvent principalement dans les classes 1 (agenda routinier et 6 jours de travail), 2 (7 jours de travail) et 3 (pas d'activité principale), et très peu dans la classe 4 (5 jours de travail et agenda assez varié) qui est surtout composée de personnes de niveau de richesse de moyen à élevé. À noter qu'aucune personne de niveau de richesse très élevé ne se trouve dans la classe 1. Ainsi, même si quelques tendances apparaissent, il n'y a pas de lien très net entre le niveau de capital économique et le déroulement des activités pour les personnes de notre échantillon.

3.2.2 Déduire des matrices de transition entre les activités

Probabilité d'effectuer une activité

À partir de ces séquences d'activités, il est possible de calculer des taux de transitions, c'est-à-dire la probabilité qu'un passage d'un état (ou activité) vers un autre soit observé dans la séquence (Gabadinho *et al.,* 2011). Ces taux de transitions peuvent permettre de créer une chaîne markovienne (Spedicato, 2017), ce qui d'un point de vue de simulation permet d'affecter à un individu de synthèse exerçant une activité donnée une probabilité d'en effectuer une autre à l'itération suivante. Ce processus markovien est largement utilisé, notamment dans le cadre de modélisations orientées sur les activités basées sur de petits échantillons (Kim et Song, 2012 ; Perkins *et al.,* 2014).

Les taux de transition, illustrés par la figure 136 sont définis en comptant pour chaque activité à un moment t de la séquence la part des activités réalisées à l'instant t+1. Une personne qui se trouve au restaurant aura respectivement 30 % et 16 % de chances de rentrer chez elle ou d'aller exercer son activité principale, contre 22 % et 13 % si elle se trouve au marché. À noter que les valeurs des auto-boucles sont fortement liées à la durée de l'activité. Par exemple une personne qui se trouve chez elle à un moment de la journée aura de forte de chance de s'y trouver lors de l'itération suivante (92 %), idem pour l'activité principale (0,82).

D'un point de vue de la génération d'agendas de synthèse, l'idée pourrait être ici d'attribuer à chaque agent une fonction de probabilité qui lui permet de passer d'une activité à l'autre (ou non) à chaque itération.

Il est également possible de créer ces chaînes markoviennes en fonction de différents groupes d'individus, en fonction du genre, de la classe sociale ou des regroupements effectués par la méthode d'appariement optimal et la CAH. La figure 137 montre ainsi que les probabilités de changer d'activités varient en fonction des clusters, ce qui est une piste intéressante pour la génération d'individus de synthèses, pour peu que l'on arrive à créer des groupes et des agendas représentatifs de la population.

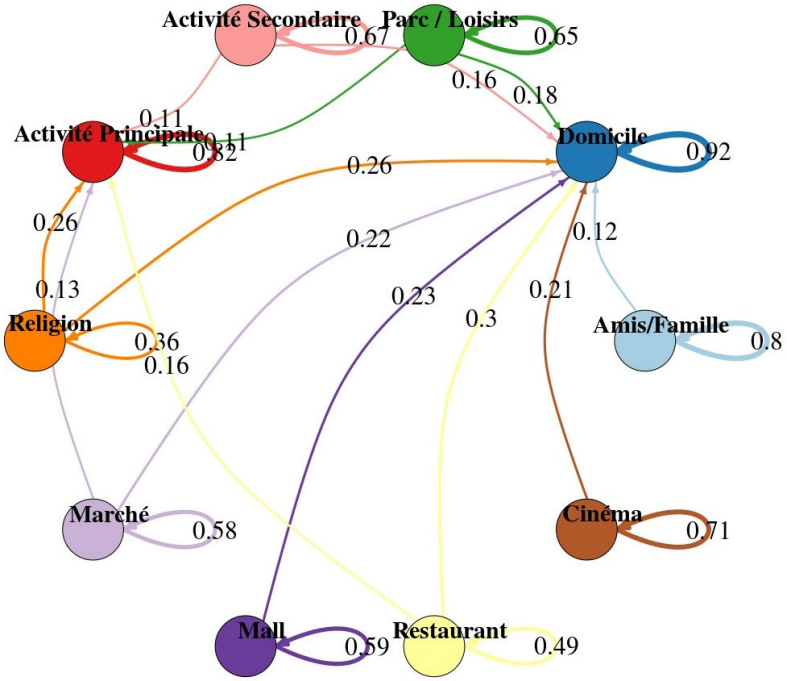


FIGURE 136 Élaboration d'une chaîne markovienne à partir des taux de transitions calculés pour l'ensemble de l'échantillon pour une simulation d'agendas. Seules les valeurs supérieures à 0,01 sont affichées pour des raisons de lisibilité. Réalisé sous R avec la librairie *igraph* - les paquets *ggplot2* ou *ggnetwork* ne permettent pas de visualiser les auto-boucles.

Cependant, compte tenu des valeurs des auto-boucles, soit la probabilité de réaliser la même activité à l'itération suivante, les personnes qui sont à leur domicile ont très peu de chance d'en sortir, et il faudrait de plus prendre en compte les plages horaires où se réalisent les activités, pour obtenir une simulation plus réaliste (éviter qu'une personne aille dans un marché la nuit, ou reste toute la journée au travail par exemple). D'après nos agendas, nous connaissons les durées et les horaires où une personne effectue une activité, et il peut être dès lors pertinent de calculer simplement les transitions entre chacune des activités, sans prendre en compte les séquences successives où une même activité est réalisée. La figure 138 montre ces nouveaux taux calculés, où une personne quittant son domicile a respectivement une probabilité de 45, 17 et 18 % d'aller exercer son activité principale, d'aller au parc ou au marché. Lorsqu'un individu quitte son travail, il a dans cet exemple 44 % de chance de rentrer chez lui, mais 15 % d'aller au parc ou au marché. Il est également possible de recalculer ces taux de transition en fonction de différentes classifications de l'échantillon.

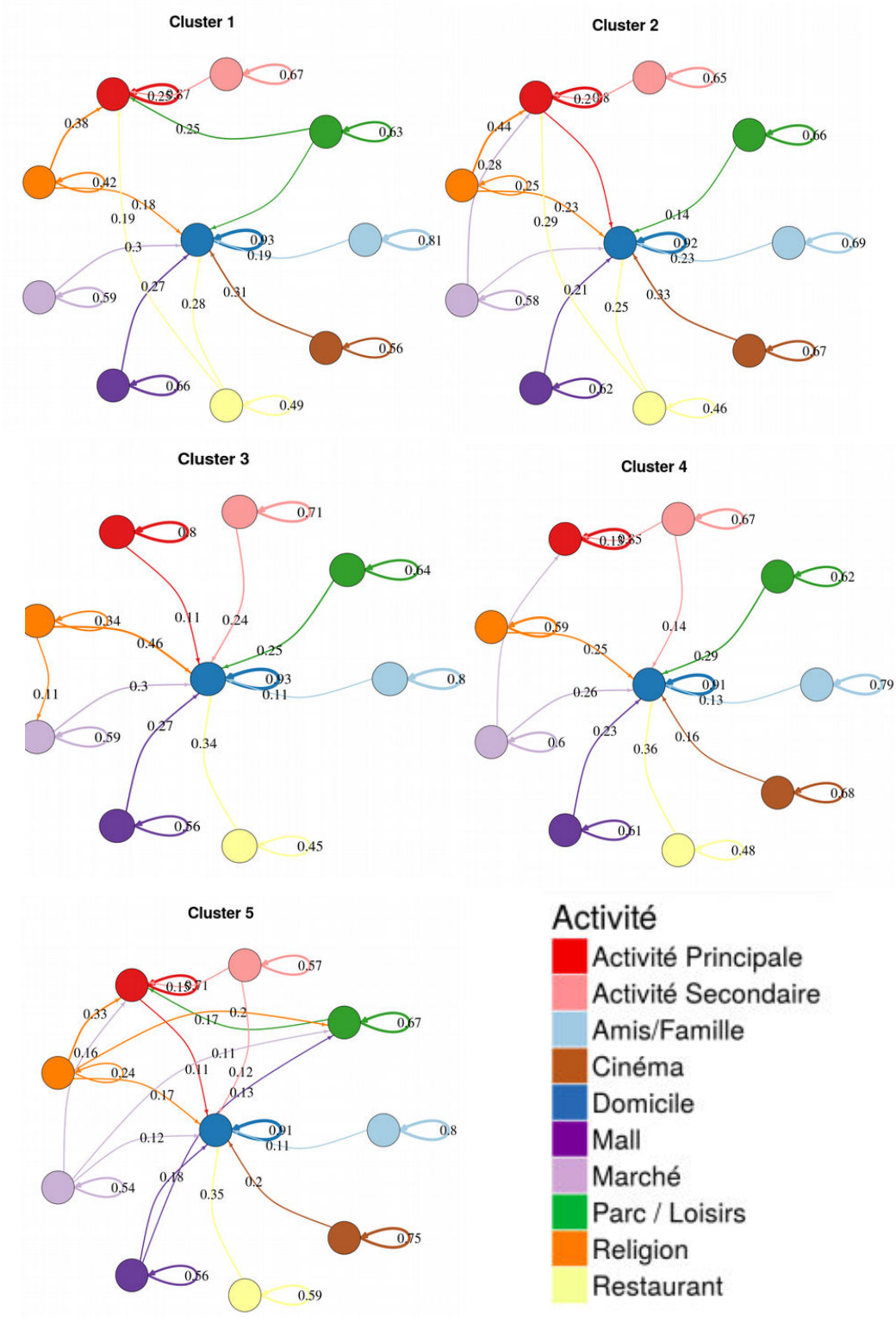


FIGURE 137 Chaines markoviennes pour les 5 clusters issus de la méthode d'appariement optimal et de la CAH.

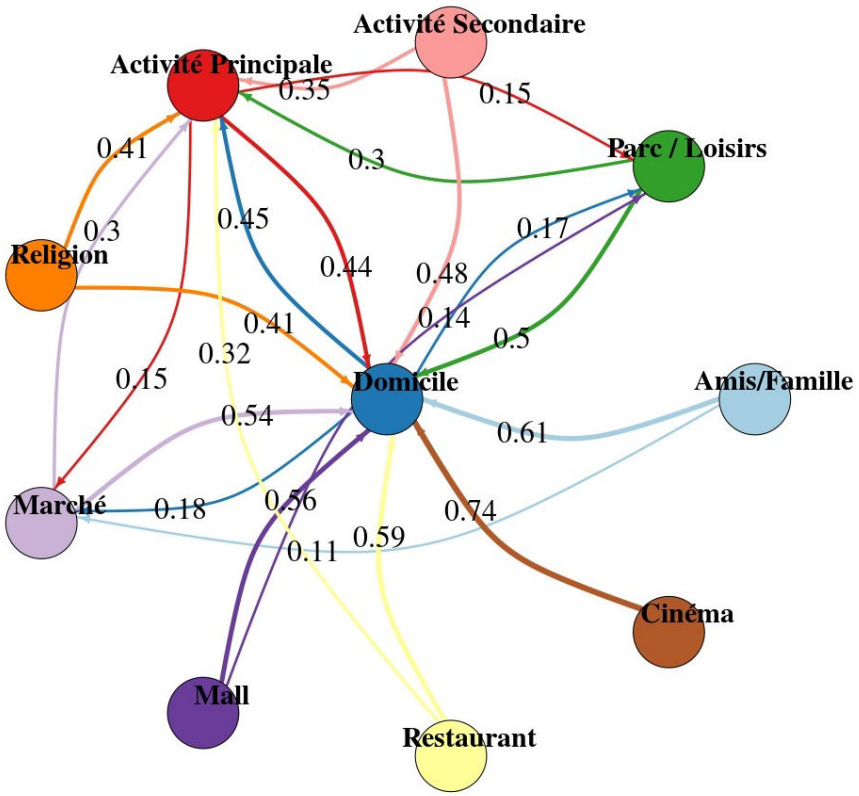


FIGURE 138 Probabilité d'effectuer une autre activité lorsque l'activité précédente se termine.

Cette approche qui requiert une connaissance des plages horaires et des durées des activités permet d'obtenir de manière simple et rapide des probabilités d'effectuer d'autres activités. Il est également possible, de définir ces matrices de transition par tranche horaires et par activité (Wu *et al.*, 2014), ce qui revient à conditionner l'activité suivante par l'heure de la journée et l'activité précédente réalisée. Par exemple, le soir, une personne aura probablement plus de chances de sortir dans un bar après être allée au restaurant que d'exercer son activité principale.

Si l'utilisation de ces matrices transitions semblent être pertinentes pour définir les séquences des activités exercées un jour donné, il convient maintenant de trouver une méthode permettant de choisir le lieu où s'exercera cette activité.

Probabilité de sélectionner un lieu donné

Une première approche pourrait être de mobiliser des informations sur les niveaux d'attractivités des lieux par tranches horaires, et d'appliquer des modèles gravitaires, radiatifs (Simini *et al.*, 2012), ou r*ank-based*, (Abbasi *et al.*, 2017 ; Noulas *et al.*, 2012) (voir chapitre 5).

Une méthode plus simple pourrait être d'utiliser les métriques de déplacements déduites des données du corpus. Nous pouvons définir pour l'ensemble de notre échantillon la distance entre chaque lieu fréquenté pour chaque individu pour une activité donnée, pondéré par la fréquence de visite du lieu. Ceci permet de générer des courbes de densités, que nous pouvons considérer comme des probabilités de tirer une distance pour passer de l'activité A à B (figure 139).

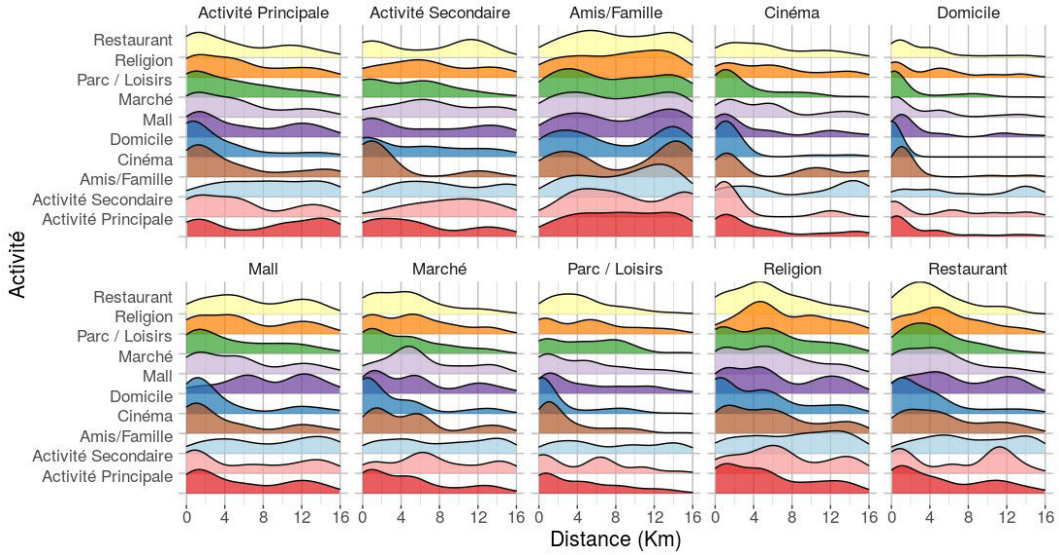


FIGURE 139 Distance entre chaque type d'activités (10 encadrés) et l'ensemble des autres activités pour l'ensemble des utilisateurs.

La figure 139 ci-dessus montre par exemple que lorsqu'une personne est dans un marché et qu'elle compte se rendre dans un restaurant, ce dernier aura une probabilité plus importante d'être situé à moins de 6 km. Si après être allée au restaurant, la personne a tiré l'activité « aller au *mall* », il conviendra de choisir un *mall* en suivant une la fréquence de distribution des distances, ici bimodale et centrée sur 4 et 12 km. Ces probabilités peuvent également être conditionnées par l'appartenance à un cluster d'individus donnés (par exemple les individus résidant dans un même quartier, ou appartenant à un même groupe social). Il est également envisageable de mettre à jours ces courbes de probabilité de parcourir une distance pour aller d'une activité à une autre en les recalculant toutes les n itérations, en fonction des déplacements des agents dans le modèle.

Synthèse

Nous avons testé une méthode pour récolter rapidement des informations sur le terrain relatives aux espaces d'activités. Une première analyse permet de rendre compte des potentiels de mobilités différents selon les caractéristiques socio-économiques des individus, sans que des tendances claires et très marquées n'apparaissent pour autant.

À partir de questionnaires sur les activités effectuées par un individu, où la localisation, la fréquence de visite, le moment de la journée, et la durée de l'activité sont soit connus soit estimés, nous avons pu créer des agendas avec une part de stochasiticité, qui nous ont permis d'obtenir des taux de transition entre activités.

Dans un contexte de création d'individus de synthèses, il est alors envisageable à partir de données les plus représentatives possible, d'affecter à chaque agent des caractéristiques sur son espace d'activité (nombres de lieux, type d'activités, fréquence de visite, etc.), auxquelles on associe une matrice de transition (chaînes markovienne) et des probabilités de distance de déplacement entre 2 activités, le tout en fonction du groupe auquel l'agent appartient (qui peut être centré sur l'âge, le genre, le niveau de richesse, le lieu de domicile, etc.). Cette approche requiert d'avoir de bonnes informations sur les pratiques des mobilités des habitants d'un quartier ou d'une ville, et peuvent passer par des données issues de traces numériques ou des enquêtes ménages déplacement.

Une base de données géographique qui répertorie les principales activités (lieux de cultes, parcs, marchés, restaurants, *malls*, cinéma, école, etc.) est également nécessaire afin de pouvoir affecter les agents à des lieux où se déroulent des activités. De plus, il convient de connaître, ou d'estimer les lieux où les personnes exercent leur activité principale. Ceci peut passer par des modèles d'attractivité et/ou d'absorption, qui peuvent être utilisés pour estimer la localisation des activités principales et les interactions entre les lieux, en fonction de la distance offrant une alternative au calcul de probabilité de parcourir une distance pour aller d'une activité A vers B.

Par ailleurs, les déplacements individuels sont souvent influencés par le réseau social (Alessandretti *et al.,* 2018 ; Calabrese *et al.,* 2011b ; Cho *et al.,* 2011), que cela soit dans la fréquentation de lieux de sorties ou dans la visite au domicile de proches (Stoddard *et al.,* 2013). Et l'affectation crédible d'un groupe d'amis localisés dans la ville à un agent requiert soit des hypothèses fortes, soit des données adaptées, ou encore l'utilisation de modèles de réseaux sociaux relativement poussés. Notre approche ne prend pas ce facteur pourtant central en compte, tout comme les modes de transport.

Notre chaîne de traitement permet théoriquement de générer en amont des agendas

stochastiques, avec des probabilités d'interactions entre activités et des choix de lieux dépendant des métriques du groupe auquel appartient l'agent. Une fois ces informations générées, les agents se déplacent de manières indépendantes selon une logique déterministe qui devrait accélérer les temps de calcul l'agent n'a pas de décisions à prendre, il suit simplement son agenda.

Le prochain chapitre évaluera si cette approche est transposable aux données individuelles issues de *Twitter*.

Chapitre VIII: Traces numériques et espaces d'activités : analyse des mobilités et génération d'agents à Delhi

Les enquêtes de terrains effectuées dans le quartier de Malviya Nagar, en plus d'acquérir une connaissance du quartier et une relative compréhension des réalités sociales dans un quartier de Delhi, nous ont permis de récolter des informations sur les lieux fréquentés par un nombre assez restreint d'individus. Ce dernier aspect est souvent montré comme une limite, notamment vis-à-vis de la représentativité dans le cadre d'études quantitatives. De nouvelles sources de données mobilités sont apparues ces dernières années, au gré des traces numériques individuelles géolocalisées et datées laissées sur les réseaux sociaux. Nous allons présenter ici une base de données de taille plus importante, constituée de messages postés par les utilisateurs du réseau social *Twitter* à Delhi.

Les données que nous avons recueillies sur le terrain étaient assez imprécises temporellement, car la fréquence et les tranches horaires de visites étaient soumises à la seule appréciation de l'interviewé. Les données *Twitter* souffrent de limites assez similaires, à savoir qu'elles sont discontinues et épisodiques, et qu'il est délicat d'estimer les fréquences de visites et les plages horaires réelles où se déroulent les activités des utilisateurs. La méthode employée pour reconstruire des agendas continus à partir de nos entretiens de terrain sera modifiée pour être adaptée aux données *Twitter*, ce qui nous permettra de reconstruire des agendas relativement plausibles pour cet échantillon. À partir de là, il est devient théoriquement possible de déduire des informations extrapolables et utilisables pour générer des agents. La modélisation des mobilités basées sur des agendas est une approche relativement récente (Schlink *et al.,* 2010 ; Zheng *et al.,* 2009, p. 20), et peu d'études utilisent des données massives issues des traces numériques.

À partir de *check-in* issus de *Weibo* à Shangaï (Wu *et al.,* 2014), ont défini des séquences d'activités réalisées lors d'un trajet quotidien. Ils ont ensuite généré des agents de synthèse en se basant sur matrices de transitions, conditionnée par l'activité exercée à l'instant t, le niveau de flexibilité de cette dernière et les plages horaires. La localisation est ensuite affectée après avoir réalisé des milliers de simulations permettant d'estimer des paramètres de rugosité des déplacements afin d'obtenir le meilleur accord entre les données observées et les données simulées.

D'autres travaux, comme ceux de Pappalardo et Simini, (2017b) à partir de données téléphonie et de GPS sur des voitures, sont basés sur une approche en deux temps. Ils génèrent tout d'abord des agendas, constitués de lieux abstraits (soit pour l'instant des séquences de labels), suivis par l'attribution d'une localisation selon un modèle d'exploration et de retour préférentiel, comparable à un modèle gravitaire amélioré (Pappalardo *et al.,* 2016b). Cette approche en deux temps nous semble pertinente, car la génération d'agendas permet d'évaluer collectivement le niveau de robustesse du déroulement temporel d'activités générées individuellement. Ne reste alors plus qu'à attribuer une localisation à chacun de ces lieux, soit en passant par des modèles théoriques, soit en se basant sur des métriques et tendances de déplacement issues d'un échantillon assez large, soit en utilisant la force brute pour calibrer les données simulées aux données observées.

Conscients que quantité ne signifie pas nécessairement qualité ou représentativité, nous évaluerons dans un premier temps la qualité de ces données *Twitter*, qui, comme vu dans le chapitre 6, ne sont déjà pas représentatives spatialement. Nous estimerons ensuite les activités qu'un individu est susceptible de réaliser, en nous basant sur des données de l'utilisation du sol, obtenue en combinant diverses sources, qu'il s'agisse du projet de cartographie libre et participatif OpenStreetMap ou de *Google Maps*. En inférant les localisations des messages envoyés au type de lieu fréquenté, nous obtiendrons un indicateur sur l'activité qu'un utilisateur de *Twitter* était possiblement en train d'exercer au moment de l'envoi de son message. Alors que la plupart des modèles basés sur les espaces d'activités ne distinguent que 3 types de lieux fréquentés (Domicile, Travail, Autre) (Jiang *et al.,* 2017, 2016 ; Karl *et al.,* 2014 ; Schneider *et al.,* 2013) nous serons par cette approche capable d'en identifier un nombre bien plus important, selon le niveau d'exhaustivité de la couche d'utilisation du sol. De plus, dans le contexte denguien, le risque de contamination et de propagation de la maladie varie selon le type de lieu fréquenté. Les espaces ouverts comme les cours d'école, les marchés ou les parcs sont le lieu de brassage d'un grand nombre de personnes, plus ou moins susceptibles de contracter la maladie. La prise en compte de ces types de lieux dans l'espace d'activité d'une personne est donc primordiale pour la compréhension des processus de diffusion de l'épidémie (Perkins *et al.,* 2014).

Le chapitre précédent a ouvert des portes sur les potentiels de simulations à partir de ces agendas dans le temps, que nous allons tenter de développer ici. Nous allons proposer un algorithme permettant de générer des agendas de synthèses à partir d'agendas reconstitués et de leurs propriétés sous-jacentes, comme les taux de transitions entre deux activités et les types de lieux fréquentés. Nous testerons ensuite cet algorithme sur les agendas reconstitués à partir des données terrain, ce qui nous permettra de discuter de sa pertinence. Nous discuterons également de l'intérêt d'opter pour une approche mixte, combinant données du terrain qualitatives et traces numériques plus quantitatives dans l'analyse des mobilités et dans l'élaboration d'un modèle.

Comme vu dans le chapitre 6, nous avons pu estimer la localisation du domicile de 4089 utilisateurs de *Twitter*, parmi lesquelles 88 habiteraient dans le secteur de Malviya Nagar. Ces effectifs sont extrêmement faibles au regard de la population de Delhi (~20 millions d'habitants). Notre démarche sera donc ici essentiellement méthodologique.

1 Vers une meilleure connaissance de l'échantillon

1.1 Qui tweet à Delhi et à Malviya Nagar ?

Connaissant la localisation et l'heure d'envoi d'un message, il est tout d'abord possible d'évaluer la répartition de l'échantillon dans la ville au cours du temps (figure 140). Trois pôles principaux apparaissent, Gurgaon au sud-ouest, le sud de Delhi, et l'est de la ville. Le centre accueille plus de personnes le mardi midi que le dimanche à la même heure, et le sud de Delhi est plus actif le dimanche soir. Nous pourrions continuer la description de cette figure, mais ces informations n'ont de valeur que si l'échantillon est représentatif spatialement, ce qui n'est pas le cas ici (chapitre 6), ou à la rigueur socio-économiquement parlant. Mais estimer cette dernière composante à partir des traces numériques laissées sur Internet n'est évidemment une chose aisée, surtout quand nous n'avons pas d'autres informations que l'heure et la localisation des messages émis.

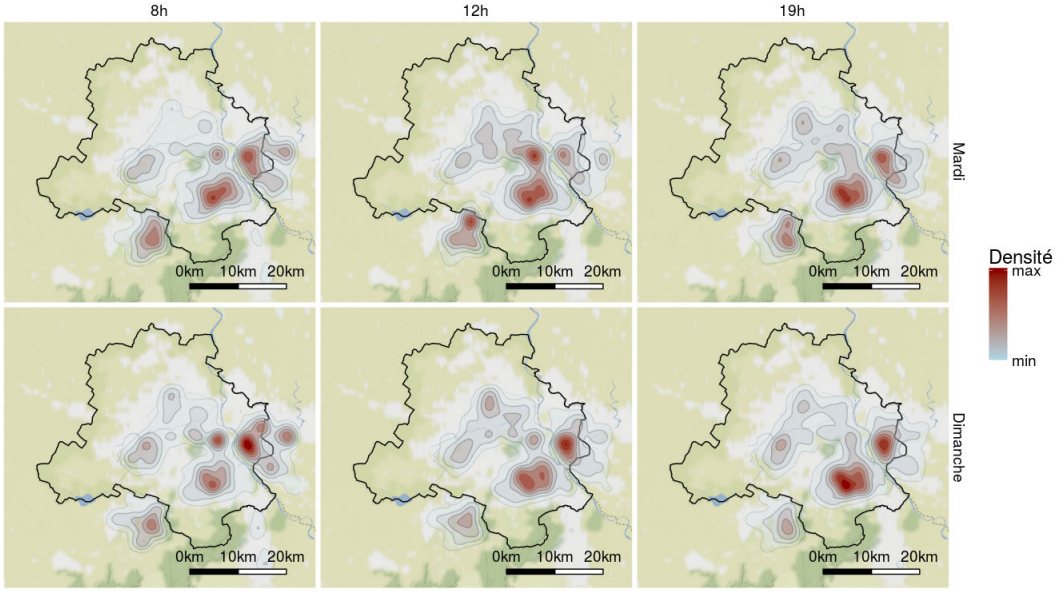


FIGURE 140 Densité de fréquentation des différentes zones de la ville de Delhi à différentes tranches horaires. (8 h, 12 h et 19 h), le mardi et le dimanche, d'après l'échantillon de *Twitter* (4089 utilisateurs).

S'il est admis que la plupart des réseaux sociaux estiment ces catégories socio-économiques en fonction de l'ensemble des données laissées par leurs utilisateurs (et pas

seulement les données publiques), nous pouvons noter que Facebook a très récemment (1er février 2018) enregistré un brevet[335] qui vise justement à définir les catégories socio-économiques de ces utilisateurs[336] (basse, moyenne, élevée) en se basant sur des données démographiques (âge, genre, niveau d'étude, « ethnicité » et zone géographique), sur le nombre et le type d'appareils utilisés, le temps passé sur Internet, ou encore les différents trajets effectués (ou localisations enregistrées)[337] (Sullivan *et al.*, 2018). N'ayant ni les données, ni les moyens et compétences techniques pour effectuer de tels traitements, ni la possibilité de savoir si leur approche fonctionne réellement, nous allons tenter ici d'apprécier le niveau de représentativité socio-économique simplement en recoupant les domiciles des utilisateurs avec la taxe foncière par colonie (figure 141) qui est *a priori* un indicateur assez correct des niveaux de richesses (chapitre 7).

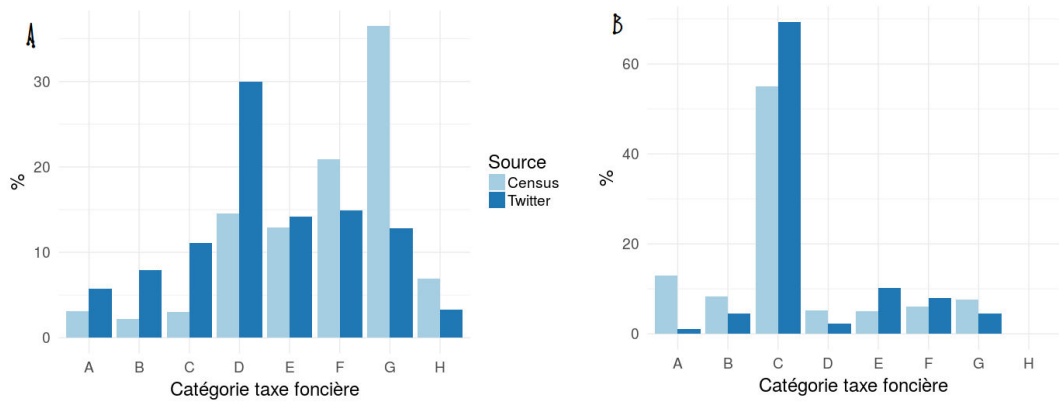


FIGURE 141 Comparaison entre la part d'habitant et le nombre de domiciles d'utilisateurs de *Twitter* estimé selon les colonies d'une taxe foncière donnée. Pour l'ensemble de Delhi (A) et pour le secteur de Malviya Nagar (B).

La figure 141.a présente la part d'habitants de Delhi par catégorie de taxe foncière (données census 2011) comparée aux domiciles estimés des 2664 utilisateurs de *Twitter* qui vivent dans une des colonies de la ville les 1425 autres résident dans les villes satellites voisines pour lesquelles nous n'avons pas ces informations. La figure 141.b se focalise sur la zone d'étude de Malviya Nagar, où nous avons détecté le domicile de 88 utilisateurs. La figure 141.a pointe clairement une surestimation des usagers de *Twitter* dans les quartiers les plus riches (taxe foncière de A à D) par rapport à la population du recensement. Alors que 25 % des utilisateurs de *Twitter* vivent dans des zones de catégories A à C, ils ne sont que 10 %

335. lu sur : http ://affordance.typepad.com/mon weblog/2018/02/cest la lutte algorithmique finale .html

336. Bien qu'applicable aux sciences sociales, ce brevet est destiné à aider les annonceurs à mieux cibler leur clientèle potentielle : "*By predicting the socio economic groups of users, the online system is able to help the third party present sponsored content to the target users.*"

337. Ce brevet utilise également des données sur le foyer, en prenant par exemple en compte le nombre de climatiseurs et de voitures, comme nous avons fait dans le chapitre précédent pour requalifier le niveau socio économique de notre échantillon.

dès lors qu'il s'agit de la population recensée par les institutions officielles. Alors que les deux tiers des Delhiites vivent dans une colonie de catégorie F à G, ce ratio passe à un tiers pour les utilisateurs de la plateforme d'envoi de message. Ceci sous-entend que les classes sociales les plus aisées sont sur-représentées dans notre échantillon issu de *Twitter*, à l'échelle de la ville.

Mais la figure 141.b montre de manière assez surprenante que les proportions par type de catégorie de taxe foncière entre les utilisateurs de *Twitter* et les données du recensement dans la zone de Malviya Nagar sont assez similaires, sauf dans les quartiers très riches (catégorie A). Il pourrait être tentant de dire que l'échantillon d'utilisateurs de *Twitter* est représentatif socialement dans notre zone d'étude, d'autant plus que les proportions sont plus proches du recensement que l'échantillon collecté sur le terrain. Mais les enquêtes et la connaissance du terrain montrent que les personnes les moins aisées, soit une grande part de la population, n'ont probablement pas le capital économique et/ou culturel caractéristique d'un utilisateur de *Twitter*. Et si 94 % de la population à Delhi a un téléphone portable (census 2011), seule 29,5 % de la population utilise internet en Inde [338], et le taux de pénétration de smartphone serait autour de 20 % [339], même si le nombre d'utilisateurs est plus important en ville [340]. Ainsi, posséder à la fois un smartphone, un forfait 3/4/5 G et vouloir partager des informations avec son réseau social lui-même connecté semble être plus l'apanage des classes moyennes / supérieures que des travailleurs précaires et des plus pauvres. Dès lors, conclure que l'échantillon est représentatif socialement reviendrait probablement à commettre une erreur écologique, car les utilisateurs de *Twitter* des quartiers les plus défavorisés moins bien notés par la taxe foncière (catégorie E, F et G) font très probablement partie des franges les plus aisées de ces secteurs. Ceci n'est pas une limite en soi, car dans l'optique d'analyser les potentiels de mobilités en fonction des niveaux de richesse, le terrain nous a donné l'impression que les personnes appartenant aux catégories socio-économiques les plus élevées étaient les plus délicates à interviewer. Il est donc envisageable de combiner les deux approches traces numériques & interview dans l'optique d'augmenter la taille de notre échantillon.

Mais auparavant, il convient d'ajouter de l'information aux données *Twitter* afin de tenter de les comparer avant d'éventuellement songer à les agréger aux données récoltées sur le terrain, ou encore leur appliquer des méthodes similaires de reconstitution d'agendas individuels à partir de données plutôt clairsemées temporellement. Les différents lieux fréquentés par les utilisateurs de *Twitter* à Delhi sont pour l'instant simplement regroupés sous forme de cluster (chapitre 6), qui contiennent les heures et les jours où chaque personne a *tweeté* dans un lieu donné. Ces informations sont cependant incomplètes au regard du concept d'espace d'activité que nous souhaitons mobiliser, car mis à part le lieu de domicile, nous n'avons pas pour l'instant affecté

338. https://www. tu. nt/en/ITU D/Stat st cs/Documents/pub cat ons/m sr2017/MISR2017 Vo ume2.pdf
339. https://newzoo.com/ ns ghts/rank ngs/top 50 countr es by smartphone penetrat on and users/
340. http://www.theh ndubus ness ne.com/ nfo tech/top 30 c t es make up 51 of smartphone market de h tops ta y/art c e8309061.ece

d'activités aux lieux fréquentés par un individu.

1.2 Dans quel type de lieu ?

Différentes études se sont basées sur l'analyse de contenu de messages pour estimer l'activité que réalise un individu (Cheng *et al.,* 2010, p. 201 ; Hiruta *et al.,* 2012), mais nous n'avons pas ces informations. De plus, une personne peut envoyer des messages non reliés à l'action qu'elle effectue. Elle peut par exemple rebondir sur un sujet d'actualité depuis son lieu de travail, de domicile ou lors de ces déplacements.

Nous allons poser dans un premier temps l'hypothèse que l'activité qu'un utilisateur de *Twitter* réalise lorsqu'il envoie un message dans un lieu donné dépend de l'utilisation du sol où se trouve ce lieu, avec les biais associés. Par exemple, si une personne envoie un message depuis un *mall,* nous considérons qu'elle s'adonne à une activité commerciale ou de chalandise. De même qu'un message envoyé depuis une route nous indique que la personne était en train de se déplacer. Nous reviendrons plus tard sur la sélection de l'activité principale parmi tous les lieux fréquentés par un individu, mais l'objectif est pour l'instant d'obtenir un jeu de donnée dont la structure est similaire à celle de nos enquêtes de terrain soit un lieu associé à une activité, des heures de présence et des fréquences de visites. Il convient donc de chercher à collecter et agglomérer des bases de données géographiques dans l'optique d'obtenir une couche d'utilisation cohérente. Cette dernière, intersectée avec les différents lieux des espaces d'activités des utilisateurs de *Twitter*, permettra d'inférer une activité potentiellement réalisée dans ces différents endroits (Cebeillac et Rault, 2016 ; Huang *et al.,* 2014).

Nous avons à notre disposition les différentes cartes des plans directeurs (MCD) ou les typologies d'utilisation du sol à Delhi (Lefebvre, 2011), mais ces dernières sont cependant plus utiles pour décrire les structures urbaines et les niveaux de ségrégation que pour faire ressortir les différentes activités susceptibles d'être exercées. De plus, ces typologies n'ont pas le degré de finesse adapté à notre thématique, car elles ne distinguent que les zones commerciales, industrielles et résidentielles. Il convient alors de rechercher d'autres sources de données qui permettent de faire ressortir des activités plus univoques, comme vu dans le chapitre 5.

OpenStreetmap

La première base de données que nous mobilisons provient de la plateforme libre OpenStreetMap (*OSM*). Il s'agit d'un « projet international fondé en 2004 dans le but de créer une carte libre du monde »[341]. Il permet à quiconque de devenir contributeur[342] et de participer à la création ou à la modification de points, de lignes (routes, barrières, etc.)

341. http://openstreetmap.fr/
342. La plateforme comptait 4,3 millions d'inscrit en octobre 2017. http://w k .openstreetmap.org/w k /Stats

ou de polygones (bâtiments, limites administratives, rivières, etc.) pour peu qu'il respecte les nomenclatures et les conventions. Le contenu de l'information géographique présent sur *OSM* change continuellement, avec plus de 3 millions d'éditions quotidiennes, pour tendre vers les cartes les plus objectives et de meilleure qualité possible, et de plus libre d'accès, conformément à l'idéal d'un grand nombre des contributeurs (Duféal et Noucher, 2017). Ainsi, il est considéré que plus le nombre de contributeurs et d'objets édités est important dans une zone donnée, meilleur est le niveau de précision (Senaratne *et al.*, 2017)*OSM* s'associe par ailleurs à des institutions publiques, comme en France avec la direction générale des finances publiques qui autorise la mise en ligne des données du cadastre utilisées pour géolocaliser les bâtiments et nommer les voies[343], ou encore l'IGN qui met à disposition sa base d'images aériennes[344].

FIGURE 142 Densité des contenus et d'édition de la base OSM en 2014. Source : Mark Graham & Stefano De Sabbata, http://geography.oii.ox.ac.uk/?page=openstreetmap

Cela dit, à l'image d'un autre projet collaboratif comme Wikipédia, seule une part infime des inscrits contribue activant, avec par exemple 21 % des 6 milliards de points GPS uploadés par seulement 49 comptes[345], et toutes les zones du monde ne sont pas cartographiées avec

343. https://w k .openstreetmap.org/w k /FR:W k Project France/Cadastre
344. http ://openstreetmap.fr/bdortho
345. https://www.openstreetmap.org/stats/data stats.htm

le même niveau de détail. La figure 142 montre deux cartes réalisées en 2014 et met en avant les différences très marquées entre les régions du monde d'un point de vue de la densité de couverture par le service et par le nombre de contenus édité. Les pays d'Europe Occidentale, les États-Unis d'Amérique, le Canada et le Japon [346] sont les régions du monde qui présentent les plus fortes densités de contenus et le plus grand nombre d'édition, du fait de la présence d'une communauté plus active, et de l'utilisation dans certains états de données institutionnelles. À titre de comparaison, 29 contributeurs étaient enregistrés à Paris, contre seulement 4 à Bangkok et Delhi [347] en février 2018. Il conviendra donc d'être relativement prudent lorsque nous utiliserons les données OSM dans les capitales Thaï et Indienne, car les bases sont plus susceptibles d'être incomplètes ou incorrectes.

En plus de leur caractère libre et facile d'accès, les données OSM contiennent des informations sur l'utilisation du sol dans un polygone, notamment via la couche « landusage » ou la variable « amenity ». 48 types sont ainsi définis pour Delhi, dont les lieux d'éducations (école, université, bibliothèques), les commerces, les lieux de sports (terrains de jeux, salles de sports, stades, etc.) les lieux de cultes, les zones végétalisées (parcs, forêts, cultures) les hôpitaux, ou encore les cinémas, les restaurants, les gares et les aéroports. Les routes principales et les voies ferrées [348] sont également récupérées sous forme de *polylignes* auxquels nous appliquons une zone tampon de 20 m, afin de simuler leur largeur.

Outre le fait que ces informations ne sont probablement pas exhaustives à Delhi (et à Bangkok), une autre limite tient dans la topologie et la géométrie des polygones constituant les entités spatiales. Il est assez fréquent que dans une même couche vectorielle, de petits polygones soient compris dans un plus grand. C'est le cas par exemple lorsque certains commerces sont renseignés au sein d'un *mall*. Dans ce cas nous faisons le choix de supprimer les petits commerces. Lorsque plusieurs bâtiments d'une université (bibliothèque, restaurant, dortoirs, parc, *etc.*) sont compris dans l'entité "Université" qui les englobe, nous estimons qu'ils font partie du campus et décidons de les regrouper sous le terme « Université ». Ainsi, pour des raisons de simplicité, nous ne souhaitons pas ici faire de distinctions entre la présence d'un individu dans un campus universitaire, et la visite d'un parc ou d'un restaurant présent dans ce campus. Nous partons du principe que les campus forment un espace complet, où les toutes activités de bases sont présentes (hébergements, visite d'amis, commerces, restaurants, éducation, parcs, *etc.*) (Martelli, 2017).

Ces universités sont parfois elles-mêmes englobées par une entité « forêt », notamment lorsque le campus se trouve dans un environnement boisé comme c'est le cas de l'université Jawaharlal Nehru (JNU) dans le sud de Delhi. Nous séparons donc temporairement les polygones

346. une réminiscence de la triade ?
347. https ://wiki.openstreetmap.org/wiki/Category :Users in * (remplacer * par Paris, Bangkok ou Delhi)
348. de type : tertiary, seconday, primary, trunk, road et rail

d'utilisation du sol qui se rapportent plus à de l'occupation du sol (végétation) de ceux associés à une activité réalisable par un individu. Pour les éléments de cette dernière couche, nous estimons que le niveau de précision des éléments enclavés est trop fin, et nous les supprimons s'ils sont contenus dans une unité spatiale plus grande. Dans le cas où un polygone est partiellement compris dans un autre, nous séparons le petit du plus grand et conservons ces deux entités spatiales[349]. Nous procédons ensuite à une agrégation successive, ajoutant à la couche d'utilisation du sol les éléments de végétation et les routes qui n'intersectent pas cette dernière. Nous obtenons ainsi une couche d'utilisation du sol où chaque point de l'espace n'appartient qu'à une seule catégorie.

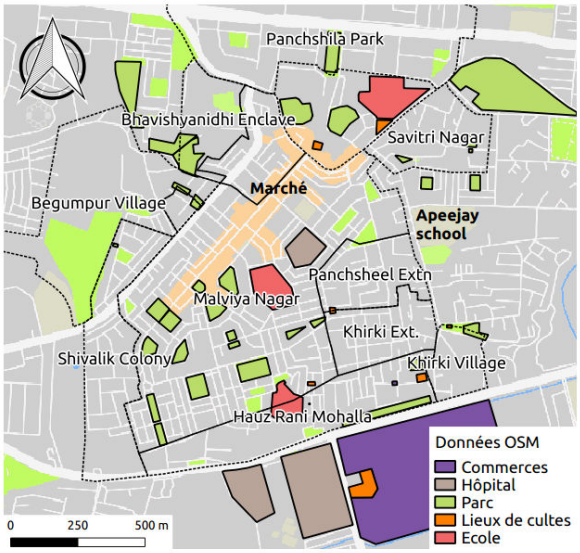


FIGURE 143 Utilisation du sol à Malviya Nagar d'après les données OSM - fond de carte : *Google Maps*

La figure 143 montre la couche d'utilisation du sol obtenue (sans les routes) dans le quartier de Malviya Nagar, avec un fond de carte de *Google Maps*. Les *malls* de Saket au sud de la zone sont bien considérés comme des zones commerciales, mais pas le marché de Malviya Nagar (en orange sur le fond de carte). Quelques parcs sont présents mais beaucoup manquent. À noter que l'école Apeejay, à la limite entre Khirki Extension et Sheikh Sarai, ne figure pas dans la base de données OSM. Comme pressenti, l'utilisation du sol issue d'OSM n'est pas exhaustive, et nous allons la compléter en mobilisant des données issues de *Google Maps*.

Utilisation de Maps (Google)

Maps, le service cartographique de *Google* a débuté en 2005, et couvre aujourd'hui

349. Ces traitements bien qu'*a priori* anodin ne sont pas directement réalisables de manière automatique dans différents logiciels de géomatique (Arcgis, Qgis, etc.) car au sein d'une même couche vectorielle. Nous mettrons prochainement notre code à disposition (sous R).

la plus grande partie des zones urbaines de la planète. Au-delà d'un simple fond de carte avec l'occupation du sol et les routes, le service met à disposition des points d'intérêts (*POI*) géolocalisés. Ces derniers représentent en général un établissement et contiennent des attributs spécifiant la nature de ce dernier (école, commerce, ou autre) et parfois leur nom. L'entreprise acquiert ces *POI* de deux manières : en collectant des bases de données officielles ou privées, et en mobilisant les utilisateurs du service, qui peuvent ajouter des *POI* en renseignant le nom et la catégorie potentiellement vérifiés par les équipes de *Google*. De plus le projet « Ground Truth [350] » de *Google* vise entre autres la validation et la création automatique de *POI* en analysant les photographies des rues réalisées dans le cadre du service « Street View » par des voitures équipées de caméras panoramiques. Il s'agit ici de détecter automatiquement les enseignes des commerces et de les localiser sur une carte (Wojna *et al.*, 2017 ; Yu *et al.*, 2015), en utilisant notamment les « *captchas* » (figure 144) pour entraîner les algorithmes de *machine learning* permettant la reconnaissance d'éléments dans des images [351]. Ce programme est lancé dans 50 pays, dont la Thaïlande depuis début 2013 [352], mais pas en inde car le service « Street View » n'y est pas autorisé [353].

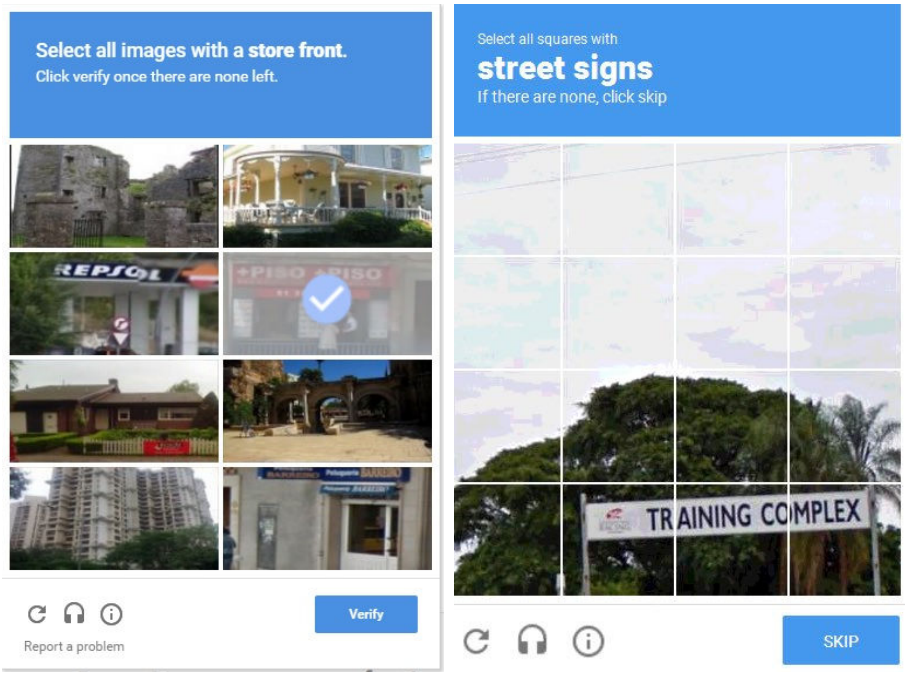


FIGURE 144 Exemple de *captcha* permettant d'entraîner les algorithmes de reconnaissances d'images de *Google*.

Ces *POI* sont facilement récupérables en passant par une *API* [354], mais les limites de

350. https://research.goog eb og.com/2017/05/updat ng goog e maps w th deep earn ng.htm
351. Les utilisateurs qui cliquent sur des images participent ainsi à la calibration des algorithmes
352. https://www.cnet.com/pictures/how *Google*s ground truth maps the world pictures/10/
353. http://www.te egraph.co.uk/techno ogy/2016/06/10/Goog e street v ew banned n nd a due to secur ty concerns/
354. https://deve opers.goog e.com/p aces/web serv ce/search?h =en

requêtes quotidiennes (2000) entraînent un temps de collecte extrêmement long sur des villes comme Delhi & Bangkok. Nous avons néanmoins récupéré environ 136 000 *POI* en août 2015 à Delhi, et 57 % d'entre eux ont un type d'établissement spécifié, dans l'une des 99 catégories renseignées qu'il s'agisse d'une école, d'un spa, ou d'un centre commercial. Nous reviendrons plus en détail les *POI* et leur utilisation des *POI* dans la typologie des quartiers de Bangkok (chapitre 9).

À noter que les résultats des requêtes peuvent varier en fonction de la langue et de la zone géographique de l'adresse IP. Ceci ajoute à l'effet de bulle filtrante (*filter bubble*) qui tend à présenter des informations aux membres de plateformes en ligne en accord avec ce que ces dernières estiment être les opinions des intéressés (Pariser, 2011), une dimension géographique, où le lieu d'émission de la requête influence de manière similaire les résultats visibles par les utilisateurs (Graham et Zook, 2013).

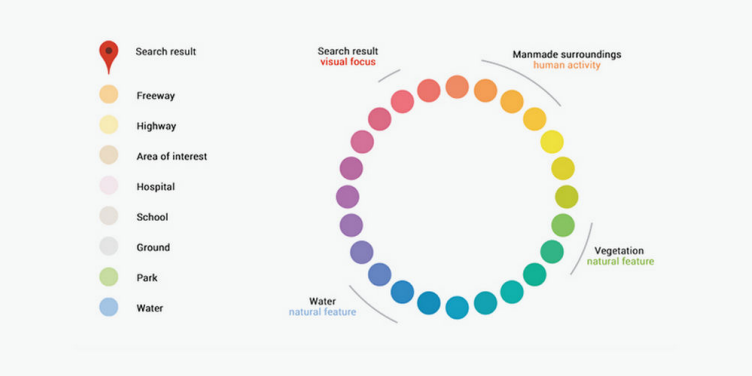


FIGURE 145 Le code couleur des cartes de *Google*. https://blog.Google/products/maps/discover-action-around-you-with-updated/

Outre la collecte de ces *POI*, la figure 145 montre que certains éléments de l'utilisation du sol sont déjà classés par *Google*. Les parcs apparaissent en vert, les écoles en gris clair et le marché en orange (figures 145, 143 et 146), couleur qui correspond aux *AOI* (Areas of Interest). Ces derniers sont définis d'après *Google* comme des « lieux où il y a beaucoup d'activités et de choses à faire » et sont déterminés par un algorithme semi-automatique qui fait ressortir les zones où les concentrations de bars, de restaurants et de commerces sont les plus fortes [355]. C'est effectivement ce que nous observons car parmi les 5512 *POI* présents dans des *AOI* (7 % des *POI* ayant un type d'activité renseigné) 41 % sont des commerces, 15 % des banques, 15 % des restaurants ou des cafés, 5 % des hôtels. Ces différentes zones définies par un code couleur peuvent être très facilement extraites, pour peu que nous puissions récupérer

355. "Areas of Interest places where there's a lot of activities and things to do (..) We determine "areas of interest" with an algorithmic process that allows us to highlight the areas with the highest concentration of restaurants, bars and shops. In high density areas like NYC, we use a human touch to make sure we're showing the most active areas" https://b og.Goog e/products/maps/d scover act on around you w th updated/

la carte géoréférencée. Cette dernière opération est rendue possible en adaptant la fonction *get_googlemap* de la librairie ggmap (Kahle et Wickham, 2013) qui accède à l'*API Google* Static Map [356], de manière à obtenir en sortie un raster sans les différents labels (nom de rue, nom de quartiers, etc) (figure 146).

La carte 146 ci-dessous montre les éléments qui ont été extraits de la zone et bien que le marché (en orange), les écoles (en gris clair) et les différents parcs (en vert) ressortent bien, les *malls* de Saket ne sont pas considérés comme des *AOI* et n'apparaissent pas au niveau de zoom choisi ici. Toutefois cette méthode d'extraction de l'information géographique des cartes de *Google* permet de récupérer très rapidement des zones de différentes catégories (zones commerciales, écoles, hôpitaux et parcs [357]).

FIGURE 146 Résultat d'une extraction des lieux d'éducation (gris), des parcs (vert) et des *AOI* (orange) de *Google Maps*, en rayé.

Bien que l'information ne soit là encore pas exhaustive, nous allons combiner ces données avec celles issues d'*OSM* afin de compléter notre utilisation du sol à Delhi [358].

Résultats

La figure 147 montre la couverture spatiale de notre couche d'utilisation du sol à Delhi, combinant les informations d'OSM et de *Google Maps*. Alors que le sud de la ville semble plutôt bien pourvu, nous pouvons noter l'absence d'information géographique dans un grand nombre de zones de la ville, tant au nord-est (Shadhara) qu'à l'ouest de l'aéroport (Dwarka), à

356. https://deve opers.Goog e.com/maps/documentat on/stat c maps/?h =fr
357. Les parcs et les zones végétalisées pourraient être simplement déduits en utilisant des images satellites et un appliquant un seuil sur un indice impliquant l'infrarouge, comme le NDVI.
358. Les conditions d'utilisation de *Google* sont défavorables à ce genre d'approche, mais nous considérons qu'elles ne s'appliquent pas à une démarche de recherche. https://deve opers.Goog e.com/maps/terms?h =fr.

l'ouest de Pashim Vihar et Rohini, et même dans Old Delhi. Ces zones sont pourtant densément peuplées, et devraient compter des commerces, écoles ou hôpitaux.

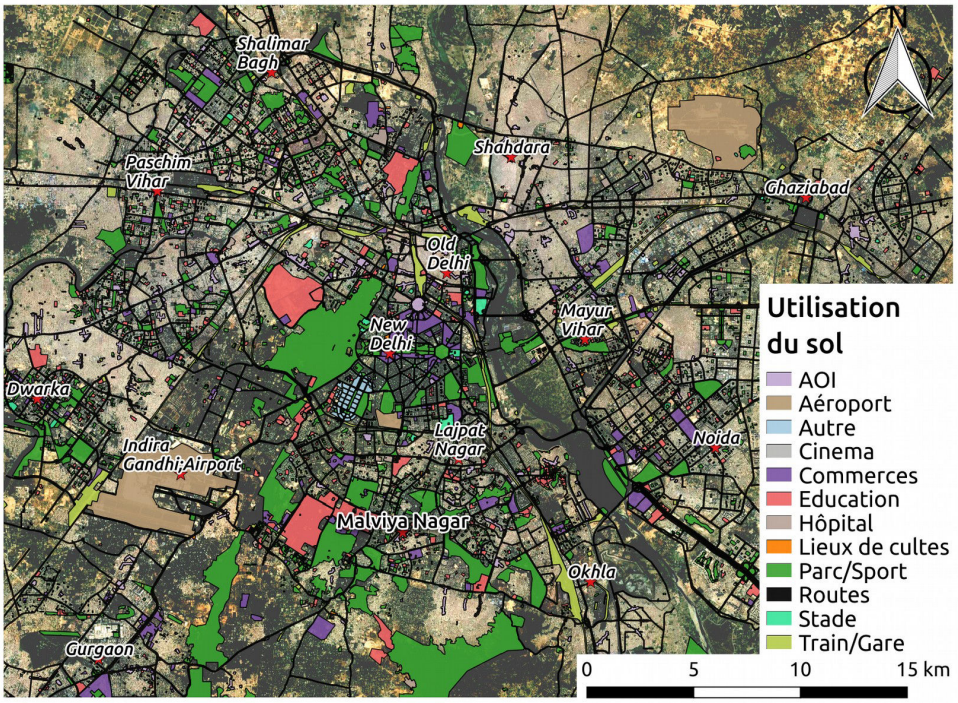


FIGURE 147 Utilisation du sol à Delhi en combinant les données d'OSM et de *Google Maps*.

La figure 148 ci-dessous est un zoom sur le quartier de Malviya Nagar. Certains petits parcs ne figurent pas sur la carte, tout comme certains lieux de cultes pourtant de taille assez importante, comme la Gurduwra ou le temple de Lakshmi prés du marché central. À noter également l'absence des commerces locaux, comme ceux disséminés dans d'Hauz Rani, ou dans Khirki extension. Ainsi, malgré l'agrégation de différentes bases de données géographique, la couverture spatiale de la couche d'utilisation du sol reste très limitée.

Néanmoins, l'utilisation du sol étant définie, avec les nombreuses limites évoquées précédemment, il ne nous reste plus qu'à intersecter les différents lieux correspondants aux espaces d'activités de chaque utilisateur avec cette dernière couche pour obtenir une activité potentiellement exercée. Nous reprenons pour cela les différents clusters définis dans le chapitre 6, et regroupons ceux qui se trouvent dans un même polygone. Dans le cas où une personne a différents lieux qui se trouvent dans un même polygone de taille importante, comme certaines universités ou parcs, ces lieux sont alors fusionnés.

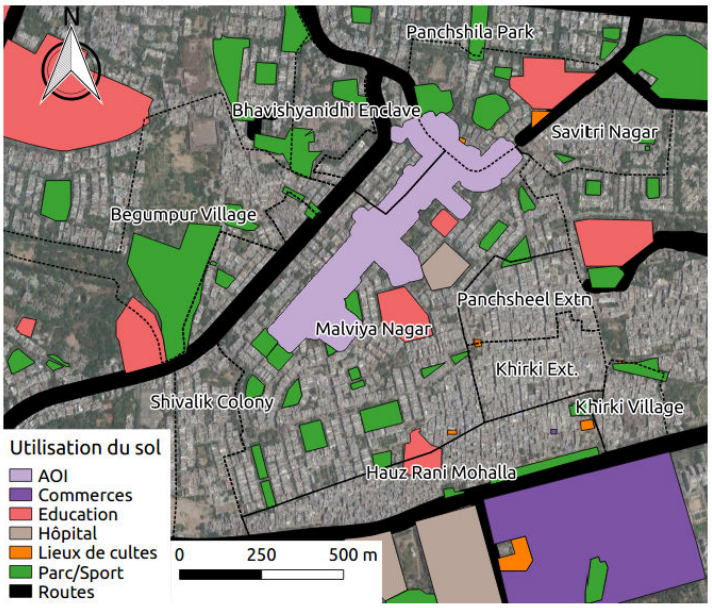


FIGURE 148 Utilisation du sol à Malviya Nagar en combinant les données d'OSM et de *Google Maps*.

1.3 Comment sont composés les espaces d'activité ?

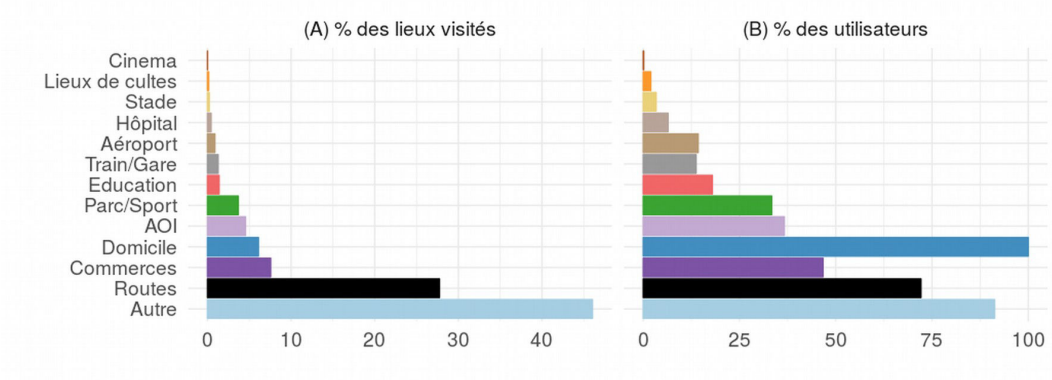


FIGURE 149 Répartition des types d'activités parmi tous les lieux fréquentés (A) et part des utilisateurs exerçant au moins une fois une des activités (B).

Nous pouvons maintenant examiner globalement comment se répartissent ces différentes activités des utilisateurs de *Twitter* à Delhi. La figure 149.a montre que plus de 45 % des lieux des fréquentés durant la période d'enregistrement (~1ans et demi) ne font pas partie de la couche d'utilisation du sol (catégorie "Autre"). Ceci peut s'expliquer par le caractère non exhaustif de notre couche géographique. Cette catégorie peut donc englober toutes sortes d'activités non référencées dans notre base de données, mais aussi des activités comme visiter des proches, se promener dans un quartier donné ou exercer son activité principale, etc. Beaucoup de *tweets* sont également envoyés depuis des routes majeures, ce qui nous donne des informations sur

les transitions entre deux activités (notamment les horaires de commutation). Les activités se déroulant dans des zones commerciales (AOI, commerces) sont très représentées, tout comme les parcs et lieux de sport (terrains de jeux, terrains de crickets, etc.). Seulement 3, 94 et 101 lieux correspondent respectivement à des cinémas, lieux de cultes et stades.

La figure 149.b indique quant à elle la part de personne qui exerce au moins une fois une activité. Quasiment tous les utilisateurs ont *tweeté* dans des zones de type "autre", près des trois quarts sur des routes. Les lieux de commerces et les parcs sont également très largement représentés. Les lieux d'éducations sont visités par près de 25 % de notre échantillon, ce qui peut nous laisser supposer qu'une grande partie des utilisateurs que nous avons enregistrés sont des écoliers ou des étudiants. Environ 10 % des personnes ont *tweeté* dans des gares et des aéroports, ce qui suggère que ces personnes ont un champ de mobilité qui dépasse l'emprise de la ville. Alors que la plupart des personnes que nous avons interviewées (chapitre 5) ont déclaré fréquenter des lieux de cultes, seul 2 % de notre échantillon *Twitter* a laissé une trace numérique dans de tels endroits, ce qui peut toujours s'expliquer par les lacunes de notre couche d'utilisation du sol, et probablement par le fait que *tweeter* dans de tels lieux peut être perçu comme inconvenant et inapproprié. Les mêmes raisons peuvent être invoquées pour expliquer que seulement 3 personnes ont *tweeté* depuis un cinéma. Compte tenu du faible nombre de *tweets* envoyés dans les lieux de cultes, les stades et les cinémas, nous regroupons ces derniers lieux sous la classe "Autre1".

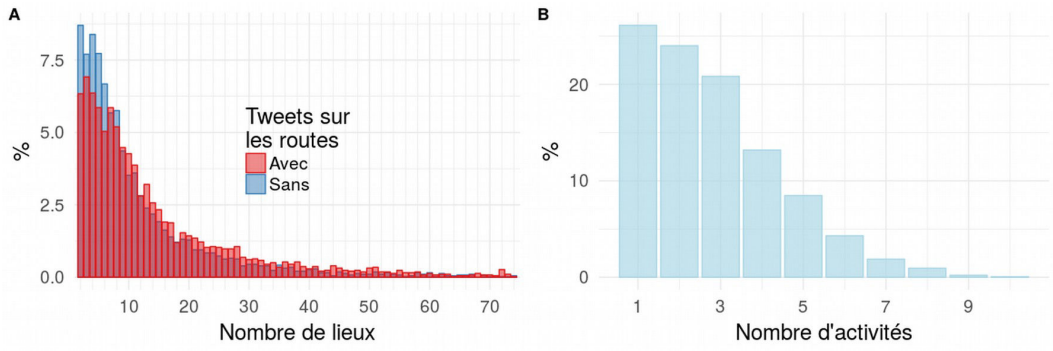


FIGURE 150 Part des utilisateurs selon le nombre de lieux fréquentés (A) ou d'activités réalisées (B). Le domicile et les *tweets* émis depuis des routes importantes ne sont pas pris en compte dans (B).

Ainsi, du fait du nombre limité de catégories présentes dans notre couche d'utilisation du sol, une personne de l'échantillon ne peut ici effectuer plus de 11 activités différentes, en comptant son lieu de domicile. La figure 150 présente la répartition des effectifs selon le nombre de lieux fréquentés et du nombre d'activités différentes réalisées sans compter la présence au domicile ou sur une route majeure.

Une étude très récente, basée sur des données téléphoniques et des données GPS a montré

qu'un individu fréquente sur plusieurs mois environ 25 lieux en moyenne (Alessandretti *et al.*, 2018). Ce n'est pas ce que nous observons sur la figure 150.a, où finalement la plupart des personnes ont laissé des traces numériques dans moins de 10 lieux différents. Si nous considérons également les *tweets* émis depuis des routes, nous notons une augmentation dans le nombre de lieux fréquentés, mais qui n'est pas si importante que cela nous sommes en tout cas bien loin d'obtenir un mode ou une moyenne à 25. N'ayant pas analysé en détail les jeux de données et les traitements de l'étude d'Alessandretti *et al.* (2018), nous ne pouvons que poser hypothèse que nos données sous-estiment la taille de l'espace d'activité réel de notre échantillon, malgré la période d'acquisition relativement longue. Hypothèse renforcée par le caractère épisodique de l'activité sur *Twitter* et dans l'envoi de messages géolocalisés, et par le fait que près de 25 % de notre échantillon ne fréquente qu'un seul lieu en plus de son domicile (sans compter les *tweets* sur les routes), figure 150.b, ce qui ne parait pas très réaliste.

1.4 À quel moment une activité est-elle effectuée ?

Malgré les nouvelles limites exposées à l'instant, nous pouvons toutefois apprécier les horaires de fréquentation des différentes activités, afin d'observer si ces temporalités sont intuitives ou non. La figure 151 montre le nombre d'utilisateurs par tranche horaire selon les catégories de lieux fréquentés. Les lieux que nous avons définis comme étant des domiciles sont les plus représentés, et montrent un premier petit pic en milieu de matinée et un pic nettement plus net en soirée. À noter que plus de messages sont envoyés en journée lors des week-ends (surtout le dimanche) et que les écarts entre les pics du matin et du soir sont moins marqués. Malgré une amplitude nettement plus faible, les lieux associés à des parcs ou à des activités sportives présentent un profil assez proche de ceux du domicile, avec un pic du soir plus tôt que celui du domicile, ce qui pourrait s'expliquer par le fait qu'il s'agit d'activités réalisées en général après le travail et avant de rentrer chez soi.

Les routes exhibent des pics de fréquentation le matin et le soir du lundi au vendredi, tout comme les trains et les gares ce qui nous donne une information sur les périodes de transitions entre deux activités et les navettes domiciles / travail du matin et du soir. Les lieux définis comme des zones d'intérêts (*AOI*) sont fréquentés de manière assez équivalente tout au long de la semaine, avec une activité légèrement plus importante le week-end, contrairement aux zones définies comme commerciales par OSM où la fréquentation baisse le samedi et le dimanche. Ceci paraît contre-intuitif et pourrait s'expliquer par des erreurs de nomenclature et de qualification de l'utilisation du sol dans la base OSM pour les catégories de type « commerce ». Les lieux d'éducation sont plus fréquentés la semaine que les week-ends, mais les universités sont aussi des lieux de vies dotés de campus, ce qui peut expliquer pourquoi un nombre assez important de messages y sont envoyés les samedi et dimanche. Les lieux qui ne sont pas dans notre couche d'utilisation du sol accueillent un grand nombre de personnes, surtout l'après-midi et le soir,

sans réelles différences selon le jour de la semaine.

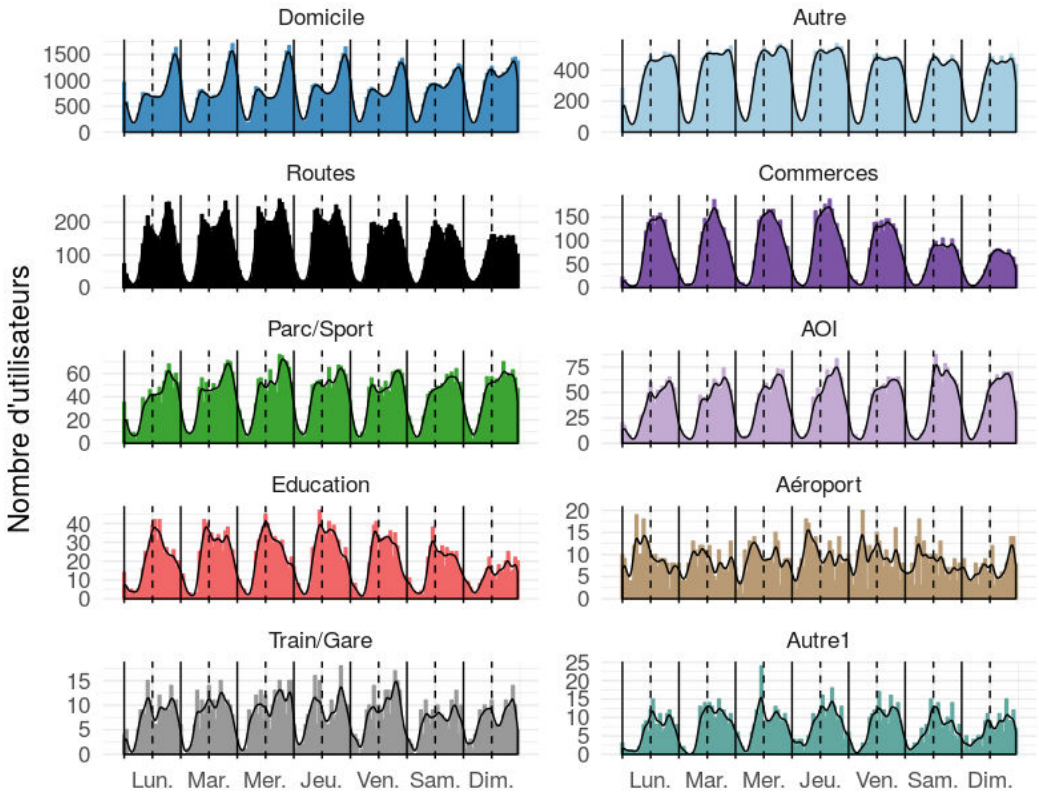


FIGURE 151 Nombre de personnes par catégorie de l'utilisation du sol, par tranche horaire sur une semaine type. La courbe noire représente une moyenne mobile sur 2 heures.

Malgré des effectifs assez faibles, nous pouvons noter que ces profils sont assez crédibles pour la plupart des activités. Dès lors, l'activité principale exercée par un individu, primordiale dans l'étude des mobilités quotidiennes pourrait elle aussi avoir une signature assez caractéristique.

1.5 Quelle est l'activité principale d'un utilisateur ?

Cette section va donc rechercher parmi tous les lieux où un individu a laissé des traces numériques celui où pourrait se dérouler son activité principale, en dehors de son lieu probable de résidence. Pour cela, nous allons poser quelques hypothèses sur les propriétés temporelles et fréquentielles (heure d'envoi des messages et nombre de messages envoyés) et le type d'utilisation du sol où se déroule une telle activité.

Les lieux d'éducations sont assez fréquentés (figures 149.b et 151) ce qui suggère qu'un grand nombre de lycéens ou étudiants font partie de l'échantillon. Nous reviendrons sur ces

individus un peu plus tard, en définissant tout d'abord les zones candidates pour accueillir l'activité principale comme faisant partie des catégories suivantes : « AOI », « Commerces », et « Autre », ce qui élimine les parcs et les transports. Nous appliquons un seuil minimum de messages, à savoir au moins 5 *tweets* envoyés en journée pendant au moins 5 % des jours où la personne a été active sur *Twitter*. Nous posons également qu'au moins deux fois plus de *tweets* y ont été envoyés entre 7 h et 19 h qu'entre 7 h et 19 h. Si plusieurs lieux sont candidats pour un même utilisateur, nous prenons celui depuis lequel le plus grand nombre de messages y a été émis entre 7 h et 19 h. 455 utilisateurs ont un lieu qui remplit ces conditions.

Nous recherchons maintenant les probables étudiants parmi les personnes n'ayant pas de lieux répondant aux critères précédents. Ce choix de commencer d'abord par trouver l'activité principale dans les lieux non dédiés à l'éducation permettra de dissocier les étudiants des gens qui peuvent déposer leurs enfants à l'école et laisser des messages sur *Twitter* par exemple. Nous posons qu'un étudiant a *tweeté* de manière cumulée au moins 5 % de ces jours actifs depuis un lieu d'éducation et au moins 3 jours différents dans un même lieu. Si plusieurs établissements scolaires sont candidats, nous choisissons celui dont le nombre de jours de *tweets* est le plus important. Nous n'appliquons pas de distinction entre des messages envoyés le jour ou la nuit, car la plupart des universités sont dotées de campus où les étudiants peuvent circuler le soir. Au final, 179 personnes ont un lieu qui remplit ces conditions.

Finalement, 634 utilisateurs (dont 15 à Malviya Nagar) ont un lieu que nous estimons comme un bon candidat pour être l'activité principale, soit 15,7 % de l'ensemble des utilisateurs dont nous avons estimé le domicile. Cet écrémage qui résulte d'un compromis entre l'application de seuils et des hypothèses de fréquentation de lieux autorise un niveau de confiance relativement correct. Il est néanmoins possible que soit considéré comme activité principale le domicile d'un tiers, ou bien un lieu où une personne s'adonne régulièrement à des activités de shopping et le partage sur Internet.

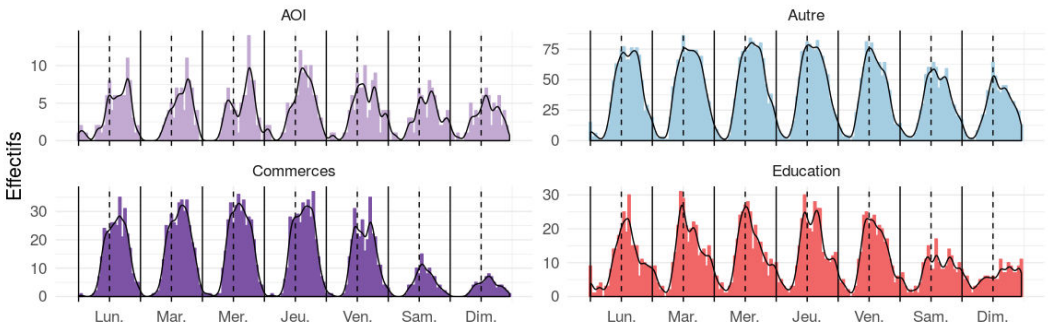


FIGURE 152 Horaire de présences à l'activité principale en fonction du type de lieu associé à cette dernière.

Nous n'avons pas posé d'hypothèses sur la distinction entre jours de la semaine et week-end. La figure 152 ci-dessous montre le nombre de *tweets* par tranche horaire en fonction du type d'utilisation du sol du lieu de travail estimé. Nous observons une diminution de l'activité le week-end nettement plus marquée que sur la figure 151 qui prend en compte l'ensemble des utilisateurs de *Twitter*. La localisation des activités principales dans la ville (figure 153) conforte encore la cohérence de notre méthode puisque mis à part la zone de South Delhi, les plus grandes densités sont enregistrées autour de la zone très commerçantes d'Old Delhi, ou encore à Connaught Place et dans les villes satellites de Noida et Gurgaon, sièges de nombreuses entreprises et commerces.

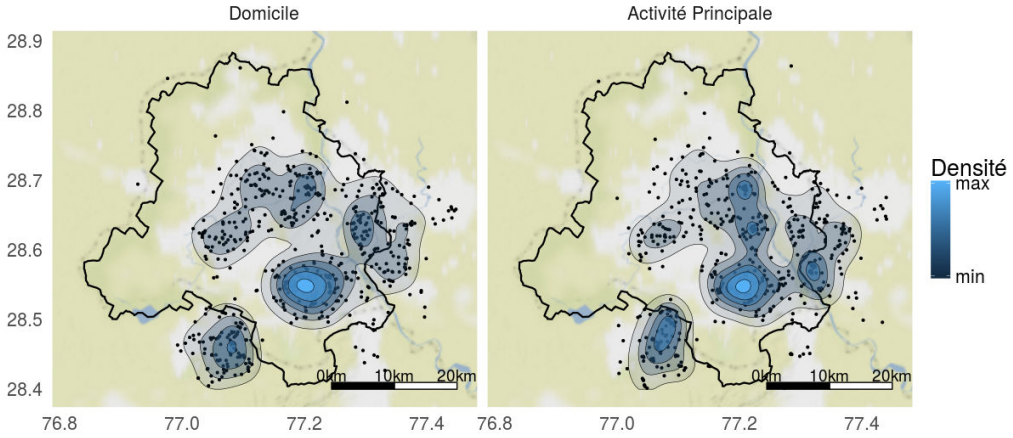


FIGURE 153 Localisation des domiciles (gauche) et des activités principales probables (droite) des utilisateurs de *Twitter* à Delhi. Sont également représentés les différents clusters de densités de chacun de ces lieux dans la ville.

La localisation des domiciles dans les quartiers les plus aisés de la ville et celle des activités principales plutôt dans les centres d'affaires confirme une nouvelle fois que cet échantillon n'est pas représentatif de la population de Delhi, mais plutôt constitué de personnes relativement aisées. Les données *Twitter* agrégées permettent de faire ressortir des pulsations urbaines associées à cette catégorie de population (figure 140). Associées à une utilisation du sol, elles permettent de dériver des informations sur les espaces fréquentés collectivement à différentes heures de la journée (figure 151). L'ajout pour chaque individu d'une localisation probable de son activité principale et de son lieu de résidence peut également servir dans l'étude des navettes domicile / travail (figure 153). Nous pourrions aller plus dans les détails, en montrant par exemple différentes métriques de déplacement et de dispersions, essayer d'estimer les interactions entre les différentes zones de la ville, tenter de faire des sous-groupes, etc. Mais comme nous l'avons vu précédemment, Delhi n'est probablement pas le bon terrain pour de telles analyses, avec seulement 4089 utilisateurs (non représentatifs de la population), dont nous avons pu estimer le domicile, parmi lesquels 634 où nous avons pu faire ressortir la

localisation probable de leur activité principale. Bangkok semble quant à elle plus propice à ce genre d'analyse, au regard de la taille et de la représentativité spatiale de l'échantillon et nous aborderons donc plus en détail ces questions d'analyses spatiales et profils individuels dans la partie D de ce manuscrit.

Néanmoins, il nous semble pertinent d'apprécier dans quelles mesures il est possible de dériver des agendas individuels à partir de données *Twitter* et d'utilisation du sol, en vue d'une modélisation à base d'agent. Ici, la finalité résidera plus dans le développement d'un modèle alimenté par ce type de donnée, que nous pourrons ensuite appliquer à d'autres zones géographies (*e.g.* Bangkok), ou à d'autres données (*e.g.* issus d'entretiens) que dans la génération d'agents représentatifs de la population.

2 Des espaces d'activités discrets à des agendas individuels continus

2.1 Limites des espaces d'activités « bruts »

Comme vu dans le chapitre 6, et dans d'autres articles traitant de données *Twitter* (Jurdak *et al.,* 2015 ; Liu *et al.,* 2015, etc.), les durées entre deux messages successifs suivent une loi puissance à large queue (*heavy tailed),* et la plupart des intervalles de temps est très court et une faible proportion est très longue. Au-delà d'impliquer des processus non-Poissonnien (Barabasi, 2005), ces irrégularités dans l'envoi de messages entraînent de longues périodes (de plusieurs heures à plusieurs jours) où aucune information n'est disponible sur la localisation d'une personne, avec des niveaux de régularité très différents entre les individus. Ainsi, le caractère épisodique et les grandes disparités de volumes et de fréquences entre utilisateurs lors de l'envoi de messages compromettent l'extraction de trajectoires individuelles fiables (Andrienko *et al.,* 2012 ; Ferrari *et al.,* 2011 ; Grossenbacher, 2014). En somme, la trajectoire réelle d'une personne dans une ville ou succession temporelle des lieux qu'elle fréquente est probablement assez éloignée de la suite chronologique de ses traces numériques prises *stricto sensu*. Il y a donc un temps et un espace propre à l'activité des réseaux sociaux qu'il faudrait analyser au regard des temporalités et des espaces réellement fréquentés par les individus, que seules des enquêtes *ad hoc* auprès d'utilisateurs de *Twitter* permettraient d'enrichir.

Une méthode pour dépasser ces limites pourrait être de partir des informations sur les espaces d'activités bruts pour générer des agendas individuels continus dans le temps relativement crédible, à l'image de ce qui a été effectué dans le chapitre précédent ou des travaux de Perkins et al, (2014). L'objectif serait d'affecter pour chaque individu une activité dans un lieu donné par tranche horaire, en mélangeant une approche probabiliste dans l'attribution des activités et des hypothèses fortes en distinguant des activités routinières (domicile / travail) des activités moins fréquentes. Bien que certaines études ont pu montrer qu'il était possible de

prédire la localisation d'un individu en prenant en compte les traces numériques de ces « amis » en ligne (Backstrom *et al.,* 2010 ; Davis Jr. *et al.,* 2011 ; Sadilek *et al.,* 2012), nous ne sommes pas ici en mesure d'estimer les lieux qu'une personne a fréquentés sans y avoir envoyé de *tweets* géolocalisés. Faute de pouvoir réaliser le type d'études précédemment mentionnées, nous posons l'hypothèse que les différents lieux de l'espace d'activité issus des données *Twitter* enregistrés sur un large corpus et sur une période temporelle relativement longue sont représentatifs de l'espace d'activité réel.

2.2 Proposition d'algorithme

La génération d'agendas et par extension d'espaces d'activités continus dans le temps requiert, *a minima* la connaissance (ou l'estimation) de 5 paramètres : le nombre de lieux fréquentés et leurs activités associées, ainsi que l'heure, la durée et la fréquence de visite de ces lieux (Perkins *et al.,* 2014). L'algorithme que nous allons présenter ici reprend en grande partie les méthodes utilisées pour générer des agendas à partir d'informations récoltées sur le terrain (chapitre 7), que nous avons adaptées à la nature des données propres à *Twitter*.

Nous appliquons une approche différenciée s'il s'agit d'une activité routinière qui sera visitée toutes les semaines (domicile et travail) ou d'une activité plus flexible dans le temps ou l'espace et raisonnerons sur agenda réalisé pour 4 semaines, en moyenne annuelle, sans distinction de périodes de vacances. Pour des raisons de simplicité, nous ne prendrons pas en compte les lieux associés à des routes, que nous considérons comme des activités transitoires entre deux visites de lieux.

Nous allons dans un premier temps définir pour chaque personne si un lieu (hors domicile et activité principale) est visité une semaine donnée (étape 1), puis nous allons choisir les jours (étape 2) et les tranches horaires où celui-ci sera fréquenté (étape 3). Nous ferons de même pour l'activité principale et le lieu de domicile (étape 4). Quelques ajustements seront effectués (étape 5), permettant la reconstitution de l'agenda.

Nous posons au préalable L, l'ensemble des lieux l fréquentés par un individu et A, l'ensemble des activités a effectuées dans chaque lieu l.

2.2.1 Étape 1 : définir une semaine de réalisation

- Fréquence de visite hebdomadaire

Nous définissons tout d'abord pour chaque lieu l de l'espace d'activité, une fréquence de visite hebdomadaire S_l comme étant le rapport entre le nombre de semaines différentes W_i où un individu a *tweeté* dans un lieu l et le nombre de semaines totales W_{tot} où l'individu a

tweeté sur la période. Soit $S_l = W_l/W_{tot}$

- Probabilité de visite une semaine s

Si S_l est égal à 1, dans ce cas le lieu sera visité toutes les semaines. Sinon, pour chaque semaine *s* nous effectuons un tirage aléatoire [359] booléen où la probabilité que le lieu *l* soit visité est égale à S_l, et la probabilité qu'il ne soit pas visité à 1-S_l. Selon que le lieu est visité ou non la semaine *w*, nous modifions la valeur de S_l pour *w+1* selon :

$$S_{i,w+1} = S_{i,w} \pm \frac{S_{i,w}}{5-w} \tag{23}$$

Avec l'utilisation de l'addition si le lieu n'a pas été sélectionné pour être visité lors de la semaine *w*, ou la soustraction s'il a été tiré au sort (avec 5 le nombre de semaines de simulation, ici 4, plus 1). Ceci permet de se rapprocher d'un tirage sans remise et d'augmenter la probabilité de visite de lieux dont les traces numériques sont très anecdotiques.

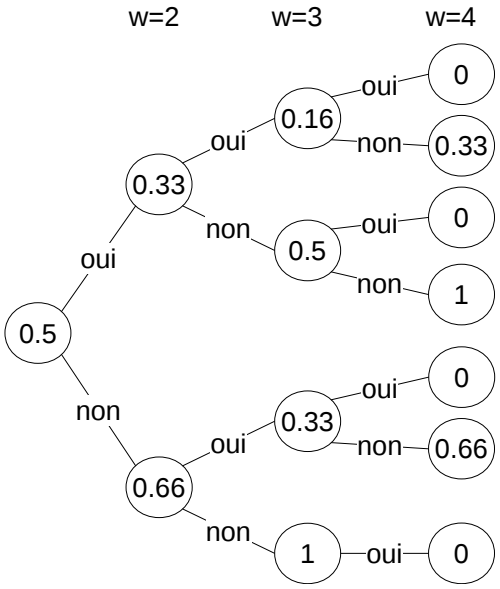


FIGURE 154 Évolution des probabilités de réalisation d'une activité ayant une fréquence de visite hebdomadaire initiale de 0.5.

La figure 154 montre par exemple que les probabilités de tirages pour un lieu ayant une fréquence de visite hebdomadaire initiale de 0.5 entraînent qu'il aura de grandes chances d'être fréquentés 2 fois, un peu moins d'être visité que 1 ou 3 fois et aucune de ne pas être visité durant le mois.

Autre exemple, avec un lieu qui a une fréquence de visite hebdomadaire initiale très faible, de 0.1 et qui n'est pas tiré les trois premières semaines. À la semaine 2, sa probabilité d'être

359. en utilisant la fonction "sample", native dans R

sélectionnée sera de 0.1+ 0.1/(5-2), soit 0,133. À la semaine 3 de 0.133+0.133/(5-3), soit 0.2, et finalement de 0.2+0.2/(5-4), soit 40 % de chance d'être visité la dernière semaine.

À la fin de cette étape, tous les lieux se voient donc affecter entre 0 et 4 semaines de visite, selon leur fréquence de visite hebdomadaire initiale et des déroulements des tirages aléatoires. Il convient maintenant de choisir les jours où un lieu sera visité.

2.2.2 Étape 2 : définir les jours de visites

Nous allons ici définir une nouvelle fréquence de visite d'un lieu *l*, F comme le rapport entre le nombre de jours J_l où une personne a *tweeté* dans un lieu *l* et J_{tot} le nombre de jours où la personne a *tweeté* globalement, que nous multiplions par 7. Si cette fréquence est inférieure à 1, nous lui affectons 1. Nous obtenons ici une valeur positive, entre 1 et 7, que nous posons être le nombre de jours qu'un lieu peut être visité une semaine donnée.

$$F_l = \frac{J_l}{\sum J_{tot}} \times 7 \quad \& \quad si \quad F_l < 1 \rightarrow F_l = 1 \tag{24}$$

Si F n'est pas un nombre entier, nous arrondissons aléatoirement soit à l'entier inférieur, soit supérieur. Nous réitérons cette opération pour toutes les semaines où ce lieu est censé être visité. Si un lieu a une fréquence de visite F de 3.5, il pourra par exemple être fréquenté lors de 3 jours une semaine donnée et 4 une autre semaine.

À chaque jour *j*, appartenant à J, la liste des jours de la semaine où la personne a *tweeté*, est associé n, le nombre de jours différents où l'utilisateur a *tweeté* un jour *j* dans le lieu *l*. Si par exemple une personne a *tweeté* lors de quatre mardis différents dans un lieu *l*, dans ce cas $n_{mardi,l}$ =4. Nous posons P_j, la probabilité de *tweeter* un jour *j* comme étant :

$$P_{j,l} = \frac{n_{j,i}}{\sum_{i}^{J} n_{i,l}} \tag{25}$$

et nous définissons ensuite P_J comme la liste des probabilités P de chaque jour *j* de J. Nous tirons ensuite J', une liste de F_l jours parmi *J* réalisés une semaine donnée, avec une probabilité d'être sélectionnée de P_J [360].

Si le nombre de jours F_l est supérieur au nombre de jours différents où la personne a *tweeté* dans ce lieu, nous effectuons un tirage avec remise, et la personne ira plusieurs fois le même jour dans le même endroit mais pas nécessairement aux mêmes heures (voir ci-après).

360. Sous R : J' = sample(x=J, n=F, prob= PJ)

2.2.3 Étape 3 : Définir les plages horaires

Après avoir posé quelle(s) semaine(s) et quel(s) jour(s) un lieu sera fréquenté, nous allons maintenant définir une plage horaire durant laquelle ce lieu sera visité. La première ligne de la figure 155 ci-dessous montre le nombre de *tweets* envoyé par heure dans 6 lieux différents pour 6 utilisateurs pris au hasard. Alors qu'un seul *tweet* fut envoyé par heure pour les lieux 3, 4, 5 et 6, les lieux 1 et surtout le 2 ont enregistré une activité plus importante.

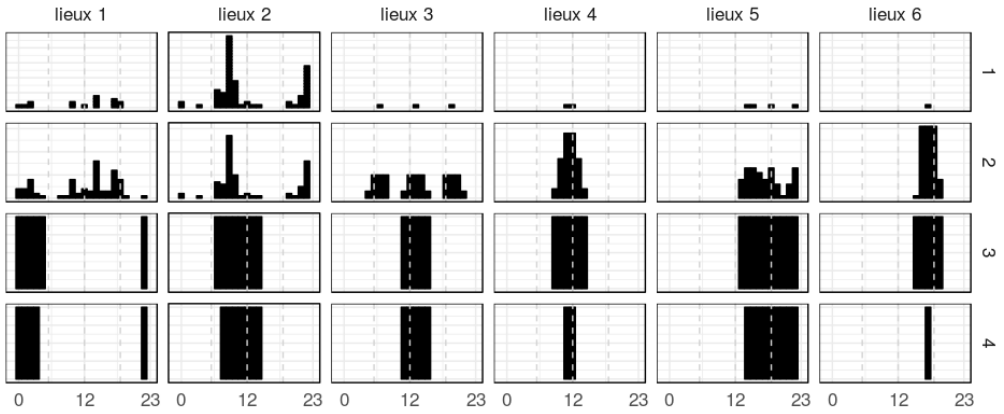


FIGURE 155 Principe de notre traitement en 4 étapes pour définir les plages horaires. Étape 1 : Les *tweets* sont agrégés par heure. 2 : Une fonction de densité carrée de portée 1 h est appliquée pour boucher les trous. 3 : Une plage horaire continue est choisie, pondérée par le nombre de *tweets* dans la plage horaire. 4 : les plages horaires sont bornées, puis les durées redéfinies en fonction de la distance au domicile pour les lieux aux plages horaires étroites et selon une fonction décroissante pour les lieux où la plage horaire est supérieure à 6h.

Plusieurs défis, inhérents au caractère épisodique de l'activité sur *Twitter* d'une personne sont présents ici. En effet, si une personne *tweet* par exemple lors de 3 jours différents, à trois moments de la journée, disons le matin, le midi et le soir. Est-ce que cela signifie que cette personne passe toute une journée dans ce lieu ou est-ce qu'elle le fréquente à différents moments de la journée selon les jours ? Il est délicat de trancher. En revanche, nous pouvons considérer qu'une personne qui *tweet* souvent l'après midi, par exemple à 14 h, 15 h et 17 h, se trouve dans ce lieu également à 16 h.

Nous décidons donc dans un premier temps d'élargir les plages horaires, en appliquant une fonction de densité de portée 1 h avec une fonction rectangulaire. Ceci a pour conséquence d'étendre plus ou moins les plages horaires selon la quantité de messages envoyés aux heures données et de supprimer les trous entre deux heures assez rapprochées (ligne 2). Après cette opération, le lieu 5 affiche ainsi des heures de présence continues tout l'après-midi, tandis que le lieu 3, du fait de l'espacement temporel des *tweets* a toujours 3 plages horaires distinctes.

L'étape suivante (ligne 3) consiste en la sélection d'une plage horaire parmi celles définies précédemment. Pour ce faire, nous allons calculer pour chaque plage horaire la somme des densités définie précédemment, que nous posons être la probabilité que cette plage horaire soit tirée. Le lieu 2 présente 3 plages horaires (une très tôt le matin, une autre entre 6 et 16 h, et une dernière entre 18 h et 1 h), et dans ce cas précis c'est celle qui se déroule entre 6 h et 16 h qui fut sélectionnée, au gré d'un plus grand nombre de *tweets*. Pour le lieux1, c'est la plage horaire nocturne qui fut choisie, même si sa probabilité d'être tirée était moindre que la plage horaire en journée. Pour le lieu trois, ce sont les heures du midi qui ont été sélectionnées, même si les heures du matin ou de l'après midi avaient les mêmes probabilités d'être tirées.

L'étape suivante (ligne 4) est le résultat de plusieurs traitements. Dans un premier temps, les plages horaires ont été bornées par leur valeur maximale et minimale, afin d'éviter que ces dernières soient trop étendues tout en conservant leur continuité temporelle établie précédemment. Deux cas se produisaient alors : soit la durée de l'activité était réduite à une heure car n'ayant qu'un seul *tweet* sur la plage horaire, soit la durée de l'activité était très longue. En effet, l'application d'une fonction de densité peut techniquement entraîner la création d'un spectre horaire très important, si par exemple tous les *tweets* sont espacés de deux heures. Nous avons donc appliqué deux traitements distincts en fonction de la durée de visite du lieu.

Lieux visités sur de courtes durées

Si la durée D dans un lieu n'est que d'une heure, avec H l'heure de la journée, nous appliquons la même méthode que dans le chapitre 7, c'est-à-dire que nous prenons en compte la distance de ce lieu au domicile et estimons que plus le lieu est éloigné, plus la durée D' de présence est importante, avec une limite à 4 heures[361]. Cette opération permet aussi de prendre en compte de manière indirecte les temps de déplacements. Nous convenons néanmoins qu'il serait plus judicieux d'estimer D' selon la distance à l'activité précédente, mais cette dernière n'est pas encore connue à ce stade.

Si D' est supérieur à D, il faut définir à quelle heure H' débute la visite de ce lieu. Commence-t-elle avant H, ou après ? Pour cela, nous tirons une valeur au hasard entre 0 et D' que nous soustrayons à H pour obtenir H'.

Si par exemple, un lieu fréquenté à H = 16 h est situé à 10 km du domicile, nous aurons, selon l'approche présentée dans le chapitre 7, alors plus de chance de tirer D' proche de 4 h que de 1 h. Admettons que nous tirons 3 h. L'heure de début H' pourra être alors équiprobablement 14 h, 15 h ou 16 h, avec une heure de départ 3 h (D') plus tard.

Lieux visités sur de longues durées

361. ou la probabilité de tirer une heure H, comprise entre 1 et 4 dépend de la distance d au domicile avec $H=1/d+d^2$ (voir chapitre 7).

Pour les activités d'une durée D supérieure ou égale à 6 heures, nous posons que la nouvelle durée D' suit une loi de probabilité décroissante de type 6+1/(D"-5), où D" est une valeur tirée au hasard dans une séquence entre 0 et D. Ainsi, D' a une durée minimale forcément égale à 6 h, et une probabilité qu'elle soit plus longue qui décroît selon une fonction inverse. La figure 156 illustre cette approche en montrant la probabilité de voir une nouvelle durée affectée en fonction de la durée initiale après 1000 itérations. Les durées des activités flexibles les plus longues sont donc réduites la plupart du temps, avec un minimum à 6 h, avec toujours la possibilité d'avoir des activités qui s'écoulent sur de longues périodes.

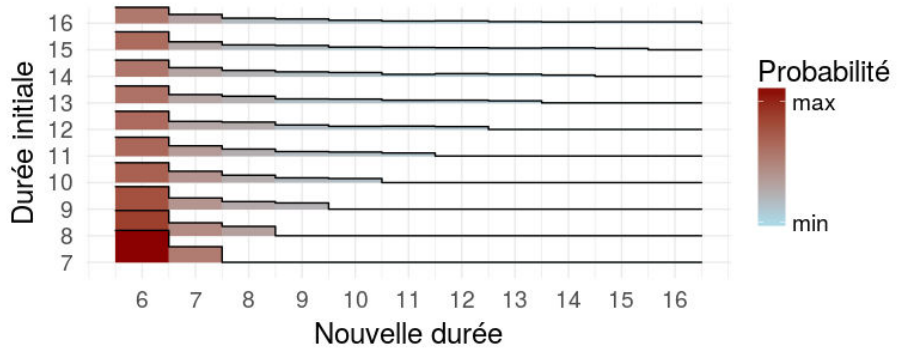


FIGURE 156 Exemples de probabilités de tirer une nouvelle durée D' selon la durée initiale D et l'application d'une fonction décroissante de type D' = 6 + 1/(D"-5).

Si nous reprenons la ligne 4 de la figure 155, alors que nous n'avions qu'un seul *tweet* enregistré dans le lieu 3, sa plage horaire s'en trouve élargie à 4 h, du fait de l'éloignement au domicile. Les lieux 4 et 6 seront visités quant à eux aux mêmes horaires où les *tweets* ont été enregistrés, et les durées des activités se déroulant dans les lieux 1, 2 et 5 sont dans ce cas présent légèrement réduites par rapport à celles définies lors de l'étape 3.

2.2.4 Étape 4 : ajout de l'activité principale et du domicile

Nous posons que l'activité principale s'exerce tous les jours de la semaine où des *tweets* ont été envoyés. Reste à définir les heures de réalisation. Pour cela, nous appliquons à l'étape 3 décrite précédemment, quelques légères variantes.

Nous agrégeons tout d'abord les *tweets* envoyés par heure de la journée, puis nous appliquons une fonction de densité rectangulaire, mais cette fois-ci d'une portée de 2 heures ce qui permet de boucher plus de trous entre deux heures de *tweets*. Si plusieurs tranches horaires sont présentes, nous en sélectionnons une avec une probabilité conditionnée par le nombre de *tweets* envoyés dans chacune des plages continues, et cette plage horaire sera valable pour toute la semaine. Nous bornons ensuite la plage horaire de telle sorte à ce que les heures de l'activité soient comprises entre la première heure où un *tweet* a été enregistré dans la zone et la plus

tardive. Si la durée de l'activité finale est supérieure ou égale à 10 h, nous appliquons la même méthode que précédemment et décrite par la figure 156, mais avec une valeur de 10 h au lieu de 6 h. Ceci permet de faire tendre la durée de l'activité principale vers 10 h, ce qui semble relativement réaliste.

Le domicile est quant à lui fréquenté tous les jours de la semaine, et nous élargissons simplement les plages horaires en appliquant une fonction de densité de portée 3 h et en bornant ces dernières de la même manière que pour les autres activités. Nous combinons ensuite les plages horaires correspondantes aux activités principales, au domicile et aux lieux visités moins fréquemment ce qui nous permet d'obtenir pour chaque individu une séquence de lieux visités de résolution temporelle d'une heure sur une période d'un mois. Enfin, les heures où aucun lieu n'est visité sont associées au domicile.

2.2.5 Étape 5 : Gestion des doublons

L'approche employée par l'attribution en cascade de semaine (étape 1), de jours (étape 2) et d'heures (étape 3) de visite pour chaque lieu est très proche de celle utilisée dans le chapitre précédent. Elle souffre donc des mêmes limites, notamment du fait que certaines activités sont susceptibles d'être réalisées en même temps. En effet, notre algorithme ne pose aucune restriction quant à la possibilité que deux lieux soient visités simultanément une même semaine, un même jour et une même heure. Il conviendra par la suite de prendre en compte ce genre de conflits horaire, que nous appellerons ici des « doublons », dans un algorithme plus poussé.

Mais pour l'instant, appliquons ici une méthode similaire à celle présentée dans le chapitre 7, en détectant dans un premier temps les lieux qui sont visités sur des mêmes tranches horaires. Parmi ces doublons, nous choisirons ensuite un lieu qui sera effectivement visité, avec des probabilités conditionnées par le type d'activité et l'heure de la journée, selon une méthode assez proche de (Banos *et al.,* 2006).

Nous posons qu'entre 6 h et 22 h, les activités autres que le lieu de résidence ont un coefficient de réalisation de 1, tandis que le domicile est moins prioritaire, avec un coefficient de 0,05. Si par exemple l'activité principale, une autre activité et le domicile sont en doublon lors d'une même tranche horaire en journée, les probabilités de tirer un de ces lieux sont respectivement de 1/2.05, 1/2.05 et 0.05/2.05 [362]. Ainsi, en cas de conflit en journée entre une présence au domicile et une autre activité, la personne aura moins de chance de se trouver à son domicile. En revanche, la nuit, entre 23 h et 5 h, nous posons que le coefficient du domicile passe à 0,5, celui d'effectuer son activité principale à 0,05 et visiter un autre lieu reste à 1. Les probabilités de tirer un de ces lieux sont alors respectivement de 0.5/1.55, 0.05/1.55 et

362. Ou 2.5 est la somme de tous les coefficients (0.05+1+1).

$1/1.55$ [363].

Ces coefficients de réalisation sont bien entendu très arbitraires, et selon les hypothèses, l'activité principale peut se voir attribuer un coefficient plus important pour être plus prioritaire en journée par rapport aux autres activités. De même, en cas de doublons nocturnes, nous pouvons aussi poser qu'une personne a en réalité plus de chance d'être à son domicile que d'effectuer une activité dans un lieu de sortie par exemple.

Cette succession de conditions permet d'obtenir pour chaque individu un agenda sur un mois. L'approche par tirage pseudo-aléatoire conditionnée par des probabilités de réalisations dépendantes des statistiques de chaque utilisateur de *Twitter* et des paramètres extérieurs comme le choix de la fonction permettant d'adapter les tranches horaires ou encore les coefficients utilisés dans le choix lors de la présence de doublon permet d'obtenir pour chaque individu des résultats proches, mais légèrement différents selon les simulations.

2.3 Présentations des agendas reconstitués

2.3.1 D'agendas individuels...

La figure 157 montre 6 agendas d'utilisateurs de *Twitter* reconstitués, donc hypothétiques, pour une simulation donnée. Il en ressort que les aspects routiniers apparaissent clairement, avec la succession des présences au domicile et des activités principales. Les utilisateurs 1,2,4 et 6 n'effectuent d'autres activités que de manière occasionnelle, contrairement à l'utilisateur 3 qui a un emploi du temps relativement chargé et varié.

La plupart des membres de notre échantillon *tweetent* peu. Ceci se traduit par un nombre relativement faible d'activités dans nos données, et peut expliquer en grande partie le faible nombre d'activités différentes qui apparaissent dans certains de nos agendas. Mais notre algorithme est probablement responsable du faible nombre d'activités différentes réalisées de manières hebdomadaires et quotidiennes. Nous nous basons en effet sur les fréquences d'envois de messages pour définir la probabilité de visite d'un lieu une semaine et un jour donné. Il est possible que cette méthode soit un peu trop naïve et sous-estime les fréquences de visites.

Concernant les horaires où s'effectue l'activité principale, l'utilisateur 4 travaillera par exemple plutôt les après-midi lors des semaines 1 et 4 et plutôt le matin lors des semaines 2 et 3. Cet aspect provient très probablement de l'étape 3 de notre algorithme, où après avoir élargi et redéfini les heures auxquelles sont susceptibles de se dérouler une activité, nous ne sélectionnons qu'une seule plage d'horaire continue. Il peut s'agir ici d'une piste d'amélioration du protocole, en sélectionnant par exemple 2 tranches horaires continues pour les activités principales, si ces dernières n'impliquent pas une durée totale trop importante. Ceci permettrait

363. Ou 1.55 est la somme de tous les coefficients (0.05+0.5+1).

de prendre en compte le fait qu'un individu ne mange pas nécessairement sur le lieu de son activité principale par exemple. L'utilisateur 2 a des durées de travail qui varient grandement d'une semaine sur l'autre. Ceci est dû au fait que nous tirons une nouvelle durée si cette dernière dépasse les 10 h et que nous l'appliquons sur l'ensemble de la semaine et non sur l'ensemble de la période.

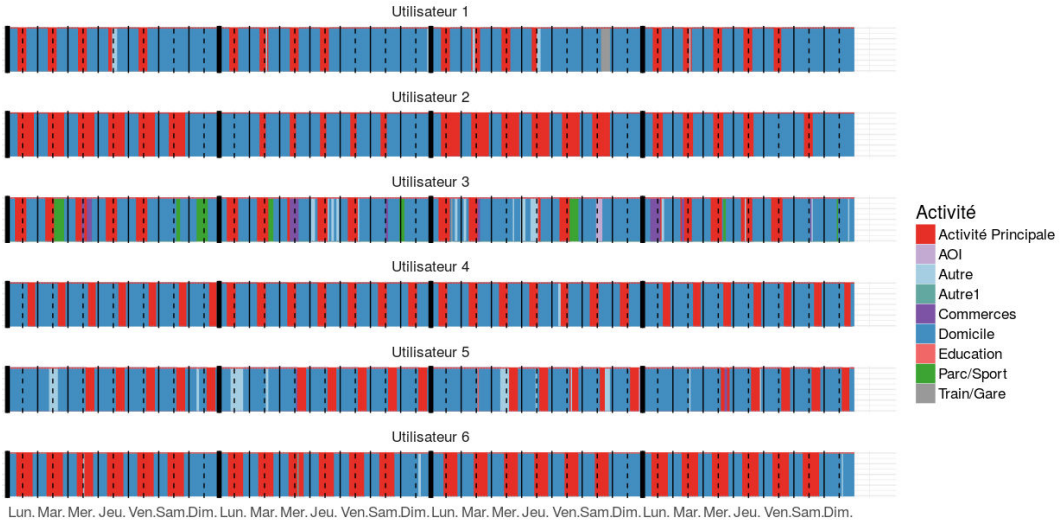


FIGURE 157 6 agendas reconstitués pour autant d'utilisateurs, sur une simulation donnée, à partir des données *Twitter* à Delhi et de notre couche d'utilisation du sol.

En somme, même si le potentiel d'amélioration de ces agendas est assez élevé, quelques critères et paramètres que nous pouvons ajuster sont déjà identifiés et feront l'objet de travaux ultérieurs (chapitre 11). Néanmoins, les agendas reconstitués demeurent assez crédibles, tant sur la succession des activités que sur leurs horaires de réalisation.

2.3.2 ...A des groupes d'utilisateurs ?

Nous allons maintenant regrouper les différents agendas générés selon la méthode de l'appariement optimal présenté dans le chapitre précédent. L'objectif est de voir si des groupes d'individus aux comportements (agendas) similaires ressortent et s'il existe des différences marquées entre ces groupes d'individus.

Lors du calcul des distances entre les agendas (en nous basant sur les séquences d'activités) nous pondérons certaines activités, en donnant un poids plus important aux transitions qui ont lieu hors du domicile et de l'activité principale, avec des poids de 0,1 pour ces dernières activités et de 2 pour les autres. Ceci permet théoriquement de réduire l'importance des activités routinières, communes à l'ensemble des agendas, dans le calcul des distances entre les agendas[364]. À partir de ces métriques calculées, nous appliquons une CAH, dont la

364. Il s'agit ici de compter entre chaque agenda le nombre d'ajustements minimum nécessaires – ajout,

répartition des parts d'inerties suggère un choix de 5 classes, mais la structure et la répartition des groupes nous incite à en choisir 6 (figure 158).

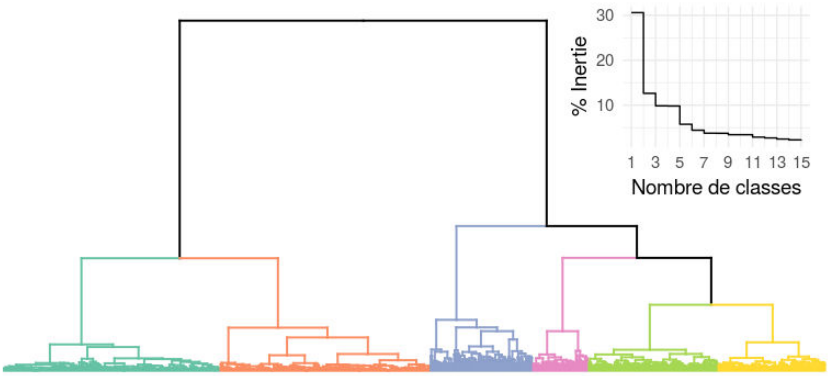


FIGURE 158 Résultat d'une CAH sur la matrice de dissimilarité des agendas reconstitués issus de *Twitter* à Delhi.

Les agendas moyens de ces 6 classes sont présentés dans la figure 159. Nous pouvons ici noter deux éléments relativement discriminants : l'importance de l'activité principale d'un côté (cluster 1, 2 et 3), et la propension à effectuer des activités autres que *principale* et à rester à son domicile. Les clusters 1 et 3 ont un profil qui ne laisse que peu de places aux activités plutôt flexibles, contrairement aux clusters 2, 4, 5 et 6. Le cluster 4 regroupe les personnes qui visitent assez régulièrement dans la semaine des zones de type non déterminées, de type « autre », alors que les individus regroupés dans le cluster 5 fréquentent, certes de manière assez épisodique, des lieux de type « parc/sports », des lieux de cultes ou encore des commerces.

