Fouille de données billettiques pour l'analyse de la mobilité dans les transports en commun

Anne-Sarah Briand

UNIVERSITÉ PARIS-EST

IFSTTAR

Thèse présentée pour obtenir le grade de docteur
Université Paris-Est

École doctorale MSTIC

Discipline : Informatique

Fouille de données billettiques pour l'analyse de la mobilité dans les transports en commun

Par : **Anne-Sarah BRIAND**

Directrice : Latifa OUKHELLOU, DR IFSTTAR
Encadrant : Etienne CÔME, CR IFSTTAR

Membres du jury:

Rapporteur : Younès Bennani, Professeur des Universités, Université Paris 13
Rapporteur : Catherine Morency, Professeur des Universités, Ecole Polytechnique de Montréal

Examinateur : Nicolas Baskiotis, Maître de Conférences, UPMC
Examinateur : Patrice Aknin, Directeur de Recherche, IRT SystemX
Examinateur : Etienne Côme, Chargé de Recherche, Ifsttar
Examinateur : Latifa Oukhellou, Directrice de Recherche, Ifsttar
Invité : Sébastien Leparoux, Kéolis Rennes

Date de soutenance : 5 décembre 2017

Remerciements

Pour commencer, je remercie l'IFSTTAR de m'avoir accordé cette bourse de thèse, et plus particulièrement ma directrice de thèse, Latifa Oukhellou, d'avoir dans un premier temps proposé ce sujet. J'ai énormément appris en travaillant à ses côtés et je la remercie pour son investissement tout au long de la thèse et ses nombreux conseils qui m'ont permis de progresser. J'adresse également mes plus sincères remerciements à Étienne Côme qui a co-dirigé cette thèse. Il m'a encouragée et soutenue quand cela était nécessaire et je lui souhaite de rester cette belle personne qu'il est à l'heure actuelle. Travailler avec eux était un réel plaisir en raison de leur bonne entente académique et humaine.

Je tiens également à remercier Keolis Rennes, et plus particulièrement Sébastien Leparoux, pour le temps qu'il nous a consacré et pour sa présence lors de ma soutenance de thèse. Cette thèse n'aurait pas vu le jour sans les données qu'ils nous ont fournies.

Je remercie le Jury d'avoir consacré du temps à l'évaluation de mes travaux et d'avoir assisté à ma soutenance et à Patrice Aknin d'avoir accepté de le présider. Je remercie tout particulièrement Catherine Morency et Younès Bennani qui, en tant que rapporteurs, ont pris le temps de lire l'ensemble de mon manuscrit et m'ont fait part de nombreuses remarques très pertinentes qui m'ont permis de mieux mettre en perspective l'ensemble de mes travaux. Enfin je remercie Nicolas Baskiotis pour ses retours sur mon travail.

J'ai eu grand plaisir à travailler avec le professeur Martin Trépanier à l'école Polytechnique de Montréal. Je le remercie pour sa gentillesse et sa bienveillance. Cette collaboration m'a permis d'élargir mes horizons. Je remercie également l'Université Paris Est pour la bourse de mobilité qu'elle m'a accordée et sans laquelle cette collaboration n'aurait pas été possible. Enfin je tiens à remercier Sylvie Cach pour son accompagnement, son aide et sa gentillesse dans toutes les démarches administratives liées à la thèse.

Je tiens à remercier l'ensemble des membres du GRETTIA actuels et passés pour l'accueil chaleureux que j'ai reçu. Je remercie Jean-Patrick, Mustapha d'être présent pour les doctorants à chaque fois qu'ils en ont besoin, Annie et Joëlle pour leur accueil lors de mon arrivée, Allou pour sa bonne humeur permanente, Laurent pour ses encouragements en cours de sport, Olivier pour son aide en toute occasion, Jean-Louis pour nos conversations sur les plantes, Régine pour son implication auprès des doctorants et tous les autres pour leurs sourires et leur gentillesse. J'adresse des remerciements tout particuliers à M.-Khalil El Mahrsi qui m'a accompagnée tout au long de ma thèse en tant que collègue mais également en tant qu'ami. J'ai énormément appris à ses côtés et je le considère comme une des personnes les plus compétentes de mon entourage. Enfin j'ai une pensée particulière pour tous les doctorants que j'ai eu la chance de côtoyer mais surtout pour Florence ma plus grande confidente et ma complice pour tout ce qui concernait la nourriture. Merci à Johanna pour m'avoir appris les rouages de l'ifsttar, Hani mon complice de bureau, Moncef le papa des doctorants, Florian qui restera toujours un petit stagiaire pour moi, Josquin pour les sessions musicales, Maxime, Nassim, Milad, Kevin, Matthieu, Dihya, Rémi, ...

Enfin je tiens à remercier ma famille et mes amis de m'avoir encouragée pendant toute ma thèse. Je remercie tout particulièrement Guillaume qui m'a toujours soutenue pour le meilleur et pour le pire.

Résumé

Les données billettiques sont de plus en plus utilisées pour l'analyse de la mobilité dans les transports en commun. Leur richesse spatiale et temporelle ainsi que leur volume, en font un bon matériel pour une meilleure compréhension des habitudes des usagers, pour prédire les flux de passagers ou bien encore pour extraire des informations sur les événements atypiques (ou anomalies), correspondant par exemple à un accroissement ou à une baisse inhabituels du nombre de validations enregistrées sur le réseau.

Après une présentation des travaux ayant été menés sur les données billettiques, cette thèse s'est attachée à développer de nouveaux outils de traitement de ces données. Nous nous sommes particulièrement intéressés à deux challenges nous semblant non encore totalement résolus dans la littérature : l'aide à la mise en qualité des données et la modélisation et le suivi des habitudes temporelles des usagers.

Un des principaux challenges de la mise en qualité des données consiste en la construction d'une méthodologie robuste qui soit capable de détecter des plages de données potentiellement problématiques correspondant à des situations atypiques et ce quel que soit le contexte (jour de la semaine, vacances, jours fériés, ...). Pour cela une méthodologie en deux étapes a été déployée, à savoir le clustering pour la détermination du contexte et la détection d'anomalies. L'évaluation de la méthodologie proposée a été entreprise sur un jeu de données réelles collecté sur le réseau de transport en commun rennais. En croisant les résultats obtenus avec les événements sociaux et culturels de la ville, l'approche a permis d'évaluer l'impact de ces événements sur la demande en transport, en termes de sévérité et d'influence spatiale sur les stations voisines.

Le deuxième volet de la thèse concerne la modélisation et le suivi de l'activité temporelle des usagers. Un modèle de mélange de gaussiennes a été développé pour partitionner les usagers dans les clusters en fonction des heures auxquelles ils utilisent les transports en commun. L'originalité de la méthodologie proposée réside dans l'obtention de profils temporels continus pour décrire finement les routines temporelles de chaque groupe d'usagers. Les appartenances aux clusters ont également été croisées avec les données disponibles sur les usagers (type de carte) en vue d'obtenir une description plus précise de chaque cluster. L'évolution de l'appartenance aux clusters au cours des années a également été analysée afin d'évaluer la stabilité de l'utilisation des transports d'une année sur l'autre.

Abstract

Ticketing logs are being increasingly used to analyse mobility in public transport. The spatial and temporal richness as well as the volume of these data make them useful for understanding passenger habits and predicting origin-destination flows. Information on the operations carried out on the transportation network can also be extracted in order to detect atypical events (or anomalies), such as an unusual increase or decrease in the number of validations.

This thesis focuses on developing new tools to process ticketing log data. We are particularly interested in two challenges that seem to be not yet fully resolved in the literature : help with data quality as well as the modeling and monitoring of passengers' temporal habits.

One of the main challenges in data quality is the construction of a robust methodology capable of detecting atypical situations in any context (day of the week, holidays, public holidays, etc.). To this end, two steps were deployed, namely clustering for context estimation and detection of anomalies. The evaluation of the proposed methodology is conducted on a real dataset collected on the Rennes public transport network. By cross-comparing the obtained results with the social and cultural events of the city, it is possible to assess the impact of these events on transport demand, in terms of severity and spatial influence on neighboring stations.

The second part of the thesis focuses on the modeling and the tracking of the temporal activity of passengers. A Gaussian mixture model is proposed to partition passengers into clusters according to the hours they use public transport. The originality of the methodology compared to existing approaches lies in obtaining continuous time profiles in order to finely describe the time routines of each passenger cluster. Cluster memberships are also cross-referenced with passenger data (card type) to obtain a more accurate description of each cluster. The cluster membership over the years has also been analyzed in order to study how the use of transport evolves.

Table des matières

Chapitre 1

Introduction générale

L'analyse de la mobilité est un sujet qui, de tout temps, a généré le plus grand intérêt, à la fois de la part des décideurs politiques qui doivent développer les politiques de mobilité de demain, et des opérateurs de transport qui cherchent à améliorer leur service aux usagers. Étudier la mobilité peut en effet révéler un grand nombre d'information sur l'organisation d'une ville, l'impact de différents événements sociaux ou culturels sur les déplacements des usagers, leurs habitudes de déplacements, ou bien l'évolution de leur utilisation des transports au cours du temps. Le champ des études possibles est vaste et pour les mener plusieurs sources de données sont disponibles : données d'enquêtes, GPS (Global Positioning System), billettiques, etc. Chacune de ces sources est riche en information, mais aucune ne peut suffire en elle-même et fournir toute l'information nécessaire à la compréhension et à l'analyse poussée de la mobilité actuelle. Dans le cadre de cette thèse, nous nous attachons à l'analyse de la mobilité dans les transports en commun avec comme source principale de données les données billettiques. Ce chapitre recense et compare les différentes sources de données actuellement utilisées pour l'analyse de mobilité en mettant en évidence le potentiel des nouvelles sources de données comparées aux données de type enquêtes, tout en montrant les traitements nécessaires préalables à la valorisation de ces données. Un panorama des travaux déjà publiés sur les données billettiques est également réalisé, afin de situer les apports de cette thèse par rapport aux études déjà existantes.

1.1 Motivation

La mobilité est un concept qui, bien qu'existant en tant que tel uniquement depuis la fin des années 1980, a parcouru les âges et évolué en même temps que les technologies. Des premières routes pavées de l'antiquité, aux lignes aériennes internationales actuelles, les infrastructures et les pratiques de mobilité se sont modifiées.

Ces dernières années, nous avons pris conscience de notre impact sur l'environnement et avons entrepris une transition énergétique afin de le limiter. L'augmentation incessante de nos besoins en mobilité n'est alors plus acceptable tel quel et il est nécessaire d'optimiser nos déplacements pour réduire leur influence sur l'environnement. Dans ce contexte les transports en commun semblent être une bonne alternative permettant de combiner une mobilité efficace avec un impact environnemental plus faible que celui généré par une multitude de véhicules individuels.

D'autre part la proportion de la population utilisant les transports étant de plus en plus importante, que ce soit pour aller au travail, sortir ou faire les courses, nos déplacements sont toujours plus nombreux et diversifiés. Il est donc indispensable que les pouvoirs publics ainsi que les opérateurs de transport aient les outils nécessaires à l'amélioration de l'offre de transport et à la mise à l'échelle des réseaux existants (tels que le projet du grand Paris en île de France).

Enfin le concept de ville durable est au cœur des recherches actuelles. Cependant ces villes ne peuvent être envisagées sans une mobilité qui soit elle aussi durable. En cela l'amélioration des transports en commun actuels est un point crucial, que ce soit pour la préservation de l'environnement, pour la santé ou pour le confort de vie.

Pour toutes ces raisons, il est important de poursuivre l'analyse de la mobilité afin de pouvoir offrir de nouveaux outils permettant de mieux la comprendre et de suivre son évolution. À terme ces analyses pourront faciliter la prise de décisions et l'anticipation des problématiques.

1.2 Sources de données pour l'analyse de mobilité

Au cours des dernières années, le nombre croissant de traces numériques collectées reflétant notre mobilité quotidienne a conduit à un renouvellement des approches de l'analyse de la mobilité [Zheng et al. 2014], allant jusqu'à renouveler les enquêtes déplacements classiques elles mêmes. De l'utilisation du GPS [Pang et al. 2013a] aux données de partage de vélo [Ahillen et al. 2015], des quantités croissantes d'informations sont disponibles. Leur traitement et leur analyse génèrent toujours plus de travaux. Dans cette section, plusieurs sources de données, leurs avantages et leurs inconvénients sont comparés.

1.2.1 Données d'enquêtes

Les données d'enquêtes sont une source de données déclaratives (Figure 1.1). Initialement la plus importante source d'information sur l'observation et l'analyse de la mobilité, elles proviennent des enquêtes de déplacements. Une enquête de déplacements est

FIGURE *1.1 – Les données d'enquêtes sont souvent obtenues en remplissant des questionnaires à l'aide des réponses d'usagers.*

un sondage sur le comportement des individus lors de leurs déplacements. La plupart des enquêtes recueillent des informations sur un individu (socio-économique, démographique, etc.), son ménage (taille, structure, relations) ainsi qu'un journal de ses déplacements sur un jour donné (heure de début et de fin de déplacement, lieu, mode de transport utilisé, but du voyage). Les enquêtes majeures sur les déplacements sont menées dans les régions métropolitaines en moyenne une fois par décennie. Certaines régions mènent une enquête par panel, qui entretient les mêmes personnes année après année, pour voir comment leur comportement particulier évolue avec le temps. Pour ce qui est des enquêtes en France, les deux plus importantes sont l'enquête nationale transport (ENT) et l'enquête ménages déplacements (EMD) [CEREMA 2013].

L'ENT, conduite tous les dix ans environ par le ministère chargé des transports et l'Insee (Institut National de la Statistique et des Études Économiques), a pour but de connaître les déplacements des ménages résidant en France ainsi que leur usage des moyens de transport collectifs et individuels [ARMOOGUM et al. 2010]. C'est la seule enquête réalisée à l'échelle nationale décrivant l'ensemble des déplacements et ce quels que soient leur motif, leur longueur ou leur durée, le mode de transport utilisé, le moment de la journée ou la période de l'année. L'EMD, menée par les collectivités locales, sert quant à elle à obtenir une meilleure connaissance des pratiques de déplacements des populations urbaines. Elle porte uniquement sur les jours ouvrables de semaine, hors vacances scolaires, entre octobre et avril. Elle recense les caractéristiques socio-démographiques des ménages (localisation, logement, motorisation, etc.) et des personnes (âge, sexe, profession, etc.) ainsi que la description précise de tous leurs déplacements réalisés la veille du jour d'enquête. Des questions sont également posées au sondé sur ses habitudes d'utilisation des différents modes de déplacements.

Bien que possédant un grand nombre d'avantages (connaissances des motifs de déplacements, données socio-économiques sur les usagers, etc.), ces données possèdent cependant un certain nombre de limitations :

— l'absence de suivi temporel des usagers
— pour l'EMD, l'absence de suivi des déplacements sur certaines périodes de temps
— le biais introduit dans la réponse des sondés, soit par leur absence de réponse, soit pas les imprécisions qu'ils y introduisent

— la vision au jour moyen de semaine qui ne permet pas de capturer la variabilité présente au sein de celle-ci.

Certains travaux tentent de pallier ces limites en complétant ces informations, notamment à l'aide de données GPS [Pham 2016]. Pour cela en plus des questions traditionnelles, un GPS est fourni aux participants afin de collecter des données sur leurs déplacements. Cependant, malgré les nombreux avantages apportés par ces informations complémentaires (diminution du biais de réponse), certains biais ont été décelés. Notamment, l'acceptabilité de ce type d'enquête par les enquêtés (qui peut être considérée comme intrusive) et les erreurs de l'utilisateur (oubli du GPS) peuvent fausser les résultats obtenus.

Enfin plus récemment des enquêtes basées sur les systèmes smartphone ont fait leur apparition. Une enquête pilote smartphone a été entreprise dans la région de Bruxelles-Capitale [STRATEC 2017]. Cette enquête origine-destination a été menée à l'aide d'une application smartphone spécifiquement dédiée aux enquêtes de déplacements. L'objectif de cette enquête était d'explorer la faisabilité technique de cette méthode d'enquête et ses performances. Elle avait également pour but de révéler ses limites et difficultés ainsi que ses avantages et inconvénients par rapport aux méthodes d'enquête classiques.

Zhao et al. 2015 ont présenté dans leurs travaux une analyse exploratoire des résultats obtenus avec une enquête de transport sur smartphone (FMS-Future Mobility Sensing) qui a été testée à Singapour entre 2012 et 2013. Les auteurs disposaient également des résultats d'une enquête transport classique (HITS-Household Inteview Travel Survey). Les auteurs ont appliqué un clustering aux données afin de révéler les habitudes temporelles quotidiennes des sondés et les résultats ont révélé une grande variabilité d'un jour sur l'autre, alors que cette variabilité ne peut pas être capturée par des enquêtes classiques. Les résultats ont également démontré la capacité des enquête FMS à réduire certaines erreurs liées aux sondages, telles que la sous-estimation du nombre de déplacements, la sur-estimation du temps de trajet ou bien les mauvaises heures ou localisations. Enfin ce type d'enquête est plus adapté au suivi de plusieurs jours d'activité.

1.2.2 Données de téléphonie

Les données GSM (Groupe spécial mobile) peuvent elles aussi fournir une bonne source d'information pour l'analyse de la mobilité. Elles sont collectées par les antennes de téléphonie (figure 1.2). Lors de chaque appel passé, l'antenne qui couvre la zone d'appel ainsi que l'antenne de destination de l'appel sont enregistrées. Dans certains pays un identifiant d'utilisateur et la durée de l'appel peuvent également être enregistrés. Une fois ces données collectées, il est alors possible de suivre les flux de mobiles qui se déplacent d'une zone géographique (zone couverte par l'antenne) à une autre et ce même lorsqu'aucun identifiant d'appelant n'est disponible.

Dans sa thèse [Milion 2015] propose des mesures de déplacements réalisées à partir de mesures de signalisation en transit. Le potentiel des sources de données GSM et de radiotéléphonie est mis en avant, notamment sa disponibilité sur l'ensemble du territoire couvert et sa grande représentativité statistique due au fort taux de recouvrement de la population. L'auteur montre que l'analyse de telles données disponibles au niveau de

FIGURE *1.2 – Les antennes captent les données émises par les mobiles présents dans leur zone.*

l'individu permettent d'estimer des flux origine-destination, des indicateurs de qualité de service ou encore de quantifier des facteurs explicatifs de choix de déplacements et de caractériser l'usage du sol.

Comme pour les données d'enquêtes, il est possible de croiser les données de téléphonie avec d'autres sources d'informations. Combinées aux données de téléphonie mobile, leur utilisation peut notamment permettre de détecter les tronçons de routes les plus actifs. WANG et al. 2014 ont développé une approche par clustering basée sur des données SIG et de téléphonie permettant d'obtenir des clusters avec les différentes distributions spatiales aux différentes heures du jour. Les auteurs ont prouvé l'efficacité de leur méthode en démontrant qu'une limitation de la vitesse ou qu'une augmentation de la capacité des routes limite la congestion dans les tronçons détectés. On peut également citer les travaux de [FAROOQ et al. 2010] qui ont pu fournir un outil d'estimation du nombre d'usagers présents dans le bus à chaque arrêt permettant ainsi d'améliorer le système de transport en croisant les données GSM avec les données obtenues à l'aide d'un GPS présent sur les bus.

Enfin, certains travaux ont comparé les résultats obtenus avec différentes sources de données [LENORMAND et al. 2014]. La comparaison des données de téléphonie avec celles de twitter et du recensement est entreprise. Trois aspects sont étudiés : la distribution spatiale de la concentration en population, l'évolution temporelle de la densité en population et les habitudes temporelles des individus. Les auteurs ont ainsi montré que les trois sources de données renvoyaient des informations similaires et pouvaient être interchangeables, bien que la complexité de traitement des données twitter soit plus élevée et leur représentativité plus faible.

1.2.3 Données GPS

Il a été souligné plus tôt que les données GPS peuvent venir en complément des données d'enquêtes, notamment grâce à des enquêtes spécifiques GPS. Cependant, elles forment en elles-mêmes une source d'information très riche. Les données GPS les plus connues du grand public sont les données utilisées par les outils d'aide à la conduite (figure 1.3). Plusieurs types de traces GPS peuvent être utilisés, les traces monomode (lorsque les traces correspondant à un unique mode de transport de transport sont les seules utilisées)

FIGURE *1.3 – Outil d'aide à la conduite ou GPS.*

et les traces multimode (les traces de plusieurs modes sont collectées).

Un type de trace monomode provient des données générées par les taxis ou les véhicules à la demande. L'IPA (Infrastructure Partnership Australia) et UBER ont développé un partenariat afin d'offrir au gouvernement australien un outil permettant de mesurer des indicateurs de mobilité dans différentes villes australiennes (Sydney, Melbourne, Brisbane et Perth) [IPA 2016]. Diverses données générées lors des déplacements UBER ont été collectées, telles que le temps de trajet et les lieux de départ et d'arrivée des courses. Des comparaisons de temps de trajet pendant et hors heures de pointe, ainsi que des études du lien entre départ/fin de courses et station de train peuvent alors être effectuées et donner des indications sur l'état du réseau. Cette métrique, qui est basée sur une collecte d'informations en des points variables grâce aux technologies présentes à bord des véhicules, offre une image plus détaillée des usages de l'ensemble du système routier, que ce qui peut être obtenu avec des données collectées en des points fixes.

Toujours à l'aide des données de course UBER, des analyses ont été menées afin de mieux comprendre les effets de l'interruption du service ferré sur la congestion [UBER 2016]. Celles-ci ont ainsi permis de révéler les zones les plus impactées par une interruption du service.

MOMTAZPOUR et RAMAKRISHNAN 2015 ont proposé une analyse permettant de caractériser les flux de taxi de la ville de New-York. Pour cela, les auteurs ont utilisé les données taxi disponibles en libre accès. Celles-ci contiennent les localisations de début et fin de course, ainsi que le temps de trajet et le prix de la course. Leurs travaux ont permis de développer de nouveaux graphiques de flux probabilistes pour représenter le comportement des flux de trafic. Ils ont également permis une caractérisation des lieux intéressants à l'aide d'une détection d'anomalie sur le graphique de flux.

Pour ce qui est des traces multimodales, les analyses ne se restreignent pas à un unique mode et les données GPS peuvent même être croisées avec d'autres sources de données. FURTLEHNER et al. 2010 ont utilisé une combinaison de données provenant de véhicules en mouvement et de données provenant de capteurs. En appliquant différents outils de data mining sur celles-ci (clustering, analyses statistiques) et en comparant les résultats obtenus avec d'autres méthodes (champs aléatoires de Markov), les auteurs ont pu obtenir

des outils leur permettant de reconstruire et prédire le trafic.

Enfin même si les données ne sont pas toujours disponibles, il est parfois possible d'avoir recourt à des données simulées pour mener des analyses. Des travaux ont été menés afin de suivre l'évolution globale d'un système routier de grande échelle [Moutarde et Han 2011]. Cependant les données étant difficiles d'accès, les auteurs ont travaillé sur des données de trafic simulées à l'aide du logiciel de simulation Metropolis. Ces travaux utilisent la factorisation de matrices non négatives (NMF) et permettent la catégorisation de l'évolution globale journalière du réseau.

1.2.4 Données billettiques

Depuis quelques années déjà de nombreuses villes en Europe et dans le monde disposent d'une source de données pertinente pour analyser l'usage des TC, les données billettiques. Celles-ci sont collectées via des cartes à puce (smartcard en anglais) qui sont identiques en taille à une carte bancaire et utilisées en remplacement (ou en complément) des titres papier (figure 1.4). Elles peuvent permettre de stocker des titres de transports (tickets) et/ou de l'argent (débité en cas de validation).

Figure *1.4 – Carte à puce (carte korriGo) et système de collecte de données billettiques de l'agglomération rennaise.*

Initialement conçues à des fins de tarification et de lutte contre la fraude, les données billettiques peuvent avoir un usage secondaire et servir à mieux comprendre le comportement de mobilité des usagers dans les transports en commun ou à caractériser le fonctionnement du réseau de transport. Elles fournissent en effet des informations riches car continues temporellement, distribuées spatialement et volumineuses (environ 200 000 validations par jour à Rennes). Elles présentent en outre un caractère longitudinale dans la mesure où elles sont collectées sur de longues périodes de temps et permettent de suivre l'évolution des déplacements.

Les premiers travaux à s'intéresser au potentiel des données billettiques [Bagchi et White 2004, 2005] s'interrogent sur le rôle des données billettiques comme nouvelle source d'information pour l'analyse des pratiques de voyage et ont étudié leur potentiel et leur faculté à compléter ou remplacer les données plus traditionnelles (enquêtes, ...). Bien qu'ayant mis en évidence un grand nombre de leurs avantages précédemment cités, ces travaux mettent également en lumière leurs limites. Comme cela a été énoncé, les en-

quêtes de transport fournissent des informations sur les lieux de départ et d'arrivée d'un déplacement, or une partie de cette information pourtant essentielle n'est pas directement captée par la billettique. En effet, les lieux de descente des usagers ne sont pas enregistrés dans la plupart des systèmes de transport collectifs. Le métro de Londres, le RER parisien et le métro de Washington font partie des rares exceptions.

D'autre part, en raison de l'anonymisation des données des usagers, aucune information socio-économique n'est disponible sur ceux-ci en dehors de leur type de titre de transport, ce qui a pour effet de limiter les analyses. De la même manière, aucune information n'est disponible sur le motif du déplacement. Enfin les usagers possédant une carte à puce ne représentent pas toute la population, une partie des usagers voyageant avec des tickets, et ne sont pas nécessairement représentatifs de l'ensemble des usagers. Ils sont par ailleurs sujet à des oublis de carte, générateurs de données manquantes. De plus les données billettiques présentent des problèmes quant à la qualité de mesure (fraude, pannes des valideurs). Les auteurs estiment donc que ces données ne peuvent jouer qu'un rôle complémentaire vis à vis des autres données.

Toutes ces informations manquantes font la faiblesse des données billettiques. De nombreux travaux ont donc été menés afin de trouver des outils pour compenser ces manques. Une partie de ces travaux est recensée dans la section suivante.

1.3 Les données billettiques et leurs applications

Les travaux qui font l'objet de cette thèse se concentrent sur les comportements de mobilité dans les transports en commun observables en utilisant les données de cartes à puce. Bien que le potentiel d'information des données de la carte à puce ait été attesté dans des études antérieures [Bagchi et White 2004 ; Park et al. 2008 ; Utsunomiya et al. 2006], leur nature incomplète (absence de localisation des descentes, manque de données socio-économiques sur les usagers, etc.) continue de motiver une quantité importante de recherches par leur enrichissement et analyse [Pelletier et al. 2011]. Les quatre principaux sujets d'analyse sont présentés ici : (i) l'enrichissement des données, (ii) la prédiction de l'activité, (iii) l'extraction des routines temporelles des usagers, (iv) le suivi temporel.

1.3.1 Enrichissement et mise en forme des données

Estimation des destinations Les systèmes billettiques ne disposent pas, sur la plupart des réseaux de transport en commun, de validation à la descente. Les premiers travaux dans ce domaine se sont donc naturellement intéressés à l'estimation des destinations. En considérant certaines hypothèses sur les comportements des usagers, les premiers travaux proposés dans [Barry et al. 2002] s'attachent à effectuer cette estimation. Pour cela un algorithme basé sur deux hypothèses est développé :

— un grand pourcentage des usagers retourne à la station de destination de son déplacement précédent pour effectuer son déplacement suivant,
— un grand pourcentage d'usagers finit son dernier voyage de la journée à la station où il a commencé son premier déplacement du jour.

Ces hypothèses ont été confirmées pour au moins 90% de la population des usagers du métro à l'aide des informations des journaux de voyage collectées auprès du conseil des transports de la métropole de New-York (NYMTC) [Barry et al. 2002]. Les sorties ont été de plus validées en les comparant aux comptages des sorties de station.

Par la suite, d'autres travaux ont développé une méthode permettant de déduire la destination avec l'origine uniquement [Zhao et al. 2007]. La différence avec l'approche précédente [Barry et al. 2002] vient principalement du fait qu'ils disposent d'une information supplémentaire, l'AVL (automatic vehicle location) c'est à dire la localisation des bus au moment de la validation. Avant de pouvoir estimer les destinations, les auteurs enrichissent leurs données à l'aide des AVL. Ils obtiennent tout d'abord les arrêts à partir des lignes de bus, ils en déduisent ensuite le numéro de bus pour fusionner les données de validation avec celles de localisation des bus. Dans ces travaux, les auteurs utilisent également des traitements spatio-temporels leur permettant d'étudier les proximités entre les différents arrêts. Une fois les données complétées, un certain nombre d'hypothèses sont faites afin d'estimer les destinations, à savoir :

— une grande partie des usagers repart de la station à laquelle son précédent voyage s'est terminé,
— les usagers n'utilisent pas de transport privé entre 2 segments dans une séquence quotidienne,
— les passagers ne vont pas marcher jusqu'à une autre station que celle à laquelle ils sont descendus,
— les usagers finissent leur dernier voyage là où ils ont commencé le premier.

D'autres travaux se sont également intéressés à développer un modèle permettant d'estimer le lieu de destination pour chaque montée dans un bus validée avec une carte [Trépanier et al. 2007]. Les auteurs mettent en avant le fait que les données doivent être attentivement corrigées avant estimation pour éviter des erreurs. Leur étude s'est basée sur les données du STO (Société de Transport de l'Outaouais). 60% des destinations ont ainsi pu être estimées, et même un taux supérieur (80%) en heure de pic. La méthode développée afin de prédire la destination des usagers est basée sur 3 étapes. La première chose à faire est d'étudier l'architecture du système de collecte des données afin d'en tirer les informations les plus intéressantes. Il faut ensuite identifier les objets nécessitant d'être analysés dans le modèle final. Enfin, le modèle analytique est développé pour estimer les lieux. Les systèmes de collecte de données billettiques regroupent ensemble plusieurs objets qui peuvent être identifiés et reliés à la base. Pour cela les auteurs ont utilisé TOOM (Transportation Object-Oriented Modeling, [Trépanier et Chapleau 2001]). Ce système classe les données en 4 méta-classes d'objets : statiques, cinétiques, dynamiques et systémiques. Quatre grands groupes d'objets peuvent être alors décrits : Les objets du réseau (éléments du réseau STO : conducteurs, bus, lignes, arrêts), les objets de l'opération (conducteurs, bus, pièces, garages), les objets administratifs (utilisés pour la carte elle-même) et les déplacements qui caractérisent la demande. Le modèle objet permet d'exprimer les relations entre les éléments des données disponibles. Le modèle d'estimation de la destination du voyage a pour but d'estimer le lieu de descente des usagers pour chaque

Littérature	Mode de transport cible	Taux d'inférence
Barry et al. (2002)	Métro	90%
Zhao et al. (2007)	Métro-Métro Métro-Bus	71.2%
Trépanier et al. (2007)	Bus	66% 80% en heure de pointe
Munizaga et al. (2012)	Bus, Métro	80.77% (Mars 2009) 83.01% (Juin 2010)
Munizaga et al. (2014)	Bus-Métro Métro-Bus	84.2%
Jung et al. (2017)	Bus	87%

TABLE *1.1 – Tableau récapitulatif des taux d'inférence des destinations dans les travaux de la littérature.*

montée et chaque usager. Il se concentre sur les objets usager, voyage, ligne et arrêt.

[MUNIZAGA et PALMA 2012] travaillent sur l'estimation d'une matrice OD (Origine-Destination). Pour l'estimation du point de descente les auteurs s'appuient sur les travaux de [BARRY et al. 2002], [TRÉPANIER et al. 2007] et [ZHAO et al. 2007]. Ils reprennent leurs hypothèses c'est à dire que l'usager retourne en fin de journée au point d'où il est parti en début de journée et que dans la journée il repartira de la station où il est sorti. Ils supposent également que l'usager en correspondance, c'est à dire l'usager ayant utilisé un premier mode de transport et nécessitant d'en utiliser un ou plusieurs autres pour arriver à destination, ne parcourra pas plus de 400 mètres et n'attendra pas plus de 5 minutes. L'information du jour suivant est également utilisée pour l'estimation. La seule variante concerne le lieu d'arrivée du soir qui est estimé en utilisant le lieu de départ du lendemain et non celui du jour même.

Enfin [JUNG et SOHN 2017] proposent une approche ne reposant pas sur les règles précédemment définies mais sur des approches de machine learning. En effet en utilisant une approche de deep learning (réseaux de neurone) s'appliquant sur les validations en entrées et les caractéristiques du territoire, une estimation des lieux de descente est proposée. Les résultats ont pu prouver l'efficacité de cette méthode comparée aux approches habituelles en testant la méthode sur le réseau de Séoul en Corée, qui dispose de validation en entrée et en sortie.

Plus récemment, des méthodes de validation de ces OD reconstituées ont été proposées dans [MUNIZAGA et al. 2014]. Le tableau récapitulatif des résultats obtenus par les auteurs dont les travaux ont été présentés est fourni (table 1.1). Le taux d'inférence présenté correspond à la proportion des destinations qui ont pu être estimées par les auteurs.

Détection des transferts Conjointement à l'estimation des destinations, des procédures d'enrichissement des données ont été développées. Celles-ci sont souvent appliquées en parallèle de l'estimation des destinations et combinent plusieurs approches (détection des transferts plus estimation de la destination).

Plusieurs travaux se sont intéressés à la détection des transferts. Une meilleure identification des transferts, peut permettre de ne pas surestimer les nombres de déplacements et d'obtenir des informations plus complètes à des fins de planification. Dans leur article Chu et Chapleau 2008 ont développé une méthodologie visant à enrichir les données billettiques. La première étape d'enrichissement des données est la reconstruction de l'itinéraire à partir des différentes validations d'un même déplacement. Un simple seuillage est ensuite utilisé pour déterminer si une validation est une correspondance ou non.

Pour mieux identifier les transferts et pour construire des profils spatio-temporels de charge des bus, le chemin spatio-temporel des bus doit être estimé. Pour cela les auteurs prennent en compte toutes les contraintes spatio-temporelles disponibles pour produire un estimateur : heures de départ et d'arrivée, distance entre les arrêts. Ils posent alors plusieurs hypothèses :

— Le départ des véhicules du terminus est celui prévu, à moins que la transaction l'indique autrement.
— L'heure d'embarquement des passagers à l'arrêt suivant sert de borne supérieure à l'estimation de l'arrêt.

En tenant compte des contraintes, une interpolation est utilisée pour estimer les heures d'arrivée aux arrêts de descente où il n'y a pas de montée. On prend donc en compte la première et la dernière transaction à chaque arrêt ainsi que la distance entre les arrêts. On utilise la même méthode pour estimer la position du véhicule entre la dernière validation et le terminus, en supposant que le bus arrive à l'heure prévue au terminus. En revanche, si la vitesse est trop faible ou trop élevée, une extrapolation linéaire est utilisée avec une vitesse moyenne. Cette extrapolation reste limitée par l'heure de départ de la prochaine course du véhicule et l'heure de montée du passager lors de l'arrêt suivant. Il est important de noter que ce système peut être amélioré si on dispose des coordonnées GPS des véhicules.

Pour détecter les transferts, chaque montée est associée à une course et un arrêt. L'heure estimée de descente est alors déduite de celle de la course puis comparée avec les heures de montée de ce même utilisateur. Le temps de transfert pris en compte est celui d'une marche à 1.2m/s entre les arrêts. Un algorithme est également utilisé pour tester si le bus pris était le premier ou non, pour cela de la variabilité est ajoutée à l'heure d'arrivée (plus 5 minutes) afin de prendre en compte les bouchons ou la vitesse de marche. De plus les auteurs font l'hypothèse qu'il y aura toujours de la place dans le bus et qu'il y aura une transaction si et seulement si le bus n'est pas dans la même direction. L'analyse de ces transferts a fait apparaître que 50% des transferts se font dans une durée inférieure à 7 minutes et 80% dans une durée inférieure à 18 minutes. L'analyse du profil de charge des bus permet d'étudier la variabilité dans l'utilisation d'une ligne. De plus, si on combine cette information avec celle des validations des voyageurs on peut obtenir le nombre de kilomètres parcourus par usager.

D'autres travaux sur la détection de transferts ont été menés par Nassir et al. 2015. L'approche développée suit une logique similaire à celle de Chu et Chapleau 2008 avec quelques différences. Les auteurs étudient la différence de temps entre le chemin le plus rapide et le temps de trajet observé, cette différence est appelée "optimalité-off". Cette

mesure sert à étudier la déviation du chemin optimal. La mesure de l'optimalité-off permet de détecter si un changement est un transfert ou une activité pour deux raisons :

— Les usagers sont plus intéressés par leur temps de parcours que par la distance parcourue et auront donc tendance à minimiser le temps d'attente en correspondance.
— L'optimalité-off capture les dépendances aux heures et aux jours du service de transport et peut donc calculer à quel point un itinéraire est raisonnable ou non en termes de temps de trajet. Elle capture également les fréquences des bus, leurs vitesses et le temps d'attente, ce qui ne peut être capturé par des critères uniquement géographiques.

Enfin, Alsger et al. 2015 ont testé les effets des différentes hypothèses posées pour la détection des transferts sur l'estimation des matrices Origine-Destination. Pour cela ils posent différentes hypothèses sur le temps d'attente (variant de 0 à 90 minutes par tranche de 15 min) et la distance de marche (400m, 800m, 1000m et 1100m) entre deux arrêts. Pour chacune de ces hypothèses les matrices OD sont estimées et le taux de transferts calculé. Les résultats ont montré que plus de 90% des usagers marchent moins de 10 minutes, et qu'un changement du seuil de transfert de 15 à 90 minutes n'a qu'un effet mineur. Ils montrent également que la plupart des usagers reviennent en fin de journée dans une zone de 800 mètres autour de leur point de départ du matin.

Inférence du motif du déplacement Comme cela a été mentionné précédemment les données billettiques, contrairement aux données d'enquêtes, ne disposent pas d'information sur le motif du déplacement. Devillaine et al. 2012 ont développé une approche pour détecter et estimer le lieu, l'heure, la durée et le motif du déplacement. Pour cela les auteurs ont utilisé les données provenant de deux villes, Santiago au Chili et Gatineau au Canada. Afin d'inférer le motif de déplacement, il est nécessaire de reconstruire les déplacements et les destinations. La méthode la plus adaptée à chaque ville est alors utilisée ([Trépanier et al. 2007] pour Gatineau et Munizaga et Palma 2012 pour Santiago). Une fois ces informations complétées, un module d'affectation des motifs de déplacements est appliqué. Il prend en compte les caractéristiques du terrain (ville, habitation, travail, etc.), l'arrêt de bus, l'heure de début, la date, la durée ainsi que l'ID d'utilisateur.

Lee et Hickman 2013 ont également proposé une approche permettant d'estimer ces motifs. Pour cela les auteurs ont développé une approche basée sur des règles heuristiques (si-alors) à partir des information personnelles, spatiales et temporelles qu'ils avaient à disposition pour chaque usager. Ces règles permettent de construire un jeu de données test des motifs de déplacements. Une classification basée sur des arbres de décision est alors conduite sur ce jeu test afin de déterminer la performance du modèle. Enfin les résultats sont comparés avec des résultats d'enquêtes pour déterminer la cohérence avec les motifs de déplacements connus.

La plupart des études menées sur les données billettiques combinent plusieurs de ces approches (par exemple, estimer la destination et détecter les transferts [Munizaga et Palma 2012 ; Zhao et al. 2007]).

1.3.2 Prédiction de l'activité

Le deuxième sujet qui a été étudié est lié à l'utilisation de données de carte à puce pour générer des informations sur les pratiques de voyage des usagers et prédire leurs déplacements, la congestion ou encore les matrices OD.

Certains travaux se sont concentrés sur la prédiction en se basant sur l'activité des usagers. Lathia et al. 2010 étudient les données de voyage individuelles sur les métros de Londres afin d'offrir des services personnalisés. Les auteurs estiment les déplacements personnels et développent une méthode de prévision des heures de voyage personnalisées pour les usagers. Cette méthode classe les stations en fonction des déplacements futurs de chaque usager. L'approche proposée par [Foell et al. 2014] est appliquée à un réseau de bus et est également utilisée pour prédire les déplacements. Elle se base pour cela sur les habitudes de l'ensemble des usagers mais également sur leurs habitudes individuelles. En comparant plusieurs méthodes, les auteurs montrent que la combinaison de ces deux types d'information offre de meilleurs résultats.

