Quantitative models of establishments location choices: spatial effects and strategic interactions

Sabina Buczkowska

Quantitative models of establishment location choices: Spatial effects and strategic interactions

By
Sabina Buczkowska

A Dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy in
Economic Science of the
University Paris-Est
OMI Doctoral School

Committee in charge:

Jean-Loup Madre, Research Director, IFSTTAR (Thesis Director)
Matthieu de Lapparent, Ecole Polytechnique Fédérale de Lausanne (Thesis Advisor)
François Legendre, U. Prof., Université Paris-Est Créteil (President)
Josep-Maria Arauzo-Carod, U. Prof., Universitat Rovira i Virgili (Rapporteur)
Sophie Masson, U. Prof., Institut Universitaire de Technologie de Perpignan (Rapporteur)
Coro Chasco Yrigoyen, U. Prof., Universidad Autónoma de Madrid (Examiner)
Laurent Gobillon, U. Prof., Paris School of Economics/INED (Examiner)

Publicly defended on March, 28, 2017

IFSTTAR

Université Paris-Est

"L'Ecole doctorale des Organisations, Marchés, Institutions de l'Université Paris-Est Créteil ne donne ni approbation ni improbation aux opinions exprimées dans cette thèse. Ces opinions sont considérées comme propres à leur auteure".

To Camil

Acknowledgments

You know how it is. You pick up a book,
flip to the dedication, and find that, once again,
the author has dedicated a book to someone else
and not to you. Not this time.

- N. Gaiman, "Anansi Boys"

IT WAS AT THE KARLSTAD UNIVERSITY library in the year 2008 when the light bulb moment struck. The Anselin's book on spatial analysis came into my hands and its contents stayed in my mind until today. The day of completing the last chapter. The day of bringing the thesis to an end.

It never crossed my mind that I'd go live in another country. But as unlikely as it was, once I went, I never looked back. After living and studying in Poland, Sweden, and Germany, it was France which conquered my heart. Yet, our love was laborious, demanding, and grueling. Exactly seven years now living in Paris. That journey, from a mute new expat/immigrant to an eventually quite fluent (or let's use with impunity a word "O.K.") French speaker, from knowing nobody to saying Hi! to all the "boulangers" and butchers on my street, from wolfing down donuts to relishing the French Pâtisserie with a capital letter, from being airy-fairy to becoming a serious mother.

That journey would not be the same without a couple of people met on my way. To Prof. André de Palma. Oh! I have already struck without even sending a warning signal. So yes, the first "Thank you" goes to this distinct, yet very particular professor for being demanding the way he is and for giving me the chance to stay in Paris after my master program ended up and my adult life began; Next, I send all my gratitude to all the Seventh EU Framework Program SustainCity Project members, mentioning Saif in particular, with whom I shared the desk at the ENS Cachan office and without whom playing music at 7 P.M. while taking a break from work would not be fun otherwise.

No thesis would be accomplished without the following smaller and bigger personæ. I am indebted to:

My international family: the Erasmus Mundus Program students who not only substituted my family but also perfectly complemented it, notably Alice for sending me nicely strong Karma coffee from the Laos mountains and lemongrass tea from Thailand to wind down from the coffee effect, Esfandiar, Yan, my Kabutar Mina for million ”miss you, miss you” words!, my awe-inspiring friend Tara, and my little Chinese sister Guodi. This one’s for you. You probably know why.

My second international family for sharing the floor in the American Fondation: my dancing Rida for her never-failing optimism, my American Polish brother Mark Zaborowski, multitalented Baidy, and to Tierra for her writing talent who helped me scrape through my very first scientific paper.

Ana, the real fighter and the Taekwondo Master for all the <3s and accepting me as I am; Gordon Bowker, Jerry Baldwin, Zev Siegl for founding Starbucks, so that we could find there a secondary office with Ana; Meno, Vadim, godfather Marouene, Adams - the tektonic specialist and the best uncle, exceptional Paola, Hamza number one and Hamza number two, the rockets builder Heric, and finally the most talented and probably the youngest Senior Researcher from the Statistics Finland, Henri. Hi Henri.