Probablement mieux adaptée à l'analyse des trajectoires de vie et/ou la mobilité sociale, cette méthode de regroupement de séquences temporelles montre ici ces limites, car seules les tendances très générales ressortent. En effet, l'importance quotidienne des activités principales et du domicile semble contribuer très fortement à la structure des groupes, au détriment des activités moins fréquentes, malgré une pondération censée mettre en avant ces dernières. Ceci est probablement dû à la durée des séquences, très longues (672 plages horaires), ainsi qu'au nombre très important de transitions dans une sous-période relativement courte (ici la journée). Mais des explications plus liées à notre méthode peuvent aussi être invoquées, comme un nombre d'activités assez important (jusqu'à 12 activités différentes), mais surtout à des agendas peut être un peu trop simple, où les navettes domicile-travail dominent largement les autres types de déplacement. D'autres méthodes de regroupement seront présentées dans le chapitre 11.

suppression ou changement d'activité à chaque tranche horaire – pour que ces derniers soient identiques.

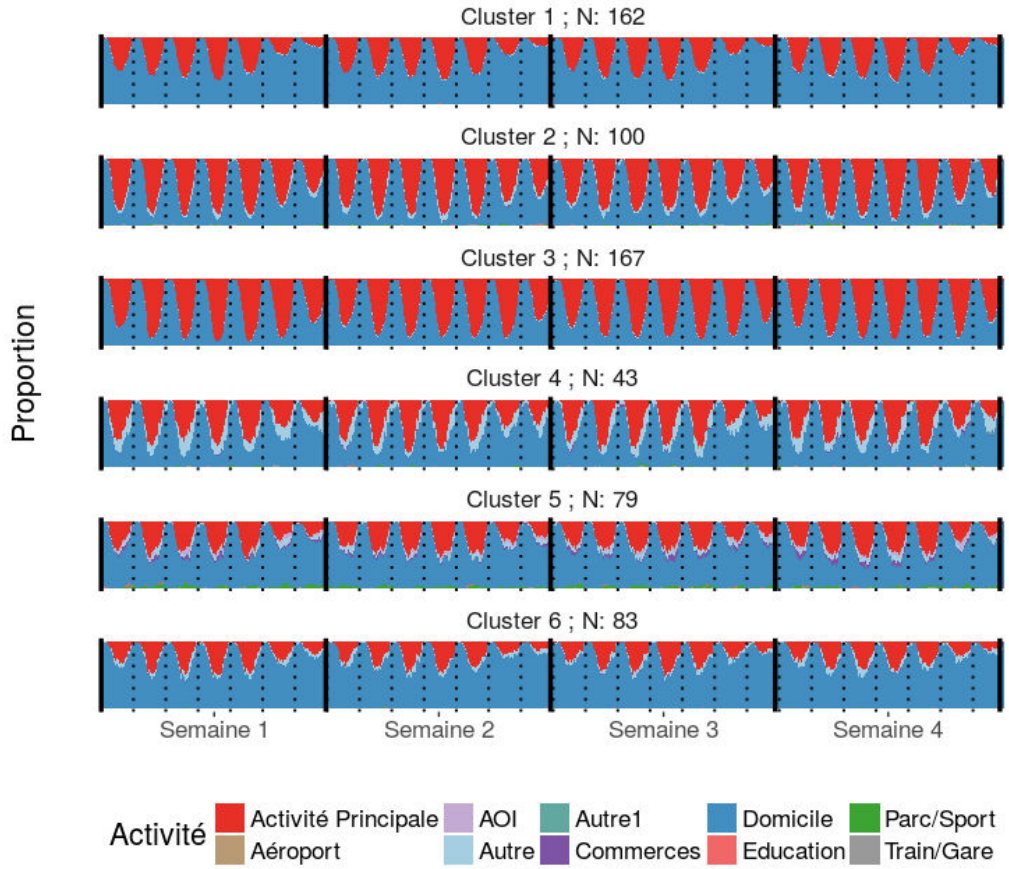


FIGURE 159 Agendas moyens sur un mois pour les six différents groupes définis précédemment.

Si cette méthode d'appariement ne nous apporte pas ici entière satisfaction, les agendas reconstitués restent au demeurant plutôt crédibles, avec bien entendu quelques améliorations à apporter (voir chapitre 11). Ces agendas nous permettent de passer de données épisodiques d'un point de vue temporel à des séquences d'activités continues, prenant ainsi quelques libertés sur les données d'origines, sans pour autant ajouter de lieux fréquentés à l'espace d'activité.

L'hypothèse sous-jacente est que les traces numériques d'un individu, et la fréquence d'envoi de messages dans différents lieux est assez proche de sa présence effective dans chacun de ces lieux, et donc de son espace d'activité réel d'autant plus si la période de collecte est suffisamment longue. À défaut d'être en mesure de vérifier cette hypothèse, nous réajustons les périodes de fréquentation des différents lieux, selon les informations que nous avons pu collecter (*e.g.* le volume horaire de messages géolocalisés dans un lieu donné). L'utilisation de tirage de valeurs conditionnées par des fréquences de distribution permet de reconstituer pour un même individu une très grande variété d'agendas plausibles, assez proches entre eux, bien que légèrement différents. Nous allons maintenant tenter d'apprécier dans quelles mesures ces agendas reconstitués peuvent servir de base à la création d'autres agendas respectant plus ou

moins leurs caractéristiques globales, tant sur les lieux et activités fréquentés que sur les heures de visite.

3 Générer des agendas synthétiques

Nous proposons maintenant une méthode qui permet de générer de nouveaux agendas à partir des agendas reconstitués précédemment, qu'ils proviennent de données collectées sur le terrain ou de données *Twitter*.

Nous optons encore pour une attribution selon une approche étape par étape des éléments nécessaires à la création d'agendas sur un mois. Nous attribuerons tout d'abord à chaque agent une liste d'activités et un nombre de lieux correspondants qu'il visitera. Nous définirons ensuite les semaines, les jours et les plages horaires de réalisation où chacun de ces lieux seront fréquentés. Ces attributs seront définis toujours selon une approche stochastique, en prenant en compte les fréquences de distributions de ces différents éléments que nous chercherons à générer. Pour des raisons de clarté de lecture, nous utiliserons l'acronyme « AR » pour désigner les agendas reconstitués obtenus en réajustant les données initiales provenant de *Twitter* (section 2) ou des enquêtes du terrain (chapitre 7), et l'acronyme « AG », pour les agendas générés à partir des AR en utilisant l'algorithme que nous allons présenter ici. Ce dernier sera illustré ici à partir d'AR d'après nos données *Twitter*, mais sera aussi appliqué aux agendas du terrain.

Au préalable, afin d'augmenter la taille de l'échantillon et d'avoir des tendances statistiques plus stables, nous combinons 5 simulations d'AR de nos données *Twitter*. Nous partons aussi du principe (contestable) que tous les agents effectuent une activité principale hors de leur domicile cette dernière pouvant être de manière indifférenciée « étudier », ou « travailler ». Nous ne ferons pas ici de distinction selon d'éventuels groupes et appliquerons ici à chaque agent un agenda dont les caractéristiques sont déduites de l'ensemble des AR.

3.1 Étape 1 : Données initiales de l'espace d'activité

3.1.1 Définir les activités qu'un agent va effectuer

Toutes les personnes n'effectuent pas le même nombre d'activités. Certains sont assez sédentaires et se cantonnent de manière routinière à des navettes domicile / travail, d'autres sont plus mobiles et sortent plus souvent, font du sport, vont dans les centres commerciaux ou voir des connaissances, etc. C'est également ce qui ressort de nos agendas de synthèses issus de données *Twitter* (figure 160.a) où finalement peu d'utilisateurs n'ont qu'une activité hors de leur domicile (Activité principale), et où l'on observe un mode à 2 activités, et une médiane à 3. Prenant en compte cette distribution, l'agent va tirer un nombre d'activité à effectuer.

La nature des activités est ensuite choisie en fonction de leur probabilité d'être effectuées

(figure 160.b), avec un tirage sans remises. Dans le cas présent, au regard de nos données *Twitter*, de notre couche d'utilisation du sol et de nos AR, si un agent n'a qu'une activité hors de son domicile et de son travail, il y aura de grandes chances qu'elle soit de type « Autre » (~65 %).

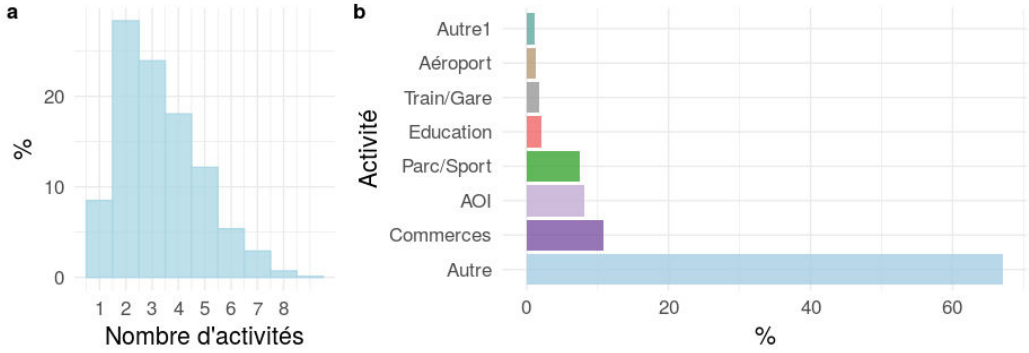


FIGURE 160 Part d'utilisateurs selon le nombre d'activités effectuées hors du domicile (a) et part des activités effectuées (hors domicile et travail) (b). D'après 5 simulations d'agendas reconstitués sur un échantillon de 634 utilisateurs de *Twitter* à Delhi.

3.1.2 Répartir les activités dans un nombre de lieux

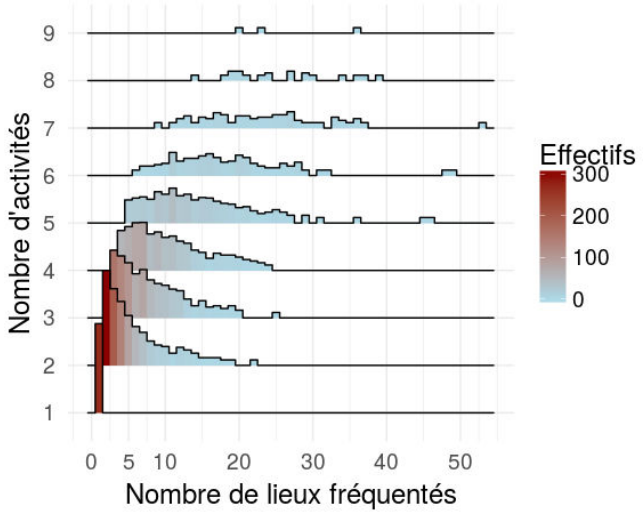


FIGURE 161 Nombre de lieux fréquentés en fonction du nombre d'activités effectuées (hors domicile et transport).

Le nombre de lieux différents fréquentés par une personne dépend de son nombre d'activités. Une personne ayant par exemple 3 activités fréquentera au moins 3 lieux, et au moins 8 si elle effectue 8 activités. Au-delà de cet aspect trivial, la figure 161 montre également que plus une personne effectue d'activités différentes, plus elle aura tendance à fréquenter un nombre de lieux proportionnellement plus importants, ce qui est assez typique de personnes

très mobiles dans la ville. Ainsi, plus une personne effectue d'activités, plus elle aura tendance à fréquenter un grand nombre de lieux, selon une relation non linéaire. Nous allons alors tirer un nombre de lieux qu'un agent pourra visiter en fonction du nombre d'activités déterminé précédemment, selon la distribution des effectifs de la figure 161.

Nous avons donc pour l'instant pour chaque agent une liste d'activité qu'il reste à répartir dans un nombre de lieux distincts. Chacune des activités sera effectuée dans au moins un lieu. Si le nombre de lieux fréquentés est supérieur au nombre d'activités, ce qui au regard de la figure 161 a de fortes chances d'arriver, nous effectuons alors un tirage avec remise, où la probabilité d'affecter une activité à un lieu dépend de la distribution des pourcentages des fréquentations de type d'utilisation du sol de la figure 160.b ci-dessus. Les activités effectuées le plus souvent comme celles de type « Autre », et dans une moindre mesure les « Commerces » et « AOI » ont donc plus de chance d'être visités dans un plus grand nombre de lieux.

À la fin de cette étape, chaque agent s'est vu attribuer un nombre d'activités qu'il réalise dans différents lieux en fonction des statistiques globales de notre échantillon [365]. Reste à définir leur moment de réalisation.

3.2 Étape 2 : Jour(s) et semaine(s) de visite

3.2.1 Sélection d'une semaine de réalisation

Nous posons dans un premier temps qu'un agent effectuera son activité principale toutes les semaines et sera présent tous les jours à son domicile.

D'après nos AR, toutes les activités ne sont pas réalisées à la même fréquence chaque semaine (figure 162). Certaines d'entre elles sont en effet effectuées systématiquement moins d'une fois par semaine (train, aéroport, autre1), et mis à part l'activité principale, seule une portion des lieux de type « Autre », « Commerces », « AOI » et « Parc/Sport » est susceptible d'être visité toutes les semaines, leur conférant un caractère plus routinier.

Nous allons donc tirer pour chaque activité a d'un lieu l une fréquence de visite hebdomadaire F selon la distribution de la figure 162.

- Si $F_{l,a}$ est supérieure à 1, le lieu l sera visité toutes les semaines.
- Si elle est inférieure à 1, nous appliquons exactement la méthode explicitée dans la section 2.2.1, à savoir que la probabilité que le lieu l soit visité la première semaine équivaut à la fréquence de visite tirée précédemment, mais cette dernière évolue la semaine suivante (augmente ou diminue) selon que le lieu ait été sélectionné ou non

365. Il est évidemment envisageable d'adapter ces statistiques en fonction de sous groupes, comme le lieu de résidence, ou le type d'activité principale, notamment s'il s'agit d'un étudiant ou non. Nous effectuerons cela dans le chapitre 11.

lors du tirage précédent.

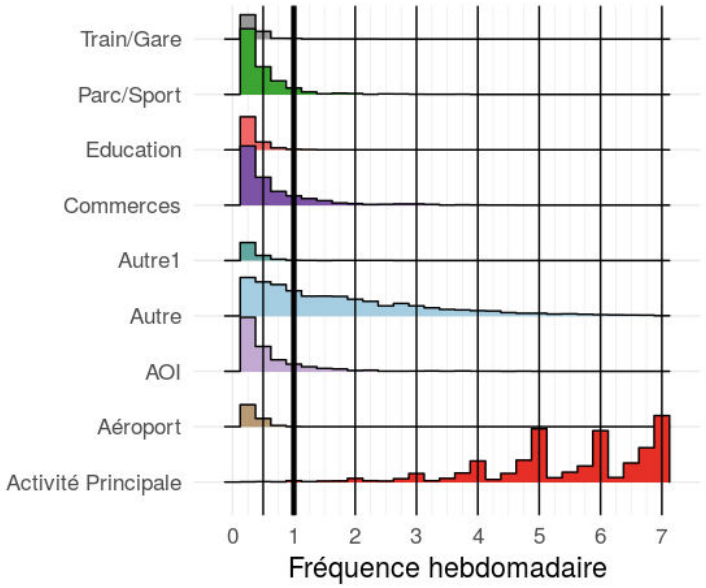


FIGURE 162 Fréquence hebdomadaire de visite dans un lieu associé à une activité. Après 5 simulations de reconstitutions d'agendas à partir des données *Twitter*

À la fin de cette étape, chaque lieu s'est vu attribuer (ou non) une ou plusieurs semaines de réalisation selon la fréquence hebdomadaire de visite $F_{l,a}$ de l'activité qui lui est associée.

3.3 Étape 3 : jour(s) de visite

3.3.1 Nombre de jours de visite hebdomadaire

Nous posons que la fréquence de visite hebdomadaire $F_{l,a}$ tirée précédemment correspond au nombre de jours où le lieu sera visité une semaine donnée. Nous arrondissons cette valeur de manière aléatoire à l'entier supérieur ou inférieur (sauf si elle est inférieure à 1), comme vu dans la section 3.1.2 du chapitre 7. Chaque lieu se voit donc attribuer un nombre de jours de visite entre 1 et 7. Reste à définir quels jours.

3.3.2 Sélections des jours de la semaine

Pour ce faire, nous nous basons sur les distributions de la figure 163 qui montre le pourcentage de visite pour chaque activité selon les jours de la semaine. Nous nommons $P_{j,a}$ la probabilité d'une activité a d'être réalisée un jour j.

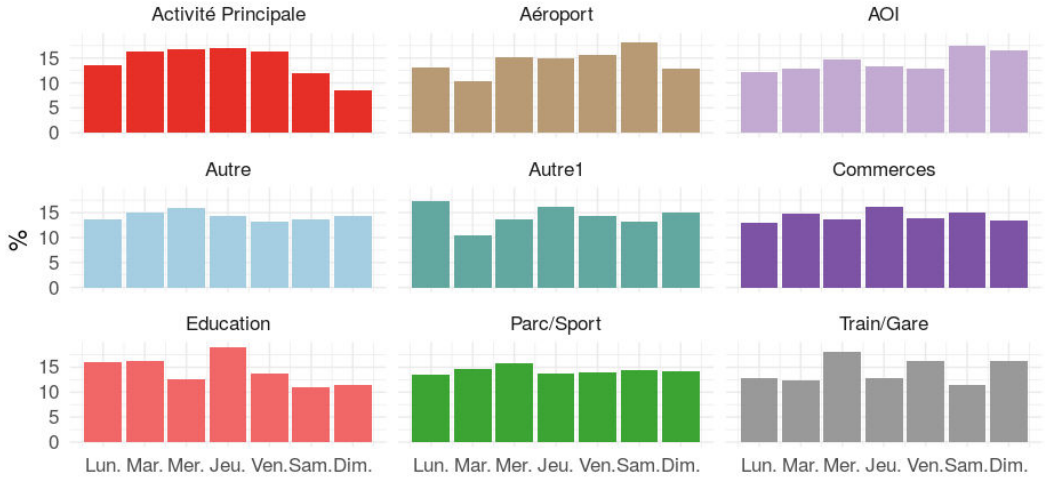


FIGURE 163 Pourcentage de fréquentation des différents lieux en fonction du jour de la semaine.

Nous pouvons noter que les activités n'ont pas toutes la même probabilité d'être effectuées un jour donnée. Par exemple les lieux d'éducations et les activités principales sont moins fréquentés le week-end, alors que nous pouvons observer l'inverse pour les AOI. Certaines variations sont cependant peu intuitives, comme le fait que les AOI soient moins visitées le vendredi, ou encore les légers pics de fréquentation dans les commerces les mardi et jeudi. Nous pouvons expliquer cela par les données initiales issues d'un échantillon relativement faible (634 individus), une couche d'utilisation du sol incomplète mais aussi par d'éventuels biais de notre algorithme. Ces données sont utilisées ici à caractère illustratif et la flexibilité de notre approche permet d'utiliser d'autres sources de données de fréquentation de lieu, comme des données issues de *check-in Facebook* par exemple.

Nous tirons ensuite pour chaque lieu l une liste de jours j de longueur $F_{l,a,}$ selon une probabilité $P_{j,a}$. Dit autrement, si un lieu associé à une activité de type « AOI » est visité une seule fois dans la semaine, il aura une plus grande probabilité d'être fréquenté un jour de week-end. De même, s'il s'agit de l'activité principale effectuée 5 jours par semaine, elle aura un peu plus de chance de se dérouler du lundi au vendredi.

Il ne reste plus qu'à définir les heures de la journée où ces lieux seront fréquentés.

Étape 4 : Organisation d'une journée

Plusieurs méthodes peuvent être envisagées pour choisir les heures et les durées où un agent visitera un lieu (voir chapitre 11). Nous proposons ici une approche inspirée des travaux de (Perkins *et al.*, 2014 ; Wu *et al.*, 2014), en mobilisant des matrices de transitions entre activités.

4.1 : Séquences des activités

Comme vu dans le chapitre précédent, les agendas reconstitués (AR), qui le soient à partir de données *Twitter* où de terrain, nous permettent de calculer la probabilité de passer d'une activité à une autre et d'obtenir ainsi une matrice de transition (figure 164).

La diagonale de la matrice présente des valeurs élevées ce qui traduit la durée d'une activité : par exemple le fait d'être à son domicile à une heure donnée et de s'y trouver encore à l'heure suivante est une observation assez récurrente, tout comme rester à son lieu de travail plusieurs heures consécutives.

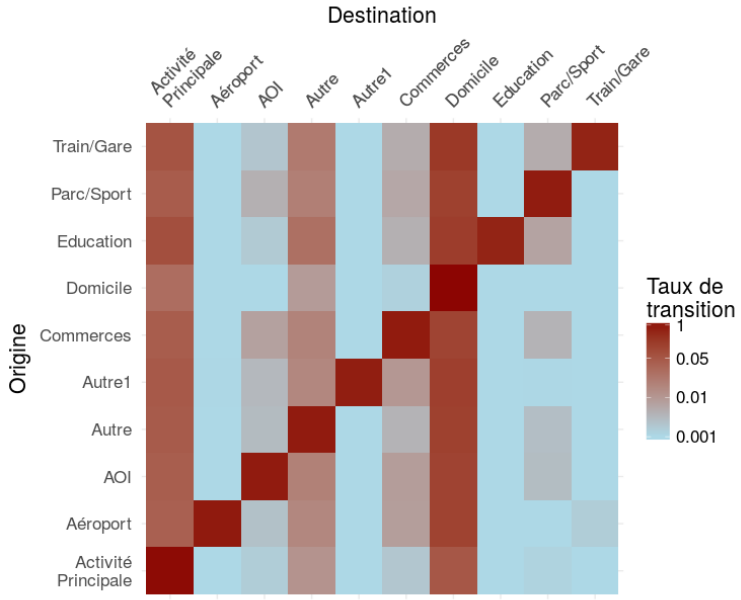


FIGURE 164 Représentation des taux de transitions entre activités sous forme de matrice. Le gradient de couleur suit une fonction logarithmique.

Cette matrice va nous permettre de définir des séquences de lieux fréquentés en une journée. Nous n'aurons alors plus qu'à leur affecter une durée (section suivante). Les séquences des activités pouvant être différentes selon les jours, nous utiliserons deux matrices de transition distinctes, calculées selon les jours de semaines ou de week-end, que nous appliquerons aux jours correspondants.

Dans un premier temps nous posons A la liste des activités $a_{l,j}$ réalisées un jour j dans un lieu l. Nous ne prenons en compte que les lignes et les colonnes de la matrice comprises dans A. Les activités a_j effectuées qu'une seule fois verront leur diagonale à 0, empêchant de fait de fréquenter plusieurs fois cette même catégorie de lieux.

Nous posons ensuite que l'agent part de son domicile et choisira $a_{l,j}$, avec une probabilité dépendant du taux de transition entre le domicile et a. De manière générale, il aura de plus grandes chances d'effectuer ensuite son activité principale (la valeur la plus élevée sur la ligne "Domicile"), mais les probabilités de choisir d'autres activités sont non nulles. Dès qu'un lieu l est sélectionné, nous retirons $a_{l,j}$ de A, et redéfinissons la matrice comme précédemment. Appliquée de manière itérative, cette approche permet d'obtenir une séquence de lieux fréquentés un jour donné, de type $\{Domicile, a_{1j}, a_{2j}, ..., a_{nj}\}$.

À noter que par conception, une personne ne peut rentrer chez elles entre deux activités ni avoir de coupure lors de son activité principale. Ces aspects devront être améliorés lors de prochains travaux.

4.2 Durée des activités

Semaines, jours et séquences des lieux fréquentés étant définis, nous allons maintenant attribuer une durée à chacune des activités $a_{l,j}$. Cependant les journées ne sont pas extensibles et plus un individu effectue d'activités lors d'une journée, plus la durée de ces dernières devrait être courte afin de pouvoir toutes les enchaîner.

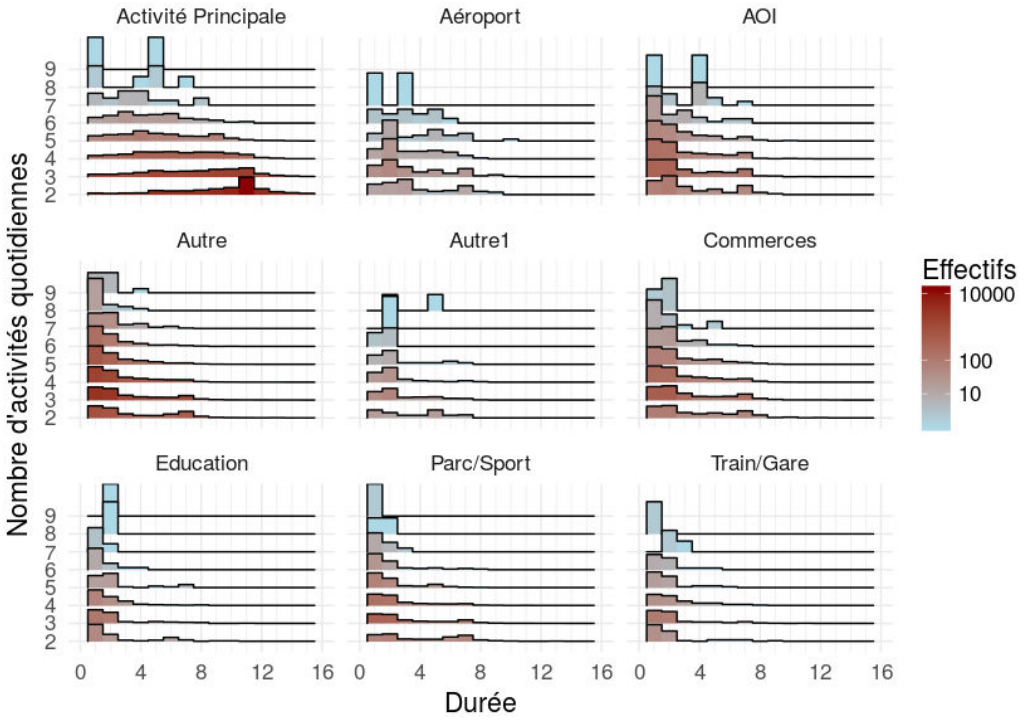


FIGURE 165 Distribution des durées des activités selon le nombre d'activités quotidiennes, d'après les AR de *Twitter*.

La figure 165, présente la distribution des durées des activités selon le nombre d'activités

quotidiennes obtenues d'après les AR de *Twitter*. On y observe une baisse des fréquences des activités de longue durée avec l'augmentation du nombre d'activités effectuées. À noter que cette baisse est plus visible pour les activités principales, du fait de l'algorithme de reconstitution des agendas et du processus de gestion des doublons.

Nous tirons pour chaque $a_{l,j}$ une durée en fonction de l'activité a et du nombre d'activités réalisées en j, selon les distributions des durées de la figure 165. Une journée commence donc par une présence au domicile puis se compose d'une succession de lieux fréquentés pendant une certaine durée. Reste à définir l'heure du départ du domicile.

3.3.3 Plages horaires

Nous tirons donc une heure de départ du domicile en nous basant sur la courbe de distribution du premier départ du lieu de résidence, hérité des AR (figure 166). Cette valeur dépend du jour, s'il s'agit d'un jour de semaine, d'un samedi ou d'un dimanche. La figure 166 montre néanmoins que les différences sont assez faibles entre les jours et que la plupart des personnes sortent de chez eux entre 5 h et 10 h, ce qui à défaut d'être conforté par d'autres types de données que nous avons en notre possession reste très cohérent. Nous ajoutons ensuite les différentes activités effectuées, avec leur durée correspondante ce qui nous permet d'obtenir une journée de l'agenda d'un agent.

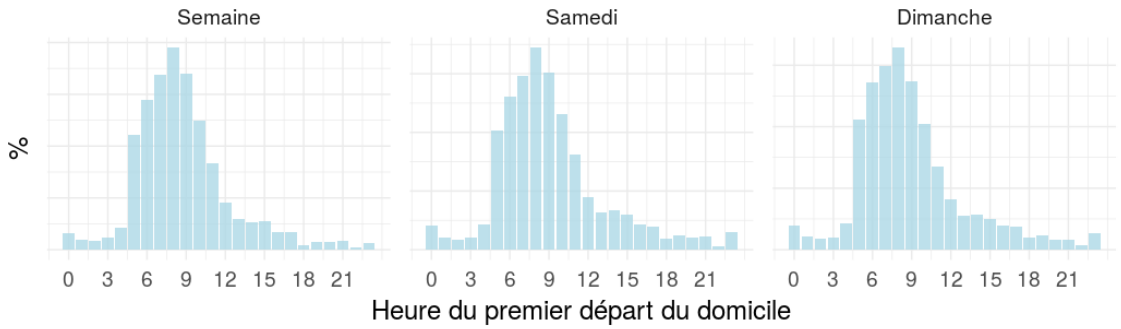


FIGURE 166 Probabilité horaire du premier départ du domicile.

À noter que si un agent effectue beaucoup d'activités un jour donné, il est possible qu'au gré des tirages aléatoires, ce dernier passe un temps excessif hors de chez lui. Compte tenu de l'approche séquentielle que nous avons adoptée ici, une durée hors du domicile trop longue, associée à une heure de départ du domicile tardive impliquerait un nombre non négligeable de lieux fréquentés la nuit, ce que nous souhaitons limiter. Nous allons donc effectuer ici le seul ajustement de cet algorithme, à savoir optimiser les durées des activités qui ne se déroulent pas au lieu de résidence.

3.3.4 harmonisation

Toujours en partant des AR, nous dressons le profil du temps passé hors du domicile en fonction du jour de la semaine (figure 167). Globalement, les courbes forment une gaussienne centrée sur 11 h, avec une croissance assez lente et une chute rapide. Ainsi, très peu de gens passent plus de 15 h hors de leur domicile. À noter que les jours de la semaine ont des profils similaires, avec des durées concentrées entre 6 h et 14 h, tandis que ceux du week-end sont caractérisés par une pente plus douce, tant avant 11 h qu'après, ce qui suggère que l'échantillon est partagé entre des personnes qui restent plutôt chez eux, et des personnes qui auront tendance à sortir plus longtemps le samedi et le dimanche.

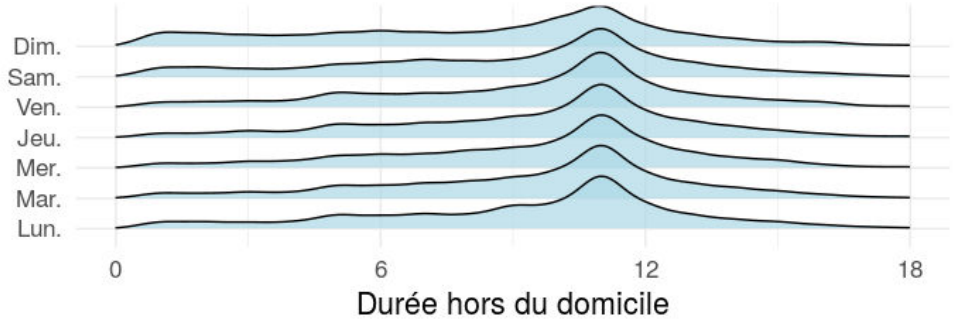


FIGURE 167 Répartition (sous forme de densité) des durées hors du domicile (en heure) en fonction des jours de la semaine. L'axe des y varie entre 0 et la valeur maximale.

À partir des courbes de distribution de la figure 167, nous tirons une durée référence d_{ref} que nous rallongeons de 25 % et nous la comparons avec la durée hors du domicile d_{out} résultant de notre algorithme.

- Si d_{out} est inférieure à d_{ref} nous conservons d_{out}.
- Si d_{out} est supérieure à d_{ref}, nous retirons une durée d_{ref}, toujours selon une probabilité issue des courbes de distribution, mais dans un intervalle plus restreint, compris entre un tiers de d_{out} et d_{out}.

Cette dernière opération permet d'augmenter les chances d'avoir une durée plus courte, sans que cette dernière ne le soit trop, afin de conserver autant que faire se peut un temps hors du domicile relativement long, sans pour autant être excessif. Nous appliquons ensuite un produit en croix sur les durées des activités concernées afin d'en réduire la durée.

Toutes ces étapes nous permettent finalement d'obtenir pour chaque agent un agenda individuel, dont nous analyserons les propriétés temporelles dans la section suivante.

4 Discussion des résultats

Nous décrirons dans cette section les différents résultats obtenus notamment en comparant les heures de déroulement des activités issues des AG et celles provenant des AR à partir des données *Twitter* et de l'enquête de terrain.

4.1 À partir des AR Twitter

Nous avons généré ici à titre d'illustration 5 000 AG à partir de l'algorithme précédent [366], en nous basant sur 5 simulations d'agendas reconstitués (AR) (634*5 individus). La figure 168 montre la distribution du nombre d'activité en fonction du nombre de lieux fréquentés pour l'ensemble de nos agents et est sans surprise très similaire à celle issue des agendas reconstitués à partir des données *Twitter* (figure 161, voir plus haut).

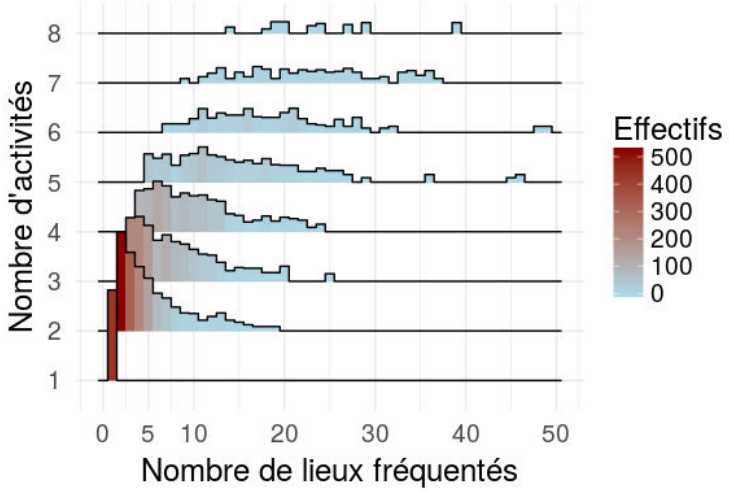


FIGURE 168 Nombre de lieux fréquentés en fonction du nombre d'activités, d'après nos AG.

La figure 169 montre les fréquentations horaires sur une semaine agrégées par type d'utilisation du sol. Les histogrammes de couleurs correspondent aux agendas générés (AG) tandis que les lignes noires représentent les valeurs des agendas reconstitués (AR) qui ont servi de base aux AG et servent ici de points de comparaison. Les heures de présence au domicile correspondent quasiment parfaitement, et nous observons une bonne concordance pour les activités principales, même si les données générées ont produit moins d'agents travaillant le dimanche. Si les tendances globales en termes d'amplitude sont relativement bien respectées pour toutes les activités, il existe toutefois des décalages dans les horaires de réalisation.

366. Loin d'être optimisé et finalisé, entre 30 et 50 secondes sont nécessaires pour créer 100 agendas pour autant d'agents, sur une période de 672 heures (4 mois), à partir d'un ordinateur portable classique.

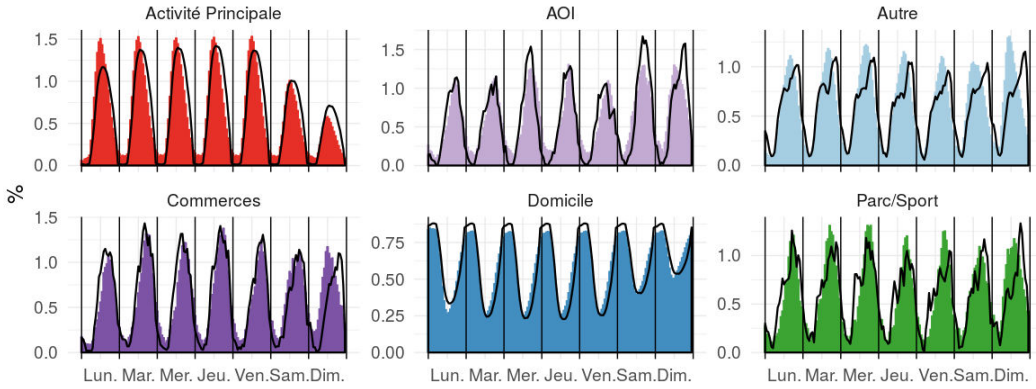


FIGURE 169 Comparaison des parts de fréquentations horaires par type de lieu sur une semaine, selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogramme).

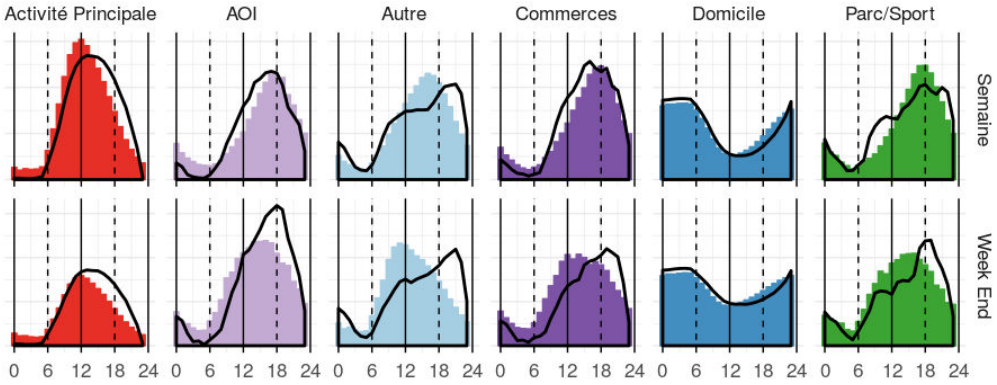


FIGURE 170 Comparaison des parts de fréquentations horaires par type de lieu les jours de semaines (haut) et le week-end (bas), selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogramme).

La figure 170 montre les fréquences moyennes de visites horaires pour les jours de semaine et de week-end, et permet de pointer plus précisément les écarts entre les deux types d'agendas. Même à ce niveau de précision temporelle, les tendances de fréquentation des lieux de résidence sont très similaires. Nous pouvons noter que les distributions des fréquences de visite des activités principales forment dans les deux cas une courbe en cloche, mais plus étroite pour les AG et avec un mode un peu plus tôt (12 h contre 13 h). Pour les AG, les lieux de type « Autre » voient leur fréquentation augmenter de manière linéaire entre 6 h et 18 h les jours de semaines, et entre 6 h et midi les week-ends. Ils ne présentent pas de pic entre 19 h et 21 h, comme c'est le cas pour les AR. Les profils des AG des « AOI », des « commerces » et des « Parc/Sport » sont très similaires, tant les jours de week-end que de semaines, avec un mode à 18 h que l'on observe aussi pour les AR. S'agissant des *AOI* et des *Commerces*, les

AG tendent à réduire la présence en journées la semaine et à l'augmenter le week-end, ce qui n'a rien d'extravagant, même si les commerces sont fréquentés un peu plus tard le week-end d'après les AR. Nous n'observons pas de pic pour les activités "Parc/Sport" le matin entre 6 h et 12 h, contrairement à ce qui ressort des agendas reconstitués à partir des données *Twitter*.

Cet algorithme réajuste également les durées des activités qui se déroulent hors du domicile lorsque ces dernières sont trop longues. La figure 171 compare les distributions des durées des activités issues des AR et des AG et sans surprises, ces dernières sont en générales plus courtes, surtout pour l'activité principale dont les distributions des durées sont plus lissées et ne présentent pas de pic à 11 h pour les AG. Il conviendra d'adapter cet algorithme pour réduire les écarts entre les durées observées et simulées, ce qui pourrait passer par l'application de contraintes supplémentaires lors de l'attribution des durées aux activités de manière à limiter l'enchaînement de visites trop longues sur une même journée. Nous pourrions également envisager de n'appliquer le réajustement que sur les activités qui ne sont pas l'activité principale, afin de préserver le caractère plus figé temporellement de ces dernières.

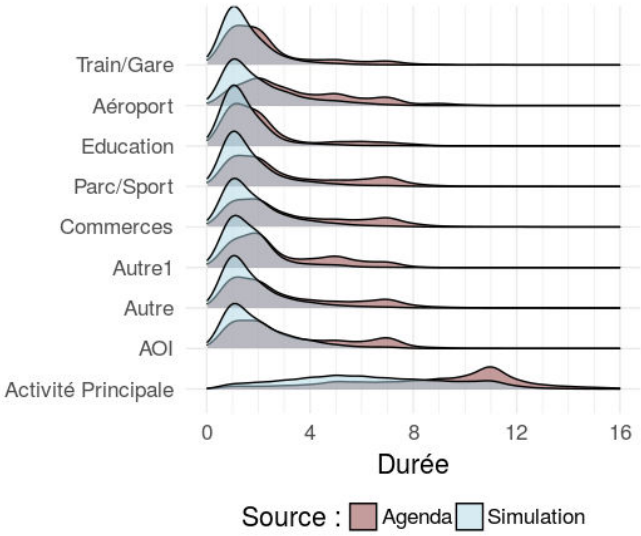


FIGURE 171 Comparaison des distributions des durées entre les données issues des agendas reconstitués (rouge) et des agendas générés à partir de ces données (bleu ciel).

Néanmoins, les profils de fréquentation temporelle des différents types de lieux des AG sont tout de même assez proches des AR, sachant que l'attribution des plages horaires s'est faite simplement en tirant une heure de départ du domicile, en appliquant une matrice de transition pour définir des séquences d'activités et en estimant une durée pour chaque activité. Les profils sont différents entre la semaine et le week-end, et les pics de 18 h apparaissent bien pour les lieux de type « Commerces », « AOI » et « Parc/Sport ». Les activités de type « Autre », ont en revanche des profils très différents des données sources, le pic de 20 h n'apparaissant pas.

Il conviendrait par la suite d'arriver à mieux discriminer ce type d'activité, en mobilisant une couche d'utilisation du sol plus complète. Il pourrait aussi être envisageable de subdiviser les lieux de type « Autre », en d'autres catégories selon leurs horaires de réalisation, par exemple ceux qui sont visités le matin, le soir ou durant toute une journée.

La figure 172 ci-dessous présente 10 agendas générés à partir des AR. Nous pouvons noter une grande variété de profils, certains étant nettement plus réguliers que d'autres. Les semaines des agents sont relativement variées, au gré du nombre d'activités différentes qu'ils effectuent.

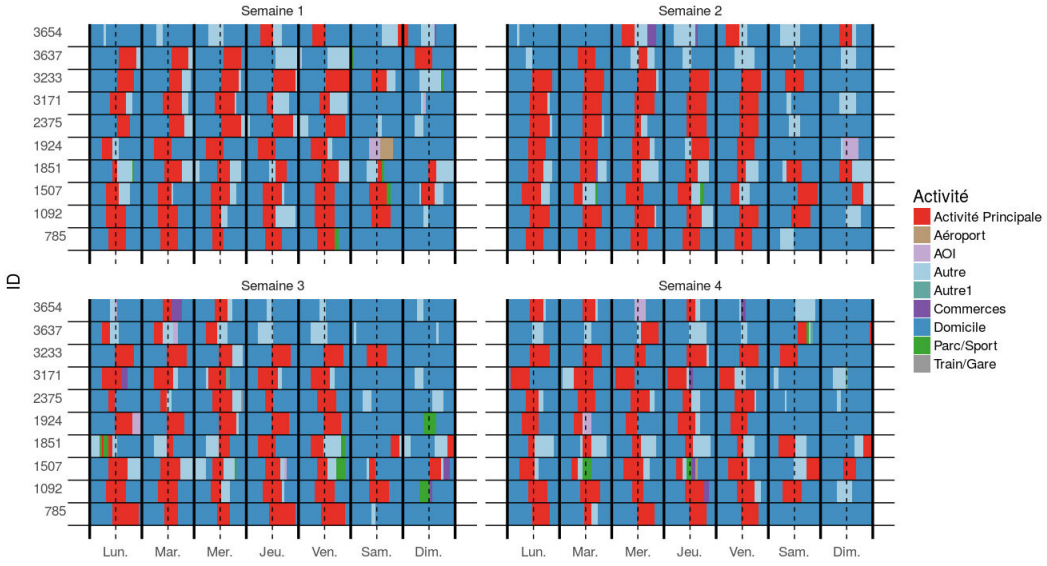


FIGURE 172 Exemple d'agendas pour 10 agents.

Néanmoins, ces résultats sont obtenus à partir d'agendas qui ont eux-mêmes leurs propres limites, du fait de la difficulté de poser des hypothèses correctes pour passer de données épisodiques et sporadiques à continues. La section suivante présente l'application du même algorithme aux agendas reconstitués à partir des données de terrain et permettra d'apporter de nouveaux éléments de comparaison et de réflexion.

4.2 À partir des données du terrain

Nous allons ici nous baser sur 10 simulations d'agendas reconstitués à partir de nos données terrain et de notre échantillon de 99 personnes. Contrairement aux AR à partir des données *Twitter*, les personnes interrogées effectuent plus d'activités différentes, environ 65 % d'entre eux ayant entre 3 et 6 activités (figure 173.a). Ceci peut s'expliquer simplement par le fait que la nature des lieux fréquentés est connue, contrairement aux utilisateurs de *Twitter*, où plus de 60 % des lieux entraient dans la catégorie « Autre », qui peut agréger des activités

très variées. La répartition des activités fréquentées est également plus équilibrée, les marchés comptant pour 22 % des lieux visités, et les autres activités entre 5 et 16 % (figure 173.b).

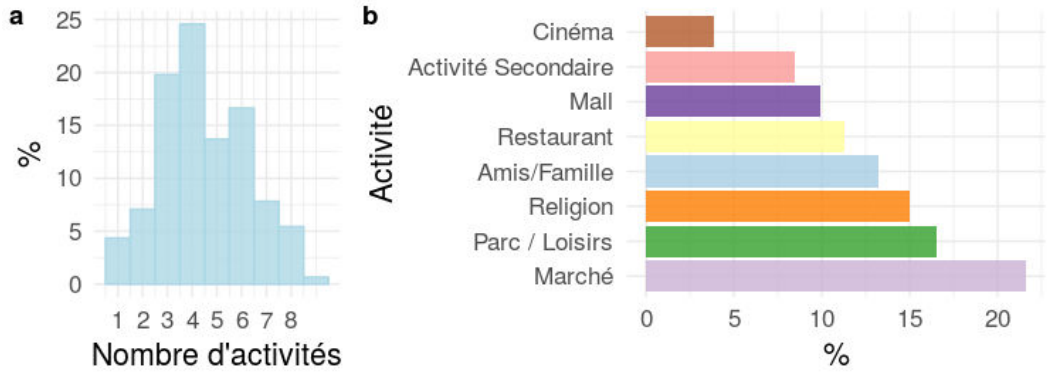


FIGURE 173 Nombre d'activités (autre que le domicile) par utilisateur (a) et pourcentages des activités effectuées par l'ensemble des utilisateurs (certaines activités sont effectuées plusieurs fois, dans différents lieux). À partir des agendas reconstitués issus des données récoltées sur le terrain.

Nous pouvons faire également la même observation que précédemment concernant le lien entre le nombre d'activités effectuées par un individu et le nombre de lieux qu'il fréquente, à savoir que plus un individu effectue d'activités, plus il fréquente de lieux différents (figure 174).

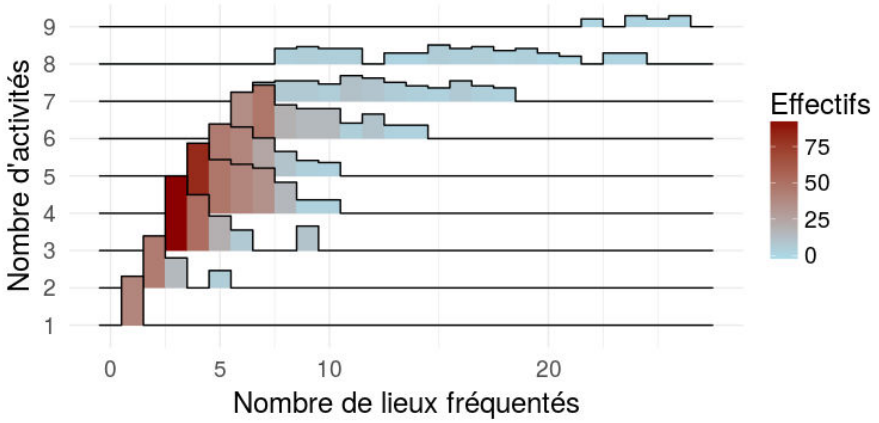


FIGURE 174 Nombre de lieux fréquentés en fonction du nombre d'activités effectuées (hors domicile), d'après les données terrain.

L'heure du premier départ et les jours où s'effectuent les différentes activités ont été fortement contraints lors de la reconstitution de ces agendas. Ceci peut expliquer les pics dans les heures de départ entre 7 h et 9 h, visiblent sur la figure 175, et la dichotomie semaine / week-end pour la fréquentation des différents types de lieux (figure 176). Ces paramètres jouant un rôle important dans notre algorithme, il conviendra d'obtenir des réponses plus précises de la part des interviewés lors d'enquêtes de terrains ultérieures.

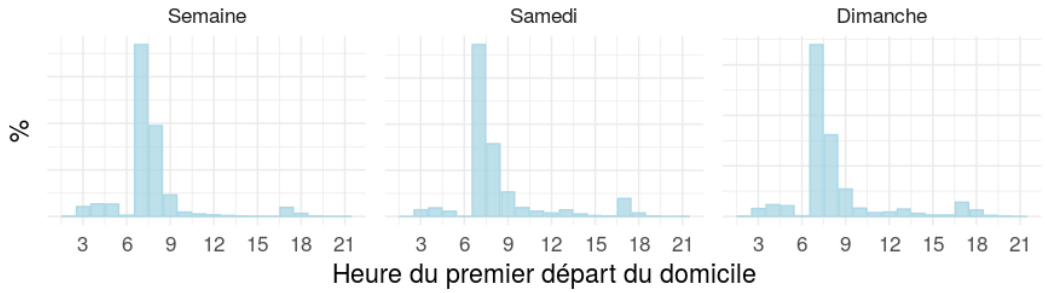


FIGURE 175 Probabilité de l'heure du premier départ du domicile. À partir des agendas reconstitués issus des données récoltées sur le terrain.

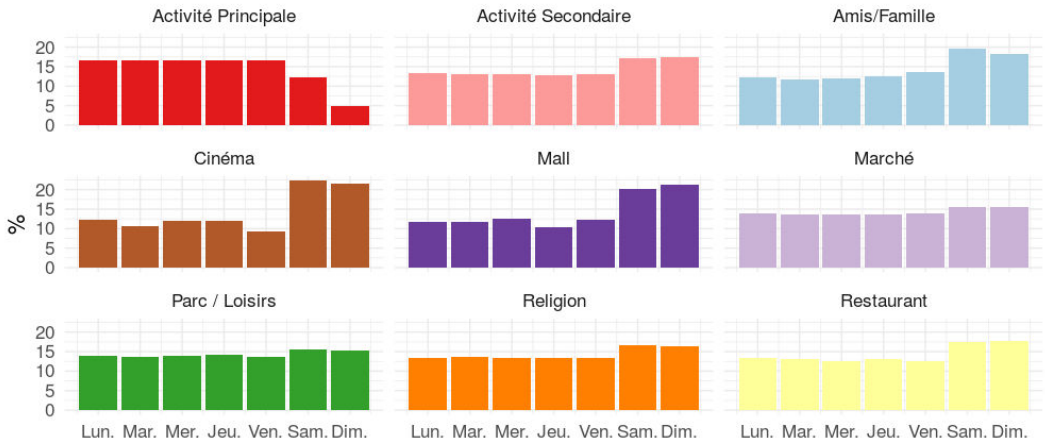


FIGURE 176 Probabilité d'effectuer une activité un jour donné. À partir des agendas reconstitués issus des données récoltées sur le terrain.

Un autre paramètre en entrée du modèle est la fréquence de visite hebdomadaire d'un type de lieu (figure 177). Elle permet d'estimer si une activité sera effectuée une semaine donnée, et le nombre de jours où cette dernière sera réalisée. Les fréquences déduites des données *Twitter* ne faisaient pas ressortir d'activités effectuées très fréquemment (plus de 3 fois par semaine) qui ne soient pas l'activité principale. Nous pouvons noter ici que des marchés, des lieux de cultes, des parcs ou encore des connaissances sont visités très régulièrement par une part non négligeable de l'échantillon. Ceci suggère qu'il faudrait peut-être d'augmenter artificiellement les fréquences de visites de certains lieux pour les données *Twitter* qui ne sont ni l'activité principale, ni le domicile, en utilisant des coefficients d'ajustement, afin de reproduire des comportements plus routiniers, en plus des navettes domicile / travail. Si cette approche permettrait d'avoir des distributions des fréquences de visites globales qui se rapprochent plus de la réalité, le risque d'affecter une fréquence de visite plus élevée à un lieu en réalité fréquenté de manière très occasionnelle est très élevé, ce qui ajouterait encore des biais.

De la même manière que précédemment nous produisons une matrice de transition

différenciée entre les jours de semaines et de week-end afin de déterminer les séquences quotidiennes des activités. La durée de ces dernières est également conditionnée en fonction du nombre d'activités qu'un agent effectuera dans une journée.

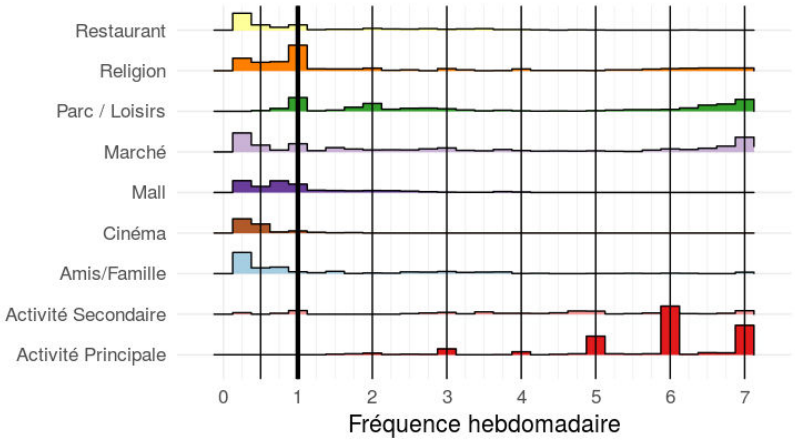


FIGURE 177 Distribution des fréquences de réalisation hebdomadaires selon le type d'activité effectuée.

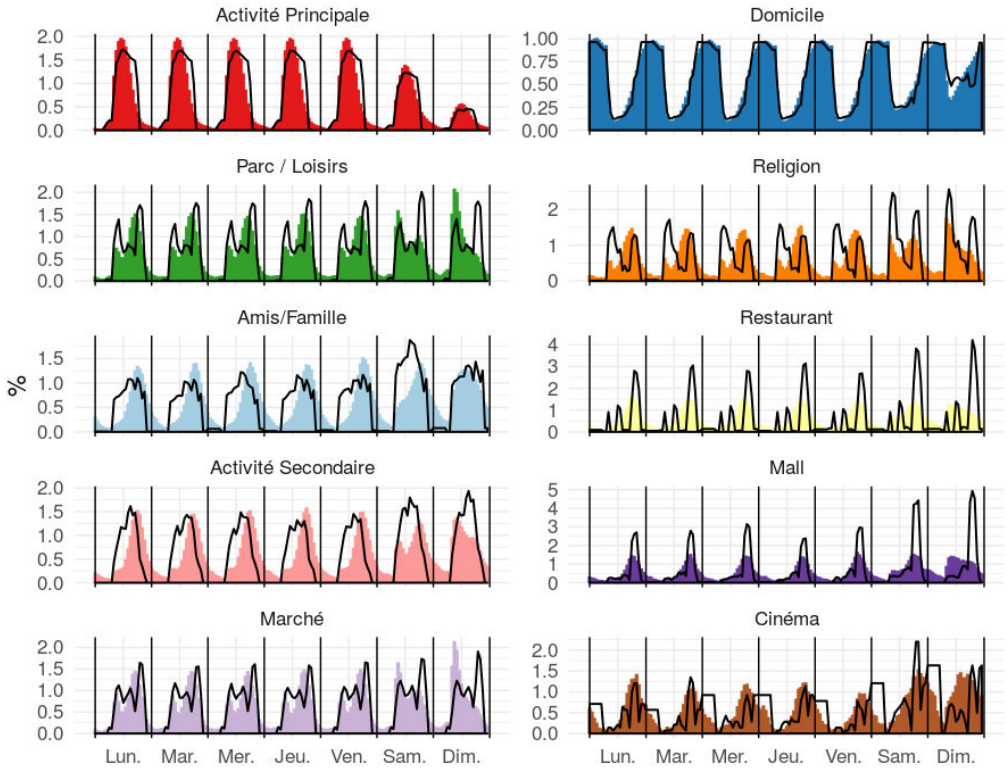


FIGURE 178 Comparaison de la part de fréquentations par activités par tranche horaire sur une semaine selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogrammes)

À partir de ces informations extraites des AR nous avons généré 5000 agendas en utilisant exactement le même algorithme que précédemment. Nous pouvons noter dans un premier temps que les niveaux d'accord entre les fréquences horaires des AR et AG varient grandement selon le type de lieu (figure 178). Comme précédemment, les présences au lieu de domicile sont quasiment identiques entre les deux jeux de données. Les courbes de fréquentation horaire de l'activité principale sont assez similaires, bien que la largeur de la plage soit plus étroite pour les AG. En revanche, pour les autres activités, les dissimilarités sont assez importantes, bien que les ordres de grandeur des amplitudes maximales semblent assez bien respectées pour les activités secondaires, marchés, parcs, ou encore les lieux de cultes. La lecture des figures 179 et 180 permet d'apprécier les niveaux de fréquentation selon l'heure de visite les jours de semaines ou de week-end.

Les jours de semaines

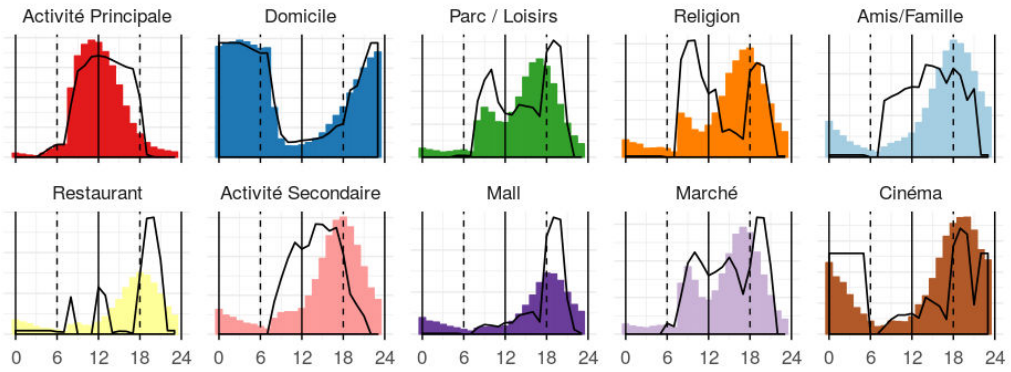


FIGURE 179 Comparaison de la part de fréquentations par activités par tranche horaire sur un jour de semaine type, selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogrammes).

Bien que les jours de semaine le mode de l'activité principale soit le même entre les AR et les AG (12 h), la largeur de la plage horaire est plus étroite pour ces derniers. Alors que le fait de visiter des proches ou d'effectuer son activité secondaire se déroule plutôt en journée d'après les AR, notre algorithme a généré un pic vers 18 h en semaine. Même si l'algorithme ne reconstitue pas correctement les heures de déroulement de ces activités issues de nos agendas, nous pouvons tout de même noter que ces pics sont toutefois cohérents. En effet, une grande partie des personnes qui effectuent des activités secondaires sont des jeunes qui vont dans des « tuition », sortes de cours du soir mais qui se déroulent juste après l'école. Et d'autre part, il n'y a rien d'étonnant à rendre visite à des amis après le travail plutôt qu'en journée. Les heures de fréquentation des cinémas sont assez chaotiques dans nos AR, mais les ordres de grandeur sont toutefois assez bien respectés chez les AG (pic vers 19 h). Le pic la nuit dans les AR est probablement dû au fait que certaines personnes interviewées ont répondu « nightime »

pour dire une plage horaire de fréquentation des cinémas, que nous avons mal retranscrite. Les horaires de visite au restaurant sont par constructions très contraintes pour les AR, où seules trois plages horaires existent : le matin, le midi et le soir. Notre algorithme de génération d'AG n'a pas su prendre en compte de telles contraintes. Conformément aux AR, nos AG présentent un pic le matin et un autre l'après-midi pour les activités de type « marché », « Parcs » et « lieux de cultes ». Néanmoins, si l'heure du pic du matin correspond bien, les amplitudes divergent grandement pour les lieux de cultes, et l'heure du pic de l'après-midi est avancée d'une à deux heures chez les AG. L'ajout d'une catégorie de type « transition » entre deux activités qui correspondrait au temps de transport permettrait de décaler ces pics dans la soirée. Si les pics de fréquentation des *malls* correspondent bien, les niveaux d'amplitudes sont très différents, ou proportionnellement beaucoup plus de personnes fréquentent les *malls* entre 18 h et 20 h d'après les AR.

Les week-ends

Concernant les horaires de présences aux domiciles et dans les activités principales, nous pouvons faire le même constat que précédemment, à savoir une bonne cohérence entre les AR et les AG, même si l'activité principale se déroule sur des tranches horaires plus entendues. Les lieux de type « Parc / Loisirs », « Religion », ou encore « Marché » présentent un pic le matin et le soir pour les AR. Seul le premier pic est reproduit chez les AG. Rappelons que notre algorithme ne prend pas pour l'instant en compte les aller / retour entre le domicile et les autres activités, ce qui peut expliquer l'absence de ce second pic pour ces activités, dont les interactions sont probablement très liées au lieu de domicile les jours de week-end.

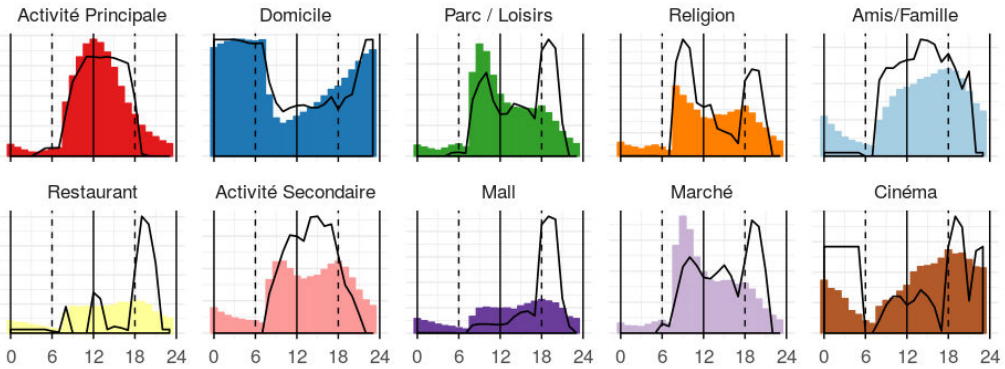


FIGURE 180 Comparaison de la part de fréquentation par activités par tranche horaire sur un jour de week-end type, selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogramme).

Malgré des courbes de fréquentation assez différentes pour les visites de proches ou la réalisation de l'activité secondaire, l'écart n'est pas si important, les AR comme les AG attribuent

bien le plus grand nombre de personnes à des heures de journées. Les *malls* sont fréquentés de manière assez équivalente toutes les heures de journée du week-end pour nos AG, tandis que les AR présentent un pic très marqué entre 18 h et 20 h. Là encore, si la différence est assez importante entre les deux types d'agendas, le fait d'estimer des présences dans les *malls* entre 10 h et 22 h n'a rien d'incohérent.

L'algorithme qui permet de reconstituer les agendas d'après les données récoltées sur le terrain contraint fortement certaines activités à s'appliquer à certaines plages horaires. Concernant les parcs, lieux de cultes ou les marchés, un grand nombre d'interviewés ont répondu les fréquenter le matin ou le soir, ce qui ressort bien sur les figures ci-dessous. Alors que nous ne prenons pas en compte les plages horaires, mais seulement les séquences de successions d'activités, ces pics ressortent assez bien pour les jours de semaines chez les AG, moins bien les week-ends où le pic de l'après-midi est nettement moins marqué. Nous pourrions envisager de déduire des AR un nombre de fois qu'un individu repasse par chez lui par jour et intégrer cet élément dans la définition des séquences des activités.

Si nous comparons les durées des activités entre les AG et les AR (figure 181), nous pouvons faire ici la même remarque que pour l'application aux données issues de *Twitter*, à savoir que notre algorithme tend à réduire les durées des activités, du fait de notre méthode d'ajustement qui contraint le temps passé hors du domicile.

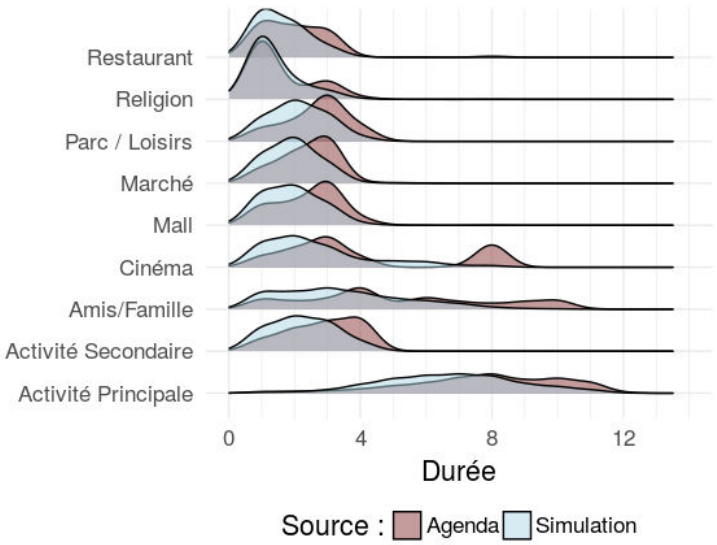


FIGURE 181 Comparaison des distributions des durées dans un lieu visité entre les données issues des agendas reconstitués (rouge) et des agendas générés à partir de ces données (bleu ciel)

Si nous regardons maintenant un échantillon de nos agendas générés (figure 182), nous pouvons noter que notre algorithme a été à même de créer des agendas globalement très

différents selon les agents, tout en faisant ressortir des variabilités d'un jour et d'une semaine sur l'autre dans les agendas individuels. Bien évidemment, tout ceci est conditionné par les paramètres initiaux, et une personne qui fréquente peu de lieux et effectue moins d'activités aura un agenda plus stable et moins varié. La distribution des heures de départ du domicile, principalement centrée autour de 7 h et 9 h (figure 175) entraîne que peu d'activités principales sont effectuées la nuit comparée aux AG de *Twitter*.

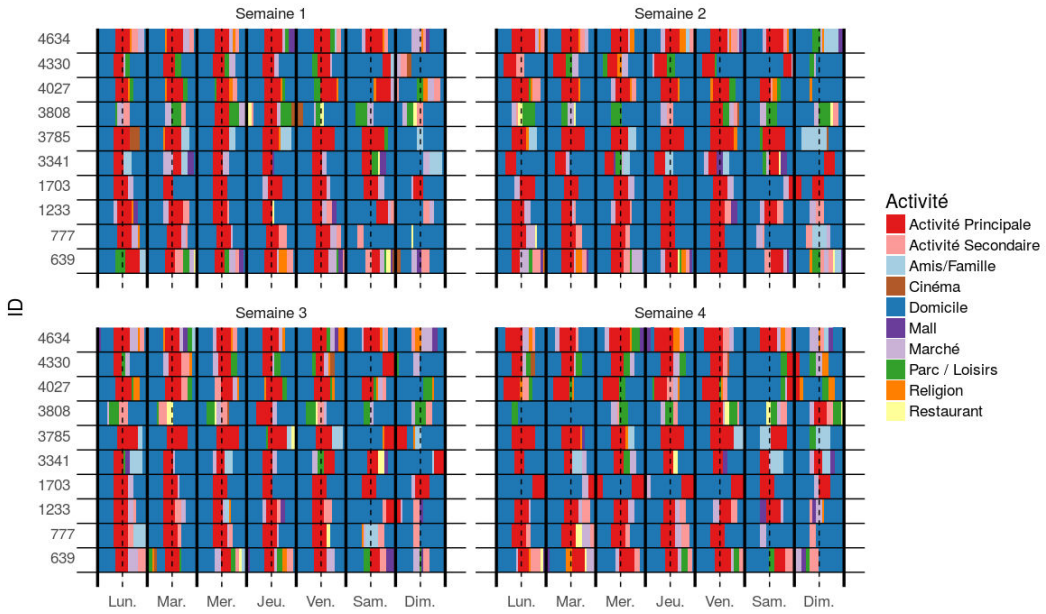


FIGURE 182 Exemple d'agendas individuels pour 10 agents, d'après les données du terrain.

Individuellement, les agendas générés semblent très cohérents, font ressortir différents profils d'individus, et les aspects fixés et flexibles de certaines activités. Globalement, ils permettent de reproduire assez fidèlement les horaires de fréquentation de certains types de lieux, notamment l'activité principale et le domicile. Les AG issus de *Twitter* reproduisent bien les pics de fréquentations en semaine pour les parcs, les AOI et les commerces, mais ne donnent pas de résultats satisfaisants pour les activités de type « Autre ». Il faut dire que ces dernières sont de natures probablement très différentes et représentent la plus grande partie des activités effectuées par nos agents. Les AG issus des données de terrains reproduisent bien les différents pics de fréquentation du matin des parcs, des marchés et des lieux de cultes pour les jours de semaine, avec des amplitudes qui diffèrent et une légère avance dans le temps.

Les séquences des activités journalières sont un point central de cette approche, et permettent d'attribuer de manière indirecte les horaires où un agent effectuera une activité. Nous nous sommes basés sur la matrice de transition entre activités calculée sur l'ensemble des utilisateurs, avec une distinction entre les jours de semaines et de week-end. Certaines

activités, comme aller au restaurant sont très contraintes dans le temps, et peut être faudrait-il utiliser une matrice de transition conditionnée aussi par des plages horaires, comme proposent par exemple (Jiang *et al.,* 2016 ; Wu *et al.,* 2014). Il pourrait être aussi intéressant d'appliquer des matrices différentes selon des catégories d'agents. Ces dernières pourraient être définies selon les regroupements d'agendas basés sur la méthode d'appariement optimale, sur une zone géographique, ou sur des critères liés au type et au nombre d'activités effectuées (chapitre 11).

Mais nous avons montré ici que notre algorithme pouvait s'appliquer à des données de sources très différentes, et la génération d'agendas individuels donnait collectivement des horaires de fréquentations de lieux assez crédibles et parfois très proches des AR. Si notre algorithme peut sans conteste être amélioré, il permet suivant les étapes d'injecter des données d'origines différentes, pour peu que la typologie des activités soit la même. Par exemple si nous avons des jours de fréquentations de lieux, des fréquences et des durées de réalisation d'activités ou encore des heures de départ du domicile provenant d'autres sources, il est possible de les utiliser lors des diverses étapes. Et si le déroulement temporel des activités d'un agent est établi par les agendas, les différents lieux qu'il visitera ne sont pas encore localisés. Nous proposerons quelques pistes et perspectives dans la dernière section du chapitre 11.

Conclusion

Reconstitution d'agendas

Individu de 1 à I		
Lieu de 1 à n	Données Terrain	Données Twitter
Localisation	Connue	Connue
Activité effectuée	Connue	Estimée à partir de l'utilisation du sol
Heure et jours de visites	Approximatifs	Connues
Durée de visite	Approximative	Estimée
Fréquence de visite	Connue	Estimée à partir des données

N simulations

Génération des agendas

Espace d'activité	
Paramètres	*Données / méthode*
Nombre d'activité	Tirage pondéré par les données des agendas reconstitués
Nombre de lieux fréquentés	Tirage pondéré par le nombre d'activité
Nature des activités	Tirage pondéré par la réalisation des activités
⇒ *n* lieux associés à *A* activités	

Lieu de 1 à n	**Temporalités**	
	Paramètres	Données / méthodes
	Heures et jours de réalisation	- Fréquence de visite par activité
	Séquence quotidienne des activités	- heure de départ du domicile - Matrice de transition entre activité - Durée des activités

FIGURE 183 Résumé de la démarche permettant de reconstituer des agendas à partir de données épisodiques, puis d'en générer de nouveaux pour des agents.

En partant de données épisodiques récoltées sur le terrain ou via *Twitter* nous avons reconstitué des agendas (AR), à partir desquels nous avons généré des agendas de synthèses (AG), comme résumé dans la figure 183.

L'étape de reconstitution d'agendas a permis de passer d'espaces d'activités discrets à des agendas continus. La comparaison entre les AR obtenus à partir des données terrain et des données *Twitter* a montré les grandes différences entre ces sources de données. Ceci nous permet d'insister sur l'importance d'effectuer des études sur le terrain, afin d'avoir de collecter des informations permettant d'établir des points de comparaison, mais également de s'imprégner en tant que chercheur des spécificités locales, qui n'apparaissent pas nécessairement dans des bases de données en ligne.

Notre algorithme fonctionne avec divers types de données, pour peu qu'elles contiennent des lieux de nature identifiée, associés *a minima* à des fréquences de visites et des heures et des jours de fréquentation. Si un nombre important d'études proposent des modèles « universels » qui permettent de limiter les données en entrée, que cela soit pour la prédiction des commutations (Lenormand *et al.,* 2012), ou des mobilités intra-urbaines (Noulas *et al.,* 2012 ; Simini *et al.,* 2012 ; Yan *et al.,* 2014) nous choisissons ici une approche centrée sur les données, partant du principe que leur obtention n'est plus un frein si important comme c'était le cas il y a quelques années, et qu'elles peuvent permettre de prendre en compte les particularités géographiques des différentes zones d'études. De plus notre algorithme est assez libre quant à la nature des données qu'il est possible d'injecter. Néanmoins, nous pourrons utiliser certains de ces modèles « universels » pour estimer les déplacements de certains agents, ou l'attribution d'une localisation pour l'activité principale.

L'utilisation conjointe de données terrain et de données *Twitter* a permis la mise en place d'un protocole de modélisation qui fonctionne avec ces deux types de données, et la confrontation des divers résultats nous permet de faire ressortir un grand nombre de pistes pour l'amélioration de notre algorithme. Nous pourrons par exemple envisager de prendre en compte les distances entre les lieux pour réajuster les agendas en intercalant des activités de type « transitions » correspondant au transport, et dont la durée serait définie par le temps de transport entre deux activités. Il conviendra également de réfléchir à une répartition des domiciles des agents en prenant en compte la constitution de leur espace d'activité et/ou leur potentiel de dispersion. Il faudrait aussi adapter par la suite les distances parcourues entre deux activités selon la zone de la ville où ont été placés les lieux de résidences des agents. Par exemple les habitants de Malviya Nagar ont à proximité de chez eux un marché important (AOI) ainsi que les *malls* de Saket (commerces). Les distances nécessaires pour se rendre dans de tels lieux sont donc moindres que pour les habitants de quartiers moins pourvus de ce type de commerces. Notre modèle ne prend pas encore en compte les retours non-définitifs au domicile, par exemple

pour manger le midi.

Contrairement à la plupart des modèles à base d'agent reprenant le concept d'espace d'activités (Jiang *et al.,* 2016 ; Schneider *et al.,* 2013), nous ne nous cantonnons pas à seulement trois activités, (Domicile, Activité principale et Autre), ce qui permettrait une analyse nettement plus fine des types de lieux qui jouent un rôle plus important que d'autres dans la propagation de la dengue.

Pour l'instant, notre modèle ne différencie pas les agents selon l'âge, le sexe ou le statut socio-économique. Dans le cas de la dengue, l'âge est un paramètre important à prendre en compte, les plus jeunes étant plus susceptibles d'être touchés par la maladie (Limkittikul *et al.,* 2014). Il conviendra de travailler sur cet aspect par la suite, mais notre approche par la création d'agendas spatialisée autorise théoriquement la création d'un agent « enfant » qui se rend à l'école la plus proche de chez lui selon des horaires qui découlent des heures d'ouvertures des établissements scolaires, puis effectue ou non des activités extra-scolaires, selon des critères qu'il conviendra de déterminer.

Concernant l'imbrication dans *MO3*, notre approche semi-stochastique permet de reconstituer des agendas légèrement différents pour chaque individu, au gré de tirages pondérés par des fréquences de réalisations, parfois conditionnés par des hypothèses. Pour l'instant, les agents générés n'interagissent pas entre eux. Si cet aspect mérite d'être pris en compte par la suite, ne serait-ce que pour étudier les mécanismes de transmission au sein d'un même réseau social. Mais le caractère indépendant des mobilités des agents et leur aspect déterministe (lors d'une simulation chaque agent effectuera ces activités dans un lieu donné à des heures prédéfinies) autorise la génération des agendas en amont des modèles *MODE* et *MOMA* qui constituent *MO3*, permettant de gagner du temps de calcul.

Si les données disponibles à Delhi ne sont pas représentatives de la population, elles ont tout de même permis de poser les bases théoriques de notre modèle, que nous appliquerons dans la partie suivante consacrée à Bangkok, où nos bases de données sont plus complètes.

Partie D:

Mobilités et activités à Bangkok

« Ma maison est située dans un soi[367]*, semblable aux 56 000 autres soi de Bangkok et de Thon Buri, étroit et populeux. [...] Pendant la journée, ce soi est bruyant. Mais quand vient la nuit, il est désert à faire peur. Il faut dire que seule la première parie du soi est bordée de bâtiments et de boutiques ».*

Nous tenterons dans cette dernière partie d'apprécier et de quantifier cet écrit de (Wanich, 1983), *(cité par Pichard-Bertaux, (2011))*. Mais notre connaissance du terrain est nettement moins développée à Bangkok qu'à Delhi, et nous n'avons pas effectué d'enquêtes *in situ*. Néanmoins, nous y avons collecté un plus grand nombre de données géographiques permettant d'y analyser les mobilités et faire potentiellement écho au texte présenté de Wanich. Ainsi, notre échantillon de *tweets* géolocalisés est représentatif spatialement à Bangkok (chapitre 6). Mais comme précédemment, il conviendra de mobiliser une couche d'utilisation du sol afin de pouvoir associer un type d'activité potentiellement réalisée à une trace numérique, et reconstituer ainsi des espaces d'activités individuels, approche centrale de ce travail. Mais avant d'aborder comme dans la partie précédente la reconstitution et la génération d'agendas individuels, nous ferons un petit détour sur le potentiel des données collectées sur Internet dans l'étude et la compréhension du système urbain de Bangkok, selon un angle d'attaque bien entendu centré sur les mobilités et les temporalités de la ville.

Nous présenterons dans le premier chapitre quelques types de quartiers commerciaux, afin de prolonger la présentation de Bangkok (chapitre 2) et de dresser un tableau plus complet sur l'organisation de la ville. Nous réaliserons ensuite une cartographie de l'utilisation du sol en mobilisant des bases de données géographiques diverses. La mise en relation de cette couche avec des *tweets* et des *check-in* datés et géolocalisés nous permettra alors d'apprécier les temporalités des différentes activités dans la ville.

367. ruelle

Le chapitre suivant explorera davantage les différentes rythmiques et discontinuités des mobilités dans la ville, sous le prisme des données et des méthodes qui impliquent des variations et des nuances dans les résultats. Nous commencerons par présenter les conditions et moyens de circulation à Bangkok, soit le socle des mobilités et des déplacements urbains. Puis nous étudierons les interactions entre les différents secteurs de la ville, qui nous permettront de définir des sous-régions fonctionnelles, ou voisinages.

Si les deux premiers chapitres ont pour objet principal la ville de Bangkok, nous changerons d'échelle dans le dernier chapitre et aborderons les mobilités sous leurs aspects individuels. Dans la lignée des chapitres 7 et 8 consacrés à Delhi, nous créerons différents groupes selon des potentiels de mobilités et appliquerons et améliorerons nos algorithmes de génération d'agendas.

Chapitre IX: Temporalité des activités à Bangkok

1 Différentes facettes des activités commerciales

Bangkok est une ville éminemment tournée vers le commerce, et ce dernier revêt différentes formes, qu'il s'agisse de grands quartiers commerciaux ouverts faits de petits stands, de *malls* de tailles et de standings très variés, ou encore de zones où s'installent temporairement des vendeurs de rue Bangkok en comptabilisait 280 000 en 2003 ((Yasmeen et Nirathron, 2014) notamment le soir pour vendre de la *street-food,* dont Bangkok fait figure de capitale mondiale[368].

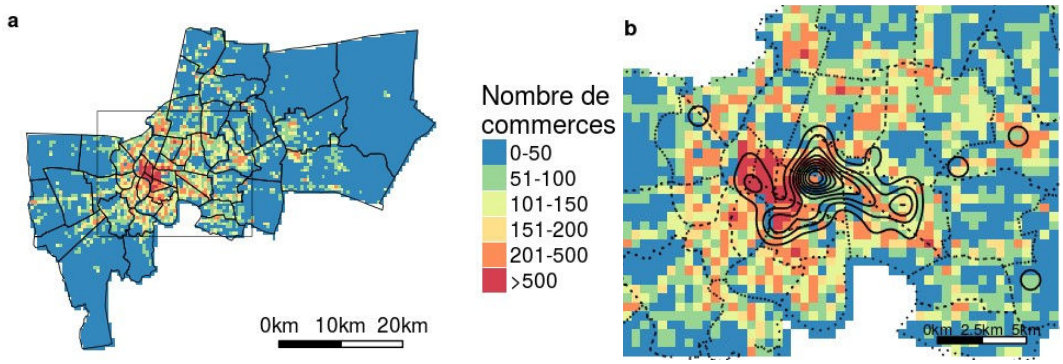


FIGURE 184 Nombre de commerces selon une grille de 500 m en 2009 (AIT) (a), et densité des *malls* à Bangkok (d'après des données de Wikipédia, géolocalisées via geocode de *Google*).

La figure 184.a montre que la plupart des zones qui présentent le plus de commerces en 2009 sont clairement localisées dans le centre-ville, même si certaines poches de rang inférieures jalonnent ponctuellement la ville, toujours selon les grands axes de communication. La figure 184.b est un zoom de la précédente, avec en addition le nombre de *malls* répertorié par Wikipédia en 2018[369], présenté sous forme de densité (portée de 2,5 km). Les données sont de sources et de dates différentes, ce qui peut suggérer soit qu'un nouveau pôle de *malls* s'est créé à l'est de la tache rouge, soit que la densité de commerces est plus importante hors des *malls*. La figure 185 suivante localise les quelques exemples de zones marchandes que nous allons présenter.

368. https://ed t on.cnn.com/trave /art c e/best c t es street food/ ndex.htm
369. La liste des malls à Bangkok a été récupérée sur https://en.w k ped a.org/w k / L st of shopp ng ma s n Tha and puis géolocalisé à partir de leur nom en utilisant le service *geoname* http://www.geonames.org/

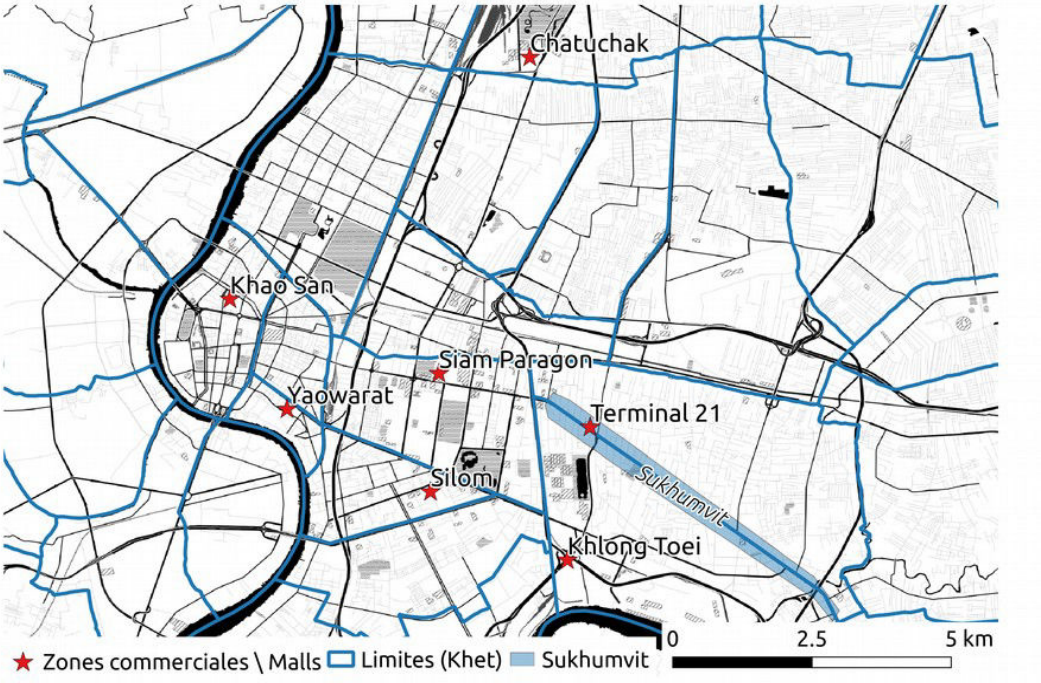


FIGURE 185 Localisation des zones commerciales dont nous allons parler dans la section. Fond de carte : Stamen.

Le quartier de Yaowarat (figures 185 et 186), considéré comme le *ChinaTown* de Bangkok, correspond globalement à la grande tache rouge au centre de la figure 184. Il s'agit du quartier commercial traditionnel par excellence, dans le sens où il est constitué d'une succession de petits restaurants de rues et de petits commerces, vendant de tout, de l'électronique et des gadgets, aux fruits et légumes, aux peluches et aux jouets (figure 186), et ce sur au moins 20 hectares. Ce quartier, très attractifs tous les jours de la semaine est extrêmement fréquenté par la population locale et est une destination prisée des touristes. C'est également dans le district de ce quartier que sont enregistrés les taux d'incidences les plus importants de dengue (chapitre 2).

Au nord-est du centre de Bangkok se trouve le grand marché de Chatuchak qui prend place tous les week-ends. Il s'agit d'un des plus importants marchés d'Asie, qui draine un grand nombre de visiteurs [370] (figure 187.d). Le quartier de Khlong Toei, considéré comme l'un des plus grands bidonvilles de Bangkok accueille également l'un des plus grands marchés de la ville (figures 187.a et 187.b), spécialisé dans la vente de produit frais et ouvert de 6 heures à 2 heures du matin [371].

370. http://www.bangkok.com/shopp ng market/popu ar markets.htm
371. http://www.bangkok.com/shopp ng market/ oca markets.htm

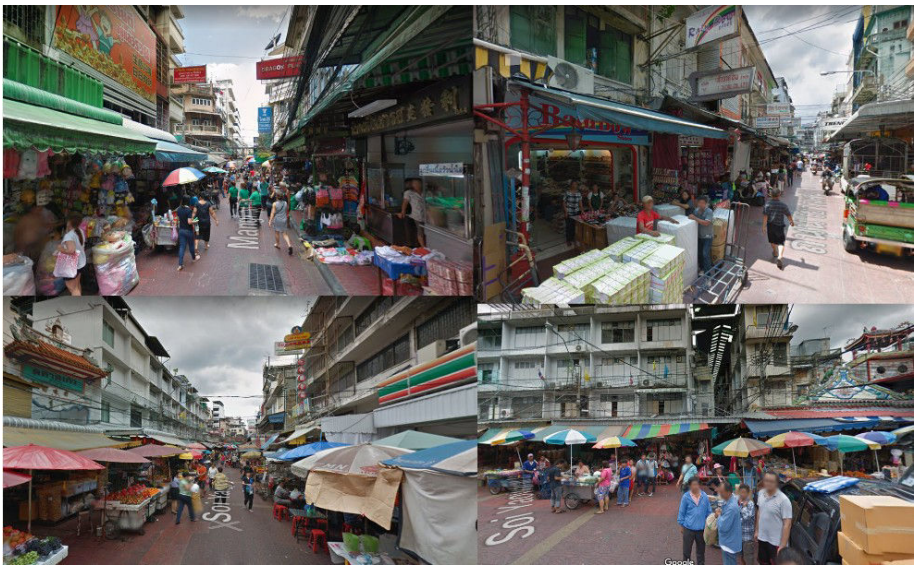


FIGURE 186 Photos du quartier de Yaowarat. Source : *Google Streetview*, (2011,2017).

FIGURE 187 Quelques photos de marchés à Bangkok. Marché de Khlong Toei (a & b, source *Google* street view). Petit marché au pied du métro wutthakat (à proximité de Bang Wa) (c), et le grand marché du week-end de Chattuchak (d).

À ces grandes zones marchandes, s'ajoutent également des petits marchés, comme celui au pied du métro de Wutthakat, dans le secteur de Bang Wa (figure 187.c), où des marchés temporaires, composés de vendeurs ambulants qui s'installent le soir aux croisements de certaines rues. Bangkok a également conservé sa tradition de marché flottant, où des petits bateaux s'amassent ponctuellement dans différents khlongs[372] pour vendre divers produits de bases.

372. canaux

FIGURE 188 Présence à Khao San road à différents moments de la journée. Photographie Brenda Le Bigot, dans Cebeillac et Le Bigot, (à paraitre)

Pour ce qui est des sorties, Bangkok n'est pas en reste, qu'il s'agisse d'une partie des quartiers qui longent la grande artère de Sukhumvit, ou encore le secteur de Khao San Road. Ce dernier offre des tarifs relativement raisonnables et est le produit d'un brassage de population important, qu'il s'agisse de touristes, d'une certaine jeunesse Bangkokoise ou encore de vendeurs de rue ambulants. Ce quartier revêt différents visages au cours d'une même journée. Quatre phases ont déjà été identifiées (figure 188) : l'installation des commerces et l'arrivée des touristes entre 12 h et 18 h, puis une période de saturation jusqu'à 22 h. S'ensuit alors une polarisation où les groupes se stabilisent autour des bars et se dispersent progressivement entre minuit et 2 h (Cebeillac et Le Bigot, 2018).

En plus de ces zones commerçantes ouvertes sur la rue, Bangkok accueille un nombre croissant de *mall*, l'équivalent de nos galeries marchandes, mais sur plusieurs étages et sur des surfaces généralement bien plus importantes. Quasiment toujours situés à proximité des stations de métro, ce qui les rend accessibles, ces lieux de consommation jouent un rôle primordial dans la vie des Bangkokois. Un article de Wangtechwat pour le Mekong Review[373] intitulé « life as a shopping mall » et repris par le courrier international[374] décrit d'ailleurs leur importance :

« L'activité la plus pratique, accessible à tous, c'est une promenade dans un centre

373. https://mekongrev ew.com/ fe shopp ng ma /
374. N°1416 17 18 21 décembre 2017

commercial. On y trouve tout ce qu'on veut au même endroit. On peut par exemple faire ses courses, se restaurer, regarder un film ou jouer au bowling. Tout habitant de Bangkok attestera que nous passons la plus grande partie de notre temps au centre commercial. C'est là qu'on voit du monde, qu'on se promène, qu'on mange, qu'on fait ses courses. C'est là qu'on se rend pour être vu, ou juste histoire d'aller quelque part ».

FIGURE 189 Photos de quelques *malls* à Bangkok. Le MBK Center (A) et le Siam Paragon (C), dans le quartier de Siam, et le Terminal 21 (B), plus à l'est.

152 *malls* étaient recensés à Bangkok par Wikipédia en mars 2018[375], et si leur prolifération accompagne la montée de la classe moyenne (Sophorntavy, 2017), tous n'offrent pas le même standing et ne visent pas la même clientèle. Le quartier de Siam-Ratchaprasong, en face de l'université de Chulalongkorn, accueille deux des vingt plus grands *malls* du monde en 2015[376], notamment Central World (10e) et Siam Paragon (13e, figure 189.c). Ce dernier, inauguré en 2005, fut le lieu le plus cité sur *Instagram* en 2012 et 2013[377]. D'une architecture extérieure moderne et luxueuse, il contraste avec le vieil MBK Center ouvert en 1985 (figure 189.a) situé à deux pas (les deux *malls* sont d'ailleurs reliés par des plateformes piétonnes). Un peu plus à l'est sur la même ligne de métro aérien, se trouve l'imposant Terminal 21 (figure 189.b), lui aussi directement connecté au métro.

Les *malls* de grandes tailles ont en général une offre très diversifiée, et permettent aux chalands d'accéder à un grand nombre de services et commerces dans une même enceinte.

375. https://en.w k ped a.org/w k /L st of shopp ng ma s n Bangkok
376. https://www.courr er nternat ona .com/grand format/ nfograph e es p us grands centres commerc aux du monde
377. https://www.huffingtonpost.ca/2013/12/13/s am paragon most nstagrammed n 4441058.htm

Par exemple, Siam Paragon est divisé en plusieurs sections dédiées au luxe (figure 190.c), aux technologies et au numérique, à des restaurants (figure 190.d), et autres commerces et loisir comme des cinémas et l'un des plus grands aquariums d'Asie du Sud au moment de l'inauguration (McGrath, 2006). Mais tous les *malls* n'ont pas la même diversité et le même standing, et si le Silom Complex (figure 190.a) tente de rivaliser avec Siam Paragon et Central World, malgré une taille plus modeste, le MBK Center présente un agencement intérieur moins ouvert et parait moins moderne (figure 190.b). Si ces *malls* se veulent généralistes, d'autres sont plus spécialisés, notamment dans l'informatique comme le vieillissant Pantip Plaza.

FIGURE 190 Quelques photos prises à l'intérieur de *malls*. Le Silom complex (a), le MBK center (b), ainsi que deux secteurs du mall de Siam Paragon. L'un dédié au luxe (c), l'autre à la gastronomie (d).

Aux quartiers commerçants de tailles et de finalités diverses, qu'il s'agisse de vente de détail, de *street-food*, de bars ou d'un mélange des trois, s'ajoutent donc ces grandes galeries marchandes fermées et climatisées, qui de par la grande variété des activités qu'elles proposent en font des lieux attractifs par excellence. Au-delà de ces descriptions sommaires mais nécessaires, un véritable enjeu revient à localiser ces différents secteurs, en sachant que nous n'avons pas eu accès aux bases de données commerces de la ville.

La prochaine section proposera donc d'utiliser des sources de données non-institutionnelles afin de réaliser une cartographie des différents types de zones marchandes, notamment les lieux plus orientés vers les restaurants et les sorties, et ceux plus centrés sur la vente au détail. Nous prendrons également en compte d'autres types de lieux, comme les écoles, les universités, les

gares et les lieux de cultes, qui sont importants à la fois dans le quotidien des Bangkokois et comme lieux de contamination privilégiée de la dengue (Wen *et al.*, 2015).

2 Apports des données *Google* et *OSM* à la cartographie de l'utilisation du sol à Bangkok

Nous avons déjà abordé dans le chapitre 8 la collecte et le couplage des données issues de *Google* et d'OpenStreetMap pour obtenir une couche d'utilisation du sol. Après un bref rappel sur ces données et sur les méthodes de collecte, nous proposerons ici des méthodologies plus poussées permettant d'obtenir *apriori* des couches spatiales plus adaptées.

2.1 Collecte des données

POI Google

Comme vu dans le chapitre 8, il est possible d'accéder et donc de collecter ce que le service de cartographie de *Google* appelle les Points d'intérêts (ou *POI*) à partir de simples requêtes [378]. Ces *POI* proviennent de différentes bases, comme l'équivalent des pages jaunes [379], de l'office du tourisme, soit par la contribution d'utilisateurs *lambda*, ou encore par la collaboration avec d'autres entreprises. Des informations générales sont par exemple fournies par "space miner" ou "where in thailand" [380], tandis que des POI concernant des restaurants ou des bars proviennent de "everydaydiningdelight.com", ou encore d'"openrice" [381]. *Google* s'appuie en plus sur le projet *Ground truth*, qui vise à créer automatiquement des *POI* à partir de l'analyse d'image de son service *Street View* (voir chapitre 8.1.2), bien établi à Bangkok.

Nous avons donc élaboré un code sous python qui utilise les outils disponibles dans l'*API* de *Google Places* [382] pour récupérer les différents *POI* dans une fenêtre spatiale mobile. Il suffit pour cela de définir des coordonnées spatiales, ainsi qu'un rayon de recherche relativement faible (ici 200 m), et de télécharger tous les résultats de la requête, puis de décaler le centre de la fenêtre vers l'ouest (ici de 100). Une fois une ligne complétée lorsque la longitude maximale que nous avons définie est atteinte nous réaffectons la longitude initiale et déplaçons la latitude de 100 m vers le nord et reproduisons la procédure. Ceci permet d'obtenir *a priori* l'ensemble des *POI* de *Google*, avec un grand nombre de doublons, du fait de la superposition des fenêtres de recherche. L'utilisation d'un rayon de recherche relativement faible est motivée par un aspect de l'algorithme de *Google* qui va tendre à renvoyer certains points plutôt que d'autres favorisant ainsi certains établissements et au fait qu'il n'affiche au maximum que

378. Mais les conditions d'utilisations et notamment les quotas ont été restreints par *Google* en juin 2018.
379. http://www.ye owpages.co.th/en
380. http://www.where n tha and.com/document/ ndex.php
381. https://th.openr ce.com/en/bangkok
382. https://deve opers.Goog e.com/p aces/web serv ce/search?h =fr

20 pages de 10 résultats. Un rayon de recherche trop important écarterait donc certains *POI* si la zone en contenait plus de 200.

Le nombre de requêtes quotidiennes est limité à 2000 par clés fournies par le service, plusieurs jours ont donc été nécessaires pour collecter les 257 924 *POI* répertoriés sur la zone [383] en juillet 2015. Parmi ces *POI*, 43 % d'entre eux (107 450) ont un type d'établissement spécifié, dans l'une des 99 catégories renseignées autre que "Points d'Intérêt", qu'il s'agisse d'une école, d'un spa, ou d'un centre commercial [384]. Nous réduisons ensuite l'emprise géographique à Bangkok et les districts limitrophes pour obtenir 93 293 *POI* exploitables.

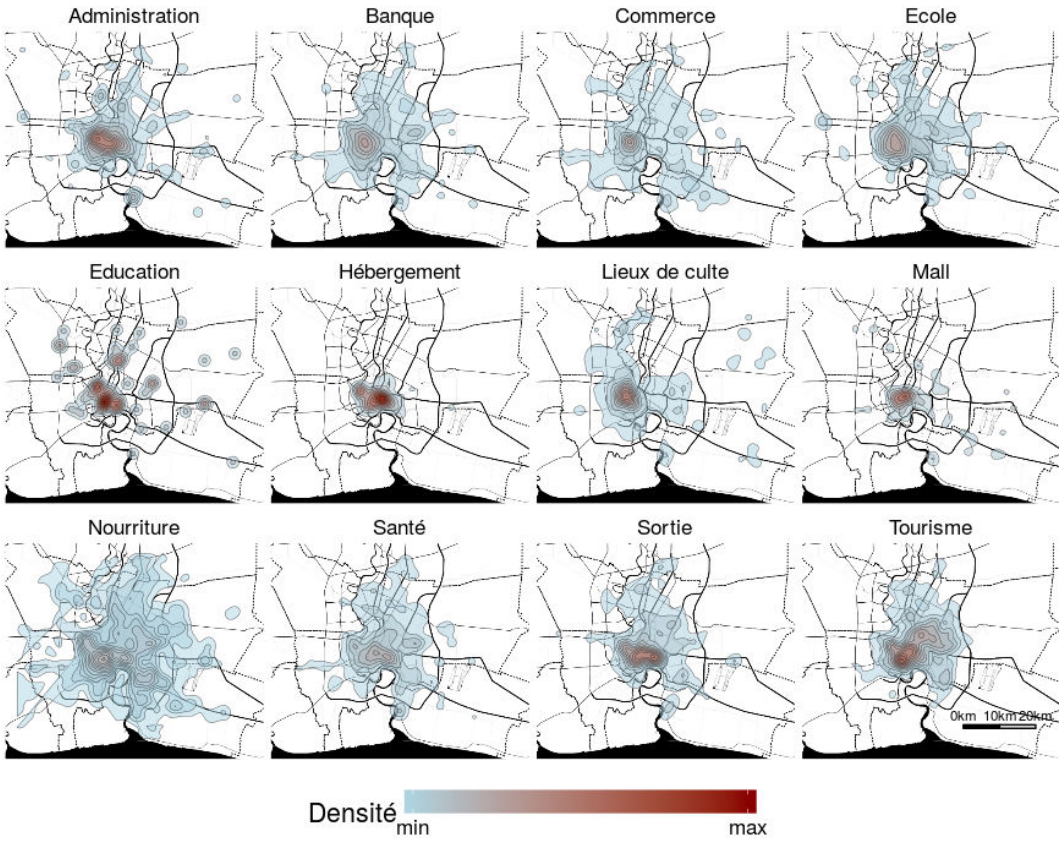


FIGURE 191 Répartition de 12 catégories de POI dans Bangkok, selon une fonction de densité de portée 5km.

Nous avons procédé à des regroupements de catégories, afin d'obtenir moins de classes. Ainsi, tous les *POI* considérés comme des « bar », « pub », « café », ou encore « restaurant » ont été reclassés comme des lieux de sorties ; les « hôtels » ou « *guest house* » comme des lieux d'hébergements ; les universités et autres bibliothèques comme des lieux d'éducation. Certaines classes de *POI* sont très vagues, notamment celle « Food », qui peut s'apparenter à

383. longitudes comprises entre 100,1 et 101°, et latitudes entre 13.45 et 14.5
384. https://deve opers.Goog e.com/p aces/supported types?h =fr

n'importe qu'elle commerce proposant de la nourriture, sans plus de distinction il peut s'agir de restaurants de rue ou de stands vendant des chips. Nous avons décidé de conserver cette catégorie sous le nom de « nourriture ». Pour plus de détails voir l'annexe E. La figure 191 montre la localisation de 12 catégories dans Bangkok, sous forme de densité (d'après une *kernel density estimation,* selon une portée d'environ 5 km 0,047°). Il ressort tout d'abord que les plus fortes densités de chacune des catégories se trouvent dans le centre-ville. Ensuite, selon les types de lieux, nous observons une concentration plus ou moins marquée dans la ville. Par exemple, les lieux de type « commerce », ou « nourriture », sont répartis de manière à peu près homogène, alors que les *malls* et les lieux d'hébergements sont regroupés principalement dans le centre. Les lieux de sorties, les écoles, les lieux de santé (hôpitaux, pharmacies), administrations, lieux de cultes et touristiques présentent en plus d'un niveau de densité très marqué dans le centre-ville, des clusters secondaires éparpillés en périphérie. Les lieux d'éducations correspondent ici à des bibliothèques ou des universités et la figure 191 fait ressortir la localisation des principaux campus.

Cette répartition des commerces et services dans la ville met en avant les disparités entre le centre et la périphérie. Il est à noter que ces résultats obtenus avec les *POI Google* sont cohérents avec d'autres études (Vichiensan, 2009, 2007) conduites à une résolution spatiale plus grande, en mobilisant des données institutionnelles qui ne distinguent pas de catégorie de commerces.

AOI et autres données surfaciques de Google

Mais l'information spatiale provenant de *Google* ne se limite pas à des points d'intérêts, et comme nous l'avons vu précédemment (chapitre 8), *Google Maps* raisonne également en termes surfaciques. Y figurent notamment la notion d'AOI (*Area of Interest*) où l'entreprise présente d'une certaine couleur (orange clair) des zones probablement commerciales susceptibles d'être attractives pour la population, auxquelles s'ajoutent d'autres types d'utilisation du sol comme des parcs (vert), des écoles ou des universités, ou encore des hôpitaux[385]. Nous avons appliqué la même méthode que dans le chapitre 8 sur Delhi pour récupérer les *AOI*, les lieux d'éducations, les hôpitaux, et les parcs. À noter que ces données ne sont probablement pas exhaustives, car certaines zones commerciales comme le grand *mall* de Central World pourtant l'un des plus grands au monde et le marché du samedi de Chatuchak, ne sont pas considérées comme des AOI, mais nous y reviendrons par la suite.

Données OSM

De manière similaire au travail effectué précédemment sur Delhi, nous avons récupéré les

385. Pour le code couleur, voir chapitre 8, ou https://b og.Goog e/products/maps/d scover act on around you w th updated/

polygones d'*OpenStreetMap* situés à Bangkok, notamment ceux concernant les écoles, les lieux de cultes, les hôpitaux, les universités, les gares, les aéroports, les parcs et complexes sportifs. Nous avons également récupéré les routes majeures, qualifiées de « trunk », « major road », « primary », et « secondary ». Nous avons fait le choix de ne pas prendre en compte les zones commerciales, tout d'abord parce qu'elles ne nous paraissaient pas forcément bien renseignées, et d'autre part parce que nous pensons que la base de données de *Google* relative à ce type de lieux est probablement plus complète, compte tenu de la nature et les objectifs du service, largement orienté vers le business.

Création de couches d'utilisation du sol

À partir de ces différentes données, nous allons créer une couche d'utilisation du sol. Nous mobiliserons ultérieurement la base de données de *Facebook Places*, à titre de comparaison (section 2.4). Les données relatives aux fonctions des lieux sont de natures très différentes : certaines sont ponctuelles (*POI*), d'autres surfaciques (*AOI*, *OSM*). Les commerces sont en général d'une étendue spatiale assez limitée, de l'ordre de quelques dizaines de mètres carrés. Si les *POI* renseignent également les *malls* qui occupent une grande superficie, ces derniers peuvent être définis par l'agrégation des commerces et autres *POI* (lieux de sorties, hébergements) qui les composent. Mais ce n'est pas le cas pour les lieux d'éducations, de santé (hôpitaux), de loisir (parc d'attractions, parc, complexes sportifs) ou encore religieux (grands temples) et administratif. En somme, une distinction peut également se faire par le type d'activité associé à un lieu, à savoir s'il s'agit plutôt d'une activité commerciale (commerce, lieu de sortie), ou d'un type de lieu à vocation non marchande, évoqué précédemment. Nous allons ici présenter deux méthodes permettant d'obtenir une typologie des activités commerciales, que nous compléterons ensuite avec l'ajout de données surfaciques représentant d'autres utilisations du sol (lieux d'éducations, de cultes, transports, etc.).

2.2 Typologie des activités commerciales à partir de POI

Peu d'études ont à notre connaissance utilisé des *POI* issus d'Internet pour réaliser des cartes d'utilisation du sol. Nous pouvons néanmoins citer les travaux de (Phithakkitnukoon *et al.*, 2010) qui ont appliqué un Kmeans sur des *POI* de *Yahoo !* regroupés dans des mailles de 500 m, ou encore (Zhan *et al.*, 2014) qui ont utilisé le même algorithme, mais sur des POI issus de Foursquare dans des cellules de 200 m. (Hu *et al.*, 2016) ont réalisé une classification supervisée sur des données *OSM* en créant au préalable des couches de densités pour les différentes catégories. Pour plus d'exemples, nous vous renvoyons à la section 5.2 du chapitre 5 relatif à l'état de l'art.

Choix de l'unité de base et agrégation des POI

Il apparaît néanmoins que l'approche la plus intuitive reste d'agréger des *POI* dans des mailles puis de réaliser des classifications (supervisées ou non) de ces-dîtes mailles. Le découpage spatial et donc la taille des cellules influence les résultats des traitements et analyses statistiques, entraînant un effet de *MAUP* (Modifiable Area Unit Problème) (Openshaw et Taylor, 1979). De trop petites cellules risquent de contenir des assemblages bien spécifiques de type de *POI* qui ne ressortiraient pas à une résolution moins fine. Nous n'avons pas fait ici d'analyse systématique, et nous avons décidé de travailler sur des cellules de 180 m², ce qui est à la fois un multiple de l'unité de base utilisée par d'autres travaux liés à cette thèse (Maneerat, 2016 ; Misslin, 2017)- un pixel d'une image *Landsat* 8 de 30 m de côté tout en constituant une nouvelle entité spatiale de la taille d'un pâté de maisons. Dans un premier temps, nous regroupons les *POI* de *Google*classés en activités commerciales et de services, soit environ 66 800 points, en 6 classes :

- Les lieux d'hébergement (hôtel, *guest house*) ;
- Les lieux de restauration et de loisirs bien définis (restaurant, café, pub, discothèques)
- les lieux de restauration moins bien définis (nourriture)
- Les magasins non définis, où seul l'attribut 'shop' figurait
- Les magasins liés à l'apparat (salon de coiffure, magasins de vêtements, vendeur d'appareils électroniques, magasins pour la maison, etc.)
- les entreprises diverses (réparation de voiture, entreprise spécialisée dans les appareils électriques ou électroniques, etc.)

Ces *POI* sont ensuite agrégés dans un carroyage de cellule de 180 m².

Méthode : Kmeans & CAH

La typologie des fonctions commerciales et de services est ensuite réalisée en deux étapes : (i) application d'un Kmeans et (ii) classification ascendante hiérarchique (CAH). La réalisation d'une classification non supervisée suivant l'algorithme des nuées dynamiques (Kmeans) en 500 classes (2000 itérations, 100 échantillons aléatoires) permet de faire un premier regroupement des mailles. Nous réduisons ainsi le nombre d'individus afin d'accélérer les traitements ultérieurs, et le nombre élevé de classes permet d'optimiser le regroupement sans perdre trop d'information. En nous inspirant de Fleury *et al.*, (2012), nous effectuons ensuite une CAH selon la méthode de la variance minimum de Ward (1963) (figure 192), en utilisant la distance du diamètre maximum afin de maximiser la distance entre deux individus appartenant à deux classes distinctes. Compte tenu du fait que dans la plupart des cas, les branches du dendrogramme ne sont pas équilibrées en termes de nombre de classes regroupées, nous avons décidé d'adapter ce dernier. L'objectif

est de conserver la structure hiérarchique et d'associer les branches en fonction d'un nombre de mailles maximum par cluster.

La première étape est une homogénéisation des niveaux, avec un maintien de la structure hiérarchique globale. On ne raisonne plus en distance, mais en rang à partir du premier nœud. Les différents niveaux sont ici gardés à titre indicatif, mais ne reflètent plus la distance réelle entre les différents groupes.