D'autres travaux ont quant à eux développé des approches pour l'analyse et la prédiction de la congestion. L'étude de la congestion peut par exemple servir à prioriser la maintenance des routes ou à améliorer le placement des arrêts de bus pour éviter les bouchons lors de leurs arrêts. En utilisant les données de la carte à puce, Ceapa et al. 2012 concentrent leurs travaux sur la congestion du trafic et établissent qu'il existe une certaine régularité spatiale et temporelle dans la congestion qui facilite la prévision. Ces auteurs montrent également que la congestion se résorbe rapidement. Dans leur article Fuse et al. 2010 développent des méthodes d'analyse détaillées des pratiques des usagers. Pour cela les données de bus de Tokyo sont utilisées. Les auteurs analysent les origine-destination, estiment le temps moyen de trajet en bus, et cherchent le lien entre temps de trajet et congestion du trafic (heure du trajet).

La prédiction de matrice OD et de flux de voyageur a également été étudiée. Li et al. 2017 effectuent une prédiction court-terme des flux dans le métro. Pour cela une approche différente de ce qui est traditionnellement proposé est présentée. Au lieu de s'intéresser aux demandes régulières de transport, l'impact d'événements atypiques sur le réseau est étudié, permettant ainsi de prédire les fluctuations irrégulières dans l'activité. Toqué et al. 2016 ont également développer une approche utilisant des réseaux de neurones récurrents pour la prédiction court terme (15 minutes) des matrices OD sur la ligne de RER A de Paris.

1.3.3 Extraction de routines temporelles des usagers

L'utilisation des enregistrements de validations pour l'analyse de mobilité dans les transports en commun, et plus particulièrement de l'analyse temporelle de mobilité, a généré un grand nombre de recherches. Une partie de ces travaux portent sur les habitudes temporelles de déplacements des usagers sur un réseau de transport en commun. Ces études cherchent en particulier à identifier les facteurs extérieurs qui peuvent influencer l'utilisation du réseau, tels que la météo [Arana et al. 2014], les dynamiques spatio-temporelles [Tao et al. 2014], et/ou l'évolution de l'activité des usagers au cours du temps [Chu 2015].

En appliquant des outils de data mining aux données billettiques, plusieurs aspects de la mobilité dans les transports en commun ont pu être étudiés. Le partitionnement des usagers en groupes homogènes au sens de leurs habitudes temporelles est un des volets permettant d'extraire de l'information sur la mobilité, et notamment sur leur régularité et leurs trajets récurrents. La première analyse par clustering est présentée [Morency et al. 2006]. Les auteurs étudient la régularité des déplacements des usagers en agrégeant les transactions appartenant à une même carte en un profil journalier, chacun indiquant l'intervalle de temps (càd l'heure) auquel l'usager a effectué au moins une validation. En utilisant une approche k-means les auteurs identifient alors des clusters de jours similaires en fonction des heures de validation. Une analyse similaire est menée sur les comportements hebdomadaires dans [Agard et al. 2006] : les voyages sont agrégés en profils qui résument l'activité des jours de la semaine des usagers et un clustering hiérarchique, ainsi qu'un k-means sont appliqués afin d'étudier les habitudes des groupes. Dans [Ma et al. 2013], en vue de révéler les trajets récurrents, un DBSCAN est appliqué aux chaînes de déplacements individuelles. Un k-means++ (méthode d'initialisation du k-means afin d'obtenir une erreur et un temps de calcul plus faibles) est également utilisé pour regrouper les usagers en fonction de leur régularité.

Certaines méthodes de machine learning plus avancées ont également été développées dans des travaux récents, notamment pour avoir une description plus précise des stations avec un usage similaire. Par exemple, la NMF (Nonnegative Matrix Factorization) est utilisée dans [Poussevin et al. 2014] pour découvrir un dictionnaire d'atomes de comportement permettant de décrire les usagers en se basant sur leurs validations quotidiennes dans le métro. La distribution de ces atomes sur les stations est alors utilisée pour effectuer un clustering multi-échelles et obtenir des groupes de stations avec des habitudes similaires. Dans l'approche décrite par [El Mahrsi et al. 2014], les chaînes de déplacements individuelles sont agrégées en profil hebdomadaire d'usager, chacun comprenant le nombre de voyages d'un usager sur une plage temporelle donnée, pour chaque jour de la semaine. Un modèle de mélange d'unigrames est alors estimé sur les profils temporels dans le but d'obtenir des clusters d'usagers possédant des habitudes temporelles de déplacements similaires.

Des travaux plus récents [Zhong et al. 2016] ont cherché à évaluer à quelle échelle la variabilité qui est détectée peut être considérée comme stable, explicable et durable. Pour cela plusieurs mesures ont été proposées, afin de pouvoir définir la stabilité de la régularité des usages en s'intéressant plus particulièrement aux changements d'échelle temporelle. Pour évaluer leurs résultats, les auteurs ont appliqué leur méthode sur trois villes : Londres, Beijing et Singapour.

1.3.4 Suivi temporel

Des travaux se sont penchés sur un jeu de données de 5 ans afin d'étudier la durée pendant laquelle les cartes restent actives [Trépanier et al. 2012]. [Langlois et al. 2016] ont effectué une analyse sur plusieurs semaines des habitudes temporelles en utilisant le clustering. Les auteurs ont proposé une représentation des séquences d'activités longitu-

dinales en utilisant une activité spatio-temporelle. Pour cela ils ont segmenté l'espace en différentes zones de validation. Chaque usager est ainsi caractérisé par ses heures de déplacements, ainsi que la zone dans laquelle il se situe en fonction de l'heure de la journée. Les groupes créés par le clustering sont associés avec des structures de séquences distinctes, permettant ainsi une meilleure connaissance des 4 semaines d'activité des usagers.

1.4 Contributions

Dans la section précédente, un état de l'art des travaux effectués pour l'analyse de mobilité dans son acceptation générale, ainsi que de la mobilité dans les transports en commun à l'aide des données billettiques a été présenté. Cette section a pour but de situer cette thèse par rapport à l'état de l'art. Le choix des données billettiques en tant que source d'analyse de la mobilité dans les transports en commun est justifié. Les verrous scientifiques qui ont été traités sont développés.

1.4.1 Choix des données billettiques

Différentes sources de données ont été abordées. Que ce soit les données d'enquêtes, les données GSM qui permettent un suivi de la mobilité d'une masse de personnes plus importante mais pas à l'échelle de l'usager ou bien les données billettiques, chaque source a ses richesses et ses faiblesses. Cette thèse se plaçant dans le cadre de l'analyse de la mobilité dans les transports publics, les données billettiques semblent être une bonne source de données pour effectuer un suivi rapproché de l'activité des usagers, mais également de l'activité sur l'ensemble du réseau. Elles offrent un autre regard sur la mobilité puisqu'elles permettent une étude fine de la variabilité. Les enquêtes traditionnelles se plaçant du point de vue de l'usager, aucune information sur le réseau, ses dysfonctionnements et ses spécificités ne sont disponibles, de plus elles fournissent uniquement une vision du jour moyen de semaine. Les données GSM n'offrent quant à elles aucune information permettant un suivi plus resserré de l'usager, puisque ces données ne suivent en général pas les usagers mais les flux.

D'autre part, bien que les données billettiques ne disposent pas d'information sur les lieux de descente, les nombreux travaux effectués sur celles-ci permettent de les estimer. De nouvelles études paraissent chaque jour afin d'estimer les différentes informations manquantes (lieu de destination) [Devillaine et al. 2012]. Dans un tel contexte, les données billettiques semblent être une source de données riche capable de pallier certaines de ses faiblesses. Il paraît donc cohérent de travailler sur elles dans le cadre de cette thèse.

1.4.2 Verrous scientifiques

Cependant bien que de nombreux travaux se soient déjà penchés sur l'enrichissement de ces données, ainsi que sur leur utilisation, de nombreux points restent encore à développer.

Comme les différents travaux effectués sur les données billettiques ont pu le montrer, l'enrichissement de ces données reste une tâche prépondérante. En effet, celui-ci permet de reconstituer une partie des informations manquantes (destinations, profil de charge,

etc.). Il peut également, par des croisements avec d'autres sources de données, fournir de nouvelles informations. Cependant les travaux menés sur ces données, que ce soit pour l'analyse de l'utilisation ou la prédiction de l'affluence dans les transports en commun sont encore perfectibles. La variété des analyses pouvant être conduites nécessite de nouveaux travaux. Des analyses à une échelle plus fine peuvent également être menées. Que ce soit pour l'enrichissement ou l'analyse des données billettiques, une base de données fiable est nécessaire afin de pouvoir travailler, or ce n'est pas toujours le cas. Les problèmes de remontées d'information par certains valideurs, les événements culturels ou sociaux générant une activité atypique ou encore les dysfonctionnements sur le réseau peuvent fournir des données biaisées. Des travaux permettant l'identification et l'analyse de telles données sont donc nécessaires pour améliorer la fiabilité des autres travaux également menés sur ces données.

Cette thèse vise à proposer de nouvelles approches permettant d'enrichir l'information fournie par ces données. Les contributions apportées par cette thèse sont les suivantes :

Une approche de détection des événements atypiques se déroulant sur le réseau de transport en commun de la ville de Rennes a été développée. La mise en qualité ainsi qu'une première exploration succincte est bien souvent la première étape lors de la prise en main d'un jeu de données billettiques. Celle-ci vise à détecter d'éventuels problèmes de mesure afin de nettoyer le jeu de données de leurs effets ainsi qu'à repérer d'éventuels comportements anormaux au sein de ce jeu de données. Pour aider à réaliser aisément cette étape, nous proposons une approche de détection des événements et mesures atypiques enregistrées par le réseau. L'activité sur le réseau dépendant du type de jour (forte affluence pour un jour de semaine, faible affluence pour les dimanches et jours fériés), cette approche doit prendre en compte un certain nombre de variables de contexte. La méthodologie proposée se décompose donc en deux étapes, la première s'attache à sélectionner automatiquement les variables de contexte, et la deuxième détecte à proprement parler les anomalies pour chaque contexte précédemment défini.

Une méthodologie de classification des usagers liée à leur utilisation temporelle des transports est également proposée. Le but est de développer des outils d'analyse centrés sur les usagers. Celle-ci modélise de façon continue les habitudes temporelles des usagers et les classe dans différents clusters en concordance avec leur activité. Cette répartition nous a également permis d'entreprendre un suivi longitudinal de l'évolution de la classification dans le temps. Les usagers changeant de clusters d'une année sur l'autre et donc changeant de comportement dans les transports en commun, ont aussi pu être analysés.

Une étude poussée et une interprétation des résultats obtenus sur données réelles sont présentées pour chacune des approches mises en place.

1.5 Plan de la thèse

La thèse se découpe en 3 chapitres.

— Le chapitre 2 présente les différents jeux de données utilisés et les villes dans lesquelles ils ont été collectés. Une analyse descriptive de ces différents jeux de données est également présentée.

— Le chapitre 3 définit l'approche méthodologique permettant de détecter les événements atypiques ayant lieu sur le réseau, mais également les défaillances des systèmes de validation, ceci afin de ne pas prendre en compte des données erronées lors d'analyses ultérieures. Les résultats obtenus sur la ville de Rennes sont analysés et croisés avec les informations dont nous disposons sur les événements sociaux et culturels ayant lieu dans la ville.
— Le chapitre 4 établit une méthodologie de classification des usagers dépendant de leur utilisation temporelle des transports. Elle comporte également une comparaison des résultats obtenus sur deux villes différentes, mettant ainsi en relief les différences de chacune et prouvant la réplicabilité du modèle sur d'autres données. Enfin un suivi de l'évolution de l'utilisation des transports par les usagers est également entrepris en utilisant la classification précédemment définie.

1.6 Publications

Journaux internationaux

[1] A.-S. Briand, E. Côme, M. K. El Mahrsi et L. Oukhellou, "A mixture model clustering approach for temporal passenger pattern characterization in public transport," dans *International Journal of Data Science and Analytics* vol. 1, nb 1, pp. 37-50, 2016.

[2] A.-S. Briand, E. Côme, M. Trépanier et L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," dans *Transportation Research Part C : Emerging Technologies*, vol. 79, pp. 274-289, 2017.

[3] A.-S. Briand, E. Côme et L. Oukhellou, "Automatic detection of atypical events on a public transport network using smart card data," dans *IEEE Intelligent Transportation Systems Transactions*, en cours de soumission.

Chapitres d'ouvrage

[4] M. K. El Mahrsi, A.-S. Briand, E. Côme et L. Oukhellou, "Utilité des données billettiques pour l'analyse des mobilités urbaines : le cas rennais," dans *Données urbaines* ECONOMICA, 2015.

[5] A.-S. Briand, E. Côme, N. Coulombel, M. K. El Mahrsi, E Munch, C. Richer et L. Oukhellou, "Projet MOBILLETIC Données billettiques et analyse des mobilités urbaines : le cas rennais.," dans *Big data et politiques publiques dans les transports* (C. Vrain, A. Péninou, and F. Sedes, eds.), ECONOMICA, 2017.

Conférences internationales avec acte

[6] A.-S. Briand, E. Côme, M. K. El Mahrsi et L. Oukhellou, "A mixture model clustering approach for temporal passenger pattern characterization in public trans-

port," dans *IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015* (Paris, France), pp. 1-10, DSAA, 2015.

[7] A.-S. Briand, E. Côme, M. K. El Mahrsi et L. Oukhellou, "Classification à base de modèle de mélange pour l'identification de profils temporels types d'usagers de transport public," dans *AAFD & SFC'16 - Francophone International Conference on Data Science*, Marrakech, Maroc, 2016.

Conférences internationales sans acte

[8] M. Trépanier, A.-S. Briand, E. Côme et L. Oukhellou, "Variations annuelles des comportements des usagers de transport collectif," dans *51e congrès de l'association québécoise des transports*, Québec, Canada, 2016.

[9] A.-S. Briand, E. Côme, M. Trépanier et L. Oukhellou, "Understanding temporal passenger habits in public transport by smart card data analysis," dans *Young Researcher Seminar*, Berlin, Allemagne, 2017.

[10] A.-S. Briand, E. Côme, M. Trépanier et L. Oukhellou, "Smart Card clustering to extract typical temporal passenger habits in Transit network. Two case studies : Rennes in France and Gatineau in Canada," dans *3rd International Workshop and Symposium : "Research and applications on the use of passive data from public transport"*, Santiago, Chili, 2017.

Chapitre 2

Données

Dans ce chapitre les deux cas d'étude utilisés tout au long de la thèse, Rennes en France et Gatineau au Canada, sont décrits. Le but est d'offrir au lecteur une meilleure connaissance de ces villes, de leur agencement spatial, mais aussi des statistiques exploratoires sur les données de transports utilisées. Cela facilitera la compréhension des résultats développés dans la suite de ce manuscrit.

Chacun des deux cas d'étude donnera lieu à une rapide présentation du réseau de transports en commun, puis à quelques statistiques descriptives de son utilisation. Une première exploration de suivi temporel de l'activité est également entreprise sur le cas d'étude de Gatineau qui permet une profondeur temporelle plus importante.

2.1 Cas d'étude : réseau de transport en commun de la ville de Rennes en France

Le premier cas d'étude, le plus fréquemment utilisé dans cette thèse, est celui de la ville de Rennes. Rennes est une commune située dans l'ouest de la France et d'une population de 213 454 habitants en 2014. Elle est le chef-lieu de la région Bretagne mais aussi la huitième ville universitaire française en 2016 avec près de 66 000 étudiants. Au premier janvier 2015, la ville de Rennes et les 42 autres communes de l'agglomération rennaise ont été regroupées au sein d'une communauté d'agglomération sous l'appellation Rennes métropole. Rennes métropole comptait une population de 426 502 habitants en 2013.

2.1.1 Présentation du réseau de transport

le réseau de transport public STAR (Service de Transport en Commun de l'Agglomération Rennaise) de Rennes Métropole assure la desserte de l'ensemble de l'agglomération rennaise. Ce réseau comprend une ligne de métro (ligne A), ainsi que 149 lignes de bus. La ville de Rennes a pour particularité d'être polycentrique, si la majorité de l'activité rennaise se concentre autour du centre-ville situé à la station *République*, elle dispose d'une multitude de petits centres dans les communes avoisinantes. L'ensemble des lignes de bus forme un réseau en étoile, chaque ligne partant d'un de ces centres et se terminant à République ou à proximité de l'une des 14 autres stations de la ligne de métro A. Cette répartition spatiale est visible sur la figure 2.1, qui représente le nombre de validations enregistrées sur le réseau rennais le lundi 2 juin entre 7H et 8H. Les stations les plus actives sont en effet celles placées dans le centre de Rennes et le long de la ligne de métro, mais des centres secondaires répartis tout autour dans les différentes communes de l'agglomération peuvent également être observés.

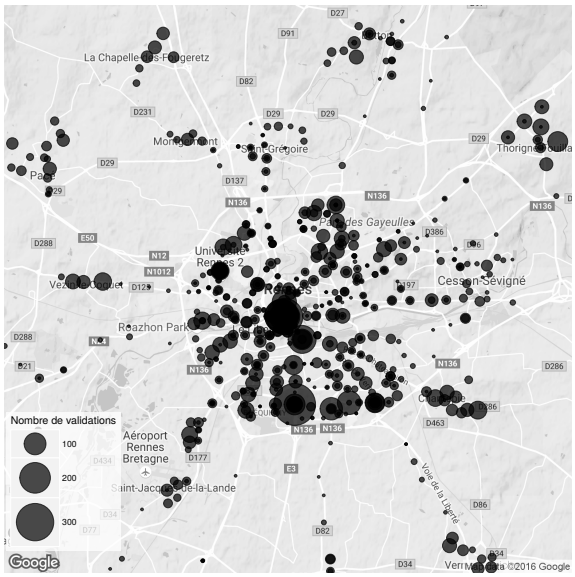


FIGURE *2.1 – Nombre de validations enregistrées sur le réseau Rennais le lundi 2 juin 2014 entre 7H et 8H.*

En 2006, la ville a introduit une carte à puce (carte KorriGo) que les usagers peuvent

utiliser pour se déplacer sur le réseau STAR. Les usagers doivent valider leur carte uniquement lorsqu'ils embarquent dans un bus ou entrent dans une station de métro (par conséquent, les lieux de descente ne sont pas collectés). Le système AFC (Automated Fare Card - validateurs) enregistre ces validations. Chaque enregistrement d'une validation contient un identifiant unique de carte anonymisé, l'heure (date et heure précise), un emplacement (le bus embarqué ou la station de métro) et le type de tarif de la carte à puce.

Le réseau STAR dispose d'un très grand nombre de titres de transports différents, ceux-ci allant du simple ticket aux abonnements couplés avec les lignes à grande vitesse. Sur l'année 2014, 108 titres différents ont été observés. Le nombre de validations mensuel pour les différents titres révèle que la majorité des validations est effectuée, par ordre décroissant, avec les tickets, les abonnements standards (abonnés -20 ans, 20-26 ans et 27-64 ans) et les abonnements subventionnés (gratuité). Les autres titres bien que nombreux restent donc minoritaires. Pour des raisons de lisibilité et à cause du grand nombre de titres, le panel a été réduit en classant les usagers en 8 catégories en fonction de leur titre : les abonnés, les abonnés jeunes, les abonnés âgés, les abonnés de courte durée (pass 1 à 7 jours), les agents Keolis Rennes (KR), les subventionnés (personnes bénéficiant de la gratuité pour raison sociale ou scolaire) et les usagers utilisant des tickets.

Cette thèse utilise un jeu de données collectées sur le réseau STAR et enregistré sur la période allant d'avril 2014 à mars 2016. Il permet d'illustrer l'application des différentes méthodes proposées. Il est, en fonction de son application, utilisé dans son entièreté (2 ans de données) ou uniquement sur la période du mois d'avril 2014.

2.1.2 Prétraitement des données

Afin de mieux appréhender la suite de la thèse, une distinction doit être faite entre deux notions : la notion de segment de déplacement, correspondant à une validation, et celle de chaîne de déplacement. Une validation constitue l'enregistrement du passage d'un usager à un instant t. Un déplacement est défini par l'ensemble des segments de déplacement d'un usager, de son point d'origine à sa destination finale, et contient de potentielles correspondances. Le déplacement est donc constitué d'une ou plusieurs validations. Il est important de noter que, comme dans la plupart des réseaux de transport urbains Français, aucune information sur la destination des déplacements n'est disponible dans ce jeu de données brutes puisque l'usager n'a pas à valider en sortie du réseau. Or, cette information s'avère être essentielle pour pouvoir reconstituer les déplacements et effectuer ensuite des traitements avancés. Il est primordial d'utiliser le déplacement et non pas les segments de déplacement et ce, afin de ne pas introduire de biais dans les résultats (les usagers dont les déplacements se composent d'une ou plusieurs correspondances pouvant, dans le deuxième cas, être perçus à tort comme ayant une utilisation plus fréquente des transports que les usagers dont les déplacements ne contiennent pas de correspondances).

Les déplacements sont reconstruits en utilisant une approche en deux étapes similaires aux travaux antérieurs proposés dans Gordon et al. 2013 ; Munizaga et Palma 2012 ; Trépanier et al. 2007 ; Zhao et al. 2007. La première étape consiste à déduire

l'emplacement des destinations de chaque transaction en fonction de deux hypothèses :

— l'hypothèse d'arrêt le plus proche : pour une transaction donnée, l'usager s'arrête probablement à l'arrêt ou à la station la plus proche de l'endroit où se déroule sa prochaine transaction
— l'hypothèse de symétrie quotidienne : pour la dernière transaction du jour où l'usager s'arrête à l'arrêt ou à la station la plus proche du lieu où s'est déroulée sa première transaction de la journée.

Pour chaque transaction, les distances entre les différentes stations de descente possible et l'emplacement d'embarquement de la transaction suivante sont inspectées. La station la plus proche de la prochaine transaction est retenue comme destination candidate de la transaction en cours. Si la destination candidate et l'emplacement d'embarquement suivant se trouvent à une distance de marche raisonnable (fixée à 500 m dans ce cas d'étude), la destination candidate est affectée comme lieu de descente de la transaction en cours. Sinon, l'inférence échoue et aucune destination n'est affectée. Le même processus est appliqué à la dernière transaction de la journée à l'exception de l'utilisation de l'emplacement d'embarquement à partir de la première transaction du jour au lieu de la transaction ultérieure. En outre, un temps d'arrivée d'estimation (allongement) est attribué à chaque transaction pour laquelle un emplacement de descente a été estimé. Déterminer si les transactions sont des transferts ou non est effectué dans un deuxième temps. Là encore, les transactions de chaque usager sont inspectées séquentiellement. Pour qu'une transaction soit marquée comme un transfert, les conditions suivantes doivent être remplies :

— l'emplacement de descente de la transaction précédente a été déduit avec succès
— la connexion entre les deux transactions (c'est-à-dire le temps écoulé entre l'estimation de l'heure d'arrivée dans la transaction précédente et le temps d'embarquement de la transaction en cours) se produit dans un délai de 30 minutes.

Sinon, la transaction est marquée comme un premier embarquement afin d'indiquer le début d'un nouveau déplacement. En outre, les transferts le long de la même ligne sont interdits : si deux transactions consécutives sont effectuées sur la même ligne, la seconde est automatiquement marquée comme première embarcation, même si les deux conditions sont remplies.

En raison de l'absence d'informations de vérité terrain sur les destinations et le comportement de transfert dans l'ensemble de données, l'efficacité de l'approche d'enrichissement des données appliquée et son influence sur les analyses postérieures effectuées sur les données ne peuvent être évaluées directement. Néanmoins, des études récentes où de telles informations sont disponibles Alsger et al. 2015 ; Gordon et al. 2013 ; He et al. 2015 ont démontré la robustesse de ces approches. He et al. 2015 s'intéressent à la calibration de la distance de transfert, c'est à dire la distance séparant l'arrivée d'une validation et le départ d'une autre. Un seuil doit être défini, toute distance inférieure à ce seuil signifiant qu'une correspondance est possible. Pour ce faire, différentes valeurs de seuils ont été testées. Une baisse de la précision étant observée avec une augmentation de la distance seuil, les auteurs ont choisi le seuil le plus adapté pour améliorer la précision. De même Alsger et al. 2015 cherchant à définir la distance de marche pendant le transfert, mais

également le temps de transfert. Le nombre de transferts effectués par déplacement est analysé. Ce nombre de transferts n'augmentant pas au-delà d'un certain seuil, celui-ci est retenu comme distance de marche. Dans notre cas, le seuil de distance de marche raisonnable (500 m) et le seuil de temps (30 minutes) ont été résolus lors de la discussion avec des experts de l'opérateur de transport public STAR et sont en accord avec les valeurs de seuil généralement utilisées dans la littérature. La figure 2.2 illustre les distributions des temps de transfert pour chacun des trois types de transfert possibles dans le réseau STAR. La plupart des temps de transfert se produisent avant le seuil de 30 minutes par définition (95% des transferts se produisent dans les 25min), les transferts au métro se produisant dans une période plus courte (90% se produisent dans les 10min) par rapport aux transferts au bus.

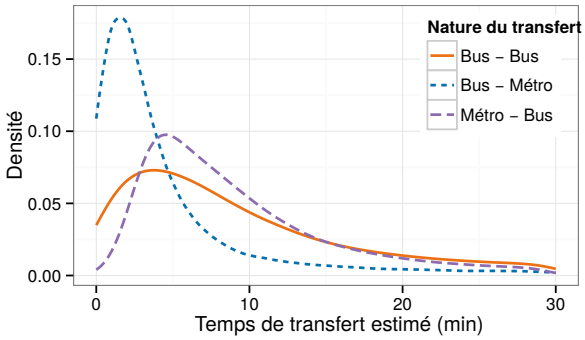


FIGURE *2.2 – Densités des temps de transfert estimés pour chaque type de transfert (bus-bus, métro-bus et bus-métro).*

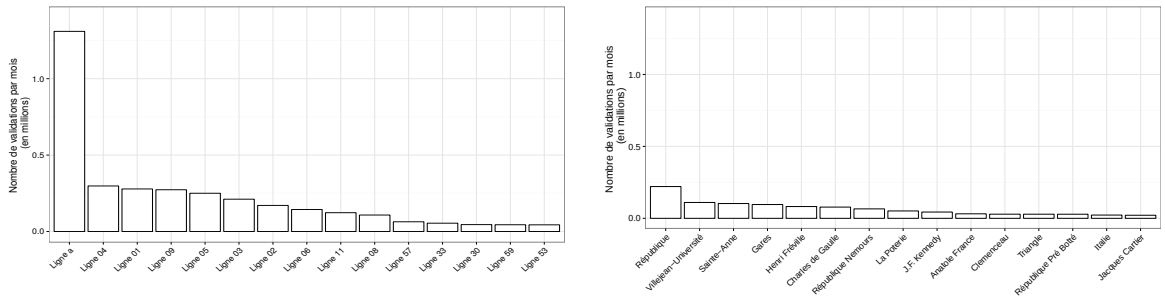
À la lumière de ces résultats, qui sont cohérents avec ceux rapportés dans la littérature, nous sommes confiants dans les valeurs de seuil choisies et nous nous attendons à ce que les modifications légères n'aient qu'un impact marginal sur les résultats du regroupement.

2.1.3 Le réseau rennais

Sur un mois de données (avril 2014), 5 404 096 validations sont enregistrées, parmi lesquelles plus de 80% ont été réalisées par environ 135 000 cartes à puce.

Sur une période d'analyse allant d'avril à octobre 2014, la ligne la plus utilisée est l'unique ligne de métro, ou ligne A. Le nombre de validations effectuées sur cette ligne est presque cinq fois supérieur au nombre de validations effectuées sur la seconde ligne la plus fréquentée (figure 2.3a). Considérons maintenant la fréquentation des différentes stations du réseau STAR en différenciant les stations par leur nom (figure 2.3b). Ceci permet de regrouper les stations bus et métro se trouvant au même endroit sous la même étiquette. La station République concentre la majorité des validations en raison de la forme en étoile du réseau de transports avec pour centre la station République. On notera qu'un petit groupe de stations reçoit un grand nombre de validations et celles-ci correspondent presque toutes à des stations de métro.

Un autre point qui mérite d'être examiné concerne la répartition géographique de ces stations. Sur la figure 2.4, on notera le grand nombre de validations de la station



(a) Lignes les plus fréquentées *(b) Stations les plus fréquentées*

FIGURE *2.3 – Nombre moyen de validations par mois pour les 15 lignes les plus fréquentées (a) et les 15 stations les plus fréquentées (b).*

République. La station Villejean Université, bien qu'excentrée attire, elle aussi, un grand nombre d'usagers en raison du caractère "ville étudiante" de Rennes. Enfin, les stations au sud du centre-ville attirent elles aussi de nombreux usagers. Ces stations correspondant en effet à des ensembles résidentiels (Triangle, Italie, ...), leur forte fréquentation est sans doute liée au fait qu'il s'agit aussi d'importantes stations de rabattement de bus.

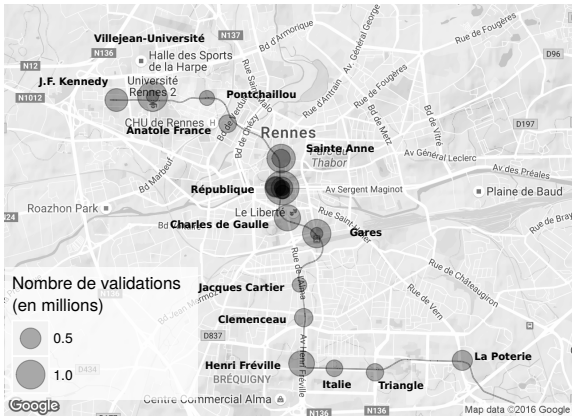
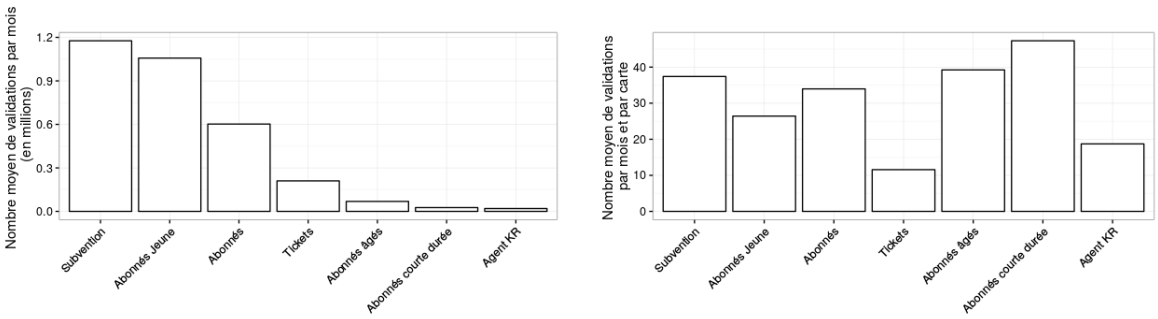


FIGURE *2.4 – Carte du réseau représentant les 15 stations bus et métro les plus actives (Nombre de validations enregistrées sur la période Avril-Octobre 2014, 7 mois).*

La figure 2.5a fait apparaître clairement que les validations proviennent essentiellement des subventionnés, des abonnés jeunes, des abonnés et des détenteurs de tickets. Rennes étant une ville connue pour sa forte population estudiantine, il n'est pas étonnant de retrouver un nombre aussi élevé de validations d'abonnement de type jeune. La figure 2.5b permet quant à elle d'observer le nombre de validation relativement au nombre de cartes actives pour chaque type de titre. Il est ainsi possible de voir que certains types de cartes, tels que *Abonnés courte durée* ou *Agent KR*, génèrent un nombre plus important de validations par carte.

Le nombre de segments de déplacements par heure agrégé pour chaque jour de la semaine du 7 avril au 14 avril est indiqué dans la Figure 2.6. On retrouve le motif typique d'une semaine de travail avec un motif similaire pour les jours de semaine, une baisse d'activité le samedi et un usage plus diffus le dimanche. Un motif différent est observable



(a) Nombre de validations *(b) Nombre de validations par carte*

FIGURE *2.5 – Nombre moyen de segments de déplacement enregistrés par mois pour les différents types de titre de transport.*

le mercredi avec une augmentation des validations de midi. Cette activité élevée s'explique par le fait que les écoliers et les lycéens en France n'ont souvent pas de cours le mercredi après-midi. La journée avec le plus d'activité est mardi. Certains jours (jeudi, vendredi et dimanche), une activité de nuit peut également être observée avec des segments de déplacements enregistrés à la fin et au début du service.

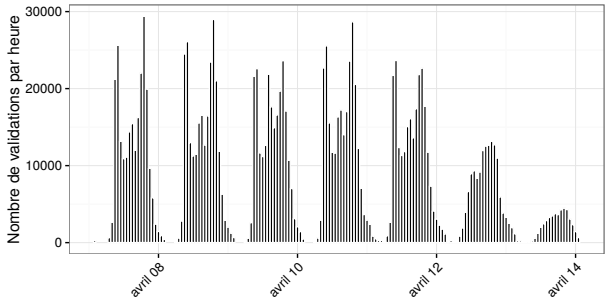
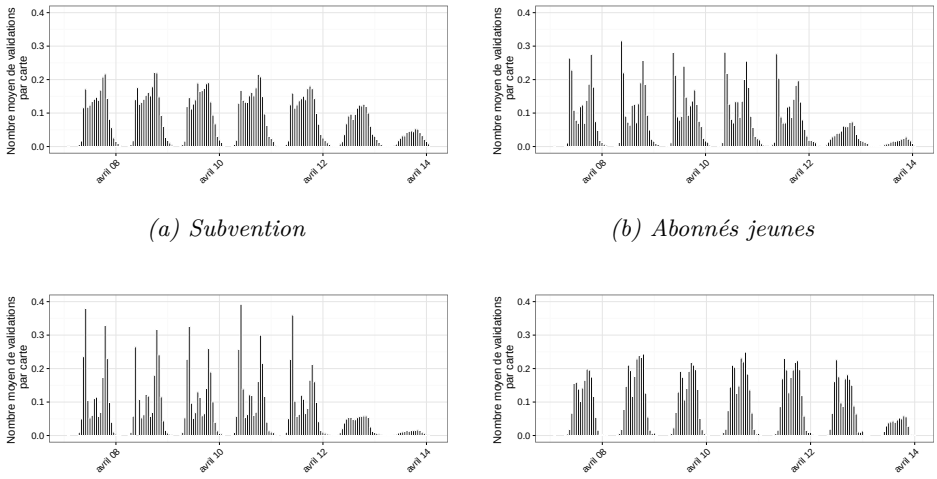


FIGURE *2.6 – Nombre de segments de déplacements par heure enregistrés sur le réseau STAR pendant la semaine du 7 avril 2014*

Mais derrière cette régularité apparente de l'activité d'un jour à l'autre en semaine, se trouve une multitude de comportements différents. Les titres de transport offrent un premier outil afin d'analyser les différences d'activité d'un groupe d'usagers à l'autre. Le nombre moyen de validations par carte enregistré sur une semaine est représenté sur la figure 2.7 pour les 4 types de titres de transports les plus utilisés.

Ce découpage plus fin de l'activité révèle des différences d'activité d'un groupe à l'autre. En effet certains groupes, comme les abonnés jeunes vont générer plus de validations, tandis que d'autres, comme les abonnés âgés vont en générer moins. Ensuite il peut être observé que le motif à trois pics vus précédemment n'apparaît pas pour toutes les populations. Les abonnés jeunes, mais surtout les abonnés, sont des populations qui ont des comportements plus pendulaires que les autres.

Toutefois, même si l'utilisation des titres de transport permet de révéler un usage plus fin des transports en commun, elle masque certaines différences d'activité présentes au sein même de ces groupes. Pour cela il est nécessaire d'effectuer une analyse plus poussée



(a) Subvention *(b) Abonnés jeunes*

(c) Abonnés *(d) Abonnés âgés*

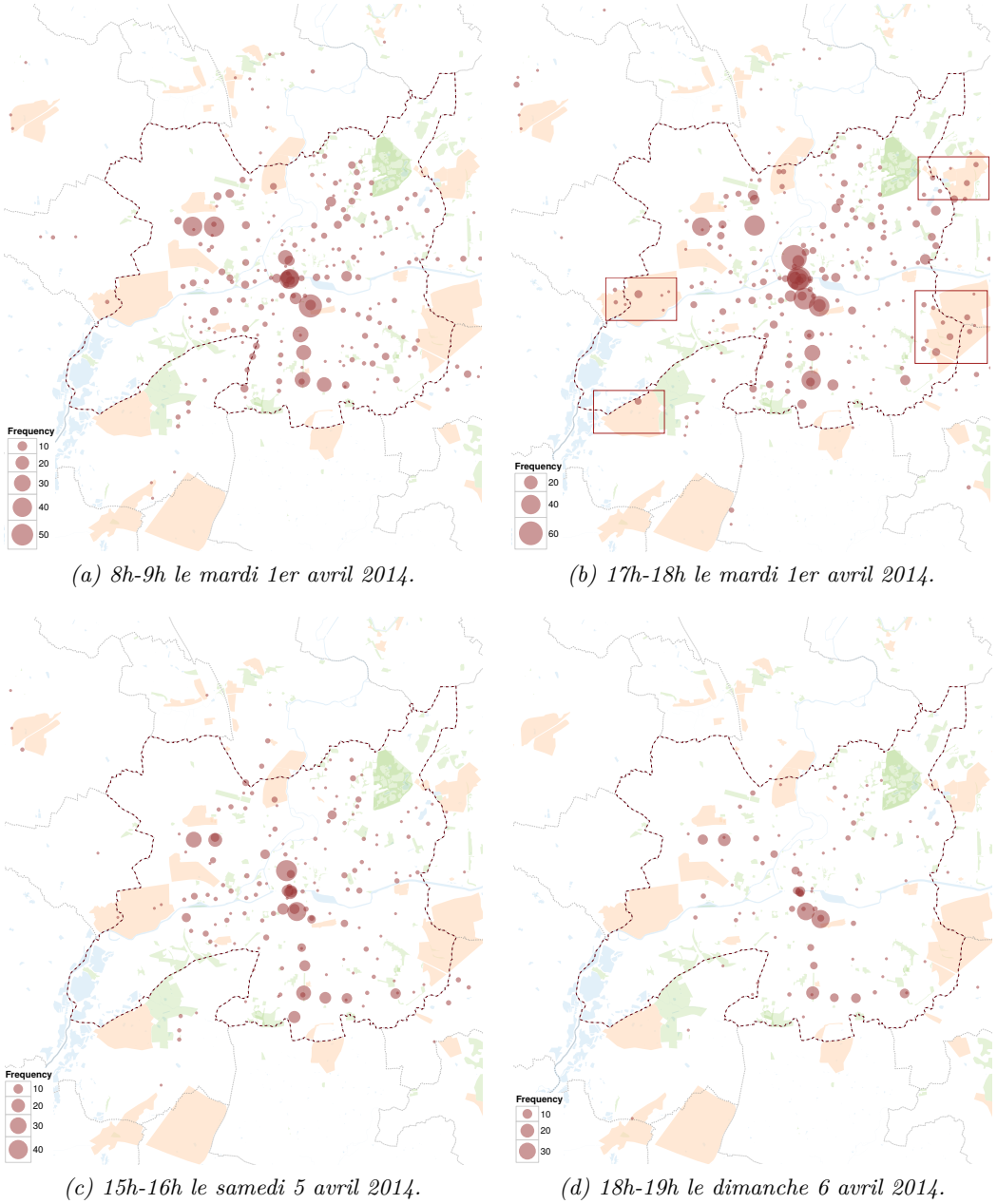
FIGURE *2.7 – Nombre de validations moyen par carte sur le réseau de transport rennais pendant la semaine du 7 avril 2014 pour différents types de titre de transport.*

afin d'avoir une idée plus précise des comportements des usagers.