Friends from IFSTTAR and the DEST Laboratory: Benoît for showing me how cool French people can be; ”Irritating” David who kept on reminding me not to take myself too seriously; Katia, Christine, and Hoai-Thu, the Ifsttar super mothers; and the most original and the best coiffé Kevin from the Ifsttar cafeteria for providing the strong enough coffee.

Prof. Michel Bierlaire who let me join his team at the École Polytechnique Fédérale de Lausanne for a couple of months during my thesis and who gave me the chance to answer correctly maybe two out of his hundred super challenging questions; and all the members of his team: always in trouble Yousef, Shadi, Marija, the genius Iliya, eating porridge at 4 P.M. Flurin, Antonin, running fast Tomáš, Evantia, strong Riccardo, Anna, and Stefan.

My heroes: Prof. Anselin who I never met in my life but who is responsible for my love to spatial analysis and whose work in many ways motivated this project; My idol, Prof. LeSage who I had a tremendous honor to meet and to talk to about my research, whose comments helped clarify my own thinking, and who almost convinced me that Bayesian analysis is the remedy to all the World’s problems; Prof. Krzysztof Malaga, a definition of hard work and a great example to students.

Slowly heading to an end of dedication, I will definitely not forget to thank: Nicolas, with whom looking for new ideas and writing articles was an unspeakable joy of creation, for staying attentive after not closing an eye during a couple of nights. It is not an easy task to be a multi-task young father-researcher;

Matthieu de Lapparent, the thesis advisor, to whom I have a huge respect for not only being a great professional mentor, but especially a great human, who over the years has taught me more than anyone;

Jean-Loup Madre, the thesis director, whose encouragement and supreme support of all my projects were essential;

and all the Ph.D. Committee Members for your time and patience while reading this thesis.

Finally, I am most deeply thankful to the closest ones: My father, the best advice-giver and the fantastic professor, who installed within me a love of science, who encouraged me to change my study field from finance and banking to econometrics. Thank you for being stubborn enough at that moment of my life; My father (the same one) to whom I promised that I will never ever go for a Ph.D. He waited for a moment and said... ”Oh yes, darling, you will”; My mom for her never-ever-ending support (literally), for these thousand Whatsapp, Facebook, phone... messages per day. Dziekuje; My grandma, the most energetic grandmother on Earth, I guess; My older (and of course more clever) sister Kamila for being an inspiration in all the life domains; The best-est Marta. A word ”best” would not be enough in her case. Adel, the biggest Cyril Lignac eater, the worst swimmer and bike rider I’ve got to know, the most dedicated father, and the best husband. You are perfect to me.

... and to Camil who decided to cheerfully teethe when I was trying to finish my thesis. To Camil, without whose unvarying emotional realm of love and a constant smile this thesis would have been completed in half the time.

After these couple of words before the story even begins... 3,5 year that this thesis was in the making, here’re six chapters. Nearly everything I intended is in it, and still the thesis is not full. Good and bad ideas are in it along with only a little bit of despondency and enormous pleasure of discovering. Here’re six chapters through which I tried to briefly and gracefully communicate a number of things. Each chapter tells one story. All the stories create an image which I could definitely keep on painting further. But let’s close it for now and let’s raise the curtains.

Paris, February 7, 2017
Sabina Buczkowska

Quantitative models of establishment location choices: Spatial effects and strategic interactions

Abstract

This thesis is breathing new life into the location choice models of establishments. The need for methodological advances in order to more realistically model the complexity of establishment decision-making processes, such as their optimal location choices, is the key motivation of this thesis.