La seconde étape est un regroupement des branches de manière à ce qu'elles aient un nombre de cluster maximum, ce qui permet d'éviter qu'une « coupe » du dendrogramme entraîne la création de classes contenant un nombre trop faible ou trop grand d'individus (figure 192).

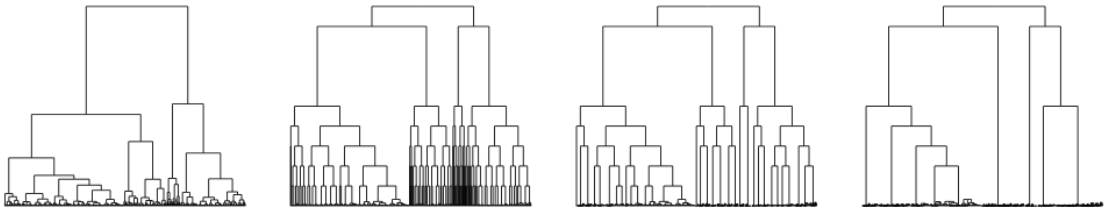


FIGURE 192 Dendrogramme original (gauche), et dendrogrammes avec niveaux homogénéisés avec au maximum respectivement 1, 20 et 100 clusters par branches finales.

Cette méthode, même si elle fait perdre la notion d'inertie et donc la part d'information résumée, permet de garder la structure des classes de niveau inférieur, et donc d'éviter d'obtenir des classes contenant un faible nombre d'individus d'un côté, et un regroupement de classes présentant un grand nombre d'individus de l'autre. On s'affranchit en quelque sorte des distances inter-classes trop faibles, en gardant la structure en grappe originale. Nous avons ensuite défini de manière empirique un critère d'agrégation (taille minimum du cluster, ici 100) et un nombre de classe désiré (ici 11).

Résultats

Les 11 classes obtenues sont ensuite regroupées en 6 grandes classes caractérisées respectivement par une surreprésentation (i) des lieux d'hébergements, (ii) des lieux de loisirs et de shopping, (iii) des lieux de shopping largement majoritaires, (iv) des commerces indifférenciés, (v) des activités du secteur secondaire (activités liées aux voitures, entreprise d'électronique, etc.) et (vi) des zones indifférenciées contenant peu de commerces. Voir l'annexe K pour l'interprétation des différentes classes. La validation a été effectuée en confrontant les résultats de la classification à la connaissance du terrain. Cette méthode, qualifiable de semi-supervisée est applicable à d'autres types de données matricielles (*e.g.* images satellites).

La typologie des activités commerciales (figure 193) montre l'organisation de Bangkok sous forme d'assemblages, diffus en périphérie et plus dense et spécialisé dans le centre-ville. Globalement, les zones contenant quelques commerces forment une sorte de bruit de fond de la ville, que nous pouvons interpréter comme des commerces de proximités, auxquels se juxtaposent sporadiquement quelques zones où les activités commerciales sont plus importantes. Mais dans le grand centre apparaît des niveaux de spécialisation nettement plus marqués, qu'il s'agisse des lieux d'hébergements caractéristiques des lieux touristiques (Khao San, Sukhumvit) ou d'affaire (Silom), des zones où les densités de commerces sont élevées et où leurs types sont variés, comme dans les malls de Siam Paragon ou de Central World, ou dans l'un des plus grands marchés d'Asie qu'est le marché du week-end de Chatuchak. Le quartier chinois de Yaowarat est quant à lui divisé en deux, une partie contenant des commerces indifférenciés, et l'autre une représentation des ateliers et petites entreprises, ce qui reflète la réalité du terrain.

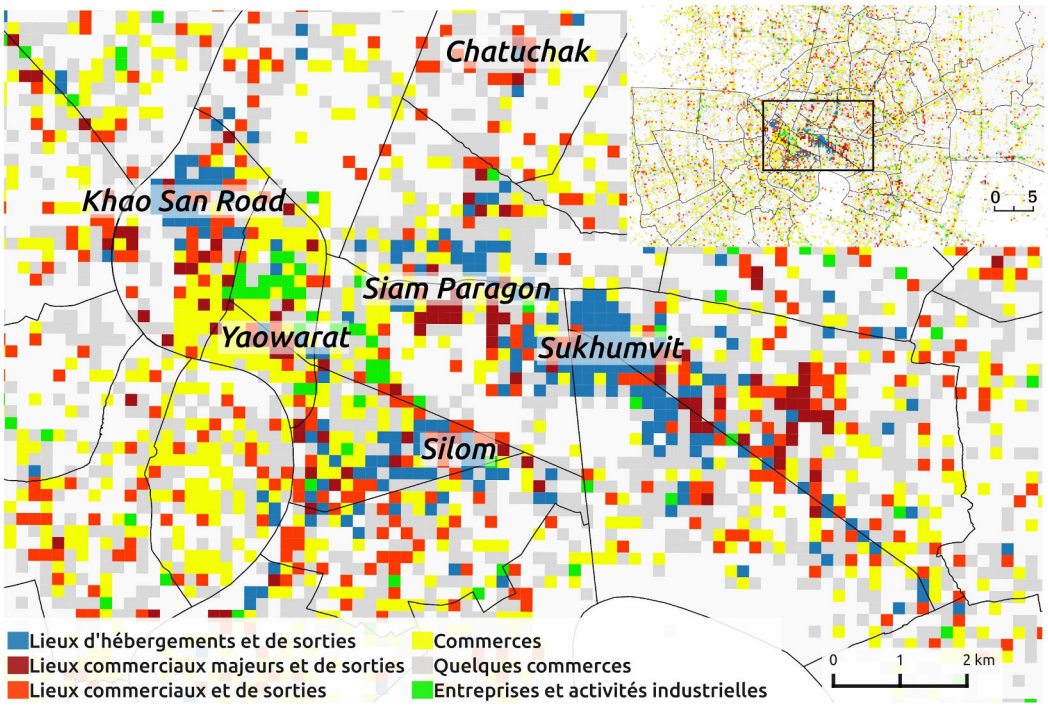


FIGURE 193 Typologie des activités économiques à Bangkok et dans le centre-ville.

Néanmoins si l'approche est assez originale et conduit à une typologie d'assemblage des différentes zones commerciales de la ville, cette méthode est relativement complexe, sujette au *MAUP* et nécessite un grand travail d'interprétation. Nous proposons maintenant une approche plus simple, en mobilisant les *AOI* de *Google*.

2.3 Mobilisation des AOI

L'avantage des *AOI* est qu'ils constituent déjà *a priori* les zones potentiellement attractives de la ville. Néanmoins, comme dit précédemment, de grands *malls* et zones commerciales ne sont pas considérés comme tels. Nous allons dans un premier temps analyser les *POI* qui constituent ces *AOI*. Comme ces zones sont parfois séparées par des routes, alors qu'elles forment un même ensemble, nous les regroupons en effectuant une dilatation de 50 m, puis une érosion de 40 m. Nous obtenons ainsi 184 zones, qui contiennent 8058 *POI* (8,6 % du total). La figure 194 montre le lien entre le nombre de *POI* et la surface des *AOI*. Si ces zones peuvent être caractérisées par une densité relativement constante de *POI* (figure 194.a), la plupart d'entre elles sont de superficie relativement faible (figure 194.b), inférieure à 1 hectare.

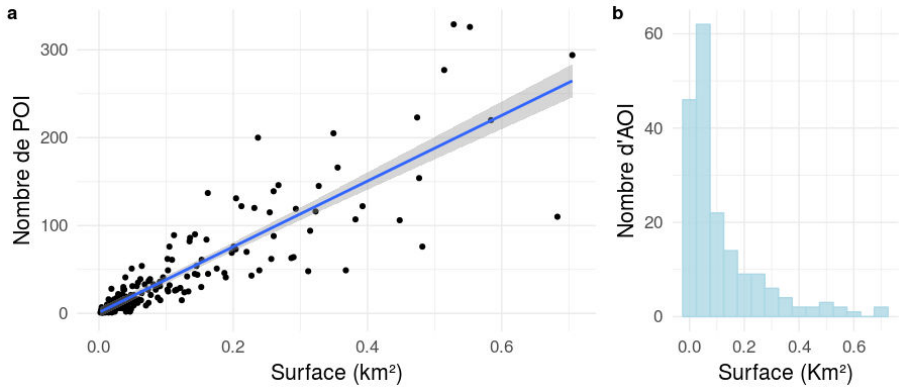


FIGURE 194 Lien entre le nombre de *POI* et la surface des *AOI* (a) et répartition du nombre d'*AOI* en fonction de leur superficie (b).

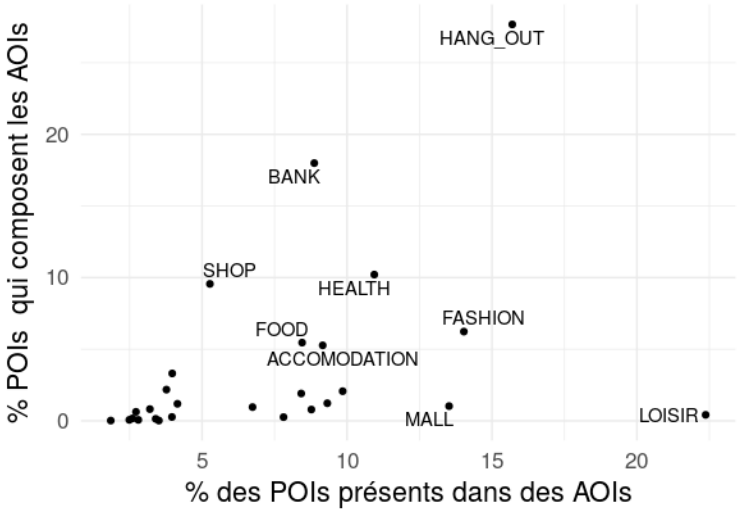


FIGURE 195 Répartition des différentes catégories de *POI* dans des *AOI*.

La figure 195 montre le pourcentage de chaque catégorie de *POI* dans les *AOI* en fonction

de la part que ces *POI* représentent dans les *AOI*. Plus clairement, cela signifie par exemple que 25 % des *POI* de type « Loisir » (musées, stades, parcs, etc.) sont compris dans des *AOI*, mais qu'ils ne représentent que 0,4 % des points présents dans des *AOI*. De même, les lieux de sorties (« Hang out ») composent 28 % des *POI* des *AOI*, et 13 % d'entre eux se trouvent dans de telles zones.

Les *POI* de *Google* en notre possession, et la composition des AOI plus ou moins révélée, il est donc possible de créer nos propres zones d'intérêts pour les fusionner ensuite avec celles de *Google*. Les sections précédentes de ce chapitre ont montré qu'il existe grossièrement trois types de zones commerciales : les zones de sorties où les bars, restaurants et discothèques sont très présents, les zones de commerces de détail, et les zones mixtes, mélangeant toutes les précédentes catégories.

Au regard de ces informations, nous allons d'abord sélectionner dans l'ensemble les *POI* correspondant à des commerces de détail, soit les commerces (« shop »), les commerces liés à l'apparat (« fashion », qui englobent les magasins de chaussures, d'habits, et autres salons de beautés et coiffeurs), et les *malls*. Nous allons ensuite regrouper sous le fanion « lieux de sorties et de loisirs » les *POI* de type « Hang out », « Accomodation », « Loisir » et « Food ». Nous effectuons ensuite un regroupement de ces différents *POI*, en appliquant l'algorithme *dbscan* déjà utilisé pour créer les espaces d'activités des utilisateurs de *Twitter*. Nous définissons de manière assez arbitraire un rayon de recherche de 50 m ainsi qu'un nombre minimum de 3 points, afin de ne garder que les zones les plus densément dotées de *POI* de chacune de ces classes. Nous obtenons ainsi nos propres zones marchandes, que nous fusionnons avec les *AOI* de *Google* pour obtenir 2101 polygones, ou "néo-*AOI*", composés de 10 832 lieux de sorties et 7222 commerces de détail. Les *malls* et marchés qui nous semblaient manquants dans les *AOI* font maintenant partie de ces néo-*AOI*.

Définir la catégorie du néo-AOI

Nous allons classer ces néo-*AOI* en 3 types de catégories : selon s'ils sont dominés par des commerces de détails, des zones de loisirs ou s'il s'agit de zones mixtes. Nous allons effectuer ce travail en affectant une de ces catégories à une zone selon les *POI* majoritaires qui la composent. Mais à partir de quel seuil définir une zone mixte ? Lorsque ces zones sont constituées du même nombre de *POI* de chaque catégorie, à x% près ? Afin d'éclairer notre choix, nous mobilisons un indicateur de diversité parmi d'autres, l'indice de Shannon (Shannon, 1948), communément utilisé en écologie pour estimer le niveau de dominance d'une espèce dans une zone donnée. Plus sa valeur est importante, plus la diversité dans la zone étudiée est grande, et inversement lorsque l'indice se rapproche de 0, où une classe domine plus largement les autres. L'entropie de Shannon H(x) s'écrit :

$$H(x) = -\sum_{i=1}^{n} P_i \ln(P_i) \tag{26}$$

Où P le pourcentage que représente l'une des catégories *i* (commerces de détail ou lieux de sorties, loisirs), dans chaque néo-*AOI*. L'entropie standardisée de Pielou, normalise entre les valeurs entre 0 et 1 (Pielou, 1966). Il se définit comme H(x)/ln(n).

Dans notre cas, l'indice de Pielou est égal à celui de Shannon, puisque nous n'avons que 2 catégories (ln(2) = 1). Nous aurions pu utiliser d'autres indices (Gini, Simpson, α de Fischer, etc.), mais c'est ici l'occasion d'introduire l'entropie de Shannon, qui est également utilisée pour définir des niveaux de régularités dans des données spatio-temporelles où une trajectoire routinière est associée à une faible entropie (e.g. Cranshaw *et al.,* 2010).

La figure 196 ci-dessous présente la part cumulée de néo-*AOI* en fonction de leur entropie. Si 30 % des zones ne sont composées que d'une seule catégorie (entropie à 0), nous pouvons observer une augmentation importante des effectifs pour une entropie à 0,81 et encore plus importante à 0,91. Nous décidons de considérer les lieux comme étant des zones mixtes lorsque leur entropie est supérieure à 0,89, soit avant le saut important. Les zones où l'entropie est inférieure à ce seuil seront qualifiées en fonction de leur catégorie majoritaire (commerces ou sorties).

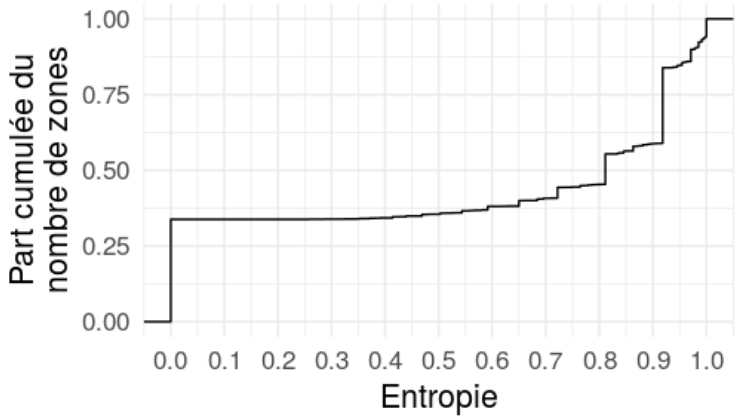


FIGURE 196 Part cumulée du nombre de zones (ou néo-AOI) en fonction de l'entropie

Si nous regardons maintenant le rapport entre le nombre de lieux correspondants à des sorties sur le nombre de lieux de commerces de détail pour chacun des néo-*AOI* et que nous le mettons en relation avec leur entropie respective (figure 197), nous pouvons noter que les zones mixtes correspondent aux néo-*AOI* où l'une des deux catégories est présente moins de 2,2 fois plus que l'autre (rapport entre 1/2.2 et 2.2). Concrètement, si une zone contient moins de 2.2 fois plus de commerces que de lieux de sorties (et inversement), nous la considérons comme

une zone mixte, ce qui nous parait cohérent.

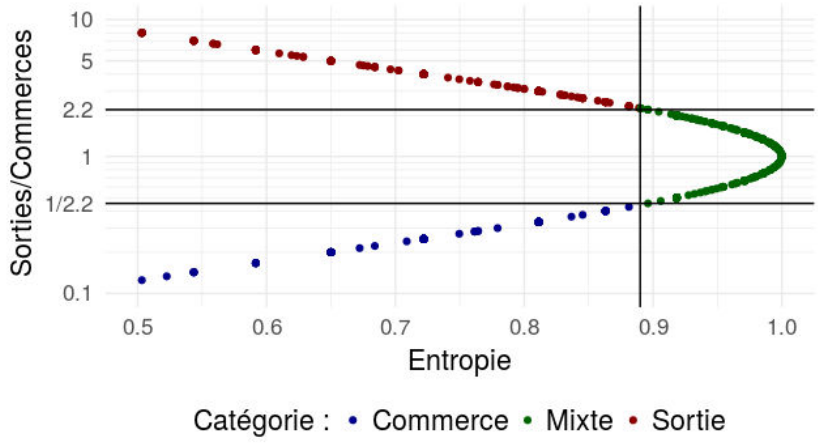


FIGURE 197 Rapport entre le nombre de lieux correspondants à des sorties sur le nombre de lieux de commerces de détail et leur entropie associée.

Densité des néo-AOI

Les catégories pour nos zones commerciales étant définies, nous allons maintenant déterminer le niveau de densité (faible, moyen, fort), en nous basant simplement sur la part cumulée des densités de *POI* dans chacun de ces néo-*AOI* (figure 198). Nous appliquons trois seuils selon les effectifs égaux, où le premier tiers sera qualifié de zone de faible densité, le second de densité moyenne, et le troisième de forte densité.

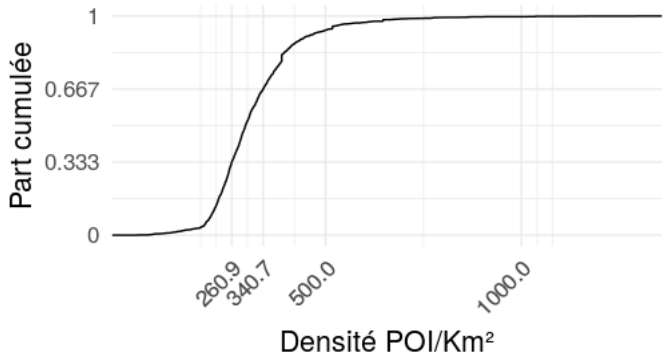


FIGURE 198 Part cumulée du nombre de néo-*AOI* en fonction de leur densité en *POI*.

Cette approche relativement simple à mettre en œuvre permet de faire une double discrétisation : selon les catégories dominantes (commerces de détail, sorties ou mixtes) et selon leur densité de *POI*. Il en résulte la figure 199, où le marché du week-end de Chatuchak ressort bien en zone commerciale dense, comme le quartier au sud de Khao San Road, au bord de la Chao Praya. Le quartier chinois de Yaowarat, tout comme les *malls* du quartier de

Siam Paragon, ainsi que les alentours du *mall* Plaza Pinklao (au nord-ouest de Khao San) sont considérés comme des zones mixtes denses, ce qui fait sens.

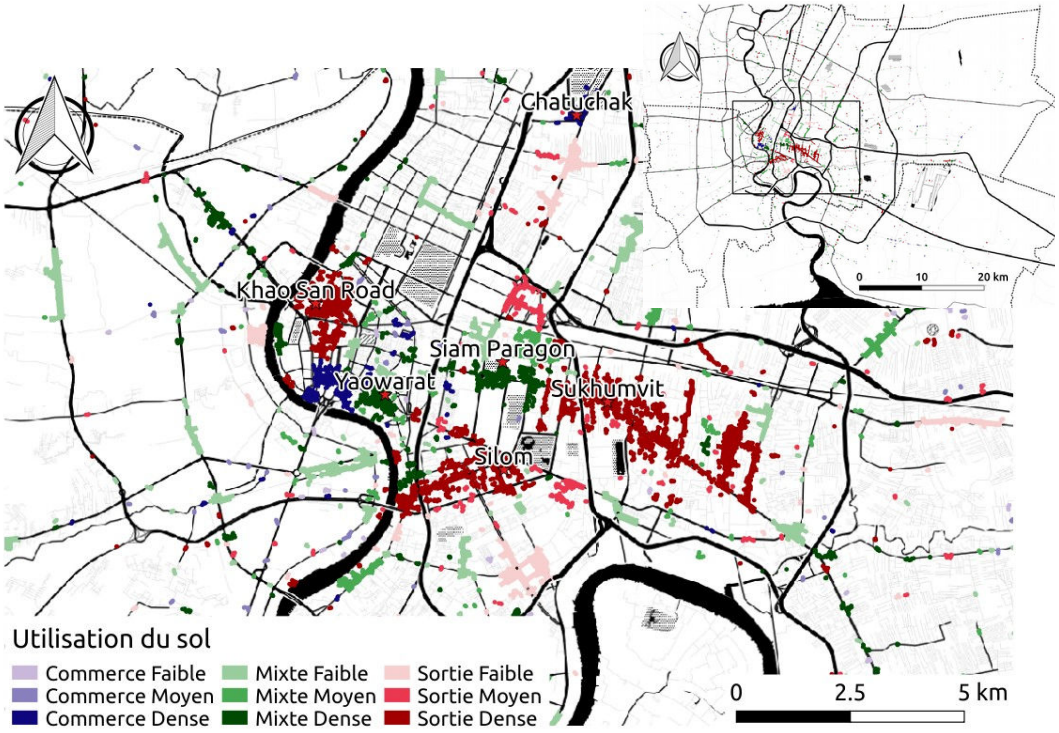


FIGURE 199 Répartition des activités commerciales à Bangkok selon notre deuxième méthode. Sont mobilisés ici les *AOI* et les *POI* de *Google*.

Les quartiers de Silom, Sukkhumvit et Khao San sont bien classés comme des lieux de sorties denses, même si la nature de ces lieux est très différente. Enclave de Backpacker aux guest-houses et bars aux tarifs abordables pour l'un (Khao San road), lieux de sorties très variés mais ou l'on trouve un grand nombre de restaurants et de discothèques pour un autre (Sukkhumvit), ou encore restaurants et hôtels, dont une proportion importante est de haut standing dans le quartier d'affaires animé de Silom[386]. D'autres petites poches à forte densité de *POI* apparaissent également localement, par exemple la RCA Alley au nord de Sukkhumvit, qui bien que ne faisant pas partie des *AOI* est pourtant un haut lieu de sorties et de boites de nuits[387]. Au-delà de ces secteurs aux fortes densités commerciales assez connus du centre, les zones de densités moins importantes se trouvent surtout en périphéries, au cœur de certains quartiers où le long d'axes routiers. Si le marché de Chatuchak est bien considéré comme une zone de commerces denses, d'autres marchés, comme celui de Khlong Toei est quant à lui classé en tant que zone mixte de faible densité, alors qu'il s'agit d'un marché plutôt très dense. La plupart des petits stands n'ont très probablement pas été pris en compte dans la base de

386. http://v vre en tha ande.com/s om quart er affa res an me/11932/
387. http://www.bangkok.com/ratchadap sek/n ght fe.htm

données de *Google*.

Discussion sur les deux méthodes

Sans palinodier la première méthode développée[388], la seconde approche est plus rapide à mettre en œuvre et permet d'obtenir une typologie simple en trois grandes catégories subdivisibles, qu'il s'agisse du type d'activité qui s'y déroule ou de l'importance de l'offre commerciale. Les résultats sont tout aussi cohérents au regard de notre (faible) connaissance du terrain, et plus simple à interpréter. Si la première méthode permet d'affecter une catégorie à une plus grande portion de l'espace Bangkokois, la majorité des zones qualifiées concernent des secteurs où les commerces sont présents en très faible proportion (quelques *POI* par mailles), ce qui n'apporte finalement que peu d'information, car Bangkok est jalonnée de petits commerces de quartiers. La seconde méthode, qui se focalise sur les zones présentant un plus grand attrait potentiel n'est donc pas désuète.

Si nous comparons ces deux typologies d'un point de vue formulation au regard concept d'espace d'activité, il parait plus naturel d'attribuer à un individu synthétique la visite d'une zone « mixte dense » plutôt qu'un secteur où « les lieux commerciaux sont très majoritaires avec quelques lieux de sorties », même si ces dernières zones sont souvent les mêmes dans les deux typologies. Si la première approche développée permet d'obtenir pour chacune des mailles une caractérisation des assemblages commerciaux, elle fait perdre la notion de quartier, maintenu par la deuxième méthode. Car si les résultats sont relativement similaires dans les deux cas en termes de localisation des assemblages des activités, la seconde approche, de par une création en deux temps définition des néo-*AOI* puis caractérisation de ces derniers permet de raisonner en termes de voisinages et de continuum d'activités ce qui permet d'obtenir des zones plus homogènes spatialement. Nous préférerons donc la seconde typologie des activités commerciales pour la suite de nos travaux.

2.4 Finalisation de la couche d'utilisation du sol

Au-delà des aspects commerciaux, nous allons utiliser des informations d'*OSM* et de *Google* pour compléter notre couche d'utilisation du sol.

À partir d'*OSM*, nous allons extraire les routes définies comme importantes, c'est-à-dire ayant les attributs : « primary road », « secondary road », « trunk », « road », « motorway ». Nous récupérons également les différentes lignes de métro, ainsi que les gares et les aéroports. Nous récupérons les polygones d'*OSM* qui se réfèrent à des activités en extérieur ou de loisir (parc, attractions, zoo, sport center) que nous fusionnons avec les polygones de *Google Maps*

388. au coeur de l'article "Discontinuités spatiales, santé et mobilités. Analyses et typologies de *Google* POI et de *tweets* pour caractériser les structures spatiales et les dynamiques d'attractivités de Bangkok (Thaïlande)" qui a reçu le prix du meilleur papier lors de la conférence SAGEO 2017.

correspondant à des parcs, pour définir une catégorie de type « parc / loisir ». Pour les lieux de cultes, nous récupérons simplement tous les polygones d'*OSM* qui se réfèrent à des temples d'une quelconque religion. La couche « Hôpital » est créée en combinant les polygones d'*OSM* définis comme « clinic » ou « hospital », auxquels nous ajoutons les hôpitaux récupérés via *Google Maps*.

Finalement, l'information qui nous pose le plus de problèmes est relative aux lieux d'éducations, et surtout aux écoles. En effet, ce dernier terme est assez ambigu en anglais, car il peut définir une école primaire, un département d'une université, et finalement n'importe quel lieu d'étude (école de musique, de kung-fu, etc.). Les temples bouddhistes présentent parfois des écoles d'apprentissages religieux dans leur enceinte, faut-il les distinguer du reste du complexe ? Nous avons fait ici le choix de regrouper tous les polygones de *Google* et d'*OSM* faisant référence à un lieu d'éducation (« school », « university », « college », « library ») dans une même classe.

Mais nous n'avons pas mis de côté les écoles primaires et les collèges pour autant, du fait de leur importance comme lieu de propagation des épidémies, notamment de dengue (Anderson *et al.*, 2007 ; Yoon *et al.*, 2012). Nous avons pour cela mis en place une méthode assez lourde, en mobilisant d'une part les plus de 2900 *POI* de *Google* définis comme « école », ainsi qu'une couche de bâtiment provenant d'un opérateur privé. S'il est possible d'extraire des bâtiments de *Google Maps* dans une ville comme Delhi, ce n'était pas le cas à Bangkok, ces derniers n'apparaissant pas sur les cartes.

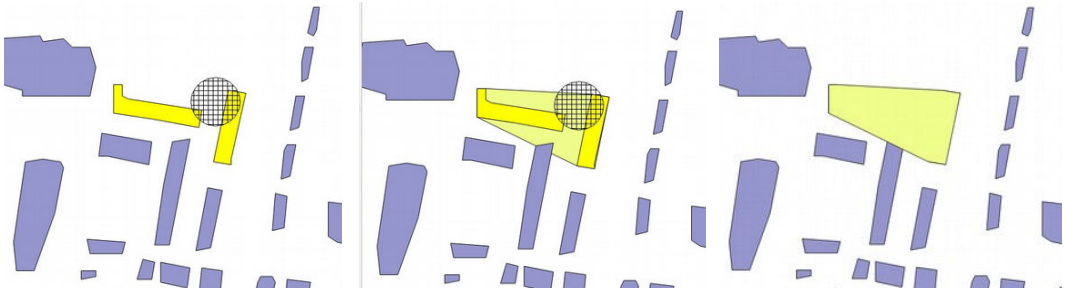


FIGURE 200 Principe adopté pour définir une école à partir de la base de données *Here*. Un *buffer* de 10 m est appliqué à un *POI Google* correspondant à une école (cercle hachuré) et nous intersectons ensuite avec les bâtiments de la base (gauche). Nous créons ensuite l'enveloppe convexe de ces bâtiments (centre) et nous obtenons notre zone que nous définissons comme une école (jaune pâle, droite).

Nous avons donc absorbé la base de données de *Here*, qui fournit par exemple *Yahoo Map*[389], selon une méthode très proche de celle utilisée pour collecter les *AOI* de *Google*, pour obtenir plus d'1,4 million de bâtiments. Nous avons ensuite sélectionné les bâtiments situés à moins de 10 m d'un *POI* de type école. Si un ou plusieurs bâtiments sont associés à un

389. https://maps.yahoo.com/b/

même *POI*, nous définissons l'emprise spatiale de l'école comme étant l'enveloppe convexe de ces bâtiments, comme présenté dans la figure 200 ci-dessus.

Le choix d'une recherche dans un rayon de 10 m est défini après plusieurs observations empiriques, car des rayons plus grands impliquent, notamment dans des quartiers très denses, qu'un grand nombre de bâtiments sera sélectionné, ce qui induit une enveloppe convexe de taille excessivement importante pour un établissement scolaire. Enfin, si aucun bâtiment ne recoupe le *POI* de l'école, nous appliquons arbitrairement une zone tampon de 50 m de rayon. Nous pourrons toujours élargir par la suite chacune de ces zones correspondant à des écoles, soit en appliquant un *buffer* d'un rayon donné, soit en agrégeant directement les *POI* dans des mailles d'une dimension donnée. Il ne nous reste plus qu'à agréger les différentes couches que nous venons de créer pour obtenir notre utilisation du sol à Bangkok.

Création de la couche d'utilisation du sol

La création de la couche d'utilisation du sol va se faire selon une logique de priorité d'un type d'utilisation sur d'autres. Nous découpons dans un premier temps tous les éléments qui intersectent un moyen de communication majeure (routes et voies ferrées). Nous considérons que les écoles qui croisent un lieu d'éducation ou un lieu de culte comme faisant déjà partie de ladite entité, et nous les supprimons de la base. Nous partons ensuite du principe que les lieux d'éducations (sauf écoles), de cultes, hôpitaux, gare et aéroport sont prioritaires par rapport à la couche des commerces. Autrement dit, cela signifie que si une de ces entités intersecte par exemple une zone de sortie moyenne, la zone de chevauchement prendra l'attribut de la première couche. Nous compilons ensuite l'ensemble pour obtenir notre couche d'utilisation du sol à Bangkok (figure 201), que nous utiliserons pour la suite de nos travaux.

Nous pouvons déjà noter que cette carte est loin d'être exhaustive et qu'il existe énormément de secteurs où nous n'avons que peu d'information. Néanmoins, nous avons fait notre mieux pour obtenir la carte la plus complète possible, à partir de données de nature distinctes (ponctuelle ou surfacique) à partir de sources différentes (*Google*, *OSM*, et dans une moindre mesure *Here*). Nous avons pu définir des zones où les brassages de population sont potentiellement importants, comme les gares et aéroports, mais aussi différentes zones commerciales que nous avons pu hiérarchiser en termes de densité. Cela dit, Bangkok est jalonné de petits commerces, que cela soit au rez-de-chaussée de certaines maisons ou immeubles (les « *shophouses* »), en passant par les marchés épisodiques, ou des vendeurs de rue s'installent à des moments donnés de la journée, surtout en début de soirée. Nous n'avons pas pu récupérer l'ensemble de ces informations, même si certaines zones commerciales de carte peuvent contenir certains de ces lieux.

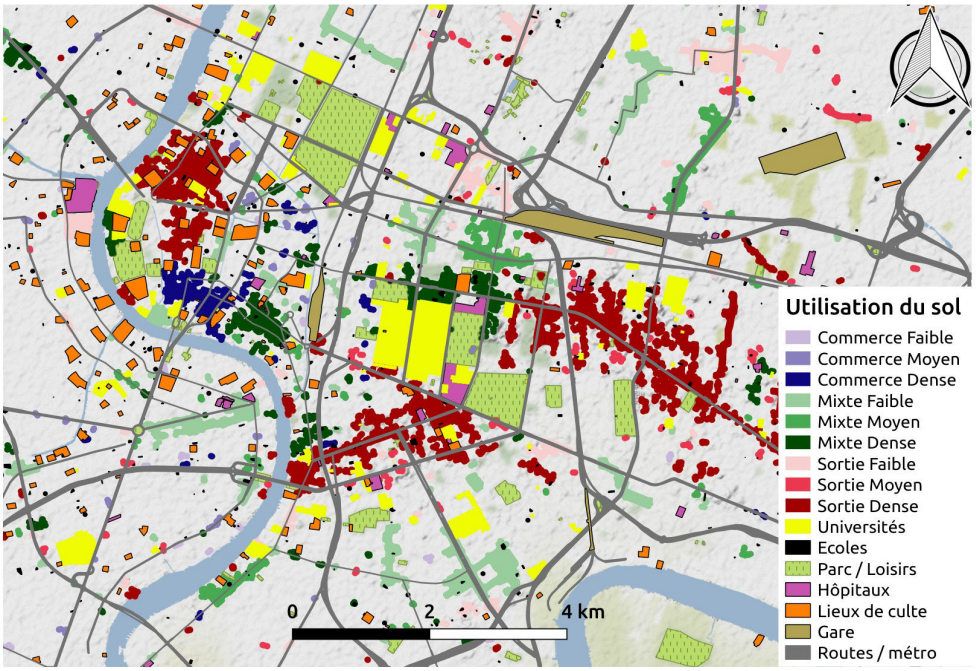


FIGURE 201 Utilisation du sol dans le centre de Bangkok.

Nous avons probablement spatialisé la plupart des universités et lycées, ainsi qu'un grand nombre d'écoles. Ces lieux sont particulièrement sensibles en termes de transmission d'épidémies, puisqu'essentiellement fréquentés par des jeunes, plus susceptibles d'être infectés par la dengue, notamment les moins de 25 ans en ce qui concerne la Thaïlande (Limkittikul *et al.*, 2014). Les principaux parcs, que nous aurions également pu définir par le calcul d'un indice de végétation d'après des images satellites, peuvent également être très fréquentés les matins et le soir, horaires où les moustiques *aedes* sont les plus actifs. Certes, nous n'avons pas pris en compte les différents lieux touristiques, mais nous estimons de manière très réductrice que le tourisme à Bangkok est surtout accès sur la visite de temples, de parcs et la fréquentation de certaines zones commerciales. La localisation des principaux hôpitaux et cliniques pourrait également avoir un intérêt dans le cadre d'une modélisation des mobilités, où un agent malade pourrait se rendre dans un hôpital pour se soigner, et potentiellement contaminer d'autres individus. Nous avons ainsi la plupart des types de lieux qui jouent un rôle important dans la transmission de la dengue (Wen *et al.*, 2015).

Cohérence des résultats

Les *check-in* de *Facebook* sont associés à une localisation et une catégorie. Il est donc tout à fait possible d'intersecter ces données avec notre couche d'utilisation pour en évaluer la cohérence. La figure 202 présente pour chaque catégorie simplifiée des lieux de la base de données de *Facebook*, le nombre de *check-in* selon nos catégories de notre couche d'utilisation

du sol. Quasiment toutes les catégories de *Facebook* croisent notre couche « Transport », ce qui est certainement dû à l'utilisation d'une zone tampon sur la couche de polyligne d'*OSM* et à la présence d'un grand nombre de lieux aux bords des routes principales. Plus de 3 millions de *check-in* sont associés à une catégorie « Shopping & Retail », soit des lieux de commerces. La plus grande partie d'entre eux sont situés dans des zones commerciales mixtes, et une plus faible proportion dans des lieux de sortie, ce qui n'est pas incompatible. Les restaurants, bars, hébergements (Accomodation), ou cafés enregistrent respectivement 2, 1,5, 0,88 *et* 0,82 millions de *check-in*, principalement dans des zones de sorties ou mixtes, ce qui fait tout à fait sens. Comme pressenti, les lieux considérés comme touristiques sont principalement associés à des lieux de cultes, de sorties, des zones commerciales mixtes, mais surtout à des parcs. Les universités et les écoles sont aussi bien classées comme des lieux d'éducation, tout comme les lieux de santé sont principalement associés à la classe hôpital.

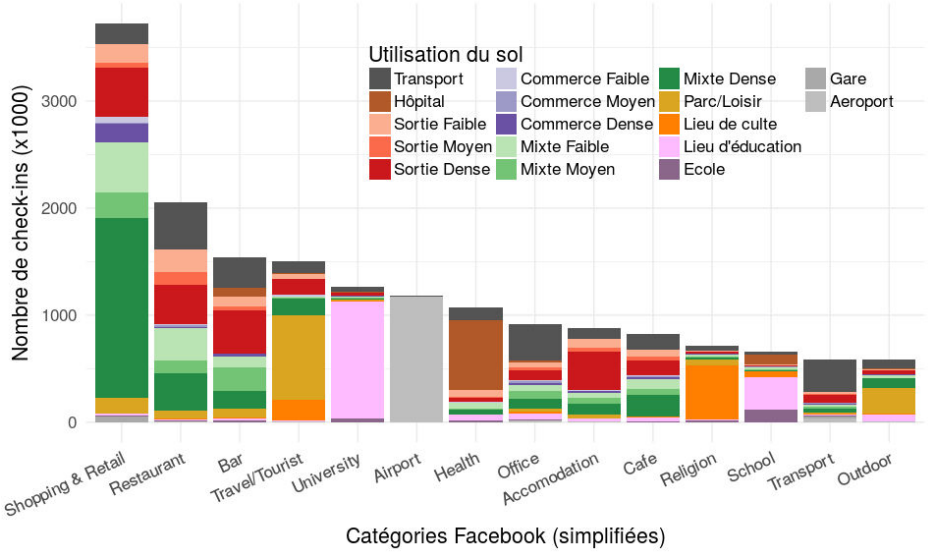


FIGURE 202 Nombre de *check-in Facebook* par catégorie selon la classe de l'utilisation du sol.

Le croisement de ces différentes informations tend donc à valider la cohérence de notre couche d'utilisation du sol. L'avantage d'une telle approche par rapport à l'utilisation directe des activités correspondants aux *check-in* est que ces derniers ne donnent qu'une information ponctuelle, alors que notre couche d'utilisation est surfacique. Ainsi, l'intersection de cette couche avec des données spatio-temporelles de toutes natures devrait donc nous fournir directement des indications sur les temporalités de ces différentes activités.

3 Fréquentations temporelles

Avant d'effectuer cette simple opération, nous allons au préalable regarder les différents profils de fréquentations temporels de 20 types d'activités issus de nos données *Facebook*, en comptant simplement le nombre moyen *check-in* enregistrés par tranche horaire (figure 203).

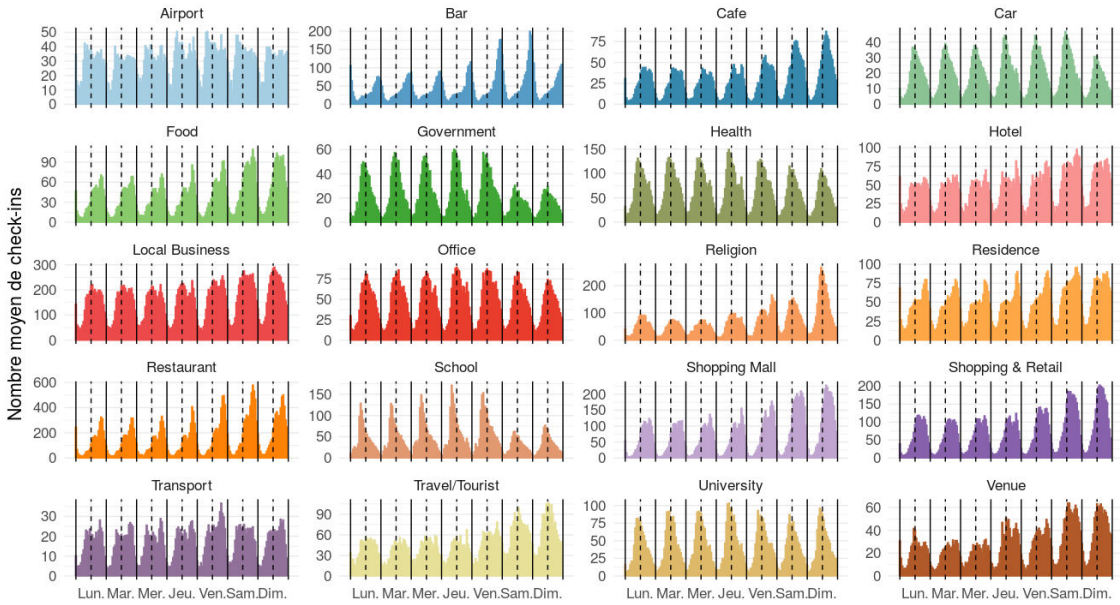


FIGURE 203 Nombre de *check-in* par tranche horaire sur une semaine moyenne pour les 20 catégories les plus représentées sur *Facebook* à Bangkok

Les lieux correspondants à des restaurants ou associés à de la nourriture (food) enregistrent en moyenne plus de personnes le soir, et d'autant plus les vendredis, samedis et dimanches. Les bars ont une activité surtout nocturne et plus marquée les vendredis et samedis soir que les autres jours de la semaine. Les écoles présentent un pic de fréquentation le matin des jours de semaines, vers 8-9h. Leurs profils sont d'ailleurs assez proches de lieux associés à l'administration (*government*). Les universités et les lieux de travail (*office*) ont un profil relativement similaire, c'est-à-dire un maximum de fréquentation entre 11 h et midi et un nombre moyen de *check-in* assez constant sur tous les jours de la semaine. Les lieux associés aux transports présentent les jours de semaine un premier pic le matin et un second plus marqué en fin d'après midi, ce qui peut être relatif aux heures d'embauche et de débauche, mais aussi aux embouteillages. Beaucoup d'activités présentent des profils où la fréquentation augmente énormément les week-ends par rapport aux jours de semaines. Il s'agit par exemple des cafés, des commerces divers (*shopping and retail*), des lieux touristiques, des galeries marchandes (*malls*), ou d'événements divers (*venue*). À noter la présence d'un pic le vendredi après midi dans les lieux associés à des activités religieuses et d'un pic encore plus marqué le dimanche matin.

Ainsi, les différents profils présentés précédemment sont globalement très intuitifs, donc plutôt crédibles, avec cependant des niveaux de *check-in* étonnamment élevés les week-ends dans des lieux de travail (*office*) et dans les universités. Mais ces dernières abritent des campus où un grand nombre d'étudiants est susceptible d'y vivre, ce qui peut expliquer ce phénomène. En revanche, pour ce qui est des lieux de travail, il faut garder à l'esprit qu'il s'agit de données agrégées et qu'un grand nombre de personnes peuvent travailler le samedi et le dimanche à Bangkok, ce qui est assez commun en Asie. De plus, se rendre à son lieu de travail des jours habituellement non-ouvrés peut susciter chez un individu la volonté de le signaler sur les réseaux sociaux. Enfin, il est aussi possible que le type soit mal renseigné. Par exemple le lieu de type « *office* » qui a enregistré le plus de *check-in* sur la période (38 153) était le centre de don du sang, ouvert les samedi et dimanche. Néanmoins, presque 24 000 lieux sont de types « office », soit 20 % de la base et représentent plus de 8 % (plus de 3 millions) de *check-in*. Nous pouvons toutefois estimer que les erreurs peuvent être lissées au regard de ces volumes.

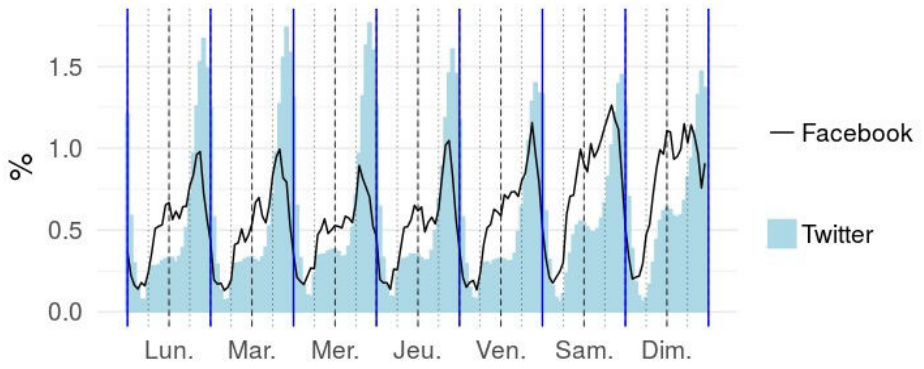


FIGURE 204 Profils horaires des lieux définis comme le domicile, sur *Facebook* et *Twitter*

La figure 203 montre aussi la présence d'une catégorie de type « Residence », qui correspond aux lieux de catégorie « Home ». La figure 204 ci-dessus présente les *check-in* enregistrés dans de telles catégories, mis au regard du nombre de *tweets* envoyés dans le domicile estimé des utilisateurs par tranche horaire (chapitre 6). Certaines similitudes apparaissent, comme une activité plus importante le soir et les après-midi des *week-ends*. Mais ces *check-in* correspondent à moins de 0,45 % des traces enregistrées et représentent moins de 1 % des lieux de la base de *Facebook*, d'où les irrégularités du profil. En revanche, avec plus de 38 000 lieux (utilisateurs) et 32 % des *tweets* envoyés, le profil de l'activité au domicile des utilisateurs de *Twitter* est très régulier. Pour rappel, notre algorithme de détection du domicile à partir des données *Twitter* ne posait pas d'hypothèses sur les jours de la semaine fréquentés, seulement sur l'activité nocturne dans certaines zones de la ville. Ces profils plutôt crédibles tendent encore à valider notre méthode présentée dans le chapitre 6.

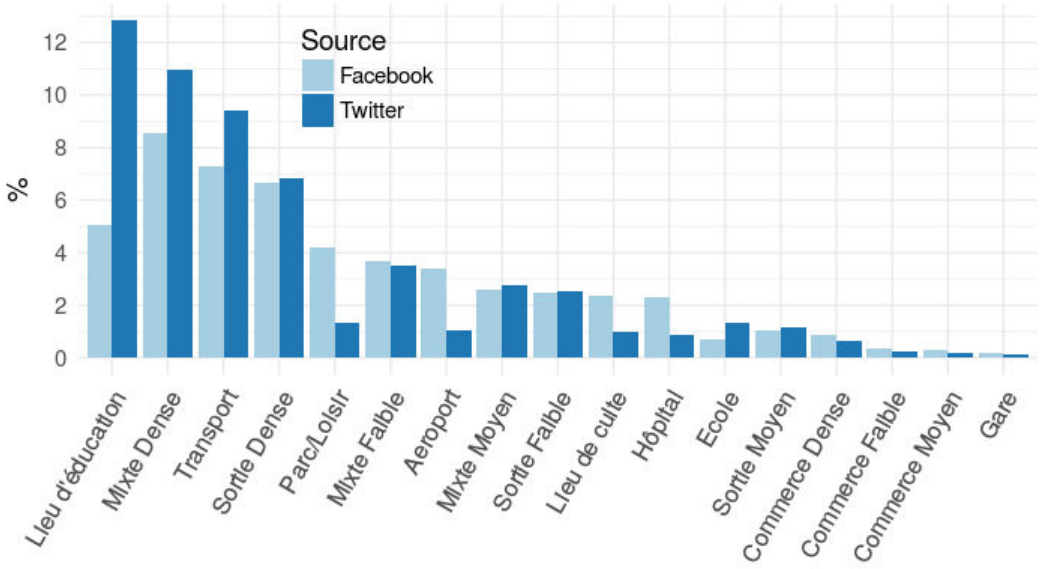


FIGURE 205 Comparaison des parts des traces numériques enregistrées dans les différentes catégories de l'utilisation du sol, selon le réseau social (*Facebook* ou *Twitter*).

Une rapide analyse des pourcentages de traces numériques laissées dans les différentes catégories de notre utilisation du sol selon les sources de nos données (*Facebook* ou *Twitter*, figure 205 ci-dessus) montre d'une part que plus de *tweets* sont émis depuis un lieu d'éducation (école comprise) que de *check-in*, ce qui confirme la part importante de personnes plutôt jeunes dans l'échantillon *Twitter*. Concernant *Facebook*, plus de *check-in* sont enregistrés dans des lieux qui se rapportent à des parcs (qui peuvent aussi être des lieux touristiques) et des aéroports, ce qui souligne l'importance de la contribution des personnes de passage à Bangkok, notamment des touristes, dans la base de données que nous avons collectée. Sinon, la répartition des traces numériques est assez équivalente dans les autres catégories, mis à part dans les hôpitaux où en proportion plus de *check-in* sont effectués que *tweets* sont émis. Toutefois, il convient de noter qu'environ 50 % des *tweets* et des *check-in* sont réalisés hors de notre couche d'utilisation du sol. Nous pouvons aussi observer que les zones de type « commerces » enregistrent finalement peu de traces numériques. À noter aussi que les zones de faibles densités de *POI* (mixte, commerce ou sortie faible) enregistrent globalement plus de *tweets* et de *check-in* que celles appartenant aux densités moyennes. Nous pouvons expliquer ceci d'une part par la potentielle absence de *POI* dans les zones de forte densité de commerces, comme c'est notamment le cas pour le marché de Khlong Toei. D'autre part, notre classification ne prend en compte que la quantité de services (POI) et non leur attractivité, c'est-à-dire qu'une zone pourrait avoir beaucoup de commerces, sans pour autant drainer un grand nombre de personnes. Des enquêtes de terrain dans ces différentes zones pourraient permettre d'apporter plus d'explications. Nous pourrons envisager par la suite soit de regrouper ces deux catégories en une seule, soit de définir leur niveau de "densité" (faible, moyen, élevé) non pas par le nombre de *POI*, mais par l'importance

des traces numériques enregistrées dans ces lieux.

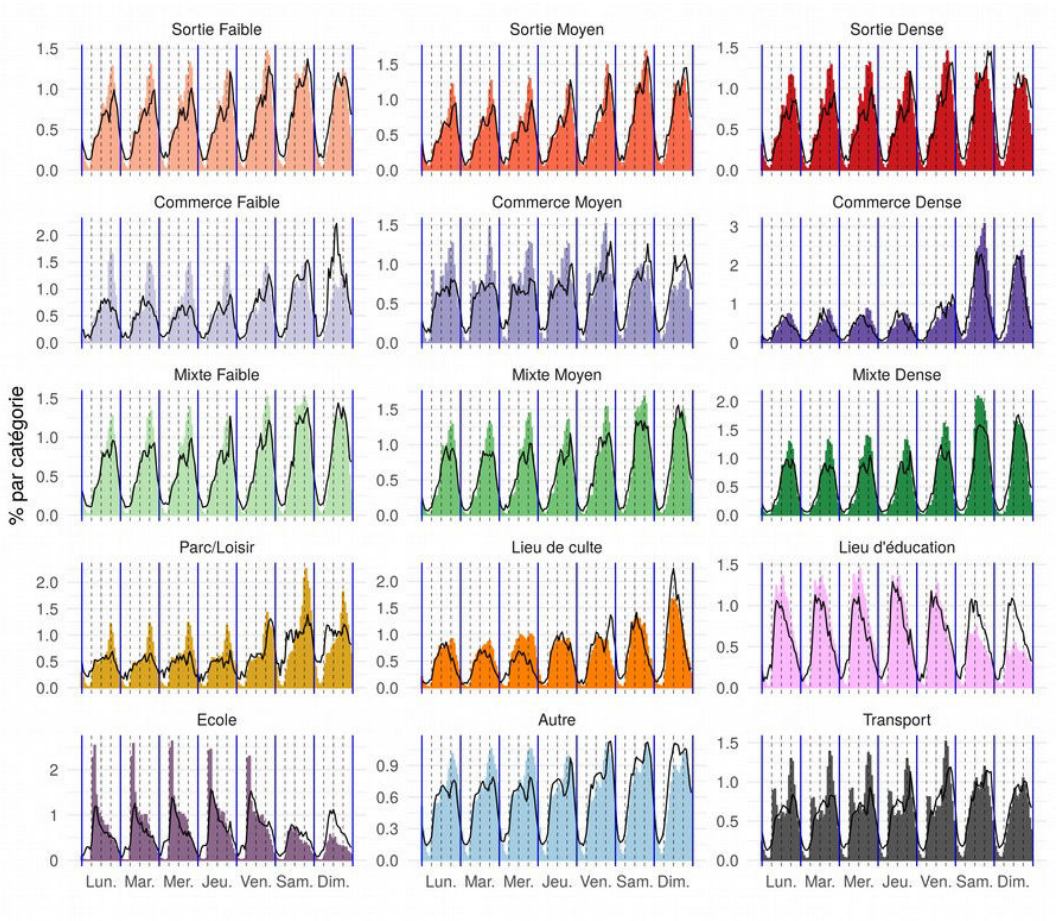


FIGURE 206 Répartition des traces numériques pour chacune des catégories de l'utilisation du sol par tranche horaire sur une semaine type. Les histogrammes correspondent aux données *Twitter*, tandis que les traits à *Facebook*.

La comparaison de la part horaire des traces numériques dans chaque catégorie de l'utilisation du sol (figure 206, ci-dessus) montre un bon accord entre données *Twitter* (histogramme coloré) et *Facebook* (trait noir), avec des tendances générales globalement similaires, même si les profils issus de *Twitter* sont plus réguliers que ceux provenant de *Facebook*. Ceci peut s'expliquer par des biais de localisation et lié à la correspondance partielle entre les catégories des lieux de *Facebook* et celles de notre utilisation du sol (section précédente). Pour toutes les catégories, une plus grande part de *check-in* est effectuée la nuit, entre minuit et 6 h, ce qui peut provenir du fait que les personnes peuvent signaler leur présence dans un lieu de donnée sur *Facebook* sans pour autant être réellement dans ce lieu précis.

Sinon, comme attendu, les lieux de sorties présentent un pic de fréquentation très marqué le soir (plus important sur les données *Twitter*), et un second moins franc le midi dans les deux jeux de données. Ces lieux sont aussi plus fréquentés les vendredis et samedis soir que les autres

jours. Si les zones de commerces et les zones mixtes de faible densité enregistrent un pic de *tweet* très marqué vers 18 h les jours de semaine, c'est moins le cas avec les *check-in.* Les zones de commerces denses enregistrent largement plus de *tweets* et de *check-in* les week-ends, ce qui peut s'expliquer par la contribution probablement importante du marché de Chatuchak, ouvert que les samedis et dimanches et membre de cette catégorie. Les *tweets* envoyés depuis des routes (Transport) présentent deux pics, l'un le matin et l'autre plus marqué le soir, correspondant parfaitement aux heures de pointe, comme nous le verrons dans le chapitre suivant.

Une plus grande activité sur les réseaux sociaux est observée dans les parcs les après-midi et les week-ends dans les deux jeux, avec des pics plus marqués d'après *Twitter*. Les écoles et les universités présentent des profils temporels assez proches, avec une plus grande proportion de *tweets* que de *check-in* enregistrée le matin dans les écoles, et une activité nettement moins importante les week-ends dans les universités est observée sur le réseau social *Twitter*.

Les différents profils de fréquentation horaires des activités présentés ci-dessous sont donc plutôt crédibles et confirment des intuitions, qu'il s'agisse d'une plus grande par de personne fréquentant les zones mixtes denses (typiquement des *malls* ou des grands marchés) les soirs et les week-ends, les routes aux heures de pointe et les écoles le matin des jours de semaine. Ces tendances de fréquentations sont très similaires entre les deux jeux de données, les nuances se faisant sur l'amplitude de certains pics de visites dans certains types de lieux. Il est donc envisageable de considérer que ces profils permettent de quantifier les probabilités pour une personne d'effectuer une activité à un moment donné de la journée.

Néanmoins, il est aussi possible que les aspects d'intentionnalité dans la création de traces numériques géolocalisées soient les mêmes sur les deux réseaux sociaux. En d'autres termes, les utilisateurs de *Twitter* ou *Facebook* auraient la même propension à indiquer à leurs contacts leur présence dans certains types de lieux (*malls* chics, bar "sympas", etc.), pour des raisons de désirabilité sociale et/ou de projection de soi dans l'espace social numérique. Mais nous pouvons considérer que cet aspect jouerait plus sur les volumes de données enregistrés globalement dans une zone que sur les temporalités en elles-mêmes.

Pour revenir à une logique de modélisation à base d'agents, ces volumes horaires par types de lieux peuvent servir dans la génération d'agendas et à l'attribution d'une activité en contraignant la visite d'une zone donnée par ces tendances de visite agrégées.

Comme nous venons de le voir, les différents secteurs de la ville présentent une signature temporelle propre à l'activité qui s'y déroule. Il parait donc tout à fait envisageable d'employer l'approche réciproque, c'est-à-dire définir l'utilisation du sol selon les signatures spatio-temporelles des différentes zones.

4 Définir l'utilisation du sol en fonction les profils temporels des traces numériques

D'un point de vue purement formel, le fait d'agréger par tranche horaire des traces numériques géolocalisées à une grille revient à créer une matrice à *n* bandes, compatible avec les traitements classiques en télédétection. Chaque maille de la grille peut être donc perçue comme un pixel possédant des valeurs dans chacunes des *n* bandes, correspondant ici aux différentes plages horaires. Partant de ce constat, certaines études ont utilisé un algorithme de classification non-supervisée, le *Kmeans*, sur des données issues de la téléphonie mobile (Soto et Frías-Martínez, 2011) ou sur des *check-in* de Foursquare (Zhan *et al.,* 2014), afin de caractériser l'utilisation du sol. (Lenormand *et al.,* 2015c) ont employé une approche un peu différente, puisqu'ils ont créé un réseau de mailles où la distance entre chaque cellule est définie par le coefficient de corrélation des profils spatio-temporel des cellules. Ce qui revient astucieusement à définir un niveau de proximité entre chaque cellule selon les formes de leurs profils. Ils ont ensuite appliqué un algorithme de détection de communautés (voir chapitre 10) pour regrouper les différentes cellules entre elles et obtenir une partition de l'espace en différentes catégories d'activités.

Nous allons ici simplement appliquer un *Kmeans* sur chacune des mailles de 180 m auxquelles est agrégé le nombre de *tweets* enregistrés par tranche horaire. Nous définissons un nombre de classes assez élevé (9), quitte à effectuer des regroupements manuels par la suite. La figure 207 présente les profils moyens de chacun de ces groupes. Les classes 1, 5 et 8 ont des profils assez similaires, à savoir une activité moyenne en journée (plus importante les jours de week-end) et une activité nettement plus importante en soirée. Intuitivement, ou en se basant sur les profils montrés précédemment, il peut donc s'agir de lieux de domiciles ou alors de lieux de sorties. En plus de ces profils, une distinction peut aussi être faite selon les volumes de données enregistrés, les zones correspondant à la classe 5 présentent plus de traces numériques que les cellules de la classe 8, elle-même plus fréquentée que les mailles de la classe 1.

Les classes 2, 4 et 7 enregistrent plus de *tweets* en semaine que le week-end, et notamment le matin. Il devrait donc s'agir de lieux où se déroule une activité principale, et notamment des lieux d'éducations. Les classes 3 et 9 présentent un grand nombre de traces numériques, plus importantes le week-end qu'en semaine, et surtout réparties en journée et principalement l'après-midi et au début de soirée. Il devrait donc s'agir de zones commerciales assez importantes. Moins de *tweets* sont enregistrés dans la classe 6, mais ils sont relativement bien équilibrés tous les jours de la semaine. On note un petit pic vers midi et un pic plus important le soir, pouvant correspondre à la fréquentation de restaurants ou de commerces.

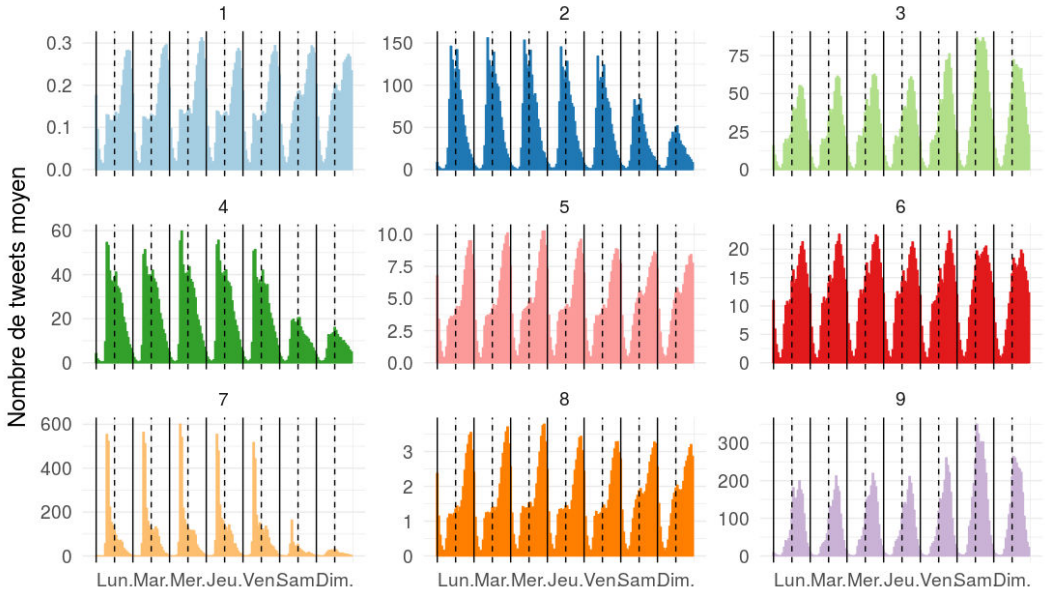


FIGURE 207 Nombre de *tweets* moyens enregistrés par tranches horaires pour les 9 classes du *kmeans*.

Afin de compléter cette analyse et d'avoir un point de comparaison autre que les profils spatio-temporels définis dans la section précédente, nous allons maintenant intersecter chacune des mailles à la couche d'utilisation du sol créé précédemment et apprécier la part de *tweets* émise depuis chacune des catégories (figure 208, gauche). À noter qu'une maille peut appartenir ici à plusieurs catégories d'utilisation du sol. Par exemple, près des 3/4 des *tweets* du *cluster* 1 sont émis depuis la classe « Autre » de notre utilisation du sol. Si l'on trouve des lieux d'éducation dans tous les groupes, ces derniers sont toutefois largement majoritaires dans les clusters 2, 4 et 7 (si l'on ne prend pas en compte les catégories de type « Autre »). Les mailles 1,5 et 8 sont principalement composées de zones où l'utilisation du sol n'a pu être définie par nos bases de données géographiques (type « Autre »), et peuvent donc, comme pressenties, être rapprochées à des zones d'habitations. Les classes 5 et 8 présentent toutefois une part assez importante de zones commerciales mixtes et de sorties et les plus grandes proportions de lieux de sorties et de commerces sont bel et bien enregistrées dans les classes 3, 6 et 9.

Avant de cartographier ces résultats, il convient de jeter un œil sur les effectifs de chacune des classes (figure 208, droite). Une première remarque est que le nombre de mailles par cluster suit un très large spectre. Ainsi la classe 1 représente 47 284 cellules, contre seulement 3 pour la classe 7. Une autre distinction peut se faire sur le nombre de *tweets* moyens enregistrés par mailles, de moins d'une centaine pour la classe 1 à plus de 15 000 pour la classe 10.

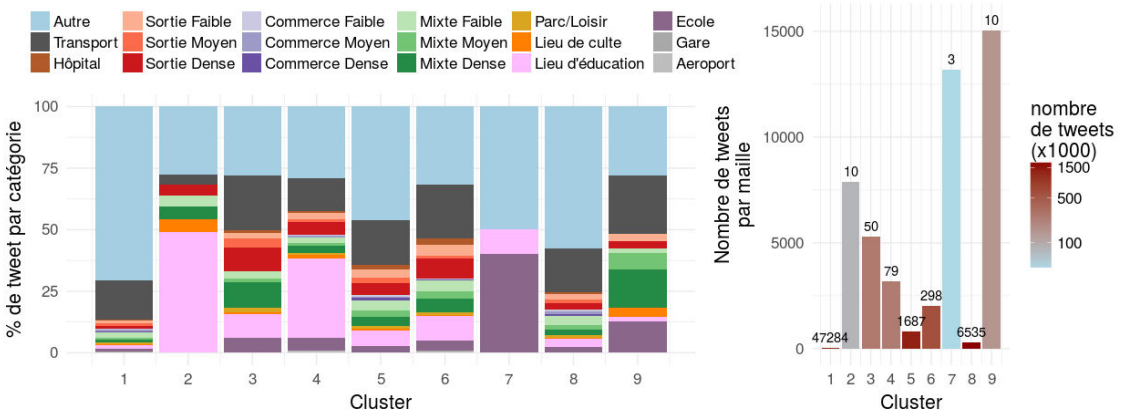


FIGURE 208 Répartition des parts de *tweets* émis depuis une classe de la couche d'utilisation du sol par cluster (gauche) et nombre de tweets par maille par cluster du *kmeans* (droite).

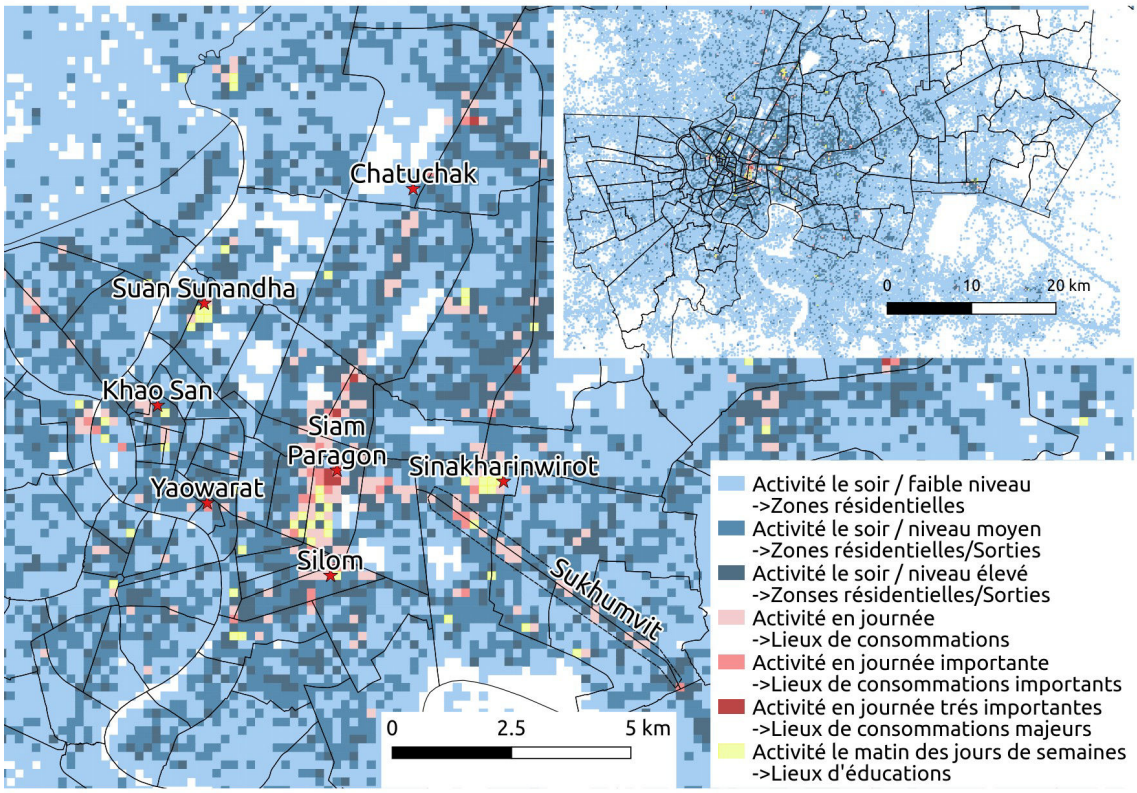


FIGURE 209 Carte de l'utilisation du sol à Bangkok, d'après un *Kmeans* appliqué sur les profils temporels des *tweets* enregistrés dans des mailles de 180 m

Ces écarts dans les effectifs sont bien visibles sur la carte 209 ci-dessus. La majorité des secteurs de la ville sont constitués de zones résidentielles, les plus denses se trouvant proches des grands axes de communications. La figure 210 ci-dessous met en relation la population par maille selon les classes associées à des zones résidentielles (1, 8 et 5), sous forme d'une fonction

de densité (intégrale égale à 1). La classe 1 correspond bien aux zones les moins densément peuplées de la ville, tandis que les classes 8 et 5 à des zones plus peuplées.

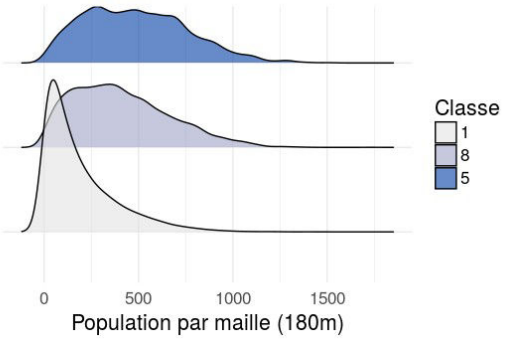


FIGURE 210 Répartition des populations dans les clusters associés à des zones résidentielles (1, 5 et 8) selon une fonction de densité (intégrale égale à 1).

Les campus universitaires apparaissent bien sur la carte 209, qu'il s'agisse des universités de Suan Sunandha Rajabat, Sinakharinwirot Prasanmit ou encore Chulalongkorn, au sud de Siam Paragon. Si les quartiers de Silom, Siam Paragon, Sukhumvit ou Khao San et dans une moindre mesure Yaowarat sont bien classés comme des lieux de consommations, ce n'est pas le cas du marché de Chatuchak. La classification par nuée dynamique n'a pas saisi la forte activité des jours de week-end.

Globalement, cette méthode permet de détecter les grands ensembles structurels de la ville (zones de consommation, d'éducations et résidentielles). Mais leur emprise spatiale est nettement plus faible comparé à la couche d'utilisation du sol présentée précédemment (figure 201). Nous n'avons pas pu détecter les zones commerciales plus locales et de moindre importance par exemple. Il conviendrait cependant d'étudier l'influence de la taille de la maille et autres effets de *MAUP* (Openshaw et Taylor, 1979), en utilisant par exemple la méthode développée par Louvet *et al.* (2016), qui proposent d'évaluer la stabilité des résultats à différentes échelles. Mais il est clair que des cellules de plus grandes tailles entraîneraient un lissage de l'information et potentiellement une moins bonne discrétisation des éléments.

Il est aussi envisageable d'appliquer une méthodologie plus subtile, comme celle employée par (Lenormand *et al.,* 2015c), ou encore utiliser des outils propres à la télédétection à partir d'images hyper-spectrales. Cela pourrait passer par une sélection des plages horaires de manière à garder les informations les plus discriminantes (Martinez-Uso *et al.,* 2007) comme les pics d'activités le soir, en semaine, etc. et tester diverses méthodes de classification propre à cette discipline (pour une revue des méthodes en imagerie hyper-spectrale voir par exemple Mountrakis *et al.,* 2011 ou Plaza *et al.,* 2009). L'application de classifications non-dirigées est aussi une piste de recherche, en prenant quelques signatures temporelles dans différents lieux de

références et en appliquant certains algorithmes, comme le *Spectral Angle Mapper*[390], il serait possible d'extrapoler l'utilisation du sol à l'ensemble de la ville.

Si nous pouvons envisager de réaliser sans trop de contraintes ce genre d'étude, cette approche qui vise à caractériser l'utilisation du sol n'est pertinente que dans des zones géographiques présentant une bonne disponibilité de traces numériques, qu'il s'agisse de *tweets* géolocalisés ou de données téléphoniques, dans un contexte d'absence d'informations sur l'utilisation du sol. La mobilisation de données de téléphonie mobile dans l'étude de grandes métropoles en développement où les bases de données géographiques sont lacunaires, comme à Delhi par exemple, pourrait permettre de mieux caractériser l'utilisation du sol. En revanche à Bangkok, même si les volumes des *tweets* géolocalisés sont très importants, la couche d'utilisation du sol que nous avons obtenue est suffisamment précise et détaillée, ce qui réduit l'intérêt d'une telle approche : autant utiliser les bases de données géographiques disponibles.

390. Cet algorithme se focalise plus sur la détection des pics et la comparaison des angles entre des profils que sur les niveaux amplitudes.

Synthèse

Création une couche d'utilisation du sol

À partir de différentes bases de données (*Google Maps* et *OpenStreetMap*), nous avons créé et comparé deux couches d'utilisations du sol cohérentes au regard de la connaissance du terrain et compatible entre elles. Des zones fonctionnelles ont été mises en évidence, où des discontinuités entre les types d'activités sont visibles dans le grand centre, tandis que la périphérie est plus fragmentée en termes de services.

Nous avons néanmoins sélectionné la couche d'utilisation du sol qui présente la typologie la plus intuitive et respectueuse des voisinages. Ceci permettra d'inférer une activité à une trace numérique géolocalisée, afin de compléter les espaces d'activités.

Temporalité des activités & profils de fréquentation

Nous avons ensuite observé les temporalités de réalisations de ces activités en inférant les données issues des *check-in* de *Facebook* ou des *tweets* géolocalisés. Les profils temporels sont relativement similaires entre ces deux sources de données, même si les données de *Twitter* sont plus à même de rendre compte des activités au domicile, du fait de leur nature individu-centrée.

Mais le corollaire est aussi vrai, c'est-à-dire qu'il est possible d'inférer une activité potentiellement réalisée dans une zone donnée en se basant sur les profils de fréquentation agrégés par plage horaire.

Ces profils de fréquentation par type d'activité forment une connaissance importante dans le cadre de la simulation à base d'agents, pouvant contraindre un agent à effectuer des activités à des moments précis de la journée.

Chapitre X:
Les mobilités à Bangkok : Variations sur le thème des données et des méthodes

Le chapitre précédent s'est principalement focalisé sur la définition de l'utilisation du sol à Bangkok, et par l'intermède d'un grand volume de données géolocalisées, sur la temporalité des réalisations de différentes activités. Si nous prenons une métaphore picturale ou musicale, il s'agit en quelque sorte de la couleur des différents secteurs de la ville, dont l'intensité et la profondeur varient au cours du temps.

Nous allons maintenant nous intéresser à la manière dont les mobilités à Bangkok se ressentent sur le rythme de la ville. Nous étudierons comment le choix des données et les méthodes employées entraînent, toujours avec une métaphore musicale, des arrangements et orchestrations qui divergent plus où moins fortement selon les objets étudiés.

Il s'agira ici de rendre compte dans un premier temps des potentiels de déplacements en étudiant le système de transport de Bangkok. Puis nous analyserons l'influence des méthodes et des données sur la mesure de la pulsation urbaine et les interactions entre les différents secteurs de la ville.

Plus prosaïquement, l'objectif de ce chapitre est, après une description des conditions de circulations, d'étudier s'il est possible de combiner des données *Twitter* et *Facebook* dans l'analyse des mobilités globales à Bangkok, et d'observer les éventuelles affinités entre les différents quartiers.

1 Le rythme de la ville

1.1 Se déplacer dans une ville congestionnée

Bangkok est considérée comme la deuxième ville la plus congestionnée au monde, derrière Mexico, d'après une étude réalisée par l'entreprise Tomtom [391]. Ceci entraîne des temps de trajets aller-retour moyens relativement longs, allant de 54 minutes pour effectuer une activité sportive, à 1h42 pour se rendre à son travail (figure 211)- soit une augmentation de 11 minutes depuis 1999 (Gakenheimer, 1999).

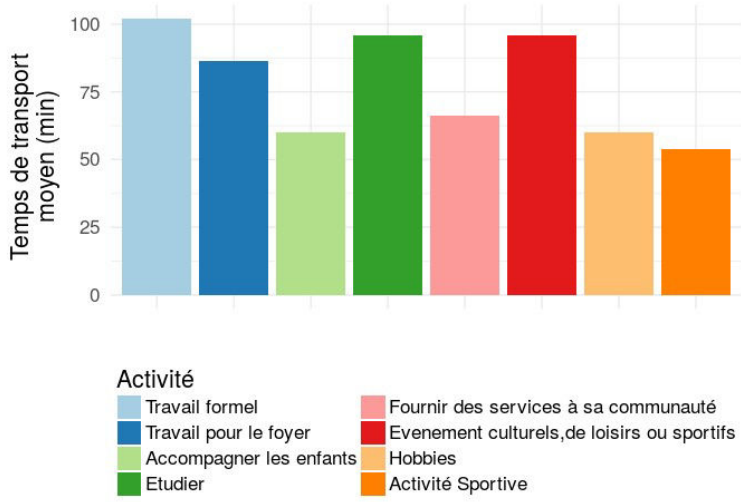


FIGURE 211 Temps de transport moyens (aller-retour) pour se rendre sur le lieu d'une activité (NSO, 2009)

Les conditions de circulation et les modes de transports utilisés ont un impact sur les mobilités des individus et nous présenterons dans cette section les transports en commun à Bangkok, ainsi qu'une analyse plus détaillée des conditions du trafic routier et des temps de transports selon la localisation dans la ville et les modes de transports.

1.1.1 Les transports en commun à Bangkok : le fleuve, l'avenue et le rail

De son passé de ville à caractère fluviale (Pichard-Bertaux, 2011; Shinawatra, 2012), Bangkok garde encore un réseau de transport par bateau relativement développé, qu'il s'agisse de lignes régulières qui empruntent certains khlongs (figure 212, bas), ou de navettes permettant de traverser le fleuve. Mais les transports publics ou privés se sont progressivement détournés de la Chao Praya ou des différents canaux et s'effectuent principalement sur les voies terrestres (Hanaoka, 2007). Néanmoins les navettes fluviales express ne sont pas soumises aux bouchons routiers lors des heures de fortes affluences, leur conférant un attrait certain (*ibid.*).

391. https://www.tomtom.com/en_gb/*Traffic*index/ Les temps de transports sont en moyenne supérieure à 61 % face à une situation où la circulation est fluide.

FIGURE 212 Les modes de transports à Bangkok. Sur l'image du haut nous pouvons voir un tuk-tuk au premier plan, ainsi que quelques taxis (en rose et en vert) ainsi qu'un bus de la ville au second plan. Sur la photo du milieu, nous pouvons voir des petites camionnettes, faisant office de minibus qui permettent de relier des zones de courtes distances (˜moins de 5 km). La dernière image montre un bateau-bus qui circule sur un *khlong*.

Plus de 77 000 taxis, 9 300 tuk-tuk (figure 212, haut) et près de 100 000 moto-taxis

(pratiques pour circuler dans les *soi*) étaient immatriculés dans la ville en 2017 (DLT, 2018). Le réseau de bus est également très développé, avec 14 907 bus circulant sur 458 lignes (BMTA, 2016). Parmi ces bus, 2774 sont directement gérés par la Bangkok Mass Transit Authority, (BMTA) et 800 d'entre eux sont gratuits (emprunté annuellement par 232 500 passagers uniques). 866,879 tickets étaient vendus quotidiennement en 2016 par la BMTA. S'ajoutent également 3621 bus privés et 8512 minibus, pour des trajets plus locaux (figure 212, centre) (*ibid.*). Il a été estimé que 37 % des trajets quotidiens étaient effectués en bus en 2007 (World Bank, 2007).

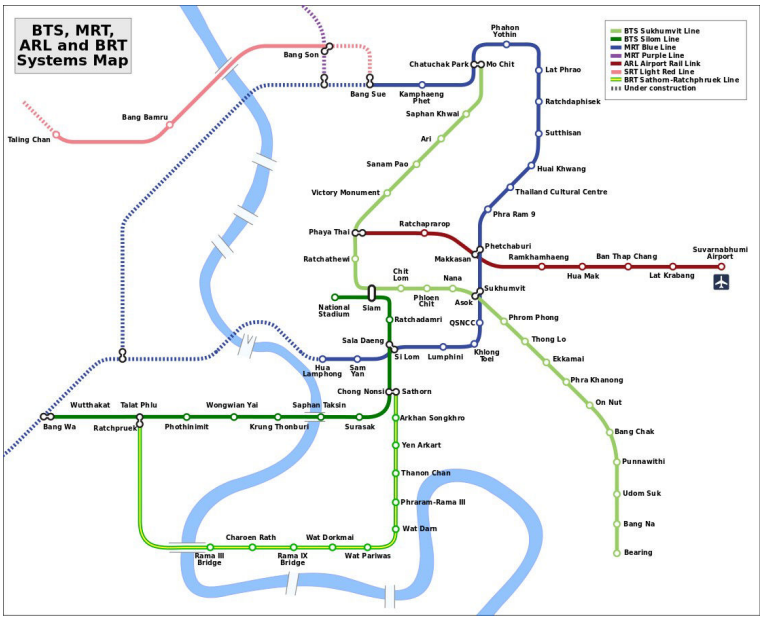


FIGURE 213 Le réseau de métro à Bangkok. Source : Globe-trotter own work pour Wikipedia.

Transit System (operator)	Number of Stations	Length (km)	Opening Year	Average Weekday Ridership
Bangkok Transit System (BTS)	34	30.95	1999	650,000
Mass Rapid Transit Authority (BMCL)	18	21	2004	240,000
Airport Rail Link (SRTET)	8	28.6	2009	47,000
Bangkok Bus Rapid Transit (KT)	12	15	2010	20,000

FIGURE 214 Nombre moyen de passagers quotidiens par ligne de métro (ou équivalent) à Bangkok. Source : Chalermpong et Ratanawaraha, (2015) et BMA.

La première ligne de métro de Bangkok a été inaugurée en 1999. Il s'agit d'un métro aérien, le BTS (Bangkok Transit System), aussi appelé "Sky Train", aujourd'hui composé de deux lignes, qui se rejoignent à la station Siam (figure 213 ci-dessus). Un métro sous-terrain, le MRT (Mass Rapid Transit) a quant à lui ouvert ces portes en 2004, et une ligne

desservant l'aéroport est rentrée en service en 2009. L'année suivante, une ligne de bus avec une voie réservée (Bus Rapid Transit) a été inaugurée. Au total, ces différents réseaux accueillent quotidiennement 957 000 personnes en moyenne (figure 212, et Chalermpong et Ratanawaraha, (2015)).

FIGURE 215 Quelques photos du métro aérien (ou *sky train*), avec en bas a gauche la station Siam, nœud du réseau, vers 16 h, et à droite la station Surasak, moins remplie, vers 17 h.

Bien qu'étant récents, propres, qu'ils permettent d'éviter les embouteillages et qu'ils véhiculent une image de modernité (Chalermpong et Ratanawaraha, 2015; Richardson et Jensen, 2008), ces métros ne sont que peu utilisés au regard de la taille de la population de Bangkok. Une première explication serait leur coût, un peu trop élevé pour les personnes les plus pauvres (Punpuing et Ross, 2001; Ratanawaraha and Chalermpong, 2016; Richardson et Jensen, 2008). Un autre aspect, rarement pointé, serait plus de l'ordre de l'accessibilité comme le montre la figure 216.a ci-dessous. Cette figure a été réalisée en comptant la population de Bangkok selon la distance à une station de métro. Il ressort que moins de 10 % de la population réside à moins de 500 m d'une station, 20 % à moins d'un kilomètre, et la moitié des habitants de Bangkok habitent à plus de 3,5 km d'une de ces stations. A contrario, le réseau de bus est nettement plus dense puisque près des trois quarts de la population de Bangkok résident à moins d'un kilomètre d'une station de bus (hors minibus) (figure 216.b).

Ainsi, ce réseau de métro moderne souffre surtout d'un problème d'accessibilité, contraignant les gens à adopter un comportement multi-modal, ou alors à se tourner vers des transports individuels.

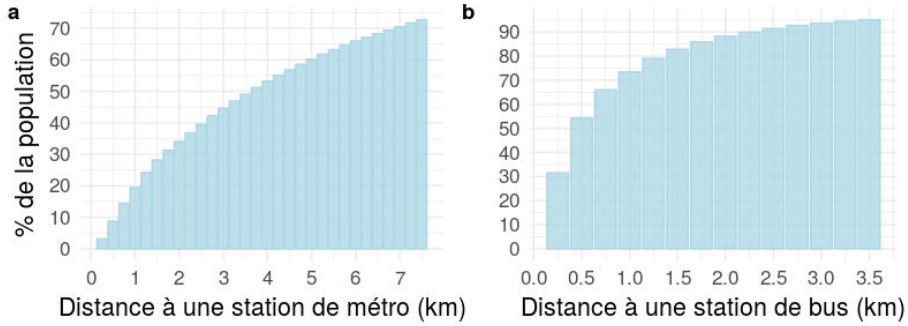


FIGURE 216 Part de la population de Bangkok selon la distance à une station de métro (a) ou de bus (b).

1.1.2 Sur la route

« Je transpire, pire je suis en transe. Les transports bloquent, ça klaxonne... Ça fait deux heures que je n'avance qu'à petit feu » Sérigne M'Baye Gueye, 2000

FIGURE 217 Photo d'un embouteillage dans le centre de Bangkok, vers 16h30. La municipalité signale le niveau de congestion des autres voies.

De ce fait, 46 % des déplacements à Bangkok sont effectués par des modes de transports privés (voitures ou motos) (World Bank, 2007). En décembre 2017, 4 242 556 voitures, 3 521 127 motos et 1 322 841 vans et pick-up étaient enregistrés à Bangkok, soit plus de 9,78 millions de véhicules (DLT, 2018). Ce chiffre a doublé depuis l'an 2000, malgré l'inauguration des différentes lignes de métro ou de voies de bus réservées.

Les millions de véhicules en circulation ont bien évidemment un impact sur la qualité du trafic routier, malgré les nombreux axes de communications à 2 × 4 voies qui jalonnent la ville. Comme le faisait remarquer Supatn, (2011) la ville est très souvent congestionnée entre 6 h et 9 h puis entre 16 h et 19 h. Ce constat peut d'ailleurs être fait par n'importe quelle personne coincée quotidiennement dans les embouteillages (figure 217). Il nous paraît ici intéressant d'aller un peu plus loin dans la quantification et la localisation de ces zones de ralentissement, qui font partie du quotidien des Bangkokois et qui peuvent être un indicateur secondaire d'attractivité des différents secteurs de la ville.

Une ville congestionnée

De nombreux sites de services cartographiques donnent le trafic en temps réel, comme *Google Traffic* ou *Bing Traffic* (*Microsoft*). Pour cela, *Google* utilise les positions GPS des utilisateurs du service, en estime la vitesse, et définit un niveau de trafic en fonction des vitesses et des localisations de l'ensemble des utilisateurs[392]. En revanche, très peu d'informations sont disponibles au sujet du fonctionnement de *Bing*[393]. Il est aussi probable que *Bing* et *Google* utilisent dans certains cas des données fournies par les municipalités. Dans tous les cas, il est délicat de définir un niveau de précision de ces informations sans points de comparaison.

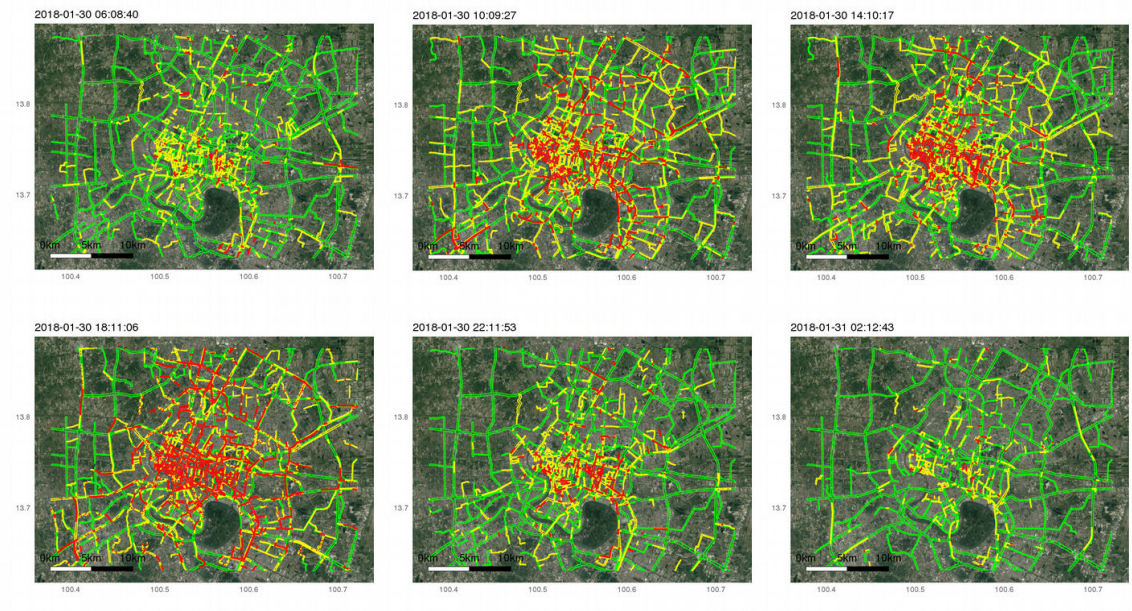


FIGURE 218 Condition du trafic à Bangkok, extrait de *Bing Traffic* pour la journée du 30 janvier 2018, toutes les 4 h entre 6 h et 2 h du matin. Le vert signifie un trafic fluide, le jaune moyennement congestionné et le rouge totalement congestionné.

392. https://goog eb og.b ogspot.fr/2009/08/br ght s de of s tt ng n traffic.htm

393. La filiale *Here* de *Nokia*, elle même filiale de *Microsoft*, fournirait ces données trafic, sans pour autant expliquer précisément leur mode d'acquisition ou de création. https://www.the nqu rer.net/ nqu rer/news/2180022/nok a supp es Traffic m crosoft B ng maps

Pour des raisons de facilité d'accès, nous nous sommes focalisés sur le service *Bing Map* Portal[394], en créant simplement un compte pour obtenir une clé permettant d'accéder aux *API* du service. Nous lançons ensuite des requêtes permettant d'obtenir une carte statique[395] avec le trafic en temps réel et sans les labels, que l'on convertit en raster[396]. Nous extrayons ensuite les pixels de couleurs verte, jaune et rouge, correspondant aux différents niveaux de congestion des routes. Connaissant la localisation du centre l'image et ces dimensions, nous en définissons sa résolution en fonction du niveau de zoom, comme explicité par *Bing*[397], pour obtenir un raster géoréférencé. Nous lançons une requête toutes les 30 min, ce qui permet d'apprécier l'évolution temporelle du trafic à Bangkok (figure 218).

La figure 218 ci-dessus montre l'état des conditions de circulation à Bangkok pour la journée du 30 janvier 2018, avec un intervalle de temps de 4 h. À 6 h du matin, le trafic est plutôt fluide en périphérie et moyen au centre-ville. Puis la circulation s'intensifie et les voies de communication sont complètement bouchées vers 18 h. À 22 h, le trafic est moyennement congestionné au centre-ville, les bouchons se concentrant vers la proche périphérie est de la ville (zone de Sukhimvit). À 2 h du matin, le trafic est enfin plus fluide. Nous avons enregistré ces informations toutes les 30 minutes entre le 22 janvier et le 28 février 2018 (5 semaines), ce qui nous permet par exemple de compter le nombre de pixels rouge, correspondant à du trafic dense par tranche horaire, et d'en tirer le profil moyen sur une semaine (figure 219).

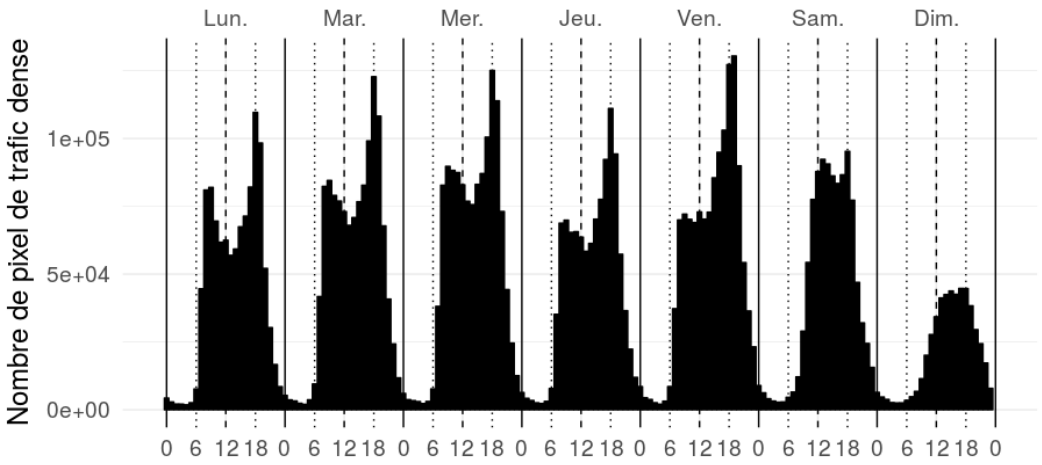


FIGURE 219 Nombre de pixels de trafic dense par tranche horaire sur une semaine type à Bangkok.

Plusieurs éléments ressortent de la figure 219, notamment le fait que les pics de congestions des jours de semaine sont bel et bien observés le matin entre 6 h et 9 h, puis

394. https://www.b ngmapsporta .com/
395. https://msdn.m crosoft.com/en us/ brary/ff701724.aspx
396. Sur R, en utilisant la fonction "download.file" pour télécharger le résultat de la requête, et la librairie "raster" pour convertir l'image.
397. https://msdn.m crosoft.com/en us/ brary/bb259689.aspx

le soir entre 16 h et 19 h, avec un maximum à 18 h. À noter que le pic du soir est nettement plus prononcé que celui du matin. Ceci peut probablement s'expliquer par le fait que les flux matinaux sont plus liés aux navettes domiciles / travail, alors que le soir les personnes pourraient fréquenter d'autres lieux après leur travail avant de rentrer chez elles, rendant plus complexe l'optimisation du trafic par les autorités de la ville. Le samedi est également une journée assez rouge en termes de conditions de circulation, avec un niveau de congestion très important entre 12 h et 18 h. Le dimanche est en général une journée où le trafic est nettement moins dense que les autres jours.

Les images brutes collectées ont une résolution de 18.5 m, ce qui est probablement trop précis pour faire ressortir clairement des zones embouteillées. Nous agrégeons alors par tranche horaire et dans une grille de 500 m le nombre pixels correspondant à du trafic dense. Il en résulte la figure 220, qui présente pour les tendances agrégées de trafic dense à 4 moments de la journée les mardi et dimanche, où l'effet centre-périphérie est particulièrement bien visible. Pour le mardi, alors que le nombre de zones embouteillées est plus important entre 7 h et 9 h qu'entre 12 h et 14 h (figure 220 ci-dessous), les zones congestionnées ressortent plus aux heures du repas, surtout dans le centre-ville, alors qu'elles sont plus diffuses aux heures d'embauches. Aux horaires où les personnes débauchent, entre 17 h et 19 h, quasiment toute la partie est de Bangkok est saturée. La soirée, entre 21 h et 23 h, le trafic est relativement plus fluide, sauf dans quelques axes majeurs situés à proximité lieux de sorties, comme à Silom ou Sukkhumvit. Les zones congestionnées le dimanche sont à peu près les mêmes que les jours de la semaine, avec néanmoins une amplitude nettement plus faible.

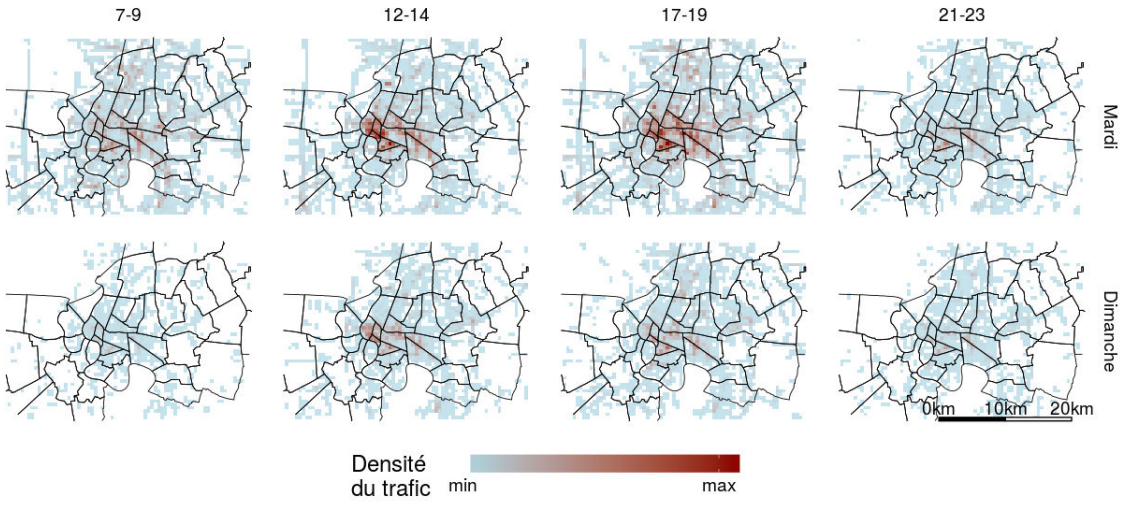


FIGURE 220 Densité du trafic à Bangkok, à 4 moments de la journée, les mardis et jeudis

Ces données trafic donnent ici une information sur les heures de départ et de retour au domicile, et de par les axes fréquentés, fournissent une information indirecte sur les zones

attractives. Et même si un axe congestionné n'implique pas nécessairement qu'il s'agisse d'un lieu de dépôt de passager, nous pouvons ici observer une forme de pulsation urbaine de type centre-périphérie. Il est d'ailleurs assez étonnant, au regard des travaux de Louf et Barthelemy, (2015) sur l'émergence de plusieurs pôles dans les villes en fonction du niveau de congestion du trafic, que le polycentrisme à Bangkok n'y soit qu'embryonnaire (chapitre 2).

Après cette rapide présentation des conditions de circulation dans Bangkok, nous allons maintenant discuter du corollaire, soit le temps passé dans les transports dans la ville.

Temps de transports

Google (décidément) met à disposition l'*API Google Direction*, qui permet de faire des requêtes entre deux lieux. Elle renvoie une distance et un temps de déplacement, selon l'heure de la journée, le mode de transport et un niveau de congestion du trafic[398]. Il est donc envisageable d'interroger l'*API* en prenant un point de départ et un grand nombre de points d'arrivée répartis de manière homogène dans la ville, ce qui permettrait d'obtenir une carte des temps de transports à partir d'un lieu donné. Appliquer cela a autant de points de départ que d'arrivée reviendrait à construire une matrice origine-destination, où les variables sont les durées. Mais comme pour l'*API Places*, le nombre de requêtes quotidiennes est limité (ici à 2500 par clés) ce qui perturbe l'application de la méthode sur un grand nombre de points. Si on choisit par exemple 5000 points dans la ville, ce qui n'est pas extravagant compte tenu de la superficie de Bangkok, le nombre de requêtes nécessaires sera de 25 millions (5000^2).

Nous avons décidé d'utiliser un maillage moins dense, en prenant un point tous les deux kilomètres, en partant du principe qu'une interpolation linéaire permettra d'estimer les données manquantes et qu'il ne s'agit ici que d'un travail exploratoire. À partir d'un code en python, nous avons récupéré les durées des trajets dans Bangkok pour le mercredi 13 décembre 2017 à 7 h et 19 h (heures de pointe), selon un trafic normal ou dense (estimé par *Google*), pour un déplacement en voiture ou en transport en commun.

À partir de chaque point de départ, nous avons appliqué une interpolation linéaire[399] sur l'ensemble des points d'arrivée. Nous avons ensuite utilisé le même algorithme, mais cette fois à partir de chaque nouveau point d'arrivé et leur temps de transports estimés précédemment, sur les points de départs associés. Ceci permet d'affiner la résolution spatiale, selon une grille que l'utilisateur peut définir. Les cartes suivantes (figure 221.a et b) présentent quelque temps de transports pour des cellules d'un kilomètre carré, à partir d'un point de départ défini par un point rouge, selon des conditions de circulation standard.

398. https://deve opers.goog e.com/maps/documentat on/d rect ons/ ntro
399. Sous R, la fonction interp de la librairie "akima" (Akima et Gebhardt, 2016)

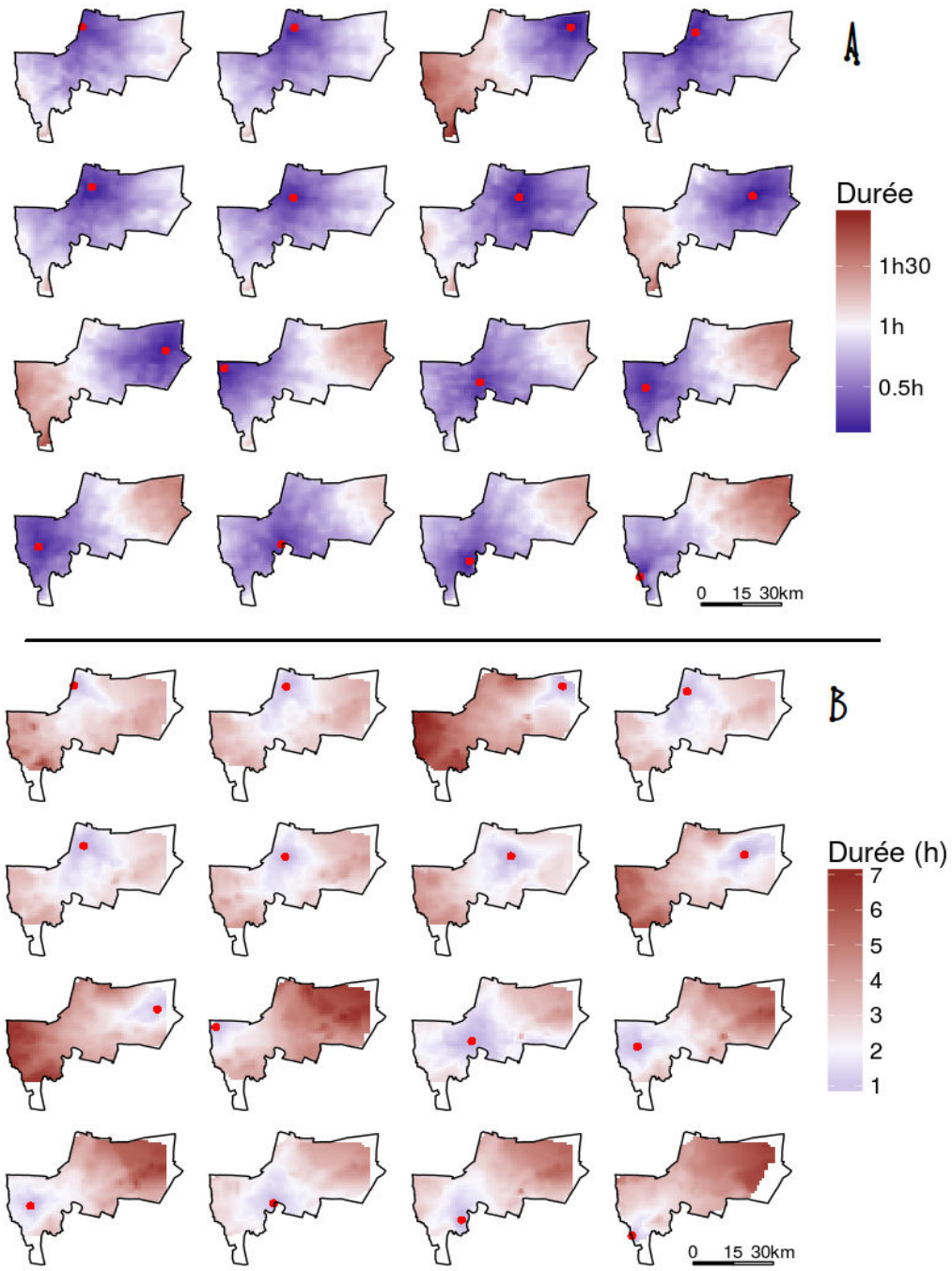


FIGURE 221 Les temps de transports dans Bangkok un mardi à 7h du matin en fonction du point de départ (point rouge) et du mode de transport. La figure du haut (A) correspond à un transport en voiture en situation de trafic normal. La plage de couleur diverge après un temps de transport d'une heure. La figure du bas (B) correspond à l'utilisation de transports en commun et la plage de couleur diverge après un temps de transport de deux heures.

Les temps de déplacements estimés pour un mode de transport en commun sont excessivement longs, surtout si le point de départ est excentré. L'impact de la proximité au

réseau apparaît également sur la figure 221.b où se dessine dans certaines situations un corridor où les temps de transports sont plus courts. Traverser la ville depuis une zone périphérique en transport en commun parait néanmoins inadapté car cela peut prendre plusieurs heures. Ainsi, il n'est pas étonnant que 46 % des déplacements se fassent en voitures, car même si les temps de trajets sont relativement longs, il est possible de traverser la moitié de la ville en moins d'une heure, si le niveau de congestion est standard.

Dans tous les cas, comme le résume (Clément-Charpentier, 2011) :« Les habitants consacrent une partie de leur journée au transport dans des bus surpeuplés ou, selon leurs revenus, dans des voitures climatisées, mais de toute façon les uns et les autres immobilisés dans des bouchons ».

1.2 La pulsation urbaine : des tempos différents selon les données

Sans chercher pour l'instant de lien avec les faits exposés précédemment, nous allons maintenant observer les pulsations urbaines selon des données issues de *Twitter* ou des *check-in* de *Facebook*. Pour limiter certains biais dans le dernier jeu de données, nous supprimons les lieux correspondant à des aéroports ou des lieux touristiques, où le nombre de *check-in* parait surreprésenté (chapitre 6).

1.2.1 Des « hotspots » de traces numériques

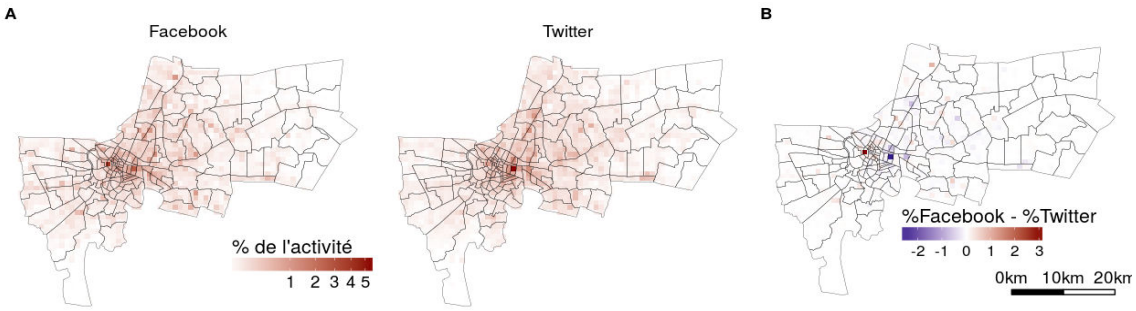


FIGURE 222 Pourcentage de traces numériques laissées dans des mailles d'un kilomètre sur le réseau *Facebook* et *Twitter* (a), et différence entre ces niveaux d'activités (b).

Une première analyse statique est effectuée en regroupant les traces numériques issues de chacun de ces réseaux sociaux dans des mailles d'un kilomètre carré. Afin de pouvoir comparer ces deux jeux, nous raisonnons en pourcentage de traces numériques émises selon les mailles (figure 222.a). À première vue, il semble qu'il y ait une bonne correspondance entre ces deux jeux. La réalisation d'une soustraction permet de ressortir les différences (figure 222.b) et nous pouvons noter qu'elles sont relativement minimes, sauf dans le centre-ville où plus de traces

numériques sont enregistrées dans le centre-ville de Bangkok sur *Facebook* (en rouge), tandis que les *malls* du quartier de Siam sont clairement surreprésentés sur *Twitter* (en bleu foncé).

Nous allons maintenant définir les « hotspots » à partir de ces différentes données en utilisant la méthode Getis-Ord Gi*, qui est un indicateur de l'auto-corrélation spatiale locale (Getis et Ord, 1992). Plus la valeur obtenue est importante dans une maille donnée (un z-score), plus les concentrations sont importantes par rapport à un voisinage. La figure 223 ci-dessous présente le résultat d'une analyse réalisée sur des données agrégées à des mailles de 250 m, avec une portée de recherche de 1 km. Comme une grande partie de la ville présente des valeurs de l'indice élevées (et très significatives), nous avons décidé de ne pas discrétiser. Une fois encore les résultats sont très similaires entre les sources, mis à part que les données de *Facebook* font ressortir deux centres majeurs, le quartier de Siam et le centre historique, alors que ce dernier est moins important dans les données de *Twitter*.