2.1.4 Suivi de l'activité spatiale

Afin d'avoir une meilleure compréhension de la variabilité spatiale dans les données, le nombre de déplacements pour quatre périodes distinctes est représenté sur la Figure 2.8. Cette figure est composée de 4 cartes représentant le nombre de déplacements par station enregistrés le mardi 1er Avril entre 8 h et 9 h (Figure 2.8a), le mardi 1er avril entre 17 h et 18 h (Figure 2.8b), le samedi 5 avril entre 15h et 16 h (figure 2.8c) et enfin dimanche 6 avril entre 6 h et 7 h (figure 2.8d). Différents usages peuvent être observés selon le jour et l'heure. Par exemple, une différence dans le nombre de validations entre le mardi, qui est un jour de semaine typique, et les samedi et dimanche peut être notée : il y a plus de déplacements et de stations actives pendant la semaine que pendant le week-end. En outre, quelques différences spatiales entre le mardi matin et le soir sont observables : alors qu'elles sont inactives pendant la première moitié de la journée, les stations situées dans des zones industrielles et commerciales (mises en surbrillance en couleur beige) deviennent actives pendant la soirée. Enfin, l'activité du dimanche est principalement concentrée autour du métro et du centre-ville.

Les points susmentionnés illustrent la nécessité d'analyser les habitudes temporelles des passagers, car leur activité évolue constamment à la fois en ce qui concerne le nombre de déplacements et leur localisation spatiale et temporelle.



(a) 8h-9h le mardi 1er avril 2014.

(b) 17h-18h le mardi 1er avril 2014.

(c) 15h-16h le samedi 5 avril 2014.

(d) 18h-19h le dimanche 6 avril 2014.

FIGURE *2.8 – Nombre de déplacements enregistrés par station pour différents jours et heures de la semaine.*

2.2 Cas d'étude : réseau de transport en commun de la ville de Gatineau au Québec

La ville de Gatineau au Canada constitue le deuxième cas d'étude de cette thèse. Gatineau est la quatrième plus grand ville de la province du Québec, avec une population de 278 589 habitants. Elle a pour particularité d'être située sur la rive Nord de l'Outaouais, face à Ottawa (Ontario) qui est la capitale du Canada. La région métropolitaine d'Ottawa-Gatineau constitue la quatrième agglomération du Canada.

2.2.1 Présentation du réseau de transport

La Société des Transports de l'Outaouais (STO) basée à Gatineau, au Canada est une autorité de transport en commun de taille moyenne, le STO exploite 310 bus et dessert 291 000 habitants. Une caractéristique importante du réseau de transport de Gatineau est, comme mentionné précédemment, sa proximité avec Ottawa. Beaucoup de lignes de bus de Gatineau servent par conséquent à Ottawa, qui abrite une population plus étendue de 883 391 habitants (selon le recensement de 2011) et offre une grande quantité d'activités, en particulier les activités liées au travail.

Le STO exploite son système de cartes à puce depuis 2001. Aujourd'hui, plus de 80% de tous les passagers STO utilisent des cartes à puce. Chaque observation correspond à une validation comportant un identifiant de carte anonymisé, un type de carte, une date et une heure, l'arrêt, la ligne et le type de transaction (transfert ou non-transfert). Comme pour Rennes, les emplacements de descente ne sont pas connus. Cependant, les emplacements des validations, correspondant aux lieux de validations, apportent une information sur l'organisation de la Ville de Gatineau et ses interactions avec les villes voisines. Le nombre de validations enregistrées en février 2005 est présenté dans la figure 2.9.

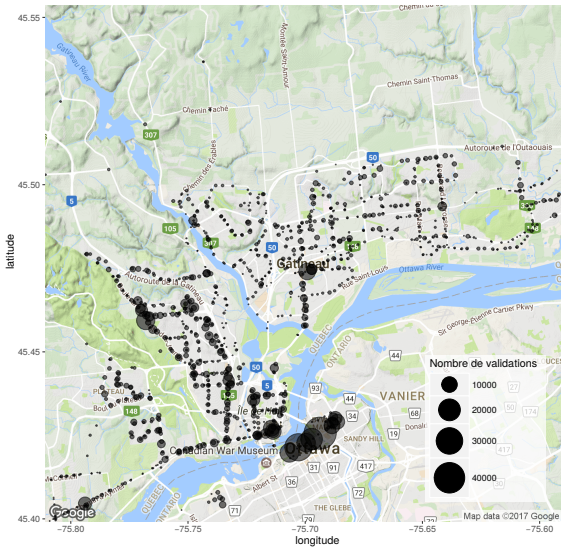


FIGURE *2.9 – Carte représentant le nombre de validations sur le réseau de la STO pendant le mois de Février 2005.*

Bien que la grande majorité des stations actives se répartissent sur le territoire de Gatineau, il est à noter que les stations générant le plus grand nombre de validations se trouvent sur la rive sud du fleuve, dans la ville d'Ottawa.

Pour mieux interpréter les résultats, les types de cartes sont regroupés en 11 types de tarifs : AP Express adulte, AP Interzone adulte, AP Adulte régulier, AP Senior, Univ., Express adultes, Interzone adulte, Autre, Adulte régulier, Senior et Étudiant. AP signifie que les cartes sont liées à un compte bancaire pour le paiement automatisé. Les utilisateurs express peuvent utiliser des itinéraires express, alors que les autres utilisateurs ne peuvent pas. Il en va de même pour les utilisateurs de Interzone qui se déplacent à partir des banlieues extérieures.

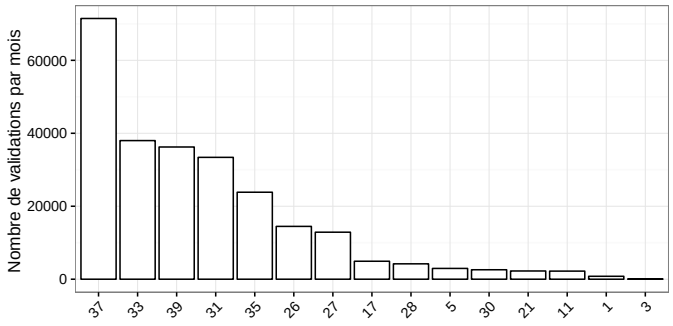


FIGURE *2.10 – Nombre de validations enregistrées sur les 15 lignes les plus fréquentées du réseau de la STO, sur le mois de février 2005.*

Cinq jeux de données différents ont été utilisés dans cette étude. Les données ont été enregistrées entre le 1er et le 28 février pour la période 2005-2009. Entre 644 614 (2005) et 740 883 (2009) observations composent chacun des jeux de données, ce qui représente un total de 3 492 310 validations faites par 82 223 cartes. Ces ensembles de données n'ont pas été filtrés avant l'étude, ce qui signifie qu'ils incluent toutes les cartes engagées dans au moins une transaction pendant la période d'étude.

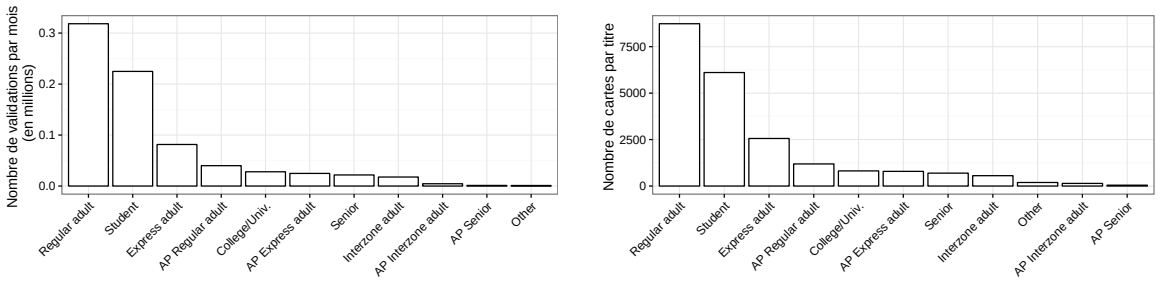
2.2.2 Gatineau au travers des données billettiques

Comme cela a été fait pour Rennes, une présentation de l'activité de Gatineau est établie, au travers de statistiques descriptives. Pour cela le jeu de données enregistrées pendant le mois d'avril 2005 est utilisé.

Contrairement à la ville de Rennes, la ville de Gatineau ne dispose pas de ligne de métro. Sur la figure 2.10 sont représentés le nombre de validations enregistrées sur la période d'étude pour les 15 lignes de bus les plus actives. Bien qu'une ligne semble attirer plus de validations que les autres (la ligne 34), la différence est moins marquée qu'à Rennes où le métro attire la plus grande partie des validations.

L'étude du nombre de cartes par type de titre (figure 2.11a) et du nombre de validations par type de titre (figure 2.11b) révèle que les titres possédant le plus grand nombre de cartes en activité, sont également ceux générant le plus grand nombre de validations sur le réseau. Les abonnements d'adultes réguliers sont les plus nombreux, suivis de près par les abonnements étudiants.

Comme pour Rennes le nombre de validations enregistrées sur le réseau pour une semaine d'activité est présentée en figure 2.12. Trois pics d'activités apparaissent, les deux premiers très marqués, en début et en fin de journée, et un troisième de plus faible amplitude qui s'étend sur toute la journée. Il est à noter que l'activité en milieu de journée est en proportion bien plus faible qu'elle ne l'est à Rennes. Il en est de même pour l'activité en week-end, qui est très faible. Contrairement à ce qui avait pu être observé sur Rennes avec l'activité du mercredi après-midi, aucun jour de semaine ne semble avoir une activité différente des autres. Le comportement pendulaire du système de transport est donc encore plus sensible qu'à Rennes.



(a) Nombre de validations par titre *(b) Nombre de cartes par titre*

FIGURE *2.11 – Nombre moyen de validations enregistrées par mois et nombre de cartes pour les différents types de titre de transport.*

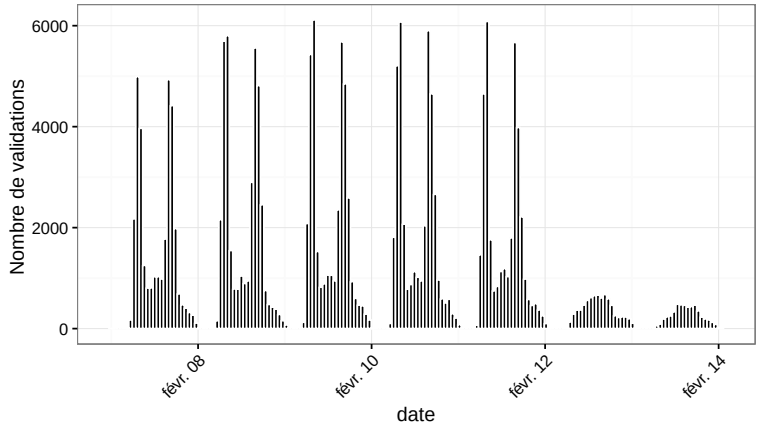
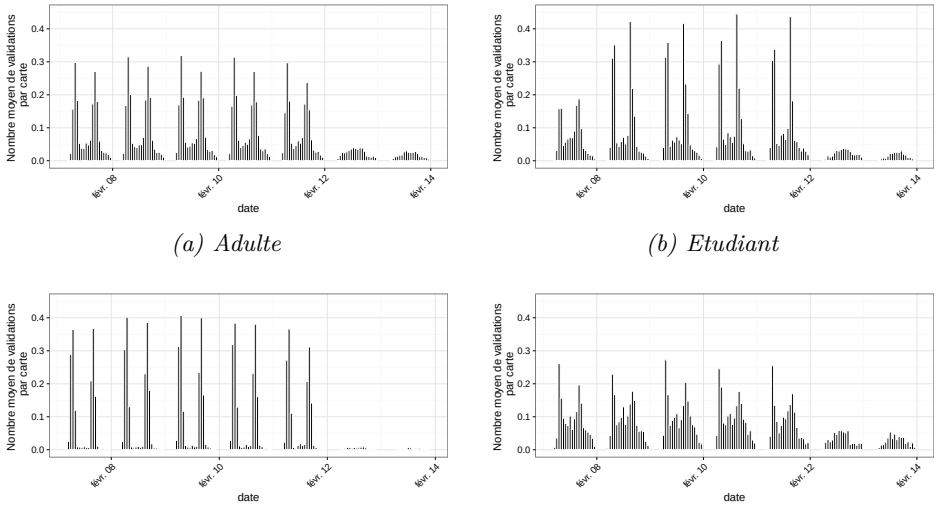


FIGURE *2.12 – Nombre de validations sur le réseau de transport de l'Outaouais pendant la semaine du 7 avril 2014.*

Afin d'investiguer les différences d'usage d'un titre à l'autre, le même histogramme est représenté en le différenciant par type de titre (figure 2.13). Il est alors observable que chaque titre semble correspondre à une utilisation différente du réseau.

Il peut être observé que les abonnés express semblent utiliser les transports uniquement pour leurs déplacements domicile-travail, et à des heures très régulières. Le reste des validations est résiduel. À l'opposé les Univ. (étudiants du supérieur) sont ceux qui utilisent le plus les transports en dehors des heures de pic. C'est également ceux qui ont la proportion la plus élevée d'activité en week-end. Les étudiants et les adultes ont quant à eux une utilisation plus classique avec toutefois une activité le week-end légèrement plus importante que pour les cartes express.

On peut voir ici que le croisement des données avec les types de cartes permet d'obtenir une représentation plus fine de l'utilisation des transports en commun. Cependant, une certaine variabilité est toujours présente au sein de ces données que des travaux plus poussés peuvent révéler.



(a) Adulte *(b) Etudiant*

(c) Express adulte *(d) Univ.*

FIGURE *2.13 – Nombre de validations sur le réseau de transport de l'Outaouais pendant la semaine du 7 avril 2005 pour différents types de titre de transport.*

2.2.3 Suivi de l'activité spatiale

Le nombre de validations enregistrées en février 2005 est présenté dans la Figure 2.14. Deux types de validations ont été considérées - les validations du matin avant 12h00 et les validations de l'après-midi après 12h00.

Comme pour Rennes, mais de manière plus marquée, les données révèlent qu'il existe une différence de lieu entre ces deux types de validations. Les validations de l'après-midi concernent surtout Ottawa, alors que les validations du matin se sont surtout produites à Gatineau, car la plupart des utilisateurs de STO vivent à Gatineau et travaillent à Ottawa.

2.2.4 Évolution de l'utilisation des transports

Il a été vu précédemment que les données disponibles pour la ville de Gatineau se composent de 5 jeux de données collectées sur les 5 mois de Février des années 2005 et 2009. Ces données offrent une profondeur temporelle permettant de conduire des analyses longitudinales sur l'activité des usagers.

Tout d'abord le nombre de validations enregistrées pour chaque mois de Février de 2005 à 2009 est inspecté (figure 2.15). Une tendance à la hausse est observable puisque le nombre de validations passe de 750 000 en 2005 à 900 000 en 2009.

L'étude de la proportion de type d'abonnements enregistrée sur chacun de ces mois (figure 2.16) montre également une évolution au cours des années. En effet la proportion de carte de type Univ. est plus importante en 2009, tandis que la part des cartes de type adulte régulier et Etudiant diminue. De manière générale, les abonnements de type AP (avec prélèvement automatique) ont une proportion qui tend à augmenter au fil des ans.

Enfin la part relative de chaque type de carte pour l'ensemble des cartes, ainsi que pour les cartes actives sur les 5 ans de données sont présentées sur la figure 2.17. On peut noter l'absence de cartes étudiantes pour les cartes actives sur les cinq années. Cela s'explique

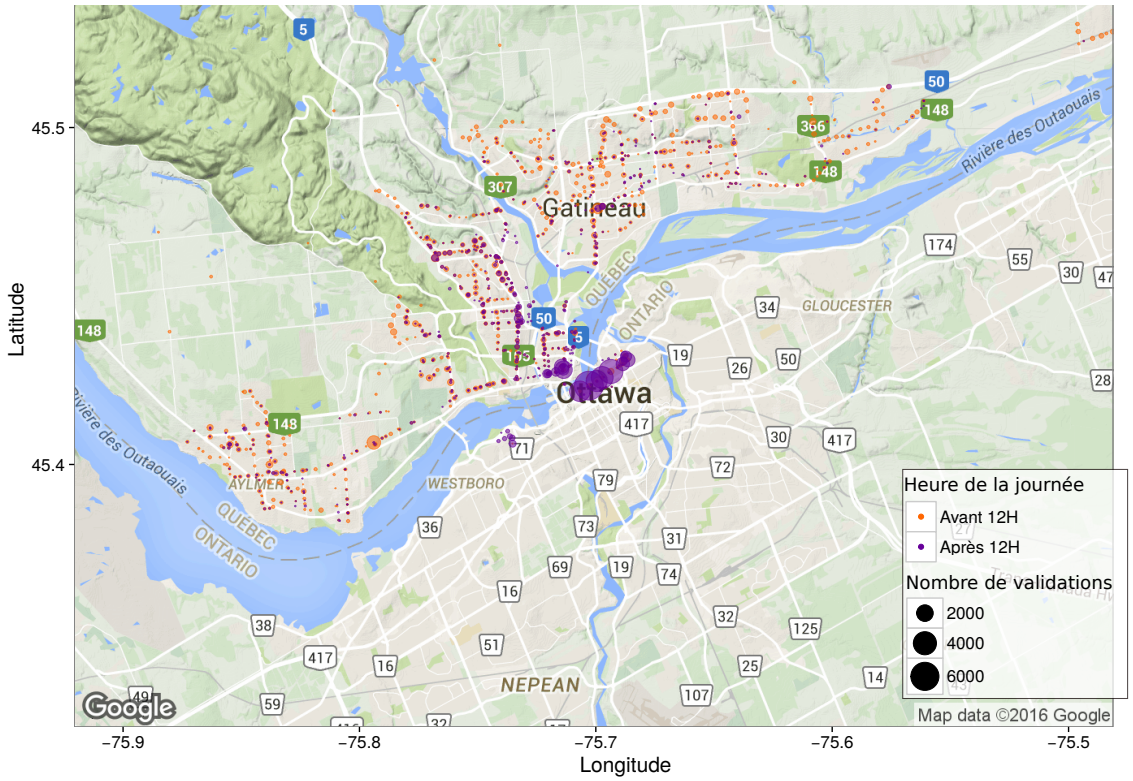


FIGURE *2.14 – Carte représentant le nombre de validations sur le réseau de la STO pendant le mois de Février 2005, avant et après midi.*

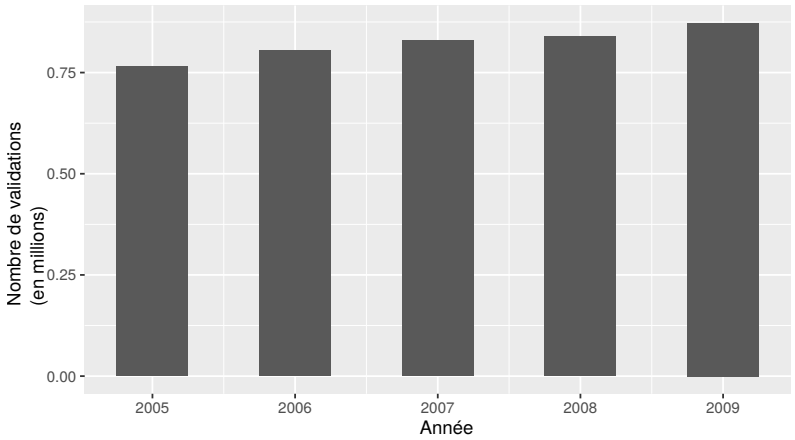


FIGURE *2.15 – Nombre de validations enregistrées pour le mois de Février des années 2005 à 2009.*

par le fait que les étudiants ne peuvent pas garder la même carte, ils doivent la changer tous les ans, ce qui n'est pas nécessairement le cas pour d'autres types de tarifs. Ainsi, il sera intrinsèquement impossible de suivre l'activité étudiante d'une année sur l'autre.

Conclusion de chapitre

Dans ce chapitre les deux cas d'étude utilisés ont été présentés. Rennes et Gatineau sont deux villes de taille moyenne mais qui se différencient par de nombreux aspects.

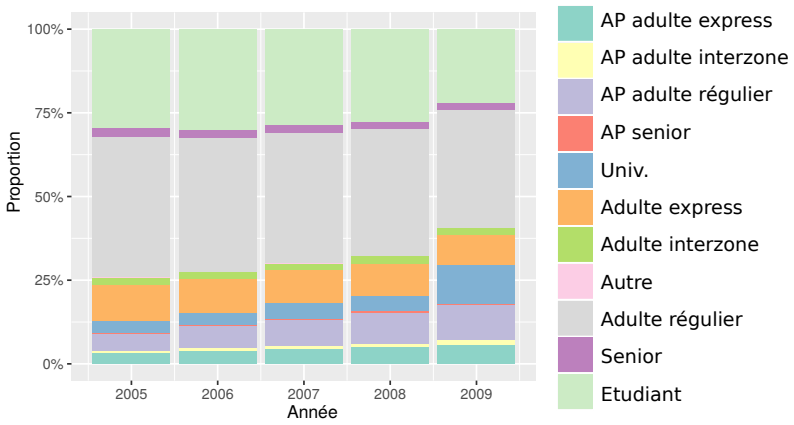


FIGURE *2.16 – Proportions des types de carte sur les années de 2005 à 2009.*

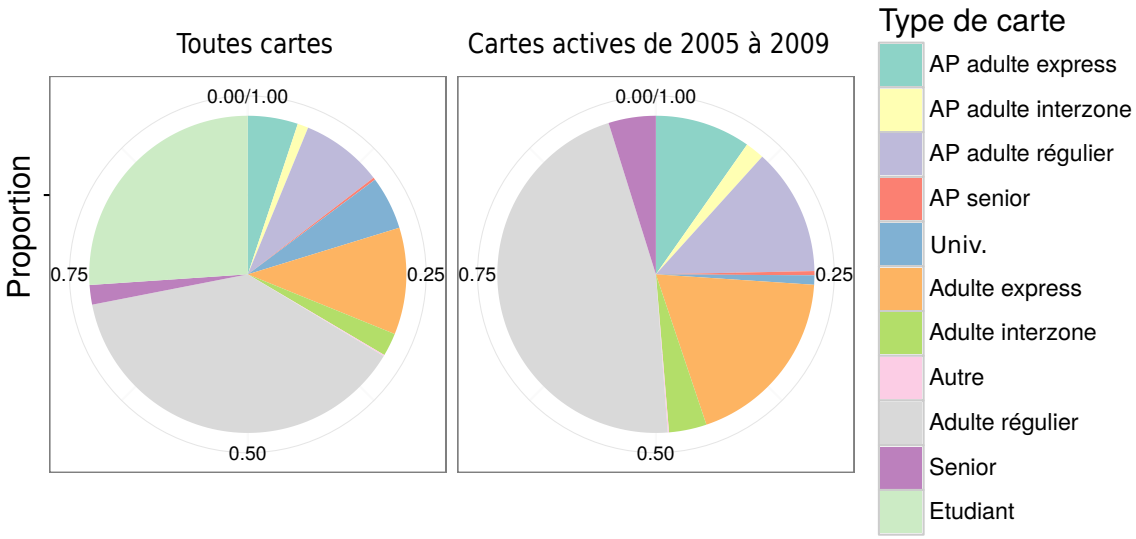


FIGURE *2.17 – Proportions des types de carte pour l'ensemble des cartes et pour les cartes actives sur toutes les années de 2005 à 2009.*

Rennes possède une ligne de métro qui concentre une grande partie de ces validations, tandis que Gatineau n'est desservi que par des lignes de bus. De plus, Rennes est au centre de l'agglomération rennaise et à ce titre concentre une grande partie de l'activité des transports, tandis que Gatineau est située dans la banlieue d'Ottawa et possède donc une activité plus restreinte. Ces deux villes offrent donc deux situations différentes qui permettront d'évaluer nos travaux.

L'analyse des jeux de données disponibles a permis de mettre en évidence certains points qui nécessitent d'être approfondis. Le premier est la variabilité de l'activité, qui n'est pas visible sur les données agrégées. Celle-ci peut se traduire par des variations au cours de la journée, ou bien par des différences entre différents usagers. Une partie des travaux présentés dans cette thèse approfondit cette question en proposant une vision non agrégée des heures de validations et en allant au-delà des groupes prédéfinis par l'opérateur de transport (type de carte). D'autre part, le suivi du nombre de validations sur les 5 mois de données disponible sur Gatineau ne permet pas d'étudier de manière précise les

évolutions dans les habitudes des usagers. Un suivi des usagers et des évolutions de leurs habitudes de mobilité est donc également proposé dans cette thèse.

Chapitre 3

Outils pour l'aide à la mise en qualité des données

Comme cela a été énoncé dans le chapitre 1, les données billettiques possèdent une richesse spatio-temporelle. Toutefois leur utilisation nécessite la plupart du temps un travail d'enrichissement afin, par exemple, d'estimer les transferts et les destinations ou les motifs du déplacement. Conjointement à ces méthodes d'enrichissement, il est nécessaire de détecter les éventuels problèmes de mesure qui pourraient être présents. Ces problèmes peuvent être de plusieurs natures : données fausses (absence ou remontées partielles des données), données inhabituelles (événements sociaux ou culturels).

De tels problèmes ne sont pas toujours facilement détectables. En effet l'activité varie en fonction du contexte (type de jour, période de l'année, etc.) et rend leur détection plus complexe. Afin de détecter ces problèmes de mesure, il est alors nécessaire de prendre en compte les variables de contexte pouvant influer sur l'activité de mobilité dans les transports en commun.

Dans ce chapitre, une approche en deux étapes est proposée pour effectuer les prétraitements cités. Elle a pour but, dans un premier temps, de partitionner les jours homogènes au sens de l'utilisation des transports, puis en s'aidant de cette classification, de détecter les journées et les quarts d'heure pour lesquels l'activité est atypique. Cette approche est appliquée sur les deux ans de données disponibles pour Rennes, puis les résultats sont croisés avec les informations disponibles sur l'activité du réseau, notamment les tweets de l'opérateur.

Le but et le contexte des travaux sont présentés. Puis, un état de l'art des travaux menés sur la problématique de détection d'événements atypiques est détaillé aussi bien du point de vue applicatif que méthodologique. La méthodologie proposée pour la détection d'événements atypiques est ensuite présentée. L'évaluation de l'approche proposée sur données réelles fait l'objet de la section 3.4 où l'ensemble des résultats obtenus est analysé.

3.1 But et contexte de la mise en qualité des données billettiques

Ce chapitre décrit une approche permettant de détecter des événements atypiques en analysant les données billettiques collectées sur un réseau de transport. Ces variations inhabituelles par rapport à un comportement nominal se traduisent en général par une hausse ou une baisse du nombre de validations, lesquelles peuvent correspondre à plusieurs situations, à savoir :

- un pic d'activité dû à un événement exceptionnel (concert, match, etc),
- un problème d'exploitation côté opérateur (incident technique, pannes) générant un nombre plus ou moins important de validations,
- une anomalie de la chaîne de valideurs (portique ouvert, validateur hors d'usage) se traduisant par une baisse du nombre de validations.

Plus formellement, il s'agit de détecter des "outliers" dans les données billettiques à l'aide d'approches statistiques. Ce travail s'effectue dans un contexte non supervisé, aucune base d'événements labellisés n'est mobilisée pour construire les règles de détection des événements atypiques. Une fois ces événements identifiés, les résultats de détection sont croisés avec des événements ayant lieu dans la ville extraits du fil tweeter de l'opérateur de transport de manière à pouvoir interpréter et évaluer les résultats obtenus.

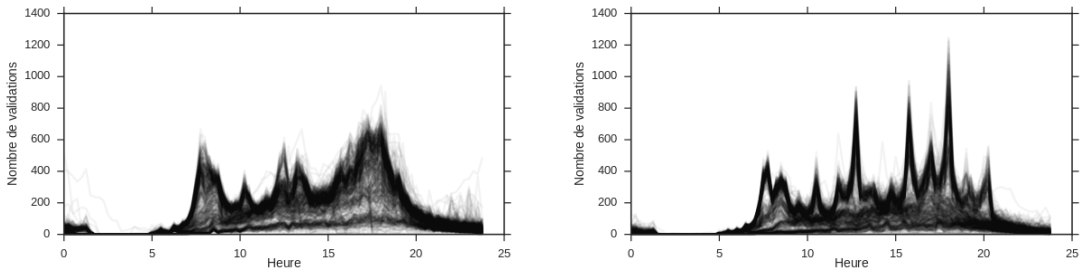
Une telle détection suppose de bien définir ce qu'est un fonctionnement nominal du système de transport et de la demande de mobilité. Les données de validations possèdent en effet une grande variabilité qui dépend de la station et du jour d'étude. Par conséquent, il est pertinent de définir au préalable des groupes de jours pour lesquels l'activité est similaire, puis d'appliquer une méthode de détection sur chacun de ces groupes. Une approche classique consiste à regrouper les jours selon le type de jour de la semaine (lundi, ..., dimanche) Tonnelier et al. 2017 puis d'effectuer la détection pour chaque jour de la semaine. Ici dans l'approche proposée, le regroupement se fait non pas selon le type de jour de la semaine mais plutôt à l'aide d'un clustering permettant de regrouper au sein d'un même cluster les journées dont l'activité est semblable. Dans ce chapitre deux types de méthodes basés sur le concept de boxplot seront utilisés et comparés.

Le nombre de validations enregistrées par station agrégées au quart d'heure est utilisé ici. L'ensemble des traitements présentés dans ce chapitre ne nécessite donc pas le stockage des identifiants anonymisés des usagers. Sur la période d'analyse (du 1er Juin 2014 au 29 Mai 2016), chaque station est décrite par un ensemble de D vecteurs (un par jour) où chaque vecteur d'observation, est constitué du nombre de validations $x_{s,d,i}$ enregistrées pour chaque quart d'heure de la journée :

$$X_{s,d} = (x_{s,d,1}, \ldots, x_{s,d,N})$$

$s \in \mathcal{S} = \{1, \ldots, S\}$ désigne l'ensemble des stations, $d \in \mathcal{D} = \{1, \ldots, D\}$ l'ensemble des journées de la période d'étude et $i \in \{1, \ldots, N\}$ le quart d'heure dans la journée. Soit un total de 4×24 enregistrements. Les journées pour lesquelles aucune validation n'a été enregistrée sur une des quinze stations entre 8H et 20H ont été retirées du jeu de données.

L'ensemble des courbes de validations est tracé sur la figure 3.1 pour les deux stations considérées.



(a) Station République *(b) Station Villejean Université*

FIGURE *3.1 – Nombre de validations enregistré par quart d'heure durant la période du 1er juin 2014 au 29 Mai 2016 pour les stations République et Villejean-Université.*

L'observation de ces courbes met en évidence la présence :

— d'une variabilité inter-stations (différence d'activité entre stations différentes) qui traduit les différences de comportements dans l'usage du réseau de transport en commun. Pour le réseau de Rennes, cette variabilité a déjà été mise en évidence dans les travaux de MAHRSI et al. 2016. Plusieurs profils d'activités des stations selon l'usage qu'il en est fait (domicile-travail, travail-domicile, loisirs, loisirs nocturnes, etc.) ont été identifiés. En considérant les stations République et Villejean-Université (figure 3.1), on peut constater que la station République présente majoritairement des courbes de validations avec trois pics d'activité qui correspondent aux pics d'activité du matin, du midi et du soir, tandis que l'activité de la station Villejean-Université est constituée d'un nombre plus important de pics plus proches reflétant l'usage majoritaire de cette station par des étudiants à la fin des différentes périodes de cours.

— d'une variabilité intra-station (différence d'activité au sein d'une même station). Celle-ci est due à un usage différent selon le type de jour (jour de semaine, jour de week-end, période de vacances, ...). Par exemple, un usager n'aura pas la même activité un jour travaillé de semaine qu'un jour de week-end ou en période de vacances BRIAND et al. 2016. La station République présente différents groupes de courbes correspondant à différents types de jours. On observe notamment un groupe avec un nombre plus faible de validations tout au long de la journée et un groupe avec une activité plus forte.

— d'événements isolés que l'on distingue par des courbes différentes de toutes les autres. Il s'agit de comportements que l'on peut qualifier d'atypiques, c'est à dire des courbes isolées ne faisant partie d'aucun groupe "nominal". Certaines courbes se détachent visuellement sur la figure 3.1. Toujours sur la station République, on observe notamment des courbes isolées aux alentours de 1h. Cependant toutes les courbes atypiques ne sont pas aisément observables et on peut supposer qu'un grand nombre d'entre elles est masqué par l'ensemble important de courbes.

C'est sur ce troisième point, c'est à dire la détection de comportements atypiques que ce

chapitre se focalise. Pour être en mesure de détecter les courbes atypiques, il est cependant intéressant de regrouper au préalable les courbes similaires afin de s'affranchir des deux variabilités observées : inter et intra-stations. En ce qui concerne la variabilité inter-station, il suffit de travailler indépendamment sur chaque station. En revanche, pour ce qui est de la variabilité intra-stations, il est nécessaire de déterminer au préalable les facteurs influents (type de jour : férié, travaillé, ...) et d'identifier les comportements nominaux propres à chaque groupe de jours. Dans ce chapitre une méthodologie de détection en deux étapes est étudiée : la première étape repose sur une classification non supervisée des types de jours à partir de l'activité de l'ensemble des stations, la seconde étape de détection d'événements atypiques à proprement parlé est effectuée station par station et utilise la typologie des jours précédemment mise au point. Des outils de détection d'outliers classiques (boxplot et boxplot fonctionnel) sont mobilisés pour effectuer cette tâche.

3.2 Travaux existants pour la détection d'événements atypiques

La détection de données atypiques a fait l'objet de plusieurs travaux de recherche de par son utilité dans une grande variété d'applications allant du nettoyage de données bruitées à la détection d'intrusion dans un réseau. Si plusieurs travaux ont été consacrés à la détection d'anomalies sur des données de trafic routier, on note très peu de travaux dans le domaine du transport public et notamment sur les données billettiques. On peut mentionner les travaux récents menés par Pang et al. 2013b qui, à partir de données GPS, proposent d'adapter le test du ratio de vraisemblance aux trajectoires des taxis en vue de détecter des situations de trafic atypiques ou encore les travaux de Wang et al. 2016 où la détection est basée sur des descripteurs du flux, tels que l'accélération des véhicules. La détection peut également se baser sur des images enregistrées sur le réseau, comme c'est le cas dans [Li et al. 2016]. Des modèles de distribution gaussienne appliqués sur les images collectées sur le réseau routier permettent d'assurer la détection en considérant que toute donnée éloignée de la moyenne de la distribution correspond à une anomalie. Cette section présente plusieurs travaux portant sur la détection d'outliers. Les méthodologies peuvent se répartir en plusieurs catégories, quelques-unes sont présentées ici.

Un grand nombre d'approches a été développé dans un cadre supervisé, c'est à dire pour des données disposant d'une vérité terrain (les situations atypiques sont annotées). Il s'agit dans ce cas, de construire un modèle de classification par apprentissage sur données labellisées en vue de discriminer entre deux classes de fonctionnement, normale et en anomalie. Dans ce cadre, on peut citer les méthodes de classification supervisée [Steinwart et al. 2005], les approches modélisant les données à l'aide de séries temporelles [Yamanishi et Takeuchi 2002] ou encore des approches par estimations de densités [Hido et al. 2011].

Lorsque l'on ne dispose pas de vérité terrain comme c'est le cas ici, il est naturel de s'orienter vers un cadre d'apprentissage non supervisé où l'on cherche à reconnaître automatiquement les situations anormales sans utiliser de labels. La distance ainsi que la densité locale de points sont des critères communément employés pour détecter des

outliers [Knorr et Ng 1999] dans ce cas de figure. Différentes définitions peuvent être utilisées pour détecter de telles situations [Bay et Schwabacher 2003] : (i) les points qui possèdent moins de p autres points à une distance d, (ii) les n points dont la distance au k voisins les plus proches est la plus élevée, (iii) les n points dont la distance moyenne aux k plus proches voisins est la plus grande. L'un des algorithmes les plus répandus pour l'identification basée sur la densité des outliers est l'algorithme LOF (Local Outlier Factor) [Breunig et al. 2000] qui affecte à chaque point un indicateur sur son niveau "d'isolement" par rapport à son voisinage. Une extension de cet algorithme a été proposée dans [Latecki et al. 2007]. Les auteurs ont modifié un estimateur de densité non paramétrique avec une variable à noyau pour obtenir une estimation de la densité localement robuste. Un point est alors considéré comme outlier sur la base de comparaison de sa densité avec celle de son voisinage.

Dans ce chapitre les approches de type boxplot sont utilisées. Le boxplot originellement introduit par Tukey 1977, est largement utilisé pour détecter des outliers. D'autres travaux ont suivi la même idée mais en appliquant le principe du boxplot à des fonctions et non plus à des points. On parle alors de boxplot fonctionnel [Sun et Genton 2011]. De nombreuses autres approches ont été développées afin de mieux détecter les outliers dans un ensemble de données. Le lecteur intéressé pourra se référer aux états de l'art existants, notamment celui proposé dans [Zhang 2013] qui décrit de manière très complète les différentes approches de détection d'outliers à la fois pour les données de faible dimension, les données de grande dimension et les flux de données. Plus récemment, on peut mentionner la revue de littérature des méthodes de détection d'outliers basées sur des approches de data mining [Agrawal et Agrawal 2015], ou encore l'article de Ahmed et al. 2016 où le focus est mis sur la détection d'anomalies sur des réseaux informatiques.

S'agissant du domaine du transport et des données cartes à puces en particulier, on peut mentionner les récents travaux basés sur la factorisation par matrice non négative [Tonnelier et al. 2017]. Dans cet article les auteurs se placent sur des journées de 24H et traitent séparément chaque jour de la semaine. Des modèles de référence robustes pour chaque couple $(jour, station)$ sont ensuite construits et la détection se base sur les différences entre des situations particulières et la référence estimée sur des données de base et d'autres débruitées par NMF. La méthodologie proposée est évaluée en croisant les résultats de détection avec les vérités terrain provenant de données twitter de l'opérateur. Cependant une telle approche ne permet pas de valider entièrement la labellisation obtenue. En effet aucune information sur les problèmes de remontées de données ou les événements atypiques n'apparaissent dans les tweets de l'opérateur.

Notre travail s'inscrit dans la même visée applicative que ces travaux, à savoir la détection d'événements atypiques à partir des données billettiques dans un cadre non supervisé. En revanche, plusieurs points de différence peuvent être notés. Alors que l'approche proposée dans [Tonnelier et al. 2017] est contrainte par le dictionnaire de module de la NMF, les deux approches de détection au quart d'heure qui vont être présentées ici distinguent les événements ayant un nombre plus important de validations de ceux ayant un nombre plus faible de validations. Des types de jours définis par un clustering de l'activité des

stations sont utilisés à la place des jours de la semaine. Ne disposant pas de labels fournis par une autre source de donnée, l'évaluation des résultats reste qualitative et repose essentiellement sur le croisement des journées détectées en anomalie avec des événements majeurs culturels et sociaux ayant lieu à Rennes durant la période d'étude.

3.3 Méthodologie

Dans cette section la méthodologie en deux étapes proposées pour la détection d'anomalies est détaillée. La première étape de clustering pour l'identification des types de jour fera l'objet d'une attention particulière et sera décrite dans un premier temps. Les indicateurs proposés pour l'évaluation de la méthodologie seront ensuite détaillés.

3.3.1 Description de la méthodologie

Considérant les données journalières de comptage en entrée de station agrégées au quart d'heure, par chaque couple $(station, jour)$ un vecteur des validations par quart d'heure $X_{s,d}$ est défini. L'ensemble des vecteurs $X_{s,d}$ forme le jeu de données initial.

La figure 3.2 illustre le principe de la méthodologie de détection. A partir des données journalières, une aggrégation sur l'ensemble de la ligne de métro est appliquée afin d'obtenir les comptages de validations au quart d'heure sur toute la ligne de métro et non plus par station. Chaque journée est alors définie par un vecteur X_d :

$$X_d = (x_{d,1}, \ldots, x_{d,N}) = \sum_s X_{s,d}.$$

La première étape consiste en un clustering de l'ensemble des journées étudiées. Le but est de regrouper dans un même cluster les journées ayant un profil d'activité similaire. L'étape de détection s'effectue ensuite non pas sur les données globales mais par cluster et par station. Cette première étape permet de simplifier grandement la problématique de détection d'anomalie car chaque cluster est composé de courbes globalement homogènes. Deux méthodes classiques ont été appliquées pour la détection, le boxplot (aussi appelé boxplot classique ou boxplot ponctuel dans le cadre de ces travaux) et le boxplot fonctionnel. Elles sont détaillées en A.1.