First, location choice models use geo-referenced data, for which choice sets have an explicit spatial component. It is thus critical to understand how to represent spatial aspect in location choice models. The final decision of an establishment seems to be related to the surrounding economic landscape. When accounting for the linkage between neighboring observations, the decision on the spatial weight matrix specification must be made. Yet, researchers overwhelmingly apply the Euclidean metric without realizing its underlying assumptions and its alternatives. This representation has been originally proposed due to scarce data and low computing power, rather than because of its universality. In areas, such as the Paris region, where high congestion or uncrossable physical barriers problems clearly arise, distances purely based on topography may not be the most appropriate for the study of intra-urban location. There are insights to be gained by mindfully reconsidering and measuring distance depending on a problem being analyzed. Rather than locking researchers into a restrictive structure of the weight matrix, this thesis proposes a flexible approach to intimate which distance metric is more likely to correctly account for the nearby markets depending on the sector considered. In addition to the standard Euclidean distance, six alternative metrics are tested: travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances.

Second, what makes these location choices particularly interesting and challenging to analyze is that decisions of a particular establishment are interrelated with choices of other players. These thorny problems posed by the interdependence of decisions generally cannot be assumed away, without altering the authenticity of the model of establishment decision making. The conventional approaches to location selection fail by providing only a set of systematic steps for problem-solving without considering strategic interactions between the establishments in the market. One of the goals of the present thesis is to explore how to correctly

adapt location choice models to study establishment discrete choices when they are interrelated.

Finally, a firm can open a number of units and serve the market from multiple locations. Once again, traditional theory and methods may not be suitable to situations wherein individual establishments, instead of locating independently from each other, form a large organization, such as a chain facing a fierce competition from other chains. There is a necessity to incorporate interactions between units within the same and competing firms. In addition, the need to state a clear difference between the daytime and nighttime population has been emphasized. Demand is represented by pedestrian and car flows, the crowd of potential clients passing through the commercial centers, train and subways stations, airports, and highly touristic sites. The Global Survey of Transport (EGT 2010), among others, is of service to reach this objective.

More realistically designed location choice models accounting for spatial spillovers, strategic interaction, and with a more appropriate definition of distance and demand can become a powerful and flexible tool to assist in finding a befitting site. An appropriately chosen location in turn can make an implicative difference for the newly-created business. The contents of this thesis provide some useful recommendations for city planners, plan developers, business owners, and shopping center investors. We asked a number of burning questions through this thesis and attempted to open the door to possible answers or to stimulate further reflection on location choice models.

Keywords: location choice, count data model, discrete choice model, hurdle model, excess zero problem, Tobit model, spatial spillovers, Euclidean distance, travel time, congestion, physical barriers, strategic interactions, Bayesian Nash, discrete game, interdependence of sectors, within- and inter-industry interactions, daytime population, demand, multi-store firms, nearest neighbor distance, rival- and same-chain stores, Paris region

Modèles quantitatifs de choix de localisation des établissements : effets spatiaux et interactions stratégiques

Résumé

Dans un contexte de carence méthodologique, cette thèse vise à apporter un nouveau souffle aux modèles de choix de localisation jusqu'ici incapables d'appréhender de manière réaliste la complexité des processus décisionnels des établissements tels que leurs choix de localisation optimale.

Les modéles de choix de localisation utilisent des données géoréférencées, pour lesquelles les ensembles de choix ont une composante spatiale explicite. Il est donc essentiel de comprendre comment représenter l'aspect spatial dans les modèles de choix de localisation. La décision finale d'un établissement semble être liée au paysage économique environnant. La quantification du lien entre les observations voisines implique une prise de décision sur la spécification de la matrice spatiale. Pourtant, la grande majorit□ des chercheurs appliquent la métrique euclidienne sans considérer de les hypothèses sous-jacentes et ses alternatives. Cette démarche a été initialement proposée en raison de données et de puissance informatique limitées plut□t que de son universalité. Dans les régions comme la région parisienne, oû la congestion ainsi que les problèmes de barrières physiques non traversables apparaissent clairement, les distances purement basées sur la topographie peuvent ne pas être les plus appropriées pour l'étude de la localisation intra-urbaine. Il est possible d'acquérir des connaissances en reconsidérant et en mesurant la distance en fonction du problème analysé. Plutôt que d'enfermer les chercheurs dans une structure restrictive de la matrice de pondération, cette thèse propose une approche souple pour identifier la métrique de distance la plus susceptible de prendre en compte correctement les marchés voisins selon le secteur considéré. En plus de la distance euclidienne standard, six autres mesures sont testées : les temps de déplacement en voiture (pour les périodes de pointe et hors pointe) et en transport en commun, ainsi que les distances de réseau correspondantes.