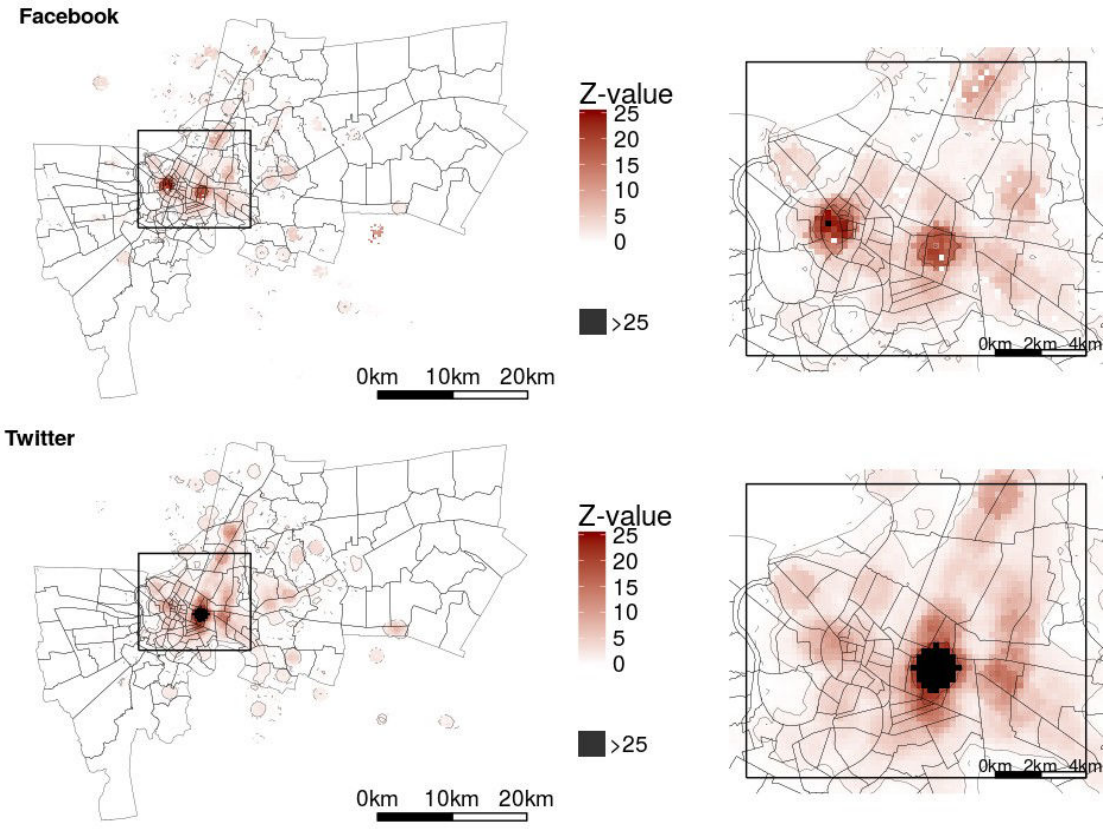


FIGURE 223 Application d'un Getis-Ord Gi* sur les données *Facebook* (haut) et *Twitter* (bas). Les Z-values supérieures à 25 apparaissent en noir.

Ainsi, si à l'échelle de la ville nous pouvons considérer que Bangkok est plutôt monocentrique, malgré quelques pôles périphériques secondaires, le fait d'effectuer un zoom permet de nous rendre compte que le centre est en fait constitué de plusieurs pôles, plus ou

moins marqués selon les données, mais où se regroupent le quartier historique, Siam et Silom, ainsi que le début de Sukkhumvit (au sud-est), Hua Kwaeng à l'est, et Bangkok Noi (au nord-ouest). Apparaît également le grand marché de Chattuchak (au nord-est de l'encadré), pourtant ouvert que le week-end. Ces différents pôles attractifs sont aussi susceptibles d'évoluer au cours du temps (figure 224).

1.2.2 Répartition spatio-temporelle des traces numériques

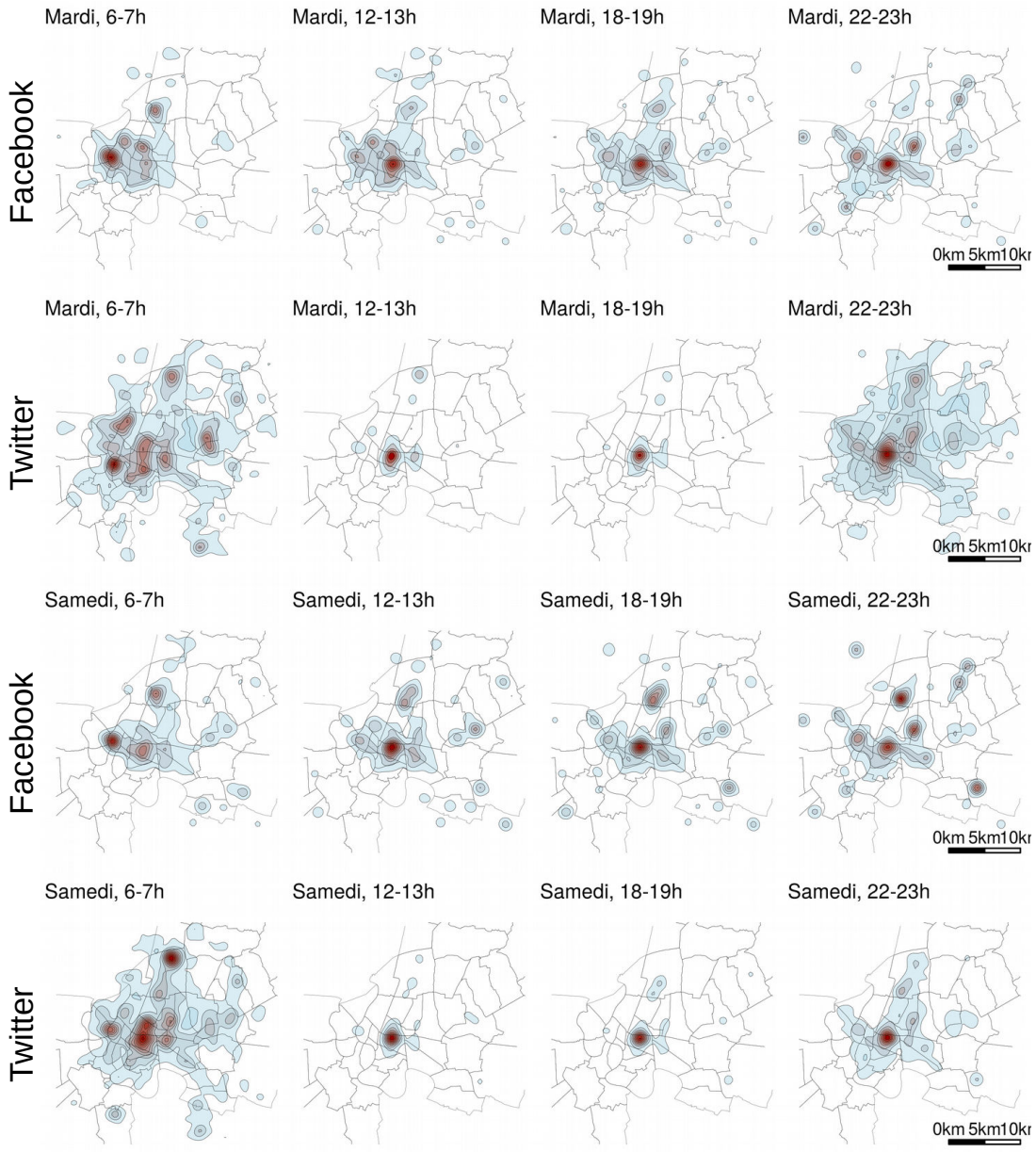


FIGURE 224 Densité des traces numériques géolocalisées (*tweets* et *check-in* à Bangkok à différents moments de la journée (6 h, 12 h, 18 h et 22 h) et différents jours de la semaine (le mardi et le samedi), calculée selon une portée de 2,5 km.

La figure 224 montre la densité du nombre de *check-in* et de *tweets* par utilisateurs à différents moments de la journée (6 h, 12 h, 18 h et 22 h) et différents jours de la semaine (le mardi et le samedi), calculée selon une portée de 2,5 km. Nous pouvons d'abord noter globalement que les zones de fortes densités sont nettement plus concentrées le week-end que la semaine . Alors qu'une réelle dichotomie apparaît entre la répartition des présences des utilisateurs de *Twitter* selon le moment de la journée, cet aspect n'est pas aussi marqué avec les données de *Facebook*. Concernant *Twitter*, nous pouvons observer clairement un effet centre périphérie, où l'activité enregistrée sur le réseau est relativement bien répartie dans l'ensemble de la ville le matin (entre 6 h et 7 h) et le soir (entre 22 et 23 h), avec bien entendu des pôles de plus fortes densités, évoqués précédemment. En revanche, la journée, le centre-ville et surtout le quartier de Siam dominent clairement les autres zones.

Si la distinction n'est pas aussi marquée lorsque l'on regarde les *check-in* de *Facebook*, il n'en demeure pas moins qu'une évolution est perceptible au cours de la journée. En effet, alors que c'est le centre historique et le quartier de Khao San Road que le plus de traces numériques sont enregistrées le matin, le pôle le plus attractif se déplace ensuite vers l'est et Ratchatewi (Siam). Les pôles de Bangkok Noi, Sukhumvit, Dusit et Hua Kwaen apparaissent également, mais sont globalement moins marqués. L'activité à Chattuchak est revanche nettement plus importante le week-end, comme attendu. Le même constat peut aussi se faire avec les données *Twitter*, mais visible dans une moindre mesure du fait de la domination écrasante de la zone de Siam.

1.2.3 Quantifier la dilatation urbaine

La figure 224 permet de localiser les différents pôles attractifs à divers moments de la journée, mais ne permet pas une comparaison quantitative des niveaux de concentration des populations. Il convient dès lors d'utiliser des indicateurs pouvant synthétiser au mieux ces phénomènes.

« Distance Venables »

Une première approche peut être d'utiliser la « distance Venables » (Louail *et al.*, 2015a), inspiré de « l'indice Venables » (Pereira *et al.*, 2013), et défini comme :

$$Dv(t) = \frac{\sum\limits_{i \neq j} s_i(t) s_j(t) d_{ij}}{\sum\limits_{i \neq j} s_i(t) s_j(t)} \tag{27}$$

Soit la somme des interactions entre les zones *j* et *i*, définies comme la multiplication des parts de traces numériques *s* enregistrées en zone *i* et *j* à l'instant *t*, et la distance entre les deux zones considérées, puis divisé par la densité des interactions. Il s'agit en quelque

sorte d'une distance moyenne prenant en compte les niveaux d'attractivité entre chaque zone (Louail *et al.*, 2015). Une distance Venable très faible signifie une forte concentration des traces numériques dans la ville, tandis que l'inverse indique une répartition plus homogène. Cet indice peut ensuite être divisé par la racine carrée de l'aire étudiée, afin de permettre une comparaison entre différentes villes (ibid.).

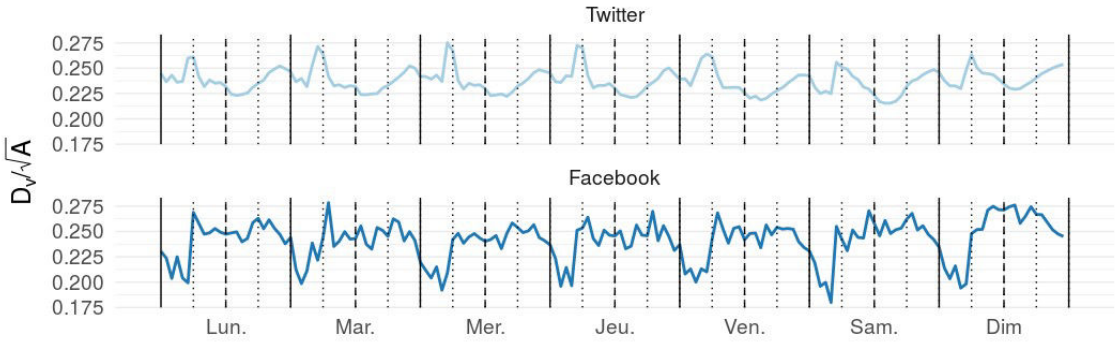


FIGURE 225 Variation de la "distance Venables" par tranches horaires sur une semaine type, selon les données issues de *Facebook* et de *Twitter*.

Nous avons calculé cet indice en définissant dans un carroyage de maille de 250 m la part de *tweets* et de *check-in* enregistrés par tranches horaires (figures 225 et 226). Il ressort dans un premier temps une grande différence entre les profils de *Twitter* et de *Facebook*, sur une semaine (figure 225), ou sur un jour de la semaine moyen (figure 226) , tant sur la forme des courbes que sur leur amplitude. Les distances Venables pour les données *Twitter* sont moins bruitées que pour les données *Facebook*, ce qui peut s'expliquer par la durée de collecte de l'échantillon (environ 1 an et demi pour *Twitter* contre quelques semaines pour *Facebook*) qui entraîne un meilleur lissage spatial. Les indices dans ces deux jeux de données sont assez faibles la nuit, du fait d'une sous utilisation de ces différents services. C'est d'ailleurs entre minuit et 6 h que sont enregistrées les plus faibles valeurs de cet indice pour *Facebook*, peut être parce qu'il y a relativement peu de lieux de la base de données ouverts la nuit, et que les *check-in* sont alors concentrés dans certains quartiers nocturnes. La dispersion est maximale pour les données *Twitter* entre 5 h et 6 h du matin, mais elle ne l'est qu'une heure plus tard pour les données de *Facebook*.

Le profil des données *Twitter* des jours de semaine est assez proche de ceux définit par Louail et al, (2015) dans une dizaine de villes espagnoles, et présente une valeur de dispersion maximale entre 5 h et 6 h du matin, probablement lorsque les premiers utilisateurs envoient des messages sur la plateforme depuis leurs lieux de domiciles, répartis dans la ville. On observe ensuite une tendance à la concentration jusqu'à 8 h, soit lorsque les personnes se rendent à leur travail. Cette tendance est stable jusqu'à midi, puis s'intensifie entre 13 h et 16 h. Finalement,

on note une lente dispersion jusqu'à 23 h, probablement due au fait que les personnes ne rentrent pas systématiquement directement chez elles. Ces profils sont assez proches les jours de week-end, même si plateau entre 8 h et 12 h est soit absent (le samedi) soit moins marqué (le dimanche), ce qui signifie que les utilisateurs de *Twitter* se concentrent plus progressivement dans le temps vers quelques zones de la ville, où le maximum est observé en milieu d'après-midi.

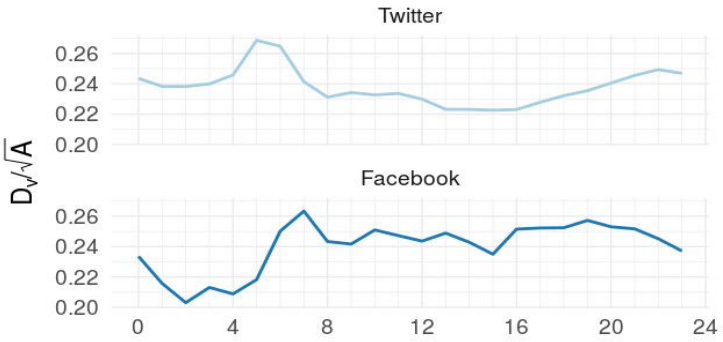


FIGURE 226 variation de la "distance Venables" par tranches horaires sur une journée type, selon les données issues de *Facebook* et de *Twitter*.

Ces distances calculées à partir des de données *Facebook*, malgré leur caractère bruité, présentent un tout autre tableau de Bangkok. Si l'on se focalise sur les données d'un jour de semaine moyen (figure 226), on observe après un pic de dispersion à 7 h, un replat entre 8 et 9 h, puis un effet en dent de scie de faible amplitude, impliquant de légères variations dans les concentrations dans la ville, avec un minimum de dispersion en journée enregistré à 15 h. À partir de 16 h, les niveaux de concentration diminuent jusqu'à 19 h. Comme les données *Facebook* ne fournissent pas d'informations sur les domiciles des personnes (qui pourrait entraîner une dispersion dans la ville après une certaine heure), il n'est pas étonnant de voir une tendance à la concentration des traces numériques jusqu'à 2 h du matin les personnes effectuant des *check-in* probablement dans des lieux de sorties, dont l'emprise spatiale est relativement réduite au regard de la taille de la ville.

Si nous comparons les amplitudes entre les deux jeux, nous pouvons dire qu'en journée, entre 6 h et 20 h, les variations de concentrations sont plus faibles sur *Facebook* que sur *Twitter*, mais que sur l'ensemble des heures, la dichotomie est plus importante entre le jour et la nuit pour les données de *Facebook*.

Emprise spatiale

Une autre manière de synthétiser les temporalités des fréquentations présentées dans la figure 227, est de compter tout simplement le nombre de mailles accueillant un nombre minimum

de personnes par tranche horaire. Cela permet de s'affranchir du poids peut-être exagéré de certains secteurs, comme le centre historique ou le quartier de Siam. Cela revient en quelque sorte à lisser la surreprésentation de certains lieux, où le fait de laisser une trace numérique peut être connoté plus positivement dans la construction de l'identité numérique des individus. Ainsi les endroits à la mode, ou « tendance », qui sont susceptibles d'enregistrer un plus grand ratio entre le nombre de traces numériques et le nombre de visiteurs réels verront leur importance minorée.

La figure 227 montre le pourcentage de maille de 250 m contenant respectivement toutes les traces numériques et plus de 25, 50 % de ces dernières, enregistrées par tranche horaire. Pour ce faire, nous calculons le pourcentage cumulé que représente chaque maille par tranche horaire et ne sont gardées que celles qui dépassent chacun de ces seuils. Cette figure représente donc l'évolution de la surface où sont enregistrées des traces numériques selon des niveaux d'attractivités.

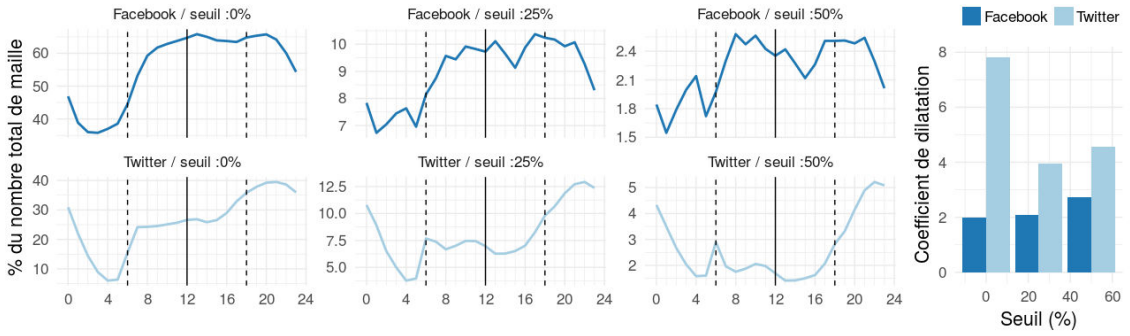


FIGURE 227 Variation de l'emprise spatiale selon les données *Facebook* et *Twitter*, avec différents seuils de concentration (gauche). À droite, figure un coefficient de dilatation selon les différents seuils.

Nous pouvons dès lors observer les évolutions des niveaux de concentration dans la ville. Un pourcentage élevé signifie une répartition des traces numériques sur une plus grande superficie, tandis qu'une valeur plus faible entraîne une plus grande concentration. Le seuil de 0 sert de référence les profils sont d'ailleurs assez proches de ceux sur le nombre de traces numériques géolocalisées enregistrées par tranches horaires sur *Twitter* ou *Facebook* (chapitre 6). Plus le seuil est important, plus on se focalise sur les mailles qui concentrent le plus grand nombre de *check-in* ou de *tweets*, ce qui permet de prendre en compte la répartition des traces numériques dans les zones plus attractives à un moment donné.

À noter qu'avec l'absence de seuil, où toutes les mailles sont considérées, le pourcentage du nombre de mailles où l'on enregistre des *tweets* ou des *check-in* ne dépassent pas respectivement 40 et 70 %. Cela signifie que même lorsque le niveau de dispersion est maximal, en journée pour les données *Facebook* et le soir pour les données *Twitter*, ces zones ne

recouvrent pas l'ensemble de la ville. Il y a donc toujours des zones non-fréquentées, et ces dernières varient au cours du temps. Aussi, le pourcentage minimum de nombre de mailles est plus important pour *Facebook* que pour *Twitter*. Ce qui peut s'expliquer par le fait que les données *Twitter* ne concernent qu'un peu plus de 38 000 personnes, alors que plusieurs dizaines de millions de *check-in* de *Facebook* sont répartis dans plus de 120 000 lieux, offrant potentiellement plus de zones avec au moins une trace numérique.

Pour les données *Facebook*, les moments où les plus fortes concentrations sont enregistrées (nombre de mailles moins important) sont surtout observés la nuit. Pour rappel, si un individu souhaite signaler sa présence dans un lieu sur *Facebook*, il crée un message avec un « Je suis là », ou *check-in,* associé à un lieu dans une base de donnée[400]. Ainsi, les *check-in* de *Facebook* reflètent plus les heures d'ouverture des différents lieux et commerces, d'où la plus grande diffusion spatiale en journée, tandis que la concentration nocturne est probablement symptomatique d'une activité dans des lieux de sorties relativement restreints spatialement. La dispersion dans la ville augmente rapidement entre 5 h et 8 h, puis plus lentement jusqu'à 13 h. On observe ensuite une phase de concentration entre 14 h et 16 h, d'autant plus marquée si l'on ne considère que les mailles les plus attractives (seuil de 25 et 50 %).

Pour ce qui est des données *Twitter*, si nous considérons toutes les mailles, nous pouvons noter une pente très forte entre 5 h et 7 h, où les utilisateurs sont dispersés dans la ville probablement vers leur lieu de domicile. S'ensuit une pente très faible entre 7 h et 14 h, puis un niveau de concentration plus important entre 15 h et 16 h précédent la phase de dispersion de la fin d'après-midi. Des seuils plus élevés (25 % et 50 %) permettent de mieux faire ressortir la répartition des utilisateurs dans les mailles les plus attractives. La première phase de dispersion à 6 h du matin est suivie d'une phase de concentration à 8 h. Le nombre de mailles visitées augmente ensuite jusqu'à 11h-12h, et diminue fortement entre 13 h et 16 h, où la population est principalement concentrée dans le centre-ville.

Le calcul d'un coefficient de dilatation permet de rendre compte de l'importance de la variation de la répartition de la population au cours de la journée. Nous définissons cet indice, de la même manière que Louail *et al.,* (2015) en faisant le ratio entre le nombre de mailles maximum et minimum observées dans chaque situation (figure 227, droite). Ce coefficient est plus important avec les données *Twitter* qu'avec celles de *Facebook*, ceci suggère des pulsations urbaines plus marquées pour le premier jeu de donnée. Nous faisons aussi le constat inverse qu'avec l'analyse de l'amplitude de la « distance venables » vue précédemment. Néanmoins, si l'indice est plus important pour les données *Twitter*, cela peut s'expliquer par le fait que très peu de messages sont enregistrés la nuit entre 2 h et 5 h, et dans un faible nombre de mailles, ce qui induit une concentration très importante des utilisateurs de *Twitter* actifs la nuit. Ceci

400. Si un utilisateur peut également créer un lieu, nous n'avons gardé que ceux fréquentés assez régulièrement (chapitre 6).

ne reflète pas la répartition de la population au domicile, car la plupart des personnes sont inactives sur le réseau social à ces heures considérées.

Les deux jeux de données et les différents indices montrent que la population se disperse le matin, à partir de 6 h pour les données *Twitter* et un peu plus tard avec les données de *Facebook*. On observe également une forte concentration entre 14 h et 16 h. Néanmoins, l'activité sur *Facebook* est répartie de manière plus homogène dans la ville en journée.

La visualisation des pulsations urbaines et le calcul d'indices de dispersion permettent une description de la répartition de la population à différents moments de la journée, et d'apprécier les niveaux d'attractivités de certains secteurs de la ville. Pour revenir à la modélisation des mobilités urbaines, ces indicateurs de concentration temporels peuvent faire office de point de comparaison entre des données observées et simulées pour apprécier la qualité de la modélisation.

Néanmoins, ils ne fournissent pas d'informations sur les interactions et flux entre chacune de ces zones, aspects que nous allons développer dans la section suivante.

2 Les interactions dans la ville : données, méthodes et nuances

Les flux entre les différentes zones, présentés généralement sous forme de matrice d'origine-destination sont à la base de certains modèles épidémiologiques qui prennent en compte les mobilités, qu'il s'agisse d'un modèle en réseau ou de modèle métapopulation (chapitres 3 et 5). Connaissant le déroulement d'une épidémie dans une zone donnée (nombre de vecteur, de personnes infectées, susceptibles ou guéries, etc.), il est alors possible d'estimer dans quelles zones la maladie est susceptible de se propager.

En raisonnant d'un point de vue individu-centré et en mobilisant les données de *Twitter*, il devient possible de quantifier les flux entre différents secteurs de la ville. Mais les *tweets* d'un utilisateur sont de natures épisodiques, c'est-à-dire que la succession temporelle des messages géolocalisés prise *stricto sensu* ne traduit pas nécessairement pas la trajectoire individuelle réelle d'un individu (chapitres 6, 8 et 11). Il faut donc trouver des méthodes permettant de définir et quantifier interactions entre les différents secteurs.

2.1 Différentes méthodes pour concevoir ces matrices

2.1.1 Approche par les lieux fréquentés

Nous présenterons ici différentes méthodes permettant de concevoir des matrices origine-destination (OD) à partir des messages individuels, selon une approche centrée sur les lieux

fréquentés. Nous utiliserons de manière interchangeable les termes de matrices OD et de réseaux, considérant que les lignes et les colonnes d'une matrice OD correspondent aux nœuds d'un réseau, et les valeurs à l'importance du lien entre deux nœuds. Le choix de l'unité surfacique de base influence les résultats, créant un effet de *MAUP* (Modifiable Area Unit Problem, (Openshaw et Taylor, 1979)). Étant donné que nous avons des informations précises sur la population au niveau du sous-district (*khwaeng*) nous avons décidé de travailler dans un premier temps à cette échelle.

La première méthode est entièrement inspirée d'un papier de Poorthuis (2018), qui cherchait à définir les différents quartiers et voisinages à New-york à partir de *tweets* géolocalisés. Pour cela, il a reconstitué pour chaque utilisateur un graphe individuel, selon une approche qu'il nomme « lieux par lieux » [401], qu'il a ensuite agrégé.

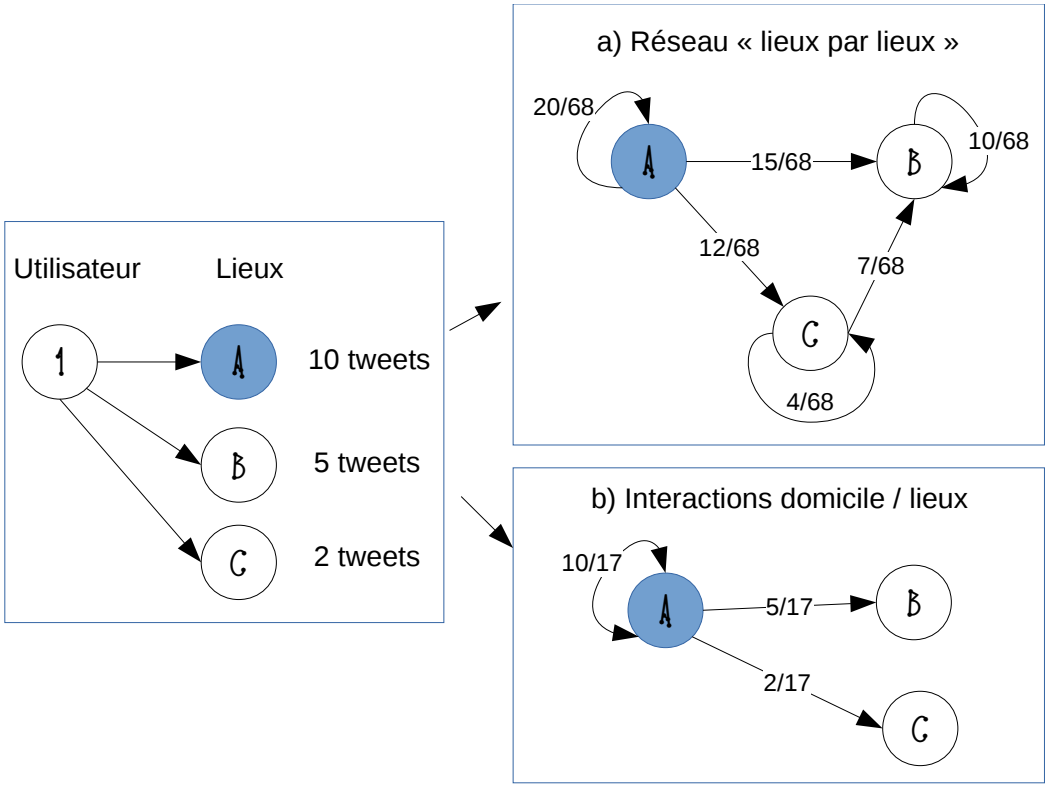


FIGURE 228 Différentes manières de concevoir un réseau à partir des lieux fréquentés. Le lieu "A" correspond au sous-district du domicile d'un utilisateur. Il est pris en compte dans la méthode b), mais pas dans la méthode a).

Le poids (ou niveau de l'interaction) W entre deux lieux i et j est défini ici comme la somme de toutes les interactions des utilisateurs u, membre de l'ensemble U entre ces deux zones. Ces dernières sont calculées comme étant la somme des *tweets* envoyés en i et j, divisé par l'ensemble des *tweets* envoyé par l'utilisateur dans toutes les zones L qu'il a visité :

401. "Location by Location network"

$$W_{ij} = \sum_{u=1}^{U} \frac{T_{iu} + T_{ju}}{(L+1)\sum l = 1^{L} T_{lu}} \tag{28}$$

La figure 228. a illustre la méthode de création d'un graphe pour un utilisateur, avec la prise en compte des interactions locales (de A vers A par exemple). La somme de ces interactions individuelles vaut 1, chaque utilisateur a donc le même poids dans ce réseau.

Nous proposons ensuite une deuxième méthode, relativement plus simple à mettre en œuvre. Il s'agit simplement de diviser pour chaque utilisateur le nombre de *tweets* envoyés dans une zone donnée par le nombre total de *tweets* émis (somme à 1). L'ensemble de ces parts de messages envoyés par secteur est ensuite associé à la zone de domicile de l'utilisateur (figure 228.b) que nous avons estimé précédemment (chapitre 6). La somme des fractions des flux sortants pour une zone donnée (en prenant en compte les flux internes, ou locaux) est donc égale à la population de l'échantillon dans ce secteur. Nous appellerons ce réseau « Domicile / Sous-districts ».

2.1.2 *Interactions entre les utilisateurs de Twitter, selon une approche sur les lieux fréquentés en commun*

Les deux méthodes précédentes permettent de reconstruire des réseaux selon les présences des utilisateurs dans différentes zones de la ville, indépendamment des relations entre les individus. La méthode de Poorthuis prend en compte toutes les interactions, tandis que la seconde méthode se focalise sur les interactions entre le lieu de domicile de chaque personne et les autres zones de la ville.

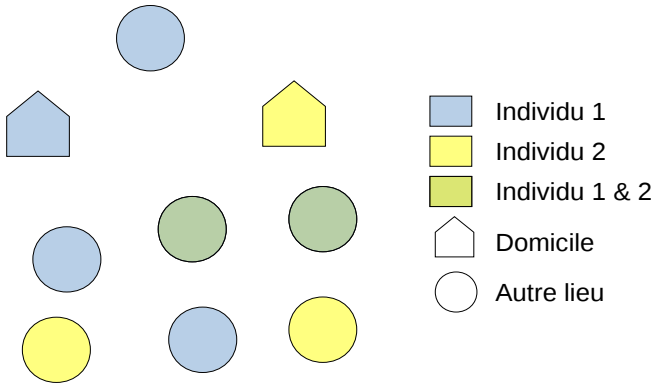


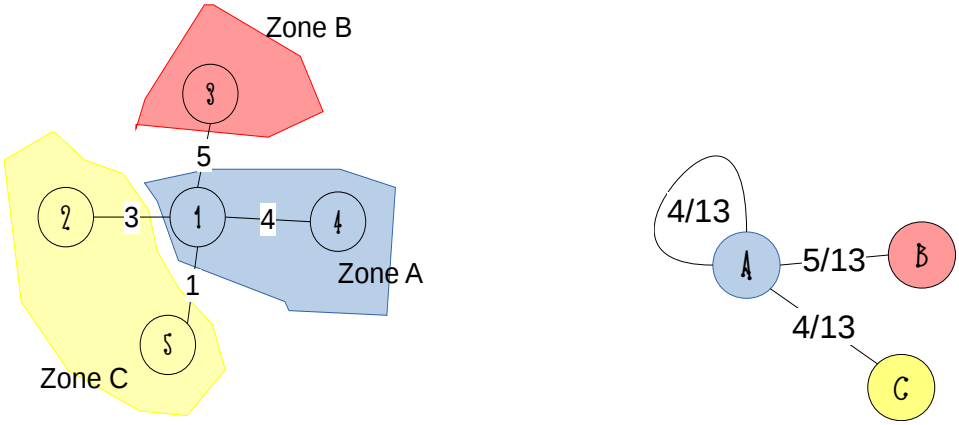
FIGURE 229 Exemple de lieux fréquentés par deux personnes. Certains de ces lieux sont visités conjointement.

En contexte épidémique, il est important de prendre en compte spécifiquement les zones de brassages et de co-présences, car c'est dans ces dernières que les agents pathogènes sont les plus susceptibles d'être transmis entre les hôtes. Nous proposons donc ici une troisième

approche, qui prend en compte pour chaque individu le nombre de lieux en commun avec les autres personnes de l'échantillon. Aussi, les cas de dengues sont enregistrés au domicile, il faut donc considérer les zones de résidence de chaque individu. Le réseau sera donc défini en prenant en compte le domicile de chaque personne qui fréquente les mêmes lieux.

La figure 229 est le scolie qui résume le début de la démarche, avec deux individus, 1 et 2, qui fréquentent différents lieux dans la ville. S'il y a transmission d'agents pathogènes entre ces deux personnes, elle se fera dans l'un des deux lieux qu'ils ont en commun, et les cas symptomatiques déclarés seront enregistrés dans le secteur du domicile.

Pour construire notre réseau, nous reprenons les espaces d'activités créés dans le chapitre 6, où chaque lieu est associé à une maille de 180 m de côté. Nous ne gardons que les mailles où chaque utilisateur a été actif pendant plus de 5 % des jours et où il a envoyé au moins un *tweet*, afin de ne garder que les zones les plus fréquentées pour chaque individu. Nous comptons ensuite pour chaque utilisateur le nombre de mailles en commun avec chacun des autres utilisateurs de *Twitter*.



(a) (b)

FIGURE 230 Principe de construction d'un réseau orienté sur les lieux fréquentés en commun pour un utilisateur. Le nombre de lieux en commun entre l'utilisateur 1 et les utilisateurs 2, 3, 4 et 5 (a) est respectivement de 3, 5, 4 et 1. Le tout est ensuite agrégé au sous-district où sont domiciliés les utilisateurs. Un poids est alors donné entre chaque zone de domicile (ou nœud) (b).

La figure 230.a illustre la construction du graphe pour un individu 1. Ce dernier partage 4 lieux avec l'utilisateur 4, qui habite dans la même zone A que lui. Il partage 5 lieux avec l'utilisateur 3 qui vit dans la zone B et respectivement 1 et 3 lieux avec les utilisateurs 5 et 2 qui habitent en zone C. Nous reconstruisons ensuite notre graphe, de la même manière que la seconde méthode évoquée précédemment, en comptant les interactions entre les différentes zones (ici les sous-districts), centrées sur le domicile de l'utilisateur concerné (figure 230.b). Nous agrégeons ensuite ces graphes individuels, ce qui nous permet d'obtenir des interactions

indirectes entre les différentes zones de domiciles des utilisateurs de *Twitter* à Bangkok. La somme des flux sortants et des flux internes (dans le même sous-district) d'un *khwaeng* est ici égale au nombre d'utilisateurs résidant dans le sous-district. Nous appellerons ce réseau « Domicile/Domicile ».

Ces trois réseaux, créés selon des méthodes très différentes, nous semblent tous aussi valables d'un point de vue conceptuel. Le premier réseau, dit « lieux/lieux » suit une approche assez naïve et est créé en prenant en compte toutes interactions entre les zones fréquentées par un individu. Le second réseau, « Domicile/Sous-district » résume les interactions entre les lieux de domicile et les autres sous-districts, mais ne prend pas en compte, à titre individuel, les interactions entre les sous-districts qui ne sont pas ceux du domicile. Enfin, le réseau « Domicile/Domicile », part des lieux fréquentés en commun par différents utilisateurs pour définir de manière indirecte les interactions entre les *khwaengs* où ces derniers sont domiciliés.

2.2 Comparaison des résultats

Il existe différentes manières de comparer des matrices OD, notamment en calculant entre chaque matrice la moyenne des pourcentages d'erreur absolue bornée (MCAPE) (Commenges, 2016 ; Cools *et al.*, 2010). Cette méthode revient à calculer pour chaque couple origine-destination les différences entre les flux estimés et les flux observés, divisées par les flux estimés ces derniers étant borné entre la valeur minimale et 100. Le tout est ensuite moyenné. Mais cette approche n'a d'intérêt que si nous sommes en présence d'une matrice origine-destination de référence, ce qui n'est pas le cas. Nous allons donc simplement comparer visuellement ces trois réseaux (ou matrices OD) créés, même si le calcul de cet indicateur global ne pose aucun problème technique.

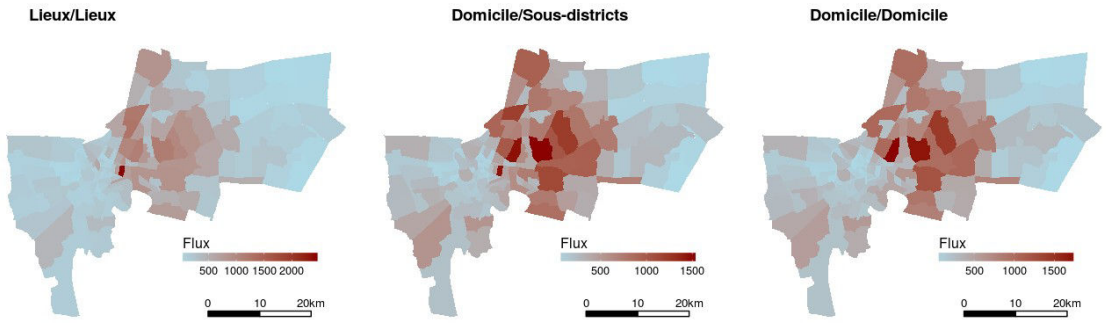


FIGURE 231 Sommes des flux entrant et sortant par sous-district selon les différentes matrices OD créées.

La figure 231 ci-dessus présente la somme des flux entrants et sortants par sous-district pour chacune des matrices, définissant ainsi un niveau de centralité pondéré (Poorthuis, 2018 ;

Scott, 1992). Nous pouvons noter une sorte de continuité entre ces différentes matrices. Le réseau « Lieux/Lieux » fait principalement ressortir, en plus des zones assez peuplées de l'est de la ville, le sous-district où se situent les principaux *malls*. Le réseau « domicile/sous-district » donne plus de poids aux zones à l'est du centre historique, tout en conservant le khwaeng ultra-central de l'hypercentre. Ce dernier est néanmoins exclu du réseau « Domicile/Domicile », qui ne prend en compte que les interactions entre les lieux de résidence des personnes qui fréquentent les mêmes lieux. Le réseau « Domicile/Sous-districts » serait donc en somme une sorte de synthèse des deux autres types de réseaux, prenant à la fois en compte les zones d'habitations de l'est du centre et le sous-district où se situent les principaux *malls* et l'université de Chulalongkorn.

Afin de conserver que les flux significatifs dans nos trois réseaux créés, nous appliquons la même méthode que Poorthuis (2018) pour ne conserver que les flux principaux, que l'auteur considère comme significatifs. Pour ce faire, nous créons 1000 matrices (ou réseaux) en répartissant les sommes en lignes et en colonne de manière aléatoire[402]. Nous ne conservons ensuite que les liens dont les poids ont été supérieurs à ceux générés plus de 990 fois sur 1000 (pour avoir un seuil à 1 %). Les résultats sont visibles dans la figure 232, ci-dessous, qui montre à la fois la somme des flux entrants par sous-district (gauche), ainsi que le ratio entre les flux entrants et les flux sortants (droite).

Une première remarque est que la suppression des flux non-significatifs entraîne une claire réduction de l'importance du sous-district de l'hypercentre où se concentrent les principaux *malls*. L'importance de ce *Khwaeng* dans la centralité des réseaux serait donc due en grande partie à une forte contribution de nombreux flux relativement faibles. Dès lors, les ordres de grandeur des centralités observées entre les différents réseaux sont plus ou moins similaires, notamment les réseaux « Domicles/Sous-districts » et « Domiciles/Domiciles » même si les valeurs absolues des flux différents.

Si nous regardons maintenant le rapport entre les flux entrants et les flux sortants, nous pouvons noter que le réseau « Lieux/Lieux » est le plus équilibré, avec finalement de faible différences entre ces différents flux. De nombreux *Khwaengs* du réseau « Domicile/Sous-districts » reçoivent beaucoup plus de flux qu'ils n'en émettent (les couleurs grisées correspondent à des ratios supérieurs à 2), notamment dans le centre, mais aussi dans un sous-district très périphérique à l'est. Enfin, le réseau « Domicile/Domicile » présentent des valeurs intermédiaires.

402. en utilisant la fonction r2table, dans R

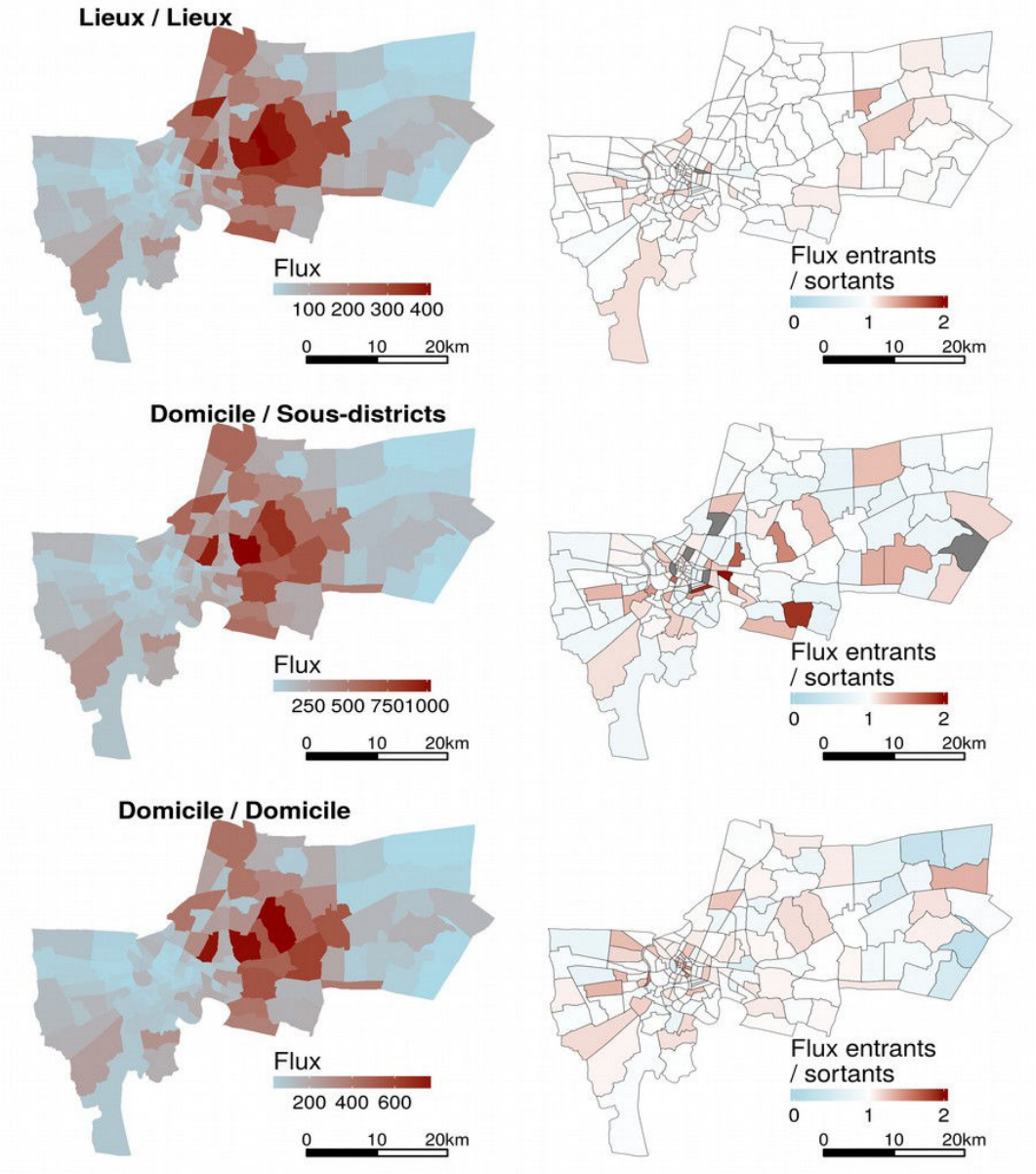


FIGURE 232 Sommes des flux significatifs entrants et sortants (gauche) et rapport entre flux entrants et sortants (droite) par sous-districts, selon les trois méthodes de construction de réseaux

Les figures 233 et 234 illustrent les interactions entre les différentes zones de la ville. Sont représentés pour chaque sous-district les deux flux (ou lien du réseau) les plus importants vers d'autres zones de la ville. Les 100 flux les plus importants en valeur absolue sont représentés par les courbes dont la couleur varie du bleu (flux relativement faible) au rouge (flux maximum) et dont l'épaisseur est proportionnelle à la valeur de ces liens. La taille des cercles correspond au volume des flux locaux. Différents pôles apparaissent, situés par rapport au centre historique, au nord, à l'est et au sud-est. Ils sont plus ou moins connectés entre eux selon les méthodes employées pour leur construction.

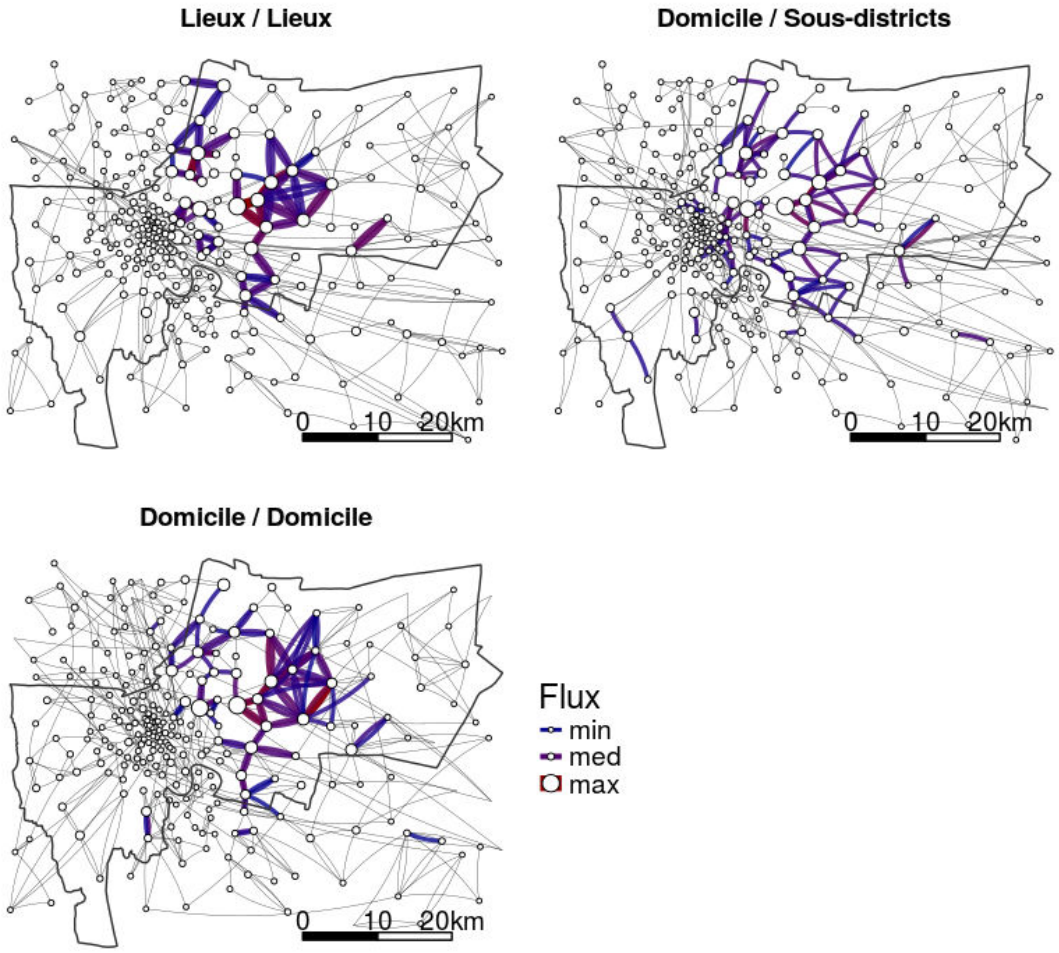


FIGURE 233 Représentation des 100 flux les plus importants et des 2 flux principaux par sous-district, pour chacune des trois méthodes.

Si nous regardons les deux flux principaux par sous-districts (trait fin noir), la plupart des interactions majeures se font suivant une logique de proximité, vers les sous-districts limitrophes. Mais une part non négligeable de ces flux est aussi dirigée vers le centre-ville, du moins pour les réseaux « Lieux/Lieux » et « Domicile/Domicile », comme nous pouvons le voir sur la figure 234. Si des structures en proche périphérie du centre historique peuvent sembler assez similaires selon les méthodes employées, il n'y a pas de motifs clairs et communs qui apparaissent dans l'hypercentre. À noter cependant que seuls les deux principaux flux sont ici représentés.

Différents niveaux d'interactions apparaissent donc selon les réseaux construits. Si les réseaux « Lieux/Lieux » et « Domiciles/Sous-districts » sont, du fait de leur formalisation, probablement susceptibles de mieux retranscrire les flux entre les différents secteurs de la ville, il est impossible à notre niveau de dire lequel se rapproche au mieux des interactions réelles. Nous n'avons en effet aucune donnée institutionnelle de référence et n'avons pas effectué d'enquêtes de terrain.

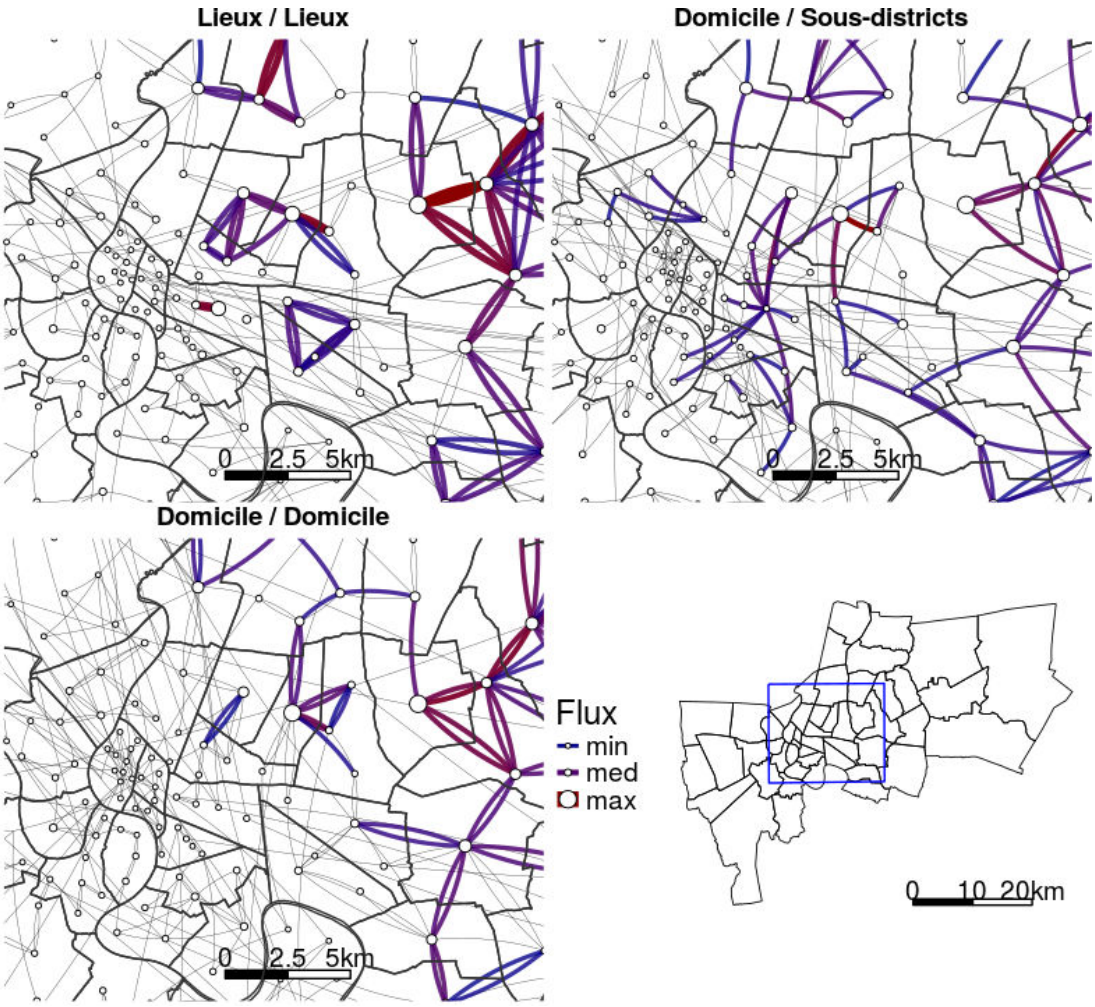


FIGURE 234 Représentation des 100 flux les plus importants et des 2 flux principaux par sous-district, pour chacune des trois méthodes. Zoom sur le centre de Bangkok.

Pour revenir au contexte de la dengue, il serait intéressant d'utiliser ces différentes matrices de flux dans des modèles métapopulations classiques ou des modèles épidémiques en réseau. Selon leur formalisation (modèles déterministes ou stochastiques, plus ou moins complexes, etc.), il devrait être possible d'estimer le nombre de personnes infectées dans chaque zone à un pas de temps donné et d'observer la propagation de l'épidémie sur les différents réseaux (voir chapitres 3 et 5).

Car les différentes matrices origine-destination (ou réseaux, graphes) présentées précédemment montrent, au-delà de leurs grandes différences issues de leurs conceptions, que certaines zones sont plus liées entre elles que d'autres. Ceci peut avoir un impact sur la diffusion des épidémies, qui auraient alors plus de chance de circuler entre des sous-districts où les flux sont importants. La prochaine section va chercher à définir des régions fonctionnelles en regroupant des sous-districts dans des zones qui interagissent plus entre eux qu'avec ceux situés dans une autre zone.

2.3 Structure et partition de Bangkok par les interactions dans la ville

2.3.1 Définir des sous-régions fonctionnelles par la détection de communautés

Transférée à la théorie des graphes, la création de régions fonctionnelles selon leur niveau d'interaction avec les autres régions revient à définir des communautés (ou des *clusters*), qui sont des groupes de nœuds (dans notre cas des sous-districts) qui partagent des propriétés similaires (ici les flux entre chaque zone) et/ou jouent un rôle équivalent au sein du graphe (Fortunato, 2010). Dans notre cas, une communauté sera un ensemble de sous-districts, réunis dans un même groupe de par leur niveau d'interaction dans le graphe. La figure 235 illustre cette approche, où 15 nœuds sont regroupés en trois catégories. Les nœuds de chaque catégorie ont des relations plus importantes entre-eux qu'avec les nœuds des autres catégories. Cela revient donc à partitionner le réseau en sous-graphes (ou communauté), et un il existe énormément d'algorithmes pour réaliser cela (Lancichinetti and Fortunato, 2009).

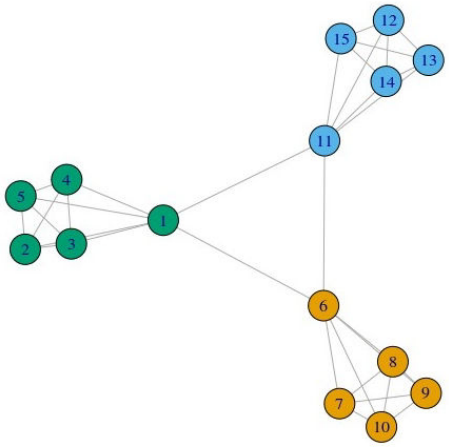


FIGURE 235 Exemple de détection de communauté dans un graphe. De par leur niveaux d'interactions, les nœuds numérotés de 1 à 5 forment un premier cluster, ceux de 6 à 10 un second, et ceux de 11 à 15 un troisième.

Comme le rappel Poorthuis, (2018), si les premiers travaux sur la recherche de sous-groupes cohérents furent effectués par des chercheurs en sciences sociales à partir des années 1950, le champ de la détection de communauté est en plein essor depuis 15 ans, avec l'augmentation des puissances de calculs et la contribution des physiciens, avec notamment l'introduction du concept de modularité par Newman et Girvan, (2004).

La modularité est une mesure qui calcule le nombre de liens (pondérés ou non) dans un même groupe, moins le nombre attendu dans un réseau équivalent (mêmes sommes marginales), mais aléatoire (Newman, 2006). La valeur obtenue varie entre -1 et 1, et plus cette dernière

est importante, plus la partition du graphe en communauté est significative (Newman et Girvan, 2004). En revanche, une faible valeur de modularité signifie que la force des liens des communautés n'est pas plus importante que dans un réseau aléatoire.

Certains algorithmes, comme celui de « *Louvain* » (Blondel *et al.,* 2008) visent la maximisation de la modularité. Il s'agit d'un algorithme itératif en deux temps. La première étape commence par associer chaque nœud à une communauté unique. Les nœuds sont ensuite affectés à la communauté voisine qui permet d'obtenir une modularité maximale, positive et supérieure à l'itération précédente. Cette étape est répétée jusqu'à ce qu'aucune amélioration de la modularité ne soit observée. La seconde étape va considérer chaque communauté comme étant un nouveau nœud, auxquelles on applique de nouveau la première étape (*ibid.)*. À noter que cette méthode entraine un partitionnement de l'espace en cluster de taille assez homogène, ce qui implique que des petites communautés seront souvent regroupées au sein d'entités plus grandes (Adam *et al.,* 2017). Cet algorithme a par exemple été employé pour définir des communautés géographiques au Portugal et en Côte d'Ivoire à partir des interactions entre les différentes antennes relais de ces pays (Amini *et al.,* 2014). Il a également été appliqué dans la région de Bruxelles sur des matrices créées à partir de données issues des migrations résidentielles, de navettes domicile-travail et de données téléphoniques (Adam *et al.,* 2017). Cette approche a permis d'apprécier l'équilibre du réseau urbain à différentes échelles spatiales et temporelles ainsi que les structures spatiales de la capitale Belge (*ibid.*).

D'autres algorithmes, comme le *Walktrap* (Pons et Latapy, 2005) sont basés sur une marche aléatoire (*random walk*, chapitre 5). La distance entre chaque nœud du réseau est calculée selon la probabilité qu'un nœud en atteigne un autre en se déplaçant dans le réseau en suivant une marche aléatoire selon un nombre fixé d'itérations. Une classification ascendante hiérarchique suivant la méthode de *Ward* (1963) est ensuite appliquée à partir de ces distances et le dendrogramme est coupé de manière à obtenir le partitionnement qui maximise la modularité. Un faible nombre d'itérations entraîne une exploration limitée du réseau, et donc un partitionnement en un plus grand nombre de classes, tandis qu'un grand nombre d'itérations tend à réduire le nombre de communautés.

Un autre algorithme, *Infomap* (Rosvall et Bergstrom, 2008), n'est pas basé sur la notion de modularité mais suit une logique de compression optimale de l'information. Il s'agit d'une approche qui vise à minimiser la description des déplacements aléatoires (de type marche aléatoire) dans le réseau. Des codes binaires sont attribués à des groupes de nœuds en fonction du nombre d'itérations effectuées dans ces groupes, en partant du principe que plus d'itération est faite dans un groupe, plus il y a de chance qu'il s'agisse de la même communauté. Si les auteurs utilisent la métaphore des cartes géographiques, qui résultent d'une optimisation du niveau de détail selon l'échelle, une bonne compréhension de cet algorithme nécessite néanmoins

des bases assez solides en traitement et compression de signaux[403]. Nous sommes conscients que la présente description d'*Infomap* peut sonner comme du galimatias, mais cet algorithme serait néanmoins l'un des plus performants dans la détection des communautés (Fortunato, 2010).

Nous testerons ici ces 3 algorithmes sur nos trois réseaux (ou matrices OD). L'un est basé sur l'optimisation de la modularité (*Louvain*), l'autre est basé sur l'optimisation d'une *random-walk* (*Infomap*), le dernier, *Walktrap*, utilise ces deux approches. Ces trois algorithmes sont implémentés sous R dans la librairie « *igraph* » développée par Csardi et Nepusz, (2006). Il existe toutefois une ribambelle d'autres algorithmiques, que nous ne testerons pas ici. Pour une revue un peu ancienne mais assez exhaustive, voir par exemple Fortunato, (2010), ou encore Yang *et al.* (2016).

2.3.2 Application à Bangkok

La figure 236 ci-dessous présente les différents regroupements des sous-districts obtenus selon les trois algorithmes[404] appliqués à nos trois matrices OD. Bien entendu, des différences sont visibles selon les configurations (matrices et algorithmes), tant sur la délimitation des voisinages que sur leur nombre. Par exemple, l'est de la ville représente une seule grande communauté dans le réseau « lieux/lieux » avec l'algorithme *Walktrap*, mais est subdivisé en 4 zones avec Louvain et 6 avec Infomap.

Néanmoins, nous pouvons observer une certaine régularité dans les regroupements des différents *Khwaengs,* avec par exemple la zone correspondant au quartier de Sathorn/ Silom, au sud du centre-ville (en rose sur les classifications issues du réseau "Domicile/sous-district", ou en violet pour le réseau "Domicile/Domicile") dont la délimitation est quasiment tout le temps la même.

Une manière de synthétiser les différents résultats de ces classifications est de définir pour chaque sous-district une séquence correspondant à la suite des classes qui lui ont été assignées dans toutes les configurations, et de regrouper entre eux les sous-districts ayant les mêmes séquences. Ainsi, la figure 237.a regroupe les *Khwaengs* qui ont été systématiquement classés de la même manière, tandis que la figure 237.b présente un découpage où les sous-districts sont classés de la même manière au moins 7 fois sur 9. Les *Khwaengs* laissés en blanc n'appartiennent à aucun groupe.

403. une démonstration de l'algorithme est disponible ici : http://www.mapequat on.org/apps/MapDemo.htm

404. Avec 100 itérations pour l'algorithme *walktrap* défini après des tests empiriques et un paramètre de contrôle de modularité p à 1 pour *Louvain* ce paramètre n'étant pas modifiable dans la fonction présente dans la librairie *igraph*.

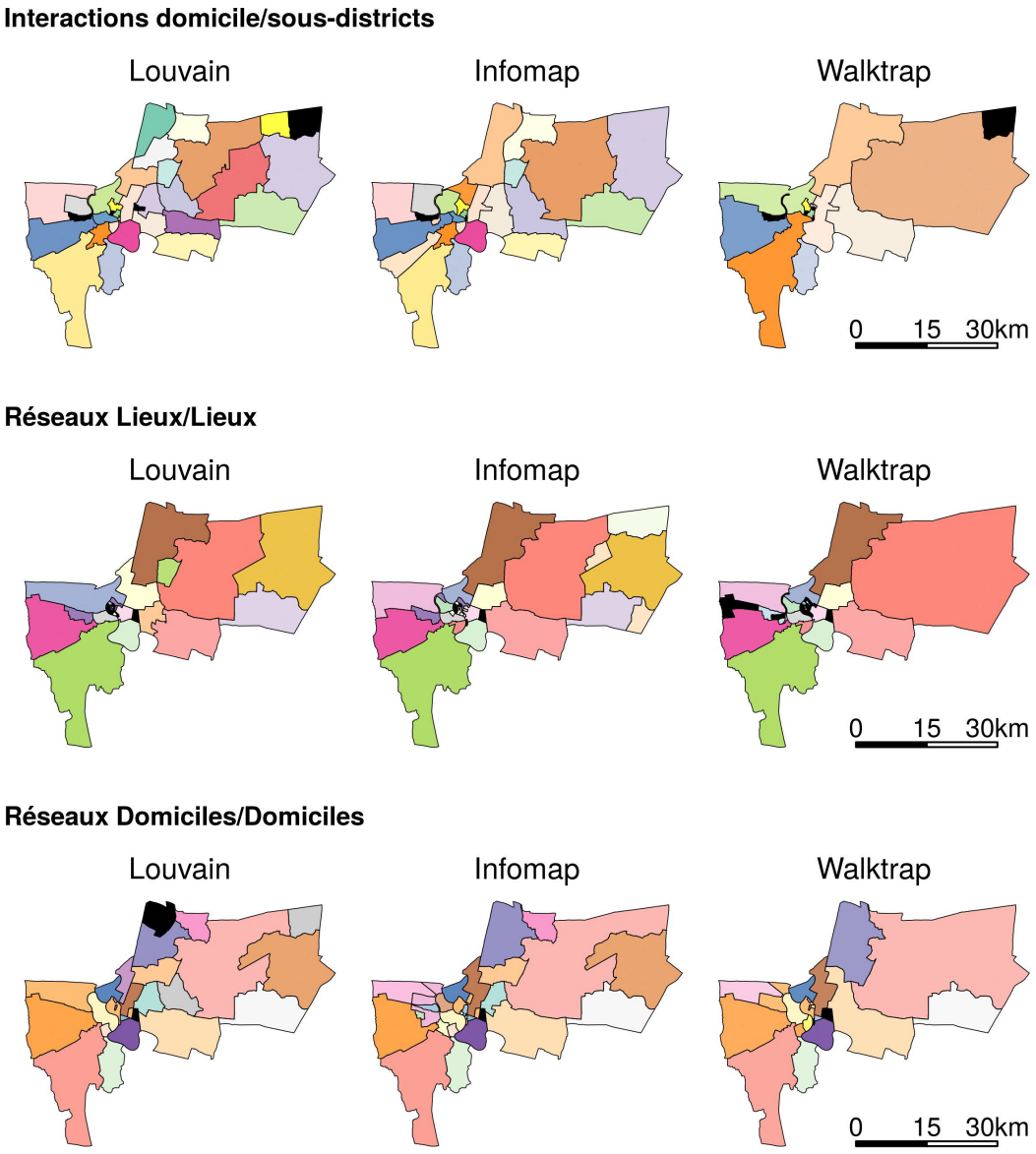


FIGURE 236 Regroupement des sous-districts en communautés, selon les différents réseaux utilisés en entrée et les trois méthodes de détection de communautés testées. Les couleurs représentent les différentes communautés obtenues après l'application des différents algorithmes et sont propres à chaque réseau. Les *Khwaengs* isolés sont représentés en noir.

La partition de la ville de la figure 237.a est le regroupement le plus fin issu des configurations évoquées précédemment. Les unités fonctionnelles sont assez nombreuses, et si certaines ne comptent que deux sous-districts d'autres, comme la zone marron au sud du centre historique regroupent jusqu'à 9 *Khwaengs*. Si nous ajoutons un peu de flexibilité dans la création de ces classes (figure 237.b), nous pouvons observer une augmentation de la taille des zones, ainsi qu'une réduction de leur nombre du fait de regroupements.

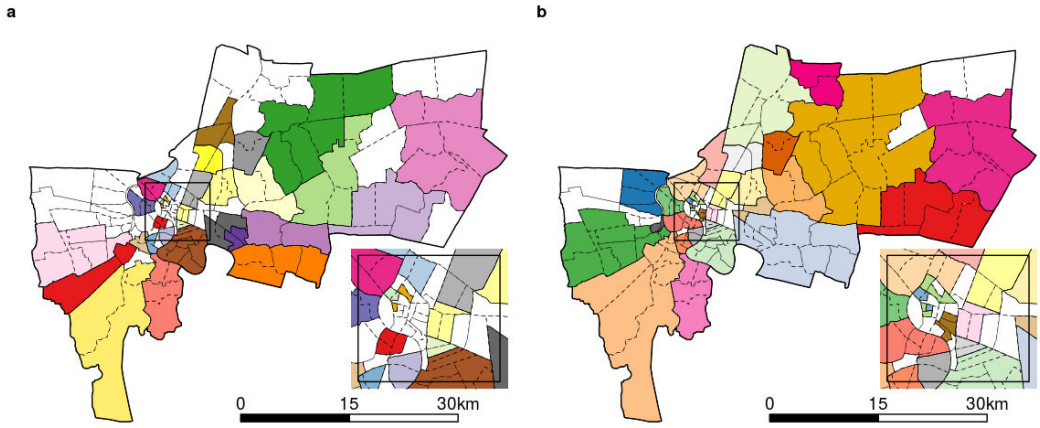


FIGURE 237 Définition de zones fonctionnelles, où tous les sous-districts sont classés de la même manière selon l'ensemble des classifications (a), et avec au maximum deux différences (b). Les *Khwaengs* laissés en blanc n'appartiennent à aucun groupe.

Ainsi, malgré des matrices OD construites de manières très différentes, avec des interactions entre zones propres à chaque réseau et des algorithmes de détection de communauté différents, nous arrivons toutefois à définir des zones fonctionnelles - ou voisinages assez robustes, où les interactions au sein de ces secteurs seraient plus importantes.

Si nous regardons maintenant les sous-districts qui n'appartiennent à aucun groupe (en blanc) nous pouvons noter qu'ils se situent soit dans l'hypercentre, soit en périphérie (nord-est ou ouest), ce qui signifie que ces zones sont assez spécifiques. Pour ce qui est des zones périphériques, relativement peu de *tweets* y sont enregistrés et peu de personnes y vive, ce qui peut expliquer le fait qu'elles ne soient pas regroupées avec d'autres sous-districts, du fait de faibles interactions. En revanche, les sous-districts de l'hypercentre sont très fréquentés, et par des personnes provenant de toute la ville. Les flux entre ces zones seraient donc relativement faibles au regard des flux provenant de l'ensemble de la ville, et leur appariement à d'autres sous-districts serait plus sensible aux matrices et aux algorithmes.

L'association de *khwaengs* entre eux sous la forme de zones fonctionnelles, caractérisées par des interactions privilégiées entre ces sous-districts est aussi une bonne piste de recherche en épidémiologie, notamment pour étudier la propagation locale d'une maladie. Une hypothèse à étudier serait qu'une épidémie (*e.g.* de dengue) qui sévit dans un sous-district donné aurait plus de chance de se propager dans un sous-district appartenant au même voisinage.

Synthèse

Bangkok est une ville extrêmement embouteillée, du fait d'un réseau de transport en commun probablement inadapté, mais aussi par son caractère principalement mono-centrique, où les zones les plus attractives sont concentrées dans l'hypercentre. Néanmoins, selon les données et l'échelle d'observation, différents pôles apparaissent, où plus ou moins de traces numériques géolocalisées sont enregistrées, et pas réparties uniquement dans l'hypercentre.

Si l'observation des pulsations urbaines tend à confirmer l'importance des flux entre la périphérie et le centre, l'étude des interactions entre les différentes zones permet de montrer qu'un grand nombre de ces flux se font entre des zones limitrophes, et sont donc plutôt locaux.

Chapitre XI: Génération d'agendas individuels : Premières notes d'un modèle à base d'agents

Les deux chapitres précédents traitaient de l'organisation spatiale et de la temporalité des activités dans Bangkok et des flux et interactions dans la ville. Nous aurons dans ce dernier chapitre une approche plus individu-centrée et élaborons des agendas individuels, en nous basant sur les données *Twitter* et la couche d'utilisation du sol. Si l'approche globale reste la même que celle présentée dans le chapitre 8, l'idée n'est pas ici d'appliquer exactement les mêmes méthodes, mais au contraire de prendre en compte les remarques et limites déjà observées afin de tenter de les dépasser.

Dans un premier temps, nous définirons l'activité principale potentielle de chacun des utilisateurs de *Twitter*, aspect central des mobilités quotidiennes. Mais ces espaces d'activités restent discontinus dans le temps, et nous proposerons une méthode de reconstruction d'agendas continus, selon une approche semblable à celle présentée dans le chapitre 8, mais simplifiée.

À partir de ces agendas reconstitués, nous définirons des groupes d'utilisateurs selon leur potentiel de mobilité, qu'il s'agisse du nombre d'activités ou de lieux fréquentés ou sur les tendances de déplacement dans l'espace (distances parcourues et niveau de dispersion dans la ville).

De ces groupes et agendas, nous générerons des agendas individuels propres à des agents synthétiques, première étape d'un modèle de mobilité basé sur le concept d'espace d'activité. Après une nouvelle critique des résultats et l'exposé de pistes d'amélioration, nous discuterons des données et des méthodes qui pourraient permettre d'affecter une localisation à chaque agent dans la ville.

1 De données épisodiques à des agendas individuels continus

1.1 Préalable : Définir l'activité principale

Pour l'instant, l'espace d'activité des utilisateurs de *Twitter* ne se compose que du lieu de domicile et d'une succession d'activités réalisées dans différents lieux. Cette section vise à déterminer parmi l'ensemble de ces lieux fréquentés, lequel peut être associé à l'activité principale de l'utilisateur. Nous emploierons la même méthode que celle utilisée dans le chapitre 8, que nous détaillons ici à titre de rappel.

1.1.1 Méthode

Comme nous l'avons vu précédemment, énormément de *tweets* émanent de les lieux d'éducations (écoles, collèges et universités), ce qui suggère déjà que les étudiants / écoliers ou personnes travaillant dans un de ces lieux représentent une part non négligeable de l'échantillon. Compte tenu que les universités sont souvent pourvues de grand campus, il est fort probable que différents lieux d'un espace d'activité (mailles de 180 m) se situent en fait dans une même enceinte universitaire (*e.g.* bibliothèque, une salle de cours, etc.). Dès lors, nous regroupons pour chaque utilisateur concerné tous les lieux associés à un même lieu d'éducation, d'après notre couche d'utilisation du sol (chapitre 9), sous un même lieu. En revanche, il est délicat de poser les mêmes hypothèses pour ce qui est des autres activités potentiellement principales. Nous nous en tiendrons alors aux lieux des espaces d'activités définis dans le chapitre 6.

Tout d'abord, nous partons du principe que l'activité principale peut se dérouler n'importe où, sauf dans un parc, un lieu de culte, ou un lieu de transport (routes, aéroports, gare[405]). L'activité principale étant effectuée régulièrement, une activité suffisante sur *Twitter* doit y être enregistrée. Nous posons qu'un utilisateur doit y avoir *tweeté* plus de 5 % des jours où il a été actif sur le réseau social, et pendant au moins 5 jours différents. Nous posons aussi que le travail est une activité essentiellement diurne, et en nous basant sur les heures de pointe des transports (chapitre 10, section 1.1.2), nous posons qu'à minima deux fois plus de messages sont envoyés depuis ce lieu entre 7 h et 19 h qu'entre 19 h et 7 h.

Si plusieurs lieux d'un espace d'activité d'un individu répondent à ces critères, nous sélectionnons celui où il a été actif sur un plus grand nombre de jours et où il a le plus *tweeté* entre 7 h et 19 h. Au final, 13 824 des 38 696 personnes de notre échantillon présentent un lieu qui correspond à ces critères.

405. Même si beaucoup de personnes travaillent dans ces types de lieux, nous estimons qu'il s'agit ici plus de lieux de passages.

Mais pour revenir aux étudiants, il est tout à fait possible que les campus soient répartis dans différents endroits de la ville. Nous sélectionnons les utilisateurs ayant émis des *tweets* depuis un quelconque lieu d'éducation plus de 10 % des jours où ils ont été actifs sur la plateforme. Parmi ces lieux, nous sélectionnons ceux où l'utilisateur a été actif pendant au moins 5 jours et où la plus grande activité est enregistrée. Nous rajoutons ainsi 3544 étudiants, pour obtenir au final 17 368 utilisateurs avec une activité principale et un domicile.

1.1.2 Analyse de l'échantillon

Dans l'algorithme exposé précédemment, aucune distinction n'est faite entre les jours de la semaine et de week-end. Dès lors, le fait d'observer les différences de fréquentations quotidiennes par type de lieux selon qu'il s'agisse d'un lieu défini comme étant l'activité principale ou non peut être un bon indicateur de validité de la méthode. Ceci est visible dans la figure 238 ci-dessous, où est représenté la différence entre les pourcentages de fréquentation quotidienne dans chaque type de lieux, selon qu'il s'agisse de l'activité principale où d'une activité parmi d'autre. Une différence positive signifie que le lieu est plus fréquenté un jour donné par les personnes dont s'est l'activité principale une valeur négative signifiant l'inverse. Les lieux correspondants aux activités principales sont systématiquement nettement moins fréquentés les week-end et davantage visités les jours de semaines[406], ce qui tend à valider notre méthode. D'autant plus que la même observation était faite lorsque nous appliquions cet algorithme sur des données *Twitter* à Delhi (chapitre 8).

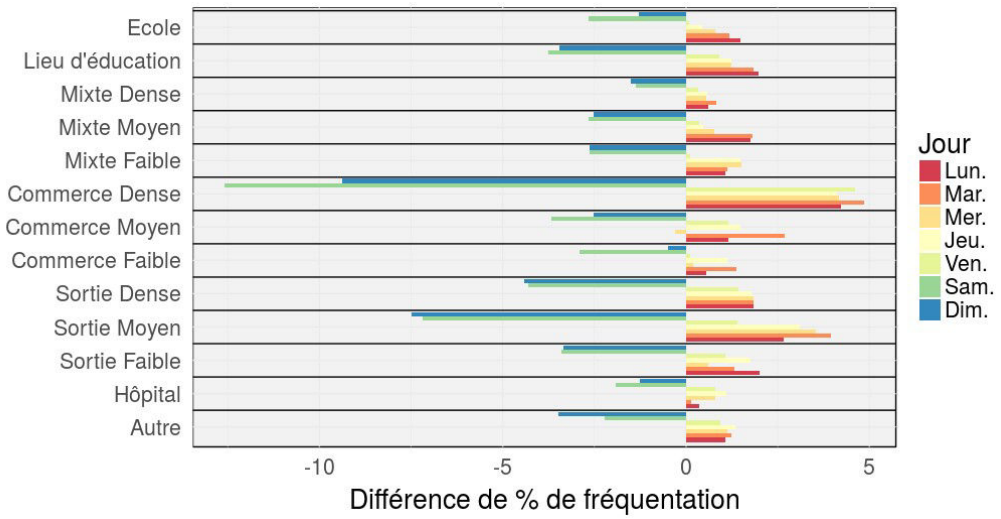


FIGURE 238 Différence de pourcentage entre le nombre d'utilisateurs uniques par jour de la semaine pour chaque type d'activité, et le nombre de personnes travaillant dans ce type de lieu.

406. Sauf le mercredi, pour les lieux de type "Commerce moyen".

La figure 239 ci-dessous présente la répartition des lieux correspondants aux activités principales de notre échantillon. Plus de 50 % des utilisateurs travailleraient dans un lieu d'éducation, et seraient donc des étudiants, lycéens ou feraient partie du personnel. Si seulement 10 % de la population de Bangkok fréquentaient des lieux d'éducations en 2009 (NSO, 2009), 39 % des 18-24 étudiaient, dont 28,4 % dans des universités (NSO, 2008). Les étudiants (ou assimilés) sont surreprésentés dans l'échantillon, même parmi les tranches d'âges concernées (18-24 ans). 25 % des utilisateurs travaillent dans un lieu non présent dans notre couche d'utilisation du sol ("Autre") et environ 10 % dans des zones de type « mixte dense ».

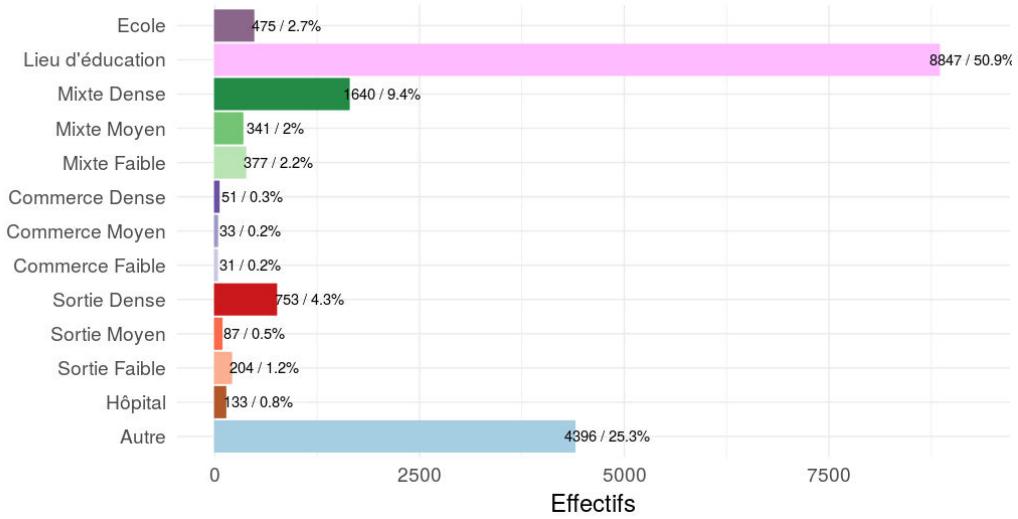


FIGURE 239 Répartition de l'échantillon selon le type de lieu considéré comme étant l'activité principale.

Si nous regardons maintenant le niveau de représentativité spatiale des domiciles des plus de 17 000 personnes dont nous avons pu estimer raisonnablement une activité principale (figure 240), nous pouvons noter différentes choses. Tout d'abord, le R^2 ajusté entre le nombre de domiciles d'utilisateurs (avec une activité principale) dans un *Kwhaeng* et la population estimée par le recensement de 2010 est inférieur à celui obtenu précédemment, lorsque nous ne considérions l'ensemble des 38 696 personnes. Mais ce R^2 est tout de même assez élevé, à 0,7 si nous écartons les étudiants. De plus, fait intéressant, alors que le coefficient de corrélation est relativement bas pour les étudiants dans Bangkok (0,57) ou dans la région (0,55), il augmente nettement lorsque l'on ne considère pas la population totale mais les 15-24 ans (0,61 et 0,64). A l'inverse, il baisse pour les personnes n'étant pas considérées comme des étudiants, ce qui suggère que ce dernier groupe est composé d'une population dont la tranche d'âge est plus ample.

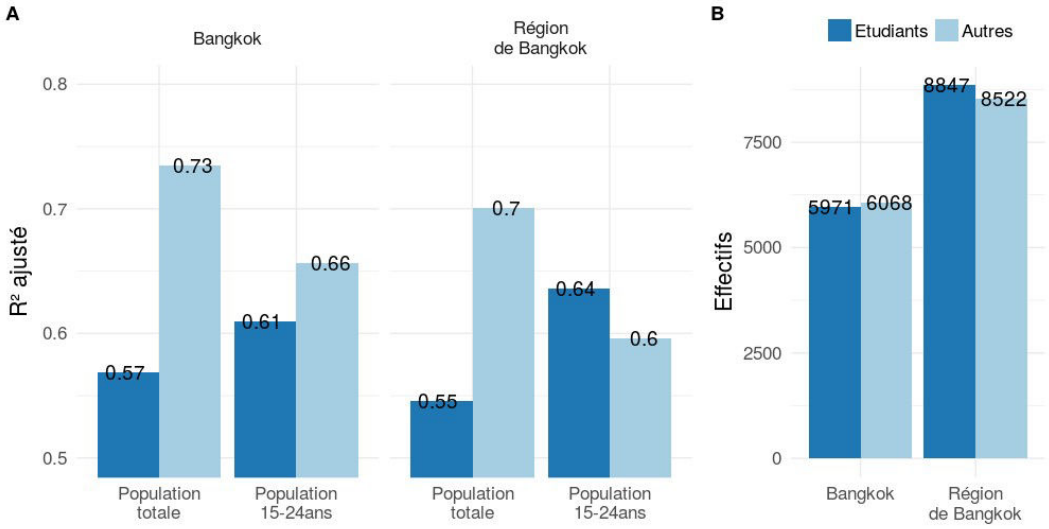


FIGURE 240 Représentativité spatiale de l'échantillon dont une activité principale a pu être estimée. N=17 368.

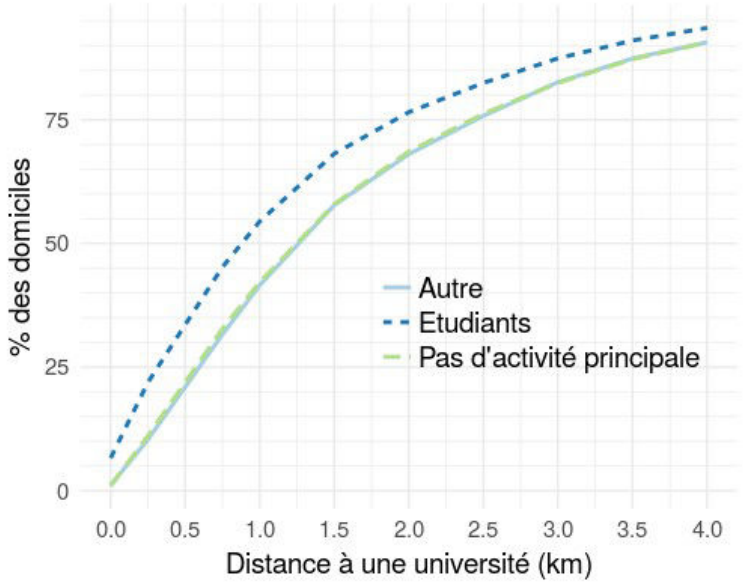


FIGURE 241 Part de chaque groupe selon la distance du domicile à un lieu d'éducation.

Pour le R^2 relativement faible chez les étudiants (ou assimilés), une explication pourrait être que la localisation de leur domicile est en partie contrainte par celle des universités. Si nous regardons la part de chaque groupe selon leur activité principale (figure 241) nous pouvons noter que les étudiants (ou assimilés) ont tendance à vivre dans une zone plus proche d'un lieu d'éducation que les autres groupes, même ceux qui n'ont pas d'activité principale. 55 % des étudiants vivent en effet à moins d'un kilomètre d'une université, contre 40 % pour les autres groupes. Cette distribution géographique, en partie contrainte par la localisation des universités

peut entraîner la formation de clusters spatiaux et expliquer cette plus faible représentativité spatiale pour les étudiants. De plus, les étudiants d'autres provinces ne sont pas nécessairement pris en compte dans le recensement.

Nous avons donc pu estimer l'activité principale de plus de 17 000 personnes, activité effectuée principalement les jours de semaine alors qu'aucune contrainte sur les jours de réalisations n'avait été posée. L'échantillon est composé d'une moitié d'étudiants (ou assimilés), et le niveau de représentativité spatiale de l'échantillon reste très acceptable et cohérent.

Nous considérons maintenant que ces espaces d'activités sont complets, en termes de lieux fréquentés et d'activités réalisées. Néanmoins, ils restent discontinus dans le temps. Nous allons maintenant reconstruire des agendas, passant de données temporellement épisodiques à des séquences d'activités que nous désirons être le plus crédibles possible. Pour cela, nous allons reprendre et apporter quelques modifications et simplifications à la méthode utilisée dans le chapitre 8.

1.2 Protocole de reconstitution des agendas

Ici encore, nous ne prendrons pas en compte les temps de transports dans l'agenda d'un individu. Ils pourront être estimés par la suite, lorsqu'une localisation sera attribuée à chaque lieu, et pourront être intercalés entre chaque lieu fréquenté. Nous raisonnons ici de la même manière que dans le chapitre 8, à savoir que nous travaillerons sur une période d'un mois et affecterons des activités et des lieux fréquentés en cascade, en définissant pour chaque lieu une semaine, un jour et des horaires de réalisation. Nous ne prenons pas en compte le lieu de domicile, en partant du principe qu'il s'agit du lieu fréquenté par défaut - lorsque aucun notre lieu n'est visité.

1.2.1 Étape 1 : Définir la probabilité de visite d'un lieu de l'espace d'activité

○ 1.1 Probabilité de réaliser une activité une semaine donnée

Précédemment (chapitre 8), nous considérions les probabilités de visite de lieux une semaine donnée selon le rapport entre le nombre de semaines où un l'individu a été actif dans un lieu, divisé par le nombre de semaines où il a été actif sur *Twitter*. Une des limites est que si un individu effectue une même activité, mais de manière assez ponctuelle dans un grand nombre de lieux, la probabilité de tirer un de ces lieux, et donc de réaliser l'activité associée devient relativement faible. Nous avions remédié à cela en augmentant la probabilité de visite d'un lieu à l'itération suivante si ce dernier n'avait pas été visité. Nous proposons ici une autre approche, centrée plus réalisation d'une activité que sur la visite d'un lieu associé à une activité.

Nous comptons tout d'abord le nombre de semaines où un utilisateur a *tweeté* dans un ensemble de lieux associés à une même activité. Nous partons ensuite du principe que le lieu de type « Domicile » ou « Activité principale » sont visités toutes les semaines, et posons que le nombre de semaines de référence n'est plus le nombre de semaines d'activités totale mais W_{ref}, soit le nombre de semaines minimum de fréquentation du lieu de domicile ou à l'activité principale. Nous divisons ensuite le nombre de semaines associées W_a à une activité a par le nombre de semaines de référence, puis nous bornons à 100 %. Une activité sera réalisée une semaine donnée selon cette nouvelle probabilité, et, les activités où des traces numériques avaient été enregistrés lors d'un plus grand nombre de semaines différentes que les lieux de domiciles ou d'activité principale, seront réalisées toutes les semaines en les considérons donc comme des activités très routinières.

Si nous prenons l'exemple présent dans la figure 242 ci-dessous, l'activité A est réalisée dans le lieu 1 les semaines 1, 2 et 4 et dans le lieu 2 la semaine 3, soit lors de 4 semaines différentes. De même l'activité B est réalisée lors de 2 semaines (1 et 2) et l'activité principale est visitée dans 3 semaines différentes et le domicile fréquenté lors des 4 semaines. Nous recalculons ensuite les probabilités de réaliser une activité une semaine donnée, en divisant la somme des semaines différentes pour chaque activité par le nombre de semaines différentes minimum entre le domicile et l'activité principale (ici 3), et nous bornons.

Lieu	Activité	Semaine
1	A	{1,2,4}
2	A	{3}
3	B	{1}
4	B	{2}
5	Domicile	{1,2,3,4}
6	Activité Principale	{1,2,4}

Activité	Semaines différentes	Probabilité
A	4	1 (4/3)
B	2	2/3
Domicile	4	1 (4/3)
Activité Principale	3	1 (3/3)

FIGURE 242 Principe de la démarche pour affecter une probabilité de réaliser une activité une semaine donnée. Le tableau de gauche représente les semaines où sont réalisées des activités (A,B, domicile, activité principale), dans des lieux (de 1 à 6). Le tableau de droite montre les probabilités de réaliser une de ces activités une semaine donnée.

1.2 Nombre de lieux associés a une activité

En admettant qu'une activité autre que le domicile ou l'activité principale soit sélectionnée pour être réalisée une semaine w donnée (*e.g.* l'activité A), il convient de définir le(s) lieu(x) qui seront effectivement visités (lieu 1 ou lieu 2, ou les deux ?).

Pour ce faire, nous définissons une probabilité $P_{l,a}$ que le lieu l soit fréquenté comme

étant le rapport entre le nombre de semaines W_l ou le lieu l a été visité divisé par le nombre de semaines total W_a où l'activité a a été réalisée.

$$P_{l,a} = \frac{W_{l,a}}{W_a} \tag{29}$$

Pour chaque lieu, nous appliquons tirage booléen, ou la probabilité de visite vaut $P_{l,a}$ et celle de non réalisation vaut $1\text{-}P_{l,a}$. Les lieux 1 et 2 ont alors respectivement une probabilité de 3/4 et 1/4 d'être sélectionnée.

Dans le cas où les probabilités sont très faibles (un grand nombre de lieux associés à une même activité mais visité ponctuellement) il est possible qu'aucun lieu ne soit tiré au sort. Dans ce cas nous sélectionnons un lieu l parmi l'ensemble des lieux L_a fréquentés d'une même activité a selon leur probabilité associée $P_{l,a}$.

À la fin de cette étape, nous obtenons une liste de lieux associés à un type d'activité qui seront visités une semaine donnée.

1.2.2 Étape 2 : Définir les jours de visite

2.1 Nombre de jours dans une semaine

Un même lieu peut être visité plusieurs fois dans une même semaine. Nous posons aussi que l'activité principale se déroule les jours de la semaine où des *tweets* y ont été enregistrés. Mais il convient de définir le nombre de jours où les autres lieux seront fréquentés.

Nous connaissons le nombre de jours différents où un individu a *tweeté* depuis chaque lieu de son espace d'activité. Dans le chapitre 8, nous posions que la probabilité qu'un lieu soit visité n jours correspondait au rapport entre le nombre de jours ou des *tweets* étaient enregistrés dans ce lieu par le nombre de jours ou l'utilisateur avait *tweeté*, multiplié par 7 (pour avoir une fréquence hebdomadaire). Nous présentons ici une variante, en partant du principe que le domicile est visité tous les jours de la semaine et définissons alors une valeur de référence D_{ref} comme étant le nombre de jours où un utilisateur a *tweeté* depuis son domicile. Nous divisons ensuite le nombre de jours D_l où un utilisateur a *tweeté* dans chacun de ces lieux l par cette valeur de référence, et multiplions par 7.

$$D'_l = 7 \times \frac{D_l}{D_{ref}} \tag{30}$$

Si une fréquence est inférieure à 1, nous lui affectons 1, afin que le lieu soit visité au moins une fois dans la semaine. Si la fréquence est supérieure à 7, nous lui affectons 7. Nous obtenons ainsi pour chaque lieu une fréquence entre et 1 et 7 que nous posons comme étant le

nombre de jours où le lieu sera fréquenté une semaine donnée. Comme dans le chapitre 8, nous arrondissons de manière aléatoire à l'unité inférieure ou supérieure, afin d'obtenir un nombre entier de jours fréquentés, valable pour une semaine donnée.

○ 2.2 Choix des jours

Chaque lieu se voit donc affecté un nombre de jours D'_l où il sera visité une semaine donnée. Nous choisissons ensuite les jours de la semaine selon le nombre de fois où l'utilisateur a *tweeté* un jour donné dans ce lieu en appliquant un tirage pondéré sans remise. Si par exemple nous devons choisir deux jours dans un lieu où des traces numériques ont été enregistrées les lundis, mardis et mercredis, respectivement dans 10 %, 40 % et 50 % des cas, le mercredi et mardi auront plus de chance d'être tirés que le lundi. Si le nombre de jours défini à l'étape précédente est supérieur au nombre de jours différents où un utilisateur a *tweeté* dans un lieu donné, dans ce cas nous sélectionnons tous les jours présents. Dans l'exemple précédent, si nous devions choisir 4 jours, nous n'en prendrions que 3 (lundi, mardi et mercredi). Nous ne désirons pas pour l'instant affecter plusieurs fois le même lieu un jour donné, pour des raisons de simplicité.

À la fin de cette étape, chaque utilisateur s'est vu attribuer une liste de lieux qu'il fréquentera à des jours définis. Reste maintenant à estimer les heures de réalisation.

1.2.3 Étape 3 : Définir les heures de réalisation

Comme nous l'avons vu et répété plusieurs fois dans cette thèse, les traces numériques de *Twitter* enregistrées dans un lieu sont de nature épisodique, c'est-à-dire non-continues dans le temps. La première partie de cette étape consiste donc (encore) à rendre continue ces données sur des plages horaires et à sélectionner celle où l'activité se réalisera.

○ 3.1 Choix d'une plage horaire

Pour rappel, et comme illustré par la figure 243, nous appliquons dans un premier temps une fonction de densité (traits) sur les fréquences de distribution des messages (histogramme). Nous définissons ensuite des groupes de plages horaires selon leur continuité temporelle définie par la fonction de densité (couleurs). La probabilité de tirer une plage horaire est alors égale à la somme des valeurs de densités enregistrées par groupes (˜intégrale). Dans l'exemple suivant, le groupe 1 à 40 % de chance d'être tiré, contre 60 % pour le groupe 2.

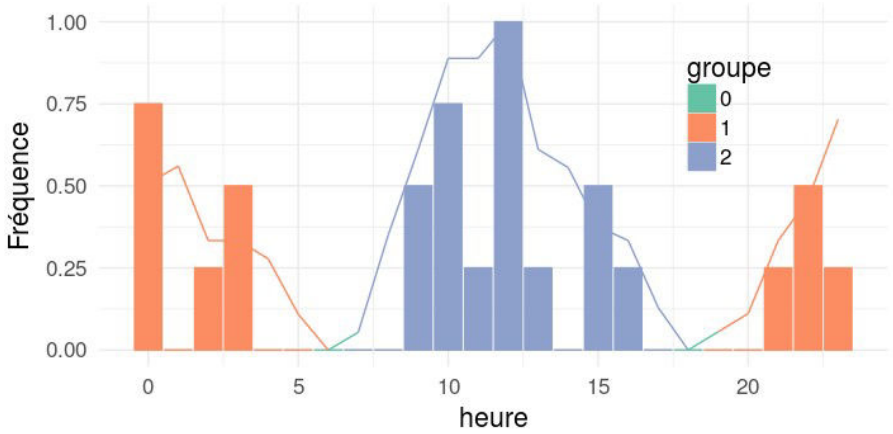


FIGURE 243 Principe de la définition et du choix des plages horaires. La fréquence d'envoi de messages à une plage horaire est illustrée par les histogrammes. Les traits correspondent à l'extension de ces plages horaires, d'après une fonction de densité.

3.2 Définition d'une durée

Une fois qu'un groupe de plage horaire a été sélectionné, par exemple le groupe 2, il nous reste à définir la durée de l'activité. Nous pourrions prendre l'ensemble de la plage horaire tirée et appliquer ensuite de nombreux ajustement, basés notamment sur la distance au domicile, comme dans le chapitre 7. Mais nous en profitons ici pour tester de nouvelles méthodes.

Car finalement, est-ce que l'ensemble de la plage horaire correspond réellement à la durée d'une activité ? Dans certains cas, comme pour les activités figées dans le temps et répétées régulièrement, comme l'activité principale, cela est tout à fait possible. Mais *quid* d'autres activités plus flexibles dans le temps, ou un utilisateur aurait globalement *tweeté* lors d'un grand nombre d'heures différentes, mais ne fréquenterait ce lieu qu'une courte durée ? Par exemple une personne peut avoir *tweeté* dans un *mall* le matin, puis un autre jour l'après midi, et enfin une fois le midi. L'approche par plage horaire tend ici à augmenter la durée de présence, alors qu'il est possible que la personne ne reste à chaque fois qu'un court moment dans ce lieu. Nous distinguerons donc ici les activités principales des autres types d'activités.

Pour les activités autres que l'activité principale

Pour ce qui est des activités autres que l'activité principale, nous allons simplement tirer deux tranches horaires, avec une probabilité définie par la fonction de densité. La figure 244 illustre les différentes plages horaires (en rouge) obtenues selon les tirages. Nous obtenons ainsi une heure d'arrivée et de départ.

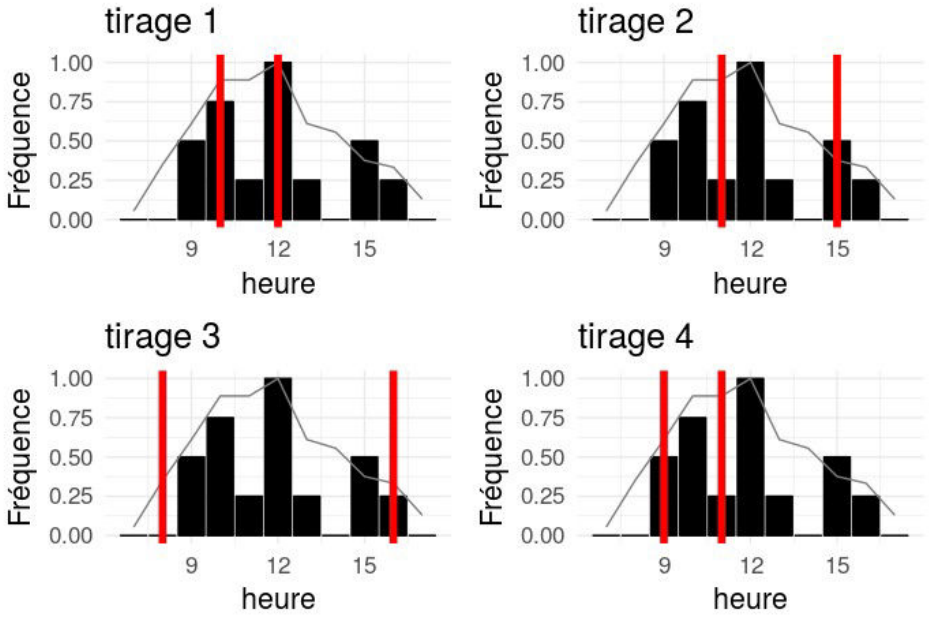


FIGURE 244 Illustration de la méthode employée pour tirer une heure de début et de fin d'activité (hors activité principale). Exemple avec 4 tirages pour un même profil de fréquentation d'un lieu donné.

- **Activité Principale**

Concernant l'activité principale, nous posons quelques hypothèses supplémentaires.

- Tout d'abord que cette activité ne commence pas avant 5 h du matin. Ensuite qu'elle dure entre 6 h et 14 h (spectre qui pourra être redéfini par la suite).
- Nous découpons ensuite la plage horaire en 3 (tierciles) selon des intervalles égaux.
- Nous posons que l'heure d'arrivée se situe dans le premier tiercile, et l'heure de départ dans le dernier.
- Nous tirons tout d'abord une heure d'arrivée, selon une probabilité définie par la fonction de densité.
- Nous appliquons ensuite des contraintes à l'heure de départ (entre 6 h et 14 h de plus que l'heure d'arrivée), et tirons une heure de fin d'activité, toujours selon une probabilité définie par la fonction de densité.

Il est toutefois possible, selon les cas de figure, qu'aucune heure de départ ne corresponde à ces critères.

- Si la plage horaire est trop étroite (inférieure à 6 h), nous tirons une durée entre 6 h et

8 h que nous ajoutons à l'heure d'arrivée pour obtenir l'heure de départ.

- Si la plage horaire est trop large, nous posons, de manière arbitraire, que l'heure de départ intervient 12 h après l'heure d'arrivée.

La figure 245 présente différents tirages d'heure de départ et d'arrivée, selon la méthode évoquée. La durée de l'activité peut parfois être assez courte (tirage 1), parfois très longue (tirage 2 & 5) mais toujours comprise entre 6 h et 14 h. À la fin de cette étape, nous obtenons pour chaque lieu une heure d'arrivée, une heure de départ, et donc une durée.

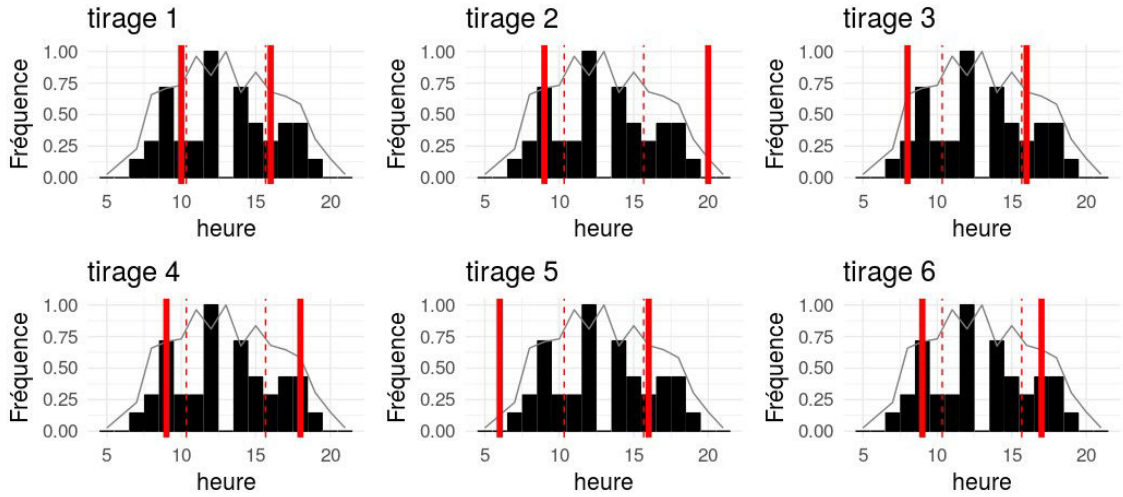


FIGURE 245 Illustration de la méthode employée pour tirer une heure de début et de fin pour l'activité principale. Exemple avec 6 tirages pour un même profil de fréquentation. Les pointillés rouges indiquent les tierciles, et les lignes rouges les heures sélectionnées selon les tirages.

1.2.4 Étape 4 : agglomération et gestion des doublons

Les étapes précédentes ont permis d'obtenir pour chaque lieu une semaine, un jour et une plage horaire de réalisation. Néanmoins, comme vu dans les chapitres 7 et 8, il est très fréquent que des « conflits » surviennent, dans le sens où plusieurs lieux peuvent être fréquentés en même temps ce que nous appellerons des doublons. Nous proposons ici une méthode simple, rapide, mais discutable, basée sur les types de lieux fréquentés, les tranches horaires et les fréquences de réalisations.

Ainsi, lorsqu'il y aura des doublons, la sélection du lieu se fera en tirant un des lieux avec une probabilité qui dépend de la part de jours de *tweets* J_l envoyés depuis le lieu par rapport à l'ensemble des jours où l'utilisateur a été actif sur *Twitter* J_{tot}, avec cependant quelques considérations.

Nous partons du principe que l'activité principale est plus figée dans le temps que les autres

activités. Mais un étudiant a probablement un emploi du temps plus flexible qu'une personne exerçant un autre type d'activité principale. Ainsi, pour les activités principales correspondant à des lieux d'éducation, nous doublons la probabilité de réalisation évoquée précédemment, tandis que nous la quadruplons pour les autres activités principales. Ces choix (doubler ou quadrupler) sont faits de manière assez arbitraire, après des observations et essais empiriques.

Néanmoins, l'activité principale n'est pas aussi figée dans le temps à tous les moments de la journée. En effet, il est possible d'embaucher plus tard ou de débaucher plus tôt, et aussi de quitter son lieu de travail aux heures des repas. Ainsi, nous posons que les heures de départs et d'arrivées peuvent être plus flexibles, tout comme l'heure du déjeuner (12 h et 13 h), et nous baissons à ces heures la probabilité de sélectionner ces lieux de 50 %. Enfin, à chaque plage horaire où des doublons existent, nous tirons un lieu parmi ceux en concurrence avec la probabilité que nous venons de définir. Finalement, à tous les horaires où aucune activité n'est réalisée, nous posons que l'utilisateur est à son domicile.

Cet algorithme n'est pour l'instant pas tout à fait optimisé d'un point de vue des temps de calculs[407] et il faut compter environ 8h pour reconstituer les 17 368 agendas de notre échantillon sur 4 semaines sur un ordinateur de bureau classique.

1.3 Résultats

Nous présentons dans cette section les principaux résultats issus de nos agendas reconstitués (AR) des 17 368 utilisateurs de *Twitter* de notre échantillon, notamment sur le déroulement temporel des activités et sur la répartition spatiale des personnes dans la ville. D'autres statistiques (nombre de lieux fréquentés, durée des activités, etc.) seront présentées dans les sections suivantes, relatives à la génération d'agendas (AG) à partir de ces AR.

Temporalités

Une lecture agrégée sur une semaine des présences horaires selon l'activité (figure 246) permet d'avoir une bonne vue d'ensemble des résultats obtenus. Tout d'abord, nous pouvons noter que la présence à l'activité principale suit une courbe, proche d'une gaussienne, centrée sur midi, avec une partie droite (heure de départ) plus lâche que la partie gauche (arrivée). Les week-ends, cette activité est moins réalisée que les jours de semaine. Une part non négligeable et relativement constante des utilisateurs resteraient chez eux durant les heures ouvrées.

Si nous regardons maintenant les autres activités, nous pouvons noter un pic de fréquentation entre 12 h et 13 h, lié à notre algorithme qui favorise la présence dans d'autres activités que l'activité principale à ces heures de la journée. Sinon, un pic d'autres activités est

407. Il conviendrait de réduire l'usage de boucle, coûteuse en temps.

à noter après 18 h les jours de semaines, moins marqué le week-end. En effet, les samedi et dimanche, les autres activités sont visitées plus tôt dans l'après midi.

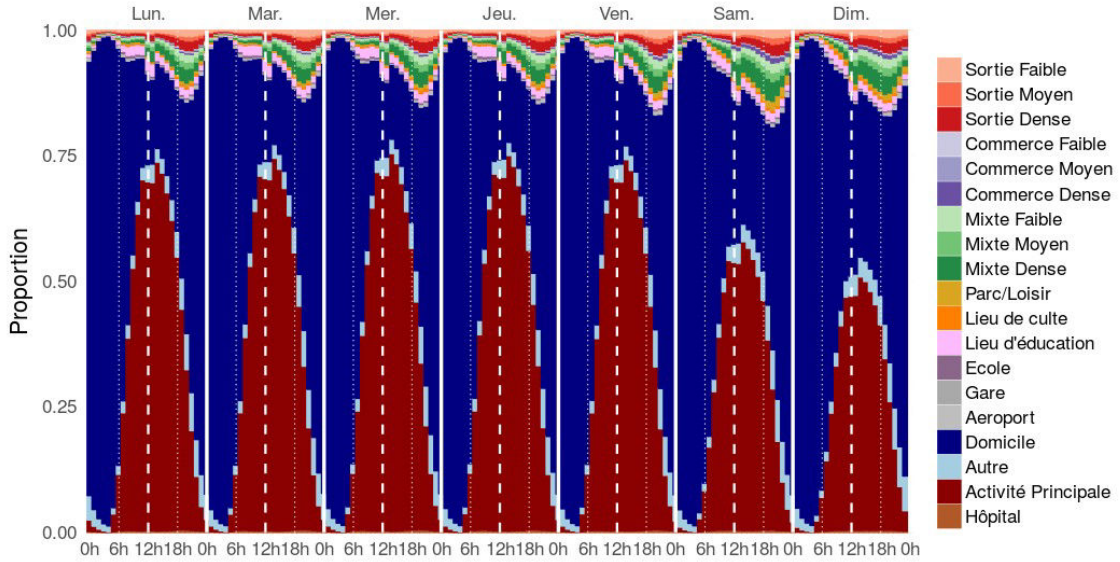


FIGURE 246 Proportion des utilisateurs effectuant une activité par tranche horaire sur une semaine. D'après les agendas reconstitués. Les lignes blanches indiquent minuit, les tirets midi et les pointillés 6h et 18h.