3.3.2 Classification des courbes de validations par Classification ascendante hiérarchique

Dans cette section les jours de la période d'étude sont classés en fonction de leur profil d'activité. Deux approches sont proposées. La première ne pose aucune contrainte et consiste simplement en un clustering des courbes représentant le nombre de validations enregistrées au quart d'heure sur l'ensemble de la ligne. La deuxième approche va quant à elle introduire des contraintes sur les types de jours et forcer les courbes correspondants à un même type de jour à appartenir au même cluster.

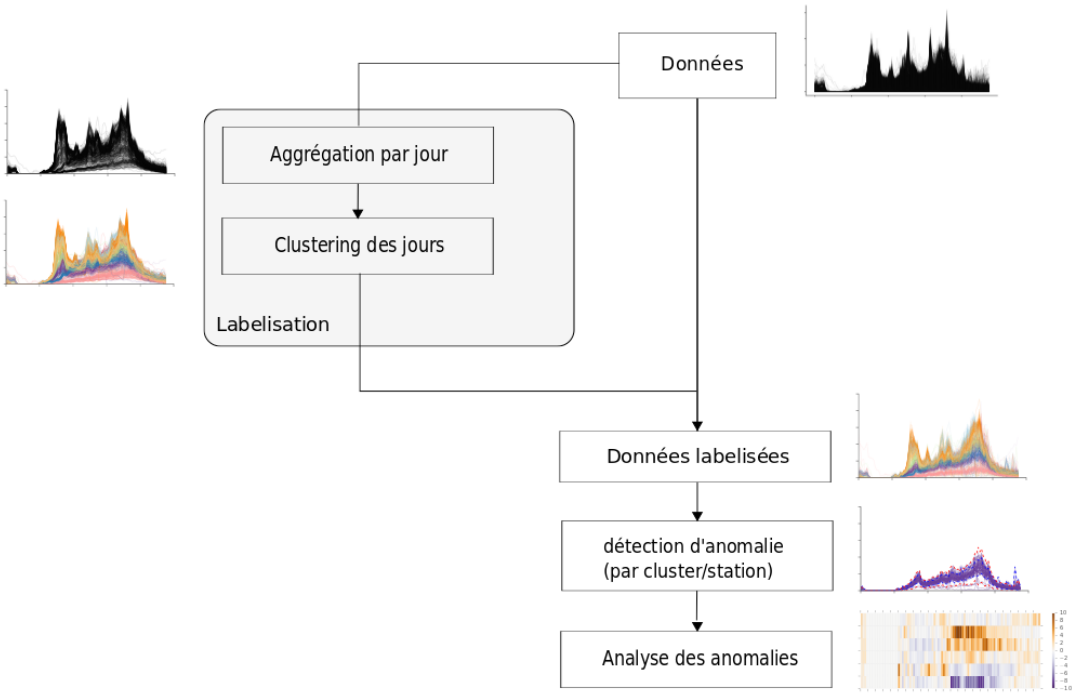


FIGURE *3.2 – Méthodologie pour la détection d'anomalie.*

Approche sans a priori : Dans ce cas, il s'agit de procéder le plus simplement possible en effectuant une classification globale sur l'ensemble des courbes de comptage de la ligne X_d. Une classification ascendante hiérarchique avec le critère de Ward est donc appliquée directement sur l'ensemble des vecteurs X_d.

Approche avec a priori sur le type de jours : Sachant que l'activité d'une journée d sur la ligne de métro, représentée par le vecteur X_d, dépend de nombreux facteurs comme par exemple le jour de la semaine (lundi, ..., dimanche), la saison, la présence ou non d'un jour férié, d'un pont ou bien de vacances scolaires, il s'agit d'effectuer une classification en prenant en compte des facteurs calendaires connus pour impacter l'activité enregistrée.

L'ensemble de ces facteurs ne peut pas être directement utilisés pour définir des groupes de courbes homogènes. En effet, par exemple le produit cartésien défini par les variables : mois, jour de la semaine, vacances, jours fériés conduit à certaines modalités avec un nombre très faible de courbes (il existe ainsi un seul jeudi férié en mai au sein des deux ans de données étudiés dans ce chapitre). Pour que les méthodes de détection d'outliers fonctionnent celles-ci doivent s'appuyer sur un nombre minimal de courbes de manière à estimer correctement la variance des situations nominales, ce qui n'est pas possible avec une approche aussi directe. L'approche adoptée ici a pour but de créer des classes avec chacune un nombre significatif de courbes. Elle se décompose en trois étapes. La première consiste à définir des groupes dépendants des facteurs précédemment énoncés (type de jour, mois, etc.). Pour chaque jour d de la période d'étude un vecteur de variables catégorielles

pouvant impacter l'activité d'une station est défini :

$$cat_d = (jour, mois, ferie, vacances, pont),$$

$jour \in \{1, \ldots, 7\}, mois \in \{1, \ldots, 12\}$, $(ferie, vacances, pont) \in \{0, 1\}^3$). Sont alors affectées à un même groupe G_{cat}, toutes les journées possédant les mêmes valeurs pour leur vecteur *cat*. La deuxième étape consiste à calculer une courbe médiane pour chaque groupe. En effet la médiane étant un estimateur robuste, elle est moins sensible aux anomalies. La médiane est calculée en chaque point (quart d'heure). Le vecteur médian $\widetilde{X_{cat}}$ d'un groupe ayant pour variable catégorielle *cat* peut s'écrire :

$$\widetilde{X_{cat}} = (M(\{x_{d,1} : d \in G_{cat}\}), \ldots, M(\{x_{d,N} : d \in G_{cat}\})),$$

avec $M(x_1, \ldots, x_n)$ la médiane de l'ensemble $\{x_1, \ldots, x_n\}$. Enfin la troisième étape consiste en une classification ascendante hiérarchique des courbes médianes ainsi obtenues.

Cette approche de classification des courbes d'activité a plusieurs avantages. Elle offre tout d'abord un outil de prévision. En effet l'hypothèse initiale est ici que les groupes formés par le produit cartésien des variables catégorielles (jour, mois, vacances, pont et jour férié) regroupent uniquement des courbes d'activité similaires (hors courbe correspondant à un événement atypique). C'est sur ces groupes qu'est appliqué le clustering, ce qui implique que pour un type de jour donné (par exemple lundi hors vacances scolaire de décembre, non férié) tous les jours correspondant à ce type de jour seront dans le même cluster. Il est alors possible d'établir un calendrier prévisionnel de l'activité d'une station en l'attribuant par avance au cluster qui lui correspond à l'aide de son type de jour. Cette approche force des courbes dont l'activité doit être similaire, car elle correspond au même type de jour, à être dans le même groupe ce qui permet d'éviter que certaines courbes correspondant à des événements atypiques soient mal classées.

3.3.3 Définition d'indicateurs d'anormalité

Les deux méthodes de détection d'anomalies étudiées dans ce chapitre (cf Annexe A.1) fournissent en sortie une liste de quart d'heure détectés comme anormaux. En plus de cette liste et de manière à simplifier l'analyse des résultats, deux indicateurs permettant de quantifier l'envergure des anomalies sont fournis. Le premier est un indicateur global sur la journée afin de savoir si celle-ci est globalement en anomalie, tandis que le second est calculé à l'échelle du quart d'heure et permet d'identifier les quarts d'heure pour lesquels le nombre de validations est éloigné de la situation nominale.

Anomalies à la journée : Deux indicateurs d'anomalies à la journée sont définis par agrégation du nombre d'anomalies détectées au quart d'heure. Si N désigne le nombre de quarts d'heure dans une journée, l'indicateur d'anomalies positives (resp. négatives) est défini comme étant le nombre de quarts d'heure détectés en anomalie positive (resp. négative) sur une journée.

$$\begin{cases} I_{pos}(s,d) = \sum_{i=0}^{N} \delta_{pos}(s,d,i) \\ I_{neg}(s,d) = \sum_{i=0}^{N} \delta_{neg}(s,d,i) \end{cases} \tag{3.1}$$

avec $\delta_{pos}(s,d,i) = 1$ si le $i^{\text{ième}}$ quart d'heure de la journée d à la station s est en anomalie positive et $\delta_{pos}(s,d,i) = 0$ sinon. De même $\delta_{neg}(i,d,s) = 1$ si le $i^{\text{ième}}$ quart d'heure est en anomalie négative et $\delta_{neg}(i,d,s) = 0$ sinon.

Anomalies au quart d'heure : Une analyse plus fine (au quart d'heure) est considérée en vue de détecter des anomalies plus ponctuelles plutôt qu'une modification de l'activité sur une journée. Contrairement au paragraphe précédent, les données ne sont pas agrégées et un indicateur d'éloignement aux données "nominales" est calculé pour chaque point (càd en chaque quart d'heure). Pour chaque quart d'heure i d'un jour d, la distance de son nombre de validations au reste des données valides est donnée par :

$$I_{fin}(s,d,i) = \frac{x_{s,d,i} - M(\{x_{s,e,i} : k(e) = k(d)\})}{\mathcal{Q}_3(\{x_{s,e,i} : k(e) = k(d)\}) - \mathcal{Q}_1(\{x_{s,e,i} : k(e) = k(d)\}),} \tag{3.2}$$

avec $k(d)$ une fonction associant à chaque jour son numéro de cluster et $\mathcal{Q}_1$ et $\mathcal{Q}_3$ les premier et troisième quartiles.

3.4 Résultats

Dans cette section les résultats sont analysés sous deux aspects. Le premier orienté méthodologie vise à comparer les deux approches de classification et les deux approches de détection d'anomalies. Le deuxième volet de l'analyse est plutôt applicatif avec en particulier, le croisement des résultats de détection avec les informations sur les différents événements (sociaux, culturels) ayant lieu à Rennes.

3.4.1 Choix des approches méthodologiques

Comparaison des deux approches de classification des jours

Les deux approches de classification ascendante hiérarchique telles que décrites précédemment ont été appliquées sur les données de validations agrégées sur l'ensemble des stations. L'objectif est d'obtenir une classification de jour commune à toutes les stations. Le choix du nombre de cluster a été porté à 9 à l'aide du dendrogramme du clustering contraint. Le premier clustering, nommé clustering "libre", est obtenu par l'application de l'approche sans contrainte, tandis que le deuxième clustering, nommé clustering "contraint", est obtenu en regroupant préalablement les jours appartenant à la même modalité du produit cartésien des variables : jour de la semaine, mois, vacances. Pour faciliter la lecture des résultats, la numérotation des clusters a été modifiée afin que deux clusters

ayant le même numéro possèdent le plus grand nombre d'éléments en commun quand cela est possible.

Comparaison de la composition des clusters

Une représentation calendaire des résultats du clustering libre est fournie sur la figure 3.3 et du clustering avec contrainte sur la figure 3.4. Une couleur a été associée à chaque couple de cluster, c'est à dire les clusters ayant le plus d'éléments en commun. Comme on peut s'y attendre, le calendrier obtenu avec une classification libre de toute contrainte appliquée aux types de jour, révèle une influence certaine de ceux-ci sur le calendrier. En effet, les mois, jours de la semaine, vacances scolaires et jours fériés jouent un rôle important dans le découpage obtenu qui reflète la demande moyenne en transport public. Ce découpage est encore plus net pour le calendrier obtenu par le clustering contraint.

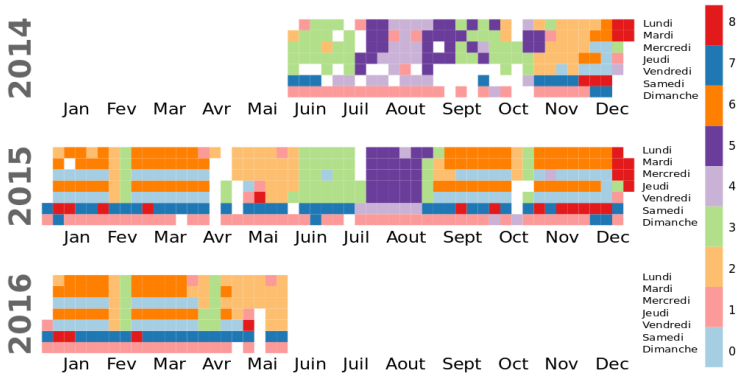


FIGURE *3.3 – Calendrier de la classification des jours obtenus à l'aide du clustering libre.*

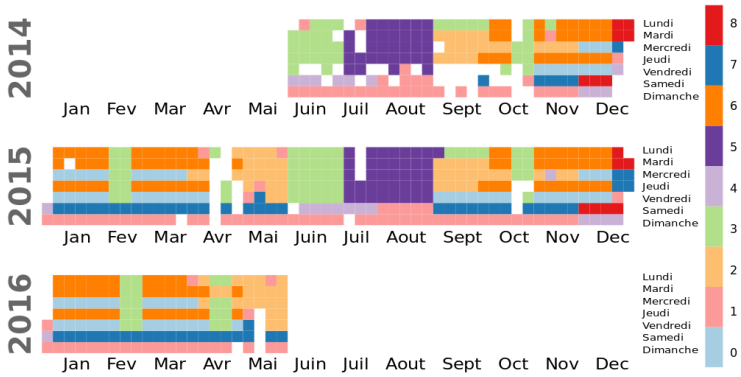


FIGURE *3.4 – Calendrier de la classification des jours obtenu à l'aide du clustering avec contrainte.*

Une certaine ressemblance entre les deux clustering ressort de l'étude de leurs calendriers, principalement sur les périodes travaillées. En effet, comme énoncé précédemment, plusieurs clusters partagent leurs éléments et ce, bien que l'un soit obtenu avec contraintes et l'autre sans. En période normale, hors vacances scolaires, les jours de semaine sont presque uniquement affectés aux clusters C6/L6 (pour les lundi, mardi et jeudi) et clusters C0/L0 (pour les mercredi et vendredi). Les clusters C7/L7 vont être composés majoritairement de samedi tandis que les clusters C1/L1 vont regrouper presque tous les dimanches. Enfin les jours travaillés des mois de mai et Juin sont affectés aux clusters C2/L2 et C3/L3

respectivement.

En période de vacances scolaires en revanche, les divergences entre les deux méthodes de clustering sont plus marquées. Le cluster C5 va s'étendre sur les mois de juillet et août en 2014 et 2015, tandis que le cluster L5 ne va contenir que quelques journées en 2014 et l'ensemble des journées de 2015 mais sur une période moins étendue. Les journées du cluster C5 non affectées au cluster L5 sont affectées au cluster L4.

Ces convergences/divergences entre les deux clustering peuvent être analysés plus finement à l'aide de la Figure 3.5. L'ensemble des courbes présentes dans chaque cluster, ainsi que les liens établis pour passer d'un clustering à l'autre sont présentés sur la figure. Chaque couleur correspond à un cluster de type de jour, avec le même code couleur que pour la figure précédente. À gauche du graphique se trouvent les courbes de validations réparties par cluster selon l'approche "libre". À droite ce sont les courbes réparties selon le clustering "contraint" qui sont tracées. La partie centrale du graphique fait le lien entre les deux classifications. Elle met en évidence la part des jours communs aux clusters des deux méthodes et ceux qui changent de cluster selon la méthode choisie.

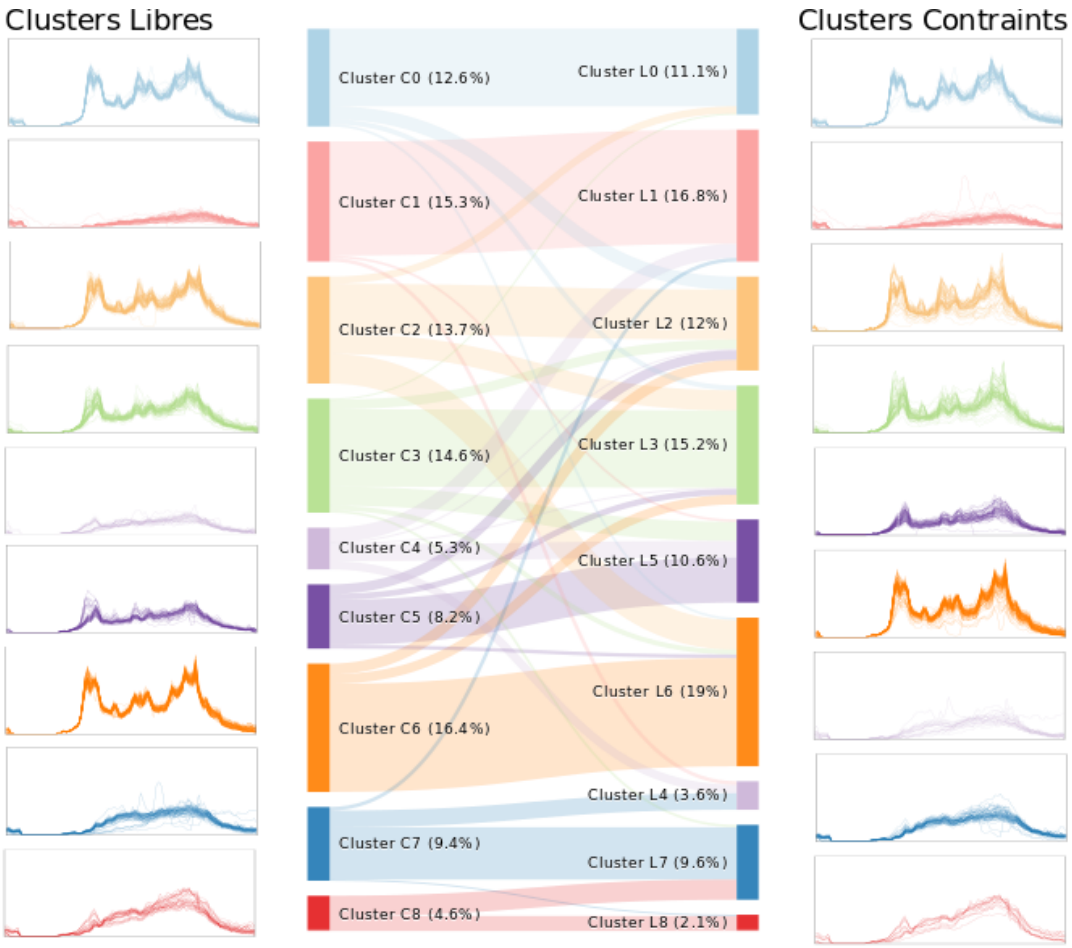


FIGURE *3.5 – Graphique représentant les correspondances entre les deux clusterings (libre et contraint par les types de jours). Les courbes de validations classées sont représentées à gauche pour le clustering libre et à droite pour le clustering contraint.*

Trois types de liens sont observables entre les clusters :

— Les plus visibles sont les liens "forts", c'est à dire lorsque la majorité des éléments affectés à un cluster est également affectée au cluster qui lui est associé dans l'autre méthodologie de clustering. C'est le cas pour les clusters libre 0 (Mercredi et Vendredi de semaine travaillée), 1 (Dimanche et jours fériés), 3 (jours de semaines de Mai et petites vacances), 5 (jours de semaines de Juillet et Août), 6 (Lundi, Mardi et Jeudi de semaine travaillée) et 7 (Samedi de semaine travaillée) qui partagent tous plus de 65% de leurs éléments avec leur cluster associé. Le restant des courbes se répartit dans les autres clusters et ajoute de la variabilité à ceux-ci. La forme générale de deux clusters associés reste donc semblable bien qu'une plus grande variabilité soit observable pour les groupes formés par clustering contraint.

— Des liens "intermédiaires" sont également observables. Ainsi, les clusters 2 (jours de semaines de Mai) et 8 (vacances de Noël et jours divers) partagent entre 43 et 47% de leurs éléments avec leur cluster contraint associé.

— Enfin, on note la présence de liens "faibles", c'est à dire des liens pour lesquels les éléments partagés par les deux clusters ne sont pas suffisamment nombreux pour que cela soit significatif. Le cluster 4 correspond aux jours de semaine du mois d'Août et à quelques samedis des mois de Juillet et Août dans le clustering libre alors qu'il est associé aux dimanches du mois de décembre et aux samedis de Juin et Juillet dans le clustering contraint. Le cluster libre 4 se sépare principalement dans deux clusters, ses éléments vont se trouver principalement dans le cluster contraint 1, composé majoritairement de dimanches et dans le cluster contraint 5, composé des jours de semaines sur la période Juillet Août. Quant au cluster 4 contraint il est majoritairement composé d'éléments du cluster 7 libre, samedis hors vacances d'été.

Pour quantifier ces liens, le tableau de contingence obtenu en croisant les deux méthodes de clustering est fourni à l'aide du tableau 3.1. Les deux classifications sont in-fine impactées par la période de forte anomalie identifiable sur l'année 2014. En effet un grand nombre de journées pour lesquelles les données sont manquantes se situent sur les mois de Juin à Octobre 2014. Les deux classifications possèdent un calendrier qui ne correspond pas à ce que l'on pourrait attendre pour une période travaillée (Septembre-Octobre). Le clustering contraint ne prenant pas en compte les différences d'une année à l'autre, il est impacté par ces anomalies sur les deux années. Cela est particulièrement visible sur la période de Septembre-Octobre, où le motif typique d'activité de semaine n'est pas apparent.

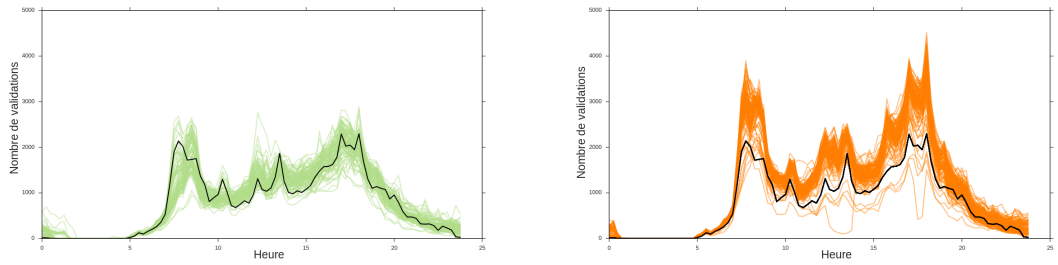
Choix d'une classification pour la détection d'anomalie

Considérant l'étape de classification comme tâche préliminaire à la détection d'anomalie, la classification des jours a pour but de regrouper les jours avec une activité semblable, afin de faciliter la détection des journées dont l'activité est atypique. Se pose alors la question du choix de la méthode de clustering : avec ou sans contrainte. En raison de la plus faible variabilité intra-clusters pour la méthode de clustering libre sur certains clusters (notamment le 6), on aurait tendance à choisir d'effectuer la détection d'anomalie sur les clusters libres. Cependant ce choix présente un inconvénient majeur lié au fait

libre/contraint	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
L0	**66**	0	11	4	0	0	2	0	0	
L1	0	**96**	0	0	3	2	0	0	0	
L2	6	0	**42**	17	0	0	25	0	0	
L3	1	0	8	**65**	0	16	4	2	0	
L4	0	**12**	1	1	7	**14**	0	0	0	
L5	0	0	8	5	0	**38**	3	0	0	
L6	0	0	9	8	0	0	**91**	0	0	
L7	0	3	0	0	14	0	0	**44**	1	
L8	0	0	0	0	0	0	0	**17**	**13**	

TABLE *3.1 – Tableau de contingence de l'appartenance des jours au clusters pour les deux approches de clustering : libre (ligne) et contrainte (colonnes).*

que les journées en anomalie vont avoir une activité qui va différer de l'activité habituelle attendue, et du fait du clustering sans contrainte, ces journées vont être classées dans le cluster ayant une activité la plus proche de la leur, masquant ainsi leur comportement atypique. Pour mettre en évidence ce problème, l'exemple de la journée du 7 Octobre 2014 est étudié. Il s'agit d'un jour standard sans vacances en anomalie. Celle-ci a été affectée au cluster libre 3 (petites vacances) et au cluster contraint 6 (Lundi, mardi et jeudi travaillés). Sur la figure 3.6 sont représentées la courbe d'activité du 7 Octobre parmi les courbes du cluster libre 3 (3.6a), et parmi les courbes du cluster contraint 6 (fig 3.6b). On peut remarquer que la journée risque difficilement d'être détectée comme étant en anomalie au sein du cluster libre 3, car très semblable aux courbes environnantes, alors qu'elle le serait pour le cluster contraint 6. Cet exemple met clairement en évidence l'intérêt du clustering avec contrainte.

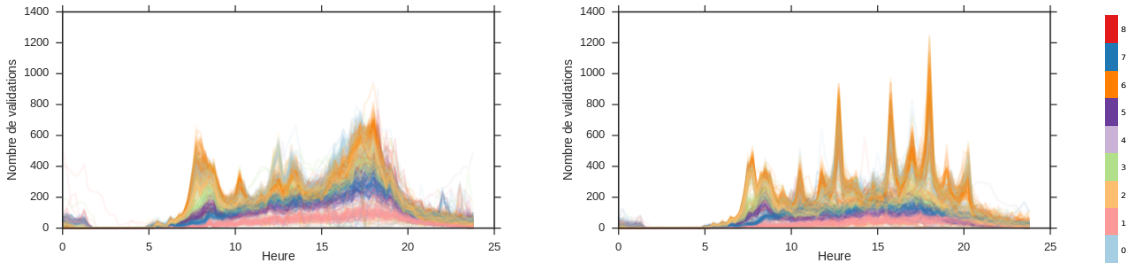


(a) Cluster Libre 3 *(b) Cluster Contraint 6*

FIGURE *3.6 – Courbe du nombre de validations enregistrées à la station République la journée du 7 Octobre 2014.*

La figure 3.7, présente les courbes du nombre de validations enregistrées pour les stations République et Villejean Université (figure 3.1) avec un code couleurs correspondant aux résultats du clustering contraint. Notons tout d'abord que les différents groupes de courbes sont bien différenciés et homogènes. Seules quelques courbes isolées correspondant probablement à des événements inhabituels, se détachent des différents groupes. Notons ensuite, que les clusters correspondant à des jours travaillés de semaine vont regrouper les courbes avec le plus grand nombre de validations tandis que le cluster avec les courbes de

plus faible activité est le cluster regroupant les dimanches et jours fériés.



(a) République *(b) Villejean Université*

FIGURE *3.7 – Nombre de validations enregistrées par quart d'heure sur les stations République et Villejean Université pour la période allant du 1er Juin 2014 au 29 Mai 2016.*

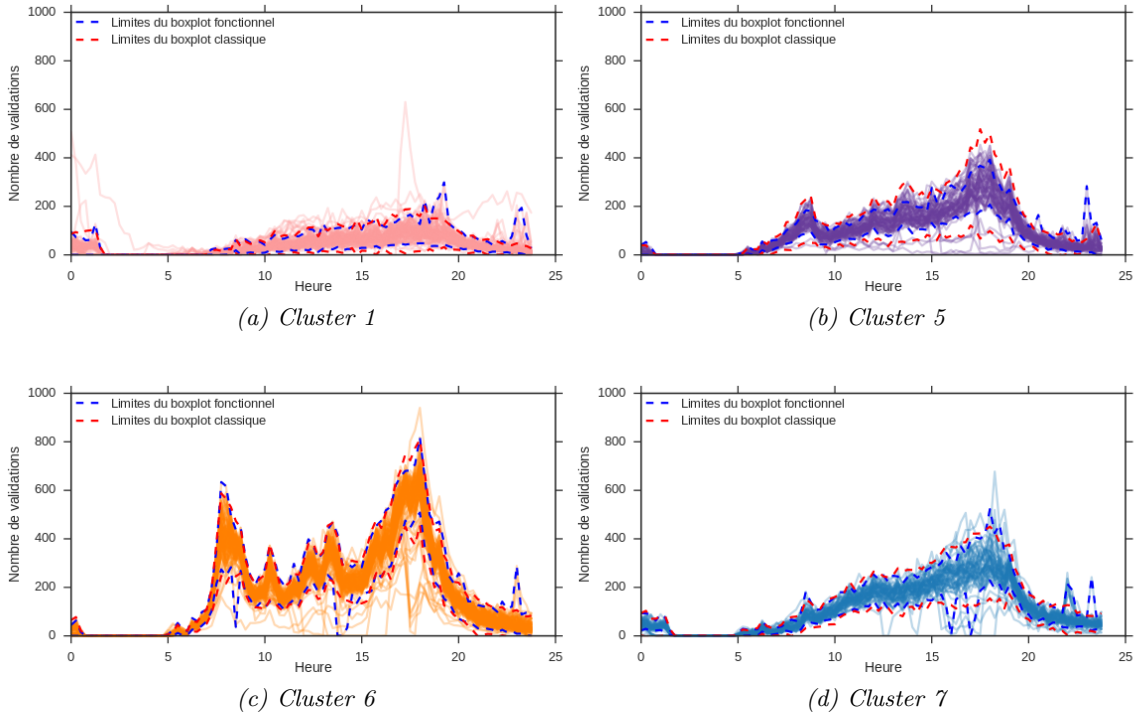
Dans la section suivante, ces clusters sont ceux utilisés pour la détection d'événements atypiques.

Comparaison des approches de détection

Les résultats des deux approches de détection (boxplot et boxplot fonctionnel) sont présentés sur la figure 3.8 pour la station République. Pour faciliter la lecture de ces résultats, seul quatre clusters sont représentés, à savoir le cluster 1 (dimanches et jours fériés), le cluster 5 (vacances d'été), le cluster 6 (lundis, mardis et jeudis travaillés) et le cluster 7 (samedis classiques). Sur chaque figure est représenté l'ensemble des courbes de validations à la journée pour le cluster concerné ainsi que les bornes supérieures et inférieures du nombre de validations considéré comme correspondant à un usage normal (courbes en pointillé). Toute courbe sortant de cette zone est considérée en anomalie et tous les points qui y sont extérieurs des outliers. Deux types de bornes sont représentées : la borne obtenue par l'approche du boxplot en chaque quart d'heure (en rouge) et la borne obtenue à l'aide du boxplot fonctionnel grâce à l'ensemble des courbes de la journée (en bleu).

On peut constater que les événements atypiques positifs, correspondant à un nombre de validations plus élevé que prévu, sont bien détectés. En revanche pour les événements atypiques négatifs les résultats semblent moins pertinents pour les clusters 7 (constitués de samedis) et 5 (vacances d'été). Lorsqu'on compare les résultats obtenus par les deux approches on observe que l'approche par boxplot fonctionnel va bien suivre le groupe de courbes où la densité est la plus forte, tandis que l'approche par boxplot ponctuel va être plus éloignée des courbes, notamment pour le cluster 5. Les anomalies du cluster 7 semblent être sous-évaluées par les deux méthodes lorsque celles-ci correspondent à un nombre plus faible de validations que la moyenne. On remarque également que bien qu'elle semble plus proche des courbes dans l'ensemble, l'approche par boxplot fonctionnel présente quelques pics qui ne semblent pas correspondre à l'usage moyen, par exemple entre 16H et 18H dans le cluster 7.

A l'analyse de ces graphiques, on note que tous les points éloignés du groupe central des validations sont bien détectés. Cependant, il reste difficile de déterminer pour les points



(a) Cluster 1 *(b) Cluster 5*

(c) Cluster 6 *(d) Cluster 7*

FIGURE *3.8 – Comparaison des deux approches sur la station République. Les courbes représentent le nombre de validations enregistrées pour une journée par quart d'heure. Les courbes en pointillées correspondent aux frontières délimitant les nombres de validations considérés comme dans la norme de ceux considérés comme en anomalie.*

en limite s'ils sont ou non en anomalie. De plus lorsqu'un quart d'heure présente une plus grande variabilité, des points qui pourraient sembler en anomalie au vu de l'ensemble du graphique, sont détectés comme normaux par le boxplot. Les proportions de quarts d'heure détectées comme standard ou en anomalie par les deux méthodes sont reportées sur le tableau 3.2. L'approche fonctionnelle détecte un plus grand nombre d'événements atypiques qui se recoupent en majorité avec ceux détectés par le boxplot ponctuel comme le montre l'analyse de l'union et de l'intersection de ces deux ensembles.

Approche	Anomalie	Standard
Boxplot ponctuel	3.93%	96.07%
Boxplot fonctionnel	5.07%	94.93%
ponctuel ∪ fonctionnel	6.44%	93.56%
ponctuel ∩ fonctionnel	2.56%	97.44%

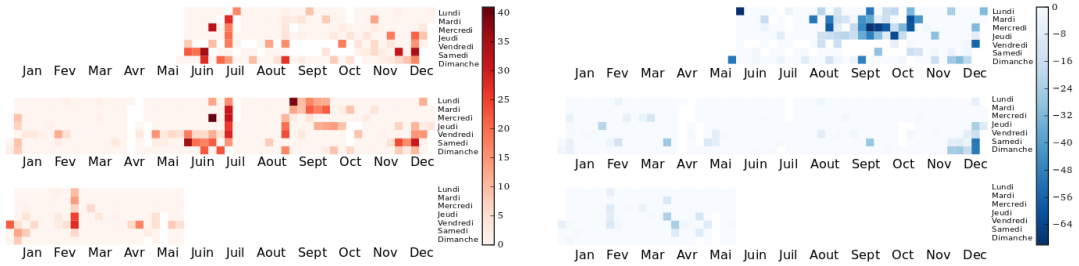
TABLE *3.2 – Tableau de contingence des proportions de quarts d'heure détectés en anomalie ou non par les deux approches : boxplot ponctuel et fonctionnel*

Dans la suite les résultats sur les données détectées par l'approche de boxplot fonctionnel sont présentés. Cette approche étant celle détectant le plus grand nombre d'événements possiblement atypiques, il a paru judicieux de la conserver.

3.4.2 Analyse détaillée des points en anomalie

Analyse des anomalies à la journée

Une vue calendaire du nombre d'anomalies positives et négatives détectées par le boxplot fonctionnel sur les deux ans de données pour la station République est reportée figure 3.9. Deux remarques peuvent être formulées. On observe des périodes sur lesquelles la station est en anomalie négative plusieurs jours durant, notamment sur la période de Septembre-Octobre 2014. L'analyse des autres stations, non présentée ici, montre que ces périodes d'anomalie sont également présentes sur celles-ci. Pour certains jours sur ces périodes il est même possible de ne disposer d'aucune donnée de validation enregistrée, notamment pour les journées du 15 au 19 avril 2014 et pour le 11 Octobre 2015. Il s'agit probablement d'un dysfonctionnement des validateurs ou d'un problème dans l'enregistrement des données. On note également des jours isolés en anomalie qui nécessitent une analyse plus particulière afin de déceler les causes de ce surplus ou cette baisse de validations. Quelques événements détectés feront l'objet de cette analyse dans le paragraphe suivant.



(a) Positives par boxplot fonctionnel *(b) Négatives par boxplot fonctionnel*

FIGURE *3.9 – Nombre d'anomalies positives (rouge) et négatives (bleu) détectées par boxplot fonctionnel sur la période allant du 1er juin 2014 au 30 mai 2016 (station République).*

Analyse des anomalies au quart d'heure

Dans ce paragraphe une analyse des anomalies à une échelle plus fine en se basant sur l'indicateur défini en section 3.3.3 est menée. Pour cela les cartes de chaleur de l'indicateur d'anomalie au quart d'heure sont tracées pour six stations et six dates différentes. Le choix des dates a été effectué en consultant, parmi les journées détectées en anomalie sur la figure 3.9, celles où a eu lieu un événement particulier dans la ville de Rennes. Les stations d'étude ont été sélectionnées de sorte qu'elles soient représentatives de l'ensemble du réseau. La localisation sur la ligne (nord ou sud de la ligne), l'activité enregistrée (faible ou élevée), la proximité à d'autres stations touchées par des événements sont les critères qui ont été retenus. Les stations d'étude sont :

— Républiques, station la plus centrale et la plus active du réseau.
— Villejean Université, deuxième station la plus active par sa proximité aux établissements d'enseignement supérieur, située au Nord-Ouest de la ligne.
— Sainte-Anne, station proche du centre-ville ayant été l'objet de fermeture.

— Gares, station située en centre-ville à côté de la gare.
— Charles de Gaulle également proche du centre-ville, cette station est souvent le point de départ de nombreux événements sportifs et culturels.
— Clemenceau, station de plus faible activité située plus au Sud de la ligne

Faible remontée de données Comme cela a été constaté précédemment, de nombreux problèmes de remontée de données sont apparus sur la période allant de juillet à novembre 2014. Afin d'illustrer l'efficacité de notre méthode pour la détection de telles anomalies, l'ensemble des courbes de validations du cluster 5 enregistrées sur la station République est représenté sur la figure 3.10. Celle-ci laisse apparaître plusieurs courbes avec une activité qui diffère de l'activité observée sur le reste des courbes. L'exemple le plus marquant est la courbe enregistrée le 19 août 2014 qui est représentée en noir et en pointillée sur la figure. Il est observable que, bien que cette journée corresponde à un mardi de semaine de vacances, aucune validation n'est enregistrée en dehors de la tranche horaire 10h-15h. Or, aucun incident n'est remonté par l'opérateur sur son fil twitter pour cette journée, ce qui suggère que l'activité de cette journée n'a pas été perturbée par des incidents réseaux.

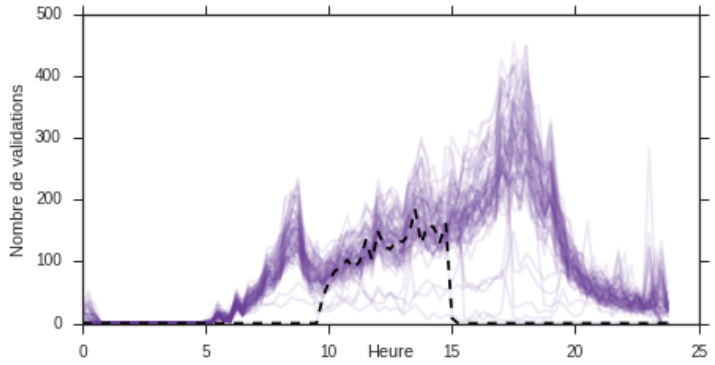


FIGURE *3.10 – Courbes de validations enregistrées sur la station république pour l'ensemble des journées du cluster 5, dont la journée du 19 août 2014.*

La visualisation des indices d'anomalies au quart d'heure sur les différentes stations d'étude (figure 3.11) confirme que ces absences de validations sont présentes sur l'ensemble des stations et doivent correspondre à un problème de remontée de données. Ces problèmes n'étant pas détectables par la simple analyse des anomalies déclarées par l'opérateur, il est nécessaire d'avoir un outil permettant de les détecter, ce qui est bien le cas ici.

Impact des événements culturels et sportifs sur le nombre de validations en station Les événements culturels et sportifs ayant eu lieu dans la ville et leurs impacts sur l'activité du réseau de transport sont analysés. Deux journées pendant lesquelles se sont déroulés des événements sportifs et culturels ont été sélectionnées : le 21 juin 2014 qui correspond à la fête de la musique et le 10 octobre 2015 qui correspond à "tout Rennes court". Les logos de ces deux événements sont présentés en figure 3.12.

En France, tous les 21 Juin a lieu la fête de la musique. C'est une soirée qui génère un grand nombre de sorties en raison du nombre important de spectacles se déroulant en

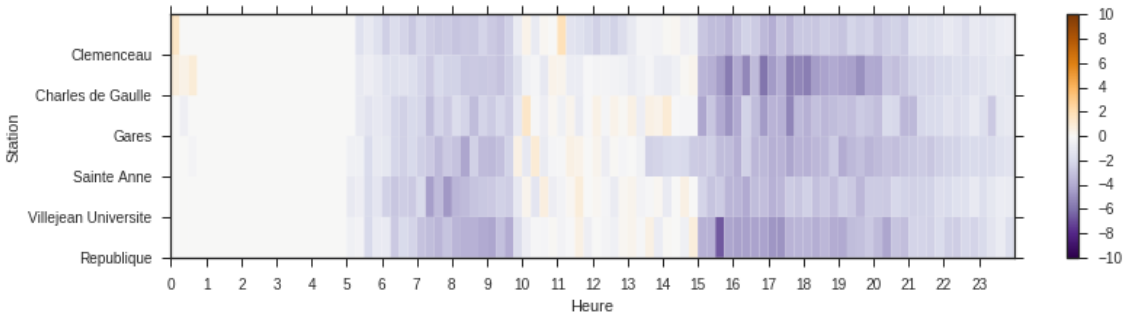


FIGURE *3.11 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 19 août 2014.*

(a) Affiche de la fête de la musique 2014 à Rennes

(b) Logo de la course "Tout Rennes court"

FIGURE *3.12 – Affiche et logo de différents événements sportifs et culturels rennais.*

extérieur. Les indicateurs d'anomalie au quart d'heure précédemment définis,

$$I_{fin}(s,d,i) = \frac{x_{s,d,i} - M(\{x_{s,e,i} : k(e) = k(d)\})}{\mathcal{Q}_3(\{x_{s,e,i} : k(e) = k(d)\}) - \mathcal{Q}_1(\{x_{s,e,i} : k(e) = k(d)\}),}$$

ont été présentés pour les 6 stations d'étude sur la figure 3.13. Pour en faciliter la lecture, les indicateurs sont tronqués au-delà de l'intervalle [-10,10], ce seuil correspondant déjà à une anomalie majeure. L'étude de la figure permet de mettre en évidence trois résultats. Tout d'abord, on observe une augmentation inhabituelle de l'activité sur l'ensemble des stations à l'exception de la station Gares. Ensuite il apparaît que le service du métro est plus tardif qu'à l'ordinaire puisque des validations sont enregistrées jusqu'à 1h30 du matin. Enfin on voit que la station Clemenceau qui est la plus au Sud des stations est également celle dont l'activité s'arrête le plus tôt dans la soirée.