Par ailleurs, les décisions d'un établissement particulier sont interdépendantes des choix d'autres acteurs, ce qui rend les choix de localisation particulièrement intéressants et difficiles à analyser. Ces problèmes épineux posés par l'interdépendance des décisions ne peuvent généralement être négligés sans altérer l'authenticité du modèle de décision d'établissement. Les approches classiques de la sélection de localisation échouent en ne fournissant qu'un en-

semble d'étapes systématiques pour la résolution de problèmes sans tenir compte des interactions stratégiques entre les établissements sur le marché. L'un des objectifs de la présente thèse est d'explorer comment adapter correctement les modèles de choix de localisation pour étudier les choix discrets d'établissement lorsqu'ils sont interdépendants.

En outre, une entreprise peut ouvrir un certain nombre d'unités et servir le marché à partir de plusieurs localisations. Encore une fois, la théorie et les méthodes traditionnelles peuvent ne pas convenir aux situations dans lesquelles les établissements individuels, au lieu de se situer indépendamment les uns des autres, forment une grande organisation, telle qu'une chaîne confrontée à une concurrence féroce d'autres chaînes. Le modèle prend en compte non seulement les interactions intra-chaînes mais aussi inter-chaînes. Aussi, la nécessité d'indiquer une nette différence entre la population de jour et de nuit a été soulignée. La demande est représentée par les flux de piétons et de voitures, la foule de clients potentiels passant par les centres commerciaux, les stations de trains et de métros, les aéroports et les sites touristiques. L'Enquête Globale Transport 2010 (EGT 2010), entre autres, est utile pour atteindre cet objectif.

Des modèles de choix de localisation plus réalistes conçus pour expliquer les débordements spatiaux, des interactions stratégiques et qui appliquent une définition plus appropriée de la distance et de la demande peuvent devenir un outil puissant et flexible pour aider à trouver un site adapté. Un lieu convenablement choisi peut également constituer un avantage pour les nouveaux établissements. Le contenu de cette thèse fournit des recommandations utiles pour les urbanistes, les promoteurs immobiliers, les propriétaires d'entreprises et les investisseurs de centres commerciaux.

En somme, nous posons dans cette thèse un certain nombre de questions brûlantes à travers lesquelles nous tentons d'ouvrir la porte à des questions possibles ou de stimuler la réflexion sur les modèles de choix de localisation.

Mots-clés : choix de localisation, modèle de comptage, modèle de choix discret, débordements spatiaux, distance euclidienne, temps de déplacement, congestion, barrières physiques, interactions stratégiques, interdépendance des secteurs, interactions interindustrielles, population de jour, entreprises multi-magasins, Île-de-France

Name and address of the research laboratory:

Department for Planning, Mobility and Environment - AME
Laboratory Economic and Social Dynamics of Transport - DEST
French Institute of Science and Technology for Transport, Development and Networks - IFSTTAR
14-20 Boulevard Newton
Cité Descartes, Champs sur Marne, F-77447 Marne la Vallée Cedex 2
France
www.ifsttar.fr

Intitulé et adresse du laboratoire de recherche :

Département Aménagement, Mobilités et Environnement - AME
Laboratoire Dynamiques Economiques et Sociales des Transports - DEST
Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux - IFSTTAR
14-20 Boulevard Newton
Cité Descartes, Champs sur Marne, F-77447 Marne la Vallée Cedex 2
France
www.ifsttar.fr

Contents

1 General introduction 1
 1.1 Location choice models of establishments . . . 1
 1.2 Key questions . . . 2
 1.3 Relevance of research . . . 3
 1.4 Literature . . . 7