Optons maintenant pour une lecture séparée des différentes temporalités des activités, visible dans les figures 247 et 248 [408]. Globalement, les lieux de type commerce faible et moyen sont peu visités, mais présentent un pic important après 18 h. Les commerces denses en revanche sont nettement plus fréquentés les vendredi soir et les durant les week-end. Pour rappel, le marché du week-end de Chatuchak appartient à cette catégorie. Les lieux de sorties ont plus ou moins le même profil, à savoir un pic très marqué le soir, notamment les vendredis et samedis, et un petit pic les matins et les midis. Les lieux de type mixte sont surtout fréquentés en fin d'après midi, et plus les week-ends que les jours de semaines.

Si nous regardons maintenant les autres activités (figure 248), nous pouvons noter une grande stabilité des présences aux domiciles et aux activités principales, du fait de leur effectif important qui tend à lisser les fréquences de visites.

Les visites de parcs se font principalement le matin et surtout en fin d'après midi, notamment les vendredi, samedi et dimanche. Nous observons deux pics sur les profils de fréquentation des lieux de cultes, le matin et le soir, même s'ils restent inégaux selon les jours de semaine. Le week-end, ces lieux sont nettement plus visités, notamment le dimanche matin.

408. Celles ci sont agrégées à la semaine comme les agendas sont générés sur un mois, il faut donc diviser les effectifs par 4 pour avoir les valeurs de fréquentation moyennes.

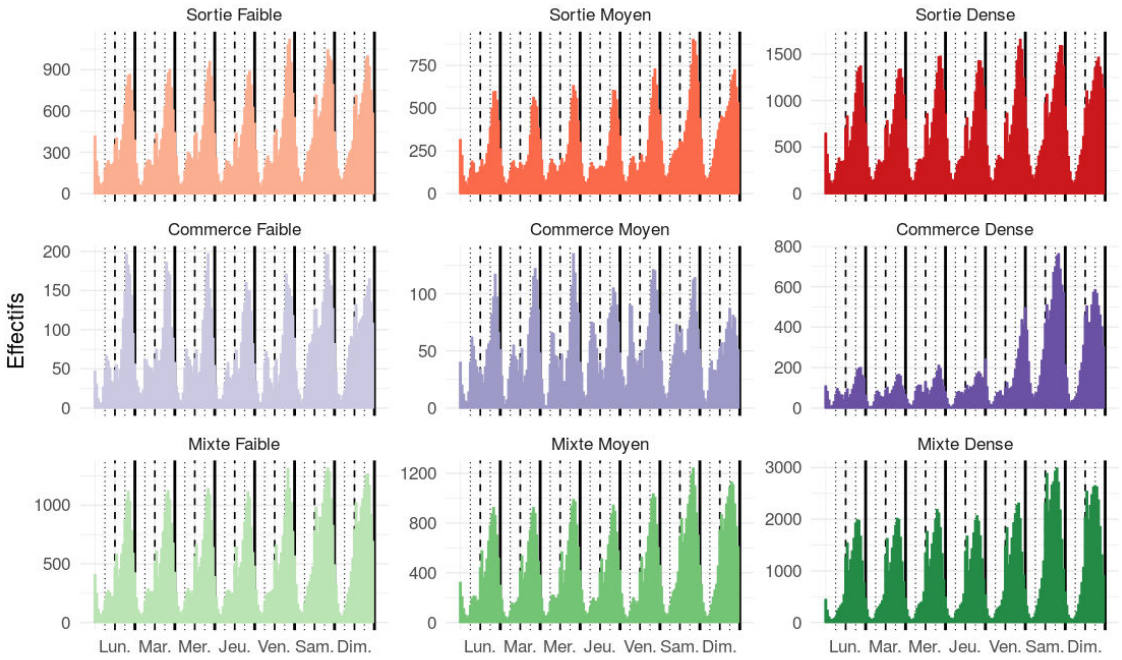


FIGURE 247 fréquentations horaires des lieux de type commerce, sortie et mixte

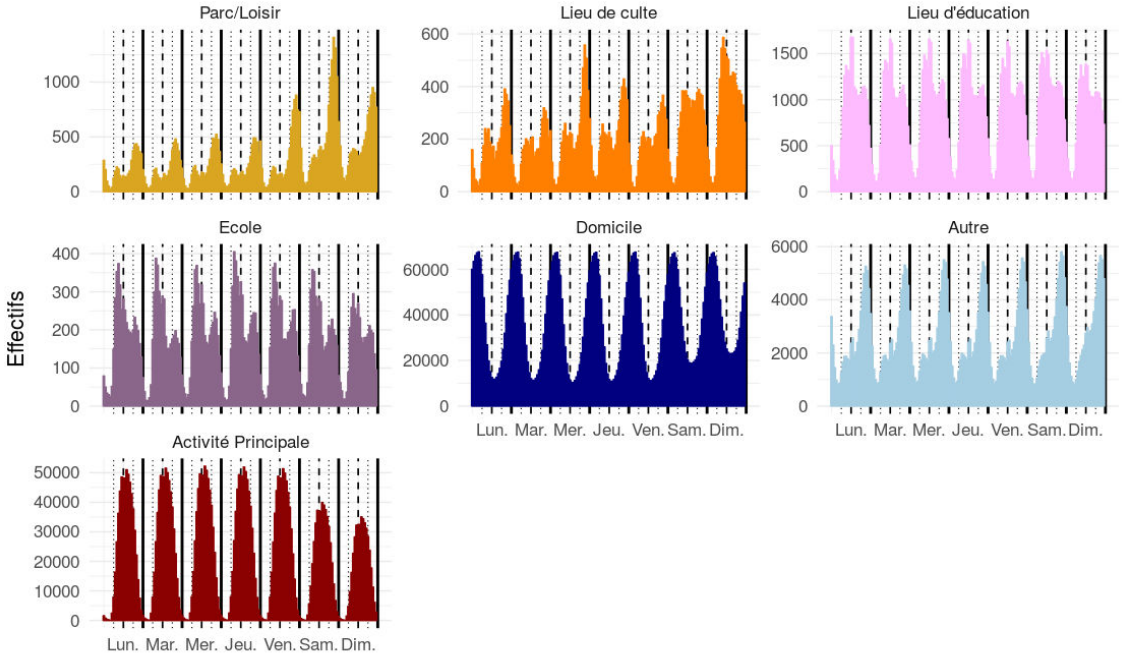


FIGURE 248 Fréquentation horaire des autres types de lieux

Les écoles ou autres lieux d'éducations présentent des profils de fréquentations similaires, à savoir un plus grand nombre de visites le matin que l'après midi. Néanmoins l'absence de différence marquée entre les jours de semaine et de week-end suggère quelques biais dans notre

algorithme. Finalement, les lieux que nous n'avons pas pu définir de type autre sont surtout fréquentés le soir.

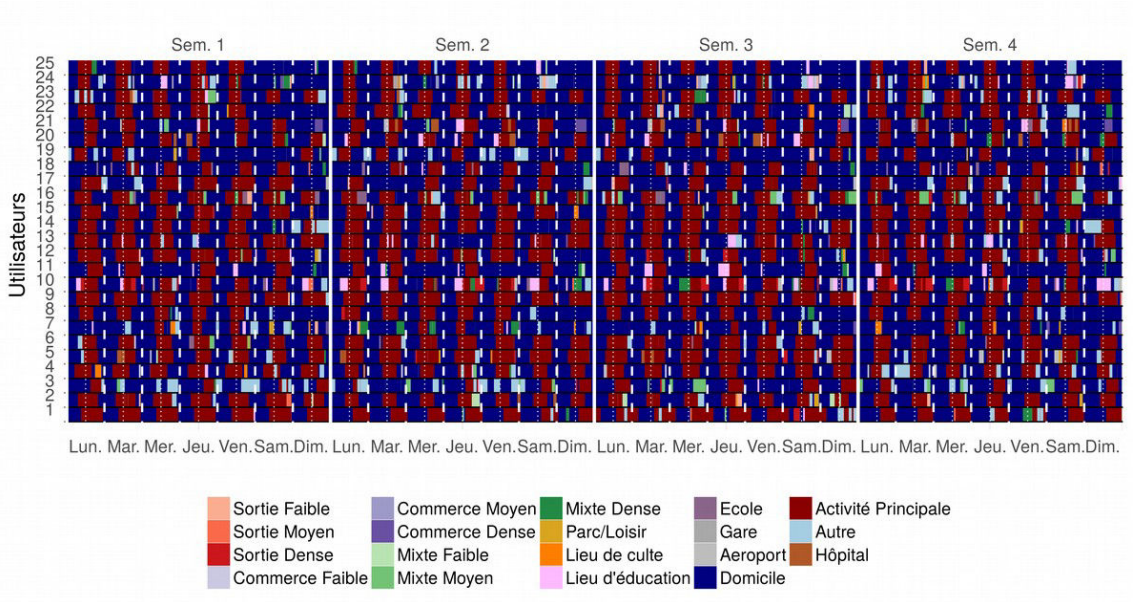


FIGURE 249 Agendas reconstitués de 25 utilisateurs sur un mois

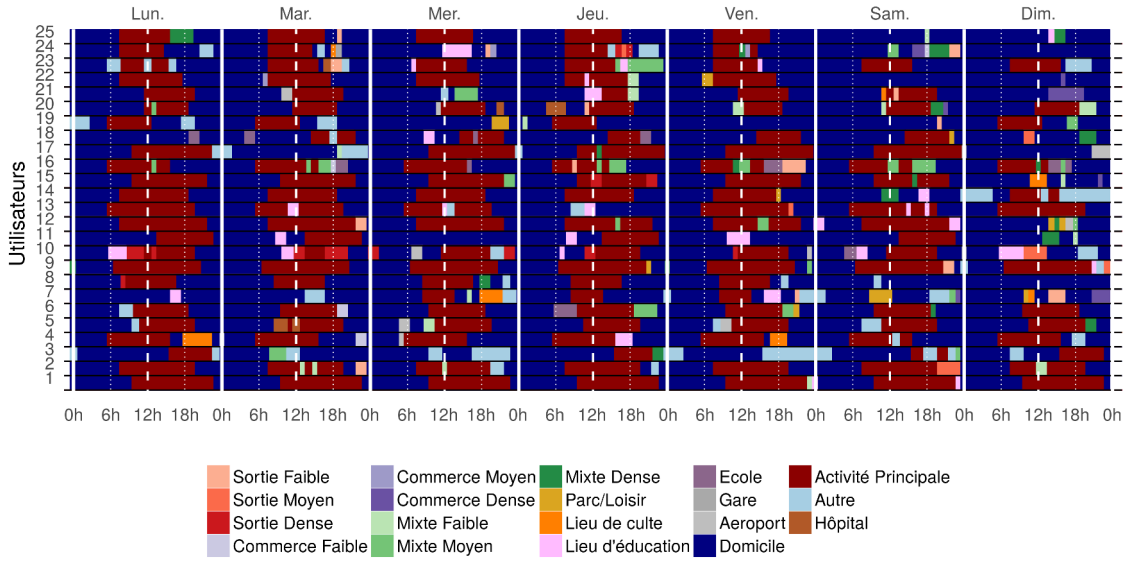


FIGURE 250 Agendas reconstitués de 25 utilisateurs pour une semaine.

Les figures 249 et 250 montrent maintenant des agendas pour 25 utilisateurs, sur une période d'un mois (figure 249) et une semaine (figure 250). N'ayant aucun point de comparaison, nous pouvons simplement dire qu'ils nous paraissent assez cohérents. Certains utilisateurs ont plus d'activités que d'autres certains ont des semaines très chargées, d'autres plus basiques.

Certains travaillent plus longtemps et sur plus de jours, d'autres ont une activité principale effectuée moins régulièrement, mais toujours aux mêmes tranches horaires. Les séquences entre les semaines de chacun sont assez proches pour ce qui est de l'activité principale, mais présentent des variations dans la visite de certains types de lieux.

Si nous regardons spécifiquement quelques agendas, l'utilisateur 25 travaille par exemple du lundi au vendredi, entre 7 h et 17 h. Le lundi il quitte son activité principale un peu plus tôt pour se rendre dans une zone mixte dense. Il passa une heure le mardi vers 19 h dans un lieu de sortie faible. Les autres jours il ne fit que navetter entre son domicile et son travail, mais sortit quelques heures le dimanche après midi, dans des lieux de type « mixte dense » ou de « sortie faible ». Les utilisateurs 1 et 2, à contrario, travaillent 7 jours sur 7, sur de longues périodes et ne fréquentent que quelques lieux après leur travail.

Répartition spatiale

Chaque lieu de l'agenda reconstitué d'un utilisateur est ici associé à une localisation, et les figures 251 et 252 ci-dessous montre la répartition de notre échantillon dans la ville les mardi et samedi, à 6 h, midi, 18 h et minuit. La figure 252 étant un zoom de la figure 251 sur le centre de Bangkok. Les résultats obtenus sont très cohérents avec ceux vus dans le chapitre précédent. Nous pouvons noter encore un effet de type centre périphérie, avec des présences mieux réparties la nuit qu'en journée, du fait de la présence des utilisateurs à leur domicile. Les activités en journées sont principalement situées dans le centre, et dans les différents pôles secondaires à l'est et au nord-est. En revanche, le samedi vers 18 h, les utilisateurs sont d'autant plus regroupés dans quelques pôles, principalement dans le quartier des *malls* (Siam Paragon, Central World, etc.).

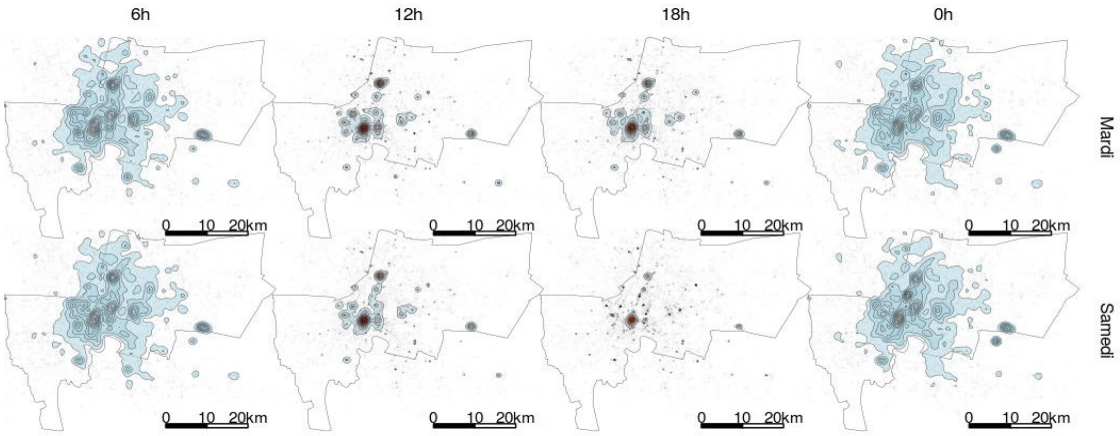


FIGURE 251 Répartition des utilisateurs selon leurs agendas reconstitués à différents moments de la journée, un mardi et un samedi. D'après une fonction de densité de portée 2.5 km.

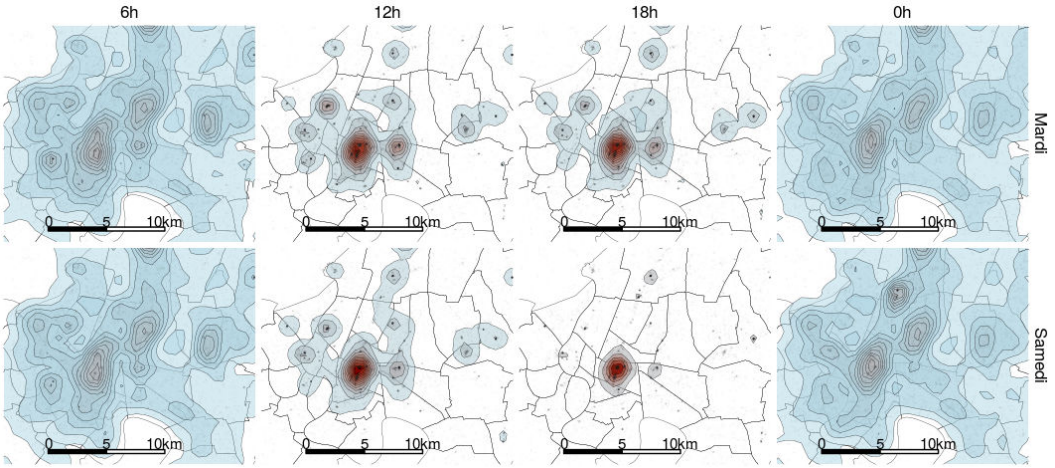


FIGURE 252 Répartition des utilisateurs selon leurs agendas reconstitués à différents moments de la journée, un mardi et un samedi, dans le centre de Bangkok. D'après une fonction de densité de portée 2.5 km.

Le calcul de la distance Venable[409]. (figure 253) sur des jours de semaine type montre la tendance à la concentration dans la ville qui débute vers 5h du matin, pour atteindre un maximum entre 12 h et 16 h. Après cette période, les personnes se dispersent plus progressivement qu'elles ne se sont concentrées le matin (pente plus faible), ce qui peut s'expliquer par le fait que les personnes ne rentrent pas nécessairement directement chez elles le soir.

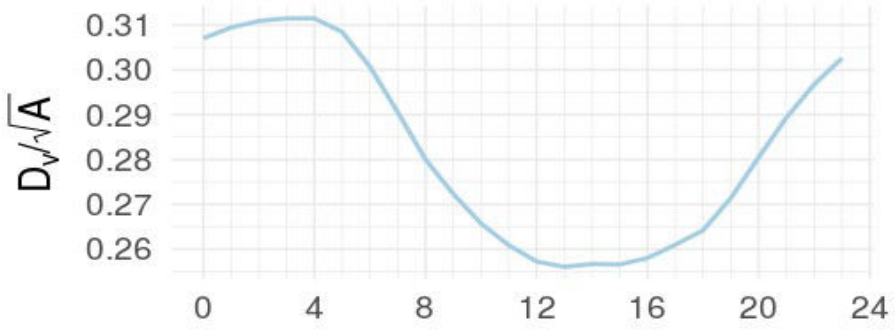


FIGURE 253 Distance "Venable" à Bangkok, d'après nos agendas reconstitués.

Ce constat peut également être fait à partir des données trafic, ou le pic d'embouteillage du matin est moins marqué que celui du soir (chapitre 10 et figure 254 ci-dessous). Le recours à ces données trafic peut aussi permettre d'avoir un point de référence pour savoir si les horaires auxquels les personnes de notre échantillon quittent leur domicile sont crédibles.

409. Si nous comparons l'amplitude de la courbe aux valeurs obtenues par (Louail *et al.*, 2015a) pour des villes espagnoles, cette dernière est très importante, ce qui suggère ici une tendance monocentrique très prononcée même si quelques pôles secondaires sont présents dans la ville

La distribution des fréquences horaires des surfaces totalement embouteillées (figure 254) est bimodale. Il est donc possible d'extraire les pics du matin ou du soir, en utilisant des méthodes de décomposition de signaux. Pour ce faire, nous utilisons l'algorithme itératif de maximisation d'espérance (EM-Algorihtm) (Hartley, 1958 ; Okafor, 1987) implémenté dans R dans la librairie "mixtools" (Benaglia *et al.*, 2009). La fonction normalmixEM permet de décomposer des séries multi-modales en plusieurs séries normales. La figure 254.a présente une décomposition en 2 séries normales. Si la courbe bleue parait être bien en accord avec le pic du soir, la rouge n'est pas bien calée sur le pic du matin. En revanche, une décomposition en 3 distributions normales 254.b, fait bien ressortir les pics du matin (rouge) et du soir (bleue), malgré une amplitude plus réduite.

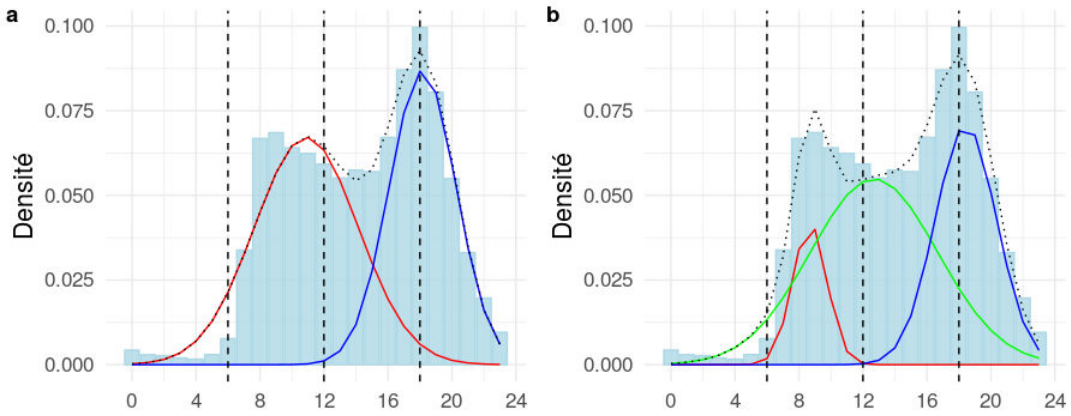


FIGURE 254 Décomposition de la distribution de la part horaire des superficies de la ville occupées par des zones de trafic très dense, selon l'algorithme EM. Décomposition en 2 (a) et 3 (b) courbes normales. La courbe en pointillé correspond à la somme des courbes rouges et bleues (a) et vertes (b).

Prenons maintenant la courbe rouge de la figure 254.b et comparons-la avec la distribution des heures de départ du domicile calculée d'après nos agendas (figure 255). La partie gauche de la courbe rouge nous intéresse plus que celle de droite car elle décrit mieux l'intensification du trafic le matin à partir de 7h, pour atteindre un maximum à 9h. La courbe qui représente les départs du domicile commence à croître un peu plus tôt, mais atteint aussi un maximum à 9 h. D'ailleurs, cette dernière suit une distribution à peu près normale et englobe bien la distribution du trafic le matin, avec une largeur à mi-hauteur double.

La confrontation de ces courbes fait apparaître un lien entre les heures de départ du domicile issues de nos agendas reconstitués et l'importance des embouteillages. Ces deniers commencent environ 2 h après les premiers départs (7 h) et s'accélèrent avec l'intensification des départs du domicile, le tout coïncidant avec un pic identique à 9 h. Ce constat permet de valider indirectement la cohérence d'une partie de nos agendas, à savoir le départ du domicile

et le début de la réalisation d'autres activités dans la ville.

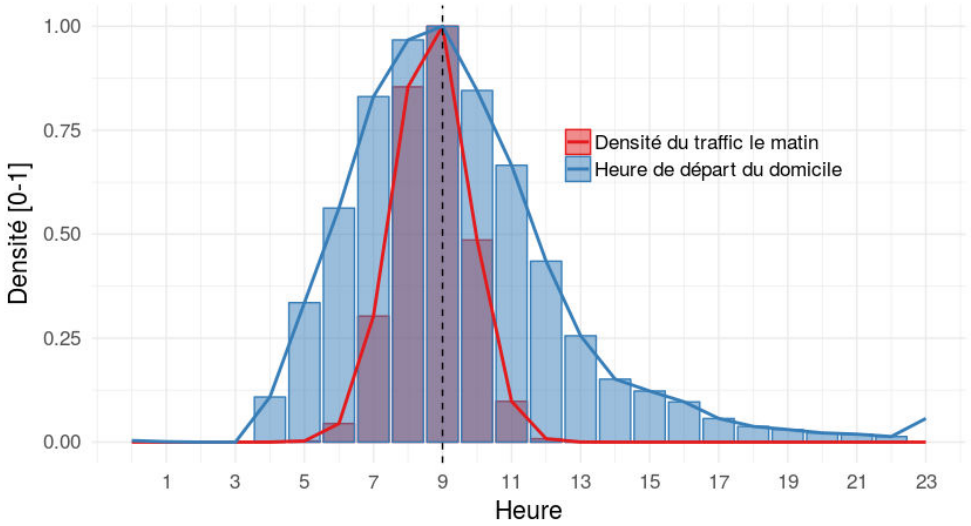


FIGURE 255 Comparaison des heures de départ du domicile et de l'intensité du trafic le matin.

À partir de ces agendas reconstitués, nous allons maintenant en générer de nouveaux, toujours en nous basant sur une approche similaire à celle développée dans le chapitre 8. Mais là encore, nous appliquerons quelques variantes, ne serait-ce que pour montrer qu'il y a énormément de manières de procéder et de paramètres à ajuster.

2 Génération d'agendas

2.1 Définir des groupes d'utilisateurs

Les agendas que nous avons reconstitués sont très différents selon les individus. Si nous devons nous servir de ces informations pour générer des agendas synthétiques, il nous parait pertinent de définir des groupes d'utilisateurs, selon leurs tendances et propriétés de visite de lieux et de déplacements.

Nous pourrions ici faire des groupes selon les séquences d'activités de leur agenda, en utilisant des méthodes d'appariement optimal, comme vu précédemment (chapitres 7 et 8). Mais les temps de calcul d'une matrice de distance entre des séquences de 672 heures (pour un mois) pour plus de 17 000 individus étaient un peu longs par rapport au temps qu'il nous restait pour finir ce travail. De plus, il est fort probable que l'aspect le plus discriminant entre les groupes soit la fréquence de réalisation hebdomadaire de l'activité principale, comme dans le chapitre 8, ce qui n'est pas une information primordiale dans le cadre d'une génération d'agendas. Il conviendra néanmoins de tester cette méthode ultérieurement.

Nous ne partons pas ici du principe que chaque personne a un capital de mobilité sujet très discutable (Borja *et al.*, 2015) et qui sort de notre domaine de compétence - mais que chaque individu peut être décrit sommairement par un potentiel de mobilité qui dépend d'énormément de facteurs (chapitre 3). Nous allons ici simplement créer des groupes d'utilisateurs, selon quelques indicateurs de potentiels de déplacement.

Banos *et al.* (2006), ont pu définir des groupes plutôt riches d'un point de vue sémantique, car ils ont utilisé des données issues d'enquêtes institutionnelles sur les mobilités qui répertoriaient les catégories socio-économiques et démographiques de l'échantillon. Malheureusement, nous n'avons pas d'informations aussi précises sur notre échantillon, mise à part une distinction entre « Etudiants » (ou assimilés) et « Autre ». Ici, nous reprendrons simplement la méthode déjà présentée dans Cebeillac *et al.* (2017), où chaque individu est décrit par un nombre de lieux qu'il visite, associé à des potentiels de dispersions comme la distance du trajet moyen et le rayon de giration (chapitres 5, 7 et 8). Ces indicateurs sont considérés comme les éléments de bases permettant une modélisation raisonnable des mobilités humaines (González *et al.*, 2008), auxquels nous ajoutons un nombre d'activités effectuées.

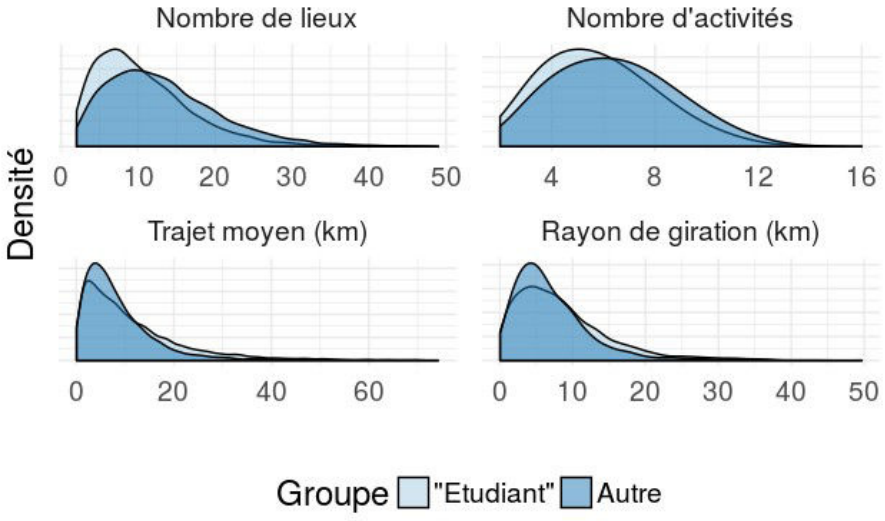


FIGURE 256 Comparaison des potentiels de dispersions selon les étudiants et le reste de l'échantillon

Si nous regardons rapidement la distribution des effectifs (sous forme de fonction de densité) de nos deux classes ("Étudiants" et "Autre") sur chacun de ces quatre critères (figure 256, ci-dessus), nous pouvons noter que globalement, les étudiants fréquentent moins de lieux et effectuent un peu moins d'activités. Une grande proportion d'entre eux effectuent des trajets moyens relativement courts, mais une partie a toutefois tendance à se déplacer sur de plus grandes distances. Mais plutôt que de se focaliser sur ces deux groupes, nous décidons d'effectuer

des classes utilisant un algorithme de classification non-dirigée, le K-means, afin de faire des regroupements selon les potentiels de dispersions.

Nous pouvons aussi noter une certaine redondance de l'information si nous regardons les liens entre nos différents indicateurs. Le nombre de lieux qu'un individu fréquente est en effet très corrélé avec le nombre d'activités qu'il exerce (figure 257, gauche), tout comme un rayon de giration élevé va en général de pair avec un temps de trajet moyen important (figure 257, droite). Cela dit, ces relations ne sont pas si linéaires que cela et les niveaux de dispersions sont assez importants. Et la redondance de l'information entre le nombre d'activités et de lieux fréquenté et la distance moyenne d'un trajet et le rayon de giration devraient tendre à fournir des classes plus stables statistiquement.

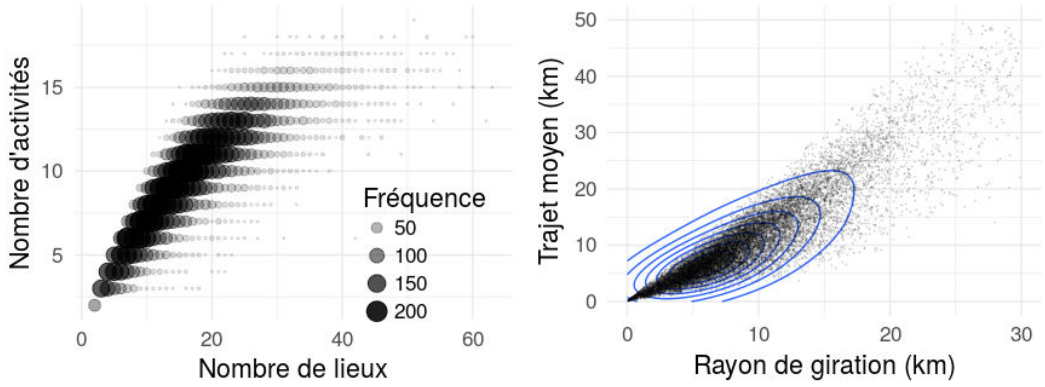


FIGURE 257 Lien entre le nombre de lieux fréquentés et le nombre d'activités effectuées (gauche) et entre le rayon de giration et le temps de trajet moyen.

Dans un premier temps, compte tenu que les unités et les amplitudes sont différentes selon les indicateurs, nous décidons de centrer et de réduire les données de chaque individu. Nous définissons ensuite un nombre de classes en calculant la variation des sommes aux carrées au sein de clusters (within-cluster sum of square, ou WSS). Nous choisissons 5 classes du fait de la présence d'un "coude" dans la distribution des WSS à ce nombre de classe (annexe L).

Caractéristiques des groupes d'utilisateurs

Nous regroupons ensuite ces utilisateurs par classes, selon le *khet* (district) de domicile, afin d'avoir un aperçu de la répartition spatiale des membres de chaque groupe (figure 258), et d'observer dans quelles mesures le lieu de domicile peut influencer les potentiels de déplacements.

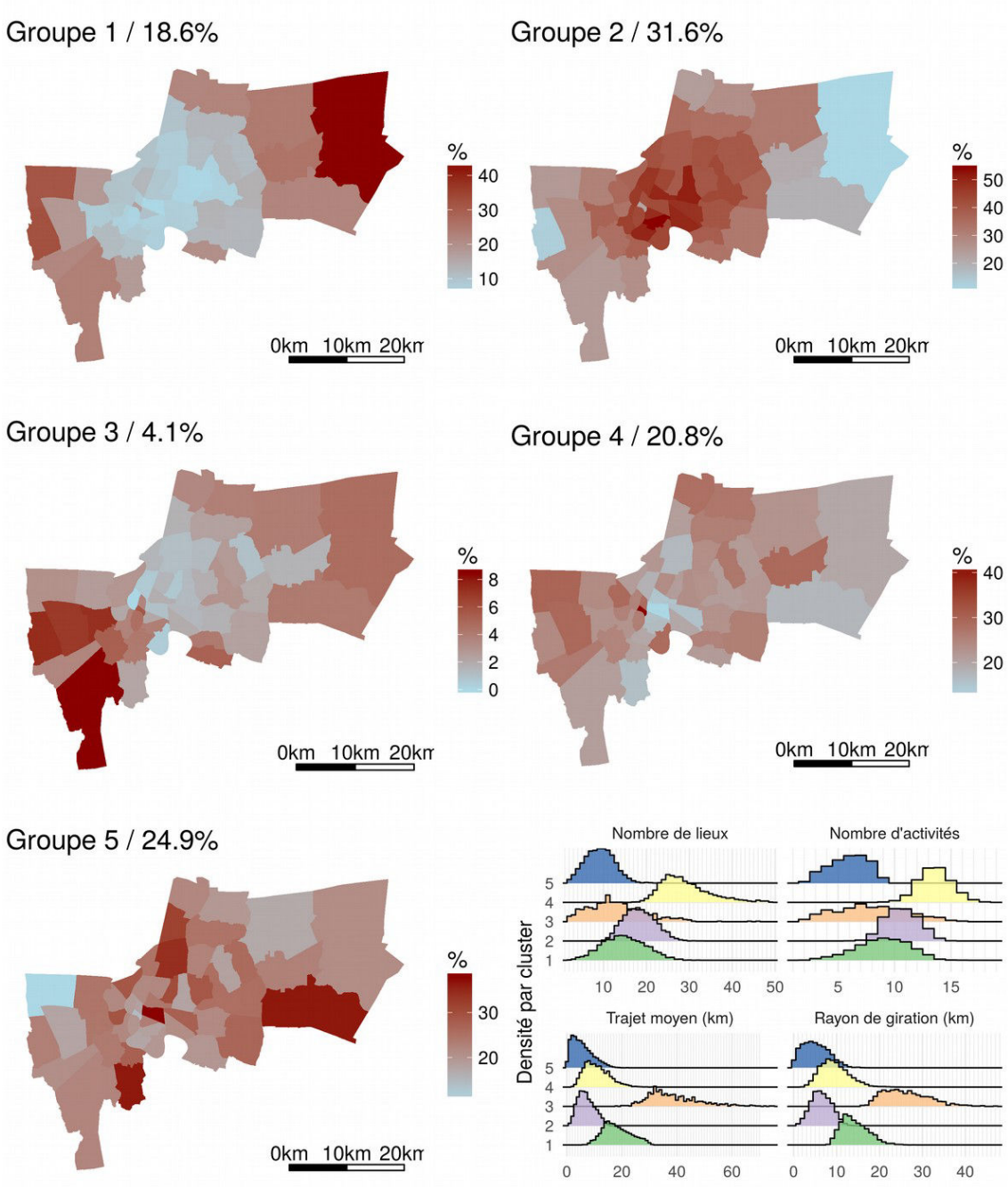


FIGURE 258 Répartition de la part de chaque groupe obtenu précédemment dans les districts de la ville. L'encadré en bas à droite montre les distributions des effectifs selon les différents critères de dispersions, par groupe.

- **Classe 1 :**

18,6 % des utilisateurs appartiennent à cette classe. Les membres de ce groupe sont caractérisés par un niveau de dispersion relativement élevé, avec un nombre de lieux visités et d'activités effectuées assez important. Ces derniers sont plutôt domiciliés dans la grande

périphérie de Bangkok.

- **Classe 2 :**

La classe 2 concerne 31,6 % de l'échantillon et regroupe des personnes qui visitent un grand nombre de lieux et effectuent beaucoup d'activités, avec cependant une faible propension à la dispersion dans l'espace. Ils habitent d'ailleurs essentiellement dans le centre étendu de Bangkok, là où les activités proposées sont les plus nombreuses (chapitres 9 et 10). Il n'est donc pas nécessaire pour ces personnes d'effectuer de très longs trajets pour effectuer leurs nombreuses activités.

- **Classe 3 :**

La classe 3 regroupe 4,1 % des utilisateurs, ceux qui se déplacent sur les plus grandes distances. Cet aspect semble plus discriminant que le nombre de lieux et d'activités qu'ils exercent, car l'amplitude de ces derniers critères sur l'axe des x est très importante dans ce groupe. Ils sont, comme pour les membres de la classe 1, essentiellement domiciliés dans la grande périphérie.

- **Classe 4 :**

Cette classe regroupe 20,8 % de l'échantillon. Il s'agit des individus qui effectuent le plus grand nombre d'activités dans un très grand nombre de lieux. Cela dit, leur dispersion dans la ville reste assez moyenne, inférieure par exemple aux membres de la classe 1. Leur répartition dans la ville n'est pas aussi claire que pour les autres classes, mais nous pouvons noter que le district qui concentre le plus de personnes de ce groupe est situé dans l'hypercentre.

- **Classe 5 :**

Les individus associés à cette dernière classe (24,9 % de l'échantillon) sont les moins mobiles. Ils fréquentent très peu de lieux, effectuent peu d'activités et se déplacent sur de courtes distances. Là encore, il n'y a pas d'opposition de type centre / périphérie claire pour les membres de ce groupe. Les trois districts qui comptent la plus grande part de personnes de ce groupe se trouvent en effet dans l'ouest, dans le sud, ou encore dans le centre.

Nous retrouvons ainsi dans certains de ces groupes (1, 2 et 4) une territorialité forte, où l'opposition centre-périphérie très marquée se traduit par des potentiels (ou nécessités) de déplacements bien distincts. Les personnes résidant dans le centre peuvent fréquenter beaucoup de lieux différents sans pour autant avoir à se déplacer sur de longues distances, ce qui n'est pas le cas pour les personnes résidant dans des zones périphériques. Néanmoins, cette logique n'est pas toujours respectée, notamment chez les plus sédentaires (groupe 5), ou ceux qui

fréquentent le plus de lieux (groupe 4), qui sont répartis dans la ville sans déterminisme spatial directement visible. Nous prendrons en compte ces considérations et spécificités lors de la génération d'agendas synthétiques.

2.2 Protocole de génération d'agendas

2.2.1 Étape 1 : initialisation des caractéristiques de l'espace d'activité

1.1 Nombre de lieux et d'activités

Dans un premier temps, nous tirons une probabilité d'appartenir à un des 5 groupes défini précédemment, selon les distributions des effectifs de ces groupes. Un agent aura par exemple 18,6 % de chance d'appartenir au groupe 1, contre 4,1 % pour le groupe 3 et 20,8 % pour le groupe 4 (figure 258).

Nous définissons ensuite l'activité principale ("Lieu d'éducation" ou "Autre") selon la répartition des effectifs dans chacun de ces groupes (figure 259). Si l'agent appartient au groupe 1 ou 3, il aura plus de chance d'être un étudiant que s'il appartenait au groupe 4.

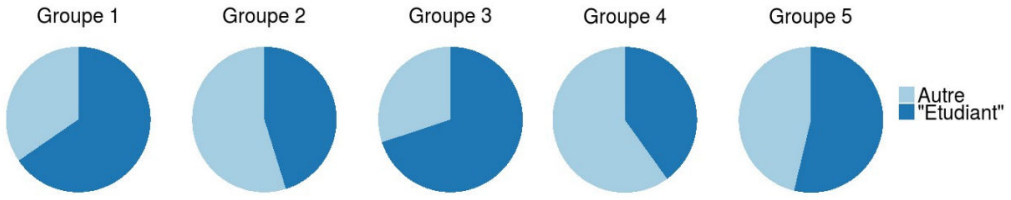


FIGURE 259 Part d'étudiants et des autres membres de l'échantillon, selon les groupes définis précédemment.

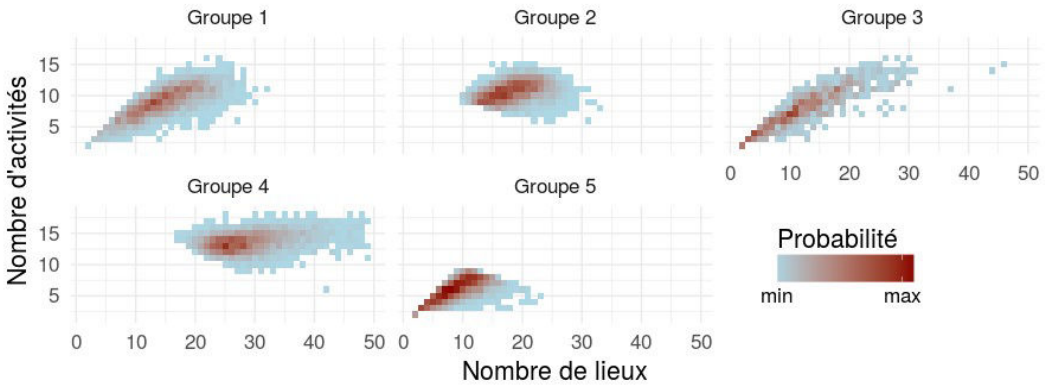


FIGURE 260 Relation entre le nombre d'activités réalisées et le nombre de lieux fréquentés, selon les différents groupes résultants de notre classification sur des paramètres de dispersions.

Il convient maintenant de définir un nombre de lieux fréquentés et un nombre d'activités réalisées. Pour cela, nous mobilisons la figure 260, qui montre la relation entre ces paramètres, en fonction des groupes. Nous tirons alors un nombre d'activités N_a, selon les fréquences de distribution du groupe auquel appartient l'agent (figure 260). Selon le nombre d'activités et le groupe de l'agent, nous tirons ensuite un nombre de lieux N_l correspondant, en suivant toujours les distributions de la figure 260.

1.2 Répartition des activités dans les lieux

Dans un premier temps, nous allons définir les activités qu'un agent va effectuer. La figure 261 ci-dessous montre la répartition des activités dans l'ensemble des lieux fréquentés par l'échantillon, d'après les agendas reconstitués. Les activités sont effectuées de manière assez similaire selon que l'activité principale soit un lieu d'éducation ou non. Et quasiment 30% des lieux visités appartiennent à des catégories de type « Autre ». Ces activités sont les plus représentées, suivies de la fréquentation de quartier « mixte dense » et « sortie dense ». N'ayant posé de contraintes sur les lieux fréquentés lors de l'élaboration des différents groupes, cette distribution sera utilisée pour tous les agents.

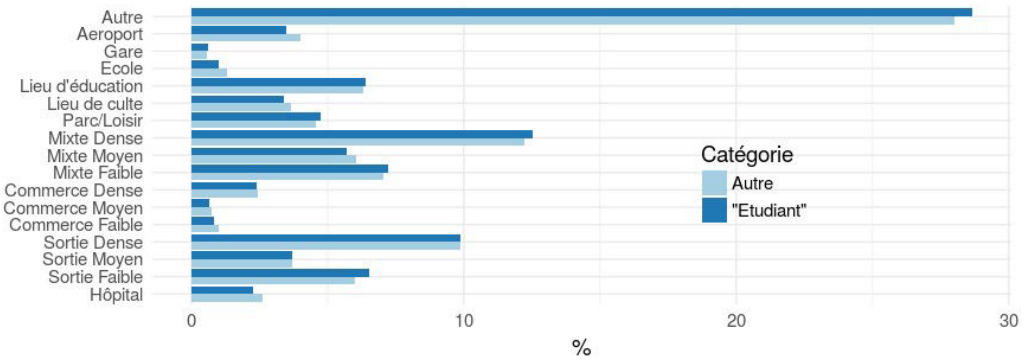


FIGURE 261 Répartition des activités dans les différents lieux fréquentés par l'échantillon.

Nous sélectionnons un nombre d'activités N_a selon un tirage sans remises parmi l'ensemble des activités disponibles, le tout pondéré par les fréquences de réalisation des activités de la figure 261. Nous obtenons ainsi une liste d'activité A que l'agent va réaliser. Les activités de type « Autre », « Mixte Moyen » ou « Sortie Dense » auront plus de chance d'être sélectionnées.

Une fois ces activités sélectionnées, nous allons leur attribuer un nombre de lieux. Par exemple, dans combien d'endroits différents un agent va effectuer une activité de type « Autre », ou « Mixte dense », si ces dernières ont été tirées lors l'étape précédente ? Pour ce faire, nous réalisons un tirage avec remise d'un nombre de lieux N_l, avec en entrée la liste des activités A, avec toujours une probabilité dépendant de la fréquence de ces activités (figure 261). Nous

obtenons en sortie une liste où chaque activité est associée à un nombre de lieux fréquentés. Les activités de types "Autre", Sortie Dense" auront ici plus de chance d'être effectuées dans plusieurs lieux qu'une activité de type "Commerce Dense".

Nous sommes aussi conscients que certaines activités sont réalisées de manières très ponctuelles, comme se rendre à l'hôpital, dans un aéroport ou une gare. Néanmoins nous les conservons pour l'instant, notamment pour observer la réaction de notre algorithme à leurs égards. Elles pourront par la suite être regroupées dans la catégorie « Autre », ou « Autre2 » si nous souhaitons maintenir une distinction.

À cette étape, chaque agent s'est vu attribuer une liste d'activité A qu'il effectue dans une liste de lieux L. Reste à définir la temporalité de leur visite.

2.2.2 Étape 2 : Fréquence et jours de visite

2.1 Fréquence mensuelle et hebdomadaire

Dans le chapitre 8, nous évaluons d'abord si un lieu est fréquenté une certaine semaine, puis nous définissions le nombre de jours de fréquentation une semaine donnée. Mais ceci peut conduire à des incohérences, dans le sens où un lieu peut être visité par exemple une seule semaine sur un mois, mais plusieurs jours cette même semaine.

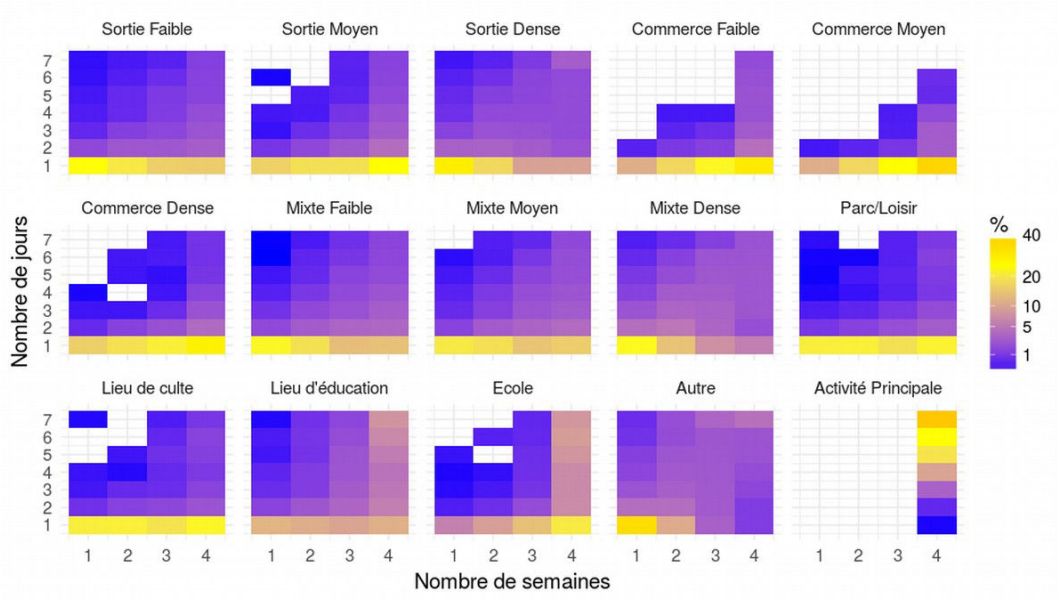


FIGURE 262 Probabilité de réalisation d'une activité un nombre de semaines et de jours donné.

Nous allons ici tenter une nouvelle approche, en tirer simultanément un nombre de semaines et un nombre de jours, selon le type d'activité concernée. Pour cela, nous nous basons

sur nos agendas et sur le pourcentage que représente chaque nombre de semaine (entre 1 et 4) au regard du nombre de jours (entre 1 et 7) pour chaque activité (figure 262). Les lieux de sortie dense ont par exemple plus de chance d'être visités une fois par semaine, et lors d'une ou deux semaines dans le mois. Si la plupart des lieux de types "Autre" sont fréquenté une fois dans un mois, certains sont visités toutes les semaines et presque tous les jours. Les écoles ont par exemple dans notre cas une probabilité non négligeable d'être visitées tous les jours et toutes les semaines. Dans ce cas, vu qu'il ne s'agit pas d'une activité principale, il pourrait s'agir implicitement d'une personne qui accompagne ses enfants à l'école.

Après cette étape, chaque lieu s'est vu associer un nombre de semaines et un nombre de jours au sein de ces semaines.

2.2 Jours de visites

Nous allons maintenant définir les jours de la semaine où une activité sera réalisée, selon le nombre de fois où elles ont été effectuées un jour de la semaine donnée d'après nos agendas reconstitués (figure 263). Nous décidons de diviser par deux cette fréquence pour les jours de week-end pour les activités principales et les lieux d'éducation.

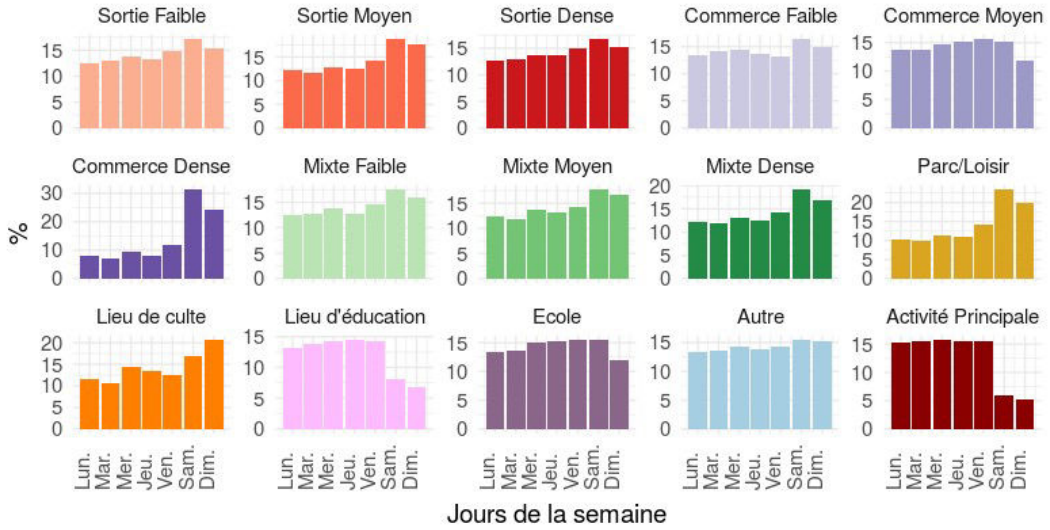


FIGURE 263 Répartition des pourcentages de fréquentation des activités selon les jours de la semaine, d'après les AR.

Ensuite, selon le nombre de jours où l'activité se réalise une semaine donnée, nous tirons des jours de la semaine selon les probabilités de réalisation de l'activité. Par exemple la visite d'un lieu de type « Commerce Dense » aura plus de chance d'être effectuée un week-end.

Si l'activité principale se déroule sur 5 jours, nous choisirons les jours de semaine. Si elle se déroule sur 6 jours, nous prendrons les jours du lundi au samedi. Sinon, nous tirons au sort,

comme pour les autres types d'activités.

2.2.3 Étape 3 : Durée d'une activité

Maintenant que nous avons défini pour chaque utilisateur le nombre de lieux qu'il fréquentera un jour donné, il ne reste plus qu'à définir une durée et une plage horaire. Dans le chapitre 8, nous définissions d'abord une séquence d'activité réalisée quotidiennement, puis nous attribuons des durées à ces activités, que nous harmonisions par la suite. Nous proposons l'approche inverse, c'est-à-dire que nous associons une durée à chaque lieu, puis nous définissons les séquences d'activités.

- ***Une durée selon le nombre de lieux visités quotidiennement***

Tout d'abord, nous commençons par estimer une durée de visite pour chaque lieu. Comme le montre la figure 264, et comme vu dans le chapitre 8, la durée d'une activité dans un lieu précis dépend en partie du nombre de lieux qu'un individu fréquente ce même jour. Plus il visite de lieux, plus les durées de visites auront tendances à être courtes. Aussi, nous posons tout d'abord qu'un individu ne peut visiter plus de 6 lieux (en plus d'être à son domicile et de réaliser son activité principale) une même journée. S'il était prévu qu'il en fréquente plus, nous en choisissons 6 de manière aléatoire. Bien entendu, les durées de visites varient aussi selon les activités concernées (figure non représentée ici). Pour chaque lieu visité, nous tirons donc une durée selon le nombre de lieux visités lors d'une journée et l'activité associée.

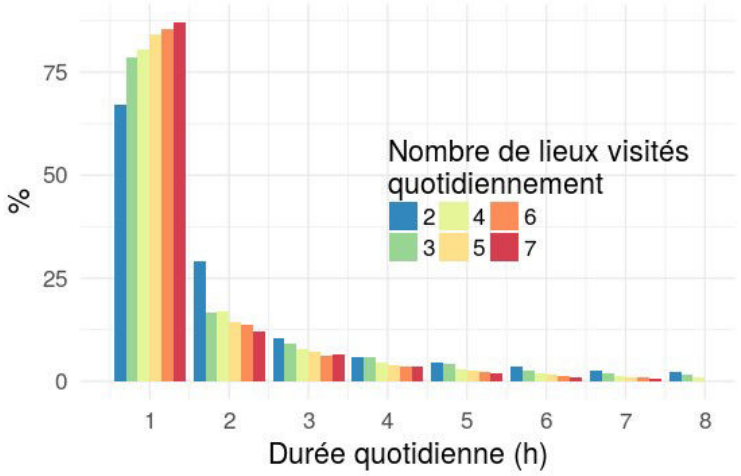


FIGURE 264 Distribution des durées quotidiennes des activités (autre que domicile et principale) selon le nombre de lieux visités quotidiennement, d'après les AR.

- ***Contraindre les durées***

L'étape suivante vise à éviter qu'un individu se voie attribuer des « journées à rallonge »,

et passe trop de temps hors de son domicile. Nous posons donc différentes contraintes.

Tout d'abord, l'activité principale doit durer entre 6 h et 12 h. Ensuite, nous effectuons des tirages de durée pour chaque lieu tant que la durée hors du domicile (somme des durées associées aux lieux fréquentés) est supérieure à 16 h. Le *time use survey* de 2009 indique en effet une durée moyenne de travail de 7h42 et un temps de sommeil moyen de 8 h (NSO, 2009).

Si au bout de 20 itérations, aucune configuration adéquate n'est apparue, nous réduisons le nombre de lieux que l'individu était censé fréquenter (nous enlevons un lieu de manière aléatoire) et recommençons la procédure. Ainsi, chaque agent a une durée de présence à son activité principale comprise entre 6 h et 12 h et reste au moins 9h par jours chez lui. Il ne nous reste plus qu'à définir les heures où les différents lieux sont visités.

2.2.4 Étape 4 : Définir la séquence quotidienne des activités

Nous allons maintenant ordonner les visites des différents lieux pour chaque journée.

- ***Heure de départ du domicile et d'arrivée à l'activité principale***

Tout d'abord, nous tirons une heure de départ du domicile pour les jours de semaines et de week-end, ce qui nous permet de déterminer à quelle heure un agent commence ses activités quotidiennes. Par souci de conservation des plages horaires de travail, nous tirons également une heure d'arrivée à l'activité principale, valable pour tous les jours où un individu est censé exercer cette activité. Nous posons que l'heure de départ du domicile doit être inférieure à l'heure d'arrivée à l'activité principale.

- ***Séquence des activités***

Nous allons maintenant raisonner d'un point de vue séquentiel.

- Un agent quitte un lieu à un instant t pour aller dans un autre lieu à l'instant $t+1$. Tous les lieux prévus d'être visités ce jour sont candidats, sauf l'activité principale. Cette dernière n'entre en jeu uniquement si l'heure d'arrivée d'une activité à t est supérieure ou égale à l'heure d'arrivée à l'activité principale.

- Lorsqu'un lieu l est sélectionné pour être visité, l'heure d'arrivée vaut t, et l'heure de départ, $t + d_l$, ou d_l est la durée de visite du lieu, et le lieu est supprimé de la liste L.

Reste à définir les probabilités d'effectuer une activité dans un lieu donné. Nous proposons ici plusieurs méthodes.

- ***méthode 1 : Utilisation d'une matrice de transition horaire***

Comme dans le chapitre 8, nous utilisons ici des matrices de transition, mais avec la particularité d'être calculées par plages horaires, à la manière de (Wu *et al.*, 2014). Les taux de transitions entre deux activités sont définis d'après les séquences d'agendas reconstituées. À chaque heure, nous ne sélectionnons que les lieux dont la fréquentation s'achève et nous comptons pour chaque activité la part des autres activités effectuées à l'heure suivante, comme illustré par la figure 265 ci-dessous.

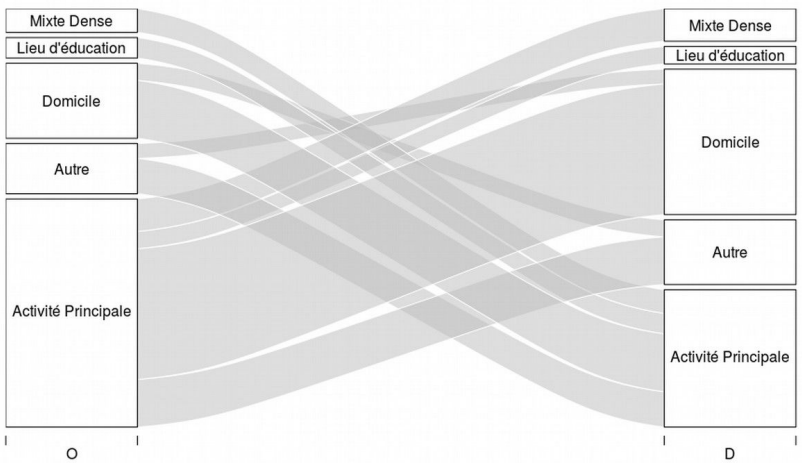


FIGURE 265 Représentation de la probabilité de passer d'une activité effectuée à 15h (gauche), à une autre à 16 h (droite), un lundi. D'après les AR.

Cette dernière présente les 10 transitions les plus importantes un lundi à 15 h. Lorsqu'une personne quitte son activité principale, elle a de grandes chances d'aller à son domicile, mais pas uniquement. La probabilité d'effectuer une activité à *t+1* est donc ici conditionnée par le type de lieu à *t*. Par la suite, nous appellerons cette méthode « Matrice de transition ».

- ***méthode 2 : heure de présence***

Nous testons maintenant une autre approche, basée sur les horaires globaux de fréquentation. Nous reprenons pour chaque activité le pourcentage de visite par tranche horaire de la figure 248, (bien plus haut). La probabilité de visiter un lieu à l'instant *t+1* ne dépend alors plus du type d'activité effectuée à *t*, mais des profils de fréquentation. Par exemple, le matin, les activités de type « Parc » ou « École », qui présentent un pic vers 6-7h, auront plus de chance d'être tirées que des activités de types « sorties », qui présentent de fortes valeurs le soir. L'avantage d'une telle approche est qu'il est envisageable d'utiliser d'autres données, comme les profils de fréquentations issus des *check-in* de *Facebook*. Par la suite, nous appellerons cette méthode « Probabilité Horaire ».

- ***méthode 3 : matrice de transition et heures de présence***

La dernière méthode est un mélange des deux précédentes, où nous multiplions pour chaque heure de la semaine les taux de transition des matrices de la méthode 1 par les profils horaires utilisés dans la méthode 2. Par la suite, nous appellerons cette méthode « Mixte ».

À noter que par construction, l'activité principale ne peut être sélectionnée qu'à partir d'une certaine heure. Afin de contraindre sa réalisation rapide après son entrée dans la séquence, nous posons qu'elle a 4 fois plus de chance d'être sélectionnée que les autres activités, sauf entre 12 h et 13 h, et ce pour les trois méthodes employées.

Il faut compter environ 35 minutes pour générer 1000 agendas sur une machine standard, et nous avons généré 40 000 agendas mensuels individuels pour chacune des trois méthodes.

2.3 Résultats

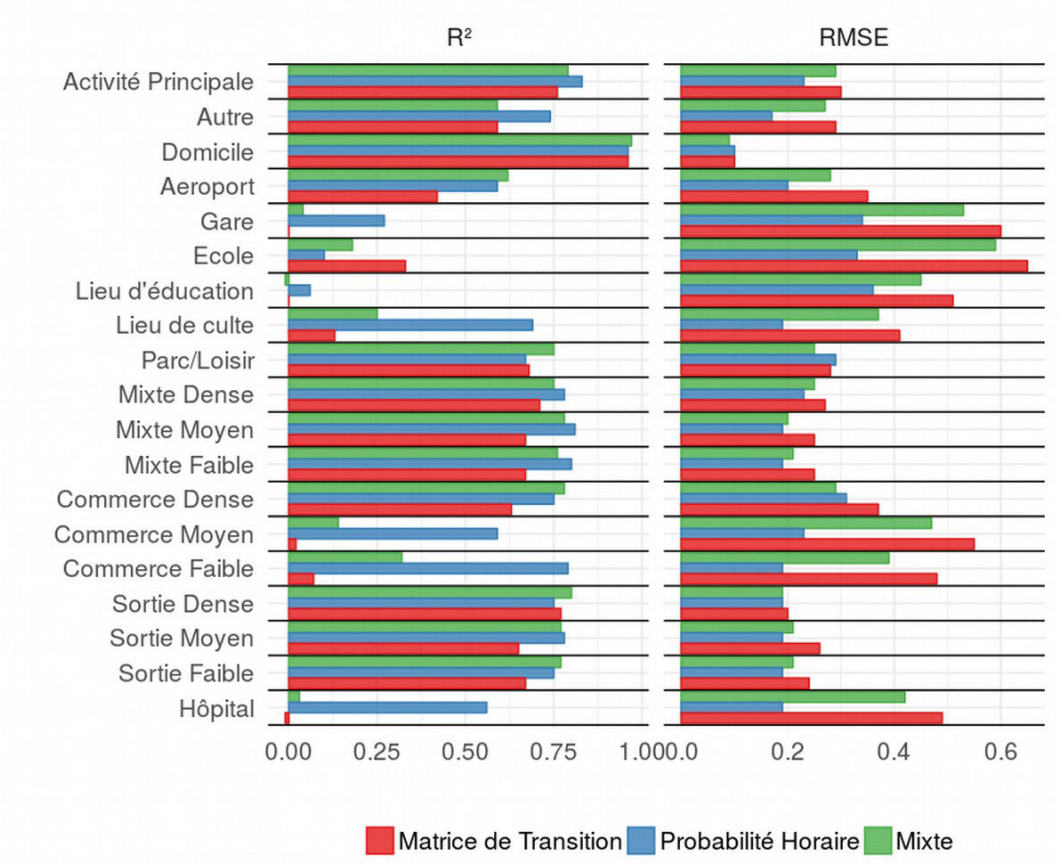


FIGURE 266 Comparaison entre le profil temporel des agendas reconstitués et les profils temporels obtenus par nos trois méthodes, par type d'activité. Sont représentés les R^2 ajustés (à gauche) et les RMSE (à droite).

Une première analyse des résultats peut être faite en comparant globalement les heures de présences dans différents types de lieux pour chacune des trois méthodes exposées ci-dessus,

avec en référence les profils de fréquentations issus des agendas reconstitués. La figure 266 présente les coefficients de corrélation (R^2 ajustés) et la racine carrée de l'erreur quadratique moyenne (Root Mean-square Error, ou *RMSE*) entre chaque méthode utilisée pour définir les séquences horaires ("Matrice de transition", "Probabilité horaire" et "Mixte") et les profils de référence. Pour rappel, un *RMSE* proche de 0 signifie une bonne concordance entre les séries.

Les différents indicateurs sont assez proches pour les activités « Domicile » et « Activité Principale », quelle que soit la méthode, avec des écarts relativement faibles avec les agendas reconstitués (R^2 très élevés, *RMSE* plutôt faible). Mais les profils issus de la méthode « Probabilité Horaire » présentent globalement les R^2 les plus élevés et les RMSE les plus faibles, dans quasiment toutes les catégories, suivie par la méthode "Mixte".

De plus, lorsque les séries obtenues sont très éloignées des profils des agendas reconstitués, comme pour des activités peu effectuées (commerces moyen ou faible, lieux de culte) ou visitées théoriquement de façon épisodique (gare, aéroport, hôpitaux), la méthode « Probabilité horaire » surpasse largement les autres. Elle est donc la plus à même à reproduire les temporalités de visites des lieux associés à des activités. Néanmoins, pour ce qui est des lieux d'éducations et les écoles, toutes les méthodes donnent de mauvaises estimations.

Pour chacune des méthodes, les profils du domicile et de l'activité principale sont très proches de ceux de références (figure 267). Néanmoins, dans chaque cas, la présence au domicile est sous-estimée la nuit. Pour ce qui est de l'activité principale, la largeur à mi-hauteur de la courbe quotidienne est plus grande sur le profil de référence, ce qui signifie globalement des temps de présence plus courts obtenus lors de la génération des agendas. Ceci peut simplement s'expliquer par nos différents algorithmes. Dans celui de la reconstruction des agendas, nous posions une durée maximale de l'activité principale à 14 h, contre 12 h dans la génération des agendas (AG).

Pour ce qui est des lieux de type « Autre », la méthode « Probabilité horaire » retranscrit bien le pic de fréquentation du soir, qui apparaît plus tôt et plus marqué dans les autres méthodes. Le pic du midi n'est pas retranscrit, du fait que nous n'avons pas scindé l'activité principale lors de la génération des agendas. Pour des raisons de simplicité, nous avions posé qu'un agent n'effectue cette activité qu'une seule fois dans une même journée. Il ne peut donc pas sortir le midi pour aller dans un lieu de sortie par exemple et retourner ensuite à son travail. La méthode « mixte » fait cependant apparaître un pic de fréquentation vers 6 h du matin, qui correspond bien avec le début de la pente de la courbe de référence.

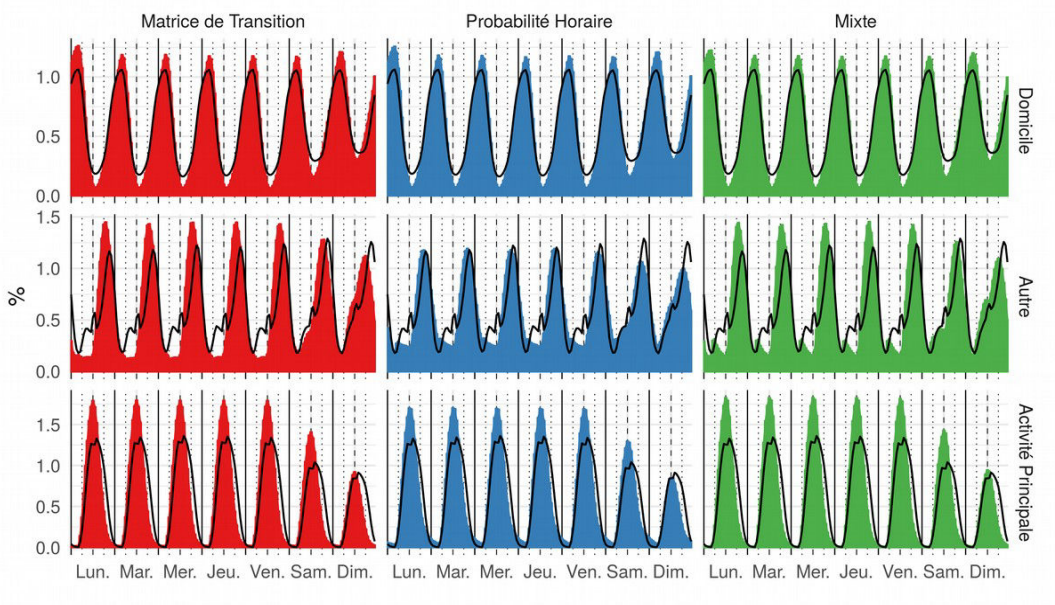


FIGURE 267 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (1/3). Sont représentées ici les activités de type "Autre", "Principale" et "Domicile".

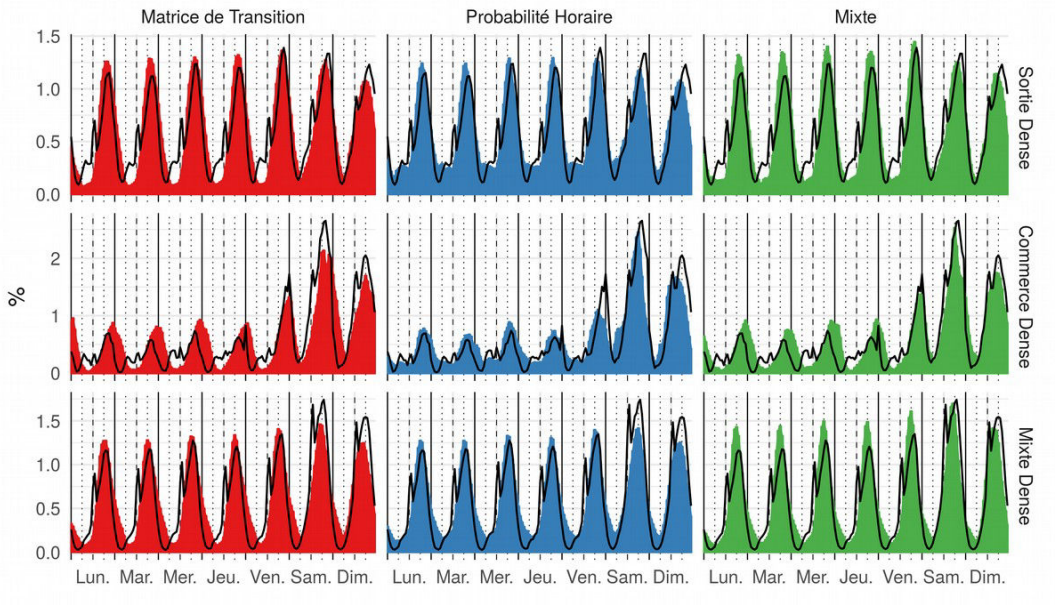


FIGURE 268 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (2/3). Sont représentées ici les activités "Commerce", "Sortie" et "Mixte" dense.

Si nous regardons les différents profils pour les activités densément dotées en *POI* (Sortie, Commerce et Mixte, figure 268), nous pouvons noter un bon accord global. Pour les lieux de sorties, les pics du soir sont bien en phase pour les méthodes « Matrices de transition » et

« Mixte », un peu moins avec la dernière approche. Le petit pic du matin est absent dans toutes méthodes, et la méthode « Probabilité horaire » tend à sur-estimer la présence entre minuit et 6 h du matin. Pour ce qui est des zones de commerces denses, les pics de fréquentations sont très tardifs les jours de semaines pour les méthodes « Matrice de transition » et « Mixte », alors qu'ils sont en phase si nous utilisons la « probabilité horaire ». Nous pouvons noter une bonne concordance dans les trois cas pour les lieux de type mixte.

La figure 269 montre les résultats pour quatre types de lieux : les parcs, les lieux de cultes, les lieux d'éducation et les écoles. Concernant les parcs, l'approche par la probabilité horaire donne les meilleurs résultats du lundi au vendredi midi, avec un pic du soir et du matin bien estimé. Dans les autres cas, les pics sont plus tardifs, vers 22 h. La méthode « mixte » fournit néanmoins les meilleurs profils le week-end.

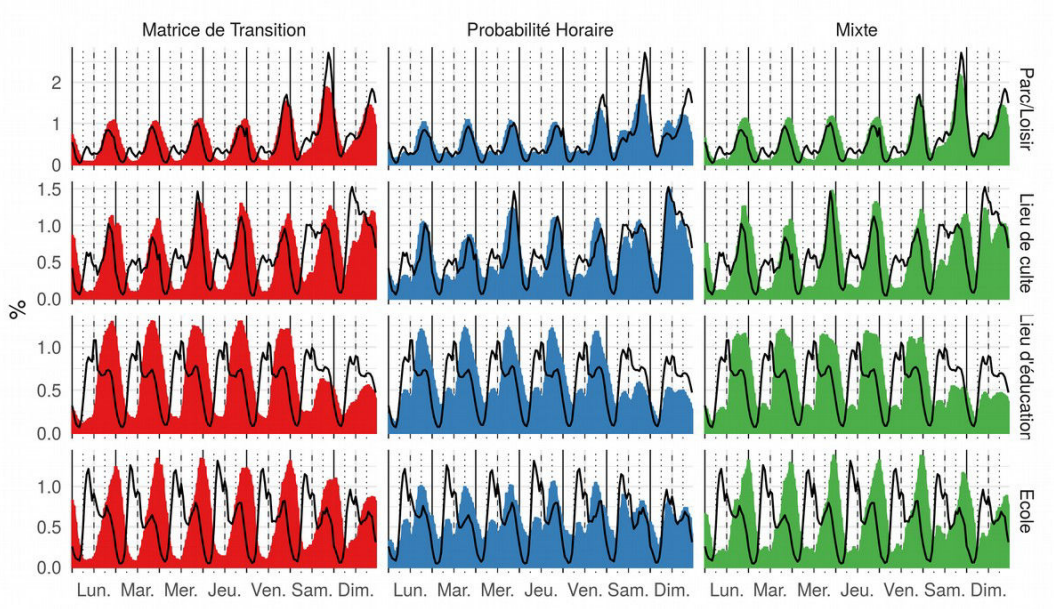


FIGURE 269 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (3/3). Sont représentées ici les activités "Parc","lieu de culte", "lieu d'éducation" et "école".

Pour les lieux de cultes, l'utilisation des probabilités horaires permet de faire ressortir les pics du soir et du matin, même si ces deniers sont moins marqués. Les autres approches entraînent des pics très tardifs et une absence de fréquentation matinale. Le même constat peut être fait pour les lieux d'éducation et les écoles, où là encore l'approche par les probabilités de fréquentations horaires fournit les meilleurs profils, même si les pics du matin ne sont pas assez marqués et ceux du soir beaucoup trop importants.

En guise de synthèse, la figure 270 compare les profils horaires pour les activités faiblement

visitées (toutes sauf l'activité principale, le domicile, les lieux de type « autres » et les zones de commerces, de sorties et mixtes denses).

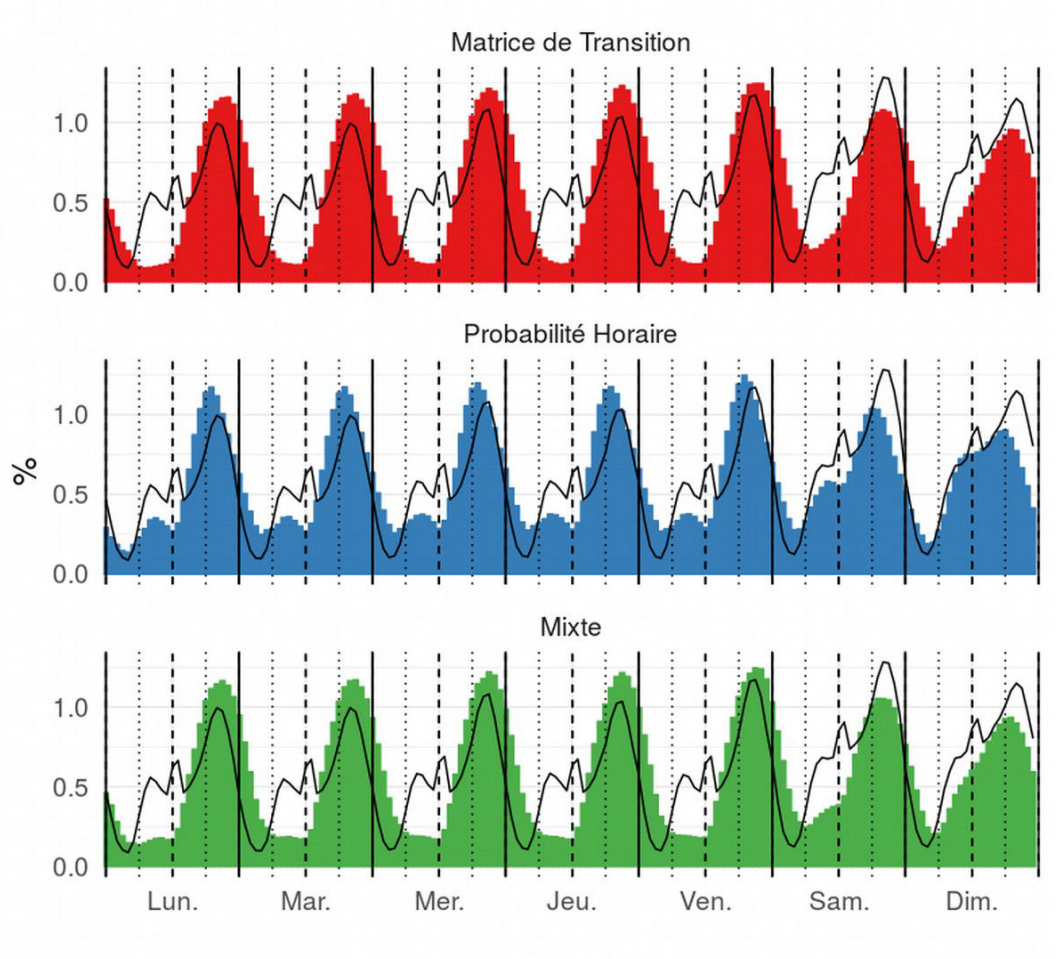


FIGURE 270 Profils temporels agrégés des activités des activités faiblement visitées (toutes sauf l'activité principale, le domicile, les lieux de types « autres » et les zones de commerces, de sorties et mixtes denses) sur une semaine.)

Les pics de fréquentation coïncident relativement bien dès lors que nous utilisons les matrices de transition, pondérées ou non par les probabilités horaires. Néanmoins, ces activités sont nettement plus fréquentées le soir, entre 22 h et minuit que lorsque nous utilisons simplement les probabilités de fréquentation horaires pour définir la séquence des activités quotidiennes. Cette dernière méthode permet aussi de faire ressortir les pics de fréquentation du matin, qui bien que moins marqués que sur les profils de références, ressortent toutefois de manière assez claire. Comme partout, les pics du midi des jours de semaines n'apparaissent pas du fait de notre algorithme (voir supra).

Les moins bons accords obtenus par l'utilisation des matrices de transition pour la définition des séquences quotidiennes sont imputables au fait que les activités les moins

fréquentées auront une plus faible probabilité d'être sélectionnées en début de journée, du fait d'une concurrence avec les activités effectuées en plus grand nombre. Par exemple, prenons une personne qui quitte son activité principale, et qui a le « choix » de visiter un lieu de type « Autre », une école et un lieu de culte. Elle aura une plus grande probabilité d'aller directement dans un lieu de type « Autre ». Le lieu de type « Autre » est ensuite retiré de la liste, laissant les écoles et lieux de culte, qui ont donc plus de chance d'être visités à la toute fin de la journée. L'approche « mixte » tend à réduire cet effet, mais l'utilisation des probabilités horaires le gomme en grande partie.

Un autre intérêt d'utiliser des probabilités horaires plutôt que des matrices de transition réside dans l'interchangeabilité des données utilisées. Il est ainsi tout à fait possible d'utiliser non pas les profils horaires obtenus par nos agendas reconstitués de nos données *Twitter*, mais par exemple les profils horaires déduits des *check-in* de *Facebook*, de simples comptages issus du terrain, voire des enquêtes institutionnelles appropriées, si ces dernières existent dans la ville étudiée.

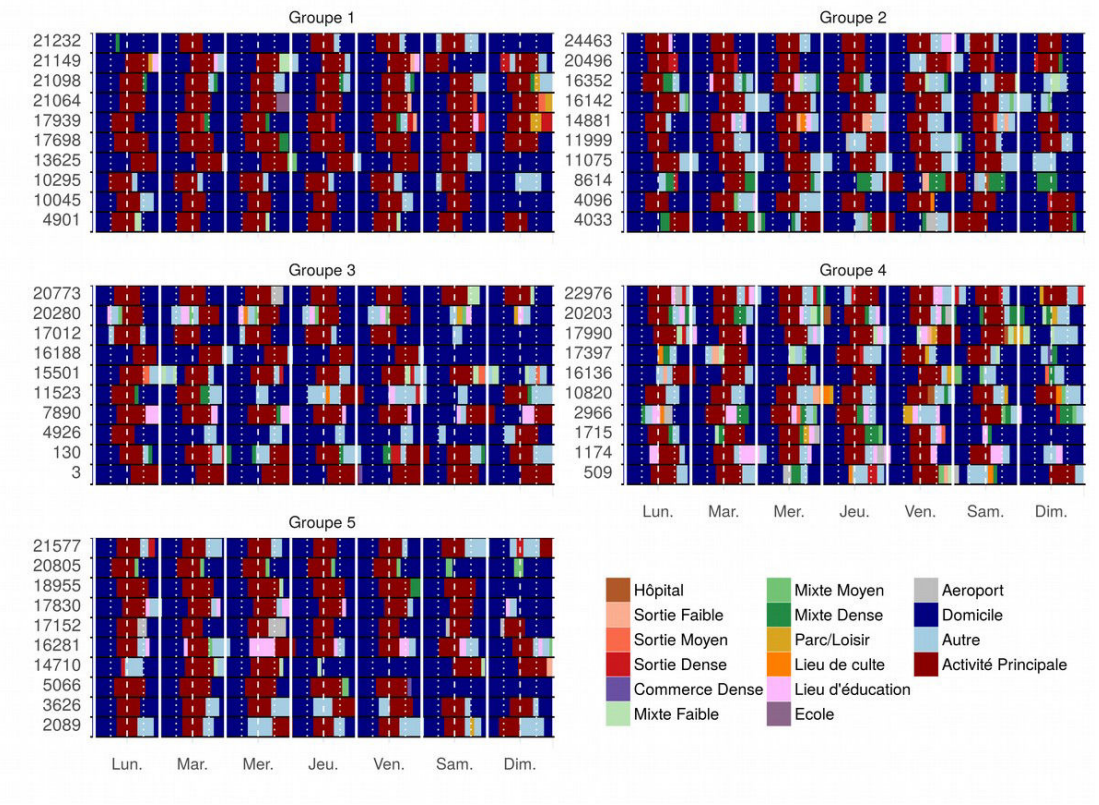


FIGURE 271 Exemples d'agendas sur une semaine par groupe d'utilisateurs.

Ainsi, l'utilisation des probabilités horaires dans la définition agendas semble appropriée pour retranscrire globalement les temporalités des activités. Examinons maintenant quelques agendas individuels issus de cette méthode. La figure 271, présente des agendas sur une semaine,

pour 10 agents de chacun des groupes définis précédemment. Comme sur les toutes les autres figures, les pointillées correspondent à 6 h et 18 h, les tirés à midi et les traits pleins à minuit.

Les membres de groupe 5 fréquentent très peu de lieux différents et ont des agendas assez réguliers. L'agent 5066 n'effectue par exemple que des navettes domicile travail du lundi au vendredi, mais fréquente une zone de type mixte moyen après son activité principale le jeudi, et une zone de commerce dense le vendredi, et reste chez lui le week-end. Aussi notre algorithme entraîne qu'un agent ne rentre chez lui qu'après la fin de ces activités, alors qu'il est fort probable qu'une personne rentre chez elles entre deux visites de lieux différents. Il faudra prendre en compte cela par la suite.

L'agent 16 281 a quant à lui une activité nocturne assez importante pour cette semaine donnée. Si la contrainte de ne pas passer plus de 16 h hors du domicile par jours est normalement respectée, du fait d'activités nocturnes prolongées un jour donné, des temps au domicile peuvent être très cours, voir absent le matin du jour suivant. Ceci est d'autant plus visible chez les membres du groupe 4, qui concerne les agents qui fréquentent le plus de lieux et ont le plus d'activités.

Il faudrait donc par la suite contraindre les présences au domicile, qui sont pour l'instant simplement attribuées par défaut, lorsqu'aucune autre activité n'est réalisée à une plage horaire donnée. Une autre possibilité serait de diminuer encore les durées passées hors du domicile ou alors le nombre de lieux fréquentés, mais cela risquerait de trop réduire le nombre de personnes ultra-mobiles. Il serait aussi envisageable d'augmenter la résolution temporelle, en raisonnant par exemple par tranche de 30 minutes au lieu d'une heure, même si cela implique de tout reprendre depuis le départ. Une autre possibilité serait aussi d'utiliser un solveur "prolog" (Carlsson *et al.,* 1997), à la manière de Banos *et al.* (2006), pour forcer la réalisation de certaines activités à des horaires précises. L'idéal serait également d'avoir des données de références fiables, récoltées sur le terrain, afin de calibrer à chaque étape les différents paramètres qui régissent la réalisation des activités.

Quoi qu'il en soit, la génération d'agendas crédibles et continus dans le temps n'est pas une chose aisée, compte tenu de la nature épisodique et de l'incertitude des données en entrée du modèle. Des biais peuvent aussi s'insérer à chaque étape et d'autres essais avec d'autres approches méritent d'être effectués. Il conviendrait également de tester cet algorithme avec les données de Delhi, qu'elles proviennent du terrain ou des données *Twitter*, ce qui pourrait nous aiguiller pour trouver des pistes d'amélioration et faire des études de sensibilité.

Dans une logique de pure calibration de modèle, où l'objectif est de forcer la corrélation entre les agendas observés et ceux simulés, il est aussi envisageable de générer un grand nombre d'agendas suivant l'une des méthodes évoquées et de n'en garder qu'une fraction qui permette

de retranscrire aux mieux les données observées.

Notre code de génération d'agendas n'est pas encore commenté et optimisé. Lorsqu'il le sera, nous le rendrons libre et public, selon une licence qu'il conviendra de définir, ce qui permettra aux personnes intéressées de pouvoir faire leurs propres tests et/ou de le modifier en ajoutant leurs propres idées quant à l'attribution de semaines de réalisation d'activités, de jours de visites, de durées ou encore de séquence quotidienne d'activités.

Aussi, l'agenda de nos agents n'est pas encore spatialisé et les temps et modes de transports non pris en compte. À défaut de présenter un modèle complet, nous discuterons dans la section suivante de quelques pistes pouvant permettre d'affecter une localisation à chaque activité effectuée par un agent.

3 Comment affecter des localisations ?

Attribuer un lieu de domicile à un agent n'a rien de très compliqué en soi. Il suffit par exemple de récupérer des résultats des différentes cartographies dasymétriques présentées dans les chapitres précédents qui estiment la répartition de la population dans un carroyage d'une maille de taille à définir. Puis à répartir les agents dans ces mailles selon leur population, en prenant en compte le groupe auquel ils appartiennent. Par exemple, un agent membre du groupe 2 aura plus de chance d'habiter dans le centre-ville que s'il était membre du groupe 1. Nous pouvons aussi considérer des informations démographiques des différents recensements, comme les tranches d'âges par sous-district et contraindre le lieu de domicile en fonction de l'âge d'un agent si nous trouvons bien entendu une méthode permettant d'associer un âge à des profils de mobilités. En revanche, affecter des localisations aux autres activités est une tache nettement plus complexe.

Utiliser des modèles d'attractivités ?

Parmi les travaux cherchant à générer des agents synthétiques mobiles à partir de données existantes, ceux de Pappalardo et Simini (2017b) sont assez proches de notre démarche. À partir de données téléphonies et d'enregistrement GPS de voitures, ils définissent des agendas, constitués de suite de lieux abstraits fréquentés par un agent à différente plage horaire. Aucune information sur la localisation où sur le type de lieu fréquenté n'est présente, juste une succession temporelle d'endroit non défini sémantiquement et spatialement. Ils attribuent ensuite une localisation à chacun des lieux fréquentés par un agent, en utilisant leur modèle *d*-EPR, (Density Exploration and Preferential Return) (Pappalardo *et al.,* 2016b). La probabilité P_{ij} qu'un agent situé en i se rende en j suit alors un modèle gravitaire de type :

$$P_{ij} = \frac{1}{N} \frac{r_i r_j}{d_{ij}^2} \tag{31}$$

Avec d_{ij}, la distance entre les deux localisations, r_i et r_j la « pertinence » des localisations i et j, soit un critère d'attractivité défini dans le cadre de l'utilisation de données téléphoniques en comptant le nombre de personnes ayant bornés avec leurs téléphones dans une antenne relais appartenant aux mailles *i* et *j*. *N* est un paramètre d'ajustement, défini comme :

$$N = \sum_{i,j \neq i} P_{ij} \tag{32}$$

Cette approche, relativement simple à mettre en œuvre est tout à fait envisageable à implémenter, mais requiert néanmoins des critères d'attractivités de lieux. Dans notre cas, nous avons les densités de *check-in Facebook* et de *tweets* dans la ville pour définir la "pertinence" *r* dans les lieux *i et j*. Il est donc possible d'utiliser une de ces bases de données. De plus, nous avons aussi ces informations par tranche horaire *t* et par type d'activité *a*, c'est-à-dire que si un agent cherche par exemple un lieu de sortie à 20 h, nous posons qu'il aura d'autant plus de chances d'aller dans de tels endroits s'ils sont très fréquentés à cette heure donnée. Nous pourrions donc ajouter des contraintes et poser que la probabilité *P* d'être dans un lieu *i* à l'instant *t* d'aller dans un lieu *j* pour réaliser une activité *a'* à *t+1* comme :

$$P_{ij,a',t+1} = \frac{1}{N} \frac{r_{i,a,t} r_{j,a,t+1}}{d_{ij}^2} \tag{33}$$

Il est aussi envisageable d'utiliser des modèles de choix de destination, comme le modèle multinomial Logit. Cette approche permettrait à un agent de sélectionner un lieu où il peut potentiellement réaliser une activité en comparant les différentes alternatives selon des critères de "choix" (distance et/ou attractivité du lieu, ou encore type de quartier et âge de l'agent). Ce modèle a par exemple été utilisé à Anvers (Hammadou *et al.*, 2008) et dans quatre villes flamandes (Thomas *et al.*, 2009), permettant d'estimer les différents facteurs susceptibles d'influencer la visite de certaines zones commerciales.

Calibrer des paramètres sur les données observées ?

Une autre approche serait de reprendre des éléments de la méthode développée par (Wu *et al.*, 2014), et de définir simultanément une localisation et une activité lors de la reconstruction d'agendas. Pour résumer de manière schématique les trois lignes d'équations qui régissent leur modèle, ils séparent tout d'abord les activités figées (domicile, travail) des activités flexibles (sorties, restaurants, autres). La probabilité d'aller dans un lieu pour effectuer une activité à l'instant *t+1* dépend de matrices de transitions horaires par activités, à la manière de ce que nous avons présenté dans la section précédente. Pour les activités flexibles, ils ajoutent

une contrainte qui dépend de l'attractivité de l'activité à l'instant t dans une zone donnée et pondérée par la distance $d_{ij}{}^{\beta}$ entre les lieux i et j, à une puissance négative β. Cette dernière est obtenue en effectuant des millions de simulations sur une large gamme de β et en comparant la distribution des distances entre deux *check-in* réalisés sur *Weibo* à celle entre deux lieux fréquentés d'après leur simulation. Ils gardent finalement le β qui permet d'avoir le meilleur accord entre les données observées et simulées.

Pris séparément, les résultats qu'ils ont présentés sont plutôt très cohérents sur la répartition spatiale moyenne de leurs agents dans la ville et sur les déroulements temporels des activités[410]. Nous sommes pour notre part un peu récalcitrants à vouloir à tout prix trouver le bon paramètre β pour caler les données simulées aux données observées, même si cette méthode est extrêmement courante.

Utiliser les distributions des distances entre deux activités d'après les données ?

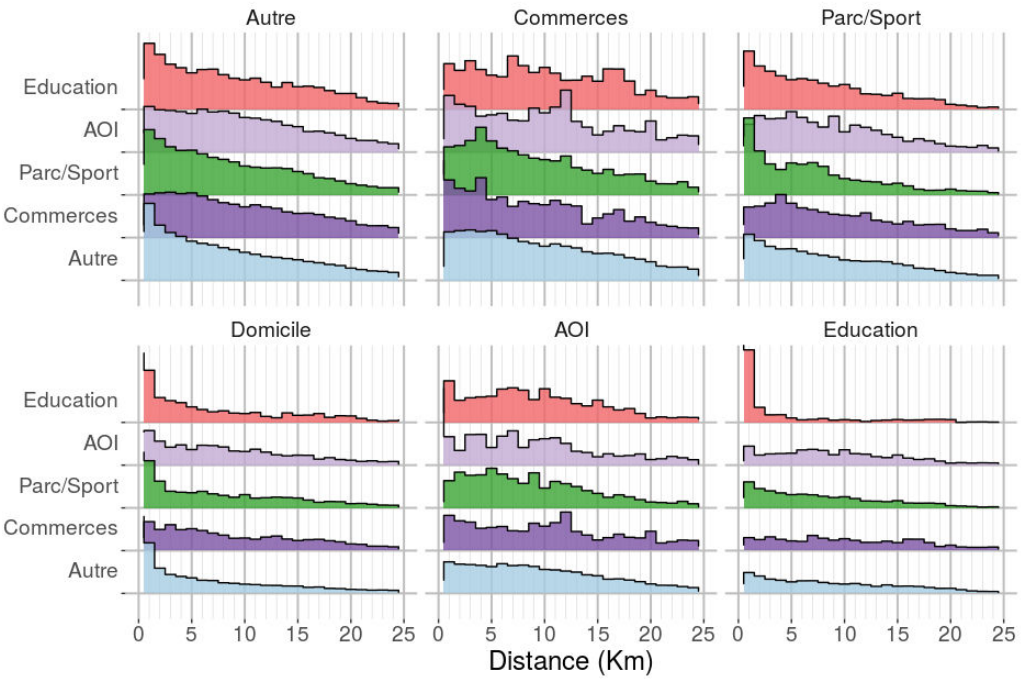


FIGURE 272 Distribution des fréquences de distances parcourues entre tous les principaux types de lieux. Chaque bloc correspond à une activité de départ, les différentes lignes aux activités d'arrivées.

En effet, si nous nous reportons aux données que nous avons collectées, nous pouvons calculer pour chaque utilisateur de *Twitter* les distances entre chacune des activités qu'ils effectuent, d'après leur espace d'activité brut, ou les AR. Et un individu ne parcourt

410. Avec les mêmes lacunes que nous avons observées lorsque nous employons des matrices de transitions horaires, à savoir des activités nocturnes beaucoup plus importantes chez les agents que dans les données observées

probablement pas les mêmes distances en fonction du niveau de contrainte d'une activité. Par exemple, aller à son travail ou rendre visite à des proches à leur domicile sont des activités figées dans l'espace, et il se peut que les distances entre le domicile et ces activités soient très grandes. D'autre part, si un individu décide d'aller dans un parc juste pour se promener, il y a de fortes chances que ce parc soit à proximité de son activité précédente. Ceci est illustré sur la figure 272, pour des données à Delhi (le même type d'information peut bien entendu être extrait à partir des données *Twitter* à Bangkok), sans prise en compte, pour l'instant de l'enchaînement temporel des activités. Pour revenir à nos agendas, si un agent se trouve à son domicile et qu'à l'itération suivante il doit visiter un parc, il aura plus de chance de sélectionner un de ces types de lieux à une distance proche de son domicile. De même, si l'agent effectue une activité de type 'autre' et qu'il doit visiter ensuite un commerce, dans ce cas la probabilité de choisir un lieu poche de son activité en cours sera un peu plus grande qu'un lieu très éloigné, car cette probabilité décroît de manière linéaire avec une pente assez faible.

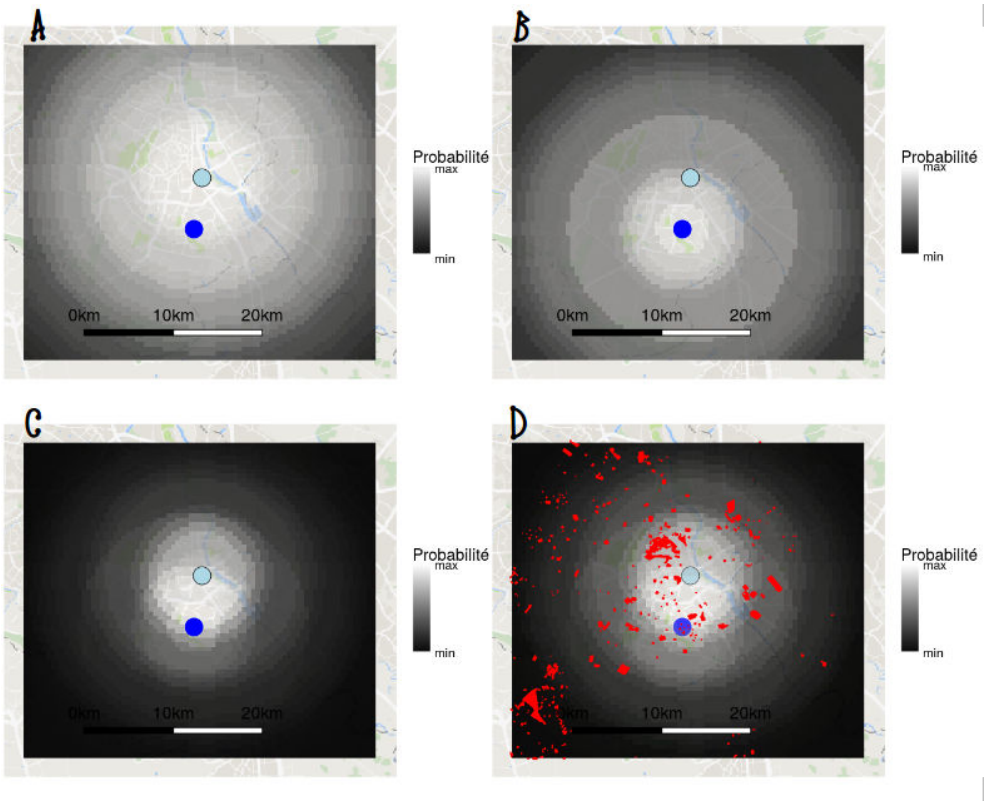


FIGURE 273 Proposition de protocole de sélection d'un lieu de catégorie "Commerce", en fonction d'une localisation pour une activité de type "Autre" (bleu ciel) et de la localisation du lieu de domicile (bleu). La distribution des distances entre un lieu de type "Autre" et un lieu de type "Commerce" est spatialisée et centrée sur le lieu de type "Autre" (a). La distribution des distances entre le domicile et un lieu de type "commerce" est spatialisée et centrée sur le domicile (b). Les distributions sont multipliées (c), et nous ajoutons ensuite les zones de type "Commerces", en rouge (d). Exemple à Delhi.

L'utilisation de ces distributions de fréquences de distance à parcourir entre deux activités pourrait donc être une alternative à l'utilisation d'un modèle gravitaire, ou à la recherche d'un paramètre permettant de caler les données simulées sur celles observées

La figure 273 est une proposition de méthode permettant d'attribuer une localisation à une activité selon les distributions des distances parcourues entre deux activités issues de la figure 272. Sur la figure 273.a, l'agent se trouve dans un lieu de catégorie « Autre » (point bleu ciel), son domicile étant le point bleu et il doit visiter à la prochaine itération un lieu de type « Commerce ». Le niveau de gris correspond à la spatialisation de la distribution des distances entre les lieux de type autre et les commerces, centrée sur sa position actuelle (point bleu ciel).

Nous pourrions poser que l'agent peut choisir n'importe quel commerce selon un niveau de probabilité qui découle du niveau de gris. Néanmoins, si nous appliquons cette méthode à chaque itération, il est possible qu'à la fin d'une journée composée de beaucoup d'activités, l'agent se retrouve assez loin de chez lui, ou au contraire très proche de son point de départ. Pour éviter cela, nous choisissons ici d'ajouter une contrainte, en prenant en compte la distribution des distances entre le domicile et les commerces (figure 273.b), afin que l'agent cherche préférentiellement un lieu pas trop éloigné de son domicile.

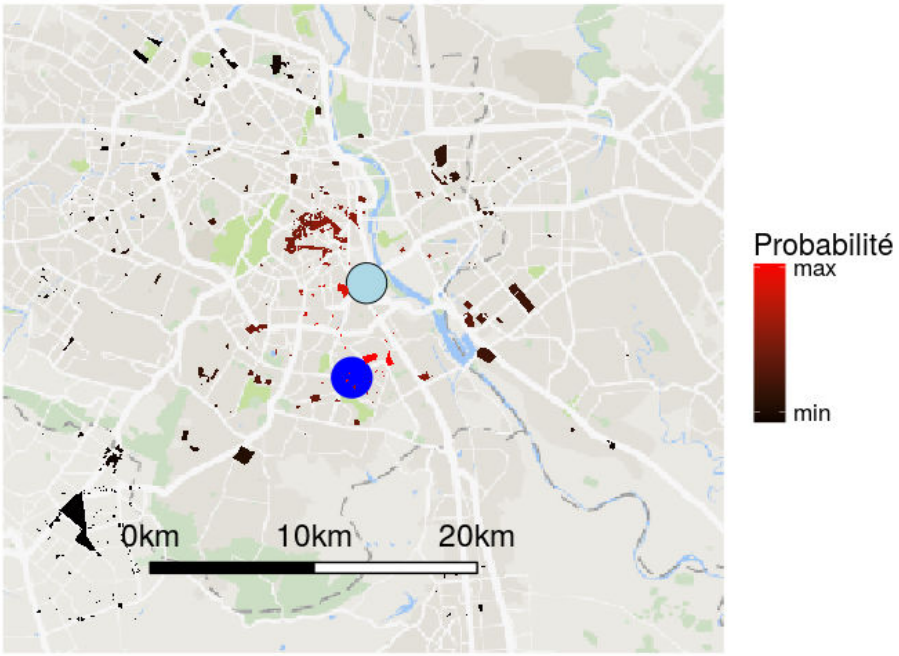


FIGURE 274 Probabilité de sélectionner un commerce dans la ville sachant que la personne se trouve dans un lieu de type « Autre » et connaissant la localisation de son domicile. Exemple à Delhi.

Nous multiplions ensuite les deux couches de probabilités (figure 273.c), en donnant arbitrairement un poids ici 2 fois plus important à la probabilité de tirer une distance pour se rendre dans un lieu de type « commerce », sachant que l'agent est dans un lieu de type « Autre », comme présenté dans la figure 273.a. Il ne reste plus qu'à ajouter les lieux de type commerces (figure 273.d), et intersecter les probabilités calculées en 273.c pour obtenir une liste de lieu dont les probabilités d'être visitées sont conditionnées par la localisation de l'activité précédente et dans une moindre mesure par la localisation du domicile (figure 274). L'agent n'a plus qu'à tirer un lieu de type « commerce » selon les probabilités définies.

Nous pourrions également ajouter une autre couche d'information géographique qui prend en compte les niveaux de fréquentation desdits commerces à l'instant t pour contraindre l'agent à visiter une zone commerciale plus populaire.

Néanmoins les mobilités des individus peuvent être contraintes par la localisation de leur domicile, plus ou moins excentrée et ayant donc une accessibilité différenciée à certains services. Nous pourrions alors nous baser sur des distributions de distances similaires à la figure 272, mais obtenues selon les différents groupes d'agents et selon leur sous-districts de domicile. Aussi, les profils de distances peuvent être différents selon la localisation de l'activité dans la ville. Nous pourrions alors tirer une distance, selon des distributions de distances dépendant du groupe auquel l'agent appartient, de la localisation du domicile, le tout selon le secteur où l'activité est réalisée.

Utiliser les temps de transports et les données sur l'utilisation du temps ?

Les lieux qui se rapportent à des loisirs ou à de la consommation de services dans des zones commerçantes sont partagés par un grand nombre d'individus. Il est alors possible, comme précédemment, d'utiliser des modèles gravitaires ou encore les distributions des distances entre activités pondérées par leur attractivité horaire pour en estimer la localisation. Mais estimer la localisation de l'activité principale revêt des aspects un peu plus subtils car elle n'est pas forcément située dans une zone dédiée (*i.e.* Comme la Défense à Paris), mais possiblement dans des zones mixtes. Aussi, si nous voulons définir la localisation de l'activité principale d'après nos données *Twitter*, et en appliquant l'une des méthodes présentées précédemment, les lieux potentiels seraient alors ceux fréquentés par notre échantillon d'environ 17 000 personnes, ce qui entraînerait un tropisme probablement trop important vers certaines zones, tandis que personne ne travaillerait dans certains secteurs de la ville. Compte tenu de la qualité des données *Facebook*, l'utilisation des *check-in* réalisés dans une catégorie de type « office » ne parait pas non plus appropriée.

Nous proposons ici une méthode basée sur un modèle d'attractivité qui pourrait permettre d'affecter plus judicieusement la localisation j d'une telle activité d'une personne, sachant son

lieu de résidence *i*.

Pour cela, nous posons que la probabilité qu'une personne habitant dans un secteur *i* travaille en *j* dépend d'un niveau d'attractivité A du lieu j, avec un coefficient d'ajustement β, divisé par un indice de friction entre *i* et *j* (une distance ou une durée), à la puissance α. Le tout est ensuite divisé par la somme des niveaux d'attractivités A dans toute la ville sur les distances entre tous les secteurs, avec les mêmes coefficients β et α, multiplié par la population du secteur d'origine N_i.

$$P_{(i \to j)} = N_i \frac{A_j^\beta / d_{ij}^\alpha}{\sum A^\beta / d^\alpha} \tag{34}$$

Ceci revient à dire que les flux de *i* vers *j* ne dépendent que de la population de *i* et du rapport entre le niveau d'attractivité de *j* pondérée par la distance (ou durée) entre *i* et *j* sur la somme de tous ces rapports. Le fait d'appliquer $(A_j^\beta / d_{ij}^\alpha) / \sum A^\beta / d^\alpha$ fait varier ce terme entre 0 et 1, ce qui implique que lorsque l'on multiplie par la population d'origine N_i nous obtenons une conservation de la population entre les zones de départs *i* et d'arrivés *j*.

La variation des paramètres α et β influence les flux entre *i* et *j*, et donc les durées ou distances moyennes de déplacements résultantes. Un β élevé signifie que plus la zone possède de *POI*, plus le potentiel d'attraction est important. Un α positif implique une relation décroissante entre la distance / durée et la répartition de la population à la destination. Les paramètres α et β peuvent être utilisés pour calibrer le modèle de telle sorte que la moyenne des durées ou distances de déplacements moyens tendent vers des données de références, comme l'utilisation du temps.

Ainsi, dans l'exemple suivant basé sur les données de Bangkok, nous définissons l'attractivité A_j, soit le potentiel qu'une personne travaille dans un lieu *j*, comme dépendant de la densité de *POI* de *Google*, sans distinction de type de lieu, dans une maille d'un kilomètre carré. Nous partons ici du principe que plus il y a de *POI* dans une maille, plus il y a de chance que de nombreuses personnes y effectuent leurs activités principales. Nous discrétisons dans cet exemple le nombre de *POI* en décile, selon des intervalles égaux, où A_j varie entre 1 et 10. Nous ajoutons une puissance α pour amplifier cette attractivité A_j. À noter que nous aurions pu utiliser d'autres données, comme les *check-in Facebook* ou les données *Twitter* globales enregistrées en journée.

Dans cet exemple, nous posons un coefficient d'attractivité β à 2, et un coefficient de rugosité α à 1 et utilisons d_{ij} comme étant la durée des temps de transports par voiture, en conditions normales, un lundi à 7 h du matin, d'après les données de *Google Transit* (chapitre 10). Nous obtenons un temps de déplacement moyen de 41.47 minutes pour se rendre à l'activité principale, soit un peu moins que les 51 minutes quotidiennes passées en moyennes dans les

déplacements pour effectuer un travail formel d'après le *time use census* de 2009 (NSO, 2009). Néanmoins, nous ne considérons pas ici les autres modes de transports (bus, métro) et les paramètres α et β sont posés de manière arbitraire, sans chercher pour l'instant à calibrer au mieux les données même si cette dernière opération repose juste sur approche itérative. La distribution des temps de trajets moyens entre chaque maille - soit la somme des temps de transports entre chaque maille d'origine (domicile) et destination (activité principale) pondérée par la population de la maille d'origine - est visible sur la figure (figure 275) ci-dessous. Dans cet exemple absolument pas représentatif, la plupart des personnes ont des temps de trajet entre leur lieu de domicile et leur activité principale proche de 35 minutes, avec cependant des personnes qui prendraient plus d'une heure pour se rendre à leur travail.

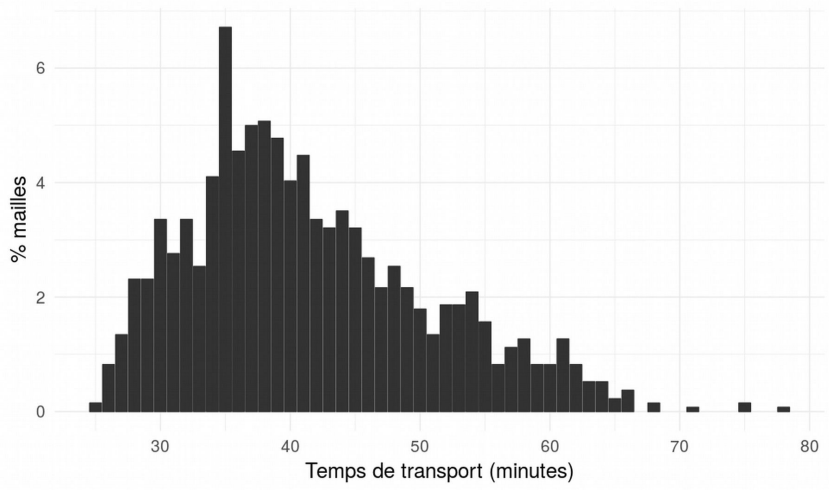


FIGURE 275 Distribution des temps de trajet entre une maille d'origine (domicile) et de destination (réalisation de l'activité principale). D'après l'algorithme exposé précédemment.