Tous les ans un événement sportif nommé "Tout Rennes Court" et regroupant plusieurs courses, dont un 10km et un semi-marathon, se déroule à Rennes. En 2015 cet événement a pris place le 11 Octobre. Sur la figure 3.14 se trouvent les indicateurs d'anomalie pour les six stations d'étude. On observe un surplus d'activité inhabituelle sur la station Charles

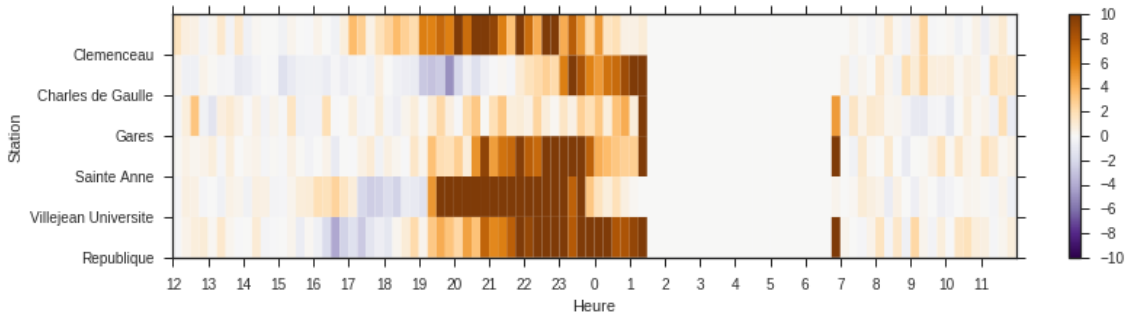


FIGURE *3.13 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations dans la nuit du 21 juin au 22 Juin 2014.*

de Gaulle, ce qui concorde avec le fait que le départ des différentes courses se situait à proximité de cette sation. Il est donc normal d'observer une recrudescence d'activité dans cette zone. On observe également une forte activité, un peu plus tôt dans la journée sur les stations Villejean Université et Clemenceau. On peut imaginer que cela correspond à des déplacements de voyageurs se rendant potentiellement sur le départ de la course.

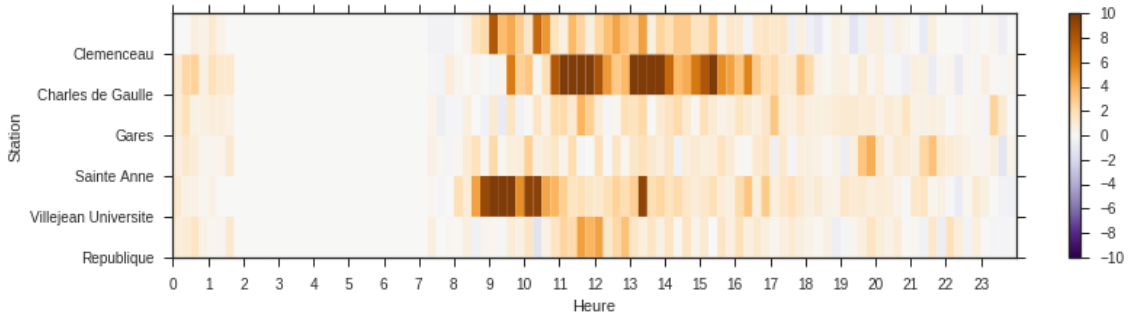
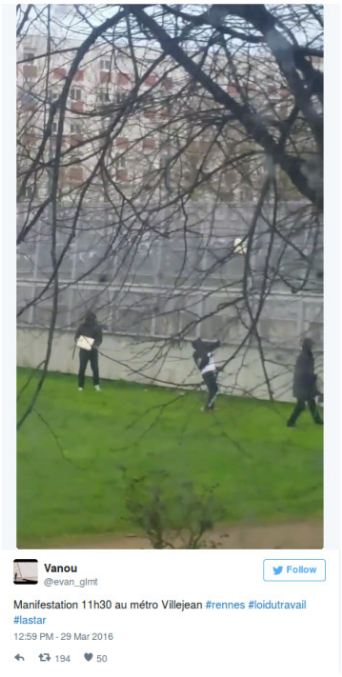


FIGURE *3.14 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 11 Octobre 2015.*

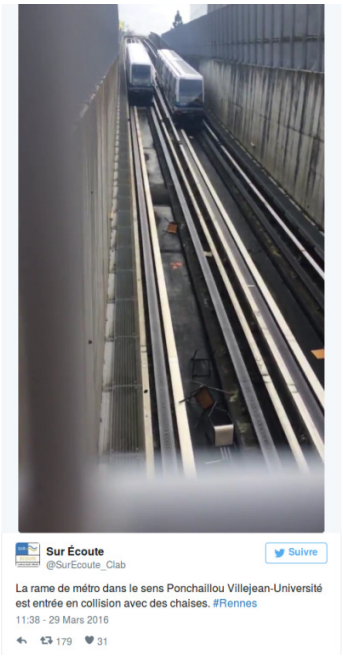
Impact des mouvements sociaux sur le nombre de validations en station À présent qu'il a été montré que des événements sportifs et culturels peuvent créer un regain d'activité dans la ville et plus particulièrement l'activité sur le réseau de transport en commun, l'attention est portée sur des événements inverses qui auront pour effet de provoquer une baisse de l'activité. Ici un intérêt particulier est porté aux journées des mois de mars, avril et mai 2016, qui sont particulières car elles correspondent à une période de grande contestation en France. En effet durant cette période un remaniement du code du travail a eu lieu, engendrant ainsi un grand nombre de mouvements sociaux et de manifestations. L'impact de ces événements sur la mobilité des usagers du réseau de transport en commun est ainsi plus visible qu'en période classique.

Le 29 mars, des étudiants ont jeté des chaises sur les voies du métro pour manifester leur désaccord avec le futur texte de loi (Figure 3.15). Cet incident a généré deux heures d'arrêt de la circulation sur l'ensemble de la ligne de métro avec une reprise du trafic vers 14H (Figure 3.16). La carte de chaleur des indicateurs d'anomalie au quart d'heure obtenus pour cette journée sur les six stations d'étude montre que l'anomalie a commencé et s'est

terminée au même moment sur l'ensemble des stations (Figure 3.17). Elle apparaît comme ayant duré 2H30 (de 11H30 à 14H). Les informations fournies par l'information trafic de l'opérateur transport, Keolis, semble indiquer une fin de l'incident aux alentours de 13h30. Cela indique que le retour à la normale de l'activité est bien détecté par l'indicateur.



(a) Des manifestants jettent des chaises sur les voies du métro à proximité de l'arrêt Villejean Université.



(b) Le métro rentre en collision avec les chaises qui ont été jetées sur les voies à l'arrêt Villejean Université

FIGURE *3.15 – Images partagées sur twitter montrant les dommages occasionnés par les manifestants sur le réseau métro de l'agglomération rennaise.*

FIGURE *3.16 – Tweet publié par l'opérateur du réseau annonçant la reprise de la desserte de la station République par le métro le 29 mars 2016.*

Le 31 mars est une journée de grève contre la loi travail. À cette occasion, une manifestation a lieu en centre-ville. En raison de sa proximité avec le tracé du défilé et donc du passage du cortège de la manifestation, la station République est fermée puis réouverte un peu plus tard dans la journée comme le signale le tweet de l'opérateur de transport (figure 3.18). Une forte baisse d'activité sur la station République, correspondant à sa fermeture est en effet mise en évidence par la carte de chaleur obtenue à cette date (cf. Figure 3.19). On note également une forte hausse de l'activité sur les stations avoisinantes Sainte Anne et Gares. Il y a donc un report de l'activité de République vers ses stations voisines.

Le 28 avril une autre manifestation a lieu en centre-ville. Pour des raisons de sécurité la station Sainte Anne puis la station République sont fermées. Elles sont ré-ouvertes plus

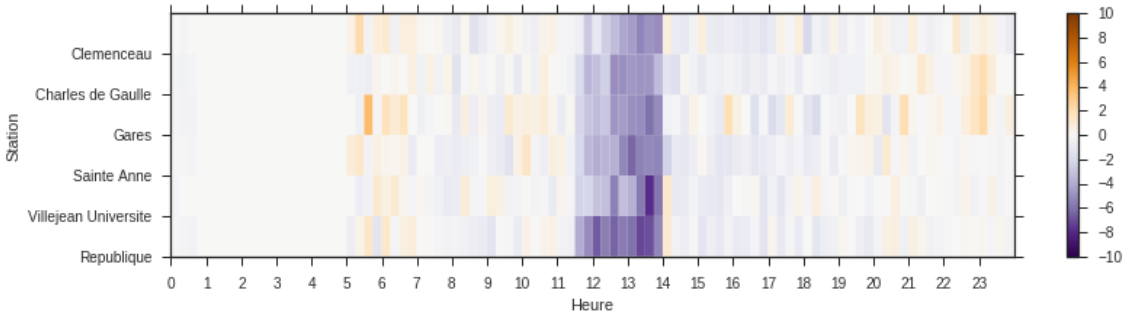


FIGURE *3.17 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 29 mars 2016.*

FIGURE *3.18 – Tweet publié par l'opérateur du réseau annonçant l'arrêt de la desserte de la station République par le métro le 31 mars 2016.*

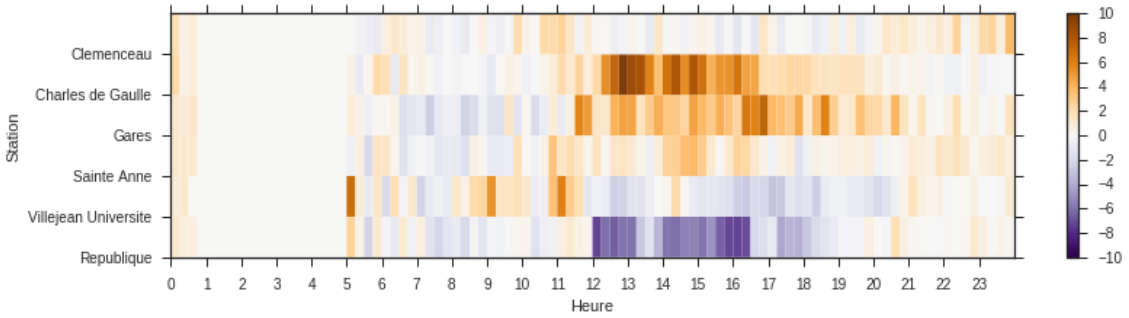


FIGURE *3.19 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 31 mars 2016.*

tard dans la journée (figure 3.20). Ces événements se traduisent encore une fois par des baisses du nombre de validations de ces deux stations (figure 3.21). On observe également le report de validations qui se crée sur les stations voisines : Gares et Charles de Gaulle.

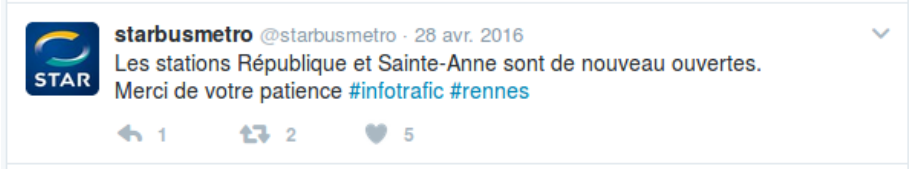


FIGURE *3.20 – Tweet publié par l'opérateur du réseau annonçant la reprise de la desserte des stations République et Sainte Anne par le métro le 28 avril 2016.*

D'autres manifestations ont lieu le 26 mai. Un dépôt de bus est bloqué une grande partie de la journée. En ce qui concerne les stations de métro, la station Charles de Gaulle, départ du cortège de la manifestation, est fermée. Une reprise du service a lieu en fin d'après-midi (figure 3.22). Ce changement de station permet d'observer un autre type de report d'activité. Sur la figure 3.23, on constate que les stations impactées sont

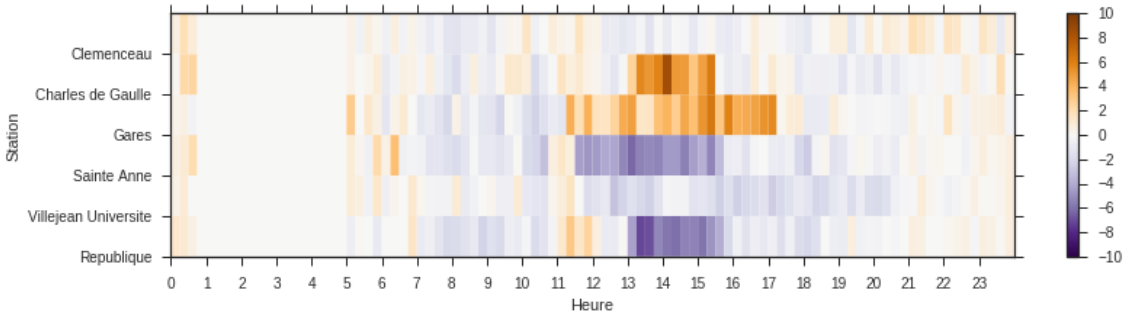


FIGURE *3.21 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 28 avril 2016.*

différentes puisque seule la station Gares a une augmentation inhabituelle de son activité.

FIGURE *3.22 – Tweet publié par l'opérateur du réseau annonçant la reprise de la desserte de la station Charles de Gaulle par le métro le 26 mai 2016.*

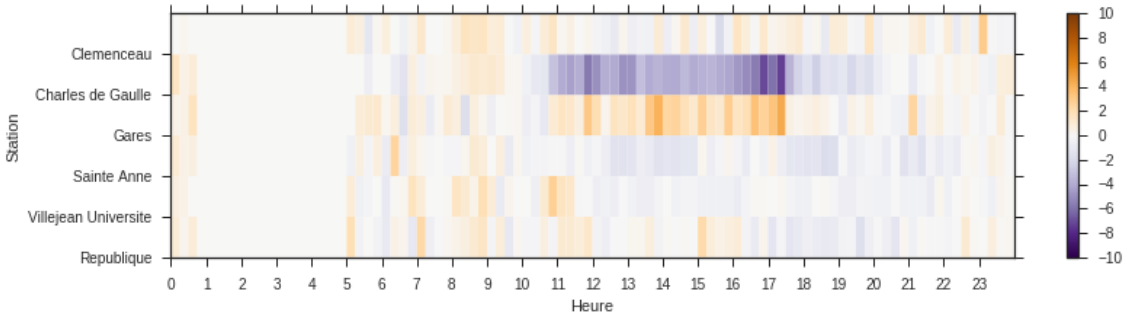
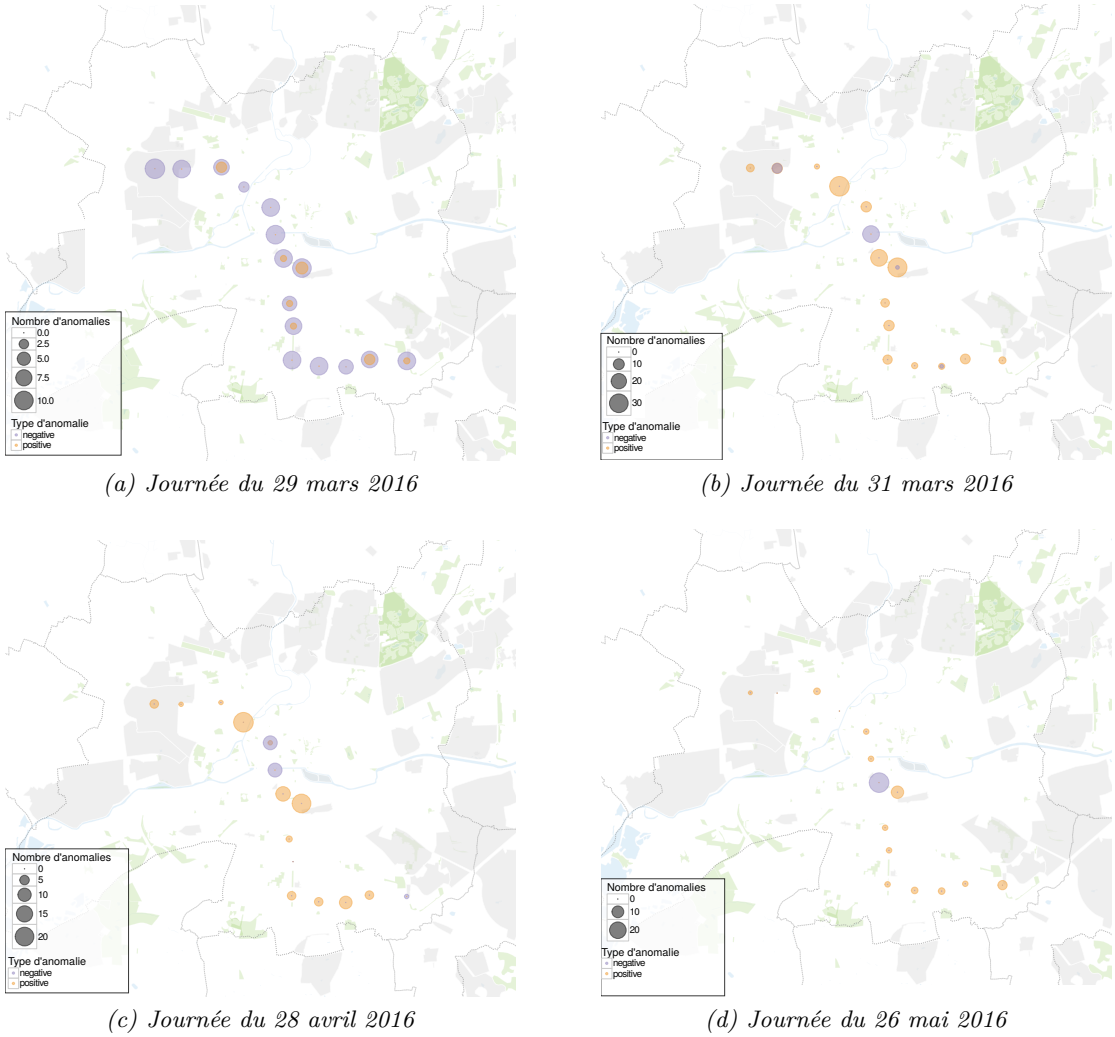


FIGURE *3.23 – Carte de chaleur des indicateurs d'anomalie positifs (orange) et négatifs (violet) enregistrés sur six stations le 26 mai 2016.*

Enfin une carte de la ville représentant le nombre d'anomalies enregistrées par station sur ces mêmes journées est tracée (figure 3.24). La figure 3.24a permet d'observer que l'arrêt de la ligne de métro a bien impacté tout le réseau. On observe également un certain nombre d'anomalies positives sur les stations les plus actives du réseau, telles que République ou Villejean-Université. Ces figures permettent également de visualiser spatialement les reports de validations dus à des fermetures de station pour la journée du 31 mars (figure 3.24b), 28 avril (figure 3.24c) et 26 mai (figure 3.24d). On constate en effet que ce sont toujours les stations à proximité de la station fermée qui ont un nombre d'anomalies positives plus élevé. On peut noter que la fermeture de la station République impacte plus fortement le reste du réseau que la fermeture de la station Charles de Gaulle.

L'analyse des résultats montre qu'un nombre important d'anomalies présentes sur le réseau de transport est souvent lié à une activité inhabituelle dans l'agglomération. Croiser les observations en anomalie avec les informations sur les événements ayant lieu dans la



(a) Journée du 29 mars 2016

(b) Journée du 31 mars 2016

(c) Journée du 28 avril 2016

(d) Journée du 26 mai 2016

FIGURE *3.24 – Carte du nombre d'anomalies positives (jaune) et négatives (violet) enregistrées par station de métro sur les journées du 29-31 mars, 28 avril et 26 mai 2016.*

ville peut donc à terme aider le gestionnaire du réseau et la municipalité à mieux anticiper la réaction des usagers aux différentes situations et ainsi leur proposer des solutions correspondant à leurs attentes.

Conclusion

Ce chapitre a présenté une méthodologie permettant la détection d'événements atypiques impactant l'exploitation d'un réseau de transport en commun à partir de données billettiques. Les événements atypiques incluent également les problèmes de mesure où la remontée de données est défectueuse alors que l'exploitation et la demande sont habituelles. Une difficulté majeure dans cette problématique est liée à la grande variabilité présente dans les données pour un fonctionnement nominal. Cette variabilité, due à des modifications de contexte (type de jour, vacances, jour férié ...) est inhérente à toute demande dans un système de transport en commun. Détecter de façon robuste une situation atypique à partir des données billettiques suppose que la méthodologie proposée soit capable de s'af-

franchir des variations dues à ces variables contextuelles. Pour répondre à ces contraintes, l'approche proposée opère en deux étapes successives : un clustering pour identifier les types de jour sur l'ensemble des stations et une détection qui opère par station et par type de jour identifié lors du clustering.

Deux méthodes de détection ont été évaluées : la détection par quantiles et la détection par boxplot fonctionnel. Des indicateurs de détection à l'échelle de la journée ou du quart d'heure ont été proposés autorisant ainsi une quantification de l'impact de l'événement en termes de durée et d'augmentation (anomalie positive) ou de baisse (anomalie négative) des fréquentations par rapport à une situation nominale. La méthode autorise également une mise en évidence de l'influence spatiale d'un événement sur le réseau. L'analyse des résultats pour certains événements importants ayant lieu dans la ville a été également menée pour montrer l'impact de ces événements sur la demande de transport.

Ces travaux peuvent être poursuivis dans plusieurs directions. Une première extension sur la brique de classification pourrait consister à créer des catégories non homogènes pour les stations ayant des spécificités. C'est le cas par exemple des stations à proximité des universités où la prise en compte du calendrier universitaire serait pertinente. Les deux approches de détection d'anomalies utilisées sont affaiblies par un trop grand nombre d'anomalies, et ce malgré la robustesse de l'estimateur utilisé. Il pourrait être intéressant d'un point de vue méthodologique de s'orienter vers des méthodes permettant de surpasser cette difficulté, par exemple en augmentant le nombre de clusters afin de limiter le nombre de courbes en anomalie par cluster. Enfin, seules les validations enregistrées sur la ligne de métro ont été prises en compte dans ce chapitre. Une extension de la méthodologie à d'autres modes de transport (bus par exemple) serait pertinente et d'un grand intérêt pour les opérateurs de transport. L'approche proposée ne nécessite pas de labellisation des données, mais il serait intéressant de croiser les résultats avec une base incidents de l'opérateur.

Chapitre 4

Classification pour l'identification de profils temporels types d'usagers de transport public

Que ce soit, pour l'opérateur, pour les décideurs politiques ou même pour les usagers de transports en commun, la connaissance des habitudes temporelles des voyageurs du réseau est intéressante. Celle-ci autorise une connaissance fine des habitudes des usagers qui pourrait être mise à profit pour adapter l'offre à la demande de certains groupes spécifiques, redéfinir les aménagements afin qu'ils soient en lien avec les flux de voyageurs, ou bien éviter les périodes de fortes affluences liées à certains groupes d'usagers. Ce chapitre décrit une méthodologie permettant d'offrir une meilleure connaissance des habitudes temporelles des usagers de transports en commun, et de leur évolution dans le temps. Cette méthodologie cherche également à rattacher ces habitudes à des zones géographiques spécifiques.

Dans un premier temps, les travaux sont contextualisés parmi les différentes approches ayant été mises en place pour l'analyse et le suivi temporel des usagers. Le but est de mettre en relief les contributions nouvelles apportées par la méthode proposée dans cette thèse. Ensuite ce chapitre présente le modèle de mélange de gaussiennes pour la classification des habitudes temporelles des usagers. Ce modèle constitue un des principaux apports méthodologiques de cette thèse. Puis la méthode est appliquée sur les deux cas d'étude, Rennes et Gatineau. L'analyse de ces résultats sur données réelles constitue une autre contribution importante de la thèse car elle permet de détailler des habitudes temporelles des usagers et de leur évolution au fil du temps. Enfin pour chacun des axes d'étude, les résultats obtenus sont croisés avec les informations spatiales disponibles, afin de caractériser les liens potentiels entre type d'usage et organisation spatiale de la ville.

4.1 Positionnement face à l'état de l'art

Le chapitre 1 section 1.3.3 a présenté un état de l'art détaillé des approches développées pour la classification des usagers selon leurs habitudes temporelles. La majorité des approches citées reposent sur une discrétisation du temps (par exemple utiliser une segmentation par heure ou sur une période d'intérêt définie, telle que le pic d'activité matinal, le pic d'activité en soirée, etc.). Un des inconvénients de cette représentation est qu'elle ne capture pas entièrement la régularité des déplacements. Par exemple, un usager qui effectue la navette domicile-travail quotidiennement entre 8h55 et 9h05 verra ses validations divisées dans deux intervalles de temps distincts (8h et 9h) si une segmentation d'une heure est définie. Cela peut être mal interprété comme un usage diffus (ce qui n'est clairement pas le cas). Ce problème est résolu par l'approche proposée dans ce chapitre. Sa nouveauté réside en effet dans le fait qu'elle considère un mélange de gaussiennes à deux niveaux pour préserver la nature continue du temps. De plus les résultats obtenus grâce au clustering sont utilisés pour suivre l'évolution des habitudes des usagers au fil du temps. Contrairement aux travaux de [Langlois et al. 2016] qui se plaçaient sur une période de 4 semaines, l'étude proposée sur le cas d'étude de Gatineau se place sur 5 ans de données, permettant ainsi de suivre l'évolution des affectations des cartes dans les différents clusters.

4.2 Méthode

Le clustering des usagers basé sur leur activité temporelle est un sujet intéressant pour l'extraction de connaissances depuis les données billettiques. En effet, la présence de groupes d'usagers avec une activité similaire peut révéler les schémas de déplacement les plus fréquents dans un réseau de transport en commun et peut contribuer ainsi à une meilleure caractérisation de la demande.

L'approche de clustering d'usagers proposée dans cette thèse est détaillée dans cette section. Le modèle génératif qui intègre une représentation continue du temps est décrit. L'algorithme proposé pour estimer ces paramètres est également détaillé.

4.2.1 Introduction aux approches de clustering

L'apprentissage non supervisé regroupe l'ensemble des méthodes permettant d'inférer la structure d'un échantillon, et ce alors qu'aucune information sur celle-ci n'est disponible. Deux des approches les plus utilisées sont la réduction de dimensions (Analyse en Composantes Principales) et la classification des observations d'un échantillon en sous-groupes homogènes (clustering).

Approches classiques de clustering. Le clustering regroupe des observations qui sont définies comme similaires. Les approches les plus classiques de clustering sont le k-means et le clustering hiérarchique. Le k-means utilise la notion de centre de classe [MacQueen 1967]. Il affecte chaque observation à la classe dont le centre est le plus proche. Après chaque étape d'affectation, le centre de chaque classe est alors recalculé en

faisant la "moyenne" des observations qu'il contient. Le clustering hiérarchique va quant à lui initialiser l'ensemble de ses classes à l'aide d'une unique observation [Jr. 1963]. À l'initialisation, autant de classes que d'observations sont donc présentes. À chaque étape, les deux classes les plus proches selon une distance définie par l'utilisateur vont être fusionnées. Le processus se continue jusqu'à ce que toutes les classes aient fusionné en une seule. Dans le cadre de ces travaux, ces approches ne suffisent pas à classer les données de validation des usagers. En effet nous souhaitons posséder une classification, mais également une modélisation de l'usage de chaque classe. De plus cette modélisation doit dépendre de certains paramètres (type de jour). Les approches par modèle de mélanges ont donc été utilisées, car elles offrent cette double compétence [McLachlan et Krishnan 2008].

Clustering par les modèles de mélange. Les modèles de mélange sont des modèles probabilistes qui identifient les sous-populations présentes dans un échantillon et les modélisent par des mélanges de distributions. Le modèle de mélange le plus couramment utilisé est le mélange de gaussiennes.

Soient $x_1, ..., x_N$ les réalisations de N vecteurs aléatoires $X_1, ..., X_N$ indépendants et identiquement distribués. Le vecteur parent de ces vecteurs suit alors une distribution qui est un mélange de K distributions caractérisées par des fonctions de densité de probabilité $f(.|\Phi_k)$, $1 \leq k \leq K$. Ces K distributions ont pour paramètres (inconnus) $\Phi_1, ..., \Phi_K$ et sont mélangées selon les proportions $\pi_1, ..., \pi_K$, $\forall k, 0 < \pi_k < 1$ et $\pi_1 + ... + \pi_K = 1$.

La densité de probabilité en $y \in \mathbb{R}^D$ s'écrit alors

$$p(y|\theta) = \sum_{k=1}^{K} \pi_k f(y|\Phi_k),$$

avec $\theta = (\pi_1, ..., \pi_K, \Phi_1, ..., \Phi_K)$ le vecteur des paramètres inconnus du mélange.

On peut calculer la vraisemblance de θ :

$$p(x|\theta) = p(x_1, ..., x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f(x_i|\Phi_k).$$

Comme cela a été énoncé plus tôt, on supposera pour la plupart des données quantitatives sur lesquelles un tel modèle va être appliqué qu'elles sont issues d'un mélange de gaussiennes. La densité en y peut alors s'écrire :

$$f(y|\Phi_k) = (2\pi)^{(-d/2)} det\,(\Sigma_k)^{-1/2}\, e^{-1/2(y-\mu_k)'\Sigma_k^{-1}(y-\mu_k)}.$$

Algorithme EM et application aux modèles de mélange. La procédure standard pour estimer les paramètres d'un modèle de mélange consiste à utiliser les algorithmes de type EM (Expectation Maximisation). L'algorithme consiste en une procédure d'estimation itérative pour laquelle on commence par initialiser les paramètres $\theta^{(0)}$ et qui à chaque itération recalculera les nouveaux paramètres $\theta^{(q+1)}$ qui maximisent $Q(\theta, \theta^{(q)})$, avec $Q(\theta, \theta^{(q)}) = \mathbb{E}\left(\log p(x, Z|\theta)|x, \theta^{(q)}\right)$ l'espérance de la log-vraisemblance complète et $p(x, Z|\theta)$ la distribution a posteriori des données manquantes Z.

L'étape E consiste donc à calculer les composantes de l'espérance $Q(\theta, \theta^{(q)})$ qui ne dépendent pas de θ. L'étape M met à jour les paramètres par

$$\theta^{(q+1)} = \ arg\max_{\theta} Q(\theta, \theta^{(q)}).$$

Dans le cadre d'une application aux modèles de mélange, les informations manquantes sont les classes Z, c'est à dire les classes des observations $(z_1, ..., z_N)$ ainsi que les paramètres du modèle $\theta = (\pi_1, ..., \pi_K, \Phi_1, ..., \Phi_K)$.

L'espérance à maximiser à l'étape $(q+1)$ s'écrira alors :

$$\begin{aligned} Q(\theta, \theta^{(q)}) &= \mathbb{E}\left(\sum_{k=1}^{K}\sum_{i\in P_k} \log(\pi_k f(x_i|\Phi_k)|x, \theta^{(q)})\right) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{E}\left(Z_{i,k}|x, \theta^{(q)}\right) \log(\pi_k f(x_i|\Phi_k) \end{aligned}$$

Lors de l'étape E on fait les calculs qui ne dépendent pas de θ, c'est à dire pour $1 \leq i \leq N$, $1 \leq k \leq K$ on a

$$t_{ik}^{(q+1)} = p(Z_{ik} = 1|x, \theta^{(q)}) = \frac{\pi_k^{(q)} f(x_i|\Phi_k^{(q)})}{\sum_{h=1}^{K} \pi_h^{(q)} f(x_i|\Phi_h^{(q)})}$$

la probabilité d'appartenance des observations aux classes a posteriori.

Pour l'étape M on cherche

$$\theta^{(q+1)} = \ arg\max_{\theta} \sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}^{(q+1)} log\pi_k + \sum_{k=1}^{K}\sum_{i=1}^{N} t_{ik}^{(q+1)} log f(x_i|\Phi_k).$$

Dans le cas des mélanges gaussiens on sera donc amené à résoudre

(1) $\pi_k^{(q+1)} = \dfrac{n_k^{(q+1)}}{N}$

(2) $\begin{cases} \mu_k^{(q+1)} = \dfrac{1}{n_k^{(q+1)}} \sum\limits_{i=1}^{N} t_{ik}^{(q+1)} x_i \\ \Sigma_k^{(q+1)} = \dfrac{1}{n_k^{(q+1)}} \sum\limits_{i=1}^{N} t_{ik}^{(q+1)} (x_i - \mu_k^{(q+1)})(x_i - \mu_k^{(q+1)})' \end{cases}$

avec $n_k^{(q+1)} = \sum\limits_{i=1}^{N} t_{ik}^{(q+1)}$

Il peut arriver de tomber sur un maximum local. Pour éviter cela on peut relancer plusieurs fois l'algorithme en changeant l'initialisation $\theta^{(0)}$.

Algorithme CEM. L'algorithme CEM est un algorithme de maximisation de la vraisemblance classifiante. Il recherche la classification Z et les paramètres θ optimaux au sens de la vraisemblance classifiante. Pour cela il reprend les étapes d'un EM classique mais recherche simultanément la classification et les paramètres θ en maximisant un certain critère.

Il utilise le critère de vraisemblance classifiante, c'est à dire la vraisemblance des don-

nées complètes :

$$C(z, \theta) = \log p(x, z|\theta).$$

Dans le cas d'un modèle de mélange on a donc

$$C_{mel}(z, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} z_{ik} log(\pi_k f(x_i|\Phi_k)).$$

Et nous avons également la vraisemblance classifiante restreinte définie par :

$$C_R(z, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} z_{ik} log f(x_i|\Phi_k) = \log p(x, z|\theta).$$

L'algorithme va donc commencer par une initialisation de θ^0. Puis on va effectuer une étape d'affectation, correspondant à l'étape E d'un EM classique (calcul de la probabilité d'appartenance t_{ik}) ainsi qu'à une étape C de classification. On aura $z^{(q+1)} = arg\max_z C_{mel}(z, \theta^{(q)})$. Enfin on effectue une étape de représentation, qui correspond à l'étape M d'un EM. On a $\theta^{(q+1)} = arg\max_\theta C_{mel}(z^{(q+1)}, \theta)$.

4.2.2 Formalisation du modèle

L'approche de clustering est basée sur l'estimation d'un modèle de mélange génératif à deux niveaux. Le premier niveau modélise la répartition des usagers dans les différents groupes (cluster usagers), tandis que le second niveau capture la distribution temporelle des déplacements des usagers pour chaque groupe.

De manière classique ces deux niveaux sont souvent modélisés à l'aide de distributions multinomiales. En ce qui concerne le premier niveau du modèle de mélange, l'appartenance de chaque usager à l'un des K clusters (K fixé a priori) modélisée à l'aide d'une variable aléatoire, notée Z^1, est considérée comme suivant une distribution multinomiale de paramètres $(\pi_1, \ldots, \pi_K)$.

De la même manière, la distribution des heures de déplacements effectués par les usagers appartenant à un cluster donné sera représentée par un mélange de distributions. Dans ce cas, un mélange de gaussiennes est un choix naturel qui s'ajuste bien à la description des habitudes des usagers lorsque la nature continue des heures de validations veut être préservée (càd quand on ne souhaite pas les discrétiser). Avec une telle distribution, les différentes heures typiques d'usage des transports en commun, ainsi que les variances autour de ces pics peuvent être extraites. Plus formellement, ce modèle génératif peut être écrit de la manière suivante :

$$\begin{aligned} Z_i^1 &\sim \mathcal{M}(1, \pi) \ , \\ Z_{ij}^2 | Z_{ik}^1 D_{ijl} = 1 &\sim \mathcal{M}(1, \tau_{khl}) \ , \\ X_{ij} | Z_{ik}^1 Z_{ijh}^2 D_{ijl} = 1 &\sim \mathcal{N}(\mu_{khl}, \sigma_{khl}) \ . \end{aligned}$$

avec :

— Z_i^1 : appartenance du $i^{\text{ème}}$ usager ($i \in \{1, \ldots, M\}$) à l'un des K clusters d'usagers
— Z_{ij}^2 : appartenance du $j^{\text{ème}}$ déplacement ($j \in \{1, \ldots, N_i\}$) effectué par ce passager à l'une des H gaussiennes décrivant le cluster
— D_{ijl} : jour de la semaine auquel le déplacement a été effectué
— X_{ij} : heure du déplacement générée en utilisant la gaussienne correspondante $\mathcal{N}(\mu_{khl}, \sigma_{khl})$.
— $\pi = \{\pi_1, \pi_2, ..., \pi_K\}$: proportion des clusters usager
— τ_{khl} : proportions des gaussiennes par cluster

Il est important de noter que la variable D_{ijl} correspond au jour de la semaine dans le cadre de ce modèle, mais elle pourrait également être remplacée par un autre type de variable catégorielle. Une représentation graphique du modèle est présentée sur la Figure 4.1.

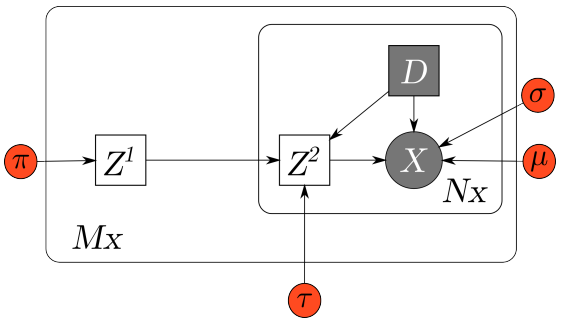


FIGURE *4.1 – Représentation graphique du modèle de mélange de gaussiennes à deux niveaux*

La probabilité conditionnelle de X_{ij} peut alors être écrite comme :

$$\begin{aligned} f(X_{ij} \quad | \quad & \{Z_{ik}^1 Z_{ijh}^2 D_{ijl} = 1\}) \\ = \quad & \sum_{h=1}^{H} \tau_{khd_{ij}} f(x; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) \ , \end{aligned} \tag{4.1}$$

avec $f(.; \mu, \sigma)$ la densité d'une distribution gaussienne de moyenne μ et variance σ^2. La vraisemblance de ce modèle est donnée par :

$$L(\theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k \left(\prod_{j=1}^{N_i} \sum_{h=1}^{H} \tau_{khd_{ij}} f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) \right) \ ,$$

avec M le nombre d'usagers, K le nombre de clusters usager, N_i le nombre de déplacements effectués par l'usager i, et H le nombre de gaussiennes.

En tout, $K + H \times K \times 7 \times 3$ paramètres doivent être estimés, 3 correspondant aux trois paramètres de la gaussienne (moyenne, variance et proportion de la gaussienne). Cependant les paramètres de cette vraisemblance, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, ne peuvent pas être estimés directement et un algorithme de maximisation de l'espérance de type CEM (Conditional Expectation Maximization) est proposé dans la section suivante pour résoudre ce problème.

4.2.3 Algorithme CEM et EM

Le but est de maximiser la vraisemblance complétée de Z^1 et Z^2 en utilisant un simple CEM avec une étape E (Espérance) pour reconstruire Z^2. Comme mentionné précédemment, la log-vraisemblance est trop complexe pour l'estimation directe des paramètres et l'utilisation de la vraisemblance complétée permet d'utiliser des algorithmes comme l'algorithme CEM et l'algorithme EM (Expectation Maximization), qui sont les méthodes les plus communément utilisées pour l'estimation des modèles de mélange (voir [McLachlan et Krishnan 2008]).