Si nous avions les temps de transports individuels des données du *time use census*, nous pourrions calibrer notre petit modèle de sorte que la distribution des temps de transports observés se rapproche de ceux simulés. Il conviendrait pour cela de nouer des partenariats avec les instituts de statistiques locaux, afin d'obtenir des données "référence" désagrégées.

Néanmoins, à partir de ces résultats préliminaires, il est maintenant possible de définir comment la population des sous-districts de la ville sera répartie dans les différentes mailles. Ceci est montré dans la figure 276, ci-dessous où les *Khwaengs* de départ sont signalés par un contour en gras. Suivant la répartition des *POI* dans la ville et au gré de l'accessibilité des différents secteurs, les habitants n'auront pas les mêmes chances de fréquenter des mailles données selon la localisation de leur domicile. Et si le centre est extrêmement attractif, plus les personnes vivent en périphérie, plus elles auront de chance de se tourner vers des secteurs plus accessibles à leur lieu de résidence.

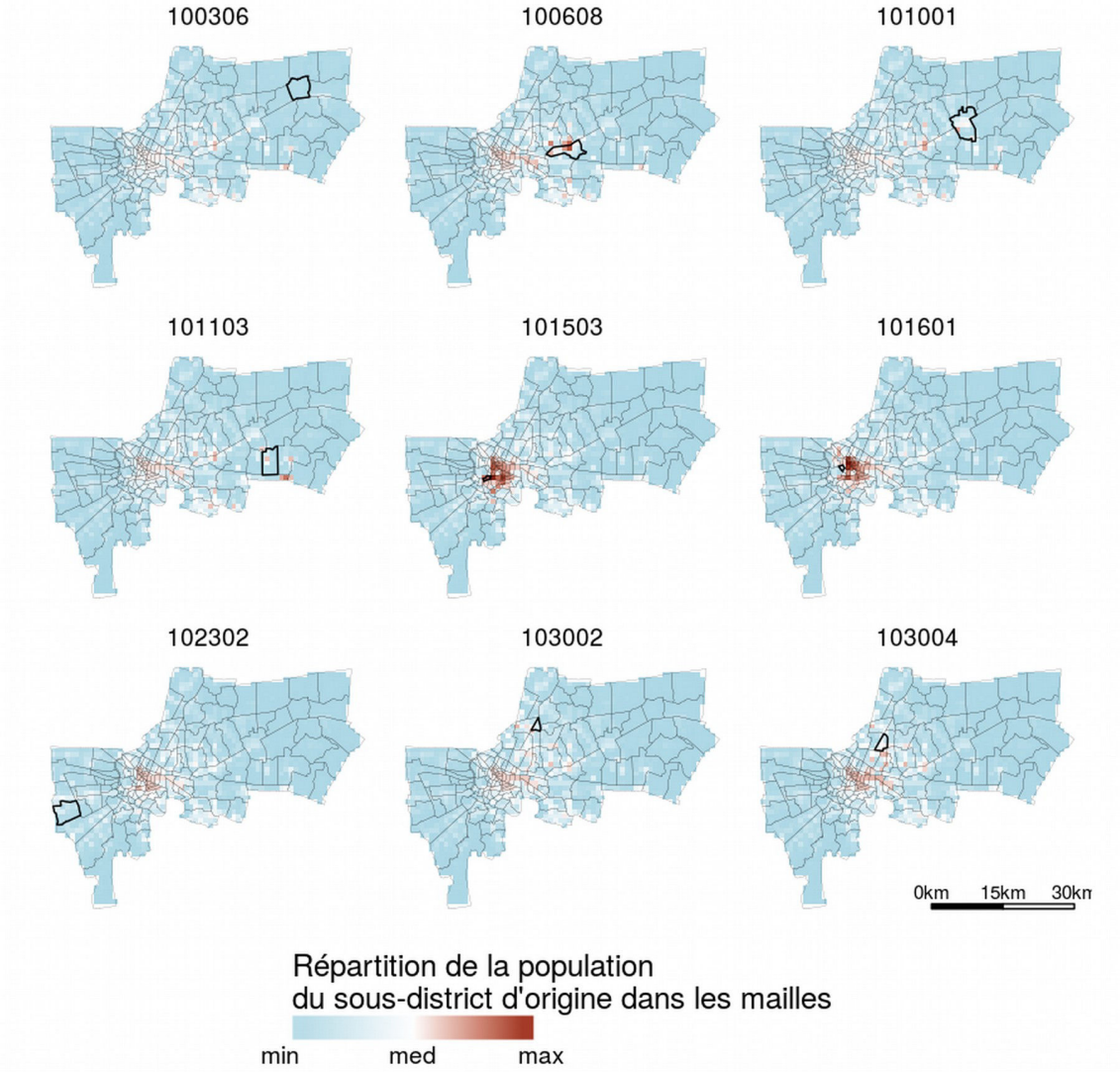


FIGURE 276 Estimation localisation des activités principales selon le sous-district de domicile (polygone noir). D'après un modèle d'attractivité, calculé ici en prenant en compte la répartition spatiale des *POI* et la durée des trajets. La population de chaque sous-district est ainsi répartie dans Bangkok selon la probabilité de réaliser l'activité principale dans une maille donnée.

Plusieurs méthodes d'attribution d'une localisation aux différentes activités sont ainsi envisageables, avec des philosophies assez variées. Leur implémentation et comparaison sera réalisée lors de prochains travaux.

Synthèse

Dans ce dernier chapitre nous avons abordé les mobilités individuelles à Bangkok, en nous basant sur les données récoltées sur *Twitter*.

Nous avons tout d'abord estimé l'activité principale des membres de l'échantillon et constaté que la moitié d'entre eux seraient des étudiants. Nous avons ensuite reconstitué des agendas probables pour chaque personne, selon une méthode proche de celle employée dans le chapitre 8, quoique plus simple.

Les individus n'ont pas tous les mêmes tendances de déplacements, et nous avons effectué des groupes selon des paramètres de dispersions et le nombre de lieux visités et d'activités réalisées. Il ressort que ces différentiels de mobilités ont un encrage géographique parfois assez marqué. En effet, les personnes ne se déplacent pas sur les mêmes distances et effectuent un nombre d'activités différent selon leur localisation dans la ville. Ainsi, la structure de la ville et la localisation de l'offre influencent grandement les mobilités urbaines et la réalisation des activités.

À partir de là, nous avons repris et modifié notre algorithme de génération d'agendas élaboré dans le chapitre 8, en prenant aussi en compte les différentes tendances de déplacements selon les groupes créés précédemment. Nous avons notamment testé différentes manières d'agencer les activités qu'un agent va effectuer lors d'une journée. Nous avons mobilisé d'une part une matrice de transition horaire entre activités dérivée des agendas et d'autre part les profils de fréquentations horaires des activités. Cette dernière méthode fournit les résultats les plus proches de ceux observés. Elle offre également l'avantage de permettre une utilisation d'autres sources de données de fréquentation que les *tweets*, comme les *check-in* de *Facebook*.

Enfin, nous avons proposé quelques pistes qui devraient permettre d'attribuer une localisation à chacun des lieux présents dans l'agenda d'un agent. Ces méthodes suivent des philosophies assez différentes et il conviendrait de les tester et de les comparer dans des travaux ultérieurs.

Conclusion générale

Cheminement

Objectif initial

L'objectif initial de ce travail doctoral était d'étudier et de modéliser les déplacements en milieux urbains pour explorer leurs effets sur la diffusion des épidémies de dengue, à Delhi et Bangkok. Nous devions pour cela nous baser sur des enquêtes de terrains et utiliser des données issues de la téléphonie mobile. Mais faute d'accords avec les opérateurs[411], nous nous sommes d'abord recentré sur la collecte de données *in situ*.

Partir d'enquêtes de terrain à Delhi

L'étude des mobilités urbaines dans des mégapoles asiatiques n'est pas une chose aisée. Ainsi, au niveau d'observation de la rue, il est très délicat de répondre aux simples questions que nous nous sommes posées, à savoir : « où vont tous ces gens et pour quoi faire ? ». Aussi, il y a-t-il des déterminants socio-économiques qui sont susceptibles d'influencer ces déplacements ? (chapitre 2, 3 et 7). Par exemple dans quelle mesure le niveau de richesse impacte les mobilités individuelles ? Les plus pauvres sont-ils plus sédentaires, ou *a contrario* contraints d'effectuer de plus longues navettes pour se rendre sur leur lieu de travail ? Lorsque l'on se promène dans certains quartiers de Delhi, comme à Nehru Place, se pose aussi la question « où sont les femmes ? ». Elles sont globalement moins nombreuses que les hommes à Delhi, mais tout de même !

Suivant l'aphorisme que l'on prête à Confucius, *« qui déplace des montagnes commence par les petites pierres »*, nous avons commencé par faire une enquête de terrain pilote, dans le sud de Delhi, à Malviya Nagar. L'intérêt de cette zone de la ville est qu'elle regroupe notamment quasiment tous les types de quartiers de Delhi, et donc potentiellement un large spectre socio-économique. Il ne restait plus qu'à interroger des personnes du quartier sur leurs mobilités dans la ville pour éclairer notre questionnement initial (chapitre 7).

Si collecter des récits de vie est tout à fait passionnant, nous avons gardé en mire le contexte de notre thèse. Cette dernière s'inscrit dans un plus large projet de modélisation

411. Ce qui toutefois peut se comprendre. Comment réagiraient les opérateurs de téléphonies français si des chercheurs indiens demandaient a accéder à leurs données pour étudier l'impact des déplacements sur la propagation des épidémies de grippe à Paris par exemple ?

à base d'agents des différents éléments qui contribuent à la propagation des épidémies de dengue en milieu urbain (chapitre 1), dont les mobilités font partie (en plus de l'environnement et du moustique). Notre démarche fut donc plus quantitative que qualitative, dans le sens où les réponses aux questions devaient pouvoir être retranscrites spatialement et analysables statistiquement.

Le virage des traces numériques et un recentrage sur Bangkok

Mais alors que nous pensions simplement faire d'autres enquêtes dans d'autres secteurs de Delhi, nous nous sommes rendu compte qu'il était possible (et finalement assez simple) de collecter des traces numériques géolocalisées individuelles provenant du réseau social *Twitter*. Nous étions alors à la fois époustouflés par ces volumes de données disponibles en temps réel et leurs potentiels évidents dans la recherche sur les mobilités, mais aussi abasourdis par ce qu'elles dévoilent sur la vie privée des personnes où comment il est possible de faire des déductions (peut être biaisées) sur les modes de vie à partir de simples *tweets* publics et géolocalisés. Cette thèse allait alors prendre une autre tournure, et se rapprocher du projet initial d'utiliser les données de téléphonie mobile pour analyser et tenter de modéliser les mobilités[412].

Il convenait tout d'abord de comprendre ces données et de ne pas se laisser impressionner par leur volume. Car finalement, la présence de bots perturbe l'analyse et se pose la question de la représentativité de l'échantillon. D'après la littérature, *Twitter* serait surtout utilisé par les plus jeunes (les moins de 35-40 ans), soit la population la plus à risque vis-a-vis de la dengue en Thaïlande et en Inde, ce qui est de bonne augure pour nos travaux. Nous avons ensuite effectué divers filtres et prétraitements afin d'épurer la base (chapitre 6). Nous réalisons alors que ces tweets ne peuvent pas être utilisés à Delhi pour modéliser les déplacements de l'ensemble de la population, car non significatifs. L'usage de *Twitter* à Delhi concernerait surtout les classes moyennes supérieures. En revanche, les niveaux de représentativités spatiales obtenus à Bangkok sont assez satisfaisants pour autoriser des analyses plus poussées.

Ainsi, plutôt que d'essayer de modéliser directement les mobilités urbaines individuelles, nous avons choisi d'insister sur les aspects méthodologiques et l'analyse des données. Ceci nous permet de produire une base suffisamment riche et significative que nous pourrons utiliser sereinement par la suite.

Les premières visualisations des pulsations urbaines à Bangkok se sont avérées très séduisantes, faisant ressortir un caractère monocentrique très marqué, avec des flux dominants de type centre-périphérie, et quelques pôles secondaires répartis dans le péricentre. Il est également possible d'étudier plus précisément les interactions entre les différents secteurs de la

412. Projet qui n'avait pu aboutir faute de trouver des accords avec des opérateurs de téléphonie tant à Delhi qu'à Bangkok

ville et de définir des ensembles de quartiers fonctionnels (chapitre 9 et 10).

Mais ces données longitudinales impliquent qu'il est possible de connaître les lieux qu'un individu a fréquentés, ce qui dégage des sentiments assez dérangeants, entre espionnage et voyeurisme. Nous avons donc pris quelques précautions en ajoutant du flou à nos données collectées afin de réduire les possibilités de ré-identification des individus, en parallèle aux recommandations de la CNIL sur la sécurisation de la base de données (chapitre 4).

Mais les traces numériques géolocalisées ne sont pas l'apanage des utilisateurs de *Twitter*. Nous sommes tous concernés, pour peu que l'on possède un téléphone portable, notamment un *smartphone*, un compte sur un réseau social en ligne, une carte de bus ou encore une carte bleue (chapitre 4 et 5). Il nous a paru nécessaire de creuser le sujet pour comprendre le mode de création de ces données et leur usage dans la recherche où ailleurs. Si l'utilisation des traces numériques dans le cadre de recherches scientifiques parait pertinente (chapitre 5), elle revêt également un intérêt économique substantiel, servant de base aux modèles économiques de toute une industrie avec les nombreuses dérives qui peuvent en découler. Sans compter leur utilisation par des états désireux de surveiller (une frange de) leur population.

Mieux connaître l'échantillon

Dans tous les cas, si un objectif est de mieux connaître son échantillon, ses clients ou ses administrés, ces traces numériques peuvent être croisées avec d'autres informations, afin de compléter les profils individuels ou identité numérique (chapitre 4).

Dans notre cas, nous avons basé notre étude des mobilités sur le concept de l'espace d'activité. Ce dernier permet à la fois d'apprécier qualitativement les différents lieux visités et activités effectuées par un individu et qui, selon une approche plus quantitative autorise une implémentation dans un système à base d'agent sans trop de contraintes. Si les enquêtes de terrain permettent d'obtenir directement ces informations au gré de questionnaires adaptés, les *tweets* géolocalisés ne nous donnaient que des indications sur les lieux fréquentés par un individu à différents moments, sans pour autant savoir quelle activité était effectuée. Dans l'optique de créer un modèle de mobilité nourris par des données compatibles entre elles, il convenait de trouver des méthodes permettant d'inférer aux membres de notre échantillon *Twitter* une activité probablement réalisée dans un lieu donné.

Notre hypothèse principale fut que le lieu depuis lequel un message est envoyé est intimement lié à l'activité que la personne est en train de réaliser. Par exemple, si un tweet est émis depuis un *mall*, la personne est probablement en train de faire du shopping, du lèche-vitrine ou d'y travailler. En l'absence de données institutionnelles accessible sur l'utilisation du sol, nous nous sommes d'abord penchés sur la base géographique libre *OpenStreetMap* Mais la

couverture de cette dernière était trop partielle à Delhi et Bangkok.

Nous réalisons aussi que la plupart des acteurs majeurs d'Internet mettent à disposition, à l'image de *Twitter*, des interfaces permettant d'accéder à une partie de leur base de données moyennant quelques ajustements et bricolages. C'est ainsi que nous avons collecté les points d'intérêts de *Google Maps a priori* assez exhaustifs, à partir desquels nous avons développé diverses méthodes permettant de caractériser l'utilisation du sol (chapitre 8 et 9). Appliquées à Bangkok, nous avons pu faire ressortir les différents assemblages de quartiers qui composent la ville.

Ne restait plus qu'à croiser les *tweets* géolocalisés à la couche géographique créée pour obtenir des espaces d'activités individuels (chapitre 8 et 11). Collectivement, nous obtenons des profils de fréquentation temporels des différentes activités à Bangkok. Si ces derniers sont assez intuitifs, montrant par exemple que les écoles sont plus fréquentées le matin que le soir, ou que les lieux de sorties sont plus visités les vendredis et samedis soirs, ils permettent de quantifier les volumes de fréquentation horaires.

Collecte de données tous azimuts

Mais à ce stade, ces tendances générales de fréquentation ne sont observées qu'avec le prisme des données *Twitter*. Il convenait de trouver d'autres données permettant d'établir des points de comparaison. À court de temps, nous n'avons malheureusement pas pu faire d'enquêtes de terrain à Bangkok.

C'est alors que nous avons jeté notre dévolu sur *Facebook*, lorsque nous avons remarqué que la compagnie propose à ses usagers d'effectuer des *check-in*. Cette action permet à un utilisateur de signaler sa présence dans un lieu de la base, par exemple un restaurant, un hôtel ou une université. Il était ensuite très aisé de récupérer les informations agrégées, permettant de savoir combien de personnes se sont enregistrées dans un lieu lors d'une tranche horaire. Après avoir collecté ces données, d'un volume bien plus conséquent que celui des *tweets*, nous avons appliqué des filtres afin de corriger divers artefacts (chapitre 6). Nous avons ainsi pu apprécier la cohérence de notre couche d'utilisation du sol tout en ayant enfin un point de comparaison avec les données *Twitter* (chapitre 9).

Si les profils de fréquentation temporels des différents types de lieux sont assez proches, il existe des écarts parfois assez importants entre les données *Twitter* et les *check-in* de *Facebook*, surtout dans les niveaux d'attractivités des principaux pôles de Bangkok (chapitre 10). Et nous ne sommes pas en mesure de savoir quel jeu fourni les indications les plus proches de la réalité. Peut-être que l'acquisition de données téléphoniques pourrait permettre de trancher.

Nous avons décidé de continuer la collecte de données auprès d'acteurs majeurs du web, car ces derniers possèdent énormément d'informations géographiques potentiellement intéressantes dans le cadre d'études sur les mobilités urbaines. Nous avons par exemple enregistré toutes les 30 minutes les conditions de circulations à Bangkok que met à disposition *Microsoft* via son service *Bing Trafic*, ainsi que les temps de transports entre les différents secteurs de la ville fournis par *Google Transit*. Ces données permettent de mieux comprendre les conditions de circulation dans la capitale thaïlandaise (chapitre 2 et 10). Elles peuvent aussi servir à calibrer ou valider des modèles de déplacement et raisonner en termes de temps de transports plutôt que sur des distances.

Modéliser

Car finalement, un des objectifs de cette thèse est de proposer des pistes pour l'élaboration d'un modèle de mobilité à base d'agents. Ce dernier sera associé à un modèle environnemental et un modèle du moustique, dans l'optique de mieux comprendre les mécanismes de propagations de la dengue (chapitre 1 et 3). Et nous avons alors à notre disposition les éléments nécessaires pour faire un tel modèle de mobilités qui suit le concept de l'espace d'activité, à savoir :

- Des données de mobilités individuelles (*tweets* ou enquêtes de terrain), qui permettent de connaître les différents lieux fréquentés dans la ville pour chaque individu. Il est ainsi possible d'en déduire des lois de mobilités, par exemple la distribution des trajets moyens ou la dispersion des individus dans le temps et l'espace.

- Une couche géographique sur l'utilisation du sol (*OSM, Google Maps*), pour estimer l'activité réalisée par les utilisateurs de Twitter.

- Des données de fréquentations agrégées (*check-in*) qui peuvent permettent de calibrer le modèle, servir d'information sur l'attractivité des différents secteurs de la ville ou encore effectuer une validation.

- Des informations sur les conditions de circulations (*Google Transit* et *Bing Traffic*) qui peuvent intervenir dans le choix des modes de transports ou des lieux qu'un individu va fréquenter.

L'idée est d'obtenir pour chaque agent créé au cours de l'initialisation du modèle multi-agent un agenda spatialisé, soit une succession d'activités de durées variables, qu'il réalise dans des lieux donnés. Il existe plusieurs approches pour atteindre un tel objectif. Tout d'abord l'activité réalisée par un agent à un pas de temps donné peut être définie de manière synchrone avec l'attribution d'une localisation dans l'espace. Nous avons ici fait le choix de raisonner en deux temps, à savoir d'abord définir un agenda sur une période (un mois), puis affecter une

localisation à chacun des lieux qu'un agent visitera.

Pour définir nos agendas, nous partons du principe que l'entrée principale du modèle doit être des espaces d'activités individuels, ce qui permet d'un point de vue théorique de s'affranchir de la nature des données (chapitre 7, 8 et 11). Il suffit que ces dernières soient structurées sous forme de séries de lieux où chaque individu effectue des activités à différents moments et fréquences. Ces données peuvent alors être des *tweets*, comme à Bangkok ou Delhi, des enquêtes de terrain, des données téléphoniques, ou éventuellement des enquêtes ménage-déplacement retravaillées.

La première étape passe par la création d'agendas continus dans le temps à partir de données épisodiques. Connaissant les heures et fréquences de visites et en estimant les durées des activités, nous pouvons alors reconstituer un agenda probable sur une période définie pour chaque membre de l'échantillon en entrée (qu'il s'agisse des utilisateurs de *Twitter* ou de personnes enquêtées).

À partir de ces agendas reconstitués, nous pouvons dériver d'autres informations afin de générer des agendas pour des agents a proprement parlé. Nous avons proposé différents algorithmes qui suivent tous la même logique, soit une attribution « étape par étape » des différents éléments de l'espace d'activité d'un agent.

Tout d'abord, il convient de définir pour chaque agent un nombre de lieux et un nombre d'activités qu'il va réaliser. Ces informations sont déduites de l'ensemble des agendas reconstitués, ou d'un sous-groupe de la population éventuellement créé, auquel l'agent appartiendra (chapitre 7, 8 et 11).

Ensuite chaque activité se verra attribuer une fréquence de visite et des jours de réalisation. Ces informations peuvent provenir des agendas reconstitués, d'enquêtes ménage-déplacement, où alors de données agrégées comme les *check-in* de *Facebook*.

À partir des agendas reconstitués, nous pouvons dériver les distributions sur les durées des activités même s'il est possible d'utiliser des sources de données extérieures, comme les enquêtes sur l'utilisation du temps[413]. Nous pouvons aussi déduire de ces agendas les taux de transition horaires entre deux activités, qui permettent de savoir par exemple quelle est la probabilité qu'une personne quitte son domicile à 6 h du matin pour faire une balade dans un parc.

Finalement, il ne reste plus qu'à organiser une journée pour chaque agent, en partant du principe qu'elle commence au domicile. Nous pouvons alors définir quelle sera sa prochaine

413. À noter que Google possède également des informations sur la durée moyenne des personnes dans les différents lieux de sa base.

activité parmi celles prévues cette journée en utilisant soit :

- Une matrice de transition, qui nous donne la probabilité d'effectuer l'activité suivante sachant que l'agent est en train de réaliser une activité donnée.

- Les profils temporels de fréquentation des différentes activités (*Twitter* et utilisation du sol, ou alors les *check-in Facebook*). Si par exemple l'agent doit choisir entre aller dans un lieu de sortie dense ou dans un parc alors qu'il est 8 h du matin, il aura plus de chance d'aller dans un parc compte tenu des profils horaires des activités.

Chaque agent se voit donc attribuer un agenda individuel dérivé des caractéristiques du groupe auquel il appartient, avec une part d'aléa dans la réalisation des activités non routinière. Ces différents groupes peuvent être des tranches d'âges, des caractéristiques socio-économiques, ou plus naïvement différents potentiels de déplacements (des groupes d'agent plus ou moins mobiles, selon les données de l'échantillon en entrée et la localisation du domicile dans la ville).

Discussion et perspectives

Les résultats du modèle sont globalement satisfaisants (chapitres 8 et 11), dans le sens où nous arrivons à reproduire relativement correctement une bonne partie des profils de fréquentations des différents types de lieux. Mais il est toujours possible d'améliorer l'algorithme. Ce dernier aspect est rendu plus aisé par l'approche « étape par étape », où le modélisateur peut apporter des modifications à tout moment sur la façon de définir les fréquences de visites, les durées des activités ou la manière d'organiser une journée.

Aussi, notre algorithme dépend plus des types d'informations collectées que de la source des données. Il peut donc être appliqué dans différentes villes du monde, pour peu que les données adaptées (e.g. enquêtes ménage-déplacement, données téléphonies, etc.) soient à disposition. Car notre méthode nécessite toutefois des données longitudinales, pour connaître le nombre de lieux fréquentés et la fréquence de réalisation des d'activités.

Concernant l'affectation d'une localisation aux activités réalisées par un agent, nous n'avons à ce stade que proposé différentes pistes (chapitre 7 et 11) qu'il conviendra de développer dans des travaux ultérieurs. Aussi, Bangkok est une ville extrêmement embouteillée, et il conviendrait de prendre en compte le mode et le temps de transport dans notre modèle de mobilité.

S'il est possible que deux agents aient des agendas très proches, voire similaires, la réalisation de ces activités de manière synchrone dans l'espace devrait avoir très peu de chance de se produire. Ceci permet de conserver une part d'ipséité, définie comme : « Ce qui fait

qu'une personne, par des caractères strictement individuels, est non réductible à une autre »[414], conférant à chaque agent une signature spatio-temporelle qui lui est a priori propre (du moins à une résolution assez fine).

Toutefois, dans le cadre de simulations pour le modèle MO3, nous ne pourrons observer des phénomènes d'émergences dans les pratiques de mobilités, au sens où notre approche implique que les actions des agents sont déterminées à l'avance et ces derniers n'interagissent pour l'instant pas entre eux. Ils n'auront donc pas de libre arbitre. Ceci rappelle le scolie de la proposition 35 de la deuxième partie de l'éthique de Spinoza : *« Les hommes, donc, se trompent en ce qu'ils pensent être libres ; et cette opinion consiste en cela seul qu'ils sont conscients de leurs actions, et ignorants des causes par lesquelles ils sont déterminés »* (Spinoza, 1677, p99).

Nous n'avons pas le bagage philosophique suffisant pour discuter cette citation lorsqu'elle s'applique à l'Homme, mais elle semble adaptée à nos agents qui sont prédéterminés à effectuer des séries d'actions au cours du temps sans avoir à prendre de décision. Lorsque ce sous-modèle sera terminé il sera combiné aux autres sous-modèles (environnements et hôtes) du modèle MO3 et cette approche, qui revient à définir en amont les mobilités individuelles des agents dans la ville devrait permettre de gagner du temps de calcul.

C'est aussi à ce moment-là que nous pourrons voir émerger des épidémies de dengue dans différents secteurs de la ville au gré des mobilités urbaines et de l'évolution des conditions environnementales et des facteurs qui influencent les comportements des moustiques. Mais se posera toujours la question de la validation, tant sur les mobilités que sur les cas de dengue simulés. Car ce dernier aspect dépend grandement des cas enregistrés par les systèmes de surveillance. De plus, la majorité des personnes sont asymptomatiques et contribuent grandement à la propagation de l'épidémie (chapitre 1).

Une tendance assez généralisée en modélisation est de calibrer les modèles de sorte à trouver les meilleurs paramètres d'ajustements qui permettent d'obtenir le meilleur accord entre les données observées et simulées. Mais compte tenu des volumes de données, il parait plus simple et réaliste d'utiliser des statistiques globales des tendances de déplacement que des lois de puissances tronquées régies par des coefficients α ou β obtenus par diverses méthodes de régressions ou après des milliers d'itérations.

Concernant les bases de données massives, elles permettent dans beaucoup de cas d'éclairer les tendances des mobilités urbaines, qu'elles soient globales ou individuelles, et peuvent être utilisées sans trop de difficultés dans des modèles. Mais ces données, dont nous n'avons fait ici qu'emprunter[415] un échantillon pour les besoins de la recherche, restent bien

414. http://www.cnrt .fr/defin t on/ psé té
415. Ou réquisitionner, selon les points de vue.

souvent la propriété de grands groupes privés côtés en bourses. En plus d'être libres de modifier leur politique de partage ou devrait-on plutôt dire de « mise à disposition à des tiers » ces derniers sont théoriquement très bien placés pour comprendre les comportements individuels et les mobilités urbaines, mais préfèrent probablement vendre des profils de leurs utilisateurs à des annonceurs et des solutions de « smart-city » à des municipalités aisées. Se posent alors des questions de souveraineté et de dépendance vis-à-vis de ces acteurs.

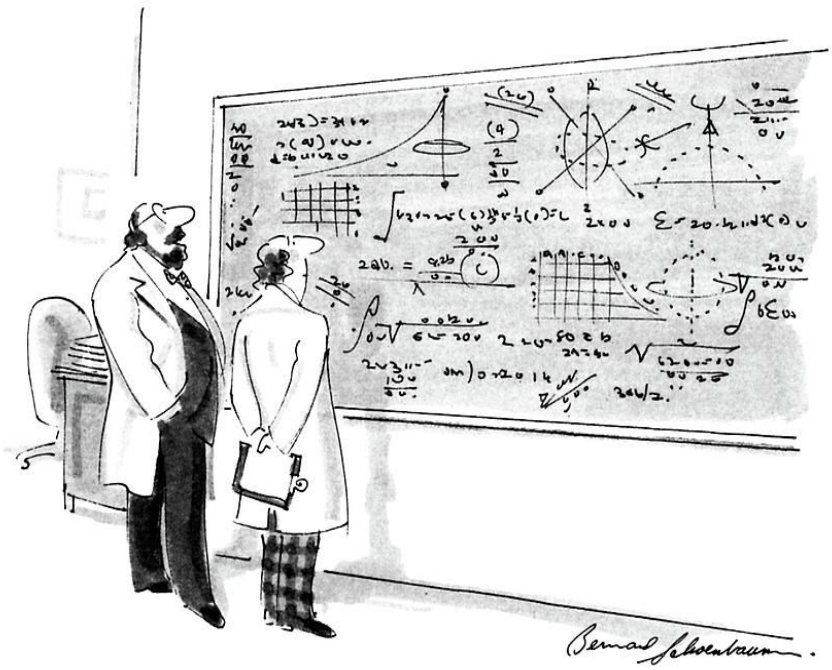
Quoi qu'il en soit, ces bases de données géographiques massives devraient être utilisées conjointement à des études de terrain, sous peine d'occulter des réalités sociales propres à certaines zones, ou d'appliquer des raisonnements valables dans la région où travaille le modélisateur, mais pas nécessairement dans la ville qu'il étudie à distance. Ainsi, notre grand regret est de ne pas avoir pu passer plus de temps à Bangkok pour faire des enquêtes dans différents secteurs de la ville, afin de questionner plus précisément le quotidien des Bangkokois et leur rapport à la ville, au travail, aux loisirs et aux *malls*, ou encore aux transports.

Mais dans tous les cas, il est extrêmement compliqué de modéliser fidèlement les phénomènes sociaux (figure 277), et par extension les mobilités urbaines individuelles, même si des tendances stables apparaissent clairement dans nos analyses à l'échelle urbaine.

La plupart des déplacements individuels sont en effet plutôt routinier (navettes domicile-travail), et les activités de loisirs sont relativement figées dans le temps et l'espace (par exemple des lieux de sorties du vendredi soir concentrés dans quelques quartiers). Collectivement, cela explique en très grande partie les régularités des fréquentations temporelles des différents secteurs de la ville.

Mais les mobilités individuelles sont aussi faîtes de micro-comportements qui vont dépendre notamment des habitudes et des potentiels de déplacement et de réalisation d'activités de chacun. Et il n'existe aucune loi universelle pouvant les décrire.

Ces singularités de mobilité à échelle fine dans un système urbain complexe, combinées à des facteurs environnementaux locaux sont très probablement à l'origine des départs d'épidémies. La prise en compte de ces facteurs dans des modèles devrait donc permettre de faire ressortir les conditions à l'origine de ces émergences. Ceci permettrait d'éclairer l'épidémio-genèse, avant que les grandes tendances (flux de déplacements massifs, conditions météorologiques) ne prédominent dans la dynamique de propagation des épidémies.



"Oh, if only it were so simple."

FIGURE 277 "Ah ! Si seulement c'était si simple !", Bernard Schoenbaum, *New Yorker Cartoon.*

Bibliographie

Abbasi, O.R., Alesheikh, A.A., Sharif, M., 2017. Ranking the City : The Role of Location-Based Social Media *check-in* in Collective Human Mobility Prediction. ISPRS Int. J. Geo-Inf. 6, 136. https://doi.org/10.3390/ijgi6050136

Abel, F., Hauff, C., Houben, G.-J., Tao, K., Stronkman, R., 2012. Semantics + Filtering + Search = Twitcident - Exploring Information in Social Web Streams, in : Proceedings of the 23rd ACM Conference on Hypertext and Social Media. pp. 285 294.

Adam, A., Charlier, J., Debuisson, M., Duprez, J.-P., Reginster, I., Thomas, I., 2018. Bassins résidentiels en Belgique : deux méthodes, une réalité ? Espace Géographique 47, 35. https://doi.org/10.3917/eg.471.0035

Adams, B., Holmes, E.C., Zhang, C., Mammen, M.P., Nimmannitya, S., Kalayanarooj, S., Boots, M., 2006. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. Proc. Natl. Acad. Sci. 103, 14234 14239. http://www.pnas.org/content/103/38/14234.short.

Aguiar, M., Stollenwerk, N., Halstead, S.B., 2016. The risks behind Dengvaxia recommendation. Lancet Infect. Dis. 16, 882. Lien : http://search.proquest.com/openview/264c9137a5fe1f8b973ed9a81bdc2297/1?pq-origsite=gscholar&cbl=44001.

Ajelli, M., Gonçalves, B., Balcan, D., Colizza, V., Hu, H., Ramasco, J.J., Merler, S., Vespignani, A., 2010. Comparing large-scale computational approaches to epidemic modeling : agent-based versus structured metapopulation models. BMC Infect. Dis. 10, 190. https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-190.

Akima, H., Gebhardt, A., 2016. akima : Interpolation of Irregularly and Regularly Spaced Data. Lien : https://CRAN.R-project.org/package=akima.

Alessandretti, L., Lehmann, S., Baronchelli, A., 2018. Individual mobility and social behaviour : Two sides of the same coin. arxiv. https://arxiv.org/pdf/1801.03962.pdf.

Alessandretti, L., Sapiezynski, P., Lehmann, S., Baronchelli, A., 2017. Multi-scale spatio-temporal analysis of human mobility. PLOS ONE 12, e0171686. https://doi.org/10.1371/journal.pone.0171686

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol. 58, 240 250. https://doi.org/10.1016/j.trc.2015.02.018

Alhazzani, M., Alhasoun, F., Alawwad, Z., González, M.C., 2016. Urban Attractors : Discovering Patterns in Regions of Attraction in Cities. ArXiv Prepr. ArXiv170108696. https://arxiv.org/abs/1701.08696.

Alonso, W., 1976. A theory of movements. Institute of Urban and Regional Development, Berkeley.

Amini, A., Kung, K., Kang, C., Sobolevsky, S., Ratti, C., 2014. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Sci. 3. https://doi.org/10.1140/epjds31

Anderson, K.B., Chunsuttiwat, S., Nisalak, A., Mammen, M.P., Libraty, D.H., Rothman, A.L., Green, S., Vaughn, D.W., Ennis, F.A., Endy, T.P., 2007. Burden of symptomatic dengue infection in children at primary school in Thailand : a prospective study. The Lancet 369, 1452 1459. https://doi.org/10.1016/S0140-6736(07)60671-0

Andersson, H., 1999. Epidemic Models and Social Networks. Math. Sci. 24, 128 147.

Andraud, M., Hens, N., Marais, C., Beutels, P., 2012. Dynamic Epidemiological Models for Dengue Transmission : A Systematic Review of Structural Approaches. PLoS ONE 7, e49085. Lien : https://doi.org/10.1371/journal.pone.0049085

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., Hecker, D., 2012. Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. KI - Künstl. Intell. 26, 241 251. Lien : https://doi.org/10.1007/s13218-012-0177-4

Angel, A., Angel, B., Yadav, K., Sharma, N., Joshi, V., Thanvi, I., Thanvi, S., 2017. Age of initial cohort of dengue patients could explain the origin of disease outbreak in a setting : a case control study in Rajasthan, India. VirusDisease 28, 205 208. https://doi.org/10.1007/s13337-017-0377-5

Arino, J., 2005. A multi-species epidemic model with spatial dynamics. Math. Med. Biol. 22, 129 142. https://doi.org/10.1093/imammb/dqi003

Arino, J., Van den Driessche, P., 2006. Disease spread in metapopulations. Fields Inst. Commun. 48, 1 13. Lien : http://server.math.umanitoba.ca/~jarino/papers/ArinoVdD_FIC.pdf.

Arino, J., van den Driessche, P., 2003. A multi-city epidemic model. Math. Popul. Stud. 10, 175 193. Lien : https://doi.org/10.1080/08898480306720

Askew, M., 2002. Bangkok : Place, Practice and Representation. Routledge, London.

Åström, C., Rocklöv, J., Hales, S., Béguin, A., Louis, V., Sauerborn, R., 2012. Potential Distribution of Dengue Fever Under Scenarios of Climate Change and Economic Development. EcoHealth 9, 448 454. https://doi.org/10.1007/s10393-012-0808-0

Aubry, M., Yoann, T., Mihiau, M., Anita, T., Marine, G., Didier, M., Van-Mai, C., 2017. High risk of dengue type 2 outbreak in French Polynesia, 2017. Eurosurveillance 22.

Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., Haupt, T., 2002. Observing the rhythms of daily life : A six-week travel diary. Transportation 29, 95 124. http://www.springerlink.com/index/XP1NDWH7NM8A9F54.pdf.

Bacaër, N., 2011. A Short History of Mathematical Population Dynamics. Springer London, London. https://doi.org/10.1007/978-0-85729-115-8

Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can : improving geographical prediction with social and spatial proximity, in : Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 61 70. Lien : http://dl.acm.org/citation.cfm?id=1772698.

Bahu-Leyser, D., 2009. Une éthique à construire, in : Traçabilité et réseaux, Hermès, La Revue. CNRS, Paris, pp. 161 166.

Bailey, N.T.J., 1975. The mathematical theory of infectious diseases and its applications. Griffin, London.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A., 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. 106, 21484 21489. Lien : http://www.pnas.org/content/106/51/21484.short.

Balestier, A., Septfons, A., Leparc-Goffart, I., Giron, S., Succo, T., Burdet, S., 2016. Surveillance du chikungunya et de la dengue en France métropolitaine, 2015. Bull. Epidémiologique Hebd. 564 571.

Ball, F., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L., Scalia Tomba, G., 2015. Seven challenges for metapopulation models of epidemics, including households models. Epidemics 10, 63 67. https://doi.org/10.1016/j.epidem.2014.08.001

Bangkok GIS, 2018. Population and housing statistics. http://www.bangkokgis.com/gis_information/population/.

Banos, A., 2013. Pour des pratiques de modélisation et de simulation libérées en Géographie et SHS. Paris.

Banos, A., Chardonnel, S., Lang, C., Marilleau, N., Thévenin, T., 2005. Une approche multi-agents de la ville en mouvement. Presented at the SMAGET, Les Arcs, p. 17.

Banos, A., Marilleau, N., Thévenin, T., Chardonnel, S., Lang, C., Mas, A.B., 2006. Génération d'emplois du temps individuels pour une simulation multi-agents des mobilités urbaines quotidiennes. Presented at the SAGEO, p. 17.

Banos, A., Thévenin, T., 2011. Generation of Potential Fields and Route Simulation Based on the Household Travel Survey, in : Modeling Urban Dynamics : Mobility, Accessibility and Real Estate Value. John Wiley & Sons, Inc, Hoboken, NJ, USA, pp. 83 102. https://doi.org/10.1002/9781118558041

Barabasi, A.-L., 2005. The origin of bursts and heavy tails in human dynamics. Nature 435, 207 2011. https://doi.org/10.1038/nature03459

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility : Models and applications. Phys. Rep. 734, 1 74. Lien : https://doi.org/10.1016/j.physrep.2018.01.001

Barbosa, H., de Lima-Neto, F.B., Evsukoff, A., Menezes, R., 2015. The effect of recency to human mobility. EPJ Data Sci. https://doi.org/10.1140/epjds/s13688-015-0059-8

Barmak, D.H., Dorso, C.O., Otero, M., 2016. Modelling dengue epidemic spreading with human mobility. Phys. Stat. Mech. Its Appl. 447, 129 140. https://doi.org/10.1016/j.physa.2015.12.015

Barmak, D.H., Dorso, C.O., Otero, M., Solari, H.G., 2011. Dengue epidemics and human mobility. Phys. Rev. E 84. https://doi.org/10.1103/PhysRevE.84.011901

Bassand, M., Brulhardt, M.-C., 1983. La mobilité spatiale : un processus social fondamental. Espace Popul. Sociétés 1, 49 54. https://doi.org/10.3406/espos.1983.902

Bassolas, A., Lenormand, M., Tugores, A., Gonçalves, B., Ramasco, J.J., 2016. Touristic site attractiveness seen through Twitter. EPJ Data Sci. 5. https://doi.org/10.1140/epjds/s13688-016-0073-5

Bastianelli, G., Bignami, A., Grassi, G.B., 1898. Coltivazione delle semilune malariche dell'uomo nell' Anopheles claviger (Sinonimo : Anopheles maculipennis Meig). Atti R. Accad Lincei 7, 313 314.

Batty, M., 2009a. Cities as complex systems : scaling, interaction, networks, dynamics and urban morphologies, in : Encyclopedia of Complexity and Systems Science. Springer, pp. 1041 1071. http://link.springer.com/10.1007/978-0-387-30440-3_69.

Batty, M., 2009b. Urban modeling. Int. Encycl. Hum. Geogr. Oxf. UK Elsevier. Lien : http://www.casa.ucl.ac.uk/rits/BATTY-Urban-Modelling-2009.pdf.

Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., Giovannini, L., 2010. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. J. Stat. Mech. Theory Exp. 2010, P05001. https://doi.org/10.1088/1742-5468/2010/05/P05001

Becker, H., Naaman, M., Gravano, L., 2011. Beyond Trending Topics : Real-World Event Identification on Twitter. Columbia Univ. Comput. Sci. Tech. Rep.

Beiro, M.G., Panisson, A., Tizzoni, M., Cattuto, C., 2016. Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Sci. 5. https://doi.org/10.1140/epjds/s13688-016-0092-2

Bell, M.G.H., 1983. The Estimation of an Origin-Destination Matrix from Traffic Counts. Transp. Sci. 17, 198 217. https://doi.org/10.1287/trsc.17.2.198

Belliappa, J.L., 2013. Gender, Class and Reflexive Modernity in India. Palgrave Macmillan UK, London. https://doi.org/10.1057/9781137319227

Benaglia, T., Chauveau, D., Hunter, D., Young, D., 2009. mixtools : An R package for analyzing finite mixture models. J. Stat. Softw. 32, 1 29.

Benedict, M.Q., Levine, R.S., Hawley, W.A., Lounibos, L.P., 2007. Spread of the tiger : global risk of invasion by the mosquito Aedes albopictus. Vector-Borne Zoonotic Dis. 7, 76 85. Lien : http://online.liebertpub.com/doi/abs/10.1089/vbz.2006.0562

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., Piarroux, R., 2015. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. Sci. Rep. 5, 8923. https://doi.org/10.1038/srep08923

Bennett, S.N., 2014. Taxonomy and Evolutionary Relationships of Flaviviruses, in : Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J., C.A.B. International (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI, Wallingford, Oxfordshire ; Boston, MA, pp. 322 333.

Bernoulli, D., 1760. Sur une nouvelle analyse de la mortalité causée par la petite vérole et les avantages de l'inoculation pour la prévenir, in : Histoire de l'Académie Royale Des Sciences. Académie Royale des Sciences, Paris, pp. 99 107. http://gallica.bnf.fr/ark:/12148/bpt6k3558n/f4.image.langFR

Berthoz, A., 2009. La Simplexité, Odile Jacob. ed.

Berthoz, A., 1997. Le sens du mouvement, Odile Jacob. ed.

Bhan, G., 2013. Planned Illegalities : Housing and the 'Failure' of Planning in Delhi : 1947-2010'. Econ. Polit. Wkly. 48.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J., Hay, S.I., 2013. The global distribution and burden of dengue. Nature 496. https://doi.org/10.1038/nature12060

Biagini, C., 2012. L'emprise numérique : Comment internet et les nouvelles technologies ont colonisé nos vies, l'Echappée. ed, Pour en finir avec.

Bian, L., 2004. A conceptual framework for an individual-based spatially explicit epidemiological model. Environ. Plan. B Plan. Des. 31, 381 395. https://doi.org/10.1068/b2833

Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. EPJ Data Sci. 4. https://doi.org/10.1140/epjds/s13688-015-0046-0

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

BMTA, 2016. Annual Report 2015/2016. Bangkok Mass Transit Authority, Bangkok. http://www.bmta.co.th/sites/default/files/files/download/annaual-report-2559.pdf

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D., 2006. Complex networks : Structure and dynamics. Phys. Rep. 424, 175 308. https://doi.org/10.1016/j.physrep.2005.10.009

Boettcher, A., Lee, D., 2012. EventRadar : A Real-Time Local Event Detection Scheme Using Twitter Stream, in : IEEE International Conference on Green Computing and Communications. pp. 358 367. https://doi.org/10.1109/GreenCom.2012.59

Borja, S., Courty, G., Ramadier, T., 2015. Les mobiles sont-ils tous motiles ? Critiques et questions autour de la motilité et de son capital, in : Kaufmann, V., Ravalet, E., Dupuit, E. (Eds.), Motilité et Mobilité : Mode d'emploi, Espaces, Mobilités et Sociétés. Alphil - Presses Universitaires Suisses, pp. 197 234.

Borker, G., 2017. Safety First : Perceived Risk of Street Harassment and Educational Choices of Women.

Boyer, F., Delaunay, D., 2017. Se déplacer dans Ouagadougou au quotidien, moyens, contraintes et pratiques de la mobilité. Monographies Sud-Nord 77.

Brady, O.J., Golding, N., Pigott, D.M., Kraemer, M.U., Messina, J.P., Reiner Jr, R.C., Scott, T.W., Smith, D.L., Gething, P.W., Hay, S.I., 2014. Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus

transmission. Parasit. Vectors 7, 338. https://parasitesandvectors.biomedcentral.com/articles/10.1186/1756-3305-7-338.

Brady, O.J., Johansson, M.A., Guerra, C.A., Bhatt, S., Golding, N., Pigott, D.M., Delatte, H., Grech, M.G., Leisnham, P.T., Marciel-de-Freitas, R., Styer, L.M., Smith, D.L., Scott, T.W., Gething, P.W., Hay, S.I., 2013. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. Parasit. Vectors 6, 351.

Brancoft, T.L., 1906. On the aetiolgoy of dengue fever. Aust. Med. Gaz. 25, 17 18.

Bretagnolle, A., Daudé, E., Pumain, D., 2006. From theory to modelling : urban systems as complex systems. Cybergeo. https://doi.org/10.4000/cybergeo.2420

Brockmann, D., Hufnagel, L., Geisel, T., 2006a. The scaling laws of human travel - supplementary information 2. Nature 439, 462 465. https://doi.org/10.1038/nature04292

Brockmann, D., Hufnagel, L., Geisel, T., 2006b. The scaling laws of human travel. Nature 439, 462 465. https://doi.org/10.1038/nature04292

Browder, J., Bohland, J., Scarpaci, J., 1995. Patterns of Development on the Metropolitan Fringe : Urban Fringe Expansion in Bangkok, Jakarta, and Santiago. J. Am. Plann. Assoc. 61, 310 327. https://doi.org/10.1080/01944369508975645

Brown, J.E., Evans, B.R., Zheng, W., Obas, V., Barrera-Martinez, L., Egizi, A., Zhao, H., Caccone, A., Powell, J.R., 2014. Human impacts have shaped historical and recent evolution in Aedes aegypti, the dengue and yellow fever mosquito. Evolution 68, 514 525. https://doi.org/10.1111/evo.12281

Brulhardt, M.-C., Bassand, M., 1981. La mobilité spatiale en tant que système. Swiss Soc. Econ. Stat. SSES 117, 505 519. https://ideas.repec.org/a/ses/arsjes/1981-iii-18.html.

Brunet, R., Ferras, R., Théry, H., 1992. Les mots de la géographie, dictionnaire critique, Reclus / La Documentation Française. ed. Paris / Montpellier.

Buckee, C.O., Tatem, A.J., Metcalf, C.J.E., 2017. Seasonal Population Movements and the Surveillance and Control of Infectious Diseases. Trends Parasitol. 33, 10 20. https://doi.org/10.1016/j.pt.2016.10.006

Buckee, C.O., Wesolowski, A., Eagle, N.N., Hansen, E., Snow, R.W., 2013. Mobile phones and malaria : Modeling human and parasite travel. Travel Med. Infect. Dis. 11, 15 22. https://doi.org/10.1016/j.tmaid.2012.12.003

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011a. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. 10, 0036 44. http://doi.ieeecomputersociety.org/10.1109/MPRV.2011.41.

Calabrese, F., Di Lorenzo, G., Ratti, C., 2010. Human mobility prediction based on individual and collective geographical preferences. IEEE, pp. 312 317. https://doi.org/10.1109/ITSC.2010.5625119

Calabrese, F., Ferrari, L., Blondel, V.D., 2014. Urban Sensing Using Mobile Phone Network Data : A Survey of Research. ACM Comput. Surv. 47, 1 20. https://doi.org/10.1145/2655691

Calabrese, F., Smoreda, Z., Blondel, V.D., Ratti, C., 2011b. Interplay between Telecommunications and Face-to-Face Interactions : A Study Using Mobile Phone Data. PLoS ONE 6, e20814. https://doi.org/10.1371/journal.pone.0020814

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. J. Phys. Math. Theor. 41, 224015. https://doi.org/10.1088/1751-8113/41/22/224015

Carlsson, M., Ottosson, G., Carlson, B., 1997. An open-ended finite domain constraint solver, in : Glaser, H., Hartel, P., Kuchen, H. (Eds.), Programming Languages : Implementations,

Logics, and Programs. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 191 206. https://doi.org/10.1007/BFb0033845
Casey, H.J., 1955. Applications to traffic engineering of the law of retail gravitation. Traffic Q. 9, 23 35.
Cebeillac, A., Daudé, Éric, Vaguet, Alain, 2017. Discontinuités spatiales, santé et mobilités. Analyses et typologies de Google POI et de Tweets pour caractériser les structures spatiales et les dynamiques d'attractivités de Bangkok (Thaïlande). Presented at the Sageo, Rouen. https://tel.archives-ouvertes.fr/SAGEO2017/hal-01649148v1
Cebeillac, A., Daudé, É., Vaguet, A., en relecture. « Spatial discontinuities, health and mobility ». *Revue Internationale de Géomatique.*
Cebeillac, A., Huraux, T., Daudé, É., 2017. Where ? When ? and how often ? What can we learn about daily urban mobilities from Twitter data and google map in Bangkok (Thailand) and what are the perspectives for dengues studies ? Netcom Réseaux Commun. Territ. https://journals.openedition.org/netcom/2725
Cebeillac, A., Le Bigot, B., à paraitre. Couplage entre enquête ethnographique et traces numériques : application aux mobilités quotidiennes d'un quartier de Bangkok, in : Meissonier, J., Vincent, S., Rabaud, M., Kaufmann, V. (Eds.), Hybridation Des Méthodes d'analyse Des Comportements de Mobilité.
Cebeillac, A., Rault, Y.-M., 2016. Contribution of geotagged Twitter data in the study of a social group's activity space. The case of the upper middle class in Delhi, India. Netcom Réseaux Commun. Territ. 231 248. Lien : http://netcom.revues.org/2529.
Cecilia, D., 2014. Current status of dengue and chikungunya in India. WHO South-East Asia J. Public Health 3, 22. Lien : https://doi.org/10.4103/2224-3151.206879
Chadee, D.D., 2009. Dengue cases and Aedes aegypti indices in Trinidad, West Indies. Acta Trop. 112, 174 180. https://doi.org/10.1016/j.actatropica.2009.07.017
Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S., Ertl, T., 2012. Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition, in : Proceedings of the IEEE Conference on Visual Analytics Science and Technology. pp. 143 152.
Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., Scott, J., 2007. Impact of Human Mobility on Opportunistic Forwarding Algorithms. IEEE Trans. Mob. Comput. 6, 606 620. https://doi.org/10.1109/TMC.2007.1060
Chalermpong, S., Ratanawaraha, A., 2015. How Land Use Affects Station Access Behaviors of Bus Rapid Transit Passengers in Bangkok, Thailand. Transp. Res. Rec. J. Transp. Res. Board 2533, 50 59. https://doi.org/10.3141/2533-06
Chang, J., Sun, E., 2011. Location3 : How Users Share and Respond to Location-Based Data on Social Networking Sites, in : Proceedings of the Fifth International AAAI Conference on Weblogs and Social Medi. Association for the Advancement of Artificial Intelligence, p. 7.
Chapuis, K., Taillandier, P., Renaud, M., Drogoul, A., 2018. Gen* : a generic toolkit to generate spatially explicit synthetic populations. Int. J. Geogr. Inf. Sci. 32, 1194 1210. https://doi.org/10.1080/13658816.2018.1440563
Chardonnel, S., 2007. Time-geography : Individuals in Time and Space, in : Sanders, L. (Ed.), Models in Spatial Analysis. ISTE, London ; Newport Beach, CA.
Charrel, R.N., Leparc-Goffart, I., Gallian, P., de Lamballerie, X., 2014. Globalization of Chikungunya : 10 years to invade the world. Clin. Microbiol. Infect. 20, 662 663. https://doi.org/10.1111/1469-0691.12694

Chastel, C., 2012. Eventual Role of Asymptomatic Cases of Dengue for the Introduction and Spread of Dengue Viruses in Non-Endemic Regions. Front. Physiol. 3. https://doi.org/10.3389/fphys.2012.00070

Chatterjee, P., 1989. The Nationalist Resolution of the Women's Question, Recasting Women : Essays on Colonial History. ed. Sangari, K. and S. Vaid, New Delhi.

Chen, C., Ma, Jingtao, Susilo, Yusak, Liu, Yu, Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285 299. https://doi.org/10.1016/j.trc.2016.04.005

Chen, W., Gao, Q., Xiong, H.-G., 2017. Uncovering urban mobility patterns and impact of spatial distribution of places on movements. Int. J. Mod. Phys. C 28, 1750004. https://doi.org/10.1142/S0129183117500048

Chen, Y., 2015. The distance-decay function of geographical gravity model : Power law or exponential law ? Chaos Solitons Fractals 77, 174 189. https://doi.org/10.1016/j.chaos.2015.05.022

Cheng, Z., Caverlee, J., Lee, K., 2010. You are where you tweet : a content-based approach to geo-locating twitter users. ACM Press, p. 759. https://doi.org/10.1145/1871437.1871535

Chevalier, P., 2018a. Observer les pratiques de mobilité quotidienne à travers un espace résidentiel et un espace de vie. Presented at the Doctorales de l'Association de Science Régionale de Langue Française, Grenoble.

Chevalier, P., 2018b. Is Long Commuting another Dimension of Inequality among the Poorest Workers ? Evidence from a demographic Perspective. Presented at the PopFest, Oxford, UK.

Chevalier, V., et al, 2005. Rift Valley Fever in Small Ruminants, Senegal, 2003-Volume 11, Number 11 November 2005-Emerging Infectious Disease journal-CDC. http://wwwnc.cdc.gov/eid/article/11/11/05-0193_article.htm.

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and Mobility : User Movement in Location-based Social Networks. Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., KDD '11 1082 1090. https://doi.org/10.1145/2020408.2020579

Choiejit, R., Teungfung, R., 2005. Urban growth and commuting patterns of the poor in Bangkok, in : Third Urban Research Symposium on Land Development, Urban Policy and Poverty Reduction. World Bank Institute of Applied Economic Research. Brasilia, DF, Brazil.

Christie, J., 1881. On epidemics of dengue fevers : their diffusion and etiology. Glasg. Med. J. 16, 167 176.

Christie, J., 1872. Remarks on "Kidinga Pepo" : A Peculiar Form of Exanthematous Disease. Br. J. Med. 577 579.

Chungue, E., Burucoa, C., Boutin, J.-P., Philippon, G., Laudon, F., Plichart, R., Barbazan, P., Cardines, R., Roux, J., 1992. Dengue 1 epidemic in French Polynesia, 1988 1989 : surveillance and clinical, epidemiological, virological and serological findings in 1752 documented clinical cases. Trans. R. Soc. Trop. Med. Hyg. 86, 193 197. http://www.sciencedirect.com/science/article/pii/003592039290568W

Clément-Charpentier, S., 2011. Bangkok, la ville à partir de ses représentations. Moussons 97 120. https://doi.org/10.4000/moussons.724

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL : A Seasonal-Trend Decomposition Procedure Based on Loess. J. Off. Stat. 6, 3 73.

Coberly, J.S., Fink, C.R., Elbert, Y., Yoon, I.-K., Velasco, J.M., Tomayao, A.D., Roque, V.J., Tayag, E., Macasocol, D.R., Lewis, S.H., 2014. Tweeting Fever : Can Twitter Be Used

to Monitor the Incidence of Dengue-Like Illness in the Philippines ? Johns Hopkins APL Tech. Dig. 32.

Coffey, L.L., Mertens, E., Brehin, A.-C., Fernandez-Garcia, M.D., Amara, A., Després, P., Sakuntabhai, A., 2009. Human genetic determinants of dengue virus susceptibility. Microbes Infect. 11, 143 156. https://doi.org/10.1016/j.micinf.2008.12.006

Commenges, H., 2016. Modèle de radiation et modèle gravitaire-Du formalisme à l'usage. Rev. Int. Géomat. 26, 79 95. http://rig.revuesonline.com/articles/lvrig/abs/2016/01/lvrig261p79/lvrig261p79.html.

Commenges, H., 2013. L'invention de la mobilité quotidienne. Aspects performatifs des instruments de la socio-économie des transports. Université Paris-Diderot-Paris VII. https://tel.archives-ouvertes.fr/tel-00923682/.

Commission Nationale Informatique et Liberté, 2004. Donnée personnelle \textbar. Lien : https://www.cnil.fr/fr/definition/donnee-personnelle.

Condomines, B., Hennequin, E., 2013. Etudier des sujets sensibles : les apports d'une approche mixte. RIMHE Rev. Interdiscip. Manag. Hommes Entrep. 5, 12. https://doi.org/10.3917/rimhe.005.0012

Conseiller Informatique et Liberté - CNRS, 2016. Qu'est-ce qu'une donnée personnelle ? - Fil d'actualité du Service Informatique et libertés du CNRS. http://www.cil.cnrs.fr/CIL/spip.php?rubrique299.

Cools, M., Moons, E., Wets, G., 2010. Assessing the Quality of Origin Destination Matrices Derived from Activity Travel Surveys : Results from a Monte Carlo Experiment. Transp. Res. Rec. J. Transp. Res. Board 2183, 49 59. https://doi.org/10.3141/2183-06

Cottineau, C., Finance, O., Hatna, E., Arcaute, E., Batty, M., 2018. Defining urban clusters to detect agglomeration economies. Environ. Plan. B Urban Anal. City Sci. 239980831875514. https://doi.org/10.1177/2399808318755146

Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N., 2010. Bridging the gap between physical location and online social networks, in : Proceedings of the 12th ACM International Conference on Ubiquitous Computing. ACM, pp. 119 128. Lien : http ://dl.acm.org/citation.cfm ?id=1864380.

Cresswell, T., 2006. On the move : mobility in the modern Western world. Routledge, New York, NY.

Csáji, B.C., Browet, A., Traag, V.A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V.D., 2013. Exploring the mobility of mobile phone users. Phys. Stat. Mech. Its Appl. 392, 1459 1473. https://doi.org/10.1016/j.physa.2012.11.040

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Systems, 1695. http://igraph.org.

Czura, G., Taillandier, P., Tranouez, P., Daudé, É., 2015. MOSAIIC : City-Level Agent-Based Traffic Simulation Adapted to Emergency Situations, in : Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014. Springer, pp. 265 274. http://link.springer.com/chapter/10.1007/978-3-319-20591-5_24

Danon, L., Ford, A.P., House, T., Jewell, C.P., Keeling, M.J., Roberts, G.O., Ross, J.V., Vernon, M.C., 2011. Networks and the Epidemiology of Infectious Disease. Interdiscip. Perspect. Infect. Dis. 2011, 1 28. https://doi.org/10.1155/2011/284909

Darriet, F., 2014. Des moustiques et des hommes : chronique d'une pullulation annoncée. IRD, Marseille.

Daudé, É., 2017. Complex Pathogenic Systems, Model and Simulation. Rev. Francoph. Sur Santé Territ.

Daudé, E., Eliot, E., 2005. Exploration de l'effet des types de mobilités sur la diffusion des épidemies, in : 7èmes Rencontres de ThéoQuant. p. 17. https://halshs.archives-ouvertes.fr/halshs-01082646/

Daudé, E., Mazumdar, S., 2016. Combating Dengue in India. Econ. Polit. Wkly. 51.

Daudé, É., Mazumdar, S., Solanki, V., 2017. Widespread fear of dengue transmission but poor practices of dengue prevention : A study in the slums of Delhi, India. PLOS ONE 12, e0171543. https://doi.org/10.1371/journal.pone.0171543

Daudé, É., Vaguet, A., 2015. Surveillance, contrôle et épidémies de dengue en Inde : Qui a échoué ? L'Espace Polit. https://doi.org/10.4000/espacepolitique.3485

Daudé, É., Vaguet, A., Paul, R., 2015. La dengue, maladie complexe. Nat. Sci. Sociétés 23, 331 342. https://doi.org/10.1051/nss/2015058

Davis Jr., C.A., Pappa, G.L., de Oliveira, D.R.R., de L. Arcanjo, F., 2011. Inferring the Location of Twitter Messages Based on User Relationships : Inferring the Location of Twitter Messages Based on User Relationships. Trans. GIS 15, 735 751. https://doi.org/10.1111/j.1467-9671.2011.01297.x

de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd : The privacy bounds of human mobility. Sci. Rep. 3. https://doi.org/10.1038/srep01376

de Montjoye, Y.-A., Kendall, J., Kerry, C.F., 2014. Enabling humanitarian use of mobile phone data. https://dspace.mit.edu/handle/1721.1/92821.

Decuyper, A., 2016. On the research for big data uses for public good purposes. Opportunities and challenges. Netcom Réseaux Commun. Territ. 305 314. http://netcom.revues.org/2556.

Degner, E.C., Harrington, L.C., 2016. Polyandry Depends on Postmating Time Interval in the Dengue Vector Aedes aegypti. Am. J. Trop. Med. Hyg. 94, 780 785. https://doi.org/10.4269/ajtmh.15-0893

Dejnirattisai, W., Jumnainsong, A., Onsirisakul, N., Fitton, P., Vasanawathana, S., Limpitikul, W., Puttikhunt, C., Edwards, C., Duangchinda, T., Supasa, S., Chawansuntati, K., Malasit, P., Mongkolsapaya, J., Screaton, G., 2010. Cross-Reacting Antibodies Enhance Dengue Virus Infection in Humans. Science 328, 745 748. https://doi.org/10.1126/science.1185181

Delisle, E., Rousseau, C., Broche, B., Leparc-Goffart, I., L'Ambert, G., Cochet, A., 2015. Foyer de cas autochtones de chikungunya à Montpellier, septembre-octobre 2014. Bull. Epidémiologique Hebd. 13 14, 212 217.

Depeau, S., Chardonnel, S., André-Poyaud, I., Lepetit, A., Jambon, F., Quesseveur, E., Gombaud, J., Allard, T., Choquet, C.-A., 2017. Routines and informal situations in children's daily lives. Travel Behav. Soc. 9, 70 80. https://doi.org/10.1016/j.tbs.2017.06.003

Derouich, M., Twizell, E.H., Boutayeb, A., 2003. A model of dengue fever. http://dspace.brunel.ac.uk/handle/2438/1692

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111, 15888 15893. https://doi.org/10.1073/pnas.1408439111

Devillers, L., 2017. Des Robots et des Hommes - Mythes, fantasmes et réalité, Plon. ed.

DLT, 2018. Number of Vehicle registered as of 31 December 2017. Department of Land Transport, Bangkok. Lien : http://apps.dlt.go.th/statistics_web/brochure/cumcar17.pdf

Dodd, S.C., 1950. The Interactance Hypothesis : A Gravity Model Fitting Physical Masses and Human Groups. Am. Sociol. Rev. 15, 245. https://doi.org/10.2307/2086789

Douplitzky, K., 2009. Le commerce du moi, modèle économique du profilage, in : Traçabilité et réseaux, Hermès, La Revue. CNRS, Paris, pp. 113 117.

Drevon, G., Jambon, F., Chardonnel, S., Christophe, S., André-Poyaud, I., Davoine, P.-A., Lutoff, C., 2014. Évaluation comparée de l'apport de l'assistance GPS aux enquêtes de mobilité. Netcom 13 34. https://doi.org/10.4000/netcom.1527

Droit Français, 1978. Loi 78-17 du 6 janvier 1978 modifiée \textbarCNIL. https://www.cnil.fr/fr/loi-78-17-du-6-janvier-1978-modifiee#Article6

Duféal, M., Noucher, M., 2017. Des TIC au TOC. Contribuer à OpenStreetMap : entre commun numérique et utopie cartographique. Netcom Réseaux Commun. Territ. 77 98. https://journals.openedition.org/netcom/263

Duong, V., Lambrechts, L., Paul, R.E., Ly, S., Lay, R.S., Long, K.C., Huy, R., Tarantola, A., Scott, T.W., Sakuntabhai, A., Buchy, P., 2015. Asymptomatic humans transmit dengue virus to mosquitoes. Proc. Natl. Acad. Sci. 112, 14688 14693. https://doi.org/10.1073/pnas.1508114112

ECDC, 2013. Communicable Disease Threats. European Centre for Disease Prevention and Control.

Edwards, A.M., Phillips, R.A., Watkins, N.W., Freeman, M.P., Murphy, E.J., Afanasyev, V., Buldyrev, S.V., da Luz, M.G.E., Raposo, E.P., Stanley, H.E., Viswanathan, G.M., 2007. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. Nature 449, 1044 1048. https://doi.org/10.1038/nature06199

Eisen, L., Moore, C.G., 2013. Aedes (Stegomyia) aegypti in the Continental United States : A Vector at the Cool Margin of Its Geographic Range. J. Med. Entomol. 50, 467 478. https://doi.org/10.1603/ME12245

Eliot, E., Daudé, É., 2006. Diffusion des épidémies et complexités géographiques : Perspectives conceptuelles et méthodologiques. Espace Popul. Sociétés 403 416. https://doi.org/10.4000/eps.1867

Enduri, M.K., Jolad, S., 2014. Dynamics of Dengue with human and vector mobility. ArXiv Prepr. ArXiv14090965. https://arxiv.org/abs/1409.0965

Eritja, R., Palmer, J.R.B., Roiz, D., Sanpera-Calbet, I., Bartumeus, F., 2017. Direct Evidence of Adult Aedes albopictus Dispersal by Car. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-12652-5

Ertzscheid, O., 2013. Qu'est ce que l'Identité numérique ?, OpenEdition Press. ed, Encyclopédie numérique.

Espín Noboa, L., Lemmerich, F., Singer, P., Strohmaier, M., 2016. Discovering and Characterizing Mobility Patterns in Urban Spaces : A Study of Manhattan Taxi Data, in : Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 537 542. http://dl.acm.org/citation.cfm?id=2890468

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-96 Proc.

Falcón-Lezama, J.A., Santos-Luna, R., Román-Pérez, S., Martínez-Vega, R.A., Herrera-Valdez, M.A., Kuri-Morales, Á.F., Adams, B., Kuri-Morales, P.A., López-Cervantes, M., Ramos-Castañeda, J., 2017. Analysis of spatial mobility in subjects from a Dengue endemic urban locality in Morelos State, Mexico. PloS One 12, e0172313. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0172313

Falkus, M., 1999. Income Inequality and Uncertain Democracy in Thailand, in : Growth, Distribution and Political Change : Asia and the Wider World. Basingstoke : Palgrave Macmillan, pp. 114 142.

Favier, C., Schmit, D., Muller-Graf, C.D.., Cazelles, B., Degallier, N., Mondet, B., Dubois, M.A., 2005. Influence of spatial heterogeneity on an emerging infectious disease : the case of dengue epidemics. Proc. R. Soc. B Biol. Sci. 272, 1171 1177. https://doi.org/10.1098/rspb.2004.3020

Feildel, B., 2014. La mobilité révélée par GPS : Traces et récits pour éclairer les sens des mobilités. Netcom 55 76. https://doi.org/10.4000/netcom.1545

Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A., 2014. The rise of social bots. ArXiv Prepr. ArXiv14075225. http://arxiv.org/abs/1407.5225

Ferrari, L., Rosi, A., Mamei, M., Zambonelli, F., 2011. Extracting Urban Patterns from Location-based Social Networks, in : Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '11. ACM, New York, NY, USA, pp. 9 16. https://doi.org/10.1145/2063212.2063226

Ffirth, S., 1804. A Treatise on Malignant Fever; with an Attempt to Prove its Non-contagious Non-Malignant Nature. University of Pennsylvania, Philadelphia.

Finger, F., Genolet, T., Mari, L., de Magny, G.C., Manga, N.M., Rinaldo, A., Bertuzzo, E., 2016. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proc. Natl. Acad. Sci. 113, 6421 6426. https://doi.org/10.1073/pnas.1522305113

Fleury, A., Mathian, H., Saint-Julien, T., 2012. Définir les centralités commerciales au cœur d'une grande métropole : le cas de Paris intra-muros. Cybergeo Eur. J. Geogr. Online Space Soc. https://doi.org/10.4000/cybergeo.25107

Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486, 75 174. https://doi.org/10.1016/j.physrep.2009.11.002

Franco, L., Di Caro, A., Carletti, F., Vapalahti, O., Renaudat, C., Zeller, H., Tenorio, A., 2010. Recent expansion of dengue virus serotype 3 in West Africa. Euro Surveill 15, 578 583. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.642.4820&rep=rep1&type=pdf#page=75

Frias-Martinez, E., Williamson, G., Frias-Martinez, V., 2011. An agent-based model of epidemic spread using human mobility and social network information, in : Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference On. IEEE, pp. 57 64. http://ieeexplore.ieee.org/abstract/document/6113095/

Frias-Martinez, V., Frias-Martinez, E., 2014. Spectral clustering for sensing urban land use using Twitter activity. Eng. Appl. Artif. Intell. 35, 237 245. Lien : http://www.sciencedirect.com/science/article/pii/S0952197614001419.

Frias-Martinez, V., Soguero-Ruiz, C., Frias-Martinez, E., 2012. Estimation of Urban Commuting Patterns Using Cellphone Network Data. Presented at the International Workshop on Urban Computing, Association for Computing Machinery. Lien : http://dl.acm.org/citation.cfm?id=2346496.

Federal Statistic Office, 2016. Switzerland's population 2015. https://www.bfs.admin.ch/bfs/en/home/news/whats-new.assetdetail.1401565.html

Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M., 2011. Analyzing and Visualizing State Sequences in R with TraMineR. J. Stat. Softw. 40, 1 37. http://www.jstatsoft.org/v40/i04/.

Gailhardou, S., Skipetrova, A., Dayan, G.H., Jezorwski, J., Saville, M., Van der Vliet, D., Wartel, T.A., 2016. Safety Overview of a Recombinant Live-Attenuated Tetravalent Dengue Vaccine : Pooled Analysis of Data from 18 Clinical Trials. PLoS Negl. Trop. Dis. 10, e0004821. https://doi.org/10.1371/journal.pntd.0004821

Gakenheimer, R., 1999. Urban mobility in the developing world. Transp. Res. Part Policy Pract. 33, 671 689. http://www.sciencedirect.com/science/article/pii/S0965856499000051

Gambs, S., Killijian, M.-O., del Prado Cortez, M.N., 2014. De-anonymization attack on geolocated data. J. Comput. Syst. Sci. 80, 1597 1614.

Gao, S., Yang, J.-A., Yan, B., Hu, Y., Janowicz, K., McKenzie, G., 2014. Detecting origin-destination mobility flows from geotagged Tweets in greater Los Angeles area, in : Eighth International Conference on Geographic Information Science (GIScience'14). https://pdfs.semanticscholar.org/5c0d/c468c0bce57483eb8d1382d7e4161b92e035.pdf.

Garcia Lopez, D., 2010. Modélisation de l'émergence de maladies infectieuses : exemple de la dengue. Pierre et Marie Curie, Paris.

Gatrell, A.C., 2011. Mobilities and Health. Ashgate (Surrey). London.

Gerbeaud, F., 2011. L'habitat spontané comme un outil de développement urbain. Le cas de Bangkok. Moussons 121 138. https://doi.org/10.4000/moussons.740

Getis, A., Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. Geogr. Anal. 24, 189 206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., Blat, J., 2008. Digital footprinting : Uncovering tourists with user-generated content. IEEE Pervasive Comput. 7, 36 43. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4653470

Girardin, F., Dal Fiore, F., Blat, J., Ratti, C., 2007. Understanding of tourist dynamics from explicitly disclosed location information, in : Symposium on LBS and Telecartography. http://www.girardin.org/fabien/publications/girardin_dalfiore_blat_ratti_lbs2007_final.pdf

Giridhar, P., Abdelzaher, T., 2017. Visualization of events using Twitter and Instagram, in : Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference On. IEEE, pp. 82 84. http://ieeexplore.ieee.org/abstract/document/7917530/

Giridhar, P., Abdelzaher, T., Kaplan, L., 2017. Social fusion : Integrating twitter and instagram for event monitoring. https://www.ideals.illinois.edu/bitstream/handle/2142/95127/paper%20(2).pdf?sequence=2

Giron, S., Rizzi, J., Leparc-Goffart, I., Septfons, A., Tine, R., Cadiou, B., Eberhart, P., Charlet, F., Lebaillif, T., Auzet-Caillaud, M., Decoppet, A., Pigaglio, L., Lopez, K., Peloux-Petiot, F., Travanut, M., Pingeon, J.-M., Teruel, I., Schaal, O., Debruyne, M., Prat, C., Flusin, O., Deniau, J., Franke, F., Noël, H., Paty, M.-C., Six, C., 2015. Nouvelles apparitions de cas autochtones de dengue en région Provence-Alpes-Côte d'Azur, France, août-septembre 2014. Bull. Epidémiologique Hebd. 217 223. http://invs.santepubliquefrance.fr//beh/2015/13-14/2015_13-14_3.html.

Gjenero-Margan, I., Aleraj, B., Krajcar, D., Lesnikar, V., Klobucar, A., Pem-Novosel, I., Kurecic-Filipovic, S., Komparak, S., Martic, R., Duricic, S., others, 2011. Autochthonous dengue fever in Croatia, August-September 2010. Euro Surveill 16, 19805. https://www.researchgate.net/profile/Ljiljana_Betica_Radic/publication/50362863_Autochthonous_Dengue_fever_in_Croatia_August-September_2010/links/09e41513644b20cc16000000.pdf.

Golledge, R.G., Stimson, R.J., 1996. Spatial Behavior : A Geographic Perspective. The Guilfor Press, New York.

Gong, li, Liu, Xi, Wu, L., Liu, Y., 2015. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartogr. Geogr. Inf. Soc.

González, J.G.O., Hernández, R.M., Suárez, A.E.F., Salas, I.F., 2001. The use of sticky ovitraps to estimate dispersal of Aedes aegypti in northeastern Mexico. Lien : http://www.biodiversitylibrary.org/content/part/JAMCA/JAMCA_V17_N2_P093-097.pdf.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779 782. https://doi.org/10.1038/nature06958

Goodchild, M.F., 2007. Citizens as sensors : the world of volunteered geography. GeoJournal 69, 211 221. Lien : http://link.springer.com/article/10.1007/s10708-007-9111-y

Graham, H., 1903. The dengue : a study of its pathology and mode of propagation. J. Trop. Med. 6, 209 214.

Graham, M., Zook, M., 2013. Augmented Realities and Uneven Geographies : Exploring the Geolinguistic Contours of the Web. Environ. Plan. A 45, 77 99. Lien : http://journals.sagepub.com/doi/10.1068/a44674. https://doi.org/10.1068/a44674

Grange, L., Simon-Loriere, E., Sakuntabha, A., Lionel, G., Paul, R., Harris, E., 2014. Epidemiological Risk Factors Associated with High Global Frequency of Inapparent Dengue Virus Infections. Front. Immunol. https://doi.org/10.3389/fimmu.2014.00280

Grassi, G.B., Bignami, A., Bastianelli, G., 1899. Ulteriore ricerche sul ciclo dei parassiti malarici umani sul corpo del zanzarone. Atti R. Accad Lincei 8, 21 28.

Griffitts, T.H.D., 1933. Air Traffic in Relation to Public Health. Am. Soc. Trop. Med. Hyg. 1 13, 283 290.

Grossenbacher, T., 2014. Studying Human Mobility through Geotagged social media content. University of Zurich, Zurich.

Gubler, D.J., 2014. Dengue Viruses : Their Evolution, History and Emergence as a Global Public Health Problem, in : Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J., C.A.B. International (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI, Wallingford, Oxfordshire ; Boston, MA, pp. 1 29.

Gubler, D.J., 2011. Dengue, Urbanization and Globalization : The Unholy Trinity of the 21st Century. Trop. Med. Health 39, S3 S11. https://doi.org/10.2149/tmh.2011-S05.

Gubler, D.J., 2006. Dengue/dengue haemorrhagic fever : history and current status. Novartis Found. Symp., Novartis Foundation symposium 277.

Gubler, D.J., 1998. Resurgent vector-borne diseases as a global health problem. Emerg. Infect. Dis. 4, 442. Lien : https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2640300/.

Gubler, D.J., 1997. Epidemic Dengue/Dengue Haemorrhagic Fever : a global public health problem in the 21st century.

Guvry, P., Huong, P.T., Thuy, T.T.T., Ngân, V.H., Lê, T.H., 2008. Bouger pour mieux vivre. Les mobilités intra-urbaines à Hô Chi Minh Ville et à Hanoi. Université Nationale d'économie, Hanoi.

Guzmán, M.G., Kouri, G., 2002. Dengue : an update. Lancet Infect. Dis. 2, 33 42. http://www.sciencedirect.com/science/article/pii/S1473309901001712

Hadinegoro, S.R., Arredondo-García, J.L., Capeding, M.R., Deseda, C., Chotpitayasunondh, T., Dietze, R., Hj Muhammad Ismail, H.I., Reynales, H., Limkittikul, K., Rivera-Medina, D.M., Tran, H.N., Bouckenooghe, A., Chansinghakul, D., Cortés, M., Fanouillere, K., Forrat, R., Frago, C., Gailhardou, S., Jackson, N., Noriega, F., Plennevaux, E., Wartel, T.A., Zambrano, B., Saville, M., 2015. Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. N. Engl. J. Med. 373, 1195 1206. https://doi.org/10.1056/NEJMoa1506223

Hägerstrand, T., 1970. What about people in regional science ? Pap. Reg. Sci. Assoc. 24, 7 21.

Haight, F.A., 1967. Handbook of the Poisson distribution. Wiley, New York.

Halstead, S. áB, Nimmannitya, S., Cohen, S.N., 1970. Observations related to pathogenesis of dengue hemorrhagic fever. IV. Relation of disease severity to antibody response and virus recovered. Yale J. Biol. Med. 42, 311. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2591704/

Halstead, S.B., 2016. Critique of World Health Organization Recommendation of a Dengue Vaccine. J. Infect. Dis. 214, 1793 1795. https://doi.org/10.1093/infdis/jiw340
Halstead, S.B., 2012. Dengue vaccine development : a 75% solution ? The Lancet 380, 1535. http://search.proquest.com/openview/b20064070c4cfeefa0ef56d695ad5fad/1?pq-origsite=gscholar&cbl=40246
Halstead, S.B., 2007. Dengue. The Lancet 370, 1644 1652. http://www.sciencedirect.com/science/article/pii/S0140673607616870
Halstead, S.B., O'rourke, E.J., 1977. Dengue viruses and mononuclear phagocytes. I. Infection enhancement by non-neutralizing antibody. J. Exp. Med. 146, 201 217. http://jem.rupress.org/content/146/1/201.abstract
Halstead, S.B., Scanlon, J.E., Umpaivit, P., Udomasakdi, S., 1969. Dengue and chikungunya virus infection in man in Thailand, 1962-1964. IV. Epidemiologic studies in the Bangkok metropolitan area. Am. J. Trop. Med. Hyg. 18, 997 1021.
Hammadou, H., Thomas, I., Verhetsel, A., Witlox, F., 2008. How to Incorporate the Spatial Dimension in Destination Choice Models : The Case of Antwerp. Transp. Plan. Technol. 31, 153 181. https://doi.org/10.1080/03081060801948126
Hammon, W.M., Rudnick, A., Sather, G., Rogers, K.D., Morse, L.J., 1960. New hemorrhagic fevers of children in the Philippines and Thailand. Trans. Assoc. Am. Physicians 73, 140 155.
Hanaoka, S., 2007. Review of urban transport policy and its impact in Bangkok. Proc. East. Asia Soc. Transp. Stud.
Hanson, S., 2010. Gender and mobility : new approaches for informing sustainability. Gend. Place Cult. 17, 5 23. https://doi.org/10.1080/09663690903498225
Hartley, H.O., 1958. Maximum Likelihood Estimation from Incomplete Data. Biometrics 14, 174. https://doi.org/10.2307/2527783
Harvey, F., 2013. To Volunteer or to Contribute Locational Information ? Towards Truth in Labeling for Crowdsourced Geographic Information, in : Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge. Springer Netherlands, Dordrecht, pp. 31 42. https://doi.org/10.1007/978-94-007-4587-2
Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Carlo, R., 2014. Geo-located Twitter as the proxy for global mobility patterns. Cartogr. Geogr. Inf. Sci.
Hazan, V., 2017. La production d'espace par la vente de rue - Le cas du quartier de Ari à Bangkok (Mémoire Master). Ecole Nationale des Travaux Publics de l'Etat, Lyon.
Heeckt, C., Gomes, A., Ney, D., Phanthuwongpakdee, N., Sabrié, M., 2017. Towards urban growth analytics for Yangon : a comparative information base for strategic spatial development.
Hempelmann, E., Krafts, K., 2013. Bad air, amulets and mosquitoes : 2,000 years of changing perspectives on malaria. Malar. J. 12, 232. https://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-12-232.
Hildebrandt, M., 2013. Slaves to big data. Or are we ?, in : IDP Revista De Internet, Derecho Y Política. Presented at the 9 th Annual Conference on Internet, Law & Politics, Barcelona. http://repository.ubn.ru.nl/bitstream/handle/2066/119975/119975.pdf
Hirsch, J.A., Winters, M., Clarke, P., McKay, H., 2014. Generating GPS activity spaces that shed light upon the mobility habits of older adults : a descriptive analysis. Int. J. Health Geogr. 13, 51. https://doi.org/10.1186/1476-072X-13-51
Hiruta, S., Yonezawa, T., Jurmu, M., Tokuda, H., 2012. Detection, classification and visualization of place-triggered geotagged tweets, in : Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, pp. 956 963.

Hong, I., 2015. Spatial Analysis of Location-Based Social Networks in Seoul, Korea. J. Geogr. Inf. Syst. 07, 259 265.https://doi.org/10.4236/jgis.2015.73020

Honório, N.A., Silva, W. da C., Leite, P.J., Gonçalves, J.M., Lounibos, L.P., Lourenço-de-Oliveira, R., 2003. Dispersal of Aedes aegypti and Aedes albopictus (Diptera : Culicidae) in an urban endemic dengue area in the State of Rio de Janeiro, Brazil. Mem. Inst. Oswaldo Cruz 98, 191 198. http://www.scielo.br/scielo.php?pid=S0074-02762003000200005&script=sci_arttext

Horton, F.E., Reynolds, D.R., 1971. Effects of Urban Spatial Structure on Individual Behavior. Econ. Geogr. 47, 36. https://doi.org/10.2307/143224

Hu, T., Yang, J., Li, X., Gong, P., 2016. Mapping Urban Land Use by Using Landsat Images and Open Social Data. Remote Sens. 8, 151. https://doi.org/10.3390/rs8020151

Huang, L., Li, Q., Yue, Y., 2010. Activity Identification from GPS Trajectories Using Spatial Temporal POIs' Attractiveness, in : Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10. ACM, New York, NY, USA, pp. 27 30. https://doi.org/10.1145/1867699.1867704

Huang, Q., Cao, G., Wang, C., 2014. From Where Do Tweets Originate ? : A GIS Approach for User Location Inference, in : Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '14. ACM, New York, NY, USA, pp. 1 8. https://doi.org/10.1145/2755492.2755494

Huy, N.T., Van Giang, T., Thuy, D.H.D., Kikuchi, M., Hien, T.T., Zamora, J., Hirayama, K., 2013. Factors Associated with Dengue Shock Syndrome : A Systematic Review and Meta-Analysis. PLoS Negl. Trop. Dis. 7, e2412. https://doi.org/10.1371/journal.pntd.0002412

Iglesias, N.G., Byk, L.A., Gamarnik, A.V., 2014. Molecular Virology of Dengue Virus, in : Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J., C.A.B. International (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI, Wallingford, Oxfordshire; Boston, MA, pp. 334 364.

Iltanen, S., 2012. Cellular Automata in Urban Spatial Modelling, in : Batty, M., Heppenstall, A.J., Crooks, A.T., See, L.M. (Eds.), Agent-Based Models of Geographical Systems. Springer Netherlands, Dordrecht, pp. 69 84. https://doi.org/10.1007/978-90-481-8927-4

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W., 2012. Human mobility modeling at metropolitan scales, in : Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services. Acm, pp. 239 252. http://dl.acm.org/citation.cfm?id=2307659

Jensen, P., 2018. Pourquoi la société ne se laisse pas mettre en équations ?, Science Ouverte. ed. Seuil, Paris.

Jha, G., 1988. Local finance in metropolitan cities : a study of Delhi. Mittal Publications, Delhi.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., Pereira, F.C., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput. Environ. Urban Syst. 53, 36 46. https://doi.org/10.1016/j.compenvurbsys.2014.12.001

Jiang, S., Ferreira, J., Gonzalez, M.C., 2017. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore. IEEE Trans. Big Data 1 1. https://doi.org/10.1109/TBDATA.2016.2631141

Jiang, S., Ferreira, J., Gonz ?lez, M.C., 2012. Clustering daily patterns of human activities in the city. Data Min. Knowl. Discov. 25, 478 510. https://doi.org/10.1007/s10618-012-0264-z

Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr, J., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces : current methods, challenges and

opportunities, in : Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM. http://dl.acm.org/citation.cfm?id=2505828

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., González, M.C., 2016. The TimeGeo modeling framework for urban motility without travel surveys. Proc. Natl. Acad. Sci. 113, E5370 E5378. https://doi.org/10.1073/pnas.1524261113

Jin, P.J., Wan, X., Li, R., 2014. Dynamic Origin-Destination Travel Demand Estimation using Location Based Social Networking Data. Presented at the TRB Meeting. https://pdfs.semanticscholar.org/af46/0f908bbd51388892ee7b45fa6a33ef343a4c.pdf

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990 993. https://doi.org/10.1038/nature06536

Juliano, S.A., Lounibos, L.P., O'Meara, G.F., 2004. A field test for competitive effects of Aedes albopictus on A. aegypti in South Florida : differences between sites of coexistence and exclusion ? Oecologia 139, 583 593. https://doi.org/10.1007/s00442-004-1532-4

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015a. Understanding human mobility from Twitter. PloS One 10, e0131469. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131469

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015b. Understanding human mobility from Twitter. PloS One 10.

Kafsi, M., Kazemi, E., Maystre, L., Yartseva, L., Grossglauser, M., Thiran, P., 2013. Mitigating epidemics through mobile micro-measures. ArXiv Prepr. ArXiv13072084. https://arxiv.org/abs/1307.2084

Kahle, D., Wickham, H., 2013. ggmap : Spatial Visualization with ggplot2. R J. 5. Lien : http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=20734859&AN=90616116&h=WodeyzOZ8FZJ9MsznSPJrnF1F23R2lnAwk280GUIRlu%2F%2B%2FvXrEqZlM8qTD782Xamk%2BPU3ikJj%2BYJPLu4djVBDA%3D%3D&crl=c

Kalra, N.L., Ghosh, T.K., Pattanayak, S., Wattal, B.L., 1976. Epidemiological and entomological study of dengue fever in an outbreak at Ajmer in 1969. J. Commun. Dis. 8, 261 279.

Kampen, H., Jansen, S., Schmidt-Chanasit, J., Walther, D., 2016. Indoor development of *Aedes aegypti* in Germany, 2016. Eurosurveillance 21. https://doi.org/10.2807/1560-7917.ES.2016.21.47.30407

Karl, S., Halder, N., Kelso, J.K., Ritchie, S.A., Milne, G.J., 2014. A spatial simulation model for dengue virus infection in urban areas. BMC Infect. Dis. 14, 447. https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-14-447

Kaufmann, V., 2012a. Mobilité. Lien : http://fr.forumviesmobiles.org/reperes/mobilite-446

Kaufmann, V., 2012b. Motilité. Lien : http://fr.forumviesmobiles.org/reperes/motilite-451

Kaufmann, V., Jemelin, C., 2004. La motilité, une forme de capital permettant d'éviter les irréversibilités socio-spatiales ? Presented at the Espaces et sociétés aujourd'hui. La géographie sociale dans les sciences et dans l'action, Rennes. http://difusion.ulb.ac.be/vufind/Record/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/18163/Home

Kaufmann, V., Schuler, M., Crevoisier, O., Rossel, P., 2004. Mobilité et motilité : De l'intention à l'action. Cah. Lasur 4, 81.

Keeling, M.J., Danon, L., Vernon, M.C., House, T.A., 2010. Individual identity and movement networks for disease metapopulations. Proc. Natl. Acad. Sci. 107, 8866 8870. https://doi.org/10.1073/pnas.1000416107

Kelley, D., Richards, C., 2017. oce : Analysis of Oceanographic Data. Lien : https://CRAN.R-project.org/package=oce

Kermack, W.O., McKendrick, A.G., 1927. A Contribution to the Mathematical Theory of Epidemics. Proc. R. Soc. Math. Phys. Eng. Sci. 115, 700 721. https://doi.org/10.1098/rspa.1927.0118

Kessous, E., Rey, B., 2009. Economie numérique et vie privée, in : Traçabilité et réseaux, Hermès, La Revue. CNRS, Paris, pp. 49 54.

Khan, S.F., Bergmann, N., Jurdak, R., Kusy, B., Cameron, M., 2017. Mobility in Cities : Comparative Analysis of Mobility Models Using Geo-tagged Tweets in Australia. https://ieeexplore.ieee.org/document/8078751/.

Kim, H., Song, H.Y., 2012. Formulating Human Mobility Model in a Form of Continuous Time Markov Chain. Procedia Comput. Sci. 10, 389 396. https://doi.org/10.1016/j.procs.2012.06.051

Kim, M., Kotz, D., Kim, S., 2006. Extracting a Mobility Model from Real User Traces., in : INFOCOM. pp. 1 13. http://eecs.wsu.edu/~nroy/courses/spring2013/cptsee555/papers/MobilityModel_Infocom06.pdf.

Kimura, R., Hotta, S., 1944. Studies on dengue fever (VI). On the inoculation of dengue virus into mice. (en Japonais). Nippon Igaku 3379, 629 633.

Kobayashi, M., Nihei, N., Kurihara, T., 2002. Analysis of northern distribution of Aedes albopictus (Diptera : Culicidae) in Japan by geographical information system. J. Med. Entomol. 39, 4 11. Lien : http://www.bioone.org/doi/abs/10.1603/0022-2585-39.1.4

Kraemer, M.U.G., Faria, N.R., Reiner, R.C., Golding, N., Nikolay, B., Stasse, S., Johansson, M.A., Salje, H., Faye, O., Wint, G.R.W., Niedrig, M., Shearer, F.M., Hill, S.C., Thompson, R.N., Bisanzio, D., Taveira, N., Nax, H.H., Pradelski, B.S.R., Nsoesie, E.O., Murphy, N.R., Bogoch, I.I., Khan, K., Brownstein, J.S., Tatem, A.J., de Oliveira, T., Smith, D.L., Sall, A.A., Pybus, O.G., Hay, S.I., Cauchemez, S., 2017. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015 16 : a modelling study. Lancet Infect. Dis. 17, 330 338. https://doi.org/10.1016/S1473-3099(16)30513-8

Kraemer, M.U.G., Sinka, M.E., Duda, K.A., Mylne, A., Shearer, F.M., Brady, O.J., Messina, J.P., Barker, C.M., Moore, C.G., Carvalho, R.G., Coelho, G.E., Van Bortel, W., Hendrickx, G., Schaffner, F., Wint, G.R.W., Elyazar, I.R.F., Teng, H.-J., Hay, S.I., 2015a. The global compendium of Aedes aegypti and Ae. albopictus occurrence. Sci. Data 2, 150035. https://doi.org/10.1038/sdata.2015.35

Kraemer, M.U.G., Sinka, M.E., Duda, K.A., Mylne, A.Q., Shearer, F.M., Barker, C.M., Moore, C.G., Carvalho, R.G., Coelho, G.E., Van Bortel, W., others, 2015b. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. Elife 4, e08347. https://elifesciences.org/content/4/e08347

Kuno, G., 1995. Review of the factors modulating dengue transmission. Epidemiol. Rev. 17, 321 335. http://epirev.oxfordjournals.org/content/17/2/321.short

Kurkcu, A., Ozbay, K., Morgul, E.F., 2016. Evaluating the Usability of Geo-Located Twitter as a Tool for Human Activity and Mobility Patterns : A Case Study for New York... Presented at the Transportation Research Board's 95th Annual Meeting, Washington, DC.

La Ruche, G., Souares, Y., Armengaud, A., Peloux-Petiot, F., Delaunay, P., Desprès, P., Lenglet, A., Jourdain, F., Leparc-Goffart, I., Charlet, F., others, 2010. First two autochthonous dengue virus infections in metropolitan France, September 2010. Euro Surveill 15, 19676. Lien : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.642.4820&rep=rep1&type=pdf#page=86.

Lamanna, F., Lenormand, M., Salas-Olmedo, M.H., Romanillos, G., Gonçalves, B., Ramasco, J.J., 2018. Immigrant community integration in world cities. PLOS ONE 13, e0191612.

https://doi.org/10.1371/journal.pone.0191612
Lampos, V., Cristianini, N., 2010. Tracking the flu pandemic by monitoring the Social Web, in : Proceedings of the 2nd International Workshop on Cognitive Information Processing, CIP '10. pp. 411 416. https://doi.org/10.1109/CIP.2010.5604088
Lancichinetti, A., Fortunato, S., 2009. Community detection algorithms : a comparative analysis. Phys. Rev. E 80. https://doi.org/10.1103/PhysRevE.80.056117
Law, R., 1999. Beyond 'women and transport' : towards new geographies of gender and daily mobility. Prog. Hum. Geogr. 23, 567 588. https://doi.org/10.1191/030913299666161864
Lederberg, J., Shope, R.E., Oaks, S.C., 1992. Emerging Infections : Microbial Threats to Health in the United States. National Academy Press, Washington, DC.
Lefebvre, B., 2011. Les services hospitaliers de Delhi : planification, privatisation et gouvernance urbaine. Université de Rouen, Rouen.
Lenormand, M., Bassolas, A., Ramasco, J.J., 2016a. Systematic comparison of trip distribution laws and models. J. Transp. Geogr. 51, 158 169. https://doi.org/10.1016/j.jtrangeo.2015.12.008
Lenormand, M., Huet, S., Gargiulo, F., Deffuant, G., 2012. A universal model of commuting networks. PloS One 7, e45985. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0045985
Lenormand, M., Louail, T., Barthelemy, M., Ramasco, J.J., 2016b. Is spatial information in ICT data reliable ? ArXiv Prepr. ArXiv160903375. https://arxiv.org/abs/1609.03375
Lenormand, M., Louail, T., Cantú-Ros, O.G., Picornell, M., Herranz, R., Arias, J.M., Barthelemy, M., Miguel, M.S., Ramasco, J.J., 2015a. Influence of sociodemographics on human mobility. Sci. Rep. 5, 10075. https://doi.org/10.1038/srep10075
Lenormand, M., Louail, T., Cantú-Ros, O.G., Picornell, M., Herranz, R., Arias, J.M., Barthelemy, M., Miguel, M.S., Ramasco, J.J., 2015b. Influence of sociodemographics on human mobility. Sci. Rep. 5, 10075. https://doi.org/10.1038/srep10075
Lenormand, M., Picornell, M., Canti-Ros, O.G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E., Ramasco, J.J., 2014. Cross-Checking Different Sources of Mobility Information. PLoS ONE 9, e105184. https://doi.org/10.1371/journal.pone.0105184
Lenormand, M., Picornell, M., Cantú-Ros, O.G., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., San Miguel, M., Ramasco, J.J., 2015c. Comparing and modelling land use organization in cities. R. Soc. Open Sci. 2, 150449. https://doi.org/10.1098/rsos.150449
Lenormand, M., Ramasco, J.J., 2016. Towards a better understanding of cities using mobility data. Built Environ. 42, 356 364. http://www.ingentaconnect.com/content/alex/benv/2016/00000042/00000003/art00005
Lesnard, L., 2006. Optimal Matching and Social Science. Centre de Recherche en Economie et Statistique.
Lesnard, L., de Saint Pol, T., 2006. Introduction aux méthodes d'appariement optimal. Bull. Méthodologie Sociol. 90. http://bms.revues.org/638
Levins, R., 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bull. Entomol. Soc. Am. 15, 237 240. http://besa.oxfordjournals.org/content/15/3/237.abstract
Levy, J., Lussault, M., 2004. Dictionnaire de géographie et de l'espace des sociétés. Belin.
Li, N., Chen, G., 2009. Analysis of a location-based social network, in : Computational Science and Engineering, 2009. CSE'09. International Conference On. Ieee, pp. 263 270. http://ieeexplore.ieee.org/abstract/document/5284112/

Li, R., Wang, W., Di, Z., 2017. Effects of human dynamics on epidemic spreading in Côte d'Ivoire. Phys. Stat. Mech. Its Appl. 467, 30 40. https://doi.org/10.1016/j.physa.2016.09.059

Liang, X., Zhao, J., Dong, L., Xu, K., 2013. Unraveling the origin of exponential law in intra-urban human mobility. Sci. Rep. 3. https://doi.org/10.1038/srep02983

Liang, X., Zhao, J., Xu, K., 2015. A general law of human mobility. Sci. China Inf. Sci. 58, 1 14. https://doi.org/10.1007/s11432-015-5402-y

Lima, A., De Domenico, M., Pejovic, V., Musolesi, M., 2015. Disease Containment Strategies based on Mobility and Information Dissemination. Sci. Rep. 5, 10650. https://doi.org/10.1038/srep10650

Lima, A., De Domenico, M., Pejovic, V., Musolesi, M., 2013. Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. ArXiv Prepr. ArXiv13064534. https://arxiv.org/abs/1306.4534

Limkittikul, K., Brett, J., L'Azou, M., 2014. Epidemiological Trends of Dengue Disease in Thailand (2000 2011) : A Systematic Literature Review. PLoS Negl. Trop. Dis. 8, e3241. https://doi.org/10.1371/journal.pntd.0003241

Lindenbach, B.D., Thiel, H.-J., Rice, C.M., 2007. Flaviviridae : The Viruses and Their Replication, in : Knipe, D.M., Howley, P.M. (Eds.), Fields Virology. Lippincott-Raven, Philadelphia, pp. 1101 1152.

Liu, J., Zhao, K., Khan, S., Cameron, M., Jurdak, R., 2015. Multi-scale population and mobility estimation with geo-tagged tweets, in : Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference On. IEEE, pp. 83 86. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7129551

Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. J. Transp. Geogr. 43, 78 90. https://doi.org/10.1016/j.jtrangeo.2015.01.016

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social Sensing : A New Approach to Understanding Our Socioeconomic Environments. Ann. Assoc. Am. Geogr. 105, 512 530. Lien : http ://www.tandfonline.com/doi/full/10.1080/00045608.2015.1018773. https ://doi.org/10.1080/00045608.2015.1018773

Liu, Y., Sui, Z., Kang, C., Gao, Y., 2014. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. PLoS ONE 9, e86026. https://doi.org/10.1371/journal.pone.0086026

Livet, P., Phan, D., Sanders, L., 2014. Diversité et complémentarité des modèles multi-agents en sciences sociales. Rev. Fr. Sociol. 55, 689. https://doi.org/10.3917/rfs.554.0689

Loebach, J.E., Gilliland, J.A., 2016. Free Range Kids ? Using GPS-Derived Activity Spaces to Examine Children's Neighborhood Activity and Mobility. Environ. Behav. 48, 421 453. https://doi.org/10.1177/0013916514543177

Lokta, A., 1923. Contributions to the analysis of malaria epidemiology. II. General part (continued). Comparison of two formulae given by Sir Ronald Ross. Am. J. Trop. Med. Hyg. 3.

Longley, P.A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of Twitter usage. Environ. Plan. A 47, 465 484. http://journals.sagepub.com/doi/abs/10.1068/a130122p

Louail, T., 2016. La seconde révolution quantitative de la géographie se fait-elle sans les géographes ? Presented at the MSFS, Marne la Vallée.

Louail, T., Lenormand, M., Cantu Ros, O.G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2015a. From mobile phone data to the spatial structure of cities. Sci. Rep. 4. https://doi.org/10.1038/srep05276

Louail, T., Lenormand, M., Picornell, M., García Cantú, O., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2015b. Uncovering the spatial structure of mobility networks. Nat. Commun. 6, 6007. https://doi.org/10.1038/ncomms7007

Louf, R., Barthelemy, M., 2015. How congestion shapes cities : from mobility patterns to scaling. Sci. Rep. 4. https://doi.org/10.1038/srep05561

Lourenço, J., Recker, M., 2014. The 2012 Madeira Dengue Outbreak : Epidemiological Determinants and Future Epidemic Potential. PLoS Negl. Trop. Dis. 8, e3083. https://doi.org/10.1371/journal.pntd.0003083

Louvet, R., Josselin, D., Genre-Grandpierre, C., Aryal, J., 2016. Impact des niveaux d'échelle sur l'étude des feux de forêts du sud-est de la France. Rev. Int. Géomat. 26, 445 466. https://doi.org/10.3166/rig.2016.00012

Luo, F., Cao, G., Mulligan, K., Li, X., 2016. Explore spatiotemporal and demographic characteristics of human mobility via Twitter : A case study of Chicago. Appl. Geogr. 70, 11 25. https://doi.org/10.1016/j.apgeog.2016.03.001

Luo, S., Morone, F., Sarraute, C., Travizano, M., Makse, H.A., 2017. Inferring personal economic status from social network location. Nat. Commun. 8, 15227. https://doi.org/10.1038/ncomms15227

Manca, M., Boratto, L., Morell Roman, V., Martori i Gallissà, O., Kaltenbrunner, A., 2017. Using social media to characterize urban mobility patterns : State-of-the-art survey and case-study. Online Soc. Netw. Media 1, 56 69. https://doi.org/10.1016/j.osnem.2017.04.002

Maneerat, S., 2016. Modélisation à base d'agents des risques vectoriels en milieux urbains : exemple d'Aedes aegypti, vecteur de la dengue, à Delhi (Inde). Université de Rouen, Rouen.

Maneerat, S., Daude, E., 2017. Étude par simulation à base d'agents des effets des discontinuités intra-urbaines à Delhi sur la dispersion des moustiques Aedes aegypti, vecteurs de la dengue, de la fièvre jaune, du chikungunya et du virus Zika. Cybergeo Eur. J. Geogr., GeOpenMod - Modèles et logiciels.

Maneerat, S., Daudé, E., 2016. A spatial agent-based simulation model of the dengue vector Aedes aegypti to explore its population dynamics in urban areas. Ecol. Model. 333, 66 78. https://doi.org/10.1016/j.ecolmodel.2016.04.012

Manson, S.M., 2001. Simplifying complexity : a review of complexity theory. Geoforum 32, 405 414. http://www.sciencedirect.com/science/article/pii/S001671850000035X

Marble, D.F., 1964. A simple Markovian model of trip structures in a metropolitan region. Reg. Sci. Assoc. West. Sect. Pap. 150 156.

Marchand, E., Prat, C., Jeannin, C., Lafont, E., Bergmann, T., Flusin, O., Rizzi, J., Roux, N., Busso, V., Deniau, J., others, 2013. Autochthonous case of dengue in France, October 2013. Euro Surveill 201, 18 50. http://www.eurosurveillance.org/images/dynamic/ee/v18n50/art20661.pdf

Martelli, J.-T., 2017. "JNU is not Just Where you go, it's What you Become" Everyday Political Socialisation and Left Activism at Jawaharlal Nehru University (JNU), New Delhi. King's College, London.

Martinez-Uso, A., Pla, F., Sotoca, J.M., GarcÍa-Sevilla, P., 2007. Clustering-Based Hyperspectral Band Selection Using Information Measures. IEEE Trans. Geosci. Remote Sens. 45, 4158 4171. https://doi.org/10.1109/TGRS.2007.904951

Masucci, A.P., Serras, J., Johansson, A., Batty, M., 2013. Gravity versus radiation models : On the importance of scale and heterogeneity in commuting flows. Phys. Rev. E 88. https://doi.org/10.1103/PhysRevE.88.022812

McGrath, B., 2006. Modernities and Memories in Bangkok. Nakhara J. Environ. Des. Plan. 1, 25 40. https://www.tci-thaijo.org/index.php/nakhara/article/view/102621

McKendrick, A.G., 1926. Applications of Mathematics to medical problems. Proc. Edinb. Math. Soc. 44, 98 130.

McKinsey & Company, 2018. Global cities of the future. https://www.mckinsey.com/tools/Wrappers/Wrapper.aspx?sid=\protect\T1\textbraceleftC84CB74F-A3B1-47B1-8265-6252F6D85B68\protect\T1\textbraceright&pid=\protect\T1\textbraceleft4F5BEDB1-6C1F-4243-A052-83ADBABE82DF\protect\T1\textbraceright

Meissonnier, J., Richer, C., 2015. Métro boulot - dodo : quoi de neuf dans nos routines de mobilité ? Espace Popul. Sociétés 8.

Meltz, R., 2008. Marc L.***. Le Tigre 28. http://pointdoc.ac-creteil.fr/IMG/pdf/77_d10_2nde_s4_article_marc_l.pdf

Merzeau, L., 2009. Du signe à la trace : l'information sur mesure, in : Traçabilité et réseaux, Hermès, La Revue. CNRS, Paris, pp. 21 29.

Messina, J.P., Brady, O.J., Pigott, D.M., Brownstein, J.S., Hoen, A.G., Hay, S.I., 2014. A global compendium of human dengue virus occurrence. Sci. Data 1. https://doi.org/10.1038/sdata.2014.4

Michael, K., Clarke, R., 2013. Location and tracking of mobile devices : Überveillance stalks the streets. Comput. Law Secur. Rev. 29, 216 228. http://www.sciencedirect.com/science/article/pii/S0267364913000587

Misao, T., Ishihara, M., 1945. An experiment on the transportation of vector mosquitoes by aircraft. Rinsho Kenkyu 22, 44 50.

Mishra, S., Ramanathan, R., Agarwalla, S.K., 2016. Clinical Profile of Dengue Fever in Children : A Study from Southern Odisha, India. Scientifica 2016, 1 6. https://doi.org/10.1155/2016/6391594

Misslin, R., 2017. Modélisation de l'environnement d'un moustique vecteur de maladies - L'exemple d'Aedes aegypti à Delhi (Inde) et Bangkok (Thaïlande). Université de Rouen, Rouen.

Misslin, R., Daudé, E., 2016. Génération d'environnements artificiels pour la simulation spatiale d'arboviroses.

Misslin, R., Huraux, T., Cebeillac, A., Vaguet, A., Daudé, E., 2017. Modélisation de l'impact des îlots de chaleur urbains sur les dynamiques de population d'Aedes aegypti, vecteur de la dengue et du virus Zika. Presented at the Sageo, Rouen, p. 16.

Misslin, R., Vaguet, Y., Vaguet, A., Daudé, E., 2018. Estimating air temperature using MODIS surface temperature images for assessing Aedes aegypti thermal niche in Bangkok, Thailand. Environ. Monit. Assess. 24.

Morin, E., Lemoigne, J.-L., 1999. L'intelligence de la complexité, L'Harmattan. ed, Cognition et Formation. Paris.

Morlan, H.B., Hayes, R.O., 1958. Urban Dispersal and Activity of Aedes Aegypti. Mosq. News 18, 137 144.

Morrison, A.C., Gray, K., Getis, A., Astete, H., Sihuincha, M., Focks, D., Watts, D., Stancil, J.D., Olson, J.G., Blair, P., others, 2004. Temporal and geographic patterns of Aedes aegypti (Diptera : Culicidae) production in Iquitos, Peru. J. Med. Entomol. 41, 1123 1142. http://www.bioone.org/doi/abs/10.1603/0022-2585-41.6.1123

Morse, S.S., 1995. Factors in the Emergence of Infectious Diseases. Emerg. Infect. Dis. 1.

Morse, S.S., Schluedeberg, A., 1990. Emerging Viruses : The Evolution of Viruses and Viral Diseases. J. Infect. Dis. 1 7.

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. Is the sample good enough ? comparing data from twitter's streaming api with twitter's firehose. ArXiv Prepr. ArXiv13065204. http://arxiv.org/abs/1306.5204.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing : A review. ISPRS J. Photogramm. Remote Sens. 66, 247 259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Muir, L.E., Kay, B.H., 1998. Aedes aegypti survival and dispersal estimated by mark-release-recapture in northern Australia. Am. Soc. Trop. Med. Hyg. 58, 277 282.