Le but étant de classer les usagers plutôt que les heures de validation, seule la vraisemblance complétée en Z^1 est utilisée comme critère de maximisation (i.e. Z^2 est exclue du processus de classification). La vraisemblance complétée en Z^1 est décrite par :

$$Lc \quad (\boldsymbol{\Theta}; \mathbf{X}, \mathbf{Z^1}) = \prod_{i=1}^{M} \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{N_i} \sum_{h=1}^{H} \tau_{khd_{ij}} f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) \right)^{z^1_{ik}} ,$$

Ce choix est motivé par la volonté de travailler dans un contexte d'estimation de densité pour Z^2 et dans un contexte de clustering pour Z^1. C'est pourquoi un algorithme CEM est utilisé, en raison de son étape classifiante qui affecte chaque observation à son cluster le plus probable (au lieu de renvoyer un vecteur de probabilités d'appartenance comme c'est le cas pour l'algorithme EM classique).

Pendant l'étape E (E1) de l'algorithme CEM, la densité conditionnelle de (4.1) est calculée. Cette espérance fournit la borne inférieure de la log-vraisemblance qui est maximisée pendant l'étape M (Maximisation) et est donnée par :

$$\begin{aligned} \mathcal{L}c \quad (\boldsymbol{\Theta}; \mathbf{X}, \mathbf{Z^1}) &= \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}^1_{ik} \log(\pi_k) \\ + \quad & \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{j=1}^{N_i} \sum_{h=1}^{H} \hat{z}^1_{ik} t^2_{ijh} \log\left(\tau_{khd_{ij}} f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) \right) , \end{aligned} \tag{4.2}$$

où t^2_{ijh} sont les probabilités a posteriori d'appartenance aux gaussiennes pouvant s'écrire :

$$t^2_{ijh} = \frac{\tau_{khd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}})}{\sum_{h=1}^{H} \tau_{khd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}})} ,$$

et $\hat{z}^1_{ik}$ les indicatrices des clusters les plus probables pour chaque usager obtenues en définissant un seuil pour les quantiles suivants :

$$t^1_{ik} \propto \pi_k \prod_{j=1}^{N_i} \sum_{h=1}^{H} \tau_{khd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) .$$

t^1_{ik} est calculé pendant l'étape E1, alors que $\hat{z}^1_{ik}$ est calculé à partir de t^1_{ik} dans l'étape de

classification C1.

$$\hat{z}_{ik}^1 = \begin{cases} 1 \text{ if } k = \arg\max_k t_{ik}^1 \ , \\ 0 \quad \text{sinon} \ . \end{cases}$$

On maximise alors de manière analytique la borne pendant l'étape M1. Ce qui donne pour les proportions π de chacun des clusters d'usager :

$$\pi_k = \frac{M_k}{M} \ ,$$

avec $M_k = \sum_i \hat{z}_{ik}^1$ le nombre d'usagers affectés au cluster $k^{\text{ième}}$ cluster. À la fin de l'algorithme CEM, une classification des usagers est obtenue. Le second EM commence avec une étape E2 qui calcule les probabilités a posteriori de t_{ijh}^2 conditionnellement à Z^1. L'étape M2 donne les estimations finales des proportions des gaussiennes et des paramètres en maximisant la borne inférieure. Les estimations sont données par :

$$\begin{aligned} \tau_{kwh} &= \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw} t_{ijh}^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw}} \ , \\ \mu_{kwh} &= \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw} t_{ijh}^2 x_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw} t_{ijh}^2} \ , \\ \sigma_{kwh} &= \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw} t_{ijh}^2 (x_{ij} - \boldsymbol{\mu}_k^{(q+1)})^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} z_{ik}^1 d_{ijw} t_{ijh}^2} \ . \end{aligned}$$

Le pseudo code de l'algorithme est présenté sur la figure 1.

4.2.4 Choix des paramètres du modèle

Dans le modèle génératif proposé, deux hyper-paramètres doivent être ajustés à savoir, le nombre de gaussiennes et le nombre de clusters usager. Pour cela, on utilise en général les critères de sélection de modèles. Dans un premier temps, différents critères de sélection de modèles sont présentés. Le choix des différents hyper-paramètres est ensuite justifié.

Critères de sélection de modèles

Plusieurs critères sont utilisés pour la sélection de modèles. Les plus connus sont l'Akaike Information Criterion (AIC) et le Bayesian Information Criterion (BIC). L'AIC est un critère qui se calcule à partir de la vraisemblance L du modèle et de son nombre de paramètres n_p. Il pénalise les modèles qui possèdent un plus grand nombre de paramètres. Le modèle considéré comme étant le meilleur est celui avec l'AIC le plus faible. Il s'écrit sous la forme

$$AIC = 2n_p - 2ln(L).$$

Algorithme 1 Algorithme EMCEM

Entrées: Données X, H nombre désiré de clusters de gaussiennes, K nombre désiré de clusters usager

Sorties: Paramètres estimés $\boldsymbol{\Theta} = (\pi, \tau, \mu, \sigma) q \leftarrow 0$

répéter

Étape E1 :

pour k in $\{1, \ldots, K\}$, i in $\{1, \ldots, N\}$ **faire**

$$t_{ik}^{1\ (q+1)} = \log\left(\pi_k \prod_{j=1}^{N_i} \sum_{h=1}^{H} \tau_{khd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}})\right)$$

Étape C1 : recherche de Z^1 par maximisation a posteriori

$$z_{ik}^{1\ (q+1)} = e_{k^*} \text{ avec } k^* = \arg\max_k t_{ik}^{1\ (q+1)}$$

fin pour

Étape M1 : maximisation : $\boldsymbol{\theta}|Z^1$

pour k in $\{1, \ldots, K\}$ **faire**

$$pi_k^{(q+1)} = \frac{M_k^{(q+1)}}{M} \text{ avec } M_k \text{ le nombre d'usagers dans le } k - ieme \text{ cluster}$$

fin pour

Étape E2 : calcule des probabilités a posteriori de $Z^2|Z^1$

pour i in $\{1, \ldots, M\}$, j in $\{1, \ldots, N_i\}$,k in $\{1, \ldots, K\}$, h in $\{1, \ldots, H\}$ **faire**

$$t_{ijh}^{2\ (q+1)} = \frac{\tau_{kd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}})}{\sum_{h=1}^{H} \tau_{khd_{ij}} \times f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}})}$$

fin pour

Étape M2 : maximisation : $\boldsymbol{\theta}|Z^1, Z^2$

pour k dans $\{1, \ldots, K\}$, h dans $\{1, \ldots, H\}$, l in $\{1, \ldots, L\}$ **faire**

$$\tau_{klh}^{(q+1)} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl} t_{ijh}^2}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl}}$$

$$\mu_{klh}^{(q+1)} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl} t_{ijh}^2 x_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl} t_{ijh}^2}$$

$$\sigma_{klh}^{(q+1)} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl} t_{ijh}^2 (x_{ij} - \boldsymbol{\mu}_k^{(q+1)})^2}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} z_{ik}^1 d_{ijl} t_{ijh}^2}$$

fin pour

jusqu'à $\boldsymbol{z}^{1\ (q+1)} \sim \boldsymbol{z}^{1\ (q)}$ et $\boldsymbol{z}^{2\ (q+1)} \sim \boldsymbol{z}^{2\ (q)}$

Le BIC est un critère qui découle de l'AIC mais qui prend également en compte le nombre d'observations n_{obs}. Il pénalise plus le nombre de paramètres que l'AIC et s'écrit

$$BIC = -2ln(L) + ln(n_{obs})n_p.$$

Ces deux critères utilisent la vraisemblance du modèle, alors que les modèles de mélanges utilisent la vraisemblance complétée. Des travaux ont démontré l'efficacité d'un autre type de critère pour ce type de modèle, l'Integrated Completed Likelihood (ICL) [McLachlan et Krishnan 2008]. Ce critère est l'équivalent du critère BIC, avec pour seule différence que ce dernier se calcule avec la vraisemblance complétée et non pas la vraisemblance

$$ICL = -2ln(L_c) + ln(n_{obs})n_p.$$

C'est ce dernier critère qui a été retenu dans l'analyse qui suit.

Sélection de modèle

Ici, les choix des paramètres a priori du modèle sont justifiés. L'analyse est conduite uniquement sur les données de Rennes. La même méthode a été appliquée pour les données de Gatineau et les résultats peuvent être trouvés en annexe A.3. Le nombre de gaussiennes H nécessaires à la représentation des profils temporels des usagers est discuté et le choix du nombre de clusters usagers K est expliqué. Pour ce modèle, deux initialisations différentes ont été développées pour l'algorithme. Elles sont décrites dans l'annexe A.2.

Pour identifier les meilleurs paramètres, l'algorithme est lancé pour différentes valeurs de H ($H = 2, \ldots, 4$) et pour différentes valeurs de K ($K = 2, \ldots, 20$). L'algorithme pouvant converger vers un minimum local, il est lancé à plusieurs reprises et le meilleur résultat est conservé. Les paramètres des gaussiennes estimés, les proportions des clusters, ainsi que la log-vraisemblance complétée obtenus pour chaque lancé sont enregistrés. L'ICL (Integrated Classification Likelihood) est alors calculé afin de choisir le meilleur modèle.

D'abord, le nombre de gaussiennes H est étudié. Le nombre de gaussiennes est important en raison de son rôle dans l'activité temporelle. Un nombre important de gaussiennes fournit une meilleure représentation de l'activité mais est plus coûteux en temps de calcul, tandis qu'un nombre plus restreint de gaussiennes est plus rapide à estimer mais peut mener à une représentation erronée des motifs d'activité par cluster. Nous pouvons prendre la Figure 4.2 comme exemple. Celle-ci montre un motif d'activité à deux pics en utilisant un mélange de trois gaussiennes. Deux des gaussiennes sont utilisées pour représenter l'activité des pics et la troisième, de variance plus importante, sert à représenter l'activité restante, ce qui peut être utile dans différents cas.

Lorsque les profils temporels des usagers sont représentés en utilisant uniquement deux gaussiennes, les valeurs prises par le critère ICL sont plus élevées que dans les autres modèles (Figure 4.3). De plus, les courbes d'activité qui en résultent montrent que les clusters issus du modèle avec $H = 2$ conduisent à une perte de précision. Cela traduit l'incapacité de ce modèle à capturer une activité à trois pics d'activité (matin, midi, soir), conduisant ainsi les pics du matin et du soir à avoir une variance plus élevée et une moyenne

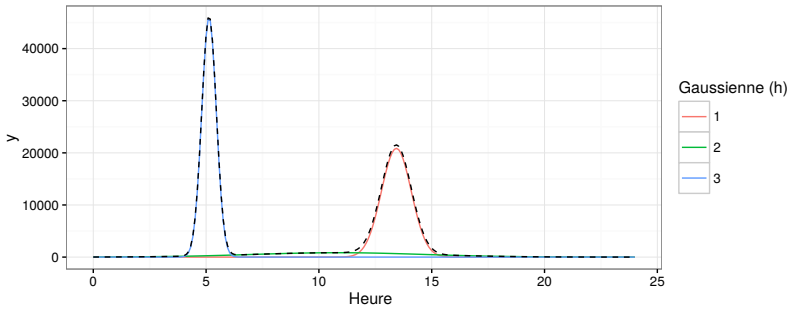


FIGURE *4.2 – Graphique d'un modèle de mélange de trois gaussiennes avec deux pics d'activité. La première gaussienne (en rouge) s'ajuste au deuxième pic. La seconde gaussienne (en vert) est applatie et représente l'activité restante. La troisième gaussienne s'ajuste au premier pic d'activité.*

décalée, dans le but de compenser l'absence du pic du midi.

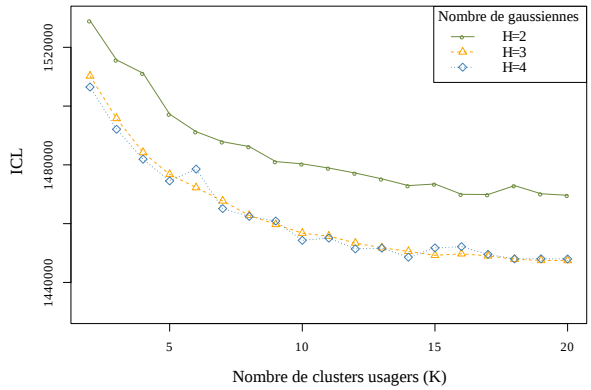


FIGURE *4.3 – Critère ICL pour les clusters usagers K variant entre 2 et 20 et le nombre de gaussiennes $H = 2, 3,$ ou 4*

Les résultats obtenus pour les deux cas restants sont proches puisque, pour toutes les valeurs de K, les modèles avec $H = 4$ ont des valeurs d'ICL similaires à celles des modèles pour $H = 3$. L'étude et la visualisation des résultats pour $H = 4$ (non développées ici) ont montré l'absence de production de quatrième pic, contrairement à ce qui pourrait être attendu intuitivement. De plus, une plus grande variabilité que dans le modèle à trois gaussiennes apparaît d'un jour à l'autre et des pics très fins apparaissent, ce qui peut correspondre à de l'overfitting. Cela suggère que le modèle à trois gaussiennes est suffisant à la représentation des heures de déplacement des usagers dans une journée. De plus il est préférable de travailler avec ce modèle, sachant qu'une gaussienne additionnelle requiert plus de paramètres mais n'apporte pas d'améliorations significatives aux résultats du clustering.

En ce qui concerne le choix du nombre de clusters usager K, le critère ICL montre qu'un nombre plus important de clusters mène à de meilleurs résultats. Plus le nombre de clusters est important, meilleure est la représentation du modèle. Cependant, dans un souci de compromis entre interprétation et précision, le choix d'un nombre de clusters plus faible que celui indiqué par le critère est nécessaire afin de garder des résultats interprétables. Pour l'étude expérimentale présentée dans ce chapitre, le choix a été fait d'analyser le modèle à 10 clusters ($H = 3$, $K = 10$) car il contient une variété de clusters différents et

intéressants et semble correspondre au bon compromis entre précision et interprétabilité. Si l'on souhaite avoir une étude plus détaillée et casser les clusters avec un pourcentage d'usagers élevé, un nombre de clusters plus élevé peut être choisi.

Dans la suite, les résultats de l'application de l'algorithme aux différents cas d'études sont présentés. Dans un premier temps, une analyse est conduite sur les données de Rennes afin de présenter le type d'analyses directement productibles avec les sorties du modèle. Une comparaison des résultats obtenus sur Rennes et Gatineau est alors menée pour révéler les différences d'usages que peut mettre en avant une telle approche. Enfin, un suivi longitudinal des appartenances aux cluster est effectué sur l'ensemble des données de Gatineau.

4.3 Analyse des résultats du modèle sur la ville de Rennes

Le modèle de mélange de gaussiennes a pour première utilité de définir des groupes d'usagers avec des habitudes temporelles similaires des lignes de transports en commun. Dans cette section, il est appliqué aux données de Rennes selon le paramétrage défini précédemment, offrant ainsi une vision des différents groupes et de leurs utilisations des transports en commun. Dans une autre section, il sera également utilisé afin de suivre l'évolution des habitudes des usagers, ce qui nous semble être un cas d'usage intéressant pour les exploitants.

Dans un premier temps, les courbes d'activité du mardi sont comparées dans le but de mettre en avant les principales différences entre leurs activités. Les quatre clusters d'usagers qui semblent les plus représentatifs sont alors présentés. Enfin, un focus est fait sur l'activité spatiale des jeunes abonnés, appartenant à un cluster spécifique, en localisant les stations où ils sont le plus actifs et en les positionnant dans le contexte urbain.

4.3.1 Spécificités du jeux de données

Le but de cette analyse est de révéler les différentes habitudes temporelles de déplacements des usagers de transports en commun. Dans un premier temps un focus sur les usagers réguliers est effectué. L'objectif est de se concentrer sur l'activité qui est susceptible de se répéter et non pas sur celle qui apparaît de manière occasionnelle. Un filtrage des données est donc nécessaire. Les usagers non réguliers seront inclus dans l'analyse du suivi longitudinal des clusters (voir section 4.5) afin de voir si leur utilisation des transports, bien que plus rare, tend à varier ou non.

Afin d'effectuer le filtrage des usagers réguliers, la distribution du nombre de jours actifs (c'est-à-dire les jours où au moins un déplacement a été effectué à l'aide d'une carte à puce donnée) pendant les 30 jours couverts par les données est représentée dans la Figure 4.4. Le graphique montre que ce nombre diminue entre 0 et 10 jours. À dix jours d'utilisation active, un premier point d'inflexion peut être observé suite à une augmentation constante du nombre d'usagers. Un deuxième point d'inflexion se produit à environ 18 jours alors que le nombre d'usagers commence à diminuer constamment. De plus, l'inspection du nombre moyen de segments de déplacement par jour effectués par les usagers sur la figure 4.5

montre que la plupart des usagers ne font qu'un ou deux segments de déplacement par jour.

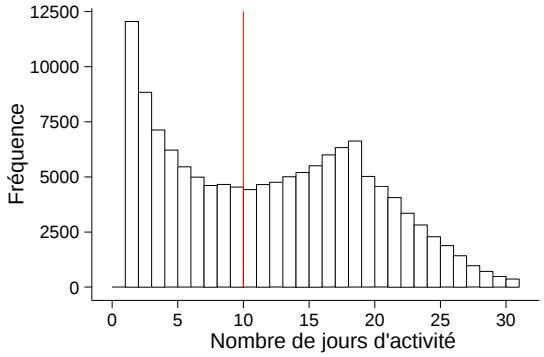


FIGURE *4.4 – Nombre de jours d'activité par usager sur le réseau de transport pendant le mois d'avril 2014.*

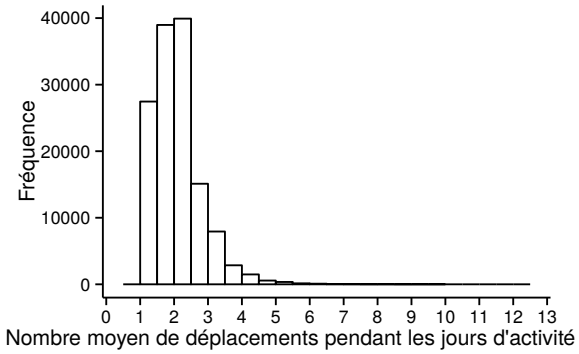


FIGURE *4.5 – Nombre moyen de segments de déplacement par usager sur le réseau de transport pendant le mois d'avril 2014.*

Par la suite, les 10 jours d'activité sont utilisés comme seuil pour distinguer les usagers occasionnels des réguliers. Seuls les usagers réguliers, i.e. ceux réalisant plus de 10 déplacements, sont retenus pour le cas d'étude rennais, l'objectif étant de mettre en évidence les motifs temporels de déplacements fréquents des usagers. Dans la section suivante, l'ensemble de données utilisé contient les déplacements effectués par 10 000 passagers échantillonnés de façon aléatoire parmi ceux ayant plus de 10 jours d'activité. Cet ensemble de données contient 28% de jeunes abonnés, 25% de tickets, 31% de subventions, 12% d'abonnés adultes et les autres types de tarifs représentent moins de 15%, il est représentatif de l'ensemble des usagers. Rennes est une ville étudiante ce qui explique le nombre élevé de jeunes abonnés.

4.3.2 Étude des clusters

Un aspect intéressant de l'analyse des différents clusters est la caractérisation de l'usage de chaque groupe par les motifs d'activité qui lui sont associés pour chaque jour de la semaine. De plus la comparaison de ces motifs entre cluster pour un jour de la semaine donné permet de mieux comprendre les différences entre chaque groupe. Dans ce but,

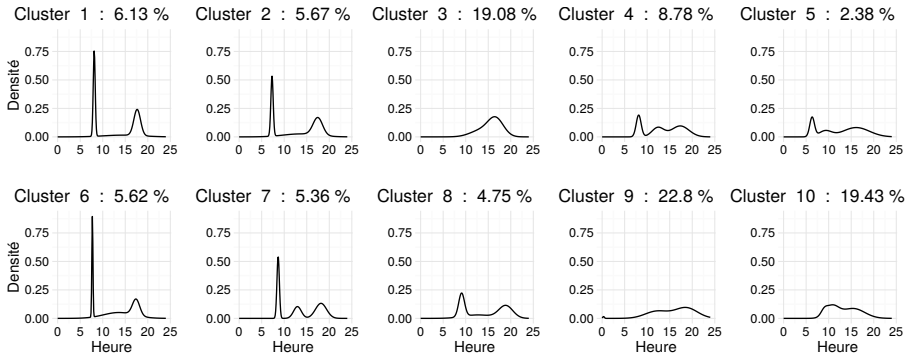


FIGURE *4.6 – Densités conditionnelles de validation pour l'ensemble des dix clusters pour la journée du mardi*

toutes les densités conditionnelles de validation des clusters pour la journée du mardi ont été tracées sur la Figure 4.6. En effet, les différences entre les clusters deviennent ainsi plus apparentes. Par exemple, bien qu'ils présentent la même forme dans leur pic d'activité du matin, le pic du cluster 2 est plus tôt que celui du cluster 7. Certains clusters (tel que le cluster 6) vont présenter une activité très régulière qui devient plus visible en la comparant aux autres.

En se basant sur la Figure 4.6, il est possible de distinguer trois catégories de clusters.

— les clusters avec deux pics d'activité (clusters 1 2, 6 et 8) : La différence entre ces clusters vient du fait que le cluster 8 a une activité plus importante le midi, même si le motif de pic n'apparaît pas. Le pic matinal du cluster 2 est un peu plus tôt que celui du cluster 1 et le cluster 6 a une variance plus importante que celle des autres clusters.

— les clusters à trois pics d'activité (clusters 4, 5 et 7) : Ils peuvent se distinguer de manière plus précise en regardant leurs heures de pic ainsi que leurs formes (par exemple le cluster 7 a des pics plus distincts).

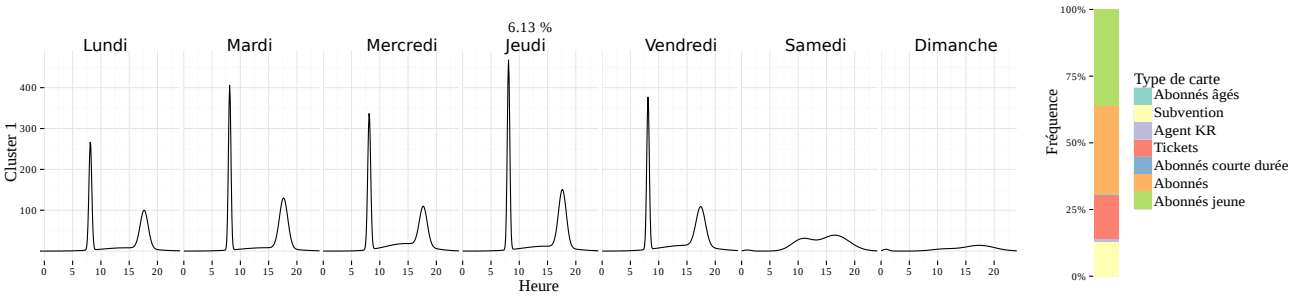
— les clusters avec une activité plus diffuse (clusters 3, 9 et 10) : Le pic matinal du cluster 10, bien qu'étendu sur un grand intervalle de temps, est tout de même visible. Dans le cluster 3, seul un pic couvre l'ensemble de la journée, et le cluster 9 présente un motif inversé par rapport aux autres clusters. Les clusters de cette dernière catégorie regroupent la majorité des usagers (61.31%).

Si nécessaire, les motifs d'activité décrit par la dernière catégorie peuvent être raffinés en augmentant le nombre de clusters (en les séparant en des clusters plus petits et plus homogènes).

Les clusters 1, 4, 7 et 9 sont retenus pour une analyse plus détaillée. Ils offrent à eux quatre un aperçu de toutes les formes d'activité précédemment observées. Afin de prendre en compte la fréquentation liée à chaque type de journée, les densités conditionnelles de validation sont multipliées par le nombre de cartes actives pour chaque type de jour. Le résultat obtenu est appelé profil d'activité temporel.

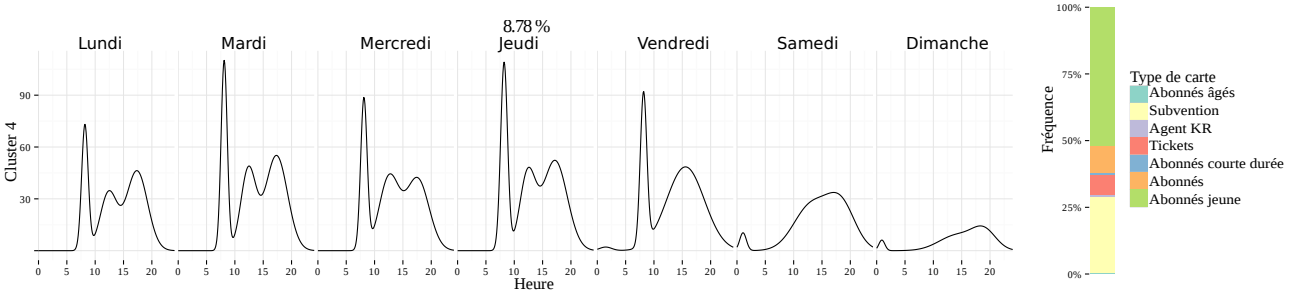
Le cluster 1 (Figure 4.7) présente un motif d'activité à deux pics qui apparaît durant les jours de la semaine. L'étude de ces deux pics montre que la variance du pic matinal est très faible puisque le pic est concentré entre 7h et 9h, tandis que le pic de l'après-midi

est plus étendu et s'étend de 15h à 20h. En dehors de ces pics, il y a très peu, voire pas, d'activité en soirée et le week-end. Cela suggère que les usagers de ce cluster sont habitués à faire la navette entre leur domicile et leur lieu de travail et ne dépendent pas des transports en dehors de leur trajet domicile-travail (c'est à dire pendant leur pause déjeuner ou pour leurs temps libre). Cette observation est appuyée par la proportion de chaque type de cartes dans le cluster (Figure 4.7b), qui montre qu'il contient un fort pourcentage d'abonnés adultes (principalement des adultes actifs) et de jeunes abonnés (principalement des étudiants et des scolaires). Ces deux types de cartes correspondent à des populations qui ont des emplois du temps bien définis que leurs déplacements reflètent.



(a) Profil temporel *(b) Type de carte*

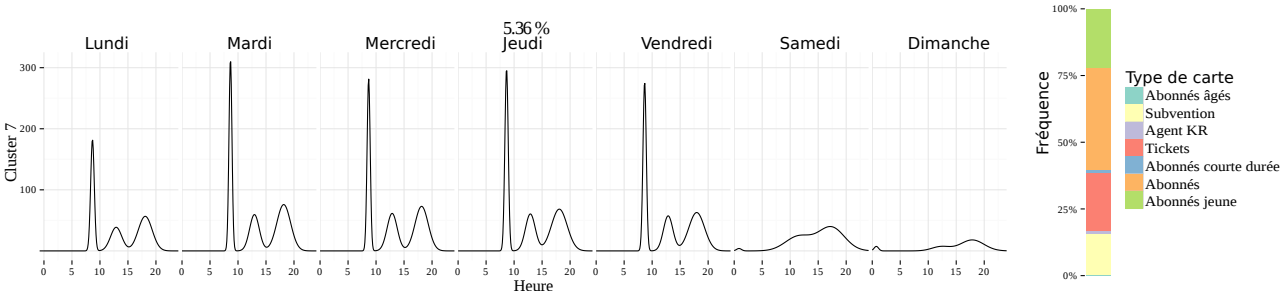
FIGURE *4.7 – Profil d'activité temporel de chaque jour de la semaine et proportion des types de cartes des usagers du cluster 1.*



(a) Profil temporel *(b) Type de carte*

FIGURE *4.8 – Profil d'activité temporel de chaque jour de la semaine et proportion des types de cartes des usagers du cluster 4*

Une tendance similaire est observable pour le cluster 4 (Figure 4.8) et le cluster 7 (Figure 4.9) pour lesquels l'activité du matin est également très régulière avec un pic très resserré. Cependant, contrairement au premier cluster, ils présentent un motif à trois pics d'activité, avec l'apparition du troisième pic entre 10h et 15h. Dans le cas du cluster 4, les pics du midi et du soir sont mélangés et seuls deux pics sont visibles pour la journée du vendredi : le deuxième et le troisième pic fusionnent complètement et forment une courbe en forme de cloche qui s'étend de 10h à minuit. Contrairement au cluster 1, les week-ends et les soirées présentent une certaine activité. Le cluster est composé à plus de 50% de jeunes abonnés et à plus de 25% de subventions, ce qui indique qu'il est majoritairement composé d'étudiants. Dans le cluster 7 les trois pics sont distincts les uns des autres et il y a moins d'activité nocturne en week-end que dans le cluster 4 (sans doute dû à la plus

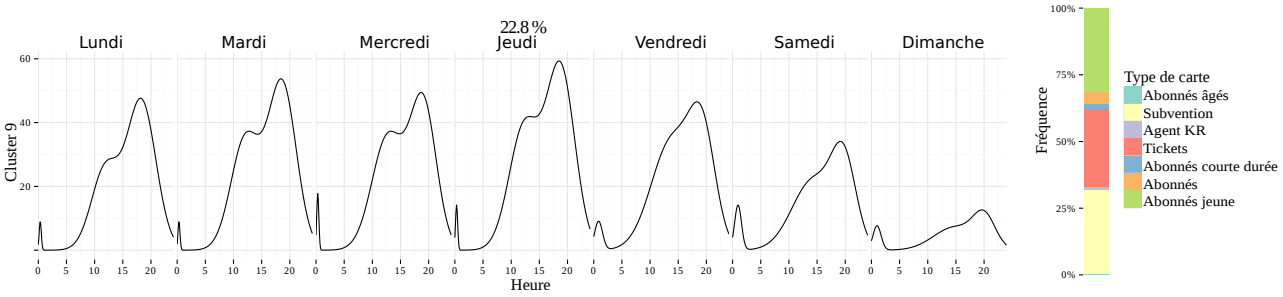


(a) Profil temporel *(b) Type de carte*

FIGURE *4.9 – Profil d'activité temporel de chaque jour de la semaine et proportion des types de cartes des usagers du cluster 7*

faible présence de jeunes abonnés).

Enfin le cluster 9 (Figure 4.10) présente certaines particularités qui sont intéressantes. Par exemple, il a une activité plus diffuse que les autres clusters. Le matin, il n'y a pas de pic précis d'activité, mais un mélange de deux pics qui s'étendent sur toute la journée. Ces pics mélangés ont un motif inversé : le second pic est plus fort que le premier, ce qui implique que l'activité est plus importante dans l'après-midi et en soirée que le matin. Cela peut s'expliquer par la forte activité nocturne présente tous les jours de la semaine : les usagers qui vont utiliser les transports en commun de manière plus tardive vont également avoir tendance à les utiliser plus tard dans la journée. Comme cela peut être observé sur la figure 4.10, les usagers utilisant des tickets et ceux bénéficiant de la gratuité (subvention) comptent pour plus de 50% de la composition du cluster. Ces types de titre de transport sont souvent utilisés par des usagers qui se déplacent de manière plus irrégulière.



(a) Profil temporel *(b) Type de carte*

FIGURE *4.10 – Profil d'activité temporel de chaque jour de la semaine et proportion des types de cartes des usagers du cluster 9*

Comparées aux résultats obtenus à partir d'une approche discrétisée du temps, les courbes d'activité issues de cette approche temporelle continue offrent une meilleure vue de l'activité des passagers : au lieu d'avoir des probabilités discrètes d'activité par heure, une probabilité d'activité continue en temps qui ne souffre pas d'un potentiel biais résultant d'une agrégation par intervalle de temps est obtenue. De plus, la moyenne et la variance de chaque pic d'activité sont connues, ce qui n'était pas possible avec l'approche agrégée. Enfin, la représentation graphique des profils temporels d'activité est aisée à comprendre et à interpréter.

4.3.3 Localisation spatiale des clusters étudiants

Dans la volonté de rattacher ces profils à des zones géographiques spécifiques, il est montré comment les activités d'un cluster donné peuvent être positionnées dans le contexte spatial de la ville. Dans ce but, le cas de l'activité étudiante est étudié. Les étudiants ont été choisis en raison de leurs générateurs de déplacements (càd les institutions académiques et les installations associées) qui sont plus regroupés que les générateurs des autres types d'usagers (par exemple les larges zones commerciales et industrielles pour les actifs adultes).

La distribution des cartes dans les différents clusters (Figure 4.11) révèle que la majorité des jeunes abonnés sont présents dans les clusters 4, 6 et 9. Comme cela a été mentionné auparavant, le cluster 9 correspond à une activité plus diffuse. Dans les clusters 4 et 6, les pics d'activité matinaux sont visibles et pour les clusters 4 et 9 une importante activité nocturne peut être observée le week-end ainsi que le jeudi soir, ce qui correspond à la nuit la plus active pour les étudiants français.

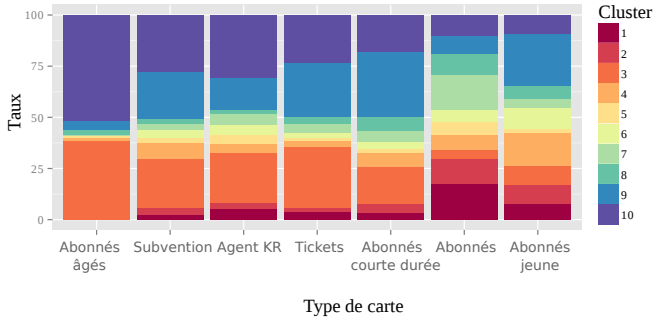


FIGURE *4.11 – Distribution des types de cartes pour les différents clusters*

En conséquence, le focus est fait sur les jeunes abonnés appartenant au cluster 4 (avec l'hypothèse, basée sur les observations précédentes, qu'ils sont pour la plupart étudiants). L'étude de l'impact de la localisation de leurs générateurs de déplacements sur leur activité dans le réseau de transport est alors étudiée. Dans ce but, une carte du réseau de transport rennais est présentée sur la Figure 4.12. Sur cette carte sont représentées les données socio-économiques des différents quartiers de la ville, la localisation des bâtiments académiques et les stations les plus actives pour les jeunes abonnés du cluster 4.

Les stations les plus actives sont localisées dans le nord-est, le sud et dans le centre-ville de Rennes. L'activité située dans le sud de Rennes est principalement observée le long de la ligne de métro. Comme la Figure 4.12 le montre, cette partie de la ville correspond à des logements collectifs avec des faibles revenus, ce qui suggère que ces stations sont probablement les stations résidentielles des jeunes abonnés. L'activité la plus forte se situe dans le centre-ville, ce qui n'est pas une surprise compte tenu du fait que le centre regroupe la plupart des activités de la ville. De plus, le réseau STAR est en forme d'étoile dont le centre est le centre-ville, ce qui crée un nombre important de transferts dans cette zone. Enfin, le nord-est de Rennes est la zone qui regroupe le plus de bâtiments académiques, ce qui explique la forte fréquentation des stations localisées à ce niveau (telle que la station de métro Villejean-Université et ses stations avoisinantes).

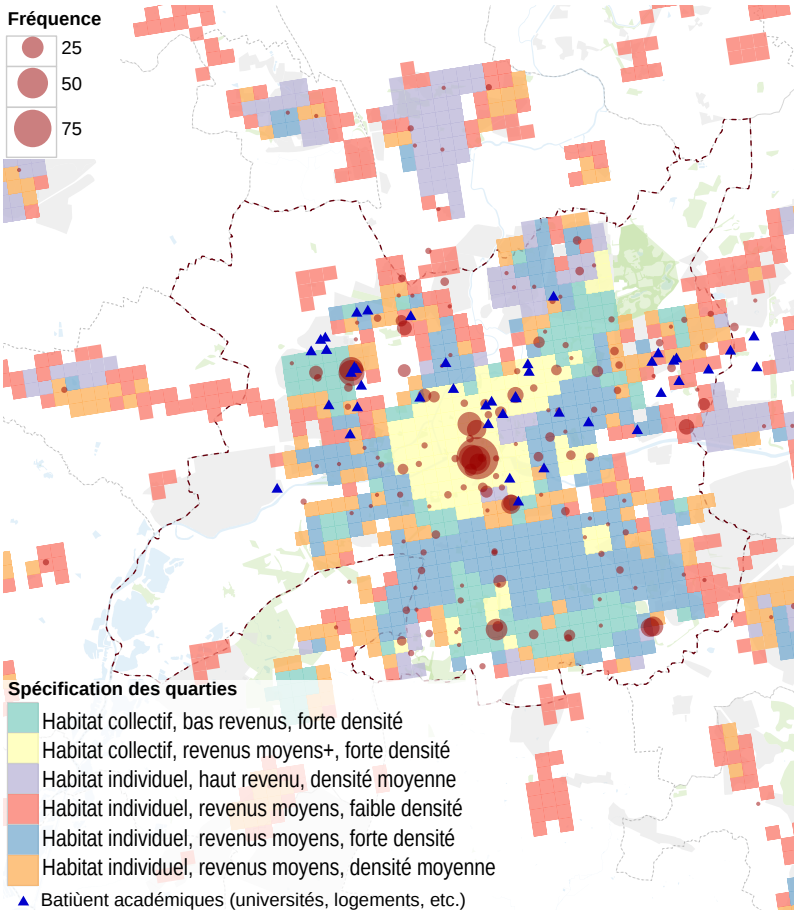
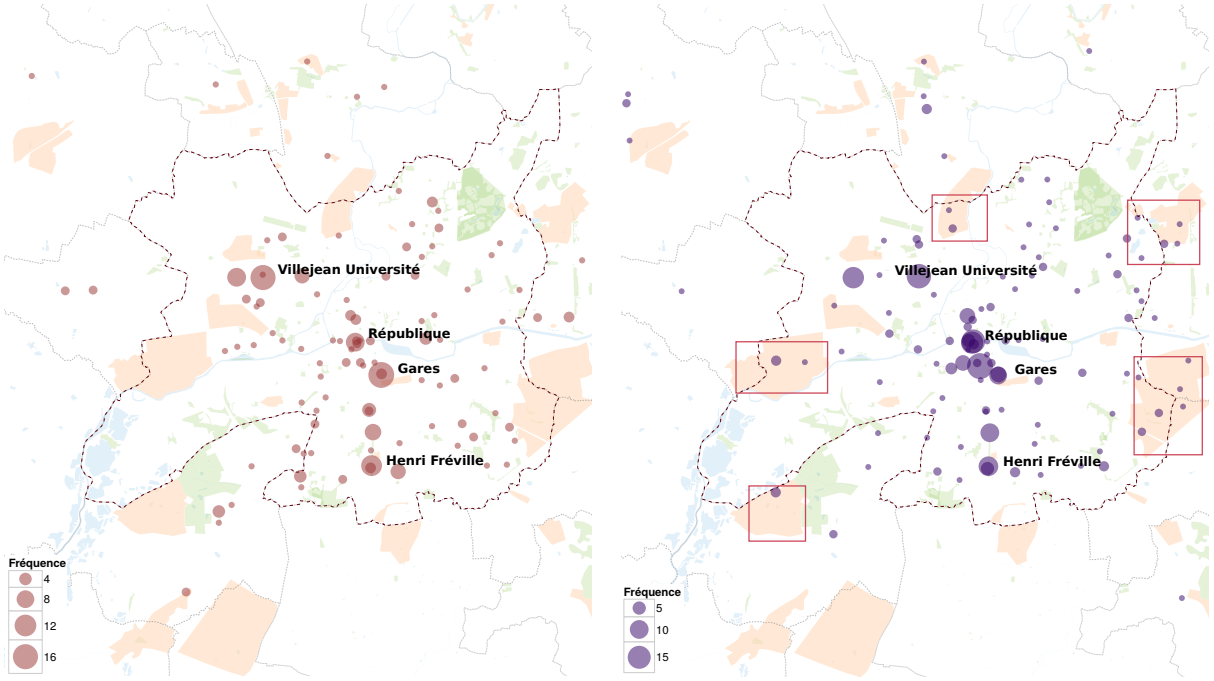


FIGURE *4.12 – Carte de l'activité des stations de Rennes générée par les jeunes abonnés du cluster 4*

4.3.4 Interprétation spatiale des clusters (Cas du cluster 1)

Dans la section précédente, la localisation spatiale de certaines populations a été étudiée en analysant leur activité par station, indépendamment de l'heure à laquelle cette activité a lieu. L'heure du déplacement joue pourtant un rôle important dans l'identification du type de déplacement effectué (domicile-travail, loisir, etc.). Afin de repérer les zones générant de l'activité, il est donc intéressant d'être capable d'identifier les zones les plus actives pour chaque pic d'activité. En travaillant sur un mélange de gaussiennes, cette approche capture automatiquement les instants les plus pertinents (et leur variabilité) pour chaque cluster. De plus, ce processus est conduit séparément pour chaque jour. En conséquence, il est possible d'affecter de façon probabiliste chaque déplacement d'un usager appartenant à un cluster donné, à la gaussienne la plus adapté du jour auquel le déplacement a été fait. Chaque groupe peut alors être étudié séparément dans le but de découvrir comment l'usage évolue en fonction du temps. On considère tous les usagers appartenant à un cluster, le cluster 1 et on projette les stations. Pour illustrer cette idée, l'activité du cluster 1 est étudiée. Comme cela a été vu précédemment, l'activité des usagers dans ce cluster est principalement divisée en deux pics d'activité. Le nombre de déplacements par station pour la première gaussienne de la journée (qui correspond à l'activité matinale) est présenté dans la Figure 4.13a, tandis que la troisième gaussienne de la



(a) Déplacements affectés à la première gaussienne du jour, avec une moyenne à 8 :07 et un écart-type de 32 minutes.

(b) Déplacements affectés à la troisième gaussienne du jour, avec une moyenne à 17 :20 et un écart-type de 80 minutes.

FIGURE *4.13 – Nombre de déplacements par station effectués par les usagers du cluster 1 pour la journée du 1er Avril 2014.*

journée (qui correspond à l'activité de fin d'après-midi) est présentée sur la Figure 4.13b. Ces deux gaussiennes concentrent la plupart de l'activité de la journée et peuvent révéler les différences d'activités entre le matin et le soir. Il est alors possible de localiser les zones que les usagers vont principalement quitter le matin (lieu de résidence) et celles dont ils vont repartir le soir (lieu de travail).

La comparaison des deux cartes de la Figure 4.13 révèle, comme attendu, certaines différences dans l'activité. La première observation qui peut être faite est que l'activité présente dans le centre-ville est plus dispersée en soirée. En effet, le matin, le centre de l'activité est principalement situé autour de deux zones (République et Villejean-Université), tandis que l'activité en soirée est plus étendue. La seconde observation qui peut être faite concerne l'apparition d'activité dans les zones commerciales et industrielles en soirée (zone beige sur la carte). Cela suggère qu'un certain nombre d'usagers travaillent dans ces zones et valident leurs cartes à la fin de la journée, quand ils retournent à leur domicile. Enfin, certaines stations sont actives à la fin de la journée alors qu'elles ne l'étaient pas le matin et inversement.

4.4 Approche comparative des résultats du modèle sur Rennes et Gatineau

Les résultats obtenus à l'aide du modèle de mélange ont pu être présentés, il est alors intéressant de transposer cette approche sur d'autres données. Tout d'abord pour montrer

la réplicabilité du modèle, c'est à dire le fait qu'il soit applicable à différentes villes. Ensuite pour montrer les différences culturelles qui peuvent être révélées au travers de l'utilisation des transports en commun par les usagers. C'est dans cette optique que le modèle est une nouvelle fois appliqué aux données filtrées de Rennes (avril 2014) et Gatineau (Février 2015). Cette fois encore, seuls 10 000 usagers de chaque ville sont tirés aléatoirement pour l'analyse.

4.4.1 Comparaison des clusters pour la journée du mardi

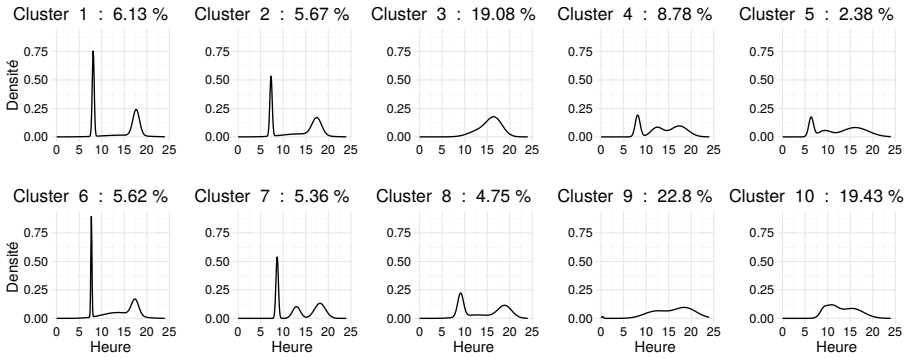
Les clusters obtenus pour les deux villes sont présentés ici. Afin de pouvoir comparer facilement les résultats obtenus sur celles-ci, le nombre de clusters a été fixé à 10 pour les deux jeux de données. [1]

La figure 4.14 présente pour chacune des villes, Rennes et Gatineau respectivement, les 10 densités conditionnelles de validation des clusters pour la journée du mardi. Comme présenté précédemment, trois types d'activités sont observables à Rennes (figure 4.14a) : des motifs d'activité à deux pics (clusters 1, 2, 6 et 8), des motifs d'activité à trois pics (clusters 4, 5 et 7) et des motifs plus diffus (clusters 9 et 10). Les clusters à deux pics correspondent à une utilisation domicile-travail des transports. Les clusters à trois pics correspondent souvent aux passagers qui utilisent les transports publics pendant leur pause déjeuner. Enfin les clusters avec une activité plus diffuse regroupent une plus grande variété d'usage des transports en commun.

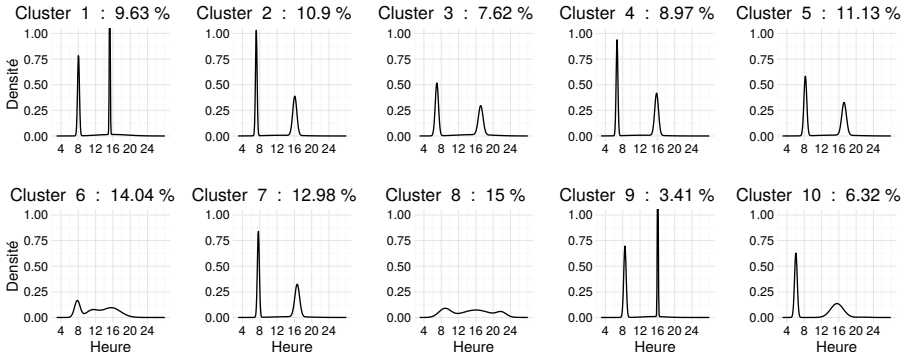
À Gatineau (figure 4.14b), trois types d'activités sont observables : des motifs d'activité à deux pics classiques (clusters 2, 3, 4, 5, 9 et 10), les motifs d'activité à deux pics plus réguliers en fin de journée (clusters 1 et 7) et des motifs d'activité à trois pics (clusters 6 et 8). Encore une fois l'activité à deux pics, correspond à des déplacements domicile-travail. Les clusters avec une activité plus concentrée en fin de journée correspondent à des trajets domicile-école et la régularité dans l'heure de prise des transports en commun est due à l'utilisation de bus scolaires. Enfin les clusters avec une activité plus diffuse regroupent l'ensemble des usagers dont l'utilisation des transports n'est pas uniquement liée au domicile-travail.

Les principales différences dans les motifs d'activité du mardi entre Rennes et Gatineau sont : la prédominance de clusters d'activité à deux pics à Gatineau, la présence de cluster avec un usage diffus des transports à Rennes et l'absence d'activité nocturne à Gatineau. Cela s'explique par la disposition spatiale des deux villes. Tout d'abord, Rennes est une ville de taille moyenne très active, y compris en soirée. C'est aussi une ville polycentrique, avec d'autres centres qui sont différents villages de petite taille autour de Rennes. Le réseau STAR sert toute l'agglomération et ces petites villes créent une activité à Rennes. Au contraire, Gatineau est situé dans la banlieue d'Ottawa. Ottawa est la capitale politique du Canada, ce qui signifie qu'elle génère une grande partie de l'activité de secteur. La majeure partie de l'activité de transport de Gatineau est donc dédiée au transfert entre Gatineau et Ottawa, car la plupart des passagers du transport habitent à Gatineau, travaillent à Ottawa

1. Les résultats de Rennes sont ceux déjà présentés dans la section 4.3. Ils sont repris ici pour faciliter la comparaison avec Gatineau.



(a) Rennes-France



(b) Gatineau-Canada

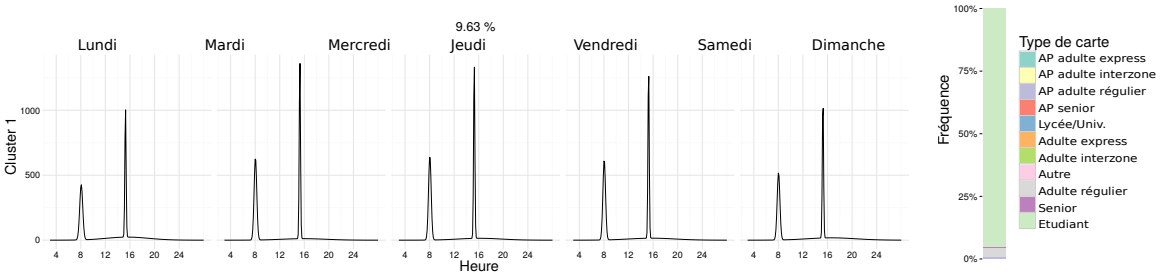
FIGURE *4.14 – Densités conditionnelles de validation des usagers pour la journée du mardi obtenus pour les 10 clusters de Rennes - France et Gatineau - Canada.*

et utilisent les transports principalement pour ce trajet. Cette différence spatiale permet de mieux comprendre le nombre élevé de clusters à deux pics d'activités de Gatineau.

4.4.2 Analyse des clusters étudiants

Une analyse plus fine des différences d'usage entre les deux villes, au travers de la comparaison de leurs populations étudiantes, est détaillée ici. Elle porte plus précisément sur le cluster 1 dans le cas de Gatineau (retenu parce qu'il contient la plus grande proportion de cartes de type "étudiant") et sur le cluster 4 dans le cas de Rennes (retenu en raison de la forte activité autour des établissements universitaires, cf. Section 4.3).

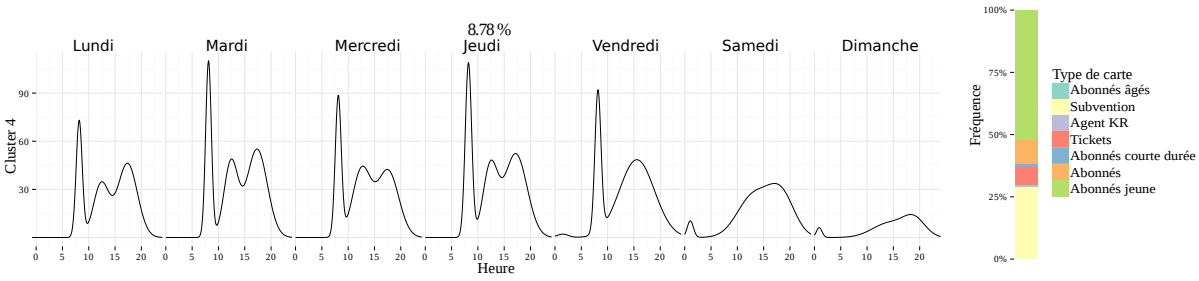
Le cluster 1 (voir la figure 4.15) présente un motif à deux pics qui se produit pendant les jours de semaine, ce qui suggère que les usagers de ce groupe n'utilisent pas le réseau de transport pendant leur déjeuner. L'étude de ces deux pics montre que la variance du pic du matin est très faible car le pic se concentre vers 8h00 et le pic de l'après-midi est plus concentré vers 16h00. En dehors de ces pics, il n'y a pas d'activité pendant la nuit et pendant le week-end. Cette faible activité en dehors des heures de pointe indique que les passagers de ce groupe ne dépendent pas des transports en commun en dehors de leurs déplacements habituels de l'école et à la maison. En ce qui concerne les types de cartes, la composition du cluster montre qu'il contient un pourcentage élevé d'étudiants qui n'ont probablement pas besoin d'utiliser les transports publics, sauf pour aller à l'école et parfois le samedi.



(a) Profil temporel *(b) Type de carte*

FIGURE *4.15 – Profil d'activité temporel pour chaque jour de la semaine et type de carte des usagers du cluster 1 de Gatineau.*

Dans le cluster 4 (Figure 4.16), l'activité du matin est très régulière avec un pic resserré. Cependant, contrairement au premier cluster, il présente un schéma à trois pics avec l'apparition d'un troisième pic entre 10h00 et 15h00. Les pics du midi et du soir se mêlent et, pour la journée du vendredi, seulement deux pics sont visibles : le deuxième et le troisième pic sont complètement fusionnés et forment une courbe en forme de cloche qui s'étend de 10h00 à minuit. Contrairement au cluster 1, le week-end et la nuit présentent une activité notable. Ce groupe regroupe plus de 50% des jeunes abonnés et plus de 25% de subvention, ce qui souligne qu'il est majoritairement composé d'étudiants.



(a) Profil temporel *(b) Type de carte*

FIGURE *4.16 – Profil d'activité temporel pour chaque jour de la semaine et type de carte des usagers du cluster 4 de Rennes.*

Les différences observées entre les deux villes sont dues à la forte utilisation des bus scolaires à Gatineau (expliquant également la très faible variance du pic de fin d'après-midi). La présence à Rennes de nombreux étudiants universitaires peut également expliquer ces différences. En effet ceux-ci sont plus susceptibles d'utiliser les transports en soirée et le week-end.

4.5 Suivi temporel des résultats du clustering (Gatineau)

Un suivi de l'évolution de l'activité des usagers au cours des années peut être effectué en suivant les changements dans leur affectation aux clusters au cours des années. Un tel suivi peut permettre de voir quel type d'usager va avoir une utilisation des transports stable dans le temps ou au contraire variable. La méthodologie permettant le suivi des cartes est introduite. Une première analyse des clusters obtenus sur ces données est ensuite

présentée afin d'avoir une meilleure compréhension des changements de cluster au cours des ans. Enfin, un suivi est effectué sur les cartes.

4.5.1 Méthodologie

Le modèle de mélange de gaussiennes est utilisé afin de permettre un suivi de l'évolution des habitudes des usagers au cours des ans. Pour cela on utilise les données du mois de février enregistrées à Gatineau sur cinq ans et non filtrées.

Pour obtenir les mêmes paramètres de cluster au cours des cinq années, ces jeux de données ont été regroupés pour former un seul et unique ensemble de données. Il est important de permettre à une carte de changer de cluster d'une année à l'autre. L'identifiant de carte d'une carte donnée est changé d'une année à l'autre, ce qui permet une différenciation entre les cartes en fonction de l'année au moment du clustering.

Une fois les clusters obtenus, les identifiants de carte originaux sont récupérés afin de permettre le suivi d'une même carte sur les 5 ans.

4.5.2 Analyse des résultats du clustering

Pour un aperçu général des clusters, les densités conditionnelles de validation de tous les clusters pour la journée du samedi sont présentés sur la figure 4.17. Les journées de semaines ne sont pas détaillées car elles sont très similaires aux résultats obtenus dans la section 4.4, seules les proportions de cartes dans chaque cluster différant légèrement.

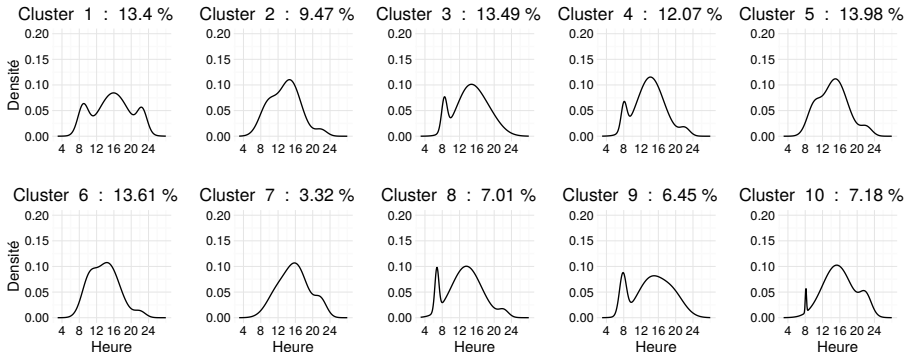


FIGURE *4.17 – Densités conditionnelles de validation des dix clusters pour la journée samedi.*

Ces clusters peuvent être plus précisément caractérisés par l'examen de leur activité du week-end. Outre le fait que tous les clusters présentent une activité inférieure à celle de la semaine et qu'elle se concentre en majeure partie de façon diffuse l'après-midi, on constate trois types d'activité. Des clusters ont encore un pic d'activité matinal le week-end (Clusters 3, 4, 8, 9 et 10), le reste de l'activité s'étendant sur toute la journée avec un pic d'activité variant entre 13 et 16h. D'autres clusters se caractérisent par une activité diffuse sur toute la journée, sans pic particulier (clusters 2, 5, 6 et 7). Enfin le cluster 1 possède une activité à trois pics à 7h, 16h et 22h. Avec les clusters 1 et 7, ce sont les seuls clusters à présenter une activité nocturne marquée.

L'examen de la répartition des cartes dans les différents clusters révèle quels clusters sont les plus liés à chaque type de carte (Figure 4.18). Le cluster 1, qui avait le profil le

plus diffus, regroupe la majorité des cartes Senior et Univ. Les clusters 2 et 8 regroupent une forte proportion de cartes Interzone. Les cartes d'étudiant sont les plus présentes dans le cluster 10. Enfin, les cartes Express Adulte sont en nombre élevé dans les clusters 2, 4 et 5, alors que les adultes réguliers semblent être répartis dans tous les clusters.

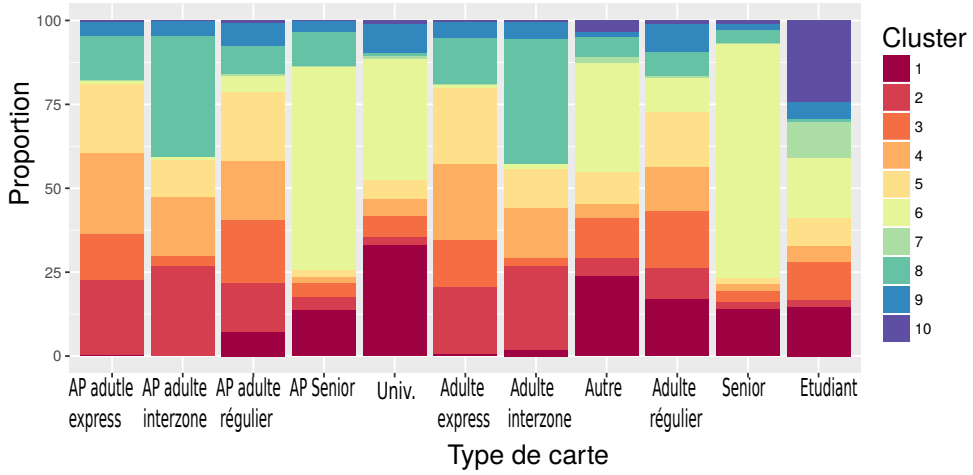
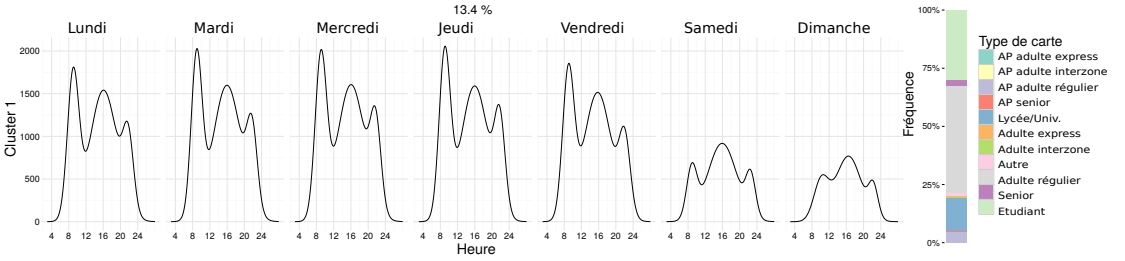


FIGURE *4.18 – Distribution des types de cartes dans les clusters usager*

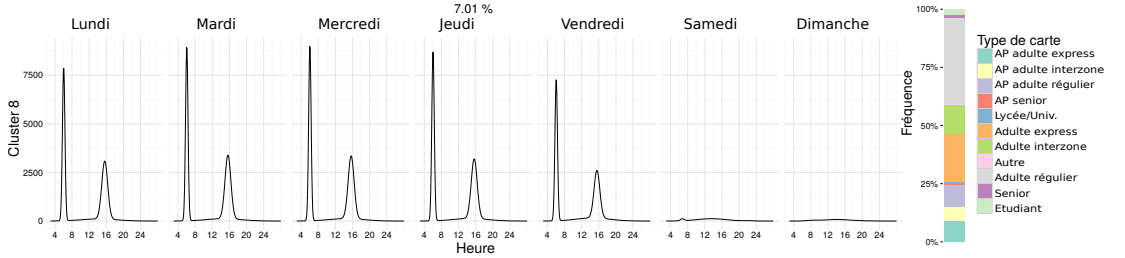
Pour une analyse plus détaillée de l'activité des clusters, celle-ci va se concentrer sur trois clusters particuliers. Comme cela a été vu précédemment, différents motifs apparaissent à partir des résultats du clustering. Le Cluster 1 est analysé plus en détail, en tant que cluster au profil d'activité diffus, le Cluster 8 comme profil classique à deux pics et le Cluster 10 comme motif à deux pics avec une activité régulière l'après-midi. Malgré son motif diffus, les pics du matin et de l'après-midi du cluster 1 (Figure 4.19b) sont visibles et le pic du matin a une variance inférieure à celle de l'après-midi. Un troisième pic apparaît pendant la soirée, ce qui peut être interprété comme un voyage secondaire. Une caractéristique intéressante de ce cluster est que l'activité de fin de semaine est plus élevée que dans les autres clusters. Ce groupe se compose principalement d'adultes réguliers (48,47%), d'étudiants (29,71%) et d'Univ. (13,74%). Le cluster 10 (Figure 4.21b) présente un motif très régulier composé de deux pics autour de 8h et 15h30. Le pic de l'après-midi est plus concentré que le pic du matin, ce qui peut être expliqué, comme cela a été discuté précédemment, par la forte proportion d'étudiants (93,5%) qui ont probablement des horaires scolaires très réguliers (lycée). Contrairement au cluster 1, il ne présente pas beaucoup d'activité le week-end. Le cluster 8 (Figure 4.20), comme le cluster 10, présente un schéma régulier à deux pics. Cependant, la variance des pics est supérieure à celle du cluster 10, en particulier l'après-midi. Le pic du matin est centré à 6 heures du matin et l'après-midi à 15 h 40. Ce groupe est plus mélangé que les deux autres. Bien que le cluster ait 39,76 % d'adultes réguliers, il a également 20,09% d'adultes express, 12,03 % d'adultes Interzone et presque pas d'étudiants. Les voyages effectués de façon plus matinale s'expliquent par la distance entre le domicile des usagers et leur lieu de travail.

À présent que la connaissance des clusters est plus précise, il est possible d'étendre l'analyse et de suivre les changements de clusters d'une année sur l'autre.



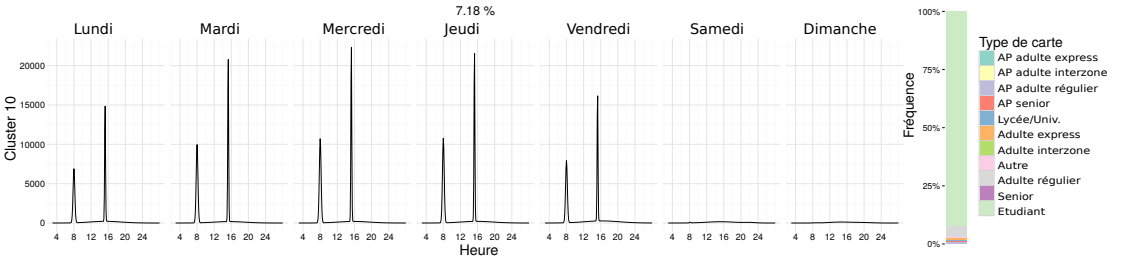
(a) Profil temporel *(b) Type de carte*

FIGURE *4.19 – Profil d'activité temporel pour chaque jour de la semaine et types de cartes des usagers du cluster 1*



(a) Profil temporel *(b) Type de carte*

FIGURE *4.20 – Profil d'activité temporel pour chaque jour de la semaine et types de cartes des usagers du cluster 8*



(a) Profil temporel *(b) Type de carte*

FIGURE *4.21 – Profil d'activité temporel pour chaque jour de la semaine et types de cartes des usagers du cluster 10*

4.5.3 Suivi des cartes

Cette section s'intéresse plus particulièrement au suivi des cartes sur les 5 ans de données et à leurs passages d'un cluster à l'autre.

Focus sur un cluster (Cluster 8)

Le focus est fait sur un cluster afin d'obtenir plus d'informations sur les changements et de voir si les cartes retournent parfois à leur cluster d'origine. Pour cette partie de l'étude, le cluster 8 a été choisi en raison de son motif typique d'activité à deux pics. La figure 4.22 montre l'évolution de l'appartenance des cartes, initialement attribuées au cluster 8 en 2005, aux clusters entre 2005 et 2009. Chaque couleur correspond à un cluster différent et l'année peut être lue de gauche à droite (de 2005 à 2009 respectivement). La

plupart des cartes qui ont quitté le cluster 8 ont été affectées au cluster 2. Les clusters 2 et 8 ont des motifs plus proches, ce qui explique pourquoi le cluster 2 a reçu la majorité des cartes perdues par le cluster 8 (6,6% en 2006). Cependant, il peut être observé que de nombreuses cartes sont retournées à leur cluster d'origine après avoir changé. La proportion de cartes quittant le cluster 8 pour des clusters d'activité plus diffuses (comme le cluster 1 ou le cluster 6) est très faible (environ 1%). Ces changements peuvent s'expliquer par des modifications d'activité ou de localisation des validations. Une analyse plus approfondie sur une base individuelle est nécessaire pour mieux interpréter ces modifications. Une analyse des cartes qui ne sont plus actives (cluster "out"), montre que le taux de rotation des cartes du cluster 8 est en moyenne de 60% sur les 5 ans.

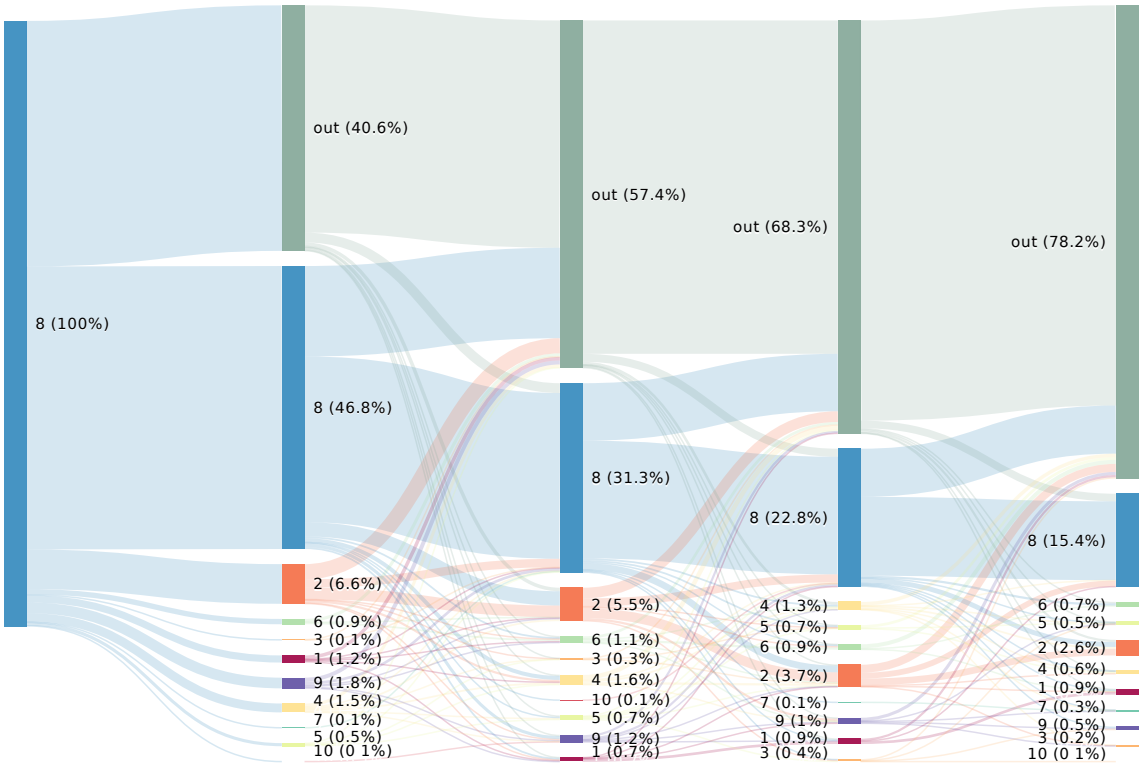


FIGURE *4.22 – Appartenance aux clusters des cartes affectées au cluster 8 (en bleu) en 2005.*

Changement de clusters au cours des années

Le flux global de cartes entre 2005 et 2009 à travers les différents clusters est illustré dans la Figure 4.23. Pour chaque année et cluster, la proportion de cartes appartenant à chaque cluster, ainsi que les clusters auxquelles les cartes sont attribuées, est présentée. Les cartes entrantes («nin»), c'est-à-dire les cartes qui n'étaient pas encore actives, ainsi que les cartes sortantes («out»), c'est-à-dire les cartes qui ne seront plus actives, figurent également sur la figure. En regardant leur flux, nous pouvons constater qu'un grand nombre de cartes n'étaient pas actives pendant plus d'un ou deux ans. Certains clusters semblent avoir perdu toutes leurs cartes rapidement d'une année à l'autre.

Afin de quantifier les cartes se déplaçant d'un cluster à l'autre, le tableau 4.1 représente les probabilités de transition moyennes estimées d'une année à l'autre entre les clusters.

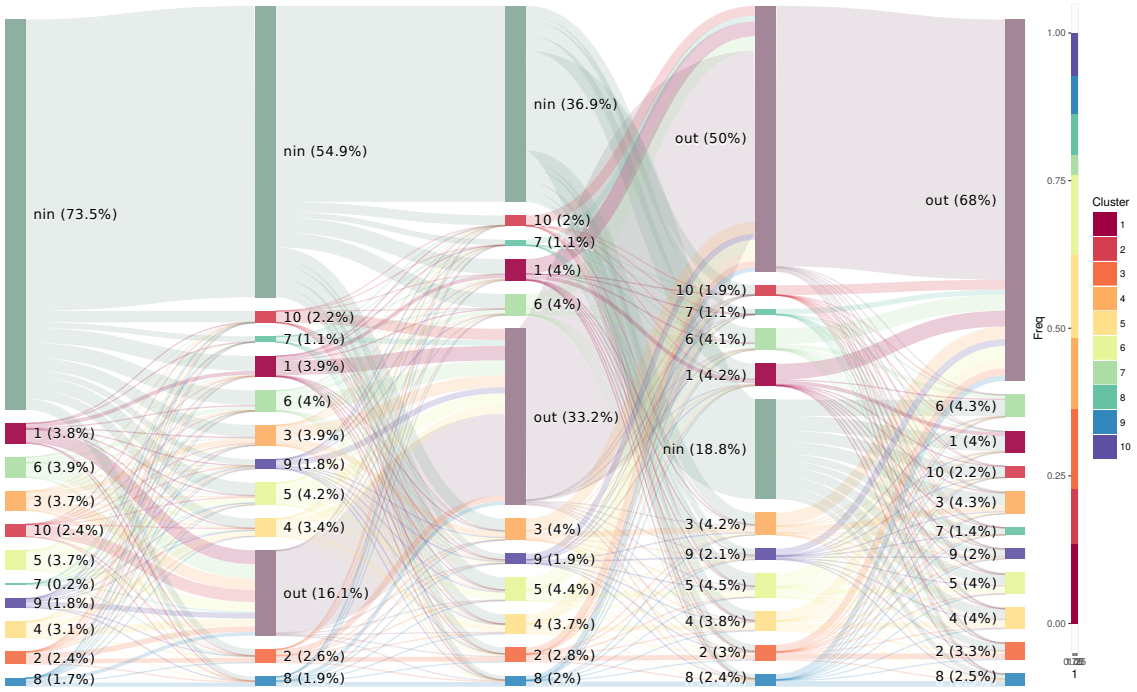


FIGURE *4.23 – Proportion des cartes par cluster et affectations aux clusters des cartes actives entre 2005 et 2009.*

Ces probabilités ont été calculées pour toutes les cartes et ont été moyennées sur les 5 ans. Chaque ligne du tableau nous permet d'observer plus clairement à la fois la probabilité qu'une carte change son cluster d'une année à l'autre et les clusters candidats où la modification du cluster pourrait se produire. Lorsqu'aucune modification de cluster n'a lieu, les éléments diagonaux sont égaux à 100%.

	1	2	3	4	5	6	7	8	9	10
1	**62.01**	1.24	9.06	1.41	2.68	**14.28**	0.21	2.04	6.29	0.78
2	0.86	**62.63**	1.03	**13.42**	3.18	1.32	0.33	12.92	3.97	0.33
3	5.83	1.31	**66.28**	2.97	**15.69**	3.48	0.07	0.33	3.33	0.71
4	0.95	**14.72**	3.00	**56.47**	**15.54**	1.69	0.61	2.12	4.51	0.38
5	1.59	2.96	**13.28**	**15.82**	**57.75**	2.58	0.26	0.76	4.49	0.51
6	**13.29**	2.17	5.15	3.49	3.68	**65.46**	0.75	1.56	3.05	1.39
7	5.83	5.27	3.17	**20.51**	**11.46**	**16.35**	**24.82**	3.05	6.11	3.43
8	1.73	**11.56**	0.42	2.02	1.24	1.61	0.25	**77.65**	3.25	0.28
9	9.83	8.62	7.48	**11.42**	**10.55**	5.77	0.76	5.91	**39.02**	0.65
10	**11.49**	2.69	**12.30**	5.47	**14.92**	**19.59**	2.18	4.95	3.43	**22.98**

TABLE *4.1 – Tableau des moyennes des probabilités de transition entre les clusters d'une année sur l'autre.*

Chaque cluster échange principalement des cartes avec un, deux ou trois clusters. Les autres changements de carte sont résiduels. Par exemple, le cluster 8 donne généralement des cartes au cluster 2 et reçoit des cartes de ce cluster. Un manque de stabilité des clusters 7, 9 et 10 peut être observé (les éléments diagonaux sont respectivement égaux à 25 %, 39 % et 23 %). Cela s'explique par leur grande proportion de cartes d'étudiants, dont

l'identifiant, comme indiqué précédemment, ne reste pas le même d'une année à l'autre et n'est pas pris en compte lors du calcul des probabilités de transition.

L'analyse ci-dessus nous amène à incorporer la relation entre les clusters (c'est-à-dire, dans quelle mesure ils sont éloignés les uns des autres) de l'année en cours et ceux vers lesquels ils vont changer. Une divergence entre tous les clusters est donc calculée en utilisant la divergence de Kullback-Leibler (KL). La divergence entre les deux clusters est ainsi définie comme la somme pour chaque jour de la semaine de la divergence de KL des deux mélanges gaussiens. Plus formellement,

$$\begin{aligned} div_{Clust}(cluster_i, cluster_j) &= \sum_{l=1}^{7} (div_{KL}(g(.;i,l), g(.;j,l)) \\ &\quad + div_{KL}(g(.;j,l), g(.;i,l))) \end{aligned} \tag{4.3}$$

où $g(.;i,l)$ est la fonction de densité de la distribution du mélange de gaussiennes de i pour la journée l et $div_{KL}()$ est la divergence de Kullback-Leibler.

En utilisant un clustering ascendant hiérarchique (HAC) sur la divergence, il est alors possible d'obtenir une vue hiérarchique des différents clusters à l'aide du dendrogramme associé (Figure 4.24). Ainsi, le custering semble être ramifié en trois groupes différents avec des motifs similaires. Le premier groupe, le groupe A, qui comprend les clusters 1, 6 et 9, est le groupe d'activité diffuse. Le deuxième groupe, le groupe B, qui comprend les clusters 3, 8 et 2, 4, 5, est un groupe de motifs à deux pics. Enfin, le dernier groupe, le groupe C, se compose des clusters 7 et 10, qui représentent eux-mêmes principalement des étudiants et présentent des motifs très réguliers dans l'après-midi.

En retournant au suivi des clusters, les changements de clusters d'une année sur l'autre deviennent alors plus compréhensibles. En effet, il peut être confirmé que les clusters échangent principalement leurs cartes avec des clusters dont l'activité est similaire. Par exemple, les clusters 2 et 4, qui sont les plus proches dans le dendrogramme, échangent un grand nombre de cartes, ce qui est probablement dû au fait que le clustering est effectué en prenant en compte chaque jour de la semaine, et que seuls quelques changements dans l'activité de la carte peuvent la faire basculer dans un cluster d'activité proche. En outre, 53,5 % des cartes de groupe du premier groupe (groupe diffus) et 80,2 % des cartes du second groupe (motif à deux pics) restent dans le même groupe pendant les cinq années d'étude. Le troisième groupe (principalement composé de cartes d'étudiants) étant presque vide, il n'est pas logique de regarder sa stabilité au cours des années.

Pour obtenir un meilleur aperçu des groupes qui restent stables et de ceux qui ne le restent pas, la figure 4.25 montre le même organigramme que précédemment en utilisant les groupes A, B et C. On peut rappeler que le groupe A se caractérise par un motif d'activité diffuse, le groupe B présente un modèle d'activité à deux pics alors que le groupe C est principalement composé d'élèves ayant des profils d'activité très réguliers dans l'après-midi. Cette figure confirme qu'un grand nombre de cartes appartenant au groupe B restent dans ce groupe, tandis que les cartes appartenant au groupe A changent de cluster plus facilement. Les cartes appartenant au groupe C qui appartiennent principalement

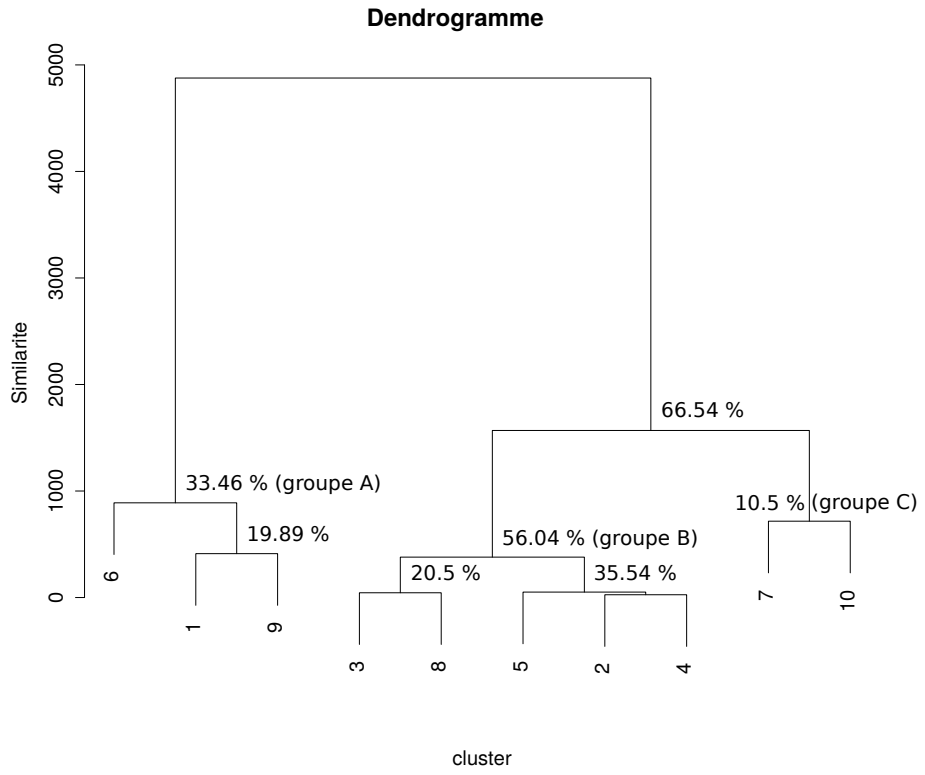


FIGURE *4.24 – Dendrogramme des divergences entre clusters obtenus par clustering hiérarchique.*

à des étudiants sont volatiles puisqu'elles doivent être renouvelées chaque année. Dans l'ensemble, nous pouvons conclure que le regroupement présente une stabilité significative d'une année à l'autre pour les cartes qui restent actives.