Murdock, V., 2011. Your mileage may vary : on the limits of social media. SIGSPATIAL Spec. 3, 62 66. http://dl.acm.org/citation.cfm?id=2047309

Murthy, D., Longwell, S.A., 2013. Twitter and disasters. Inf. Commun. Soc. 16, 837 855. https://doi.org/10.1080/1369118X.2012.696123

Mustafa, M.S., Rasotgi, V., Jain, S., Gupta, V., 2015. Discovery of fifth serotype of dengue virus (DENV-5) : A new public health dilemma in dengue control. Med. J. Armed Forces India 71, 67 70. https://doi.org/10.1016/j.mjafi.2014.09.011

National Statistical Office, 2010. http://popcensus.nso.go.th/en/

Neumayr, A., Muñoz, J., Schunk, M., Bottieau, E., Cramer, J., Calleri, G., López-Vélez, R., Angheben, A., Zoller, T., Visser, L., Serre-Delcor, N., Genton, B., Castelli, F., Van Esbroeck, M., Matteelli, A., Rochat, L., Sulleiro, E., Kurth, F., Gobbi, F., Norman, F., Torta, I., Clerinx, J., Poluda, D., Martinez, M., Calvo-Cano, A., Sanchez-Seco, M.P., Wilder-Smith, A., Hatz, C., Franco, L., for TropNet, 2017. Sentinel surveillance of imported dengue via travellers to Europe 2012 to 2014 : TropNet data from the DengueTools Research Initiative. Eurosurveillance 22. https://doi.org/10.2807/1560-7917.ES.2017.22.1.30433

Newman, M.E.J., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. 103, 8577 8582. https://doi.org/10.1073/pnas.0601602103

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69. https://doi.org/10.1103/PhysRevE.69.026113

Newton, E.A., Paul Reiter, 1992. A Model of the Transmission of Dengue Fever with an Evaluation of the Impact of Ultra-Low Volume (ULV) Insecticide Applications on Dengue Epidemics. Am. Soc. Trop. Med. Hyg. 47, 709 720.

Nguyen-Luong, D., 2012. Faisabilité d'une enquête globale transports (EGT) intégrale par association d'un GPS, d'un SIG et d'un système expert en Île-de-France. Rapp. Final Inst. D'aménagement D'urbanisme Région D'Île de France. 98p .

Nobuchi, H., 1979. The symptoms of a dengue-like illness recorded in a Chinese medical encyclopedia. Kanpo Rinsho 26, 422 425.

Normile, D., 2013. Surprising New Dengue Virus Throws a Spanner in Disease Control Efforts. Science 342, 415.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A Tale of Many Cities : Universal Patterns in Human Urban Mobility. PLoS ONE 7, e37027. https://doi.org/10.1371/journal.pone.0037027

Noulas, A., Scellato, S., Mascolo, C., Pontil, M., 2011. An empirical study of geographic user activity patterns in foursquare. ICwSM 11, 70 573. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2831/3241

NSO, 2009. Time use survey 2009. National Statistical Office, Bangkok. http://web.nso.go.th/en/survey/timeuse/time_use_09.htm

NSO, 2008. The Children and Youth Survey. National Statistical Office, Bangkok. http://web.nso.go.th/en/survey/child/data/Statistical%20%20Tables%20%20The%20Education.pdf

Office of the National Economic and Social Development Board, 2017a. Average income of households. http://social.nesdb.go.th/SocialStat/StatReport_Final.aspx?reportid=1340&template=1R1C&yeartype=M&subcatid=80

Office of the National Economic and Social Development Board, 2017b. Poor population. http://social.nesdb.go.th/SocialStat/StatReport_Final.aspx?reportid=1340&template=1R1C&yeartype=M&subcatid=80

Office of the National Economic and Social Development Board, 2017c. Income inequality coefficient Or Gini coefficient. http://social.nesdb.go.th/SocialStat/StatReport_Final.aspx?reportid=1340&template=1R1C&yeartype=M&subcatid=80

Office of the National Economic and Social Development Board, 2011. Gross provincial product chain volume measures. http://www.nesdb.go.th/nesdb_en/ewt_news.php?nid=4315&filename=index

Okafor, R.O., 1987. Maximum likelihood estimation from incomplete data. J. Appl. Stat. 14, 23 33. https://doi.org/10.1080/02664768700000003

Oliver, N., Matic, A., Frias-Martinez, E., 2015. Mobile Network Data for Public Health : Opportunities and Challenges. Front. Public Health 3. https://doi.org/10.3389/fpubh.2015.00189

Openshaw, S., Taylor, P.J., 1979. A million or so correlated coefficients : three experiments on the modifiable areal unit problem. Spatistical Appl. Spat. Sci. 127 144.

Ortúzar S., J. de D., Willumsen, L.G., 2011. Modelling Transport, Fourth edition. ed. John Wiley & Sons, Chichester, West Sussex, United Kingdom.

Otero, M., Barmak, D.H., Dorso, C.O., Solari, H.G., Natiello, M.A., 2011. Modeling dengue outbreaks. Math. Biosci. 232, 87 95. https://doi.org/10.1016/j.mbs.2011.04.006

Otero, M., Schweigmann, N., Solari, H.G., 2008. A Stochastic Spatial Dynamical Model for Aedes Aegypti. Bull. Math. Biol. 70, 1297 1325. https://doi.org/10.1007/s11538-008-9300-y

Pajot, P., 2018. La naissance d'une théorie au carrefour des disciplines. La Recherche 38 39.

Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z., Colizza, V., 2017. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. R. Soc. Open Sci. 4, 160950. https://doi.org/10.1098/rsos.160950

Pant, C., 1974. Control of Aedes Aegypti.

Pappalardo, L., Rinzivillo, S., Simini, F., 2016a. Human Mobility Modelling : Exploration and Preferential Return Meet the Gravity Model. Procedia Comput. Sci. 83, 934 939. https://doi.org/10.1016/j.procs.2016.04.188

Pappalardo, L., Rinzivillo, S., Simini, F., 2016b. Human Mobility Modelling : Exploration and Preferential Return Meet the Gravity Model. Procedia Comput. Sci. 83, 934 939. https://doi.org/10.1016/j.procs.2016.04.188

Pappalardo, L., Simini, F., 2017a. Modelling spatio-temporal routines in human mobility. ArXiv Prepr. ArXiv160705952v2.

Pappalardo, L., Simini, F., 2017b. Data-driven generation of spatio-temporal routines in human mobility. Data Min. Knowl. Discov. https://doi.org/10.1007/s10618-017-0548-4

Pariser, E., 2011. The filter bubble : what the Internet is hiding from you. Penguin Press, New York. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1118322

Parrott, L., 2002. Complexity and the limits of ecological engineering. Trans.-Am. Soc. Agric. Eng. 45, 1697 1702. https://elibrary.asabe.org/azdez.asp?AID=11032&T=2

Parry, H.J., Crossley, H.M., 1950. Validity of responses to survey questions. Public Opin. Q. 14, 61 80.

Pellis, L., Ball, F., Bansal, S., Eames, K., House, T., Isham, V., Trapman, P., 2015. Eight challenges for network epidemic models. Epidemics 10, 58 62. https://doi.org/10.1016/j.epidem.2014.07.003

Perchoux, C., Chaix, B., Cummins, S., Kestens, Y., 2013. Conceptualization and measurement of environmental exposure in epidemiology : Accounting for activity space related to daily mobility. Health Place 21, 86 93. Lien : http ://linkinghub.elsevier.com/retrieve/pii/S1353829213000117. https ://doi.org/10.1016/j.healthplace.2013.01.005

Pereira, R.H.M., Nadalin, V., Monasterio, L., Albuquerque, P.H.M., 2013. Urban Centrality : A Simple Index : Urban Centrality : A Simple Index. Geogr. Anal. 45, 77 89. https://doi.org/10.1111/gean.12002

Perez-Saez, J., King, A.A., Rinaldo, A., Yunus, M., Faruque, A.S.G., Pascual, M., 2016. Climate-driven endemic cholera is modulated by human mobility in a megacity. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2016.11.013

Perkins, T.A., Garcia, A.J., Paz-Soldan, V.A., Stoddard, S.T., Reiner, R.C., Vazquez-Prokopec, G., Bisanzio, D., Morrison, A.C., Halsey, E.S., Kochel, T.J., Smith, D.L., Kitron, U., Scott, T.W., Tatem, A.J., 2014. Theory and data for simulating fine-scale human movement in an urban environment. J. R. Soc. Interface 11, 20140642 20140642. https://doi.org/10.1098/rsif.2014.0642

Perrier, E., 2014. De la simplexité et des systèmes complexes, in : Berthoz, A., Petit, J.-L. (Eds.), Complexité - Simplexité. Collège de France, Paris, pp. 65 73.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware map : Identifying human daily activity pattern using mobile phone data. Hum. Behav. Underst. 14 25. http://link.springer.com/content/pdf/10.1007/978-3-642-14715-9.pdf#page=22

Pichard-Bertaux, L., 2011. Le tout et son contraire : une lecture de Bangkok. Moussons 149 160. https://doi.org/10.4000/moussons.757

Pielou, E.C., 1966. The Measurement of Diversity in Different Types of Biological Colledions. J. Theor. Biol. 13, 131 144. https://doi.org/10.1016/0022-5193(66)90013-0

Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., Trianni, G., 2009. Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. 113, S110 S122. https://doi.org/10.1016/j.rse.2007.07.028

Polwiang, S., 2016. Estimation of dengue infection for travelers in Thailand. Travel Med. Infect. Dis. 14, 398 406. https://doi.org/10.1016/j.tmaid.2016.06.002

Pongsumpun, P., Patanarapelert, K., Sriprom, M., Varamit, S., Tang, I.M., 2004. Infection risk to travelers going to dengue fever endemic regions. http://imsear.li.mahidol.ac.th/handle/123456789/35555

Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks (long version). arXiv :physics/0512106. http://arxiv.org/abs/physics/0512106

Poorthuis, A., 2018. How to Draw a Neighborhood ? The Potential of Big Data, Regionalization, and Community Detection for Understanding the Heterogeneous Nature of Urban Neighborhoods : How to Draw a Neighborhood ? Geogr. Anal. 50, 182 203. https://doi.org/10.1111/gean.12143

Powell, J.R., Tabachnick, W.J., 2013. History of domestication and spread of Aedes aegypti - A Review. Mem. Inst. Oswaldo Cruz 108, 11 17. https://doi.org/10.1590/0074-0276130395

Prevots, D.R., 1991. The effect of human mobility on the geographic spread of dengue fever in Mexico. University of Michigan.

Punch, K.F., 2014. Introduction to Social Research - Quantitative and Qualitative Approaches. Sage Publications, University of Western Australia.

Punpuing, S., 1993. Correlates of commuting patterns : a case-study of Bangkok, Thailand. Urban Stud. 30, 527 545.

Punpuing, S., Ross, H., 2001. Commuting : The human side of Bangkok's transport problems. Cities 18, 43 50.

Quesnot, T., 2016. L'involution géographique : des données géosociales aux algorithmes. Netcom Réseaux Commun. Territ. 281 304. http://netcom.revues.org/2545

Rainham, D., McDowell, I., Krewski, D., Sawada, M., 2010. Conceptualizing the healthscape : Contributions of time geography, location technologies and spatial ecology to place and health research. Soc. Sci. Med. 70, 668 676. https://doi.org/10.1016/j.socscimed.2009.10.035

Ratanawaraha, A., Chalermpong, S., 2016. How the Poor Commute in Bangkok, Thailand. Transp. Res. Rec. J. Transp. Res. Board 2568, 83 89. https://doi.org/10.3141/2568-13

Rault, Y.-M., Mathew, S., Cebeillac, A., 2018. The social dynamics of india's shopping malls. Bull. Assoc. Géographes Fr. 43 60.

Ravenstein, E.G., 1885. The Laws of Migration. J. Stat. Soc. Lond. 48, 167 235. https://doi.org/10.2307/2979181

RCCES, 2017. RCCES. Regional Center For Climate and Environmental Studies. http://www.rcces.soc.cmu.ac.th/.

Reich, N.G., Shrestha, S., King, A.A., Rohani, P., Lessler, J., Kalayanarooj, S., Yoon, I.-K., Gibbons, R.V., Burke, D.S., Cummings, D.A.T., 2013. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. J. R. Soc. Interface 10, 20130414 20130414. https://doi.org/10.1098/rsif.2013.0414

Reilly, W.J., 1931. The law of retail gravitation. Knickerbocker Press, New York.

Reiner, R.C., Stoddard, S.T., Scott, T.W., 2014. Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal. Epidemics 6, 30 36. https://doi.org/10.1016/j.epidem.2013.12.003

Reiter, P., 1998. Aedes albopictus and the world trade in used tires, 1988-1995 : the shape of things to come? J. Am. Mosq. Control Assoc. 14, 83 94. http://www.biodiversitylibrary.org/content/part/JAMCA/JAMCA_V14_N1_P083-094.pdf

Reiter, P., Amador, M.A., Anderson, R.A., Clark, G.G., 1995. Short Report : Dispersal of Aedes aegypti in an Urban Area after Blood Feeding as Demonstrated by Rubidium-Marked Eggs. Am. Soc. Trop. Med. Hyg. 52, 177 179.

Reiter, P., Sprenger, D., 1987. The used tire trade : a mechanism for the world wide dispersal of container breeding mosquitoes. J. Am. Mosq. Control Assoc. 3, 494 501.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S.J., Chong, S., 2011. On the Levy-Walk Nature of Human Mobility. IEEEACM Trans. Netw. 19, 630 643. https://doi.org/10.1109/TNET.2011.2120618

Richardson, T., Jensen, O.B., 2008. How mobility systems produce inequality : Making mobile subject types on the Bangkok Sky Train. Built Environ. 34, 218 231.

Riley, S., 2007. Large-scale spatial-transmission models of infectious disease. Science 316, 1298 1301. http://science.sciencemag.org/content/316/5829/1298.short

Riley, S., Eames, K., Isham, V., Mollison, D., Trapman, P., 2015. Five challenges for spatial epidemic models. Epidemics 10, 68 71. https://doi.org/10.1016/j.epidem.2014.07.001

Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., Giannotti, F., 2014. The purpose of motion : Learning activities from Individual Mobility Networks. IEEE, pp. 312 318. https://doi.org/10.1109/DSAA.2014.7058090

Ríos-Velásquez, C.M., Codeço, C.T., Honório, N.A., Sabroza, P.S., Moresco, M., Cunha, I.C., Levino, A., Toledo, L.M., Luz, S.L., 2007. Distribution of dengue vectors in neighborhoods with different urbanization types of Manaus, state of Amazonas, Brazil. Mem. Inst. Oswaldo Cruz 102, 617 623. http://www.scielo.br/scielo.php?pid=S0074-02762007000500012&script=sci_arttext.

Robette, N., Bry, X., 2012. Harpoon or Bait ? A Comparison of Various Metrics in Fishing for Sequence Patterns. Bull. Sociol. Methodol. Méthodologie Sociol. 116, 5 24. https://doi.org/10.1177/0759106312454635

Rodhain, F., 1995. Aedes albopictus : a potential problem in France. Parassitologia 37, 115 124.

Rodrigues, F., Pereira, F.C., Alves, A., Jiang, S., Ferreira, J., 2012. Automatic classification of points-of-interest for land-use analysis, in : Proceedings of Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services, January. pp. 41 49. http://ares.lids.mit.edu/fm/documents/automatic_landuse.pdf

Ross, R., 1916. An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part I. Proc. R. Soc. Math. Phys. Eng. Sci. 92, 204 230. https://doi.org/10.1098/rspa.1916.0007

Ross, R., 1911a. The Prevention of Malaria, second edition. ed. John Murray, London.

Ross, R., 1911b. Some quantitatives studies in epidemiology. Nature 87, 466 467. https://doi.org/10.1038/087466a0

Ross, R., 1908. Report on the prevention of malaria in Mauritius. P. Dutton & Company, New York.

Ross, R., 1902. Mosquito brigades and how to organize them. Longmans, Green, London.

Ross, R., 1897. On some peculiar pigmented cells found in two mosquitos fed on malarial blood. Br. Med. J. 2, 1786.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2408186/.

Ross, R., Hudson, H.P., 1917. An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part III. Proc. R. Soc. Math. Phys. Eng. Sci. 93, 225 240. Lien : https://doi.org/10.1098/rspa.1917.0015

Rossman, G.B., Wilson, B.L., 1985. Numbers and Words : Combining Quantitative and Qualitative Methods in a Single Large-Scale Evaluation Study. Eval. Rev. 9, 627 643. https://doi.org/10.1177/0193841X8500900505

Rosvall, M., Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. 105, 1118 1123.https://doi.org/10.1073/pnas.0706851105

Roth, C., Kang, S.M., Batty, M., Barthélemy, M., 2011. Structure of Urban Movements : Polycentric Activity and Entangled Hierarchical Flows. PLoS ONE 6, e15923.https://doi.org/10.1371/journal.pone.0015923.

Rush, B., 1789. An account of the Bilious remitting Yellow Fever, as it appeared in the City of Philadelphia, in the Summer and Autumn of the year. Medical inquiries and observation, Philadelphia, pp. 104 117.

Sabchareon, A., Wallace, D., Sirivichayakul, C., Limkittikul, K., Chanthavanich, P., Suvannadabba, S., Jiwariyavej, V., Dulyachai, W., Pengsaa, K., Wartel, T.A., others, 2012. Protective efficacy of the recombinant, live-attenuated, CYD tetravalent dengue vaccine in Thai schoolchildren : a randomised, controlled phase 2b trial. The Lancet 380, 1559 1567. http://www.sciencedirect.com/science/article/pii/S0140673612614287.

Sabin, A.., 1952. Research on dengue during World War II. Am. J. Trop. Med. Hyg. 30 50.

Sadilek, A., Kautz, H., Bigham, J.P., 2012. Finding your friends and following them to where you are, in : Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM, pp. 723 732.

Sadin, E., 2015. La vie algorithmique : critique de la raison numérique, l'Echappée. ed, Pour en finir avec.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors, in : Proceedings of the 19th International Conference on World Wide Web, WWW '10. ACM, New York, NY, USA, pp. 851 860. https://doi.org/10.1145/1772690.1772777

Salje, H., Lessler, J., Endy, T.P., Curriero, F.C., Gibbons, R.V., Nisalak, A., Nimmannitya, S., Kalayanarooj, S., Jarman, R.G., Thomas, S.J., Burke, D.S., Cummings, D.A.T., 2012. Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. Proc. Natl. Acad. Sci. 109, 9535 9538. https://doi.org/10.1073/pnas.1120621109

Salje, H., Lessler, J., Maljkovic Berry, I., Melendrez, M.C., Endy, T., Kalayanarooj, S., A-Nuegoonpipat, A., Chanama, S., Sangkijporn, S., Klungthong, C., Thaisomboonsuk, B., Nisalak, A., Gibbons, R.V., Iamsirithaworn, S., Macareo, L.R., Yoon, I.-K., Sangarsang, A., Jarman, R.G., Cummings, D.A.T., 2017. Dengue diversity across spatial and temporal scales : Local structure and the effect of host population size. Science 355, 1302 1306. https://doi.org/10.1126/science.aaj9384

Salje, H., Morales, I., Gurley, E.S., Saha, S., 2016. Seasonal Distribution and Climatic Correlates of Dengue Disease in Dhaka, Bangladesh. Am. J. Trop. Med. Hyg. 94, 1359 1361. https://doi.org/10.4269/ajtmh.15-0846

Sanders, L., 2006. Les modèles agent en géographie urbaine, in : Modélisation et simulation multi-agents; applications pour les Sciences de l'Homme et de la Société. Hermes-Lavoisier, pp. 151 168.

Sapiezynski, P., Stopczynski, A., Gatej, R., Lehmann, S., 2015. Tracking Human Mobility Using WiFi Signals. PLOS ONE 10, e0130824. https://doi.org/10.1371/journal.pone.0130824

Sarzynska, M., Udiani, O., Zhang, N., 2013. A study of gravity-linked metapopulation models for the spatial spread of dengue fever. ArXiv Prepr. ArXiv13084589. https://arxiv.org/abs/1308.4589.

Sattenspiel, L., Dietz, K., 1995. A structured epidemic model incorporating geographic mobility among regions. Math. Biosci. 128, 71 91. http://www.sciencedirect.com/science/article/pii/002555649400068B.

Schlapfer, M., Bettencourt, L.M.A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G.B., Ratti, C., 2014. The scaling of human interactions with city size. J. R. Soc. Interface 11, 20130789 20130789. https://doi.org/10.1098/rsif.2013.0789

Schlieder, C., Matyas, C., 2009. Photographing a City : An Analysis of Place Concepts Based on Spatial Choices. Spat. Cogn. Comput. 9, 212 228. Lien : http://www.tandfonline.com/doi/abs/10.1080/13875860903121848.https://doi.org/10.1080/13875860903121848

Schlink, U., Strebel, K., Loos, M., Tuchscherer, R., Richter, M., Lange, T., Wernicke, J., Ragas, A., 2010. Evaluation of human mobility models, for exposure to air pollutants. Sci. Total Environ. 408, 3918 3930. Lien : https://doi.org/10.1016/j.scitotenv.2010.03.018

Schneider, C.M., Belik, V., Couronne, T., Smoreda, Z., Gonzalez, M.C., 2013. Unravelling daily human mobility motifs. J. R. Soc. Interface 10, 20130246 20130246. Lien : https://doi.org/10.1098/rsif.2013.0246

Schneider, F., 2017. New Estimates for the Shadow Economies of 11 Asian Countries from 2000 to 2014, in : Rövekamp, F., Bälz, M., Hilpert, H.G. (Eds.), Cash in East Asia. Springer International Publishing, Cham, pp. 27 41. https://doi.org/10.1007/978-3-319-59846-8_3

Schwartz, E., Meltzer, E., Mendelson, M., Tooke, A., Steiner, F., Gautret, P., Friedrich-Jaenicke, B., Libman, M., Bin, H., Wilder-Smith, A., others, 2013. Detection on four continents of dengue fever cases related to an ongoing outbreak in Luanda, Angola, March to May 2013. Euro Surveill 18, pii 20488. http://www.eurosurveillance.org/images/dynamic/EE/V18N21/art20488.pdf

Scott, J., 1992. Social Network Analysis. Sage Publications, London.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., Haklay, M. (Muki), 2017. A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 31, 139 167. https://doi.org/10.1080/13658816.2016.1189556

Septfons, A., Noël, H., Leparc-Goffart, I., Giron, S., Delisle, E., Chappert, J.L., 2015. Surveillance du chikungunya et de la dengue en France métropolitaine, 2014. Bull. Epidémiologique Hebd. 13 14, 204 2011. Lien : http://fulltext.bdsp.ehesp.fr/Invs/BEH/2015/13-14/1.pdf.

Service, M.W., 1978. Review article : A Short history of early medical entomology. J. Med. Entomol. Lien : http://jme.oxfordjournals.org/content/14/6/603.abstract.

Servin, C., 2006. Réseaux et Télécoms, Dunod. ed.

Shah, H., 2018. Use our personal data for the common good. Nature 556. https://doi.org/10.1038/d41586-018-03912-z

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379 423.

Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F., 2017. The spread of fake news by social bots. Eprint ArXiv170707592.

Sheikh, S., Banda, S., 2014. The Thin Line between Legitimate and Illegal : Regularising Unauthorised Colonies in Delhi, A report of the Cities of Delhi project. Centre for Policy Research, New Delhi.

Sheller, M., Urry, J., 2016. Mobilizing the new mobilities paradigm. Appl. Mobilities 1, 10 25. https://doi.org/10.1080/23800127.2016.1151216

Sheller, M., Urry, J., 2006. The New Mobilities Paradigm. Environ. Plan. A 38, 207 226. https://doi.org/10.1068/a37268

Shen, L., Stopher, P.R., 2014. Review of GPS Travel Survey and GPS Data-Processing Methods. Transp. Rev. 34, 316 334. https://doi.org/10.1080/01441647.2014.903530

Shinawatra, W., 2012. Understanding cultural landscapes in thaï urban context : Bangkock as a neglecting water-based city. Conference, march.

Shlesinger, M.F., Klafter, J., 1986. Lévy Walks Versus Lévy Flights, in : Stanley, H.E., Ostrowsky, N. (Eds.), On Growth and Form. Springer Netherlands, Dordrecht, pp. 279 283. https://doi.org/10.1007/978-94-009-5165-5

Shoval, N., Auslander, G., Cohen-Shalom, K., Isaacson, M., Landau, R., Heinik, J., 2010. What can we learn about the mobility of the elderly in the GPS era ? J. Transp. Geogr. 18, 603 612. https://doi.org/10.1016/j.jtrangeo.2010.03.012

Silveira, L.M., de Almeida, J.M., Marques-Neto, H.T., Sarraute, C., Ziviani, A., 2016. MobHet : Predicting human mobility using heterogeneous data sources. Comput. Commun. https://doi.org/10.1016/j.comcom.2016.04.013

Simini, F., González, M.C., Maritan, A., Barabási, A.-L., 2012. A universal model for mobility and migration patterns. Nature 484, 96 100. https://doi.org/10.1038/nature10856

Sloan, L., Morgan, J., Burnap, P., Williams, Ma., 2015a. Who Tweets ? Deriving the demographic characteristics of age, occupation and socialClass from Twitter User Meta-Data. PloS ONE. https://doi.org/doi:10.1371/journal.pone.0115545

Sloan, L., Morgan, Jeffrey, Burnap, Pete, Williams, Ma., 2015b. Who Tweets ? Deriving the demographic characteristics of age, occupation and socialClass from Twitter User Meta-Data. PloS ONE. https://doi.org/doi:10.1371/journal.pone.0115545

Smith, C.E., 1956. The history of dengue in tropical Asia and its probable relationship to the mosquito Aedes aegypti. J. Trop. Med. 59, 243 251.

Smith, D.L., Battle, K.E., Hay, S.I., Barker, C.M., Scott, T.W., McKenzie, F.E., 2012. Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. PLoS Pathog. 8, e1002588. https://doi.org/10.1371/journal.ppat.1002588

Sofean, M., Mustafa, Smith, 2012. A Real-Time Architecture for Detection of Diseases using Social Networks : Design , Implementation and Evaluation. Proc. 23rd ACM Conf. Hypertext Soc. Media 309 310.

Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010a. Modelling the scaling properties of human mobility. Nat. Phys. 6, 818 823. https://doi.org/10.1038/nphys1760

Song, C., Qu, Z., Blumm, N., Barabasi, A.-L., 2010b. Limits of Predictability in Human Mobility. Science 327, 1018 1021. https://doi.org/10.1126/science.1177170

Soper, F.L., 1970. Building the health bridge. Selections from the works of Fred L. Soper, M.D. London. Bloomington : Indiana University Press, Indiana 47401, U.S.A.

Soper, F.L., 1967. Dynamics of Aedes aegypti distribution and density. Seasonal fluctuations in the Americas. Bull. World Health Organ. 36, 536. Lien : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2476418/.

Sophorntavy, V., 2017. A Meeting of Masks. Status, Power and Hierarchy in Bangkok, NIAS Monographs. ed.

Soto, V., Frías-Martínez, E., 2011. Automated land use identification using cell-phone records, in : Proceedings of the 3rd ACM International Workshop on MobiArch. ACM, pp. 17 22. Lien : http://dl.acm.org/citation.cfm?id=2000179.

Spedicato, G.A., 2017. Discrete Time Markov Chains with R. R J. Lien : https://journal.r-project.org/archive/2017/RJ-2017-036/index.html.

Steiger, E., Albuquerque, J.P., Zipf, A., 2015a. An advanced systematic literature review on spatiotemporal analyses of twitter data. Trans. GIS 19, 809 834. Lien : http://onlinelibrary.wiley.com/doi/10.1111/tgis.12132/full.

Steiger, E., Westerholt, R., Resch, B., Zipf, A., 2015b. Twitter as an indicator for whereabouts of people ? Correlating Twitter with UK census data. Comput. Environ. Urban Syst. 54, 255 265. Lien : https://doi.org/10.1016/j.compenvurbsys.2015.09.007.

Stewart, J.Q., 1942. A Measure of the Influence of a Population at a Distance. Sociometry 5, 63. Lien : https://doi.org/10.2307/2784954

Stoddard, S.T., Forshey, B.M., Morrison, A.C., Paz-Soldan, V.A., Vazquez-Prokopec, G.M., Astete, H., Reiner, R.C., Vilcarromero, S., Elder, J.P., Halsey, E.S., Kochel, T.J., Kitron, U., Scott, T.W., 2013. House-to-house human movement drives dengue virus transmission. Proc. Natl. Acad. Sci. 110, 994 999. https://doi.org/10.1073/pnas.1213349110

Stoddard, S.T., Morrison, A.C., Vazquez-Prokopec, G.M., Paz Soldan, V., Kochel, T.J., Kitron, U., Elder, J.P., Scott, T.W., 2009. The Role of Human Movement in the Transmission of Vector-Borne Pathogens. PLoS Negl. Trop. Dis. 3, e481. https://doi.org/10.1371/journal.pntd.0000481.

Stouffer, S.A., 1940. Intervening Opportunities : A Theory Relating Mobility and Distance. Am. Sociol. Rev. 5, 845. https://doi.org/10.2307/2084520

Succo, T., Leparc-Goffart, I., Ferré, J.-B., Roiz, D., Broche, B., Maquart, M., Noel, H., Catelinois, O., Entezam, F., Caire, D., Jourdain, F., Esteve-Moussion, I., Cochet, A., Paupy, C., Rousseau, C., Paty, M.-C., Golliot, F., 2016. Autochthonous dengue

outbreak in Nîmes, South of France, July to September 2015. Eurosurveillance 21. Lien : http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22485. https://doi.org/10.2807/1560-7917.ES.2016.21.21.30240

Sullivan, B.M., Karthikeyan, G., Liu, Z., Massa, W.L.P., Gupta, M., 2018. Socioeconomic group classification based on user features. Lien : http://pdfaiw.uspto.gov/.aiw?PageNum=0&docid=20180032883&IDKey=175CD64C2991&HomeUrl=http%3A%2F%2Fappft.uspto.gov%2Fnetacgi%2Fnph-Parser%3FSect1%3DPTO1%2526Sect2%3DHITOFF%2526d%3DPG01%2526p%3D1%2526u%3D%25252Fnetahtml%25252FPTO%25252Fsrchnum.html%2526r%3D1%2526f%3DG%2526l%3D50%2526s1%3D%25252220180032883%252522.PGNR.%2526OS%3DDN%2F20180032883%2526RS%3DDN%2F20180032883.

Sun, L., Axhausen, K.W., Lee, D.-H., Huang, X., 2013. Understanding metropolitan patterns of daily encounters. Proc. Natl. Acad. Sci. 110, 13774 13779. https://doi.org/10.1073/pnas.1306440110.

Sun, L.-S., Wang, S.-W., Yao, L.-Y., Rong, J., Ma, J.-M., 2016. Estimation of transit ridership based on spatial analysis and precise land use data. Transp. Lett. 1 8. Lien : https://doi.org/10.1179/1942787515Y.0000000017

Sun, Y., Fan, H., Helbich, M., Zipf, A., 2013. Analyzing Human Activities Through Volunteered Geographic Information : Using Flickr to Analyze Spatial and Temporal Pattern of Tourist Accommodation, in : Krisp, J.M. (Ed.), Progress in Location-Based Services. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 57 69. https://doi.org/10.1007/978-3-642-34203-5_4

Supatn, N., 2011. Industrial Estates, Ports, Airports and City Transport in the Greater Bangkok Area for Promoting Connectivity in the Mekong Region (No. 6), Intra - and Inter - City Connectivity in the Mekong Region. BRC Research.

Taylor, L., 2016. No place to hide ? The ethics and analytics of tracking mobility using mobile phone data. Environ. Plan. Soc. Space 34, 319 336. https://doi.org/10.1177/0263775815608851

Telle, O., 2015. Géographie d'une maladie émergente en milieu urbain endémique, le cas de la dengue à Delhi, Inde. Cybergeo Eur. J. Geogr. Lien : http://journals.openedition.org/cybergeo/26921.

Telle, O., 2011. Aedes : Analyse de l'émergence de la dengue et simulation spatiale. Université de Rouen.

Telle, O., Vaguet, A., Yadav, N.K., Lefebvre, B., Daudé, E., Paul, R.E., Cebeillac, A., Nagpal, B.N., 2016. The Spread of Dengue in an Endemic Urban Milieu The Case of Delhi, India. PLOS ONE 11, e0146539. https://doi.org/10.1371/journal.pone.0146539

ten Bosch, Q.A., Clapham, H.E., Lambrechts, L., Duong, V., Buchy, P., Althouse, B.M., Lloyd, A.L., Waller, L.A., Morrison, A.C., Kitron, U., Vazquez-Prokopec, G.M., Scott, T.W., Perkins, T.A., 2018. Contributions from the silent majority dominate dengue virus transmission. PLOS Pathog. 20.

Teurlai, M., Huy, R., Cazelles, B., Duboz, R., Baehr, C., Vong, S., 2012. Can Human Movements Explain Heterogeneous Propagation of Dengue Fever in Cambodia ? PLoS Negl. Trop. Dis. 6, e1957. https://doi.org/10.1371/journal.pntd.0001957

The Trinh, D., Wills, B., 2014. Clinical Features of Dengue, in : Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J., C.A.B. International (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI, Wallingford, Oxfordshire ; Boston, MA, pp. 115 144.

Thomas, I., Adam, A., Verhetsel, A., 2017. Migration and commuting interactions fields : a new geography with community detection algorithm ? Belgeo. https://doi.org/10.4000/belgeo.20507

Thomas, I., Verhetsel, A., Witlox, F., 2009. Incorporer l'espace dans la modélisation du choix de destination : le cas de 4 villes flamandes. Cybergeo.https://doi.org/10.4000/cybergeo.22192

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C.M., Blondel, V., Smoreda, Z., González, M.C., Colizza, V., 2014. On the Use of Human Mobility Proxies for Modeling Epidemics. PLoS Comput. Biol. 10, e1003716.https://doi.org/10.1371/journal.pcbi.1003716

Tizzoni, M., Sun, K., Benusiglio, D., Karsai, M., Perra, N., 2015. The Scaling of Human Contacts and Epidemic Processes in Metapopulation Networks. Sci. Rep. 5, 15111.https://doi.org/10.1038/srep15111

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. 46, 234. https://doi.org/10.2307/143141

Tomasello, D., Schlagenhauf, P., 2013. Chikungunya and dengue autochthonous cases in Europe, 2007 2012. Travel Med. Infect. Dis. 11, 274 284. https://doi.org/10.1016/j.tmaid.2013.07.006

Toole, J.L., Herrera-Yaque, C., Schneider, C.M., Gonzalez, M.C., 2015. Coupling human mobility and social ties. J. R. Soc. Interface 12, 20141128 20141128. https://doi.org/10.1098/rsif.2014.1128

Toole, J.L., Ulm, M., González, M.C., Bauer, D., 2012. Inferring land use from mobile phone activity, in : Proceedings of the ACM SIGKDD International Workshop on Urban Computing. ACM, pp. 1 8. http://dl.acm.org/citation.cfm?id=2346498.

Topalov, C., 1991. La ville, «terre inconnue». L'enquête de Charles Booth et le peuple de Londres, 1886-1891. Genèses 5, 4 34. https://doi.org/10.3406/genes.1991.1075

Torre, C.A., 2009. Deterministic and Stochastic Metapopulation models for Dengue Fever. Arizona State University.

Trpis, M., Hausermann, W., 1986. Dispersal and other Population Parameters of Aedes aegypti in an African Village and their Possible Significance in Epidemiology of Vector-Borne Diseases. Am. Soc. Trop. Med. Hyg. 35, 1263 1279.

Tsuda, Y., Suwonkerd, W., Chawprom, S., Prajakwong, S., Takagi, M., 2006. Different spatial distribution of Aedes aegypti and Aedes albopictus along an urban-rural gradient and the relating environmental factors examined in three villages in northern Thailand. J. Am. Mosq. Control Assoc. 22, 222 228. https://doi.org/10.2987/8756-971X(2006)22[222:DSDOAA]2.0.CO;2

Tukey, J.W., 1977. Exploratory data analysis. Reading Massachusetts : Addison-Wesley.

Tun-Lin, W., Kay, B.H., Barnes, A., 1995. Understanding Productivity, A Key to Aedes aegypti Surveillance. Am. Soc. Trop. Med. Hyg. 53, 595 601.

Undurraga, E.A., Halasa, Y.A., Shepard, D.S., 2013. Use of Expansion Factors to Estimate the Burden of Dengue in Southeast Asia : A Systematic Analysis. PLoS Negl. Trop. Dis. 7, e2056. https://doi.org/10.1371/journal.pntd.0002056

United Nations, 2015. Review of maritime transport 2015. United Nations, Place of publication not identified.

United Nations Human Settlements Programme (Ed.), 2008. Housing finance mechanisms in Thaïland, Human settlements finance systems series. United Nations Human Settlements Programme, Nairobi.

Varenne, F., 2008. Épistémologie des modèles et des simulations : tour d'horizon et tendances, in : Les Modèles, Possibilités et Limites. Editions Matériologiques, pp. 13 46. http://www.cairn.info/les-modeles-possibilites-et-limites--9782919694624-page-13.html

Vassal, J.J., Brochet, A., 1908. La dengue en Indo-Chine : épidemie a bord de la Manche en 1907. Ann. Hygiène Médecine Colon. 11, 547 572.

Vazquez-Prokopec, G.M., Kitron, U., Montgomery, B., Horne, P., Ritchie, S.A., 2010. Quantifying the Spatial Dimension of Dengue Virus Epidemic Spread within a Tropical Urban Environment. PLoS Negl. Trop. Dis. 4, e920. https://doi.org/10.1371/journal.pntd.0000920

Vazquez-Prokopec, G.M., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Kochel, T.J., Scott, T.W., Kitron, U., 2009. Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. Int. J. Health Geogr. 8, 68. https://doi.org/10.1186/1476-072X-8-68

Veloso, M., Phithakkitnukoon, S., Bento, C., Fonseca, N., Olivier, P., 2011. Exploratory study of urban flow using taxi traces, in : First Workshop on Pervasive Urban Applications (PURBA) in Conjunction with Pervasive Computing, San Francisco, California, USA. https://www.researchgate.net/profile/Carlos_Bento/publication/232175450_Exploratory_Study_of_Urban_Flow_using_Taxi_Traces/links/09e41507829c808c03000000.pdf

Vichiensan, V., 2009. Urban Mobility and Employment Accessibility in Bangkok : Present and Future. Cooperation for urban mobility in the developing world.

Vichiensan, V., 2007. Dynamics of urban structure in Bangkok based on employment cluster and commuting pattern. J. East. Asia Soc. Transp. Stud. 7, 1559 1574.

Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Murphy, E.J., Prince, P.A., Stanley, H.E., 1996. Lévy flight search patterns of wandering albatrosses. Nature 381, 413 415. http://dx.doi.org/10.1038/381413a0

Vongrattanatoh, S., 2011. Comment les Thaïs traduisent l'idée de slum, taudis et bidonville. Moussons 139 148. https://doi.org/10.4000/moussons.753

Wagner, D., Huzly, D., Hufert, F., Weidmann, M., Breisinger, S., Eppinger, S., Kern, W.V., Bauer, T.M., others, 2004. Nosocomial acquisition of dengue. Emerg Infect Dis 10, 1872 1873. http://www.drplace.com/Nosocomial_acquisition_of_dengue.17.2163.htm

Waite, H., 1910. Mosquitoes and Malaria. A Study of the Relat ion between the Number of Mosquitoes in a Locality and the Malaria Rate. Biometrika 7, 421 436. https://doi.org/10.1093/biomet/7.4.421

Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.-L., 2011. Human mobility, social ties, and link prediction, in : Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1100 1108. http://dl.acm.org/citation.cfm?id=2020581

Wang, S.S., Stefanone, M.A., 2013. Showing Off ? Human Mobility and the Interplay of Traits, Self-Disclosure, and Facebook *check-in*. Soc. Sci. Comput. Rev. 31, 437 457. https://doi.org/10.1177/0894439313481424

Wang, W., Zhao, X.-Q., 2004. An epidemic model in a patchy environment. Math. Biosci. 190, 97 112. https://doi.org/10.1016/j.mbs.2002.11.001

Wang, Y., Callan, J., Zheng, B., 2015. Should We Use the Sample ? Analyzing Datasets Sampled from Twitter's Stream API. ACM Trans. Web 9, 1 23. https://doi.org/10.1145/2746366

Wanich, J., 1983. Soi diaokan, Bangkok : Pheka.

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236. https://doi.org/10.2307/2282967

Washer, P., 2010. Emerging infectious diseases and society. Palgrave Macmillan, New York.

Weaver, S.C., Vasilakis, N., 2009. Molecular evolution of dengue viruses : Contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. Infect. Genet. Evol. 9, 523 540.https://doi.org/10.1016/j.meegid.2009.02.003

Weiler, A., Grossniklaus, M., Scholl, M.H., 2015. Evaluation Measures for Event Detection Techniques on Twitter Data Streams, in : Maneth, S. (Ed.), Data Science. Springer International Publishing, Cham, pp. 108 119. https://doi.org/10.1007/978-3-319-20424-6_11

Wellmer, H., 1983. Dengue Haemorrhagic Fever in Thailand : Geomedical Observations on Developments Over the Period 1970 1979. Springer-Verlag Berlin Heidelberg.

Wen, T.-H., Lin, M.-H., Teng, H.-J., Chang, N.-T., 2015. Incorporating the human-Aedes mosquito interactions into measuring the spatial risk of urban dengue fever. Appl. Geogr. 62, 256 266. https://doi.org/10.1016/j.apgeog.2015.05.00

Wesolowski, A., Buckee, C.O., Bengtsson, L., Wetter, E., Lu, X., Tatem, A.J., 2014. Commentary : Containing the Ebola Outbreak the Potential and Challenge of Mobile Network Data. PLOS Curr. Outbreaks.

Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. J. R. Soc. Interface 10, 20120986 20120986. https://doi.org/10.1098/rsif.2012.0986

Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., O, B.C., 2012. Quantifying the Impact of Human Mobility on Malaria. Science 338, 267 270. https://doi.org/10.1126/science.1223467

Wesolowski, A., Qureshi, T., Boni, M.F., Sundsøy, P.R., Johansson, M.A., Rasheed, S.B., Engø-Monsen, K., Buckee, C.O., 2015a. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc. Natl. Acad. Sci. 112, 11887 11892. https://doi.org/10.1073/pnas.1504964112

Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., Bousema, T., Drakeley, C., Cox, J., Buckee, C.O., 2015b. Quantifying travel behavior for infectious disease research : a comparison of data from surveys and mobile phones. Sci. Rep. 4. https://doi.org/10.1038/srep05678

Wichmann, O., Jelinek, T., 2004. Dengue in Travelers : a Review. J. Travel Med. 11, 161 170.

Wilder-Smith, A., 2014. Dengue Infections in Travelers, in : Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J., C.A.B. International (Eds.), Dengue and Dengue Hemorrhagic Fever. CABI, Wallingford, Oxfordshire ; Boston, MA, pp. 90 98.

Wilder-Smith, A., Gubler, D.J., 2008. Geographic Expansion of Dengue : The Impact of International Travel. Med. Clin. North Am. 92, 1377 1390. https://doi.org/10.1016/j.mcna.2008.07.002

Wilder-Smith, A., Murray, Quam, M., 2013. Epidemiology of dengue : past, present and future prospects. Clin. Epidemiol. 299. https://doi.org/10.2147/CLEP.S34440

Wilder-Smith, A., Quam, M., Sessions, O., Rocklov, J., Liu-Helmersson, J., Franco, L., Khan, K., 2014. The 2012 dengue outbreak in Madeira : exploring the origins. https://dr.ntu.edu.sg/handle/10220/19685

Winter, G., 1983. Deux méthodes d'investigation irréductibles mais complémentaire. Cah. ORSTOM, Série Sciences Humaines, Institut de Recherche pour le Développement (IRD) 20, 17 24. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers4/15289.pdf

Wojna, Z., Gorban, A., Lee, D.-S., Murphy, K., Yu, Q., Li, Y., Ibarz, J., 2017. Attention-based Extraction of Structured Information from Street View Imagery. ArXiv Prepr. ArXiv170403549. https://arxiv.org/abs/1704.03549.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary : Experiment to derive trip purpose from global positioning system travel data. Transp. Res. Rec. J. Transp. Res. Board 125 134. http://trrjournalonline.trb.org/doi/abs/10.3141/1768-15

World Bank, 2007. Strategic Urban Transport Policy Directions for Bangkok. http://siteresources.worldbank.org/INTTHAILAND/Resources/333200-1177475763598/2007june_bkk-urban-transport-directions.pdf

Wu, L., Zhi, Y., Sui, Z., Liu, Y., 2014. Intra-Urban Human Mobility and Activity Transition : Evidence from Social Media Check-In Data. PLoS ONE 9, e97010. https://doi.org/10.1371/journal.pone.0097010

Xie, K., Deng, K., Zhou, X., 2009. From Trajectories to Activities : A Spatio-temporal Join Approach, in : Proceedings of the 2009 International Workshop on Location Based Social Networks, LBSN '09. ACM, New York, NY, USA, pp. 25 32. https://doi.org/10.1145/1629890.1629897

Xue, L., 2013. Modeling and analysis of vector-borne diseases on complex networks. Kansas State University. http://krex.k-state.edu/dspace/handle/2097/16788

Xue, L., Scott, H.M., Cohnstaedt, L.W., Scoglio, C., 2012. A network-based meta-population approach to model Rift Valley fever epidemics. J. Theor. Biol. 306, 129 144.http://www.sciencedirect.com/science/article/pii/S002251931200210X

Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., Wang, W.-X., 2014. Universal predictability of mobility patterns in cities. J. R. Soc. Interface 11, 20140834 20140834.https://doi.org/10.1098/rsif.2014.0834

Yang, F., Jin, P.J., Cheng, Y., Zhang, J., Ran, B., 2015. Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. Int. J. Sustain. Transp. 9, 551 564.https://doi.org/10.1080/15568318.2013.826312

Yang, Z., Algesheimer, R., Tessone, C.J., 2016. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. Sci. Rep. 6. https://doi.org/10.1038/srep30750

Yardi, S., Boyd, D., 2010. Tweeting from the Town Square : Measuring Geographic Local Networks, in : Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Yasmeen, G., Nirathron, N., 2014. Vending in Public Space : The Case of Bangkok. WIEGO Policy Brief 16, 18.

Yoon, I.-K., Rothman, A.L., Tannitisupawong, D., Srikiatkhachorn, A., Jarman, R.G., Aldstadt, J., Nisalak, A., Mammen, M.P., Thammapalo, S., Green, S., Libraty, D.H., Gibbons, R.V., Getis, A., Endy, T., Jones, J.W., Koenraadt, C.J.M., Morrison, A.C., Fansiri, T., Pimgate, C., Scott, T.W., 2012. Underrecognized Mildly Symptomatic Viremic Dengue Virus Infections in Rural Thai Schools and Villages. J. Infect. Dis. 206, 389 398. https://doi.org/10.1093/infdis/jis357

Yu, Q., Szegedy, C., Stumpe, M.C., Yatziv, L., Shet, V., Ibarz, J., Arnoud, S., 2015. Large Scale Business Discovery from Street Level Imagery. ArXiv Prepr. ArXiv151205430. https://arxiv.org/abs/1512.05430.

Yuan, Y., Medel, M., 2016. Characterizing International Travel Behavior from Geotagged Photos : A Case Study of Flickr. PLOS ONE 11, e0154885.https://doi.org/10.1371/journal.pone.0154885

Zandbergen, P.A., Barbeau, S.J., 2011. Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. J. Navig. 64, 381 399.https://doi.org/10.1017/S0373463311000051

Zang, J., Dummit, K., Graves, J., Lisker, P., Sweeney, L., 2015. Who Knows What About Me ? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps. Technol. Sci. 53. http://techscience.org/a/2015103001.

Zhan, X., Ukkusuri, S.V., Zhu, F., 2014. Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. Netw. Spat. Econ. 14, 647 667. https://doi.org/10.1007/

s11067-014-9264-4

Zheng, Q., Hong, X., Liu, J., Cordes, D., Huang, W., 2009. Agenda driven mobility modelling. Int. J. Ad Hoc Ubiquitous Comput. 5, 22 36.

Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X., 2015. You Are Where You Go : Inferring Demographic Attributes from Location *check-in*. ACM Press, pp. 295 304. https://doi.org/10.1145/2684822.2685287

Zipf, G.K., 1946. The P 1 P 2 D Hypothesis : On the Intercity Movement of Persons. Am. Sociol. Rev. 11, 677. https://doi.org/10.2307/2087063

zu Erbach-Schoenberg, E., Alegana, V.A., Sorichetta, A., Linard, C., Lourenço, C., Ruktanonchai, N.W., Graupe, B., Bird, T.J., Pezzulo, C., Wesolowski, A., Tatem, A.J., 2016. Dynamic denominators : the impact of seasonally varying population numbers on disease incidence estimates. Popul. Health Metr. 14. https://doi.org/10.1186/s12963-016-0106-0

Liste des figures

1 Plan de la thèse . . . 9
2 Cycle de transmission de la dengue. . . . 13
3 Cycle de vie du moustique . . . 16
4 Probabilité de survies des femelles *Aedes* en fonction de la température. . . 17
5 Probabilité d'occurrence d'*Aedes aegypti*. . . . 19
6 Probabilité d'occurrence d'*Aedes albopictus*. . . . 19
7 Zones ayant connu des épidémies aux symptômes similaires à ceux de la dengue entre 1635 et 1950 . . . 22
8 Évolution des occurrences de dengue et de moustiques *Aedes Aegypti* et *Albopictus* entre 1960 et 2010 . . . 23
9 Évolution des occurrences de dengue et de moustiques *Aedes Aegypti* et *Albopictus* entre 1960 et 2010 en Asie . . . 24
10 Le système de la dengue, ses composants et leurs interactions. . . . 36
11 Évolution de la population à Delhi depuis 1950. . . . 42
12 Répartition de la population à Delhi. . . . 44
13 Répartition des colonies selon leur catégorie de taxe foncière à Delhi. . . . 47
14 Répartition de la population en fonction de la taxe foncière . . . 48
15 Distances parcourues à Delhi pour se rendre à son lieu de travail. . . . 50
16 Distances parcourues pour se rendre au travail en fonction du genre et du type de transport, dans le sud de Delhi . . . 51
17 Nombre de cas de dengue enregistrés entre 2008 et 2017 à Delhi . . . 51
18 Intensité de la dengue à Delhi en 2008, 2009 et 2010 . . . 52
19 Pourcentage de la population par décile de densité et nombre de cas de dengue par décile et par année . . . 53
20 Évolution de la surface bâtie à Bangkok depuis 1850. . . . 54
21 Évolution de la population à Bangkok entre 1919 et 2012. . . . 55
22 Répartition de la population à Bangkok . . . 56
23 Exemples de maisons au bord des *khlongs* dans le secteur de Bang Wa. . . . 57
24 Exemples d'habitations, entre Bang Wa et Wutthakat. . . . 58
25 Quelques ensembles d'immeubles à Bangkok . . . 59
26 Bangkok et ses quartiers très contrastés. Exemple de Sathorn . . . 60
27 Répartition des types de bâti dans Bangkok . . . 61
28 Répartition de la population à Bangkok au domicile (a) et en journée (b). . 63
29 Tendances de déplacements à Bangkok selon les activités à réaliser (NSO, 2009) . . . 65
30 Nombre de cas de dengue à Bangkok en 2013 par tranche d'âge . . . 68
31 Carte par anamorphose de la répartition des cas de dengue par khet entre 2005 et 2013 . . . 68
32 Lien entre précipitation et cas de dengue enregistrés à Bangkok entre 2005 et 2012. . . . 70
33 Décomposition temporelle des cas de dengue à Bangkok. . . . 71
34 Décomposition temporelle des précipitations à Bangkok. . . . 71
35 Corrélations croisées temporelles entre les données non bruitées des précipitations et des cas de dengue . . . 72
36 Comparaison des données débruitées et centrées-réduites pour les cas de dengue et les précipitations . . . 72

37 Comparaison des séries temporelles de la saisonnalité des cas de dengue et des précipitations . . . 73
38 Spatialisation des coefficients R^2 issus d'une régression linéaire entre les séries temporelles des cas de dengue par *Khet* et l'ensemble de Bangkok . . 74
39 Lien entre le nombre de cas de dengue enregistrés et la population par *Khet* (a) et lien entre l'incidence annuelle et la densité de population (b). 75
40 Représentation schématique de quelques éléments susceptibles d'influencer les mobilités individuelles en zone urbaine. 91
41 Exemple d'espace d'activité pour deux individus 94
42 Organisation schématique du réseau d'antennes téléphonique 111
43 Localisation d'un téléphone auprès d'antennes relais 112
44 Illustration d'un réseau social avec Géolocalisation 118
45 Extrait de la politique de confidentialité de *Twitter* concernant la collecte des données utilisateur 119
46 Illustration d'un « *check-in* » sur Swarm 120
47 « Laissez votre empreinte », une allégorie de l'identité numérique 126
48 La métaphore des éboueurs 128
49 Le projet SAFARI 129
50 Données utilisées dans les travaux de notre revue bibliographique 140
51 Répartition des taux de pénétration de téléphones portables dans le monde en 2015 141
52 Taux de pénétration d'Internet en 2016 par pays 142
53 Rang par pays pour *Facebook*, *Twitter* et Instagram 143
54 Le modèle r-EPR 157
55 Représentation schématique du modèle WHERE 158
56 Présentation du modèle de Schneider *et al.* (2013) 159
57 Principe du modèle rang / distance 160
58 Rang de *Twitter* parmi les autres réseaux sociaux par pays. 181
59 Comparaison des volumes de données *Twitter* collectés à Bangkok et à Delhi 191
60 Nombre de *tweets* par utilisateurs à Delhi et Bangkok. 192
61 Nombre de jours actifs sur la période et nombre moyen de *tweets* envoyé par jours par utilisateur (activité). 192
62 Part des distances entre deux messages successifs pour chaque utilisateur à Delhi et Bangkok. 193
63 Proportion (a) et pourcentage cumulé de *tweets* selon l'intervalle de temps entre deux messages successifs (envoyés par chaque utilisateur) à Bangkok et Delhi. 193
64 Proportion de *tweets* selon l'intervalle de temps entre deux messages successifs sur une semaine et sur un mois 194
65 Vitesse instantanée et moyenne d'envoi de *tweet* 195
66 Des utilisateurs qui *tweetent* exactement dans la même localisation à Bangkok 199
67 Nombre de messages envoyés par tranche horaire sur une semaine à Delhi et Bangkok 201
68 Principe de fonctionnement de l'algorithme *db-scan* 203
69 Densité du nombre d'utilisateurs en fonction du nombre de *clusters* (lieux) qui composent leur espace d'activité. 204

70 Pourcentage de clusters en fonction du nombre de *tweets* qui les composent (a) et pourcentage du nombre de jours où un utilisateur a *tweeté* depuis ce lieu (b), à Bangkok. . . . 204
71 Nombre de *tweets* par tranche horaire par lieu fréquenté pour deux utilisateurs. 206
72 Représentativité de l'échantillon à Bangkok . . . 207
73 Cartographie des résidus . . . 208
74 Densité des domiciles des utilisateurs de *Twitter* au regard de la densité de population à Bangkok. . . . 208
75 Densité des domiciles des utilisateurs de *Twitter* au regard de la densité de population à Delhi. . . . 209
76 Lien entre la population enregistrée au sous-district par le recensement de 2011 à Delhi et le nombre d'utilisateurs de *Twitter* habitant dans les mêmes *ward*. . . . 210
77 Création d'un *check-in* sur *Facebook*. . . . 212
78 Quelques exemples de lieux mal référencés dans la base de *Facebook*. . . . 213
79 Exemple de badges et autres récompenses qu'une personne éditant des contenus sur *Facebook* peut recevoir. . . . 214
80 Exemple de résultat d'une requête sur *Facebook Places Search*. . . . 216
81 Principe de fonctionnement de la fenêtre mobile permettant de collecter les données *Facebook*. . . . 216
82 Distribution du nombre de *check-in* (a) et du nombre de jours différents de check-in (b) par lieu . . . 218
83 Nombre de *check-in* enregistrés par tranche horaire . . . 218
84 *Facebook* et la gestion des lieux doublons . . . 219
85 Résultat de notre algorithme de suppression de pics dans 12 lieux . . . 221
86 Résultat de notre algorithme de suppression de pics sur l'ensemble des données 221
87 Répartition des *check-in* à Bangkok . . . 222
88 Nombre moyen de *check-in* effectué par tranche horaire sur une semaine . . . 223
89 Nombre de lieux et de *check-in* en fonction des catégories. . . . 223
90 Profils temporels sur une semaine des 20 catégories les plus visitées sur *Facebook*. . . . 224
91 Catégorie de taxe foncière des colonies du secteur de Malviya Nagar. . . . 234
92 Localisation des différents quartiers de la zone d'étude. . . . 235
93 Quelques photos de Malviya Nagar. . . . 236
94 Hauz Rani. . . . 237
95 Khirki. . . . 237
96 Begampur Village . . . 238
97 Extérieur du Valmiki Camp, Begumpur . . . 239
98 Le *Main Market* de Malviya Nagar . . . 240
99 Le mall Select Citywalk de Saket . . . 240
100 Les cas de dengue à Malviya Nagar. . . . 241
101 Fumigation à Khirki Extension . . . 242
102 Une ambiance aux abords d'un marché en Inde . . . 246
103 Des Hommes qui attendent . . . 247
104 Portraits de femmes . . . 247
105 Localisation des lieux d'interview et des domiciles des personnes interrogées. 248
106 Structure démographique de l'échantillon . . . 249
107 Effectifs des personnes interrogées par type d'activité principale. . . . 250

108 Localisation des domiciles des personnes interrogées 251
109 Répartition des effectifs par colonie et par catégorie de la taxe foncière. . . . 252
110 Répartition de la population par catégorie de la taxe foncière 252
111 Répartition de l'échantillon selon le revenu moyen individuel (gauche) et le revenu moyen du foyer (droite). 253
112 Résultat d'une ACP sur des indicateurs de richesse 254
113 Lien entre le niveau de richesse estimé et la catégorie de la taxe foncière du lieu de domicile. 256
114 Répartition des activités en fonction des interviewés 257
115 Distance des activités au domicile et fréquence de visite hebdomadaire. . . . 258
116 Fréquence de visite hebdomadaire d'une activité selon sa distance au domicile. 259
117 lieux déclarés comme fréquentés par les hommes et les femmes 260
118 Lieux fréquentés, en fonction du niveau de richesse. 261
119 Part de chaque groupe socio-économique déclarant effectuer une activité donnée. 261
120 Localisation des activités effectuées à Malviya Nagar, selon le groupe socio-économique. 263
121 Différentes métriques de dispersions selon le genre. 264
122 Distance parcourue pour effectuer une activité en fonction du genre. 264
123 Fonction de densité exprimant le nombre d'activités, de lieux, le rayon de giration et enveloppe convexe de l'échantillon en fonction (a) de l'âge des individus et (b) de leur niveau de richesse estimé. 265
124 Fonction de densité représentant les tendances de déplacements des différents groupes de la classification 266
125 Caractéristiques socio-démographiques des différents groupes de la classification 267
126 Probabilité d'effectuer une activité à une tranche horaire donnée (gauche) et localisation de l'espace d'activité (droite) pour deux personnes. 271
127 Probabilité d'effectuer une activité à une tranche horaire donnée (gauche) et localisation de l'espace d'activité (droite) d'une personne n'ayant pas d'activité rémunérée. 271
128 Exemple théorique montrant pour une activité de durée maximum de 4h, la probabilité de tirer une durée (entre 1 et 4h) en fonction de la distance x au domicile, suivant une fonction du type $y=1/x+x^2$, après 1000 itérations. . . 275
129 Exemple d'agenda reconstitué pour un interviewé (81). 277
130 Exemple d'agenda reconstitué pour un interviewé (33). 278
131 Exemple d'agenda reconstitué pour un interviewé (86). 278
132 Résultat d'une CAH réalisée sur la matrice de dissimilarité entre agendas . 281
133 Séquences d'activités moyennes par classes. 281
134 Séquence d'activités de chaque individu par classe. 282
135 Répartition des individus par classe en fonction de leur niveau de richesse estimé. 282
136 Élaboration d'une chaîne markovienne à partir des taux de transitions calculés pour l'ensemble de l'échantillon pour une simulation d'agendas. . . 284
138 Probabilité d'effectuer une autre activité lorsque l'activité précédente se termine. 286
139 Distance entre chaque type d'activités (10 encadrés) et l'ensemble des autres activités pour l'ensemble des utilisateurs. 287

140 Densité de fréquentation des différentes zones de la ville de Delhi à différentes tranches horaires. . . . 293
141 Comparaison entre la part d'habitant et le nombre de domiciles d'utilisateurs de *Twitter* estimé selon les colonies d'une taxe foncière donnée. . . . 294
142 Densité des contenus et d'édition de la base OSM en 2014 . . . 297
143 Utilisation du sol à Malviya Nagar d'après les données OSM . . . 299
144 Exemple de *captcha* permettant d'entraîner les algorithmes de reconnaissances d'images de *Google*. . . . 300
145 Le code couleur des cartes de *Google* . . . 301
146 Résultat d'une extraction des lieux d'éducation, des parcs et des *AOI* de *Google Maps*. . . . 302
147 Utilisation du sol à Delhi en combinant les données d'OSM et de *Google Maps*. 303
148 Utilisation du sol à Malviya Nagar en combinant les données d'OSM et de *Google Maps*. . . . 304
149 Répartition des types d'activités parmi tous les lieux fréquentés (A) et part des utilisateurs exerçant au moins une fois une des activités (B). . . . 304
150 Part d'utilisateur selon le nombre de lieux fréquentés (A) ou d'activités réalisées (B). . . . 305
151 Nombre de personnes par catégorie de l'utilisation du sol, par tranche horaire sur une semaine type. . . . 307
152 Horaire de présences à l'activité principale en fonction du type de lieu associé à cette dernière. . . . 308
153 Localisation des domiciles et des activités principales probables des utilisateurs de *Twitter* à Delhi. . . . 309
154 Évolution des probabilités de réalisation d'une activité ayant une fréquence de visite hebdomadaire initiale de 0.5. . . . 312
155 Principe de notre traitement en 4 étapes pour définir les plages horaires. . . 314
156 Exemples de probabilités de tirer une nouvelle durée D' . . . 316
157 6 agendas reconstitués pour autant d'utilisateurs, sur une simulation donnée, à partir des données *Twitter* à Delhi et de notre couche d'utilisation du sol. 319
158 Résultat d'une CAH sur la matrice de dissimilarité des agendas reconstitués issus de *Twitter* à Delhi. . . . 320
159 Agendas moyens sur un mois pour les six différents groupes définis précédemment. . . . 321
160 Part d'utilisateurs selon le nombre d'activités effectuées hors du domicile (a) et part des activités effectuées (hors domicile et travail) (b). . . . 323
161 Nombre de lieux fréquentés en fonction du nombre d'activités effectuées (hors domicile et transport). . . . 323
162 Fréquence hebdomadaire de visite dans un lieu associé à une activité . . . 325
163 Pourcentage de fréquentation des différents lieux en fonction du jour de la semaine. . . . 326
164 Représentation des taux de transitions entre activités sous forme de matrice 327
165 Distribution des durées des activités selon le nombre d'activités quotidiennes, d'après les AR de *Twitter*. . . . 328
166 Probabilité horaire du premier départ du domicile. . . . 329
167 Répartition (sous forme de densité) des durées hors du domicile (en heure) en fonction des jours de la semaine. . . . 330

168 Nombre de lieux fréquentés en fonction du nombre d'activités, d'après nos AG. . . . 331
169 Comparaison des parts de fréquentations horaires par type de lieu sur une semaine. . . . 332
170 Comparaison des parts de fréquentations horaires par type de lieu les jours de semaines et le week-end . . . 332
171 Comparaison des distributions des durées entre les données issues des agendas reconstitués (rouge) et des agendas générés à partir de ces données (bleu ciel). . . . 333
172 Exemple d'agendas pour 10 agents. . . . 334
173 Nombre d'activités (autre que le domicile) par utilisateur (a) et pourcentages des activités effectuées par l'ensemble des utilisateurs . . . 335
174 Nombre de lieux fréquentés en fonction du nombre d'activités effectuées (hors domicile), d'après les données terrain. . . . 335
175 Probabilité de l'heure du premier départ du domicile. . . . 336
176 Probabilité d'effectuer une activité un jour donné. . . . 336
177 Distribution des fréquences de réalisation hebdomadaires selon le type d'activité effectuée. . . . 337
178 Comparaison de la part de fréquentations par activités par tranche horaire sur une semaine selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogrammes) . . . 337
179 Comparaison de la part de fréquentations par activités par tranche horaire sur un jour de semaine type, selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogrammes). . . . 338
180 Comparaison de la part de fréquentation par activités par tranche horaire sur un jour de week-end type, selon qu'il s'agisse de données issues des agendas reconstitués (traits noirs) ou de 5000 agents générés à partir desdits agendas (histogramme). . . . 339
181 Comparaison des distributions des durées dans un lieu visité entre les données issues des agendas reconstitués (rouge) et des agendas générés à partir de ces données (bleu ciel) . . . 340
182 Exemple d'agendas individuels pour 10 agents, d'après les données du terrain. 341
183 Résumé de la démarche permettant de reconstituer des agendas à partir de données épisodiques, puis d'en générer de nouveaux pour des agents. . . . 343
184 Nombre de commerces selon une grille de 500 m en 2009 (AIT) (a), et densité des *malls* à Bangkok. . . . 351
185 Localisation des zones commerciales dont nous allons parler dans la section. Fond de carte : Stamen. . . . 352
186 Photos du quartier de Yaowarat . . . 353
187 Quelques photos de marchés à Bangkok. . . . 353
188 Présence à Khao San road à différents moments de la journée . . . 354
189 Photos de quelques*malls* à Bangkok . . . 355
190 Quelques photos prises à l'intérieur de *malls*. . . . 356
191 Répartition de 12 catégories de POI dans Bangkok, selon une fonction de densité de portée 5km. . . . 358

192 Dendrogramme original (gauche), et dendrogrammes avec niveaux homogénéisés avec au maximum respectivement 1, 20 et 100 clusters par branches finales. 362
193 Typologie des activités économiques à Bangkok et dans le centre-ville. . . . 363
194 Lien entre le nombre de *POI* et la surface des *AOI* (a) et répartition du nombre d'*AOI* en fonction de leur superficie (b). 364
195 Répartition des différentes catégories de *POI* dans des *AOI*. 364
196 Part cumulée du nombre de zones (ou néo-AOI) en fonction de l'entropie . . 366
197 Rapport entre le nombre de lieux correspondants à des sorties sur le nombre de lieux de commerces de détail et leur entropie associée. 367
198 Part cumulée du nombre de néo-*AOI* en fonction de leur densité en *POI*. . . 367
199 Répartition des activités commerciales à Bangkok selon notre deuxième méthode. 368
200 Principe adopté pour définir une école à partir de la base de données *Here*. . 370
201 Utilisation du sol dans le centre de Bangkok. 372
202 Nombre de *check-in Facebook* par catégorie selon la classe de l'utilisation du sol. 373
203 Nombre de *check-in* par tranche horaire sur une semaine moyenne pour les 20 catégories les plus représentées sur *Facebook* à Bangkok 374
204 Profils horaires des lieux définis comme le domicile, sur *Facebook* et *Twitter* 375
205 Comparaison des parts des traces numériques enregistrées dans les différentes catégories de l'utilisation du sol, selon le réseau social (*Facebook* ou *Twitter*). 376
206 Répartition des traces numériques pour chacune des catégories de l'utilisation du sol par tranche horaire sur une semaine type. 377
207 Nombre de *tweets* moyens enregistrés par tranches horaires pour les 9 classes du *kmeans*. 380
208 Répartition des parts de *tweets* émis depuis une classe de la couche d'utilisation du sol par cluster (gauche) et nombre de tweets par maille par cluster du *kmeans* (droite). 381
209 Carte de l'utilisation du sol à Bangkok, d'après un *Kmeans* appliqué sur les profils temporels des *tweets* enregistrés dans des mailles de 180 m 381
210 Répartition des populations dans les clusters associés à des zones résidentielles (1, 5 et 8) selon une fonction de densité (intégrale égale à 1). 382
211 Temps de transport moyens (aller-retour) pour se rendre sur le lieu d'une activité (NSO, 2009) 388
212 Les modes de transports à Bangkok. 389
213 Le réseau de métro à Bangkok. 390
214 Nombre moyen de passagers quotidiens par ligne de métro à Bangkok 390
215 Quelques photos du métro aérien. 391
216 Part de la population de Bangkok selon la distance à une station de métro (a) ou de bus (b). 392
217 Photo d'un embouteillage dans le centre de Bangkok, vers 16h30 392
218 Condition du trafic à Bangkok, extrait de *Bing Traffic* pour la journée du 30 janvier 2018, toutes les 4 h entre 6 h et 2 h du matin. Le vert signifie un trafic fluide, le jaune moyennement congestionné et le rouge totalement congestionné. 393

219 Nombre de pixels de trafic dense par tranche horaire sur une semaine type à Bangkok. . . . 394
220 Densité du trafic à Bangkok, à 4 moments de la journée, les mardis et jeudis 395
221 Les temps de transports dans Bangkok un mardi à 7h du matin. . . . 397
222 Pourcentage de traces numériques laissées dans des mailles d'un kilomètre sur le réseau *Facebook* et *Twitter* (a), et différence entre ces niveaux d'activités (b). . . . 398
223 Application d'un Getis-Ord Gi* sur les données *Facebook* (haut) et *Twitter* (bas). Les Z-values supérieures à 25 apparaissent en noir. . . . 399
224 Densité des traces numériques géolocalisées (*tweets* et *check-in*) à Bangkok à différents moments de la journée et différents jours de la semaine. . . . 400
225 Variation de la "distance Venables" par tranches horaires sur une semaine type, selon les données issues de *Facebook* et de *Twitter*. . . . 402
226 variation de la "distance Venables" par tranches horaires sur une journée type, selon les données issues de *Facebook* et de *Twitter*. . . . 403
227 Variation de l'emprise spatiale selon les données *Facebook* et *Twitter*, avec différents seuils de concentration (gauche). À droite, figure un coefficient de dilatation selon les différents seuils. . . . 404
228 Différentes manières de concevoir un réseau à partir des lieux fréquentés . . 407
229 Exemple de lieux fréquentés par deux personnes. Certains de ces lieux sont visités conjointement. . . . 408
230 Principe de construction d'un réseau orienté sur les lieux fréquentés en commun, pour un utilisateur . . . 409
231 Sommes des flux entrant et sortant par sous-district selon les différentes matrices OD créées. . . . 410
232 Sommes des flux significatifs entrants et sortants (gauche) et rapport entre flux entrants et sortants (droite) par sous-districts, selon les trois méthodes de construction de réseaux . . . 412
233 Représentation des 100 flux les plus importants et des 2 flux principaux par sous-district, pour chacune des trois méthodes. . . . 413
234 Représentation des 100 flux les plus importants et des 2 flux principaux par sous-district, pour chacune des trois méthodes. Zoom sur le centre de Bangkok. 414
235 Exemple de détection de communauté dans un graphe. . . . 415
236 Regroupement des sous-districts en communautés, selon les différents réseaux utilisés en entrée et les trois méthodes de détection de communautés testées. . . . 418
237 Définition de zones fonctionnelles. . . . 419
238 Différence de pourcentage entre le nombre d'utilisateurs uniques par jour de la semaine pour chaque type d'activité, et le nombre de personnes travaillant dans ce type de lieu. . . . 425
239 Répartition de l'échantillon selon le type de lieu considéré comme étant l'activité principale. . . . 426
240 Représentativité spatiale de l'échantillon dont une activité principale a pu être estimée. N=17 368. . . . 427
241 Part de chaque groupe selon la distance du domicile à un lieu d'éducation. . 427
242 Principe de la démarche pour affecter une probabilité de réaliser une activité une semaine donnée. . . . 429
243 Principe de la définition et du choix des plages horaires. . . . 432

244 Illustration de la méthode employée pour tirer une heure de début et de fin d'activité (hors activité principale). 433
245 Illustration de la méthode employée pour tirer une heure de début et de fin pour l'activité principale. 434
246 Proportion des utilisateurs effectuant une activité par tranche horaire sur une semaine. D'après les agendas reconstitués. 436
247 fréquentations horaires des lieux de type commerce, sortie et mixte 437
248 Fréquentation horaire des autres types de lieux 437
249 Agendas reconstitués de 25 utilisateurs sur un mois 438
250 Agendas reconstitués de 25 utilisateurs pour une semaine. 438
251 Répartition des utilisateurs selon leurs agendas reconstitués à différents moments de la journée, un mardi et un samedi. 439
252 Répartition des utilisateurs selon leurs agendas reconstitués à différents moments de la journée, un mardi et un samedi, dans le centre de Bangkok. . 440
253 Distance "Venable" à Bangkok, d'après nos agendas reconstitués. 440
254 Décomposition de la distribution de la part horaire des superficies de la ville occupées par des zones de trafic très dense, selon l'algorithme EM. 441
255 Comparaison des heures de départ du domicile et de l'intensité du trafic le matin. 442
256 Comparaison des potentiels de dispersions selon les étudiants et le reste de l'échantillon 443
257 Lien entre le nombre de lieux fréquentés et le nombre d'activités effectuées (gauche) et entre le rayon de giration et le temps de trajet moyen. 444
258 Répartition de la part de chaque groupe obtenu précédemment dans les districts de la ville. L'encadré en bas à droite montre les distributions des effectifs selon les différents critères de dispersions, par groupe. 445
259 Part d'étudiants et des autres membres de l'échantillon, selon les groupes définis précédemment. 447
260 Relation entre le nombre d'activités réalisées et le nombre de lieux fréquentés, selon les différents groupes résultants de notre classification sur des paramètres de dispersions. 447
261 Répartition des activités dans les différents lieux fréquentés par l'échantillon. 448
262 Probabilité de réalisation d'une activité un nombre de semaines et de jours donné. 449
263 Répartition des pourcentages de fréquentation des activités selon les jours de la semaine, d'après les AR. 450
264 Distribution des durées quotidiennes des activités (autre que domicile et principale) selon le nombre de lieux visités quotidiennement, d'après les AR. 451
265 Probabilité de transition entre deux activités, à 15 h un lundi. 453
266 Comparaison entre le profil temporel des agendas reconstitués et les profils temporels obtenus par nos trois méthodes, par type d'activité. Sont représentés les R^2 ajustés (à gauche) et les RMSE (à droite). 454
267 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (1/3). Sont représentées ici les activités de type "Autre", "Principale" et "Domicile". 456
268 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (2/3). Sont représentées ici les activités "Commerce", "Sortie" et "Mixte" dense. . 456

269 Profils temporels des activités sur une semaine, selon les différentes méthodes de génération d'agendas (histogramme) au regard des AR (trait noir) (3/3). Sont représentées ici les activités "Parc","lieu de culte", "lieu d'éducation" et "école". 457
270 Profils temporels agrégés des activités des activités faiblement visitées (toutes sauf l'activité principale, le domicile, les lieux de types « autres » et les zones de commerces, de sorties et mixtes denses) sur une semaine.) . . . 458
271 Exemples d'agendas sur une semaine par groupe d'utilisateurs. 459
272 Distribution des fréquences de distances parcourues entre tous les principaux types de lieux. 463
273 Proposition de protocole de sélection d'un lieu de catégorie "Commerce", en fonction d'une localisation pour une activité de type "Autre" (bleu ciel) et de la localisation du lieu de domicile (bleu). 464
274 Probabilité de sélectionner un commerce dans la ville sachant que la personne se trouve dans un lieu de type « Autre » et connaissant la localisation de son domicile. 465
275 Distribution des temps de trajet entre une maille d'origine (domicile) et de destination (réalisation de l'activité principale). D'après l'algorithme exposé précédemment. 468
276 Estimation localisation des activités principales selon le sous-district de domicile (polygone noir). D'après un modèle d'attractivité, calculé ici en prenant en compte la répartition spatiale des *POI* et la durée des trajets. La population de chaque sous-district est ainsi répartie dans Bangkok selon la probabilité de réaliser l'activité principale dans une maille donnée. 469
277 "Ah ! Si seulement c'était si simple !", Bernard Schoenbaum, *New Yorker Cartoon*. 482

Liste des tableaux

1 Différents types de mobilités géographiques. 27
2 Synthèse des travaux sur les modélisations à base d'agents des mobilités humaines. 173
3 Informations contenues dans un *tweet* concernant le message 184
4 Informations contenues dans un *tweet* concernant la géolocalisation du message 185
5 Informations contenues dans un *tweet* concernant l'utilisateur 185
6 Informations contenues dans un *tweet* concernant la page de l'utilisateur . . 186
7 Évolution du nombre de *tweets* et d'utilisateurs selon les critères de sélections. 198
8 Liste des 10 lieux les plus fréquentés à Bangkok et le nom du lieu associé . . 200
9 Répartition des effectifs selon le niveau de richesse, défini à partir de critères de possession. 255
10 Approximation des moments de la journée en plage horaire 269
11 Durée minimale et maximale pour chaque activité. 274
12 Exemple de séquences d'activités pour deux individus. 280

Acronymes

ABM - Agent Based-Model (modèles à base d'agents)

AG - Agenda Généré (à partir des AR)

AOI - Area Of Interest

API - Application Programming Interface

AR - Agenda Reconstitué

BMA - Bangkok Municipal Autority (Thaïlande)

BTS - Base Transceiver Station (Téléphonie)

BTS - Bangkok Transit System (Transport)

CDR - Call Detail Record - Statistiques d'appels

GPS - Global Positionning System

KAP - Knowledge, Attitudes, and Practices

MRT - Mass Rapid Transit

NCT - National Capital Territory (Inde)

NIMR - National Institute of Malaria Research (Delhi)

NSO - National Statistic Office (Thaïlande)

OSM - OpenStreetMap

POI - Point Of Interest

SMA - Systèmes Multi-Agents

Annexes

Annexe A
Un modèle méta-population fermé

Nous proposons un ici modèle déterministe qui permet de prendre en compte pour chaque zone la population en journée, et la population durant la nuit.

1 Formulation

Sa formulation est tout d'abord classique (voir chapitre 3) :

$$\frac{dS}{dt} = \frac{-i\beta SI}{N} + \sum_{j=1}^{p} m_{ij}S_j - \sum_{j=1}^{p} m_{ji}S_i \tag{35}$$

$$\frac{dI}{dt} = \frac{i\beta SI}{N} - i\gamma I + \sum_{j=1}^{p} m_{ij}I_j - \sum_{j=1}^{p} m_{ji}I_i \tag{36}$$

$$\frac{dR}{dt} = i\gamma I + \sum_{j=1}^{p} m_{ij}R_j - \sum_{j=1}^{p} m_{ji}R_i \tag{37}$$

mais lors des itérations impaires, décrivant les flux de retour, la matrice de flux devient alors :

$$m^r = \frac{(mN)^T}{\sum_{i=1}^{p} S_i^1 + I_i^1 + R_i^1} \tag{38}$$

Avec $\sum_{i=1}^{p} S_i^1 + I_i^1 + R_i^1$ Représentant la population des différentes zones lors des itérations impaires (typiquement la population en journée) et la transposée de la matrice des flux multipliée par la population, $(mN)^T$ L'application de ces équations dans un système à 4 patchs est illustré par la figure i ci-dessous. La figure i, haut, montre l'oscillation du système entre les allers (figure i, centre) et les retours (figure i, bas). L'intérêt d'un tel modèle est la conservation de la population lors de deux périodes de la journée.

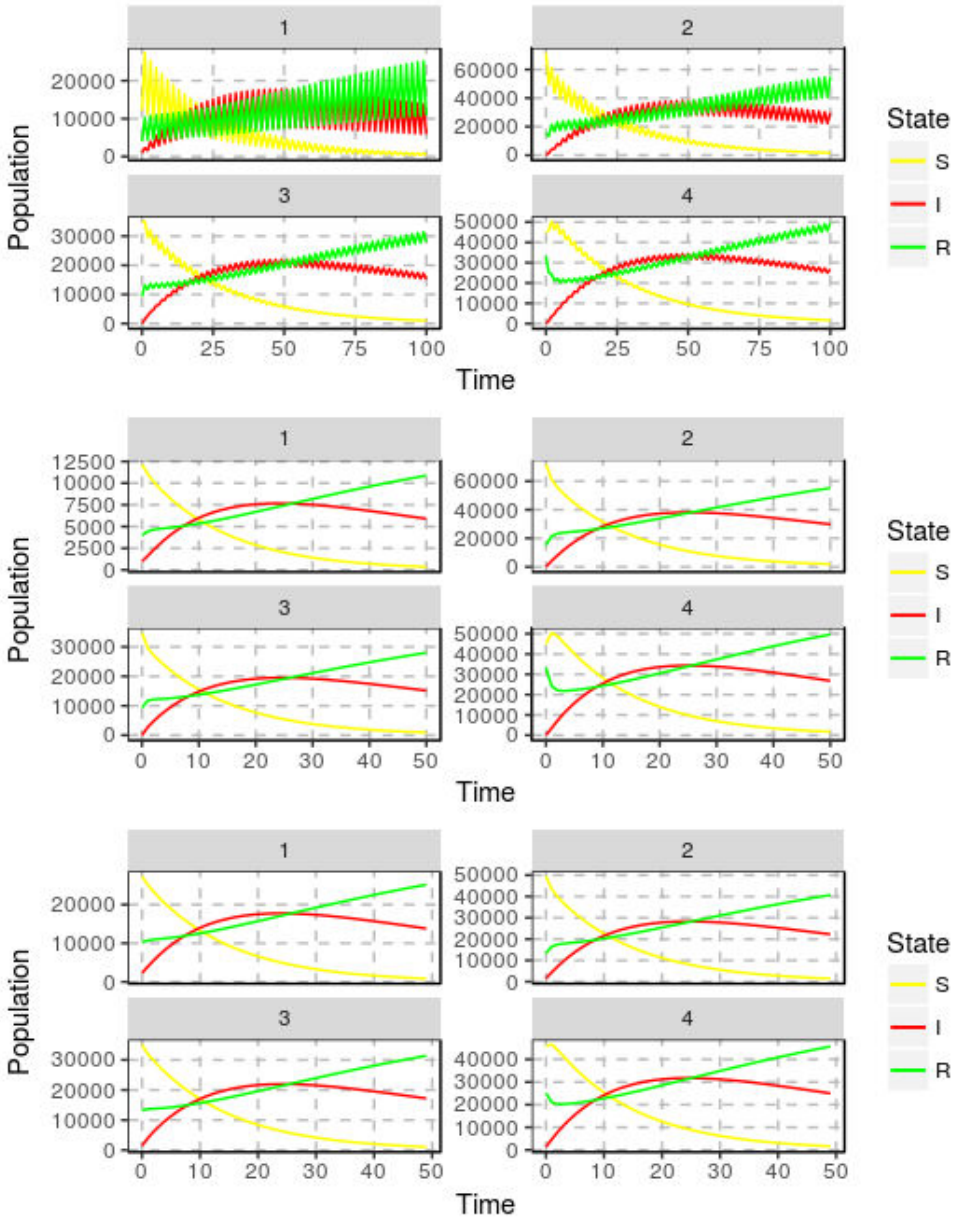


Figure i Évolution des états sérologiques dans le modèle méta-population fermé à 4 patchs. La figure du haut montre les résultats globaux. Les oscillations caractérisent l'aspect pendulaire des déplacements. La figure du milieu montre le nombre de personnes dans chaque état après les flux allé, tandis que la figure du en bas montre les états après les flux retours. Dans chaque patch, à chaque pas de temps, la somme des individus S,I ou R est constante et vaut la population en journée lors des itérations impaires (centre), et la population au domicile lors des itérations impaires (bas).

1.1 Sensibilité du modèle théorique dans un système à 4 zones

Nous allons maintenant étudier la sensibilité du modèle aux paramètres de mobilité. Nous définissons les populations initiales dans chacune des zones selon une loi normale avec une population moyenne de 60 000 habitants et un écart type de 20 000 habitants. Nous posons

que la part de personnes immunisées (R) suit elle aussi une loi normale, avec une moyenne à 25 % et un écart type de 10 %. Ces paramètres démographiques sont fixes au cours des différentes simulations, tout comme la durée de la maladie $1/\gamma$ (ici 1/5) et la force d'infection β (à 1.5).

Nous allons tester ensuite différentes matrices origine-destination générées selon des distributions exponentielles, normales ou aléatoires. Nous ferons ensuite varier la part de personnes se déplaçant hors de sa zone, et nous testerons les réactions du modèle en initialisant 1000 personnes infectées dans différentes zones. Nous ne représenterons graphiquement que les états sérologiques correspondants aux flux de retour au domicile.

1.2 Initialisation d'une matrice d'origine-destination

Dans un premier temps, nous générons trois matrices origine-destination (figure ii). Nous définissons la première matrice m_1 comme composée de flux entre les différentes zones suivant une loi exponentielle et répartis de manière décroissante (la zone 1 accueille le plus grand nombre de flux, et inversement pour la zone 4) (figure ii, gauche). Les flux de la matrice m_2 suivent une distribution normale, centrés sur les zones 2 et 3 (figure ii, centre). Enfin, la matrice m_3 suit une distribution aléatoire (figure ii, droite).

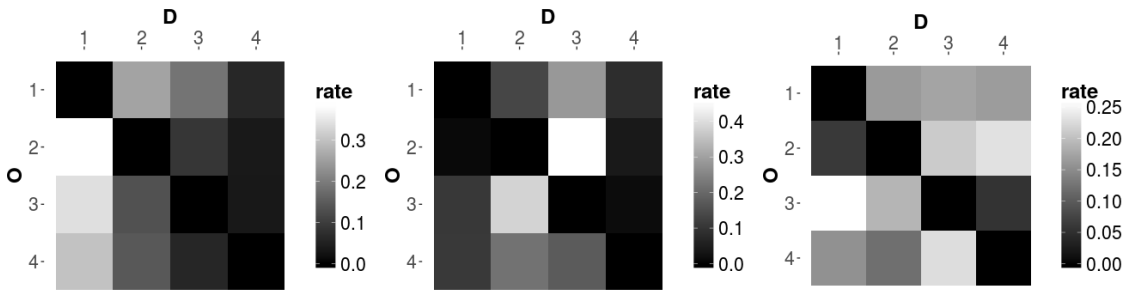


Figure ii Matrices de flux utilisées pour l'exemple du modèle méta-population fermé. À droite, les flux sortants sont décroissants de la zone 1 vers la zone 4. Au centre les flux sont surtout importants entre les zones 2 et 3. À droite les flux sont répartis de manière plus aléatoire.

1.3 Initialisation des paramètres de mobilité et des personnes infectées

Pour chacune des matrices ci-dessus, nous allons faire varier le nombre de personnes sortant de leur zone (10 %, 50 % 90 %). Nous allons également faire varier le nombre de personnes infectées par zones, en initialisant tout d'abord 1000 personnes infectées en zone 1,

puis en zone 4.

Notations

Pour des raisons de lisibilité, nous utiliserons un système de notation suivant une série de 3 lettres pouvant prendre pour valeur A, B ou C. Les résultats correspondant par exemple à la matrice m_1 (A) avec 10 % de personnes mobiles (A) et une initialisation des personnes infectées en zone 1 (A), seront nommée AAA. Les résultats pour la matrice m_1 (A) avec 50 % de mobilité (B) et une initialisation des personnes infectées en zone 4 (B) seront nommés ABB. De la même manière, dans le cas de la m_3 avec 90 % de mobilité et 1000 personnes infectées en zone 1 sera CCA. Suivant la même logique, l'expérience BAB correspond à l'utilisation de la matrice m_2, avec 10 % de personnes mobiles et 1000 personnes infectées en zone 4.

1.4 Résultats

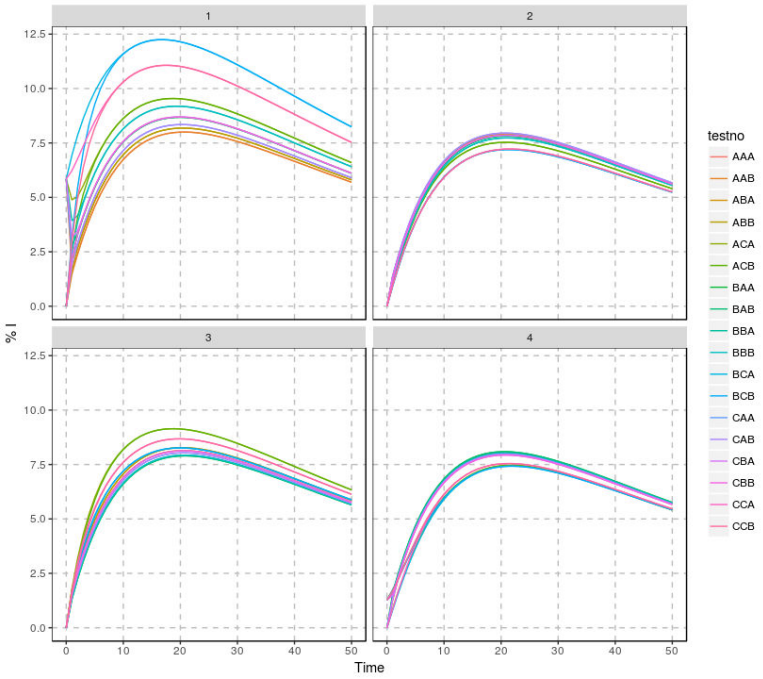


Figure iii Résultat des différents tests montrant l'évolution de la part du nombre de personnes infectées au cours des itérations.

La figure iii montre les résultats des différents tests selon le protocole expliqué précédemment. Nous pouvons déjà remarquer le niveau de similarité des profils des personnes infectées. La figure iv ci-dessous représente de manière plus lisible les résultats d'une manière peut être plus lisible, en regroupant les patchs en fonction des différents scénarios :

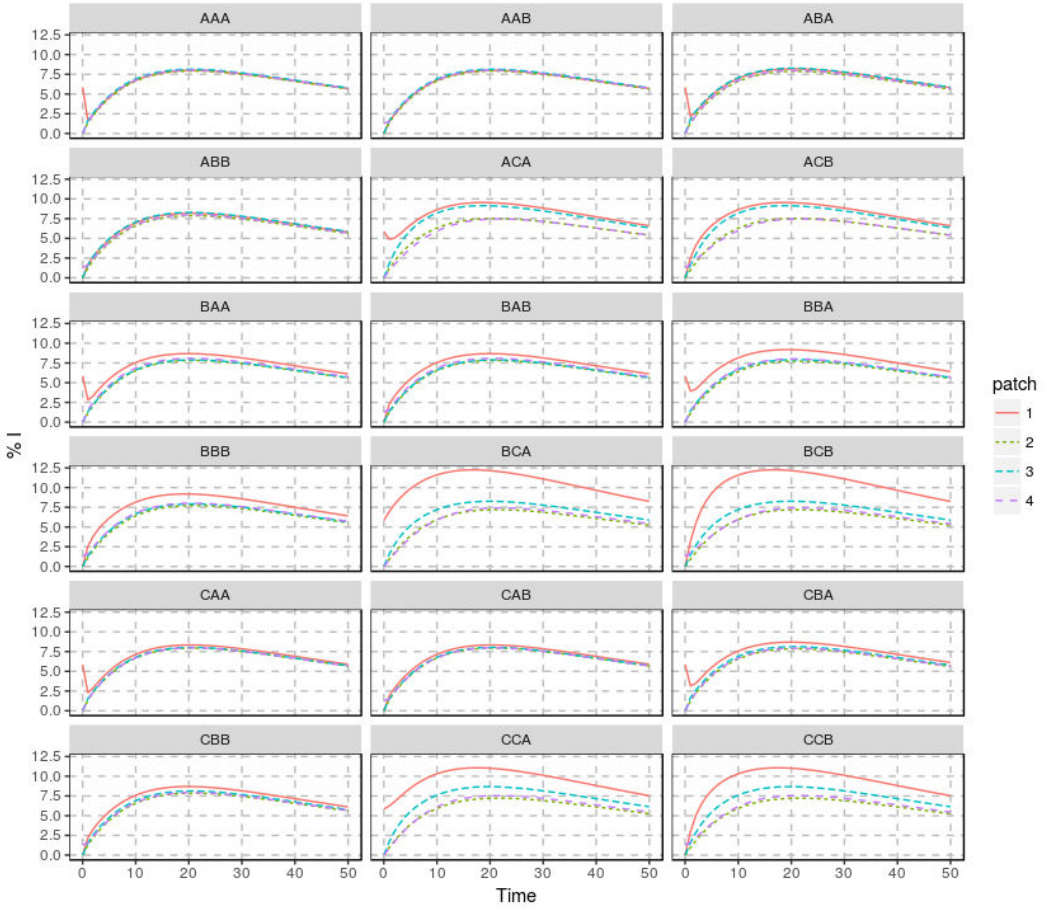


Figure iv Résultat des différents tests montrant l'évolution de la part du nombre de personnes infectées au cours des itérations, par expérimentation.

La figure iv montre que le paramètre qui va le plus influencer les résultats est surtout la part de personne qui sort de la zone, surtout lorsque cette dernière est très élevée (xCx). Vient ensuite le type de matrice OD utilisé. Nous pouvons constater dans les résultats obtenus à partir de la matrice 1, où les flux entrants sont dirigés vers une même zone (Axx), que l'épidémie se déroule de manière relativement similaire dans les autres zones. Ceci peut s'expliquer par le fait que les populations se croisent en journée au même endroit, ce qui entraîne une évolution relativement commune de la part des personnes infectées. *A contrario*, dans les résultats de la matrice 2 (Bxx),nous observons de plus grandes différences dans l'évolution de la part de personnes infectées dans les différentes zones. Ceci peut provenir de la présence de 2 pôles d'attraction (les patchs 2 et 3), ce qui complexifie l'évolution de l'épidémie.

Finalement, la zone où l'infection a été initialisée n'a ici qu'un rôle marginal dans le déroulement futur de l'épidémie. Ce paramètre n'a un impact visible que lorsqu'il est combiné avec une part importante de personnes sortant de la zone (xCx). Ceci peut s'expliquer par le rôle des mobilités et des interactions entre les patchs, et de l'aspect déterministe du modèle.

1.5 Exemple avec une modulation du taux d'infection

Nous allons maintenant modifier le modèle en prenant en compte un paramètre extérieur global pour moduler le taux d'infection. Nous allons ici poser que la force d'infection dépend d'un régime de précipitation théorique, ce qui peut simuler un nombre plus important de moustiques lorsque les précipitations sont élevées.

Nous posons que β = (a +1) / (b*$P(t)$) où P(t) est la valeur des précipitations à l'instant t, a et b étant fixes et permettant de décrire cette évolution au cours du temps.

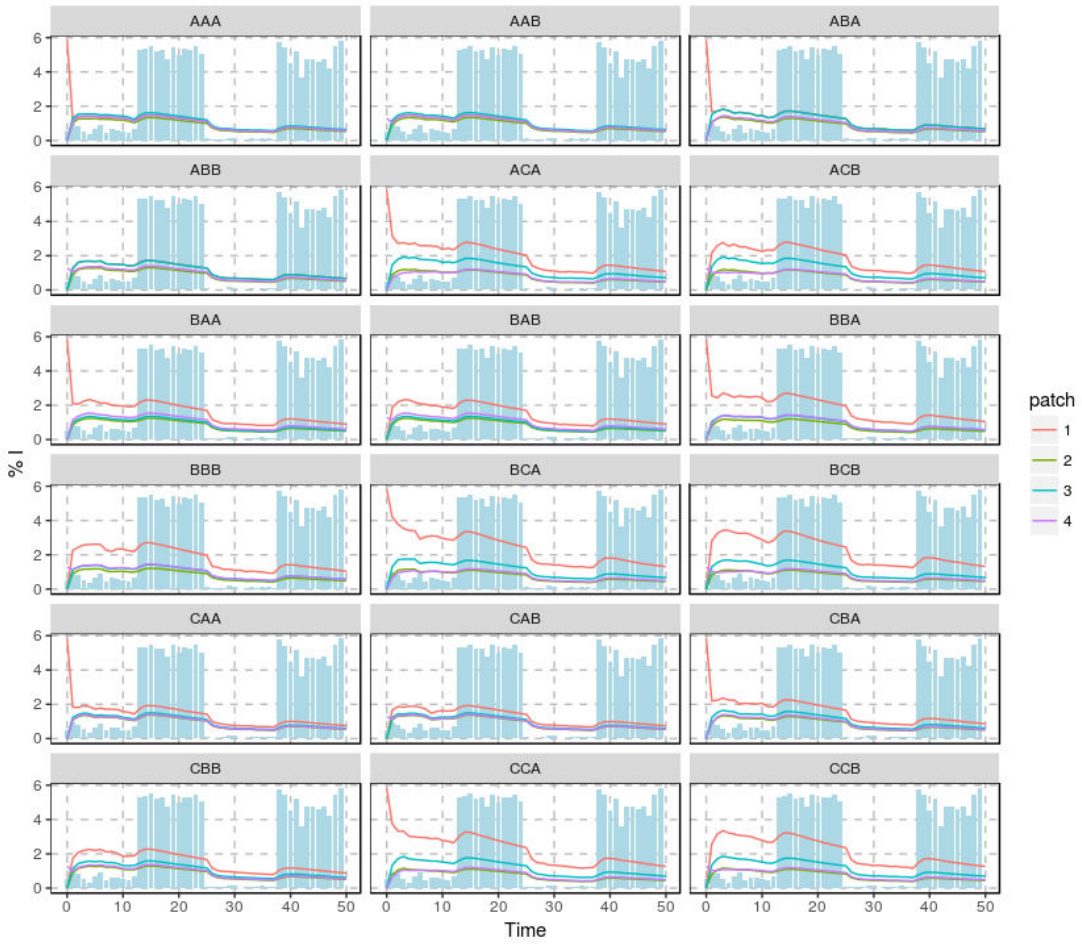


Figure v Résultats des différents scénarios lorsque la force d'infection est modulée par un paramètre extérieur global.

La figure v met bien en évidence la saisonnalité des contaminations. Elle montre également que dans des scénarios où les déplacements ne sont pas centrés vers une zone (matrice 2 et 3), plus de différences sont observées dans le pourcentage de personnes infectées dans les différentes zones. En effet, contrairement à la figure iv où il fallait un nombre important de personnes sortant du quartier pour constater des différences visibles dans le nombre de personnes

infectées, ce gradient s'observe de manière plus nette en Bxx et Cxx.

1.6 Exemple avec prise en compte du vecteur

Nous allons maintenant prendre en compte le vecteur, en nous inspirant des modèles développés par (Bailey, 1975 ; Derouich *et al.*,2003). Nous considérons ici que la population d'hôtes est stable (pas de décès, ni de naissance), et contrairement à Derouich *et al.*(2003), nous n'ajoutons pas de paramètre décrivant une probabilité d'immunité permanente.

Les paramètres C_{vh} et C_{hv} décrivent les interactions entre le vecteur et l'hôte, et inversement, μ_v est la durée de vie du moustique. Le modèle se décrit ainsi :

Hôtes :

$$\frac{dS_h}{dt} = \frac{-C_{vh}S_h}{N_h} \tag{39}$$

$$\frac{dI_h}{dt} = \frac{-C_{vh}S_h}{N_h} - (i\gamma_h + i\mu_h)I_h \tag{40}$$

$$\frac{dR_h}{dt} = i\gamma_h I_h - i\mu_h R_h \tag{41}$$

Vecteurs :

$$\frac{dS_v}{dt} = i\mu_v N_v - (i\mu_v + \frac{C_{hv}I_h}{N_h})S_v \tag{42}$$

$$\frac{dI_v}{dt} = \frac{C_{vh}I_h}{N_h}S_v - i\mu_v I_v \tag{43}$$

Nous posons N_v, la population de moustiques comme étant 2.5 fois N_h, un μ_v de 1/8, un C_{hv}=C_{vh} à 0.33, et nous définissons la zone 3 comme ayant 2.5 % de moustiques infectés. Pour une étude de la sensibilité du modèle à ces différents paramètres (sans mobilité) voir (Bakach, 2015 ; Derouich *et al.*, 2003)

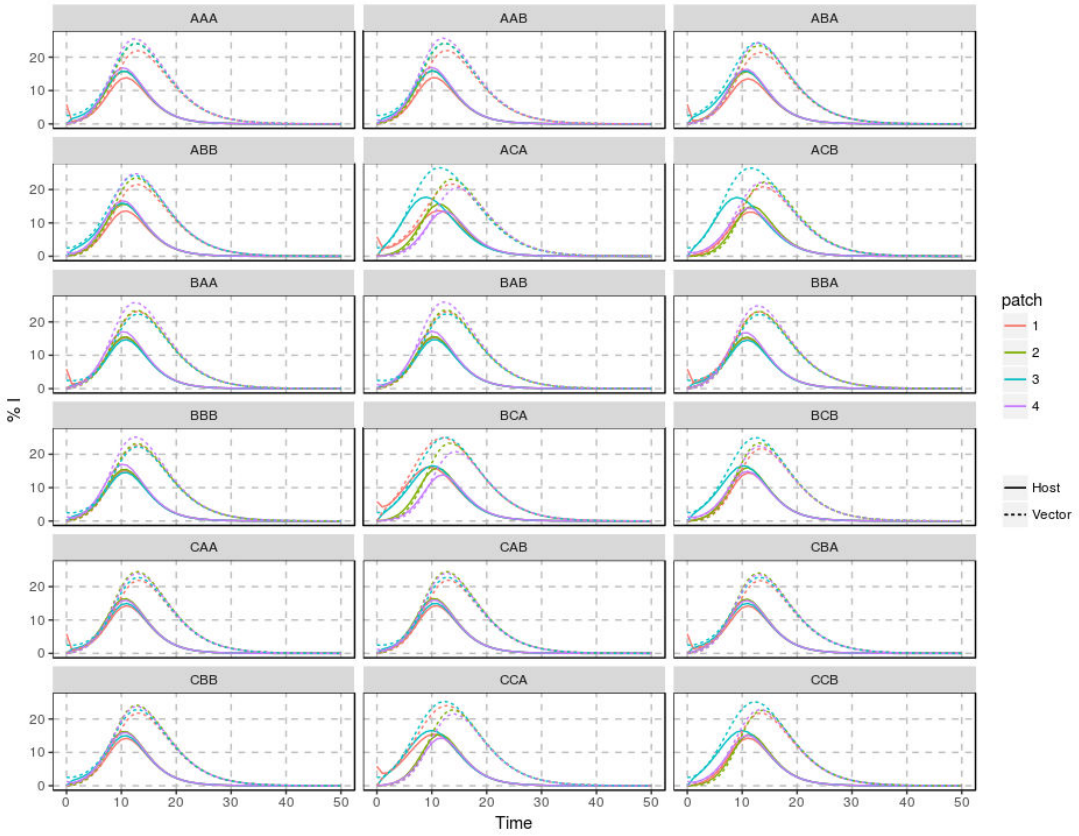


Figure vi Évolution de la part d'hôtes (traits pleins) et de vecteurs (pointillés) au cours du temps, selon les différentes configurations et les zones.

La figure vi montre qu'une fois encore, le paramètre qui semble avoir le rôle principal dans la propagation de l'épidémie est la part de personnes qui sortent de leur zone (xCx). Mais le niveau de différenciation de la part du nombre de personnes infectées est relativement faible.

Discussion sur le modèle méta-population fermée

Ce modèle méta-population fermé présente l'avantage de pouvoir décrire les états sérologiques en fonction des flux allés et retours, avec une conservation de la population. Il s'agit d'un modèle théorique SIR, dont le niveau de complexité est ajustable, notamment en le transformant en SEIR ou SEIRS, ou en influençant la force d'infection β, comme c'est le cas dans la plupart des modèles déterministes. Il ne prend cependant en compte qu'un seul sérotype.

Cette étude a montré que le modèle est plus sensible à la part de personnes sortant d'une zone, qu'au nombre de personnes initialement infectées. Le type de matrice utilisée (centrée sur une zone, sur 2 zones ou aléatoire) joue également un rôle dans la propagation. Ce rôle est plutôt secondaire lorsque l'on considère un β fixe, tandis qu'il s'accentue lorsque β est modulé

par un facteur extérieur, par exemple les précipitations. Lorsque l'on inclut le moustique selon un modèle simple, la zone où les moustiques contaminés sont injectés joue un rôle important dans l'évolution du modèle.

Comme montré dans le chapitre 3, les modèles SIR et SEIR sont extrêmement sensibles aux paramètres γ et β. Ainsi, avec d'autres paramètres, qu'ils soient définis et fixes ou modulés (par des précipitations ou des vecteurs) le modèle entraînerait des résultats très différents.

Parmi les 3 cas de modulation de β, qu'il soit fixé, dépendant d'un facteur externe comme les précipitations, ou interne et modulé par le compartiment des moustiques, l'expérience où la force d'infection est dépendante des précipitations globales donne des courbes épidémiques plutôt réaliste. Il serait intéressant d'approfondir lorsque les précipitations et la température locales influent sur la population vectorielle et donc sur les hôtes, comme effectué par (Lourenço and Recker, 2014). Finalement, ce modèle n'a été décrit que graphiquement et ces propriétés mathématiques restent encore à explorer.

Annexe B
Analyses de nos traces numériques

Nous avons beaucoup discuté des données personnelles dans ce travail, de leur création, en général involontaire et que les personnes concernées ne soupçonnent pas une telle utilisation dans la recherche sur les mobilités. La partie B aborde largement le traitement de données personnelles obtenues sur le réseau social *Twitter* à Bangkok et à Delhi, et également sur *Facebook*. Nous estimons ici que l'étude de nos propres données personnelles peut permettre une meilleure compréhension du sujet, notamment des différents biais et potentiels de traçage. L'objectif est ici est donc de faire une mise en abyme en testant différentes sources de données nous concernant et susceptibles de définir notre espace d'activité, à l'échelle locale et nationale. Il n'y a ici aucune volonté exhibitionniste, mais plutôt une approche (toute proportion gardée) à la Stubbins Ffith. Pour ce faire, nous allons dans un premier temps répondre à des questions classiques, relatives aux lieux fréquentés. Les résultats serviront de base permettant de comparer avec des données issues (1) des transactions bancaires, (2) des billets de train achetés en ligne et (3) des adresses IP enregistrées par le réseau social *Facebook*. L'étude porte sur une période allant du 1er mai au 23 juillet 2017 [416]. Nous ajouterons ensuite les métadonnées issues de nos communications téléphoniques, afin d'étudier l'évolution des interactions en fonction des lieux fréquentés. Nous partirons du global (échelle nationale) pour aller vers le local.

Présentation des données

Auto-Interview

Lieux fréquentés en France

La première partie de l'interview consistait à énumérer les différents lieux fréquentés en France (date et une durée, tableau i).

416. Le début de l'étude commence au 1er mai, car il s'agit des données les plus anciennes que nous avons pu récupérer sur notre compte *Facebook*.

Lieux	Date / Durée
Bordeaux	∼1/2 du temps | Dates sûres : - 5 → 11 mai & 18,21,23 juillet
Côtes de Bourg	∼1/6 du temps | Dates sûres : 18 juin & 17 → 20 juillet
Région Parisienne	∼1/3 du temps | fin mai → 8 juin & fin juin → début juillet
Le Havre	1 jours 8 juin
Rouen	1 jours 9 juin
Meilleray (77)	1 jours 1er juillet
Angers	3 jours 5-8 juillet
La Flèche	2 jours 8-10 juillet
Avignon	2 jours 21-23 juillet
Nîmes	4 heures -21 juillet

Tableau i Villes déclarées comme fréquentées entre le 1er mai et le 23 juillet 2017.

Nous pouvons remarquer que certaines dates et estimations de durée sont assez précises, notamment lorsqu'il s'agit de lieux fréquentés ponctuellement. En effet, certaines dates sont plus faciles à se rappeler (dates des anniversaires, des conférences et rendez-vous professionnels, ou de festivals d'art de la rue et de théâtre). Nous pouvons supposer que ce genre d'événements qui requièrent de l'organisation sont plus facilement mémorisables dans le temps. La ville de Nîmes est également citée, dû au retard d'un train de plus de 4h. En revanche, il est délicat de définir des temporalités exactes pour ce qui est des visites en région parisienne, à Bordeaux ou dans les Côtes de Bourg, puisque nous sommes partagés entre ces lieux depuis plusieurs mois, sans réelle régularité.

Les déplacements en région parisienne sont notamment motivés par la nécessité de nous rendre à l'Université de Rouen depuis Bordeaux, ce qui implique une nuit dans la capitale. Au-delà de Paris comme un hub vers la Normandie, la présence d'amis et d'événements considérés comme suffisamment importants sont des facteurs motivant les déplacements.

En somme, pour reprendre une analogie des modèles d'interactions que nous évoquons dans différentes sections de cette thèse, les déplacements vers Paris semblent dépendre (1) de raisons professionnelles impérieuses (2) du niveau d'attractivité sociale à l'instant t (3) du temps disponible. La notion (2) peut s'estimer inconsciemment en prenant en compte la distance qu'un ami a parcourue pour venir à Paris (effort), le temps écoulé depuis la dernière rencontre, la nature exceptionnelle d'un événement, la nécessité de socialisation après une longue période d'isolement, et bien entendu par les liens d'attachements vis-à-vis des individus présents à Paris. Nous pouvons noter que les spécificités de Paris (bassin d'emploi, tourisme, architecture et urbanisme, offre culturelle, mode de vie particulier, etc.) n'interviennent qu'indirectement dans l'attractivité de la ville, car exercées principalement sur les relations sociales vivant dans la capitale, et ce sont ces dernières qui ont un pouvoir attractif sur nous. Cela dit, si Paris n'était pas si joli, peut-être que nous y viendrons moins souvent.

Lieux fréquentés à Bordeaux et Paris

Pour ce qui est de l'étude des déplacements locaux, la défaillance de la mémoire ne permet pas de retrouver précisément les jours et horaires de fréquentations. Nous noterons donc une liste de lieux fréquentés, associées à une fréquence de visite, ainsi qu'à une heure de la journée.