4.5.4 Analyse spatiale à l'aide de l'entropie

Jusqu'à présent, l'analyse effectuée n'a été que temporelle. Cependant, comme indiqué plus haut, les données d'origine incluent également les lieux d'embarquement. Dans cette section, une caractérisation spatiale des résultats du clustering est présentée en utilisant un jeu de données réduit aux cartes actives sur les 5 années.

Pour effectuer cette analyse, l'entropie spatiale de chaque carte est étudiée. Cette entropie est basée sur la probabilité que chaque carte soit validée dans l'une de ses stations les plus fréquemment utilisées (c'est-à-dire les stations les plus actives pour la carte). L'activité spatiale de chaque carte est décrite en utilisant une distribution multinomiale pour ses différentes stations. L'entropie utilisée ici est l'entropie de Shannon [SHANNON 1948], qui est définie comme suit :

$$H(X) = -\mathbb{E}[\log P(X = x_i)] = -\sum_{i=1}^{n} P_i \log P_i.$$

Un diagramme violon est utilisé pour étudier l'entropie de chaque cluster sur la figure 4.26. Trois différents types de diagramme violon peuvent être observés. En outre, ces trois

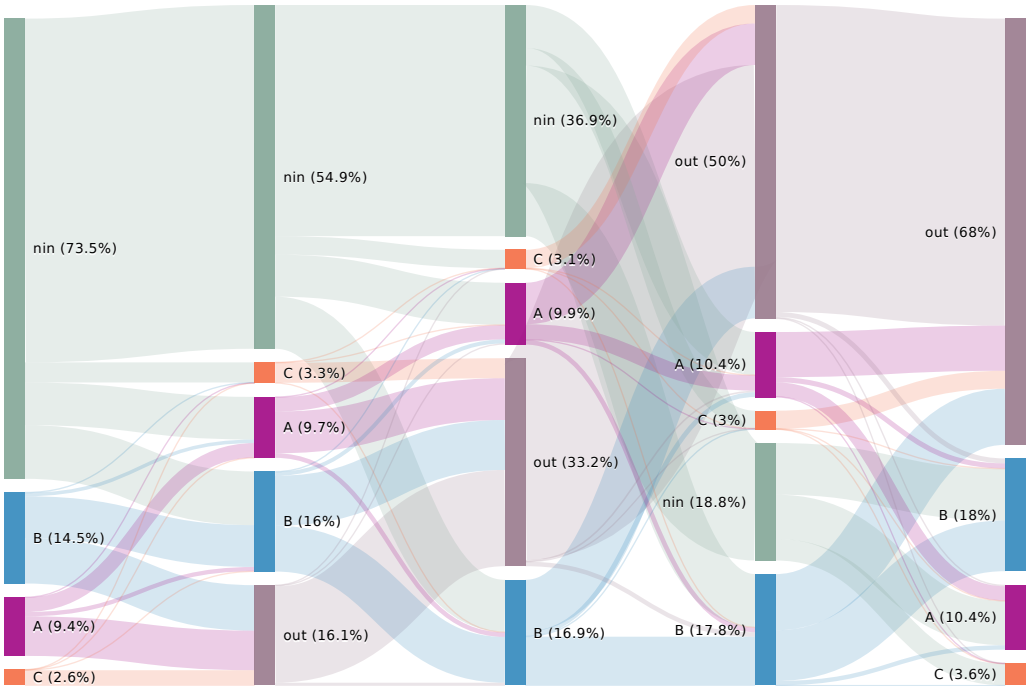


FIGURE *4.25 – Pourcentage des cartes appartenant à chaque groupe pour l'ensemble du jeu de données et affectation aux différents groupes des cartes actives entre 2005 et 2009.*

types correspondent aux trois types de clusters précédemment observés. Les deux clusters d'élèves (Clusters 7 et 10) ne peuvent être examinés plus en détail en raison du très petit nombre de cartes à puce qu'ils contiennent (entre 19 et 32 cartes selon l'année, soit 0,8 % et 1,3 % du nombre total de cartes). En effet, comme cela a été vu, ils n'ont pas assez de cartes pour présenter des résultats significatifs. Les clusters 1, 6 et 9, qui présentent une activité temporelle plus diffuse, semblent également présenter une activité spatiale plus diffuse. L'entropie moyenne des clusters 1, 6 et 9 est plus élevée que l'entropie des autres clusters.

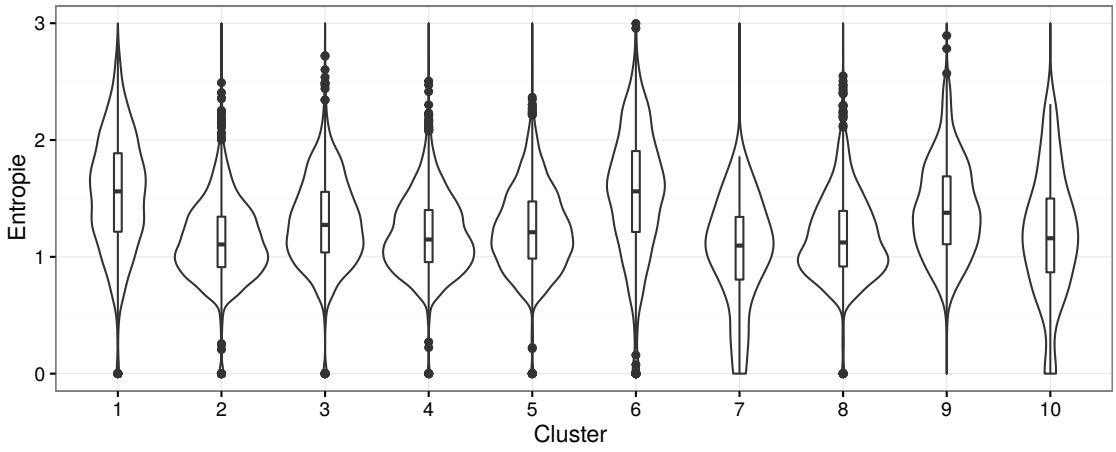


FIGURE *4.26 – Mesure de l'entropie des lieux de validation pour chaque carte dans les différents clusters entre 2005 et 2009.*

Conclusion de chapitre

Ce chapitre présente une analyse détaillée de l'activité générée par les usagers de transports en commun au travers des données carte à puce enregistrées sur les réseaux de la ville de Rennes, en France, et de la ville de Gatineau, au Canada. À l'aide d'un modèle génératif de mélange de gaussiennes, un regroupement des usagers a été effectué en fonction de leurs activités temporelles. Une comparaison des résultats obtenus pour chacun des deux cas d'étude a été présentée. Un suivi de l'appartenance des cartes aux différents clusters sur plusieurs années a également été effectué.

Une analyse détaillée des clusters utilisant leurs caractéristiques statistiques (moyenne des pics d'activité et variance) montre qu'en plus des modèles temporels classiques correspondant aux déplacements pendulaires domicile/travail, certains groupes effectuent leurs voyages plus tôt de manière plus régulière ou plus diffuse que d'autres. La comparaison des résultats sur les deux cas d'étude a mis en avant des différences fortes d'utilisation des transports, avec un usage pendulaire beaucoup plus marqué pour la ville de Gatineau, tandis que la ville de Rennes présente une variété d'activité beaucoup plus grande le midi, en week-end et en soirée.

L'analyse de l'évolution des affectations de clusters de la même carte sur une période de plusieurs années révèle que même si certains changements dans le cluster émergent au cours des années, les clusters semblent maintenir la même proportion de cartes et que la majorité des cartes qui changent de cluster migrent vers des clusters ayant une activité similaire à leur cluster d'origine.

D'autres travaux peuvent être menés pour prolonger ce travail. Tout d'abord, le nombre de clusters a été choisi en fonction d'un critère de sélection, ce qui nécessite de faire un compromis entre l'interprétation et l'ajustement du modèle à l'ensemble des données. Il serait intéressant d'étudier le choix du nombre de clusters plus en détail et ce, afin de mettre en évidence les usagers ayant des comportements atypiques. En outre le modèle proposé prend en compte les jours de la semaine, ce qui peut augmenter combinatoirement le nombre de clusters. À Gatineau, la principale différence dans l'activité de transport est liée à la différence jour de semaine / week-end. En revanche à Rennes, l'analyse a révélé certaines différences d'activité selon le jour de la semaine, en particulier dans le cas des mercredis et vendredis, en plus de la différence inhérente entre l'activité du week-end et de la semaine. Il serait intéressant dans un travail futur de comparer deux modèles, l'un prenant en compte les jours de la semaine et le second prenant en compte uniquement le type de jour (semaine ou week-end). Ce type d'étude pourrait aider à mettre en évidence le lien étroit entre le type de modélisation statistique et les habitudes temporelles dans l'utilisation des transports publics dans une ville. En outre, certaines données supplémentaires devraient être considérées sur les cartes. Les données utilisées dans notre analyse sont incomplètes (les cartes perdues et volées ne contiennent pas le même ID lorsqu'elles sont remplacées). Et enfin, une analyse approfondie est nécessaire pour mieux comprendre les motivations sous-jacentes aux changements de clusters des cartes. À cette fin, un modèle dédié devrait être développé.

Conclusion générale

Les travaux détaillés dans ce manuscrit ont pour objectif de développer des méthodes de fouille de données permettant d'extraire une information plus riche et plus fiable à partir des données billettiques qui n'ont pas été initialement conçues pour de telles analyses. Pour cela, divers outils ont été mis en place tout au long de la thèse.

Le premier chapitre expose le contexte et la motivation des travaux. Un état de l'art sur les travaux menés sur les données billettiques, mais également sur les autres sources de données (GSM, GPS, enquêtes) est présenté, permettant ainsi de mettre en relief les avantages et inconvénients intrinsèques de chaque source de données et de rappeler ceux des données de référence que sont les enquêtes.

Suit une analyse descriptive des différents jeux de données utilisés tout au long de nos travaux. Une telle analyse exploratoire est souvent nécessaire pour la prise en main d'un jeu de données. Elle permet en effet de voir leurs limites (nécessité de corrections ou de pré-traitements), mais également d'identifier les pistes de travail pour une analyse encore plus poussée avec des outils plus avancés (dans notre cas un suivi de l'activité à l'échelle de l'usager par exemple).

Une première contribution de cette thèse est la méthodologie présentée pour la détection d'anomalies. Celle-ci est en effet d'une grande importance car elle permet d'améliorer la fiabilité des données. Très peu de travaux ont porté sur le sujet dans le cadre de l'analyse des données billettiques. La méthodologie que nous proposons tient compte non pas des jours de semaine seulement mais d'autres facteurs influant sur l'activité tels que les vacances, le mois, etc. Une approche en deux étapes a alors été développée afin, dans un premier temps, de détecter les facteurs les plus influents définissant ainsi des contextes homogènes d'utilisation des transports et dans un deuxième temps, de proposer une détection d'anomalies sur chaque contexte ainsi défini. Une des forces de cette approche en deux étapes est qu'elle peut être adaptée aux besoins de l'utilisateur, les approches permettant de définir les contextes et de détecter les anomalies peuvent être modifiées afin de mieux correspondre aux données. Enfin cette méthodologie de détection d'anomalies peut offrir un bon outil à des experts afin de faciliter la labellisation des données en anomalies ou non. Une fois une telle labellisation validée par un expert, il sera alors possible de mener un apprentissage supervisé sur les données.

Une autre contribution est l'analyse et le suivi de l'activité des usagers. Bien que cela soit un sujet largement traité dans la littérature, aucune des approches proposées ne prenait en compte la continuité temporelle des données billettiques. L'approche par mélange de gaussiennes que nous avons proposée dans le chapitre 4 permet une telle extension. La prise en compte de l'aspect continu des données permet de supprimer le biais créé par une

segmentation du temps. Elle permet également une caractérisation plus précise des pics d'activité par groupe d'usagers via leur moyenne et leur variance.

D'autre part le suivi longitudinal de l'appartenance des usagers aux différents clusters présenté dans le chapitre 4 n'a encore jamais été mis en œuvre précédemment à l'échelle de plusieurs années (Langlois et al. 2016 avaient proposé une approche similaire mais à l'échelle de la semaine). Une telle profondeur temporelle permet un suivi des variations de l'activité des usagers sur le long terme.

Une des limitations de ces travaux concerne la détection d'anomalie. Comme l'ont montré les résultats, certaines stations ont une activité qui ne dépend pas uniquement des facteurs pris en compte (jour, vacances scolaires, etc.). Il serait donc nécessaire d'enrichir ces variables, notamment à l'aide du calendrier universitaire, pour avoir une vision plus juste des types de jour. De plus, l'utilisation d'un unique calendrier pour l'ensemble des stations n'est pas toujours justifiée, certaines stations ayant une activité qui leur est spécifique. Il serait intéressant d'adapter un calendrier par type de station.

Les perspectives de ces travaux sont nombreuses. Pour ce qui est de la détection d'anomalie, comme cela a été mentionné, l'une des richesses de l'approche proposée est sa flexibilité et la possibilité de faire varier les approches de classification et de détection utilisées. Grâce à cette flexibilité, il peut être intéressant de développer une nouvelle approche de détection des événements atypiques qui soit plus robuste. Un des points clés des études futures concernera la validation des résultats obtenus. Les travaux présentés ici n'ont bénéficié que d'une validation "manuelle" de certains événements. Il serait intéressant d'exploiter en plus des données billettiques des sources de données telles que les logs des anomalies enregistrées par l'opérateur, pour confirmer ces résultats. Cependant l'avis d'un expert serait également nécessaire car, comme cela a été précédemment énoncé, certaines anomalies sont liées à des problèmes de remontées de données qui ne sont pas nécessairement signalés par l'opérateur.

Le suivi de l'activité des usagers génère également des perspectives à ce travail. En effet de nombreux points méritent d'être étendus. Tout d'abord le choix du nombre de clusters pose toujours une vraie question par son difficile compromis entre précision et interprétation. D'autres critères que celui utilisé dans cette thèse pourraient être envisagés afin d'avoir une pénalisation supérieure à la pénalisation actuelle sur le nombre de paramètres du modèle. En effet un nombre important de clusters générant une augmentation du temps de calcul et une augmentation de la difficulté d'interprétation, c'est un caractère important à prendre en compte.

D'autre part les résultats plus détaillés (avec un nombre plus important de clusters) n'ont pas été étudiés. Une telle analyse, bien que plus complexe, permettrait peut-être de révéler certains usages plus intéressants même s'ils restent minoritaires.

Pour ce qui est des travaux sur le suivi temporel et l'évolution de l'appartenance aux clusters des usagers, il pourrait être intéressant d'aller plus loin dans la modélisation. Une étude poussée des causes pouvant générer ces changements d'activité pourrait être d'un

grand intérêt et permettre de mieux les anticiper et les prédire. Une modélisation des transitions entre clusters à des visées de prédiction à l'aide de Modèles de Markov Cachés est également un volet qui mériterait de plus amples travaux.

Annexe A

Compléments méthodologiques

Différents compléments méthodologiques liés aux méthodes proposés sont détaillés ici.

A.1 Détection d'événements atypiques par station et type de jour

Dans cette section les approches utilisées pour la détection d'événements atypiques sont présentées.

A.1.1 Approche par boxplot au quart d'heure

La première approche est une détection d'événements atypiques par utilisation du boxplot. Le boxplot est une boîte dont la limite inférieure est égale au premier quartile, la limite supérieure au troisième quartile et le trait central à la médiane des données étudiées. La hauteur de cette boîte représente l'espace interquartile définit par :

$$E = \mathcal{Q}_3(\{x_1, \ldots, x_n\}) - \mathcal{Q}_1(\{x_1, \ldots, x_n\}).$$

L'écart interquartile est une mesure de la dispersion des données. Afin de savoir si un point de la courbe est considéré comme atypique par rapport au reste des données, on utilise l'écart interquartile pour définir des bornes de détection. L'intervalle ainsi obtenu s'écrit :

$$I_Q = [\mathcal{Q}_1(\{x_1, \ldots, x_n\}) - 1.5 \times E; \mathcal{Q}_3(\{x_1, \ldots, x_n\}) + 1.5 \times E]$$

Tout point hors de cet intervalle est alors un événement atypique, car trop éloigné du reste des points.

L'application du boxplot à nos données de validations se fait sur un total de 15×10 sous-ensembles. Un sous-ensemble comprend l'ensemble des données pour une station s, $s \in \{1, \ldots, 15\}$ l'ensemble des stations de la ligne A, et un type de jour j, $j \in \{1, \ldots, 9\}$ l'ensemble des types de jours définis dans la section précédente. Pour chaque sous-ensemble les quartiles et l'espace interquartile ont été calculés pour l'ensemble des quarts d'heure de la journée. Tout point extérieur à cet intervalle en un quart d'heure donné est alors labellisé comme événement atypique et le quart d'heure du jour correspondant comme un quart d'heure avec un nombre de validations atypiques.

A.1.2 Approche par boxplot fonctionnel

Contrairement à l'approche précédente, le boxplot fonctionnel porte sur l'ensemble des validations d'une journée. Nous présentons ici l'approche définie dans Sun et Genton 2011. Une observation est définie comme une courbe décrite par le vecteur du nombre de validations par quart d'heure y_i s'écrivant sous la forme :

$$y_i = \{y_{i,1}, \ldots, y_{i,d}\}, \quad i \in \{1, \ldots, D\},$$

avec i un jour donné et D le nombre de jours de la période d'étude. Une bande est la zone délimitée par un ensemble de courbes. Il est possible de définir la profondeur de bande d'une courbe. Plus une courbe est proche du centre de la bande, plus sa profondeur de bande est élevée. Inversement, plus une courbe est éloignée du centre plus sa profondeur est faible. La courbe la plus profonde, et par conséquent la plus centrale, est la courbe médiane.

Dans le cadre de notre approche nous avons utilisé la profondeur de bande modifiée(MBD) définie dans López-Pintado et Romo 2009. Cette modification a pour avantage de prendre en compte le temps passé par une courbe dans la bande. En raison du caractère longitudinal des courbes à notre disposition, cet indicateur est mieux approprié.

Pour mesurer la proportion de temps qu'une courbe va passer dans la bande, la profondeur de bande modifiée est définie par :

$$BDM_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_j \leq n} \lambda_r\{A(y; y_{i_1}, \ldots, y_{i_j})\}, \tag{A.1}$$

avec $A_j(y) \equiv A(y; y_{i_1}, \ldots, y_{i_j}) \equiv \{t \in \mathcal{I} : min_{r=1,\ldots,k} y_{i_r}(t) \leq y(t) \leq max_{r=1,\ldots,k} y_{i_r}(t)\}$ et $\lambda_r(y) = \lambda(A_j(y))/\lambda(\mathcal{I})$ avec λ la mesure de Lebesgue dans $\mathcal{I}$.

Une fois la profondeur de bande calculée pour chaque courbe, il ne reste plus qu'à appliquer une approche boxplot classique sur nos données. En effet les courbes étant ordonnées selon leur profondeur de bande, il est possible de définir la courbe médiane, ainsi que les deux courbes délimitant la région centrale contenant 50% des données. On va alors définir comme étant outlier potentiel, toute courbe qui sortira de l'intervalle définie par la région centrale augmentée de 2 fois sa hauteur.

A.2 Méthode d'initialisation

Deux approches d'initialisation du clustering sont possibles. La première, nommée *seeds*, est une initialisation par k-means. Des petits groupes sont d'abord tirés aléatoirement. Un k-means est alors appliqué sur ces groupes afin d'estimer les paramètres du modèle. Dans la deuxième méthode, les clusters sont initialisés aléatoirement et les paramètres sont estimés à l'aide de ces clusters. Le but est ici de déterminer laquelle des deux méthodes est la plus efficace.

Pour les deux initialisations, l'algorithme est lancé 50 fois. La log-vraisemblance com-

plétée et le nombre d'itérations avant convergence sont enregistrés. Le choix a été fait de comparer ces deux initialisations avec un modèle de mélange de 3 gaussiennes et 15 clusters.

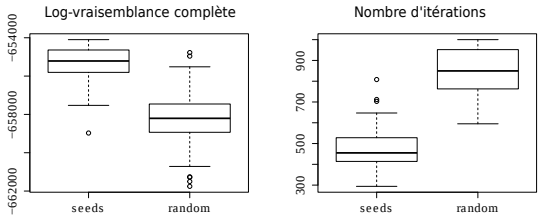


FIGURE *A.1 – Comparaison de la vraisemblance complétée et du nombre d'itérations avant convergence de l'algorithme pour les deux initialisations seeds et random*

Sur la figure A.1, les boxplots de la vraisemblance complétée et du nombre d'itérations avant convergence obtenus sont montrés. L'initialisation *Seeds* converge plus vite que l'initialisation *random* avec une moyenne égale à 260 itérations contre 493. De plus, la vraisemblance complétée est en moyenne égale à -97440 pour la méthode *seeds* et à -98300 pour l'autre méthode. Dans cette thèse le méthode *seeds* est la seule utilisée en raison de sa vitesse de convergence.

A.3 Choix des paramètres pour Gatineau

Comme cela a été fait pour Rennes, le nombre de clusters et de gaussiennes nécessaires au modèle de mélange de gaussiennes, doit être estimé pour le cas d'étude de Gatineau.

Afin d'identifier les meilleurs paramètres, l'algorithme a été lancé pour différentes valeurs de H ($H = 2, \ldots, 5$) et pour différentes valeurs de K ($K = 2, \ldots, 15$). L'algorithme pouvant converger vers un minimum local, il est lancé plusieurs fois et le meilleur résultat est conservé. Le critère ICL est utilisé.

Le nombre de gaussiennes H est d'abord étudié. La figure A.2a montre un saut dans les valeurs de l'ICL entre H égal à 2 et H plus grand ou égal à 3. Cependant il n'y a pas de différences significatives entre H égal à 3, 4 ou 5, alors que 4 et 5 gaussiennes nécessitent plus de temps de calcul pour l'estimation des paramètres. Sachant que la majorité des déplacements sont des navettes domicile/travail (avec validation le matin et le soir, soit une activité à deux pics) et qu'aucune activité à 4 ou 5 pics n'apparaît quand $H = 4\,;\, 5$, nous avons décidé de garder $H = 3$ gaussiennes.

La courbe du nombre de cluster montre que plus le nombre de clusters est élevé, meilleur est le modèle. Le but de l'étude étant de révéler les principales habitudes temporelles des usagers et l'analyse des résultats des clusters montrant des résultats intéressants, nous choisissons de fixer le nombre de clusters K à 10. Une étude plus détaillée peut être menée en augmentant le nombre de clusters.

Le boxplot présent sur la figure A.2b montre qu'en plus de la qualité des clusters révélée par le critère ICL, le résultat a une certaine stabilité.

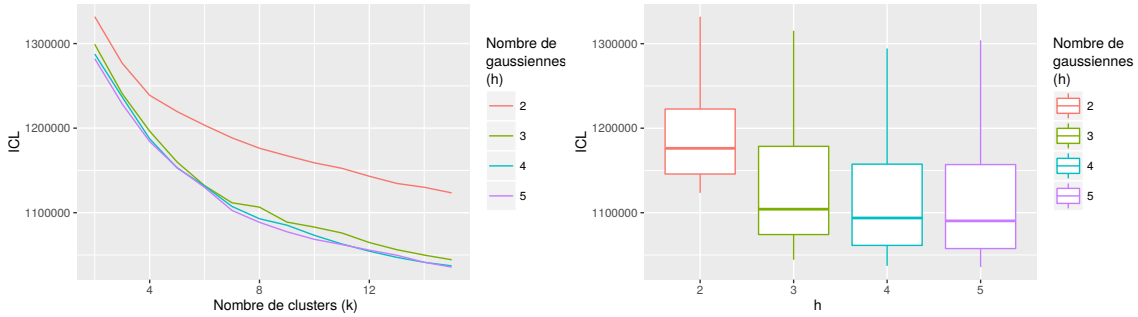


FIGURE *A.2 – Critère ICL et boxplot du critère ICL pour 10 lancé, avec différents nombre de gaussiennes,* $H = 2, \ldots, 5$*, et différents nombres de clusters usager,* $K = 2, \ldots, 15$*, dans le modèle de mélange. .*

Bibliographie

Agard, B., C. Morency et M. Trépanier (2006), « Mining public transport user behaviour from smart card data », *in* : *The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, p. 17–19.

Agrawal, Shikha et Jitendra Agrawal (2015), « Survey on Anomaly Detection using Data Mining Techniques », *in* : *Procedia Computer Science* 60, p. 708 –713, issn : 1877-0509.

Ahillen, Michael, Derlie Mateo-Babiano et Jonathan Corcoran (2015), « The Dynamics of Bike-Sharing in Washington, D.C. and Brisbane, Australia : Implications for Policy and Planning », *in* : *International Journal of Sustainable Transportation* 0.*ja*, null.

Ahmed, Mohiuddin, Abdun Naser Mahmood et Jiankun Hu (2016), « A survey of network anomaly detection techniques », *in* : *Journal of Network and Computer Applications* 60, p. 19 –31, issn : 1084-8045.

Alsger, Azalden A., Mahmoud Mesbah, Luis Ferreira et Hamid Safi (2015), « Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix », *in* : *Transportation Research Record : Journal of the Transportation Research Board* 2535, p. 88–96, eprint : `http://dx.doi.org/10.3141/2535-10`.

Arana, P., S. Cabezudo et M. Peñalba (2014), « Influence of weather conditions on transit ridership : A statistical study using data from Smartcards », *in* : *Transportation Research Part A : Policy and Practice* 59, p. 1 –12.

Armoogum, J, JL Madre, MO Gascon et D François (2010), « Les enquêtes nationales et locales sur la mobilité : sources et méthodes », *in* : *La revue du SOeS du CGDD*, p. 217–218.

Bagchi, M. et P. R. White (2004), « What role for smart-card data from bus systems ? », *in* : *Proceedings of the Institution of Civil Engineers - Municipal Engineer* 157.*1*, p. 39–46, eprint : `https://doi.org/10.1680/muen.2004.157.1.39`.

— (2005), « The potential of public transport smart card data », *in* : *Transport Policy* 12(5), p. 464–474.

Barry, J.J., R. Newhouser, A. Rahbee et S. Sayeda (2002), « Origin and destination estimation in New York City with automated fare system data », *in* : *Transportation Research Record* 1817, p. 183–187.

Bay, Stephen D et Mark Schwabacher (2003), « Mining distance-based outliers in near linear time with randomization and a simple pruning rule », *in* : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, p. 29–38.

Breunig, Markus M, Hans-Peter Kriegel, Raymond T Ng et Jörg Sander (2000), « LOF : identifying density-based local outliers », *in* : *ACM sigmod record*, t. 29, 2, ACM, p. 93–104.

Briand, Anne-Sarah, Etienne Côme, Mohamed K. El Mahrsi et Latifa Oukhellou (2016), « A mixture model clustering approach for temporal passenger pattern characterization in public transport », *in* : *International Journal of Data Science and Analytics* 1.*1*, p. 37–50, issn : 2364-4168.

Ceapa, I., C. Smith et L. Capra (2012), « Avoiding the Crowds : Understanding Tube Station Congestion Patterns from Trip Data », *in* : *Proceeding of the 1st ACM SIGKDD International Workshop on Urban Computing. ACM press*, p. 134–141.

CEREMA (2013), *Les enquêtes déplacements « standard CERTU»*, rapp. tech., Centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques.

Chu, K. et R. Chapleau (2008), « Enriching Archived Smart Card Transaction Data for Transit Demand Modeling », *in* : *Transportation Research Record : Journal of the Transportation Research Board* 2063, p. 63–72.

Chu, Ka Kee Alfred (2015), « Two-year Worth of Smart Card Transaction Data – Extracting Longitudinal Observations for the Understanding of Travel Behaviour », *in* : *Transportation Research Procedia* 11, Transport Survey Methods : Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia, p. 365 –380, issn : 2352-1465.

Devillaine, F., M. Munizaga et M. Trépanier (2012), « Detection of activities of public transport users by analyzing smart card data », *in* : *Transportation Research Record : Journal of the Transportation Research Board* 2276, p. 48–55.

El Mahrsi, Mohamed K., Etienne Côme, Johanna Baro et Latifa Oukhellou (2014), « Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data », *in* : *3rd International Workshop on Urban Computing (UrbComp), ACM SIGKDD Conference*, New York, USA.

Farooq, U., T. u. Haq, M. Amar, M. U. Asad et A. Iqbal (2010), « GPS-GSM Integration for Enhancing Public Transportation Management Services », *in* : *2010 Second International Conference on Computer Engineering and Applications*, t. 2, p. 142–147.

Foell, Stefan, Santi Phithakkitnukoon, Gerd Kortuem, Marco Veloso et Carlos Bento (2014), « Catch Me If You Can : Predicting Mobility Patterns of Public Transport Users », *in* : *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, China, p. 8 –11.

Furtlehner, Cyril et al. (2010), « Spatial and Temporal Analysis of Traffic States on Large Scale Networks », *in* : *13th International IEEE Conference on Intelligent Transportation Systems ITSC'2010*, Madère, Portugal, p. –.

Fuse, T., K. Makimura et T. Nakamura (2010), « Observation of travel behavior by ic card data and application to transportation planning », *in* : *Special Joint Symposium of ISPRS Commission IV and AutoCarto 2010*.

Gordon, Jason, Harilaos Koutsopoulos, Nigel Wilson et John Attanucci (2013), « Automated inference of linked transit journeys in London using fare-transaction and vehicle location data », *in* : *Transportation Research Record : Journal of the Transportation Research Board 2343*, p. 17–24.

He, L, N Nassir, M Trepanier et M Hickman (2015), *Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems*, rapp. tech.

Hido, Shohei, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama et Takafumi Kanamori (2011), « Statistical outlier detection using direct density ratio estimation », *in* : *Knowledge and Information Systems* 26.*2*, p. 309–336, issn : 0219-3116.

IPA (2016), *Driving change : Australia's cities need a measured response*, rapp. tech., Infrastructure Partnerships Australia.

Jr., Joe H. Ward (1963), « Hierarchical Grouping to Optimize an Objective Function », *in* : *Journal of the American Statistical Association* 58.*301*, p. 236–244, eprint : `http://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845`.

Jung, Jaeyoung et Keemin Sohn (2017), « Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data », *in* : *IET Intelligent Transport Systems*.

Knorr, Edwin M et Raymond T Ng (1999), « Finding intensional knowledge of distance-based outliers », *in* : *VLDB*, t. 99, p. 211–222.

Langlois, Gabriel Goulet, Haris N. Koutsopoulos et Jinhua Zhao (2016), « Inferring patterns in the multi-week activity sequences of public transport users », *in* : *Transportation Research Part C : Emerging Technologies* 64, p. 1 –16, issn : 0968-090X.

Latecki, Longin Jan, Aleksandar Lazarevic et Dragoljub Pokrajac (2007), « Outlier Detection with Kernel Density Functions », *in* : *Machine Learning and Data Mining in Pattern Recognition : 5th International Conference, MLDM 2007, Leipzig, Germany, July 18-20, 2007. Proceedings*, sous la dir. de Petra Perner, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 61–75, isbn : 978-3-540-73499-4.

Lathia, N., J. Froehlich et L. Capra (2010), « Mining public transport usage for personalised intelligent transport systems », *in* : *IEEE International Conference on Data Mining. Sydney, Australia*.

Lee, Sang Gu et Mark Hickman (2013), « Trip purpose inference using automated fare collection data », *in* : *Public Transport* 6.*1*, p. 1–20.

Lenormand, Maxime et al. (2014), « Cross-Checking Different Sources of Mobility Information », *in* : *PLOS ONE* 9.*8*, p. 1–10.

Li, Yang, Xudong Wang, Shuo Sun, Xiaolei Ma et Guangquan Lu (2017), « Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks », *in* : *Transportation Research Part C : Emerging Technologies* 77, p. 306 –328, issn : 0968-090X.

Li, Yanshan, Weiming Liu et Qinghua Huang (2016), « Traffic anomaly detection based on image descriptor in videos », *in* : *Multimedia Tools and Applications* 75.*5*, p. 2487–2505, issn : 1573-7721.

López-Pintado, Sara et Juan Romo (2009), « On the Concept of Depth for Functional Data », *in* : *j-J-AM-STAT-ASSOC* 104.*486*, p. 718–734, issn : 0162-1459 (print), 1537-274X (electronic).

Ma, Xiao-lei, Yao-Jan Wu, Yin-hai Wang, Feng Chen et Jian-feng Liu (2013), « Mining smart card data for transit riders' travel patterns », *in* : *Transportation Research Part C : Emerging Technologies* 36.*0*, p. 1 –12.

MacQueen, J. (1967), « Some methods for classification and analysis of multivariate observations », *in* : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, Berkeley, Calif. : University of California Press, p. 281–297.

Mahrsi, M. K. El, E. Côme, L. Oukhellou et M. Verleysen (2016), « Clustering Smart Card Data for Urban Mobility Analysis », *in* : *IEEE Transactions on Intelligent Transportation Systems* PP.*99*, p. 1–17, issn : 1524-9050.

McLachlan, Geoffrey J et Thriyambakam Krishnan (2008), *The EM algorithm and extensions*, Wiley.

Milion, Chloe (2015), « méthodes et modèles pour l'étude de la mobilité des personnes par l'exploitation de données de radiotéléphonie », thèse de doct., Université Paris-Est.

Momtazpour, Marjan et Naren Ramakrishnan (2015), « Characterizing Taxi Flows in New York City », *in* : *Proceedings of the International Workshop on Urban Computing (UrbComp)*.

Morency, C., M. Trépanier et B. Agard (2006), « Analysing the variability of transit users behaviour with smart card data », *in* : *The Ninth International IEEE Conference on Intelligent Transportation Systems, Toronto, Canada, September*.

Moutarde, Fabien et Yufei Han (2011), « A new traffic-mining approach for unveiling typical global evolutions of large-scale road networks », *in* : *18th World Congress on Intelligent Transport Systems (ITSwc'2011)*, TS17–2236.

Munizaga, M. A. et C. Palma (2012), « Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smart card data from Santiago, Chile », *in* : *Transportation Research Part C : Emerging Technologies* 24, p. 9–18.

Munizaga, Marcela, Flavio Devillaine, Claudio Navarrete et Diego Silva (2014), « Validating travel behavior estimated from smartcard data », *in* : *Transportation Research Part C : Emerging Technologies* 44, p. 70 –79.

Nassir, Neema, Mark Hickman et Zhen-Liang Ma (2015), « Activity detection and transfer identification for public transit fare card data », *in* : *Transportation* 42.*4*, p. 683–705.

Pang, Linsey Xiaolin, Sanjay Chawla, Wei Liu et Yu Zheng (2013a), « On Detection of Emerging Anomalous Traffic Patterns Using GPS Data », *in* : *Data and Knowledge Engineering*.

— (2013b), « On detection of emerging anomalous traffic patterns using {GPS} data », *in* : *Data & Knowledge Engineering* 87, p. 357 –373, issn : 0169-023X.

Park, J. Y., D.-J. Kim et Y. Lim (2008), « Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea », *in* : *Transportation Research Record : Journal of the Transportation Research Board* 2063, p. 3–9.

Pelletier, Marie-Pier, Martin Trépanier et Catherine Morency (2011), « Smart card data use in public transit : A literature review », *in* : *Transportation Research Part C : Emerging Technologies* 19.*4*, p. 557–568.

Pham, Thi Huong Thao (2016), « Apports et difficultés d'une collecte de données à l'aide de récepteurs GPS pour réaliser une enquête sur la mobilité », thèse de doct., Paris Est.

Poussevin, Mickaël, Nicolas Baskiotis, Vincent Guigue et Patrick Gallinari (2014), « Mining ticketing logs for usage characterization with nonnegative matrix factorization », *in* : *SenseML 2014–ECML Workshop*.

Shannon, Claude Elwood (1948), « A mathematical theory of communication », *in* : *Bell System Technical Journal* 27.*3-4*, 379–423 and 623–656.

Steinwart, Ingo, Don Hush et Clint Scovel (2005), « A Classification Framework for Anomaly Detection », *in* : *J. Mach. Learn. Res.* 6, p. 211–232, issn : 1532-4435.

STRATEC (2017), *Enquête pilote smartphone en Région de Bruxelles-Capitale*, conférence sur les big data.

Sun, Ying et Marc G. Genton (2011), « Functional Boxplots », *in* : *Journal of Computational and Graphical Statistics* 20.*2*, p. 316–334, eprint : `http://dx.doi.org/10.1198/jcgs.2011.09224`.

Tao, Sui, David Rohde et Jonathan Corcoran (2014), « Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap », *in* : *Journal of Transport Geography* 41, p. 21 –36.

Tonnelier, Emeric, Nicolas Baskiotis, Vincent Guigue et Patrick Gallinari (2017), « Anomaly detection and characterization in smart card logs using NMF and Tweets », *in* : *ESANN*.

Toqué, F., E. Côme, M. K. El Mahrsi et L. Oukhellou (2016), « Forecasting dynamic public transport Origin-Destination matrices with long-Short term Memory recurrent neural networks », *in* : *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, p. 1071–1076.

Trépanier, M., N. Tranchant et R. Chapleau (2007), « Individual trip destination estimation in a transit smart card automated fare collection system », *in* : *Intelligent Transportation Systems* 11, p. 1–14.

Trépanier, Martin et R. Chapleau (2001), « Linking Transit Operational Data to Road Network with a Transportation Object-Oriented GIS », *in* : *Urban and Regional Information Systems Association Journal* 2.*13*, p. 23–27.

Trépanier, Martin, Khandker M.N. Habib et Catherine Morency (2012), « Are transit users loyal ? Revelations from a hazard model based on smart card data », *in* : *Canadian Journal of Civil Engineering* 39.*6*, p. 610–618.

Tukey, John W (1977), « Exploratory data analysis », *in* :

UBER (2016), *Using Uber Movement to understand the effects of DC Metrorail service disruptions on traffic congestion*, rapp. tech., UBER Movement.

Utsunomiya, M., J. Attanucci et N. Wilson (2006), « Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning », *in* : *Transportation Research Record* 1971, p. 119–126.

Wang, Junjie, Dong Wei, Kun He, Hang Gong et Pu Wang (2014), « Encapsulating urban traffic rhythms into road networks », *in* : *Scientific reports* 4.

Wang, Youcheng et al. (2016), « A Feature-based Method for Traffic Anomaly Detection », *in* : *Proceedings of the 2Nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, UrbanGIS '16, New York, NY, USA : ACM, 5 :1–5 :8, isbn : 978-1-4503-4583-5.

Yamanishi, Kenji et Jun-ichi Takeuchi (2002), « A unifying framework for detecting outliers and change points from non-stationary time series data », *in* : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, p. 676–681.

Zhang, Ji (2013), « Advancements of outlier detection : A survey », *in* : *ICST Transactions on Scalable Information Systems* 13.*1*, p. 1–26.

Zhao, Fang et al. (2015), « Exploratory analysis of a smartphone-based travel survey in Singapore », *in* : *Transportation Research Record : Journal of the Transportation Research Board* 2.*2494*, p. 45–56.

Zhao, J., A. Rahbee et N. Wilson (2007), « Estimating a rail passenger trip origin–destination matrix using automatic data collection systems », *in* : *Computer-Aided Civil and Infrastructure Engineering* 22, p. 376–387.

Zheng, Yu, Licia Capra, Ouri Wolfson et Hai Yang (2014), « Urban Computing : Concepts, Methodologies, and Applications », *in* : *ACM Transaction on Intelligent Systems and Technology*.

Zhong, Chen et al. (2016), « Variability in Regularity : Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data », *in* : *PLOS ONE* 11.*2*, p. 1–17.

Table des figures