Région bordelaise

À Bordeaux, le nombre de lieux fréquentés entre le 1 mai et le 23 juillet paraît assez restreint. Nous sommes principalement restés à notre domicile, et les sorties principales se cantonnent alors à faire des courses au supermarché de proximité, environ 4 fois par semaine, à des horaires variables, entre 11 h et 21 h. Durant cette période, nous sommes allés 2 ou 3 fois à l'hypermarché du centre ville, en prenant le tramway, pour une durée d'environ 1h-1h30, mais sans nous rappeler de la date exacte. De même nous avons dû aller au tabac du coin de la rue environ 3 ou 4 fois, et au moins une fois à la librairie. Pour ce qui est des sorties, ces dernières furent plutôt rares. Ces sorties furent localisées dans le quartier de Saint Michel (au moins 3 fois) et au bord de la Garonne (3 fois) et une fois vers Gambetta. Nous sommes passés par la gare au moins 6 ou 8 fois afin de prendre le train Nous avons aussi dû nous rendre 4 ou 5 fois à la gare de bus de Lormont Buttinière, au nord-est de la ville. Une fois arrivé dans les côtes de Bourg, les lieux fréquentés se cantonnent à l'arrêt de bus et au domicile. Ces informations sont consignées dans le tableau ii et spatialisées dans la carte vii.

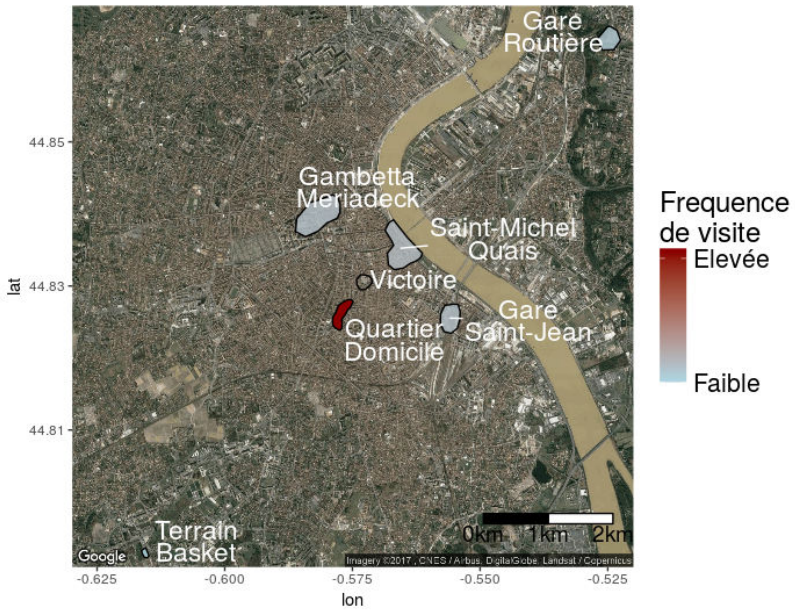


Figure vii Localisation des lieux déclarés être fréquentés à Bordeaux.

Lieux	Fréquence	Heure de la journée	Quartier
Domicile	∼100 %	-	Quartier Domicile
Supermarché de proximité	4-5 fois par semaine	variable	Quartier Domicile
Hypermarché	2 ou 3 fois	Entre 11h et 17h	Gambetta - Meriadeck
Tabac	3 ou 4 fois	variable	Quartier Domicile
Librairie	2 fois	Aprés midi	Gambetta - Meriadeck
Gare SNCF	6-8 fois	variable	Gare SNCF
Gare Routière	4-5 fois	variable	Gare Routrière
Quartier Saint Michel	4-5 fois	Soir	Saint Michel - Quais
Quai de Garonne	2 fois	Début de soirée	Saint Michel - Quais
Bar Gambetta	1 fois	Début de soirée	Gambetta - Meriadeck
Place de la victoire	Lors de chaque déplacement au-delà du quartier	Variable	Victoire
Fac de Bordeaux	1 fois	Midi	Université

Tableau ii Lieux déclarés être fréquentés à Bordeaux.

Paris

Durant les séjours à Paris, nous avons été hébergés à Gentilly et à Boulgone-Billancourt. Comme à Bordeaux, nous travaillons à la maison en journée, et allons 3 ou 4 fois par semaine au supermarché de proximité. Le soir, les sorties furent plus nombreuses, principalement prés de la Fac de Paris-7, la Butte aux Cailles, sur les quais de Seine derrière Notre Dame, vers le Panthéon, vers Science Po, vers Châtelet, dans le quartier de Strasbourg Saint-Denis, au canal Saint-Martin, et une fois vers Ménilmontant. Nous avons également été conviés à une réunion avec des collègues à l'Institut Pasteur un vendredi matin (tableau iii et carte viii).

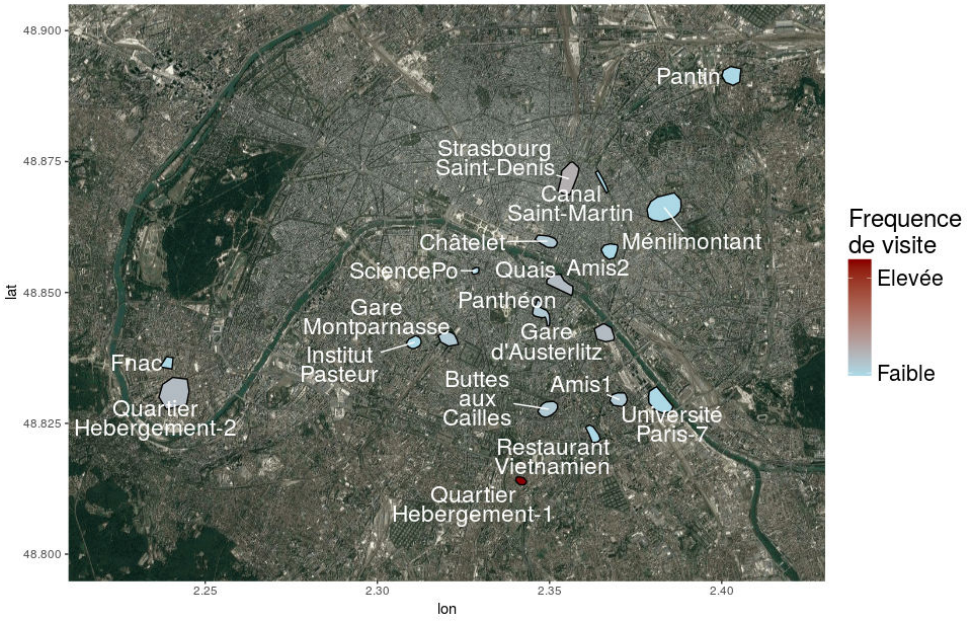


Figure viii Localisation des lieux déclarés être fréquentés à Paris.

Lieux	**Fréquence**	**Heure de la journée**
Hebergement Gentilly	∼85 %	Journée
Hebergement Boulogne	∼15 %	Journée
Supermarché de proximité	3/4 fois par semaines	Variable
Gare SNCF	6-8 fois	Variable
Paris-7	1 fois	Début de soirée
Science Po	1 fois	Début de soirée
Panthéon	2 fois	Début de Soirée / Soirée
Quai de Seine	3 fois	Début de Soirée / Soirée
Châtelet	2 fois	Début de Soirée / Soirée
Ménilmontant	1 fois	Début de Soirée / Soirée
Strasbourg-Saint Denis	4 fois	Début de Soirée / Soirée
Buttes aux Cailles	2 fois	Début de Soirée / Soirée
Fnac - Boulogne	1 fois	Aprés midi
Amis 13eme	2 fois	Début de Soirée / Soirée
Restaurant 13eme	1 fois	Midi
Concert Pantin	1 fois	Soirée
Canal Saint Martin	2 fois	Soirée
Institut Pasteur	1 fois	Matin

Tableau iii Lieux déclarés être fréquentés à Paris.

Cette rapide étude sur les lieux fréquentés montre déjà plusieurs choses. Tout d'abord, les activités en journée sont relativement similaires (domicile supermarché), indépendamment de la ville. Mais malgré le fait que plus de temps soit passé à Bordeaux, le nombre de lieux fréquentés est nettement moindre qu'à Paris. Ceci est dû principalement au nombre de relations

sociales très différents entre les villes, car ces dernières conditionnent grandement la planification des venues dans la capitale. Une volonté d'optimisation et de maximisation des rencontres va alors de pair avec une bonne organisation. De même, les lieux fréquentés dans les villes dépendent surtout des compromis géographiques liés à l'accessibilité et aux habitudes passées. Nous pouvons également noter que nous ne faisons pas de référence à une éventuelle dichotomie entre les jours de la semaine et de week-end. Finalement, il est assez délicat de dire précisément quel jour un lieu fut fréquenté et à quelle heure. Il faut également prendre en compte le contexte bien particulier de la période d'étude qui couvre la rédaction d'une thèse.

Données de Voyages-Sncf

Les données issues d'achats de billets de train en ligne sur le site de voyages-sncf sont obtenues simplement en regardant notre messagerie électronique. Elles indiquent les dates des différents trajets et serviront à combler les approximations dans les dates de départ. Même si tous les trajets ne furent pas effectués en train.

Données Ip Facebook

Depuis le combat juridique d'un jeune autrichien commencé en 2011 envers la firme américaine [417], couplé aux obligations vis-à-vis de la CNIL, il est maintenant possible d'avoir accès à nos données personnelles laissées sur le réseau social *Facebook* [418]. Ces données contiennent énormément d'informations, tant sur l'historique des liens sociaux (ajout de contacts, conversations, etc.) que sur l'historique des connexions (horodatage des connexions, associées à une adresse IP). Nous nous focaliserons ici sur ce dernier aspect, en récupérant les adresses IP enregistrées par le service et en utilisant un annuaire inversé pour estimer la localisation [419]. Un paramètre important est que n'ayant de *smartphone*, nous ne nous connectons à *Facebook* uniquement depuis notre ordinateur portable, ce qui implique que les adresses associées ne sont révélatrices que des lieux nous avions accès à la fois à un ordinateur et une connexion wi-fi sécurisée. Cela dit, lorsqu'une personne utilise *Facebook* avec la 3 g, les plages d'adresses IP qui lui sont fournies ne donnent pas d'information sur la localisation (car associé au siège social de l'opérateur).

Données transactions bancaires

Les transactions bancaires sont obtenues en téléchargeant notre historique de compte, sur lequel figure la date de l'opération ainsi que la raison sociale. Les localisations ont été estimées à partir de cette dernière information, tout comme le type d'établissement fréquenté :

417. http://www. emonde.fr/p xe s/art c e/2014/08/07/max m an schrems e but est de fa re respecter a facebook a eg s at on europeenne 4468090 4408996.htm
418. https://www.facebook.com/he p/contact/180237885820953
419. https://db p.com/ & https://www. p ocat on.net/

Alimentation (supermarché, épicerie), lieux de sorties (restaurant, café, bar), DAB (distributeur automatique de billets), ou transport (gare SNCF ou borne d'achat de ticket de transport). À noter qu'un retrait dans un DAB mentionne l'heure et la minute, ce qui n'est pas le cas lors d'un achat par carte bancaire, ou uniquement le jour apparaît.

Données téléphonies

Les données de téléphonie sont issues de notre historique d'appel présent sur les factures fournies par notre opérateur, que nous utiliserons uniquement comme proxy des interactions sociales (appels émis ou reçus) en fonction des lieux fréquentés.

Google map timeline

L'entreprise Google met à disposition le service « timeline », qui permet après activation du service de géolocalisation de retracer les historiques de déplacements[420], et les données collectées ne sont visibles que par l'utilisateur et par Google. Cet outil aurait pu être très utile pour notre étude, et fournirait sans aucun doute les meilleurs résultats[421], mais nous n'avons malheureusement pas de GPS, et encore moins de téléphone capable d'installer de telles fonctionnalités.

Comparaison des résultats à l'échelle nationale

La figure ix montre les différents lieux visités dans le temps, en fonction des différentes données. Nous avons ajouté une notion d'incertitude, qui concerne les données issues des billets de train et de l'interview. Nous pouvons déjà remarquer qu'il y a une très bonne concordance entre les résultats de l'interview et des données bancaires. À noter que nous avions oublié que nous avions effectué un séjour à La Teste. Cette omission est comblée par les données bancaires, rappelant que nous avons acheté un billet à la gare et fait quelques courses dans un hypermarché de la ville. Cependant, les déplacements dans les Côtes de Bourg ne furent pas remarqués par notre banque, étant donné que nous ne faisons pas de courses ni ne retirons d'argent lorsque nous sommes dans cette zone.

420. https://www.goog e.com/maps/t me ne?pb

421. Pour l'aide de Google : https://support.goog e.com/maps/answer/6258979?co=GENIE.P atform%3DDesktop&h =fr, pour un exemple d'application : https://www.youtube.com/watch?v=N 0 Y O 1Y0 et et pour des informations sur le paramétrage du compte : https://www.cnet.com/how to/how to de ete and d sab e your goog e ocat on h story/

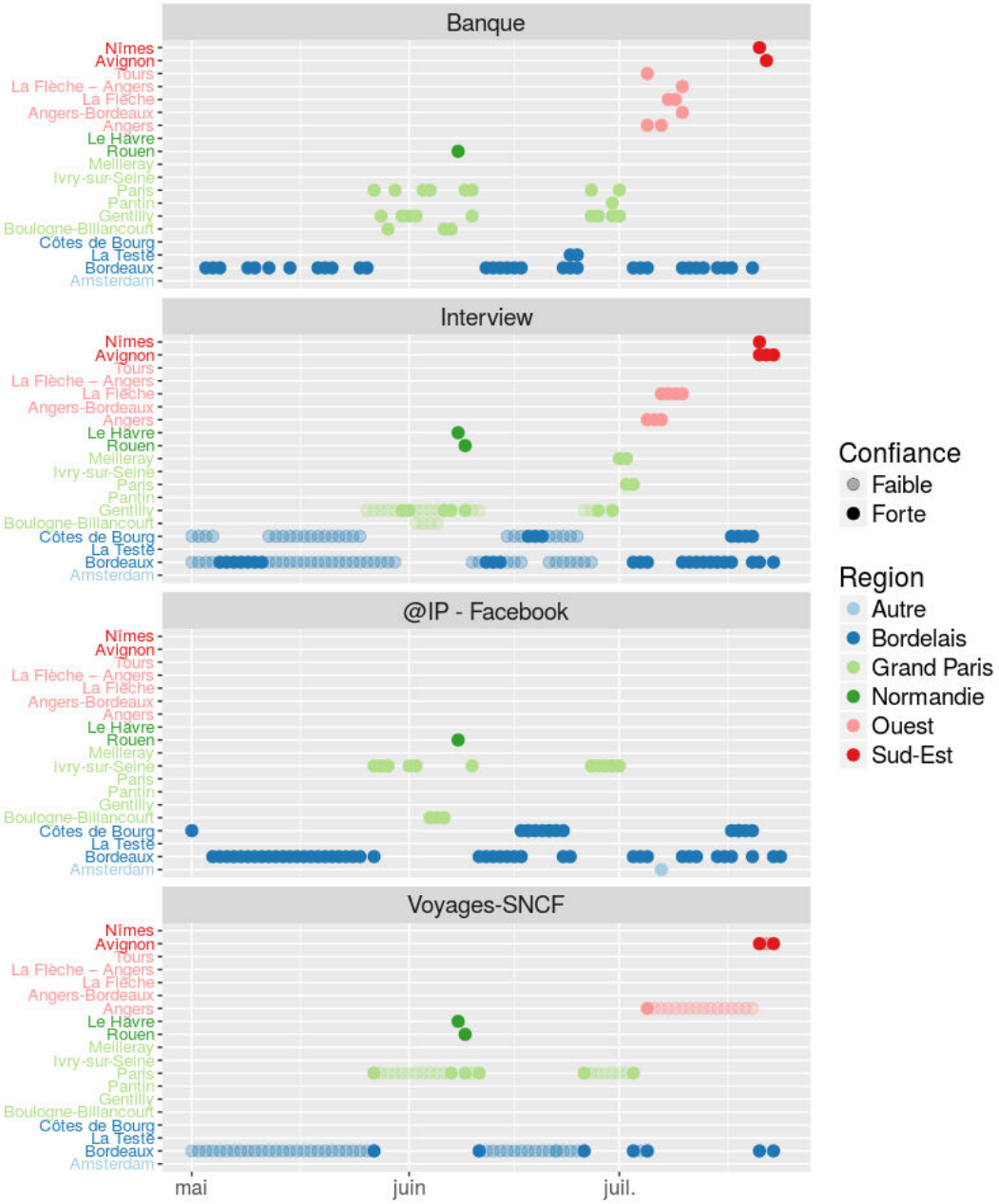


Figure ix Comparaison des localisations quotidiennes selon les différents jeux de données.

En revanche, nous utilisons *Facebook*, et les adresses IP furent bien détectées dans cette région viticole de la rive droite de la Dordogne, ce qui permet de nous rendre compte que, contrairement à ce que nous avions déclaré lors de l'entretien, nous n'avons pas été dans les Côtes de Bourg pendant la deuxième quinzaine du mois de mai. Mais *Facebook* n'a pas été utilisé tout le temps, notamment durant certaines périodes d'itinérance ce qui entraîne que les déplacements dans l'Ouest et le sud-est de la France n'ont pas été enregistrés. À noter que *Facebook* a enregistré une adresse IP provenant d'Amsterdam, alors que nous étions à

Angers pour une conférence. Ceci peut s'expliquer par l'utilisation du réseau Wifi international européen eduroam[422] qui peut consister en un VPN, n'affectant pas l'adresse IP à la véritable localisation de l'utilisateur.

De plus alors que nous étions à Gentilly, l'adresse IP renvoyée correspondait à Ivry-sur-Seine, la ville voisine, soit probablement la localisation du point d'accès à Internet. Les données issues des réservations de billet de train en ligne permettent quant à elle de valider les grands trajets à travers la France, et de préciser les inexactitudes temporelles de l'interview - sauf quand le retour se fait par un autre moyen (stop, covoiturage, achat de billet directement à la gare, etc.). En somme, mis à part la non-détection des séjours dans les Côtes de Bourg, les données bancaires sont ici la meilleure source de donnée permettant d'estimer les déplacements à l'échelle nationale. La synthèse de ces informations permet d'obtenir la figure x, qui est la reconstruction précise de notre agenda au cours de la période étudiée.

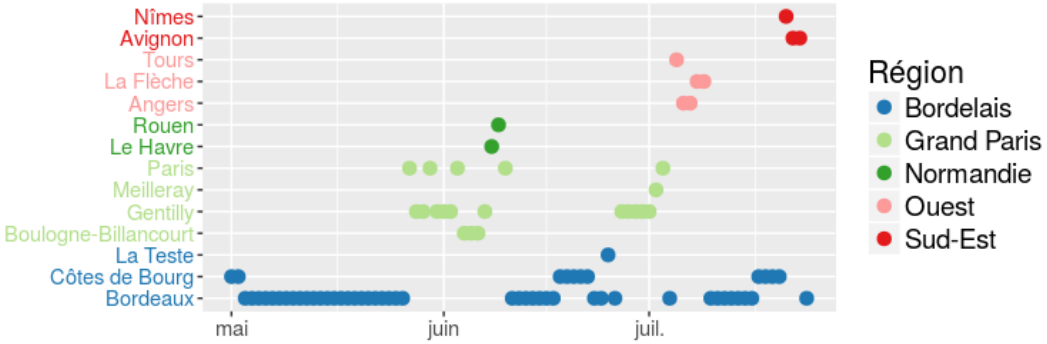


Figure x L'agenda final, obtenu par croisement des différentes sources de données.

Comme nous étudions les traces numériques laissées par les utilisateurs de Twitter dans cette thèse, nous allons essayer de voir dans quelle mesure la fréquentation d'un lieu inhabituel se manifeste sur les interactions sociales, et donc potentiellement sur la quantité d'information géographique laissée. Comme à titre personnel, nous n'utilisons pas Twitter[423], nous partons du principe que l'envoi d'un *tweet* part d'une volonté irrépressible de communiquer, et utilisons ainsi nos données d'appel téléphonique comme proxy. La figure xi illustre le nombre de personnes contactées en fonction des différentes régions visitées.

422. Contraction d'Education & Roaming (itinerance)
423. Ce qui limite les conflits d'intérêts

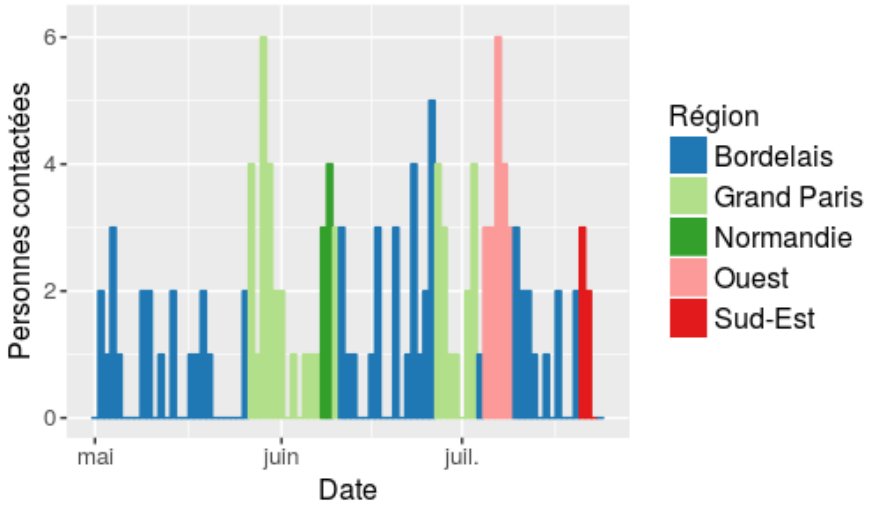


Figure xi Nombre de personnes contactées en fonction de la zone visitée

Comme visible dans la figure xi et le tableau iv ci-dessous, le nombre d'appels moyen est plus faible lorsque nous sommes dans la région de notre domicile principal. Ce nombre augmente dans les autres régions, avec des pics en début de séjour (manifestation de présence) et en fin de séjour. En raisonnant par analogie, il est donc possible qu'une personne laisse plus de traces sur les réseaux sociaux lorsqu'elle quitte son domicile pour signaler qu'elle visite des endroits plus inhabituels, peut être en compagnie d'une de ces relations sociales.

Afin d'obtenir un estimateur du niveau d'interaction sociale en fonction du lieu fréquenté, nous avons classé les contacts par rang. Plus la personne a été appelée fréquemment, plus son rang est faible, et inversement pour les personnes contactées peu de fois. Le rang varie ici entre 1 et 9, et le calcul du rang moyen (tableau iv) met en évidence le fait que nous contactons plus souvent les mêmes personnes lorsque nous sommes dans notre zone géographique de domicile que lorsque nous sommes en déplacement. Dans le second cas, il s'agirait d'une forme de « réveil d'agents dormant », soit une prise de contact virtuelle qui précède une rencontre dans la vraie vie. Bien qu'une étude montre le caractère décroissant des relations sociales en fonction de l'intervalle de temps entre les appels téléphoniques (Navarro et al., 2017), nous estimons que dans notre cas il est préférable de voir la personne directement, et les appels téléphoniques peuvent également servir à prendre des rendez-vous, lorsque nous sommes dans la même zone géographique.

Région	Nombre moyen de personnes contactées	Rang moyen
Bordelais	1.1	4.61
Grand Paris	2.05	6.24
Normandie	3.5	5.85
Ouest	3.8	5.84
Sud-Est	1.6	6.2

Tableau iv Nombre moyen de personnes contactées quotidiennement selon les zones géographiques visitées.

Comparaison des résultats à l'échelle locale (Paris et Bordeaux)

Les seules informations qui permettent une comparaison avec les données de l'entretien à l'échelle urbaine sont les données bancaires. Les lieux où des transactions ont été effectuées sont géolocalisés et le type de lieu où s'est effectuée la transaction y est répertorié. Les figures xii et xiii superposent ces deux sources de données. De manière générale, nous pouvons noter une très bonne concordance entre les lieux déclarés fréquentés et les zones où nous avons utilisé notre carte de paiement.

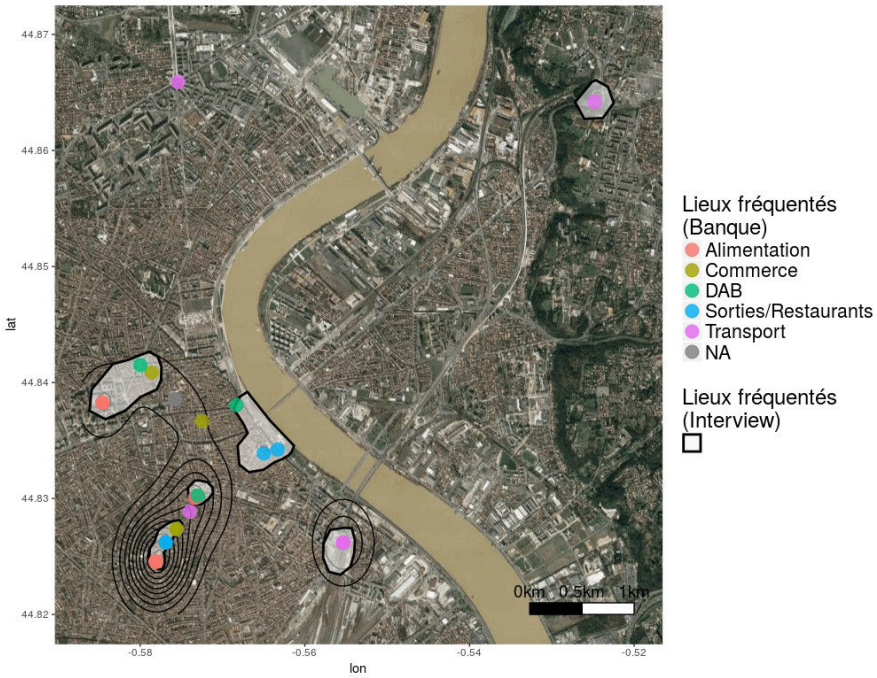


Figure xii Lieux fréquentés à Bordeaux. Données bancaires vs lieux déclarés (densité)

Nous pouvons observer sur la figure xii que toutes les zones déclarées comme visitées à Bordeaux sont enregistrées par les données bancaires, sauf la fois où nous sommes à l'Université. De plus, les données bancaires montrent certains lieux non déclarés comme fréquentés, comme la place Ravezies (au nord) ou après un covoiturage nous avons acheté

un ticket de tramway, ou encore un magasin de l'ultra-commerçante rue Sainte-Catherine que nous évitons habituellement, où nous avons acheté un cadeau d'anniversaire. Un avantage des données bancaires est également de pouvoir dater la fréquentation des lieux, de manière un peu similaire à un *tweet* géolocalisé. Cela dit, ces données ne permettent pas de faire la différence entre un lieu fréquenté de manière vraiment exceptionnelle d'un lieu fréquemment visité dans lequel on ne laisse qu'une seule trace, comme c'est le cas par exemple de la gare routière de Lormont Buttinière, où malgré 4 ou 5 visites, nous n'avons laissé qu'une seule trace bancaire.

Le même genre de remarque peut se faire pour la comparaison des lieux fréquentés à Paris (figure xiii), ou de manière générale les données bancaires ciblent bien les quartiers, avec cependant quelques lieux manquants. Nous avons globalement plus utilisé les distributeurs de billets à Paris qu'a Bordeaux, ce qui fait perdre un peu d'information sur les activités exercées. Cela dit, il paraît possible, connaissant le type de quartier à proximité d'inférer sans trop de problèmes une activité. Par exemple des retraits effectués entre 18 et 22 h dans les quartiers de Strasbourg Saint-Denis ou de la Butte aux Cailles peuvent suggérer que la personne s'apprête à aller au restaurant ou dans des bars. Cette hypothèse peut également être renforcée en connaissant les autres activités effectuées par cette personne lorsqu'elle n'est pas dans son quartier de domicile. En l'occurrence on note ici un nombre relativement important de paiements réalisés dans des lieux de sorties ou de restaurant par rapport aux autres activités.

Figure xiii Lieux fréquentés à Paris. Données bancaires vs lieux déclarés (densité)

Discussion

Les études de terrains, dans notre cas une interview, sont vraisemblablement le meilleur moyen d'évaluer les lieux fréquentés par une personne, même si de nombreux biais sont induits,

du fait de défaillances de mémoire. Ainsi, certains lieux peuvent être omis, et les horaires et fréquences de visites ne sont pas précis.

Cette rapide étude auto-centrée met en évidence la précision des données bancaires dans la captation de notre espace d'activité, à l'échelle nationale comme locale. En plus de fournir des données spatiales et temporelles, elles permettent également d'estimer les activités réalisées. Elles présentent cependant quelques biais liés à l'interprétation. En effet, leur création dépend de notre utilisation de la carte bancaire, et certains lieux peuvent ne pas être enregistrés.D'autres lieux, pourtant fréquentés assidûment peuvent se retrouver au même rang qu'un lieu fréquenté de manière exceptionnelle si l'on prend en compte simplement le nombre de retraits effectués dans une zone. Ce genre de donnée présente donc, comme nous le des similitudes avec les *tweets*, que cela soit en termes de production de données géographiques (chapitre 6) ou d'utilisation (parties C et D). En effet, dans le cas présent, ces données sont créées volontairement par la personne et cette dernière est au courant, mais leur utilisation dans le cadre d'étude sur les espaces d'activité constitue un écart clair entre le pourquoi de leur création (retrait d'argent ou volonté de communiquer) et l'utilisation potentielle qui peut en être fait.

Annexe C
D'autres données personnelles géolocalisées

Données des Taxis

A New-York, la commission des Taxis et des Limousines (TLC), met à disposition les traces GPS des Yellow et des Greens Taxis[424] enregistrés depuis 2009. La mise en place de GPS sur les taxis a pour objectif premier d'assurer l'accessibilité des citoyens ainsi que leur sécurité, et permet aux chercheurs d'analyser les déplacements dans la grande Pomme. Les objectifs sont similaires à ceux de la Delhi Police qui impose des GPS aux taxis et aux rickshaws[425]. La DIMTS[426], a développé une application, PoochhO[427], qui recense en temps réél la position des rickshaws[428] et des taxis dans la ville. Cela dit, ces données sont plus difficilement récupérables que celles des taxis de New York car non accessibles d'un simple clic. En janvier 2017, la compagnie Uber, une société mettant directement en relation des chauffeurs privés à des utilisateurs cherchant à se déplacer a annoncé qu'elle rendrait une partie de ces données publiques[429]. L'entreprise, présente dans 632 villes dans le monde[430] a depuis sa création enregistrée plus de 5 milliards de trajets[431], malgré une stratégie économique déconcertante[432]. Nous attendons toujours l'accès a leur base de données afin de l'étudier[433].

Voitures personnelles & et boitiers de l'assurance

Depuis quelques années, certaines compagnies d'assurances proposent à leurs clients dans certains pays (notamment l'Italie) d'installer des GPS dans leur voiture pour vérifier que leur conduite n'est pas à risque, et donc de réduire les coûts des-dites assurances. Outre le début de la fin d'un système mutualiste et les nombreuses connaissances que les assureurs sont susceptibles d'acquérir sur leurs clients (lieux fréquentés, fréquences de visites, type de conduite, etc.), cette approche permet de recueillir des données sur les déplacements d'utilisateurs privés, potentiellement utilisable pour gérer optimiser le trafic routier et faire des modèles de déplacement (Pappalardo and Simini, 2017).

424. http://www.nyc.gov/htm /t c/htm /about/tr p record data.shtm
425. Un moyen de transport motorisé à 3 roues très utilisé en Inde et en Asie du sud est.
426. Delhi Integrated Multi Modal Transit System Ltd, un partenariat entre le gouvernement de Delhi et la fondation de la banque IDFC.
427. https://p ay.goog e.com/store/apps/deta s? d=com.d mts.de h autojunct on&h =en
428. Quand ils activent leurs compteurs et leur GPS...
429. https://techcrunch.com/2017/01/08/uber debuts movement a new webs te offer ng access to ts traffic data/?nc d=rss
430. www.uber.com
431. https://newsroom.uber.com/5b on/
432. Avec 8.8 milliards de dollars obtenus par levées de fond et valorisé à plus de 68 milliards de dollars en 2017, https://www.zacks.com/stock/news/266715/w uber be the hottest po of 2017 mais perdant 2.8 milliards en 2016 http://www.bus ness ns der.fr/us/uber 2016 financ a numbers revenue osses 2017 4/
433. https://movement.uber.com/c t es

Annexe D

Dossier CNIL Twitter

Traitement n°	2017-UMR-IDEES-Twitter
Type	Inscription au registre de l'université
Finalité du traitement	Collecte de données publiques et géolocalisées issue du réseau social Twitter en vue d'étudier les mobilités quotidiennes dans différents contextes géographiques. Ces données individuelles permettent de définir des typologies de déplacement dans des zones urbaines qui connaissent des épidémies de dengue. Ces typologies sont ensuite utilisées pour la calibration de modèle de simulation de type multiagent. Ce modèle a pour objectif d'étudier le lien entre les mobilités journalières des populations en milieu urbain et la propagation de la Dengue.
Date de mise en œuvre	
Service chargé de la mise en œuvre	Unité Mixte de Recherche-IDEES Rouen, Université de Rouen "Identité et Différenciation de l'Espace, de l'Environnement et des Sociétés" 7 rue Thomas Becket, Bat. IRED, 76821 Mont Saint Aignan Cedex
Régime juridique applicable : Date d'inscription au registre	Inscription au registre de l'université le XXX
Date de décision de la CNIL	
Mise à jour : Date, objet	
Détails des finalités du traitement	Collecte des données : 1/ Collecte en temps réels de messages publics envoyés sur le réseau Twitter en Thaïlande, en Inde, aux Phillipines, à Tokyo (Japon), à Rio et Sao Paulo (Brésil) et dans le sud du Mexique. Ces données sont enregisrées entre le le 25 juin 2015 et le 25 juin 2019. L'ensemble de ces données sont récupérées via l'API STREAM fournie par Twitter (https://dev.twitter.com/streaming/overview). Seuls la géolocalisation (coordonnées GPS du mobile au moment de l'envoie du Tweet), la date d'envoie du Tweet (jour, heure, minute), le nom de l'identifiant (appelé pseudonyme) et la plateforme d'accés (Twitter pour Androïd, pour OS, FourSquare etc.) sont enregistrés. Le nom de la plateforme d'accès permet de supprimer l'ensemble des messages envoyés par un robot (dénommé Twitbot), c'est-à-dire n'étant pas identifié en tant que Twitter pour Androïd, pour OS, FourSquare etc. Anonymisation et traitement des données : 2/ Chaque pseudonyme a été associé à un code aléatoire, et la colonne pseudonyme a été supprimée de telle manière qu'il ne soit plus possible de remonter, possédant le code aléatoire, à son pseudonyme initial. Il n'est pas non plus possible de lier ces nouveaux identifiants (code aléatoire) à un compte twitter existant. Dès lors il n'est plus possible de poursuivre l'enregistrement des données issues de ce réseau social sur une nouvelle période et de les rattacher à leur identité (pseudonyme) de la période précédente une fois l'anonymisation effectuée.
Détails des finalités du traitement	3/ Pour chaque identifiant, les coordonnées GPS sont regroupées au centre d'une grille de cellules carrées de 180m de côté ajoutant ainsi un flou géographique. 4/ Les informations temporelles ont été filtrées de telle manière à identifer des jours de semaine (lundi, mardi, mercredi, jeudi) et de week-end (vendredi, samedi, dimanche) type, sans référence à la date exacte (par exemple le "8 avril" devient "mardi"). Ce traitement permet de créer un flou temporel sur les données. 5/ Ces données anonymisées sont recoupées avec d'autres données publiques non personnelles (recensements nationaux tels que le Census of India et le Census of Thailand, occupation du sol, etc.). L'objectif est de pouvoir qualifier les espaces de vie selon leur fonction (résidence, commerce, service) et leur fréquentation, tels que par exemple la fréquentation de lieux commerciaux le samedi. 6/ Ces données sont alors analysées pour mettre en évidence l'existence de flux préférentiels entre zones géographiques d'une part et pour révéler des typologies de mobilités d'autre part. Ces résultats sont alors utilisés pour calibrer un modèle multiagent par la génération d'individus synthétiques.
Fonction ou service auprès duquel s'exercent les droits	Siège UMR 6266 IDEES, Université de Rouen, 7 rue Thomas Becket, Bat. IRED, 76821 Mont saint Aignan CEDEX

Données de localisation	les coordonnées géographiques d'un message envoyé par un utilisateur (à la collecte)
Données de localisation : Origine de la collecte	Les données sont récupérées via l'API Stream fournie par l'entreprise Twitter
N° de sécurité sociale	
Données biométriques	Aucune information
Données biométriques : Origine de la collecte	
Données génétiques	Aucune information
Données génétiques : Origine de la collecte	
Infractions condamnations, mesures de sûreté	Aucune information
Infractions condamnations, mesures de sûreté : Origine de la collecte	
Appréciation sur les difficultés sociales des personnes	Aucune information
Appréciation sur les difficultés sociales des personnes : Origine de la collecte	
Données de santé ou de l'assurance maladie ou prélèvements biologiques identifiants	Aucune information
Données de santé ou de l'assurance maladie ou prélèvements biologiques identifiants : Origine de la collecte	
Données sensibles	Pas de données sensibles au sens de la loi informatique et libertés
Données sensibles : Origine de la collecte	
Personnes concernées par le traitement	Les personnes ayant envoyé des messages géolocalisés et publics dans la région de Bangkok (Thaïlande), en Inde, aux Philippines, au Japon, au Brésil et au Mexique, sur la plateforme Twitter, entre le 25 juin 2014 et le 25 juin 2019.
Destinataires des données	UMR IDEES Rouen et éventuellement des revues et des relecteurs. Nous ne fournirons pas de données brutes
Durée de conservation des données	Conservation des données brutes jusqu'à la fin de la recherche (soutenance de thèse, prévue en octobre 2018). Données anonymisées conservées deux ans de plus pour d'éventuels traitements ultérieurs.
Dispositions en vue d'assurer la sécurité des données	Données brutes stockées sur une machine/serveur et encryptées via ecryptfs. Cette machine est mise à disposition par http://www.huma-num.fr/ et se trouve à Lyon, dans les locaux (surveillés) de l'IN2P3. L'accés aux données se fait via SSL et SFTP. Les données anonymisées et traitées sont stockées en local sur disque dur encrypté via veracrypt + mot de passe bios et disque dur. Les serveurs Huma-num suivent les protections suivantes : - tous les serveurs utilisés appartiennent et sont administrés exclusivement par Huma-Num - ceux-ci sont hébergés physiquement et au niveau réseau au sein du Centre de Calcul de l'IN2P3, relié au réseau national RENATER - le tout est protégé par 2 niveaux de filtragé réseau - protection classique en entrée du centre de calcul - protection avancée par un pare-feu (firewall) Huma-Num avec inspection de contenu et signatures d'attaques - le tout est sauvegardé chaque nuit sur les deux robotiques de bandes magnétiques du centre de calcul - concernant le stockage des donnés, elles sont sur un serveur de fichiers NetApp, avec des règles de sécurité ad-hoc permettant d'assurer l'étanchéité entre les différents jeux de données et les serveurs qui en ont besoin

	- une mise à jour systématique et pluri-hebdomadaire de tous les systèmes est appliquée afin de les maintenir à jour - une information est donnée régulièrement aux responsables des sites basés sur des outils CMS standards (Wordpress, Drupla, Joomla, SPIP), pour leur demander de procéder à une mise à jour - des journaux des connexions reçues sur nos services sont constitués sur des serveurs dédiés - seuls les administrateurs d'Huma-Num ont les droits d'accès sur l'ensemble des données. Ils veillent à organiser les services mis à disposition, afin de respecter l'étanchéité des données entre les différents projets hébergées. Les CGU sont disponible : http://www.huma-num.fr/sites/default/files/huma-num-cgu-mars2015.pdf
Transfert hors-UE	non
Sous-traitance: Date du contrat comportant les clauses « Informatique et liberté	non
Analyse de risques: Niveau de gravité et de vraisemblance	
Utilisation de cookies	non
Remarques / Commentaires	

Annexe E
Complément d'information sur les *POI Google*

Répartition spatiale

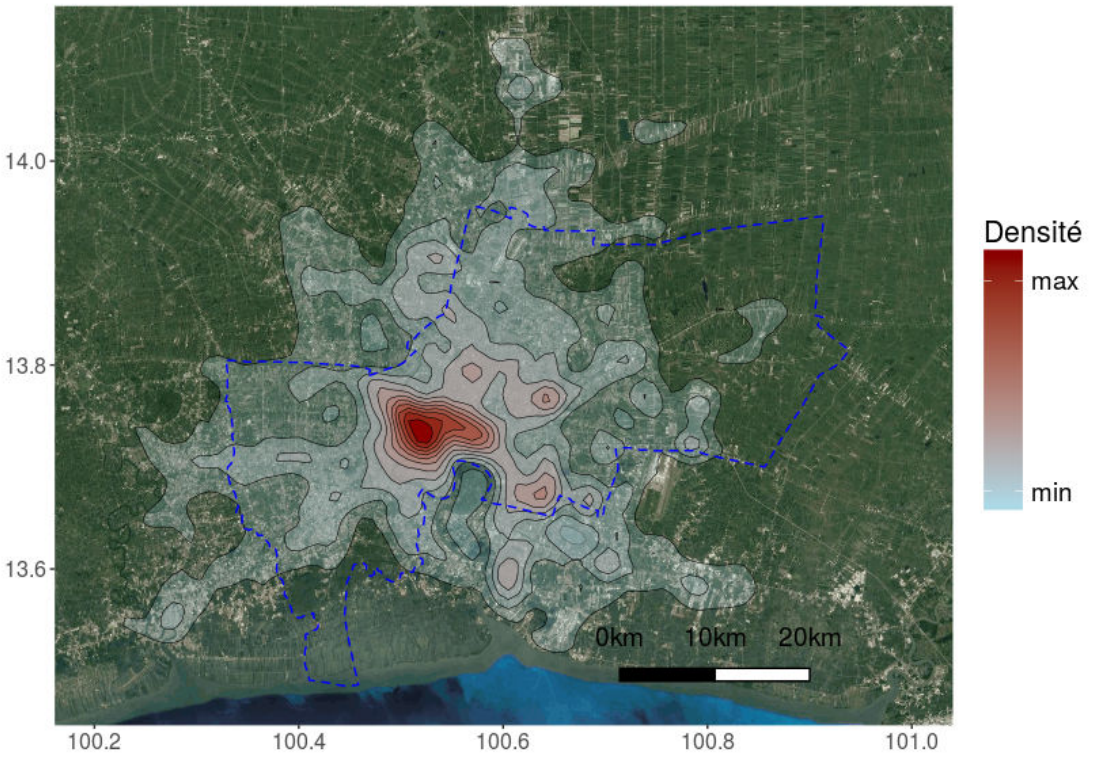


Figure xiv répartition spatiale des *POI*

Description des POI

Nous avons enregistré 99 catégories de *POI*. Outre les catégories très génériques « establishment » et « point of interest », chaque *POI* est affublé dans 41.6 % des cas d'une ou plusieurs catégories. Ainsi, la catégorie « restaurant » vient souvent avec la catégorie « food », tout comme « store » qui est associée à bon nombre de types de lieux, principalement commerciaux.

La figure xv ci-dessous exprime ces associations sous forme de corrélation, obtenues par analyse textuelle des termes caractérisant chaque *POI* selon le package tm présenté et implémenté sous R par (Meyer et al., 2008). Une corrélation à 1 signifie qu'une catégorie est toujours associée à une autre. Cette figure fait ressortir les niveaux d'association des catégories servant à définir des *POI*. Par exemple, "Restaurant" est quasiment toujours associé à "food", "ATM" à "finance".

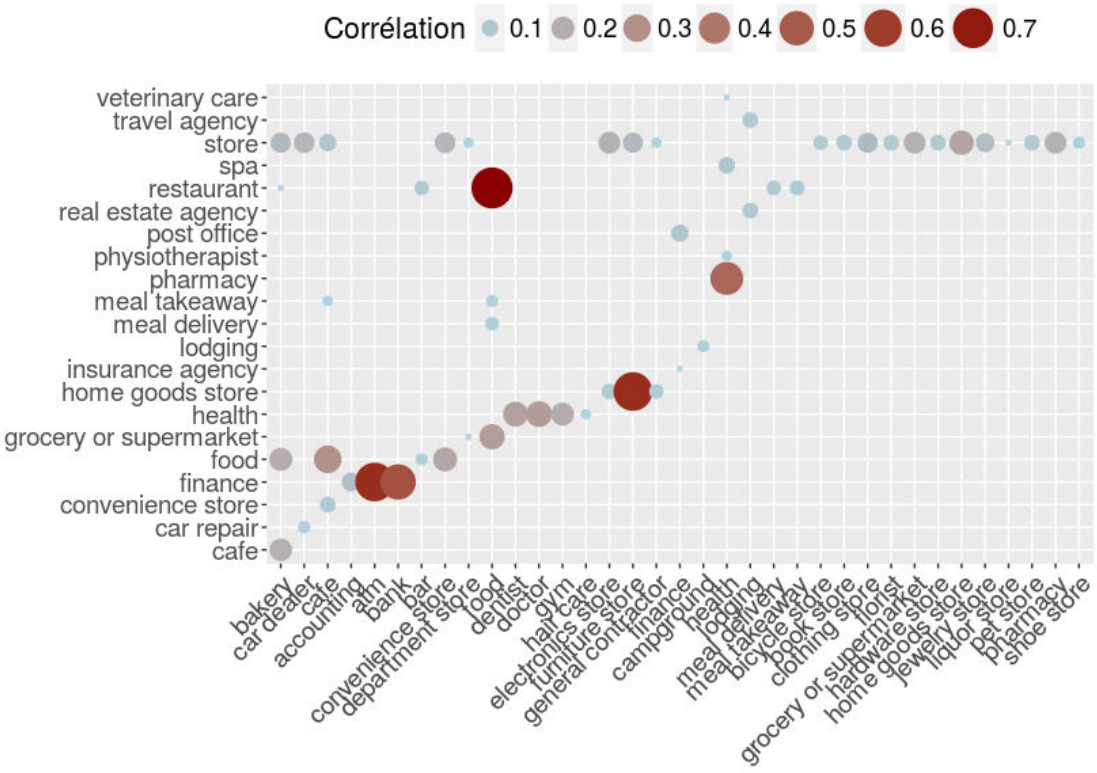


Figure xv Correspondance entre les catégories associées à un même *POI*. On ne récupère que les catégories qui sont présentes plus de 2000 fois avec un niveau de correspondance supérieur à 0.03, sans prendre en compte les « establishment » et « point of interest »

Nous avons récupéré 257 924 points d'intérêts de *Google Place* dans Bangkok et ces alentours, répartis en 99 catégories (restaurant, banques, écoles, déménageurs, avocats, réparateurs de voitures, etc.). Nous allons réaliser un premier regroupement de ces catégories selon une attribution en cascade. Par exemple les hôtels seront prioritaires sur les restaurants, et les restaurants prioritaires sur *food*. Ceci nous permet d'attribuer une seule catégorie à un *POI*.

Réduction du nombre de catégories

Nous regroupons ensuite les *POI* dans des classes (tableau xvi). La figure xvii montre quant à elle les niveaux de correspondances entre les catégories d'origines et un regroupement en 5 classes. Par exemple les restaurants sont toujours associés à la classe "Hang out" (ou sorties), tandis que "food" est partagé entre "Hang Out" et "food". La classe "Accomodation" est composée principalement de *POI* de type "lodging".

Catégorie initiale	Niveau 2	Niveau 3	Niveau 4
lodging, hostel, guest-house	ACCOMODATION	ACCOMODATION	ACCOMODATION
local_government_office, police, post_office, courthouse, fire_station, embassy	ADMIN	ADMIN	ADMIN
atm, bank, finance	BANK	BANK	BANK
library	EDUCATION	EDUCATION	EDUCATION
university	UNIVERSITY		
food	FOOD	FOOD	FOOD
gas_station	GAS_STATION	GAS_STATION	GAS_STATION
restaurant, night_club, cafe, bar	HANG_OUT	HANG_OUT	HANG_OUT
health, hospital, pharmacy, doctor, dentist, veterinary_care, physiotherapist	HEALTH	HEALTH	HEALTH
museum, art_gallery	CULTURE	LOISIR	LOISIR
movie_rental, zoo, movie_theater, aquarium, casino, bowling_alley	LOISIR		
park, cemetery, rv_park	PARK		
stadium	SPORT		
shopping_mall, department_store	MALL	MALL	MALL
place_of_worship	RELIGION	RELIGION	RELIGION
school	SCHOOL	SCHOOL	SCHOOL
car_rental, car_repair, car_dealer, car_wash	CAR	CAR	SEC
electrician	ELECTRONIC	ELECTRONIC	
moving_company, plumber, locksmith, painter	ARTISAN	SERVICE	SERVICE
general_contractor	LIBERAL		
home_goods_store	MINIMARKET	MINIMARKET	SHOP
store, pet_store, storage, book_store, florist, bicycle_store, laundry, liquor_store, funeral_home	SHOP	SHOP	
electronics_store, hardware_store	ELEC_STORE	ELEC_STORE	SHOP2
shoe_store, beauty_salon, clothing_store, jewelry_store, spa, hair_care	FASHION	FASHION	
furniture_store	FURNITURE	FURNITURE	
travel_agency	TOURISM	TOURISM	TOURISM
train_station, bus_station, airport, parking, taxi_stand, subway_station	TRANSPORT	TRANSPORT	TRANSPORT

Figure xvi Regroupement des *POI* en cascade. Les classes surlignées ont été utilisées pour la classification.

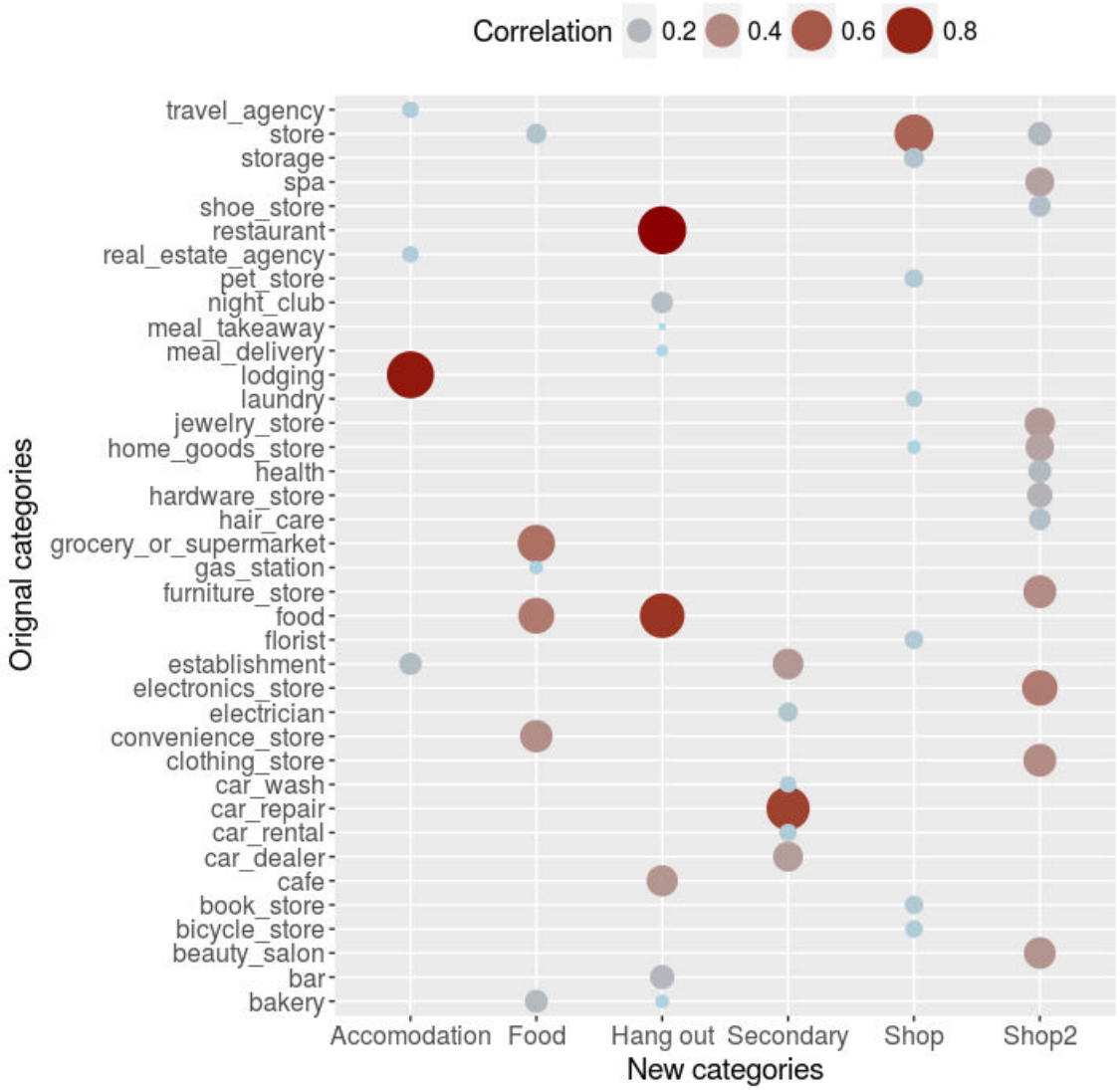


Figure xvii Niveau de correspondance des *POI* dans 5 catégories.

Annexe F
Regroupement des catégories de lieux *Facebook*

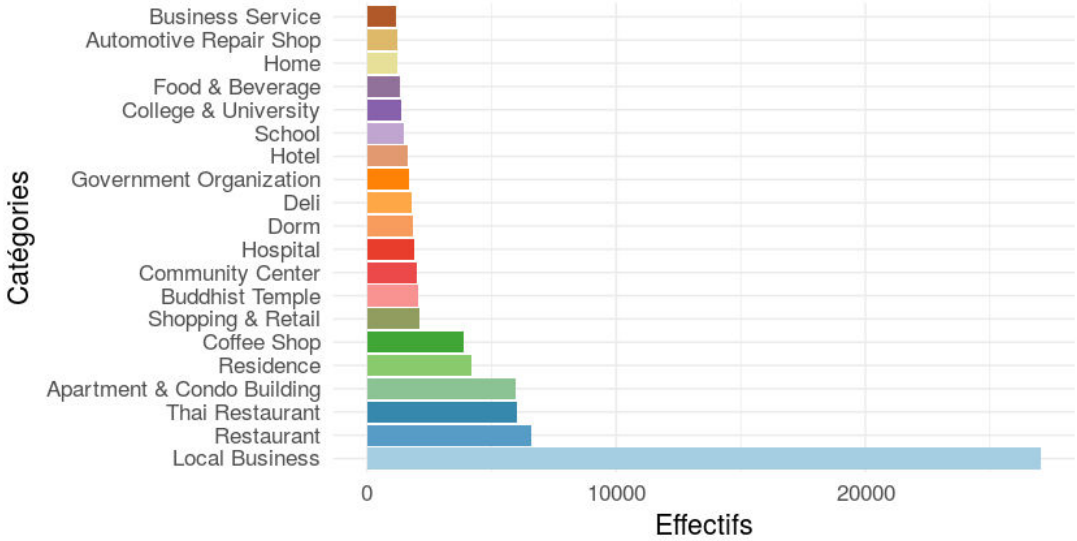


Figure xviii Nombre de lieux de *Facebook* associés à une catégorie

Catégorie Originale	**Nouvelle Catégorie**
Bank, Insurance	Bank
Bar, Pub, Beer, Night_Club	Bar
Beauty, Hair, Cosmetics, Tattoo, Nail_Salon	Beauty
Coffee, Cafe, Tea_Room	Cafe
Car, Motor, Automotive	Car
Education	Education
Deli, Food, Dessert_Shop, Ice_Cream_Shop, Bakery, Food	Food
Government	Government
Gerontologist, Gastroenterologist, Endocrinologist, Anesthesiologist, Neurologist, Doctor, Surgeon, Dentist, Dental, Medical, Hospital, Clinic	Health
Hotel, Dorm, Lodge	Hotel
Lawyer	Lawyer
Local_Business	Local_Business
Office, Company, Service, Business_Center, Business_Consultant	Office
Outdoor, Park	Outdoor
Religious, Temple, Church, Mosque	Religion
Residence, Apartment_&_Condo_Building, Home	Residence
Restaurant, Steakhouse, Diner	Restaurant
School	School
Cooking_School, Culinary_School, Religious_School, Art_School, Language_School, Cooking_School, Computer_Training_School, School_Fundraiser, Music_Lessons_&_Instruction_School, Driving_School, Trade_School, Massage_School, Medical_School, Aviation_School, Dance_School, School_Transportation_Service, Cosmetology_School, Traffic_School, Martial_Arts_School, Performing_Arts_School	School2
Shopping_&_Retail, Wholesale_&_Supply_Store, Grocery_Store, Wholesale_&_Supply_Store, Department_Store, Big_Box_Retailer, Electronics_Store, Collectibles_Store, Footwear_Store_, Clothing, Sporting_Goods_Store, Mobile_Phone_Shop, Jewelry_&_Watches_Store	Shopping_&_Retail
Shopping_District, Shopping_Mall	Shopping_Mall
Transit, Highway, Bus_Line, Railway_Station, Train_Station, Bus_Station	Transport
Travel, Tourist, Palace, Tour_Agency, History_Museum, Landmark_&_Historical_Place	Travel/Tourist
University, College, Campus_Building	University
Venue, Community_Organization, Community_Center, Convention_Center, Sports_Event	Venue

Figure xix Reclassification des catégories de lieux *Facebook*

Annexe G
Recherche de certains points dans des bases de données externes (*Foursquare* et *Instagram*)

Des coordonnées géographiques partagées par plusieurs utilisateurs

La figure xx montre la localisation des points GPS communs à plusieurs utilisateurs de *Twitter*.

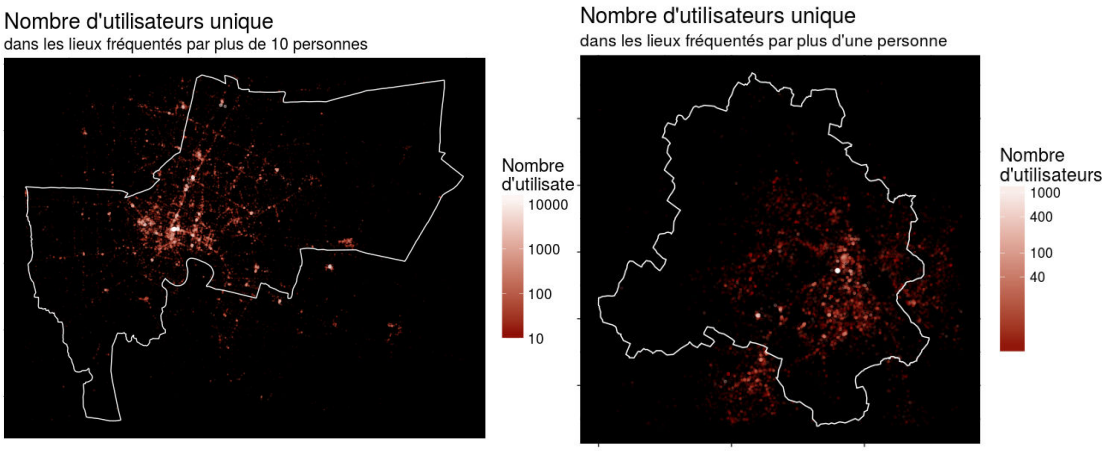


Figure xx Nombre de personnes par coordonnées partagées par plusieurs utilisateurs à Bangkok (gauche) et Delhi (droite).

Recherche dans la base de données de Facebook et de Foursquare

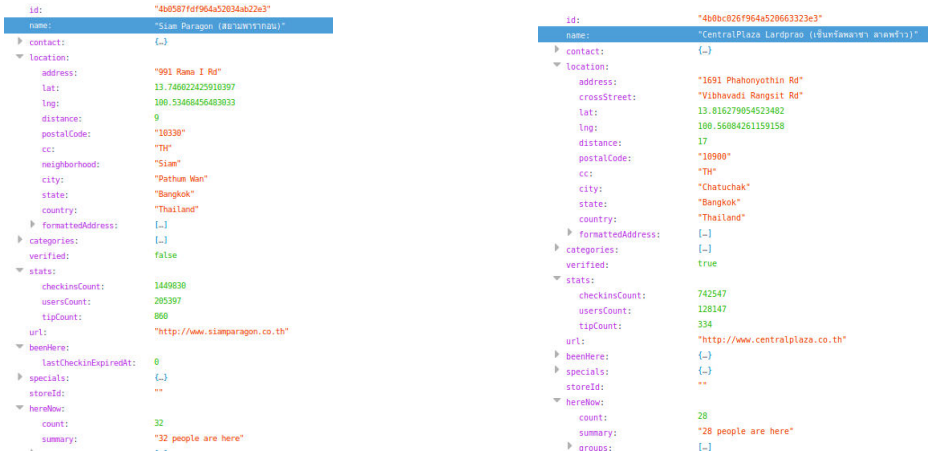


Figure xxi Résultat d'une requête de recherche de lieu sur *Foursquare* pour les coordonnées des points communs au plus grand nombre d'utilisateurs (1er et 4eme). Ces lieux correspondent respectivement à Siam Paragon et Central Plaza, deux malls extrêmement fréquentés. Figure également le nombre de '*check-in*' effectués sur *Foursquare* dans chacun de ces lieux (*checkinCount*), tout comme le nombre de personnes associées (*userscounts*) et le nombre de personne ayant effectué un checkin au moment de la requête (*herenow* ; *count*).

Figure xxii Résultat d'une requête sur *Instagram* (gauche) et *Facebook* (droite) pour les coordonnées 100.494, 13.7522. *Instagram* utilise la même base de données que *Facebook*, et la requête renvoie plus d'une vingtaine de lieux localisés au même endroit, dont "Bangkok, Thailand", catégorisé comme une ville (category_list : name ="City") qui comptabilise plus de 24 millions de *check-in* "checkins") et 2.26 millions de likes (fan_count)

Nom	**Catégorie 1**	**Catégorie 2**	**Catégorie 3**	**Nombre de check-ins sur la période**	**Nombre de checkins total**	**nombre de likes**
Bangkok	City	City		777328	24916700	2275470
Palais Royal	Landmark & Historical Place			189717	880015	10115
Suvarnabhumi Airport	Airport			165133	10466185	354822
Aéroport international Don Muang	Airport	Local Business		156140	1832025	34982
Palais Royal	Buddhist Temple			143653	180433	1648
Palais Royal - Sanam Luang	Local Business			142069	377994	6754
Suvarnabhumi Airport	Airport	Airport Terminal	Airport Lounge	110571	604497	4336
Chulalongkorn University	College & University	College & University	Workplace & Office	102125	891347	228849
Major Cineplex BangKapi	Movie Theater	Department Store	Just For Fun	87423	296990	9378
The Mall Bangkapi I	Barbecue Restaurant	Shopping & Retail	Arts & Entertainment	85372	1739296	39090

Tableau v Les 10 lieux avec le plus de *check-in Facebook*, en novembre 2017

Annexe H
Informations supplémentaires sur les traitements des données Twitter

Vitesses maximales selon différents seuils à Bangkok & Delhi

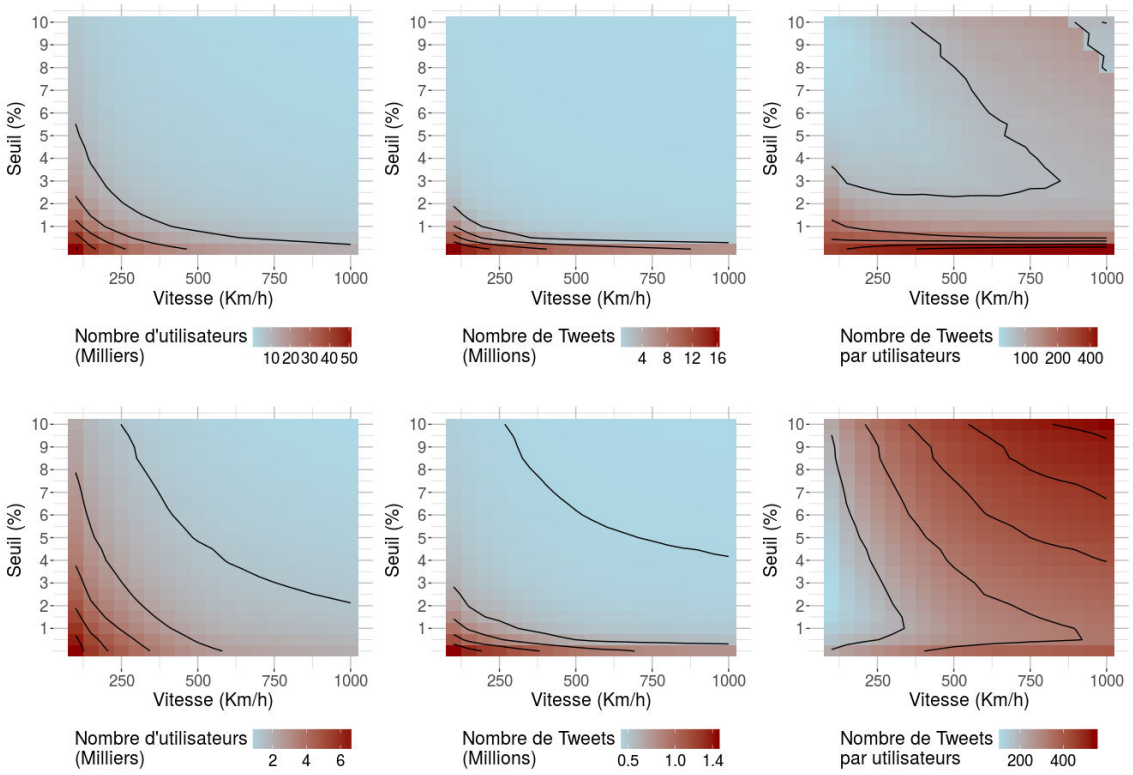


Figure xxiii influence de la vitesse moyenne et d'un pourcentage minimum de *tweets* au dessus de cette valeur sur le nombre d'utilisateurs (gauche), le nombre de *tweets* (centre) et le nombre de *tweets* par utilisateurs (droite). A Bangkok (haut) et Delhi (bas).

Influence de certains critères dans le choix de seuils pour la détection du domicile

La figure xxiv présente l'influence de différents seuils (sur le pourcentage de jours de *tweets*, le nombre de jours de *tweet* et le pourcentage de *tweet* entre 20h et 8h) utilisés dans l'algorithme de détection du domicile. En *x* figure le nombre d'utilisateurs répondant à ces différents seuils (plage de couleur) et en *y* est présenté le R^2 associé (soit la correspondance entre le nombre d'utilisateurs habitant dans un sous-district et la population de ce sous-district).

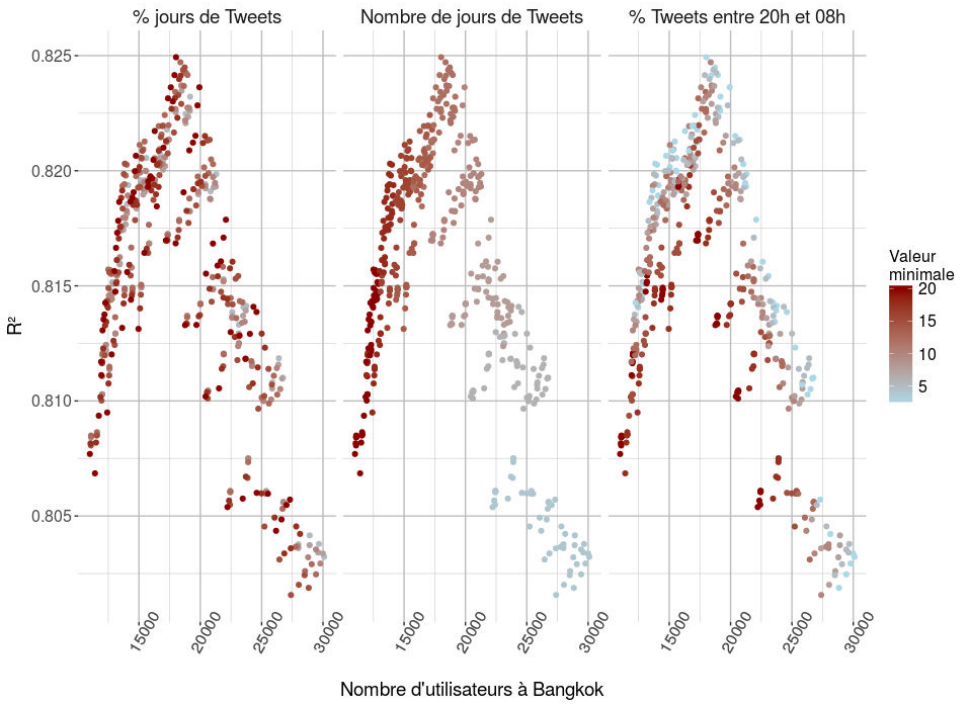


Figure xxiv influence de la valeur de certains seuils sur 3 paramètres sur le nombre de domiciles détectés et la corrélation entre la population enregistrée au sous-district et celle estimée.

Considérons maintenant toutes les configurations qui permettent de détecter entre 20 000 et 30 000 domiciles. La figure xxv, gauche montre la moyenne des écarts types entre la population estimée au domicile et la population du recensement. La figure de droite présente la moyenne des pourcentages de déviation. Les résultats de la figure xxv, droite, sont très similaires à ceux de la figure 73.

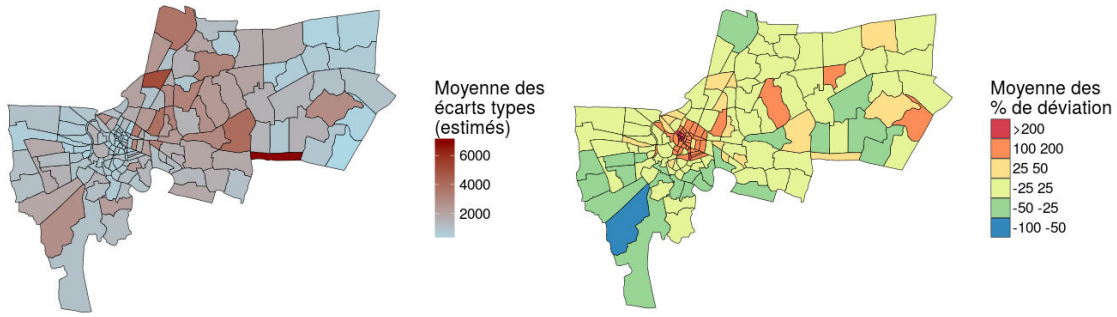


Figure xxv Moyenne des écarts types estimés (gauche) et moyenne des pourcentages de déviation (droite).

Représentativité des données Twitter dans la métropole de Bangkok

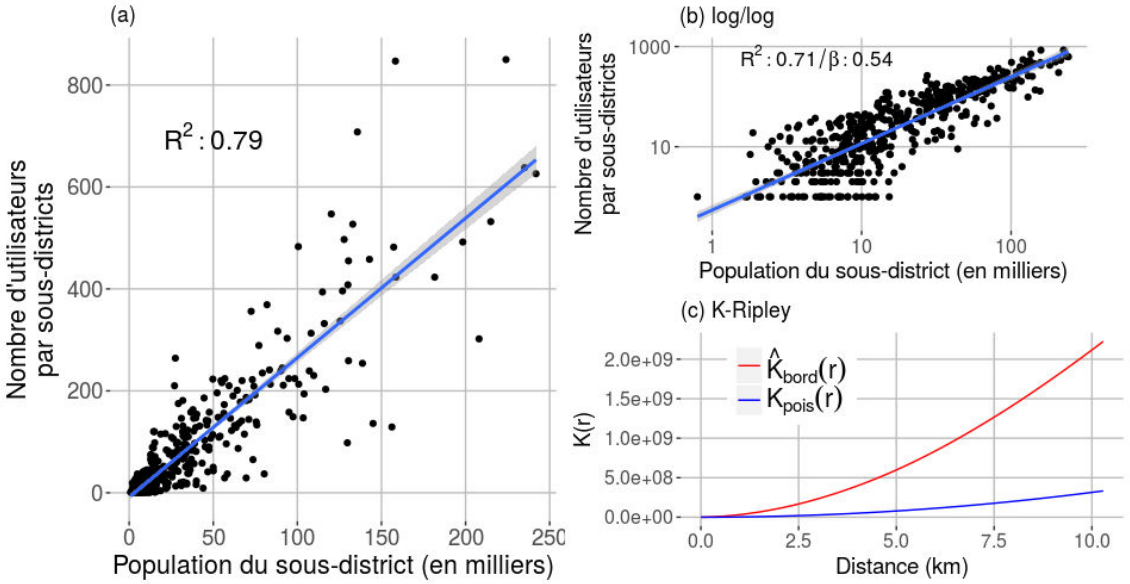


Figure xxvi Représentativité de l'échantillon à Bangkok. Lien entre la population enregistrée au sous-district par le recensement de 2010 et le nombre d'utilisateurs de *Twitter* habitant dans les mêmes sous-districts (a). En logarithme (b). Et écart à un modèle Poissonien d'après le K de Ripley (c).

Annexe I
Questionnaire de terrain

Date : Place : Time :

1. Generality

1.1 Name
1.2 Age
1.3 Gender M / F
1.4 Profession
1.5 Office Address

2. Dengue Knowledge

2.1 Have you heard of Dengue Fever? Yes / No

2.2 How? Radio / TV / Newspapers / Religious places / word of mouth / MCD /
2.3 Do you know how we can get Dengue and how it spread?

2.4 Did you get it? Yes / No
2.4.1. If yes : When ?
2.4.2. If yes : Where?
2.5 How do you know it was dengue? Diagnostic by doctor or hospital / blood test /

2.6 Did someone of your relative get it? Yes / No
2.6.1. Who ?

2.6.2. When?
2.6.3. Where?

2.7 How do you know it was dengue? Diagnostic by doctor or hospital / blood test /

2.8 Do you have problems with mosquito? Yes / No
2.9 Where? In house / at work / in transport / in public spaces /

2.10 Do you use mosquito repellent? Yes / No

2.11 Which kind? cream / spray / coil / liquid/ mosquito-net in bed / mosquito net in all doors and windows

2.12 Which frequency? Daily / weekly / monthly /

2.13 Did the MCD came to check for mosquito breeding site last year? Yes / No

2.14 Do you check your house to control mosquito breeding site? Yes / No

2.15 How?

3. Housing

3.1 Address :

3.2 Rent / Owner

3.3 Why do you live here ?

3.4 For how long?

3.5 Type of house? Individual / building

3.6 Where do you sleep at night? Closed room / court-yard / open space /

3.7 Which floor ? GF / 1/ 2 / 3 / 4

3.8 How many household? ___

3.9 How many rooms? ___

3.10 Sanitary facilities? Private Latrine / Community Latrine / No latrine

3.11 Water supply type? Tap Water / Water Storage / Borewell

3.12 How many Cooler? ___

3.13 How many AC ? ___

4. Mobility

how many XXX in your household Cycle: / motorbike: / Car:

5. Economy

How much do you spend in your rent?

What is your average income?

	Place (Colony – Landmark)	Time	Transport mode	How often	Comment
Job					
Food Shopping					
Relative visit / Parties (marriage / Birthday)					
Religion					
Park					
Restaurant					
Mall					
Theater / Cinema					
Other					

Annexe J
Graphiques supplémentaires pour le chapitre 7

Contribution des variables et des individus dans l'ACP (chapitre 7)

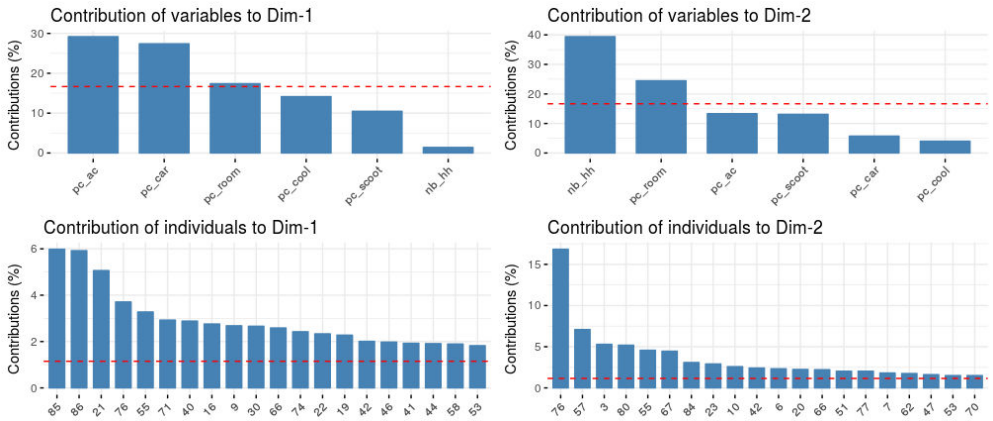


Figure xxvii Contribution des variables et des individus dans l'ACP (figure 112, chapitre 7)

Choix du nombre de classes pour un K-means sur trois critères de déplacements des personnes interrogées sur le terrain.

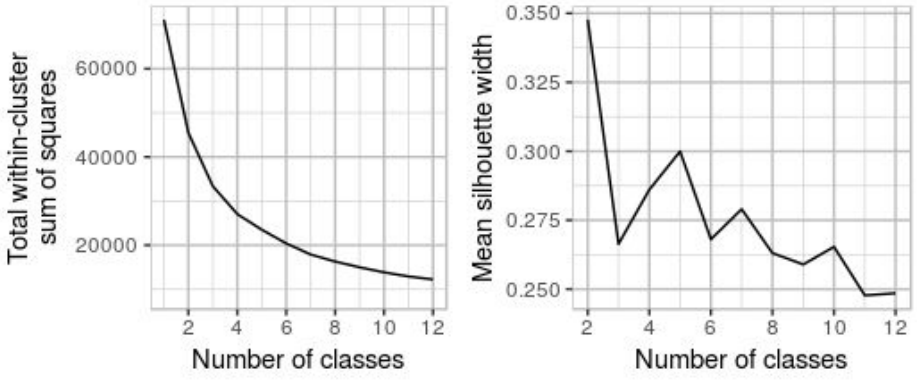


Figure xxviii Choix du nombre de classes pour le kmeans. "Elbow method" à gauche, et silhouette moyenne à droite. À noter la présence d'un coude entre 3 et 6 classes sur la figure de gauche, et de creux à 3, 6 et 9 classes sur la figure de droite.

Annexe K
Interprétation des classes de la première méthode de classification de l'utilisation du sol à Bangkok (chapitre 9)

Interprétation des 11 classes

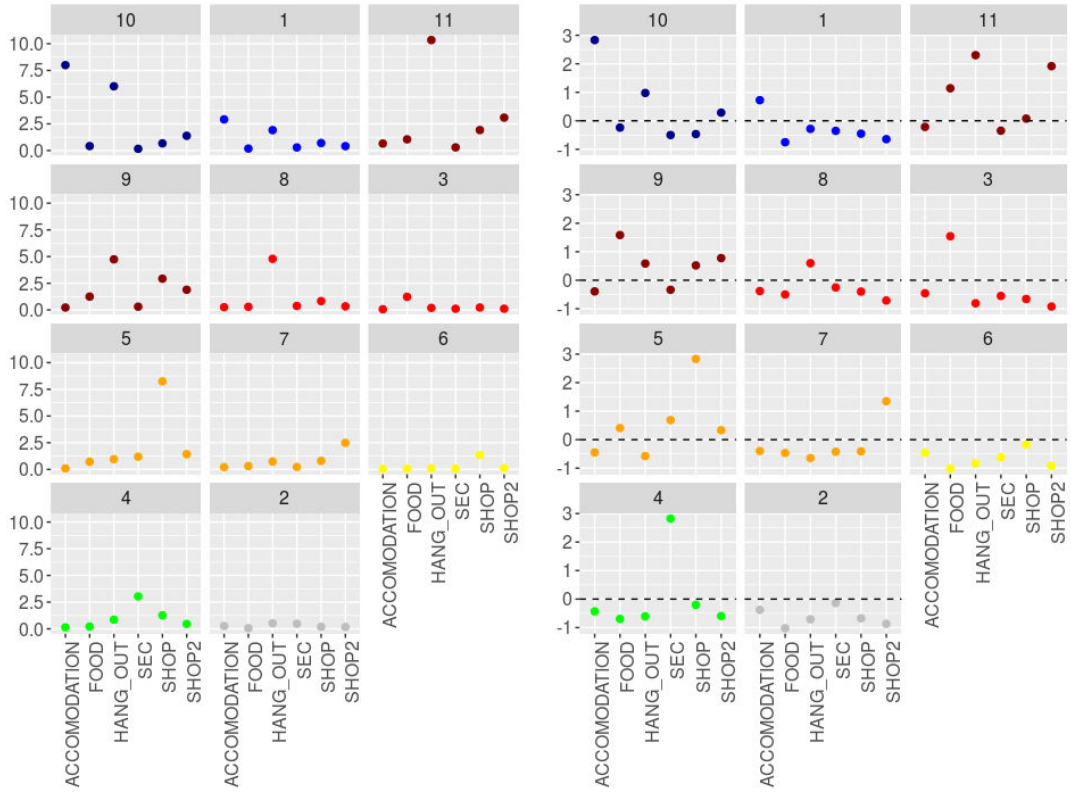


Figure xxix Interprétation des 11 classes. Valeur moyenne pour chacune des classes dans chaque catégorie. Valeurs absolues à gauche et centrées réduites à droite.

Reclassification en 6 catégories

Lieux d'hébergements et de sorties :

- Classe 10 :Très forte surreprésentation des *accomodation*, avec lieux de sorties et magasins
- Classe 1 : surreprésentation des *accomodation*, quelques lieux de sorties

Lieux de sorties majeurs et de shopping

- Classe 11 : Très forte surreprésentation HO, nourriture et shopping.
- Classe 9 : Très forte surreprésentation nourriture, fort pour shopping, HO.
- Classe 5 : Très forte surreprésentation des shops, présence de shop2, sec et food.

Lieux de sorties et de shopping

- Classe 8 : Légère surreprésentation HO.
- Classe 3 : Surreprésentation des lieux de nourritures non déterminés, le reste très faible.

Quelques commerces

- Classe 7 : Surreprésentation des lieux de shopping (shop2).
- Classe 6 : Faibles valeurs, présence de shop.

Secteur secondaire

- Classe 4 : Surreprésentation des lieux d'activités du secteur secondaire.

Faibles valeurs

- Classe 2 : Valeurs très faibles.

Annexe L
Graphiques supplémentaires pour le chapitre 11

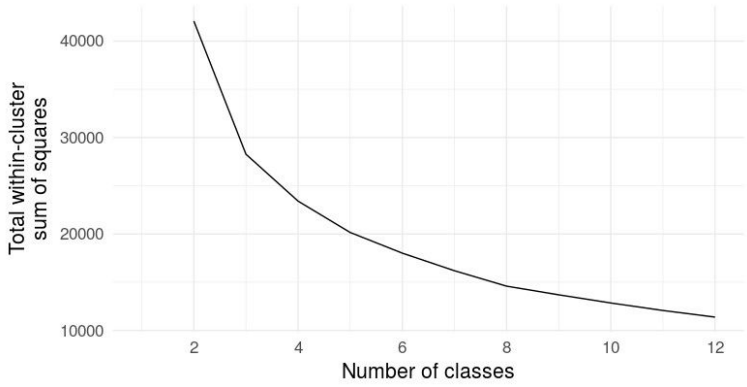


Figure xxx Choix du nombre de classes (méthode "elbow") pour effectuer un K-means sur l'échantillon des utilisateurs de *Twitter* à Bangkok d'après leurs paramètres de dispersion.

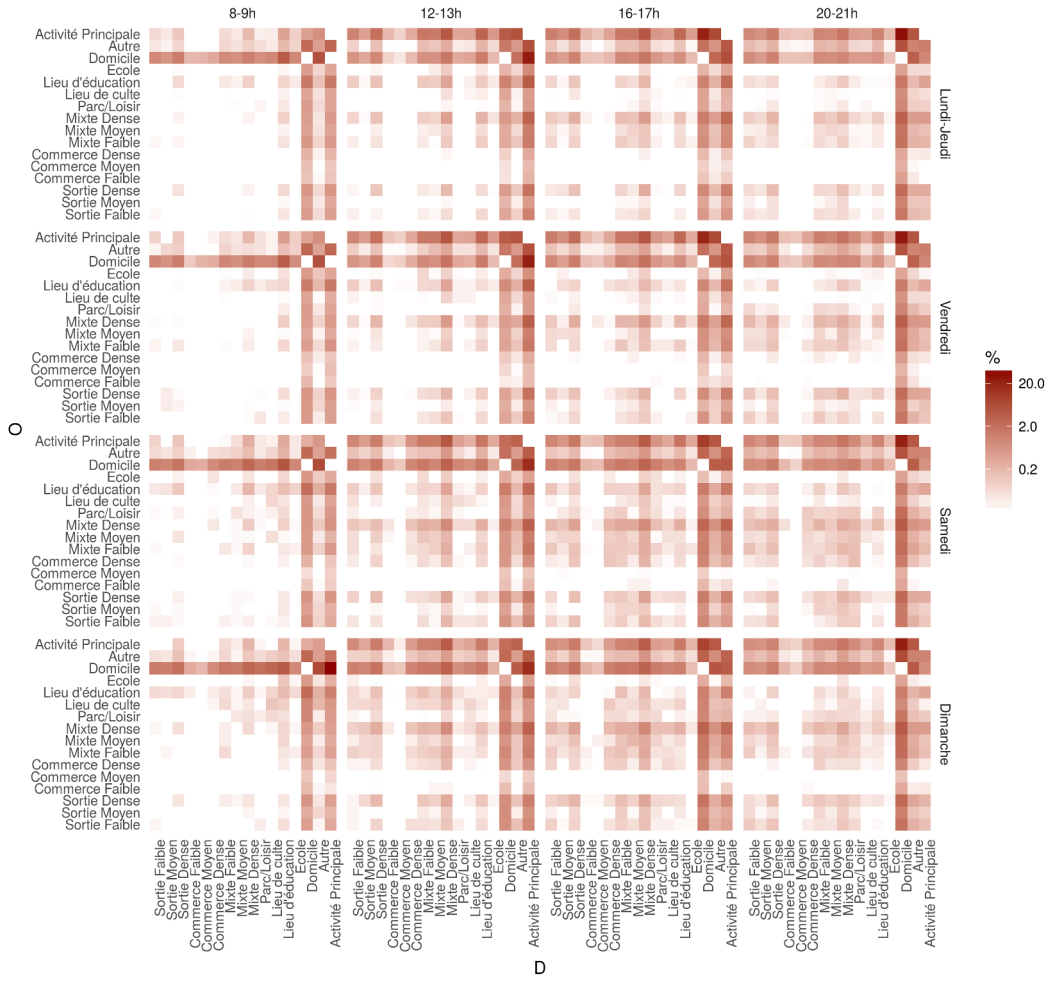


Figure xxxi Matrice de transition horaire entre activités.

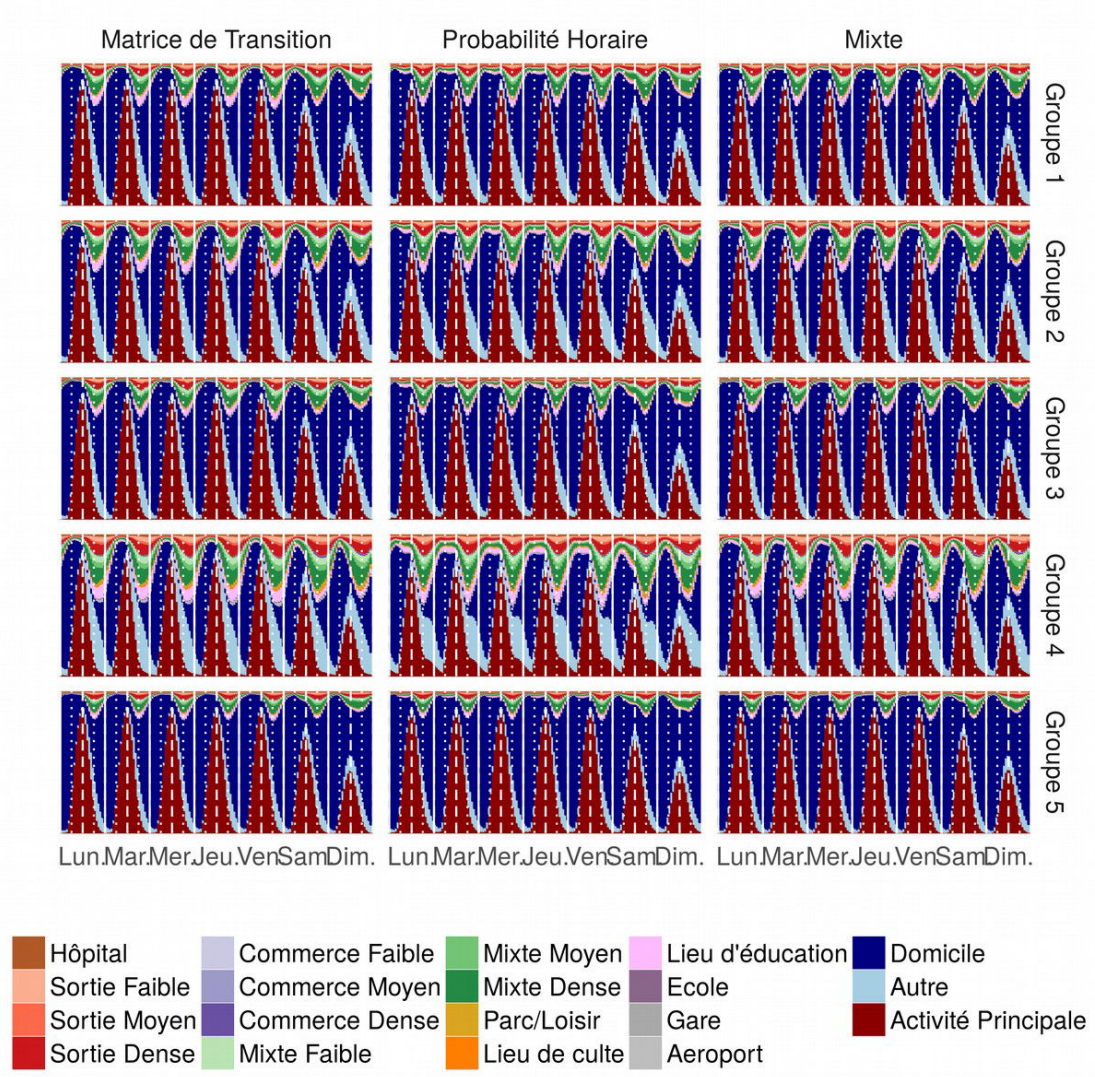


Figure xxxii Comparaison des agendas moyens pour chacun des 5 groupes définis et des méthodes utilisées pour reconstruire une journée d'un agent.

Liste des figures (Annexe)

i Évolution des états sérologiques dans le modèle méta-population fermé à 4 patchs. . . . 537
ii Matrices de flux utilisées pour l'exemple du modèle méta-population fermé. 538
iii Résultat des différents tests montrant l'évolution de la part du nombre de personnes infectées au cours des itérations. . . . 539
iv Résultat des différents tests montrant l'évolution de la part du nombre de personnes infectées au cours des itérations, par expérimentation. . . . 540
v Résultats des différents scénarios lorsque la force d'infection est modulée par un paramètre extérieur global. . . . 541
vi Évolution de la part d'hôtes (traits pleins) et de vecteurs (pointillés) au cours du temps, selon les différentes configurations et les zones. . . . 543
vii Localisation des lieux déclarés être fréquentés à Bordeaux. . . . 547
viii Localisation des lieux déclarés être fréquentés à Paris. . . . 549
ix Comparaison des localisations quotidiennes selon les différents jeux de données. 552
x L'agenda final, obtenu par croisement des différentes sources de données. . . . 553
xi Nombre de personnes contactées en fonction de la zone visitée . . . 554
xii Lieux fréquentés à Bordeaux. Données bancaires vs lieux déclarés (densité) . 555
xiii Lieux fréquentés à Paris. Données bancaires vs lieux déclarés (densité) . . . 556
xiv répartition spatiale des *POI* . . . 563
xv Correspondance entre les catégories associées à un même *POI*. . . . 564
xvi Regroupement des *POI* en cascade. Les classes surlignées ont été utilisées pour la classification. . . . 565
xvii Niveau de correspondance des *POI* dans 5 catégories. . . . 566
xviii Nombre de lieux de *Facebook* associés à une catégorie . . . 567
xix Reclassification des catégories de lieux *Facebook* . . . 568
xx Nombre de personnes par coordonnées partagées par plusieurs utilisateurs à Bangkok (gauche) et Delhi (droite). . . . 569
xxi Résultat d'une requête de recherche de lieu sur *Foursquare* pour les coordonnées des points communs au plus grand nombre d'utilisateurs (1er et 4eme). . . . 570
xxii Résultat d'une requête sur *Instagram* (gauche) et Facebook (droite) pour les coordonnées 100.494, 13.7522 . . . 570
xxiii influence de la vitesse moyenne et d'un pourcentage minimum de *tweets* au dessus de cette valeur sur le nombre d'utilisateurs (gauche), le nombre de *tweets* (centre) et le nombre de *tweets* par utilisateurs (droite). A Bangkok (haut) et Delhi (bas). . . . 572
xxiv influence de la valeur de certains seuils sur 3 paramètres sur le nombre de domiciles détectés et la corrélation entre la population enregistrée au sous-district et celle estimée. . . . 573
xxv Moyenne des écarts types estimés (gauche) et moyenne des pourcentages de déviation (droite). . . . 574
xxvi Représentativité de l'échantillon dans la région de Bangkok . . . 574
xxvii Contribution des variables et des individus dans l'ACP (figure 112, chapitre 7) . . . 578
xxviii Choix du nombre de classes pour le kmeans. "Elbow method" à gauche, et silhouette moyenne à droite. . . . 578

xxix interprétation des 11 classes. 579
xxx Choix du nombre de classes (méthode "elbow") pour effectuer un K-means sur l'échantillon des utilisateurs de *Twitter* à Bangkok d'après leurs paramètres de dispersion. 581
xxxi Matrice de transition horaire entre activités. 581
xxxii Comparaison des agendas moyens pour chacun des 5 groupes définis et des méthodes utilisées pour reconstruire une journée d'un agent. 582

Liste des tableaux (Annexe)

i Villes déclarées comme fréquentées entre le 1er mai et le 23 juillet 2017. . . . 546
ii Lieux déclarés être fréquentés à Bordeaux. 548
iii Lieux déclarés être fréquentés à Paris. 549
iv Nombre moyen de personnes contactées quotidiennement selon les zones géographiques visitées. 555
v Les 10 lieux avec le plus de *check-in Facebook*, en novembre 2017 571

Table des matières détaillée

Résumé . . . vii
Abstract . . . viii
Publications & valorisation . . . ix
Remerciements . . . xiii

Table des matières xix

Introduction générale 1

Partie A : De la prise en compte des mobilités dans l'étude des épidémies de dengue 11

Chapitre I : Extension du domaine de la dengue . . . 13

1 Épidémiologie de la Dengue . . . 13
1.1 Généralités sur la Dengue . . . 13
1.1.1 Aspects cliniques . . . 14
1.1.2 Écologie des vecteurs . . . 16
1.1.3 Distribution spatiale . . . 18
1.2 Une petite histoire de la dengue . . . 19
1.2.1 Evolution sémantique . . . 20
1.2.2 Le moustique comme vecteur de maladies . . . 20
1.2.3 Isolement du Virus . . . 21
1.2.4 Évolution des cas de Dengues dans le monde . . . 22
1.3 La dengue, une maladie émergente . . . 25
2 Mobilités et propagation des épidémies . . . 26
2.1 Définition des mobilités spatiales . . . 26
2.1.1 Déplacements sur de longues distances . . . 27
2.1.2 Les mobilités quotidiennes et urbaines . . . 28
2.2 Les mobilités comme facteur de l'extension géographique de la dengue . . . 29
2.2.1 Propagation du vecteur . . . 30
2.2.2 Importation de la dengue . . . 31
2.2.3 Propagation de la dengue en milieu urbain . . . 34
3 La dengue en milieu urbain, sous l'angle de la complexité . . . 35
3.1 Systèmes complexes et théorie de la complexité . . . 35
3.2 La dengue, une maladie complexe . . . 36
3.3 Le système des mobilités urbaines . . . 37

Chapitre II : L'épidémie et le territoire : Le fardeau de la dengue à Delhi et Bangkok . . . 41

1 Delhi, une mégapole faite de ruptures. . . . 42
1.1 Une répartition inégale des populations . . . 43
1.2 Une grande variété de types de quartiers d'habitation . . . 44
1.3 De grandes disparités socio-économiques entre les quartiers 46

1.4 Des potentiels de mobilité différents en fonction des districts et très genrés . . . 49
1.5 Situation de la dengue à Delhi . . . 51
2 Les discontinuités de Bangkok . . . 54
2.1 Généralités sur Bangkok . . . 54
2.2 Les tendances de mobilité à Bangkok, d'après la littérature 63
2.2.1 Tendances globales . . . 63
2.2.2 Temps de transport selon les genres et les âges . . . 64
2.3 Dengue à Bangkok . . . 67
2.3.1 Aspects climatiques . . . 69
2.3.2 Dengue et population . . . 75

Chapitre III : Prendre en compte les mobilités dans la modélisation des arboviroses : la possibilité de modèles . . . 79

1 Des débuts des modèles compartimentaux à la prise en compte du vecteur de la dengue . . . 80
1.1 Ross et les premiers modèles mathématiques en épidémiologie 80
1.2 Les premiers modèles compartimentaux . . . 81
1.3 La prise en compte du vecteur . . . 83
2 Modélisation d'épidémies et mobilité humaine . . . 84
2.1 Modèles compartimentaux métapopulations . . . 85
2.2 Modèle épidémiologique en réseaux . . . 87
2.3 Modélisation à base d'agents . . . 88
3 Ontologie d'un modèle de mobilité urbaine . . . 89
3.1 Quels concepts utiliser pour analyser et modéliser les mobilités humaines ? . . . 89
3.1.1 La motilité, ou les différents aspects sociaux à l'origine des mobilités . . . 89
3.1.2 L'espace d'activité et la time-geography . . . 93
3.2 Vers une formulation abstraite d'un modèle de mobilité . . . 95
3.2.1 Quels attributs pour un agent mobile dans le système de la dengue ? . . . 95
3.2.2 Attributs d'un agent mobile idéal . . . 98
3.2.3 Processus de modélisation : . . . 99

Partie B : Les traces numériques : de «nouvelles» données pour aborder les mobilités 103

Chapitre IV : L'abondance des traces numériques géolocalisées . . . 107

1 Des données protocolaires... . . . 110
1.1 ...Générées lors de l'utilisation d'un téléphone mobile... . . . 110
1.2 ...Lors de l'usage d'Internet... . . . 112
1.3 ...Ou à d'autres moments . . . 115
1.3.1 Données bancaires . . . 115
1.3.2 Données transports publics . . . 115
2 Création et collecte de données sur les réseaux sociaux . . . 116
2.1 *Twitter* ou le micro-blogging . . . 119

2.2 Sites de « check-in » . . . 119
2.3 Partage de photos . . . 121
2.4 Des données géographiques volontaires ? . . . 122
3 Collecte massive et identité numérique . . . 124
4 L'appareil législatif : la CNIL . . . 129
4.1 Contexte de création . . . 129
4.2 Rôle de la CNIL . . . 130
4.3 CNIL et données personnelles . . . 131
4.3.1 Liste des données personnelles selon la CNIL . . . 131
4.3.2 « Principes clés de la protection des données personnelles » . . . 132
Chapitre V : « Nouvelles données » et mobilités urbaines : état de l'art . . . 139
1 Représentativité des différents jeux de données . . . 141
1.1 Part de la population concernée par le jeu de données utilisé 141
1.2 Représentativité et connaissance de l'échantillon . . . 144
1.2.1 Valider avec des données de recensement . . . 144
1.2.2 Caractérisation socio-économique et démographique de l'échantillon . . . 145
1.2.3 Hybridation avec des enquêtes de terrains . . . 147
2 À la recherche de lois sur les mobilités individuelles . . . 147
3 Analyse des interactions spatiales . . . 151
3.1 Création de matrices origine-destination . . . 151
3.2 Modèle gravitaire . . . 153
3.3 Modèle d'opportunité et modèle radiatif . . . 155
4 Modélisation à base d'agents . . . 157
5 Des approches centrées sur les temporalités des activités . . . 162
5.1 Mise en relation de l'utilisation du sol et des mobilités . . . 163
5.2 Caractériser l'utilisation du sol à partir de POI . . . 165
5.3 Définir l'utilisation du sol à partir des données de mobilités 166
6 Autres études en contexte urbain . . . 167
6.1 Détection d'événements . . . 167
6.2 Rôle du réseau social dans les mobilités . . . 168
7 Utilisation en contexte d'épidémies maladies infectieuses . . . 169
Chapitre VI : Traces numériques à Delhi et Bangkok . . . 179
1 Données *Twitter* à Delhi et Bangkok . . . 180
1.1 Présentation du service . . . 180
1.1.1 Utilisation du réseau social dans le monde . . . 181
1.1.2 *Twitter* et la géolocalisation . . . 182
1.1.3 Décorticage d'un Tweet . . . 183
1.1.4 Récupération des données . . . 186
1.1.5 *Twitter* et les bots . . . 187
1.2 Données brutes à Delhi et Bangkok . . . 189
1.2.1 Collecte des données . . . 189
1.2.2 Descriptions des données . . . 189

1.3 Filtres et prétraitements . . . 196
1.3.1 Suppression des utilisateurs occasionnels et des bots 196
1.3.2 Des points GPS partagés par plusieurs utilisateurs... 199
1.4 Temporalité à Bangkok et Delhi . . . 201
1.5 Création des espaces d'activité individuels . . . 202
1.5.1 Formalisation de la méthode . . . 202
1.5.2 Détection du domicile . . . 205
1.6 Synthèse . . . 210
2 *Check-in Facebook* à Bangkok . . . 211
2.1 Présentation . . . 211
2.2 Collecte des données . . . 215
2.3 Filtres . . . 219
2.3.1 suppressions de lieux . . . 219
2.3.2 suppressions des pics . . . 220
2.4 premiers résultats . . . 222

Partie C : Approche mixte des mobilités à Delhi 229

Chapitre VII : Des enquêtes de terrain pour appréhender les mobilités à Delhi et poser les bases d'un modèle à base d'agents . 233

1 Malviya Nagar et ses alentours, un bon laboratoire pour aborder les mobilités à Delhi . . . 233
1.1 De grandes hétérogénéités socio-économiques et spatiales... 233
1.2 ...matérialisé par des quartiers très différents . . . 234
1.3 Une bonne accessibilité et des zones commerçantes attractives 239
1.4 Une zone où la dengue sévit . . . 241
2 Présentation de l'enquête de terrain . . . 242
2.1 Protocole . . . 242
2.2 Présentation de l'échantillon . . . 246
2.2.1 Déroulement de l'enquête . . . 246
2.2.2 Structure démographique et socio-économique . . . 249
2.2.3 Étude des espaces d'activités . . . 256
2.2.4 Étude des coprésences . . . 259
2.2.5 Analyse des mobilités . . . 263
2.2.6 Synthèse de l'enquête . . . 267
3 Des données terrain à des agendas individuels . . . 269
3.1 Création des agendas . . . 269
3.1.1 Harmonisation des plages horaires . . . 269
3.1.2 Affectation d'un jour pour effectuer une activité . 272
3.1.3 Estimation des durées et des plages horaires des activités . . . 274
3.1.4 Gestion des redondances . . . 276
3.1.5 Quelques exemples d'agendas de synthèses . . . 277
3.2 Analyse des agendas générés et perspectives pour la simulation à base d'agents . . . 279
3.2.1 Regrouper des individus selon les similarités de leur agenda . . . 279

3.2.2 Déduire des matrices de transition entre les activités 283
Chapitre VIII : Traces numériques et espaces d'activités : analyse des mobilités et génération d'agents à Delhi . . . 291
1 Vers une meilleure connaissance de l'échantillon . . . 293
1.1 Qui *tweet* à Delhi et à Malviya Nagar ? . . . 293
1.2 Dans quel type de lieu ? . . . 296
1.3 Comment sont composés les espaces d'activité ? . . . 304
1.4 À quel moment une activité est-elle effectuée ? . . . 306
1.5 Quelle est l'activité principale d'un utilisateur ? . . . 307
2 Des espaces d'activités discrets à des agendas individuels continus . . 310
2.1 Limites des espaces d'activités « bruts » . . . 310
2.2 Proposition d'algorithme . . . 311
2.2.1 Étape 1 : définir une semaine de réalisation . . . 311
2.2.2 Étape 2 : définir les jours de visites . . . 313
2.2.3 Étape 3 : Définir les plages horaires . . . 314
2.2.4 Étape 4 : ajout de l'activité principale et du domicile 316
2.2.5 Étape 5 : Gestion des doublons . . . 317
2.3 Présentations des agendas reconstitués . . . 318
2.3.1 D'agendas individuels... . . . 318
2.3.2 ...A des groupes d'utilisateurs ? . . . 319
3 Générer des agendas de synthèses . . . 322
3.1 Étape 1 : Données initiales de l'espace d'activité . . . 322
3.1.1 Définir les activités qu'un agent va effectuer . . . 322
3.1.2 Répartir les activités dans un nombre de lieux . . . 323
3.2 Étape 2 : Jour(s) et semaine(s) de visite . . . 324
3.2.1 Sélection d'une semaine de réalisation . . . 324
3.3 Étape 3 : jour(s) de visite . . . 325
3.3.1 Nombre de jours de visite hebdomadaire . . . 325
3.3.2 Sélections des jours de la semaine . . . 325
3.3.3 Plages horaires . . . 329
3.3.4 harmonisation . . . 330
4 Discussion des résultats . . . 331
4.1 À partir des AR *Twitter* . . . 331
4.2 À partir des données du terrain . . . 334
Partie D : Mobilités et activités à Bangkok 347
Chapitre IX : Temporalité des activités à Bangkok . . . 351
1 Différentes facettes des activités commerciales . . . 351
2 Apports des données *Google* et *OSM* à la cartographie de l'utilisation du sol à Bangkok . . . 357
2.1 Collecte des données . . . 357
2.2 Typologie des activités commerciales à partir de *POI* . . . 360
2.3 Mobilisation des AOI . . . 364
2.4 Finalisation de la couche d'utilisation du sol . . . 369
3 Fréquentations temporelles . . . 374

4 Définir l'utilisation du sol en fonction les profils temporels des traces numériques . . . 379

Chapitre X : Les mobilités à Bangkok : Variations sur le thème des données et des méthodes . . . 387

1 Le rythme de la ville . . . 388
1.1 Se déplacer dans une ville congestionnée . . . 388
1.1.1 Les transports en commun à Bangkok : le fleuve, l'avenue et le rail . . . 388
1.1.2 Sur la route . . . 392
1.2 La pulsation urbaine : des tempos différents selon les données 398
1.2.1 Des « hotspots » de traces numériques . . . 398
1.2.2 Répartition spatio-temporelle des traces numériques 400
1.2.3 Quantifier la dilatation urbaine . . . 401
2 Les interactions dans la ville : données, méthodes et nuances . . . 406
2.1 Différentes méthodes pour concevoir ces matrices . . . 406
2.1.1 Approche par les lieux fréquentés . . . 406
2.1.2 Interactions entre les utilisateurs de Twitter, selon une approche sur les lieux fréquentés en commun . 408
2.2 Comparaison des résultats . . . 410
2.3 Structure et partition de Bangkok par les interactions dans la ville . . . 415
2.3.1 Définir des sous-régions fonctionnelles par la détection de communautés . . . 415
2.3.2 Application à Bangkok . . . 417

Chapitre XI : Génération d'agendas individuels : Premières notes d'un modèle à base d'agents . . . 423

1 De données épisodiques à des agendas individuels continus . . . 424
1.1 Préalable : Définir l'activité principale . . . 424
1.1.1 Méthode . . . 424
1.1.2 Analyse de l'échantillon . . . 425
1.2 Protocole de reconstitution des agendas . . . 428
1.2.1 Étape 1 : Définir la probabilité de visite d'un lieu de l'espace d'activité . . . 428
1.2.2 Étape 2 : Définir les jours de visite . . . 430
1.2.3 Étape 3 : Définir les heures de réalisation . . . 431
1.2.4 Étape 4 : agglomération et gestion des doublons . 434
1.3 Résultats . . . 435
2 Génération d'agendas . . . 442
2.1 Définir des groupes d'utilisateurs . . . 442
2.2 Protocole de génération d'agendas . . . 447
2.2.1 Étape 1 : initialisation des caractéristiques de l'espace d'activité . . . 447
2.2.2 Étape 2 : Fréquence et jours de visite . . . 449
2.2.3 Étape 3 : Durée d'une activité . . . 451
2.2.4 Étape 4 : Définir la séquence quotidienne des activités . . . 452

2.3 Résultats 454
3 Comment affecter des localisations ? 461

Conclusion générale 473

Bibliographie 485

Liste des figures 521

Liste des tableaux 531

Liste des acronymes 532

Annexes 535

Annexe A Un modèle méta-population fermé 536

Annexe B Analyses de nos traces numériques 545

Annexe C D'autres données personnelles géolocalisées 558

Annexe D Dossier CNIL Twitter 560

Annexe E Complément d'information sur les *POI Google* 563

Annexe F Regroupement des catégories de lieux *Facebook* 567

Annexe G Recherche de certains points dans des bases de données externes. 569

Annexe H Informations supplémentaires sur les traitements des données Twitter. 572

Annexe I Questionnaire de terrain 575

Annexe J Graphiques supplémentaires pour le chapitre 7. 578

Annexe K Interprétation des classes de la classification de l'utilisation du sol à Bangkok (chapitre 9). 579

Annexe L Graphiques supplémentaires pour le chapitre 11 581

Liste des figures (Annexe) 584

Liste des tableaux (Annexe) 584

Table des matières détaillée 585