2 Location choices of newly-created establishments: Spatial patterns at the aggregate level 11
 2.1 Introduction . . . 12
 2.2 Review of empirical literature . . . 13
 2.3 Study area . . . 18
 2.4 Data . . . 19
 2.5 Descriptive statistics . . . 23
 2.6 Parametric models . . . 32
 2.7 Results . . . 37
 2.8 Conclusions . . . 42

3 Euclidean versus network distances: A mixture model of the intra-metropolitan location choice of establishments 51
 3.1 Introduction . . . 52
 3.2 Literature review . . . 56
 3.3 Data . . . 59
 3.4 Model . . . 65
 3.5 Results . . . 69
 3.6 Conclusions . . . 81

4 Location choices under strategic interactions: Interdependence of establishment types 93
 4.1 Introduction . . . 94
 4.2 Within-industry and inter-industry interactions . . . 97
 4.3 Why bother about strategic interactions? . . . 98

4.4 Push and pull factors 98
4.5 Data 100
4.6 Model 107
4.7 Results 116
4.8 Conclusions and possible applications 125

5 Locational strategies of multi-store firms: A daytime population 131
5.1 Introduction 132
5.2 Literature review 135
5.3 Data 145
5.4 Market description 149
5.5 Nearest neighbor distance 152
5.6 Model: Reduced-form profit function 159
5.7 Empirical results 161
5.8 Conclusions 180

6 Concluding remarks 187
6.1 Main findings 189
6.2 Practical recommendations 192
6.3 Future research directions 193

Listing of figures

2.1 Newly-created establishments in 2007 - rankings based on the market share - focus on Paris and the inner ring. 26

3.1 Road network (upper figure) and public transit network (lower figure) in the Paris region 62

3.2 Example 1. Google map image: Possible barriers which require to make detours: a river, islands, a lake, parks, a golf course, stadiums, highways, a national road, departmental roads, train routes. Euclidean (958 m) versus real distance (4.0 km and 12 min) from point A to B in the area of Boullereaux-Champigny, the Paris region. Source: Google My Maps 69

3.3 Example 2. Google map image: Possible barriers which require to make detours: a river, lakes, an attraction island theme park, parks, a highway, departmental roads, train routes, electric and thermic centers, cemetery. Euclidean (2.66 km) versus real distance (8.2 km and 13 min) from point A to B in the area of Champs-sur-Marne, the Paris region. Source: Google My Maps 70

3.4 Selected example of the estimates of the truncated-at-zero Poisson parts of the mixture models for the construction sector. Comparison of two cases for models that use two classes. 90

3.5 Posterior probabilities of belonging to class #1 (ED) as opposed to class #2 (TtVhMph) at the municipality level (each municipality can be treaten as an alternative in the decision-making process of an establishment): construction sector (upper figure), special, scientific and technical activities (central figure), and real estate sector (lower figure) 91

4.1 109 EGT zones of the Paris region 102

5.1 Example of data representation: Zones' characteristics. Various data sources; The most touristic sites in the Paris region with over 1 million and 0.5 million tourists each year in 2012; Locations of train stations. Source: Ministry of Crafts, Commerce and Tourism, "Memento du tourisme 2013", General Directorate for Competitiveness, Industry and Services 147

5.2 Monthly turnover index value in million euros for the fast-food and traditional restaurants over time. Situation for the metropolitan France from January 1995 until August 2015. Reference period: 2010. Source: INSEE, DGFiP 149
5.3 How the best performing fast-food chains in the Paris region share the market between each other over time. Accumulated monthly number of newly opened stores in the Paris region between 2001 and 2015, including the stock of stores registered before January 2001. Source: Calculations based on the SIRENE/INSEE database 150
5.4 Attractiveness of the Paris region for the best performing chains: The global situation which considers the total accumulated number of all pre-existing stores of these chains in 2015. Source: Calculations based on the SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains 152
5.5 The change of location strategy over time: The change of the attractiveness level of Paris, the closer and outer suburbs for the best performing chains. Focus on the situation between 2001 and 2015. Source: Calculations based on the SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains . . 169
5.6 Examples of location strategies of McDonald's versus Pomme and McDonald's versus Pomme and Paul in the central part of Paris in 2015 using Google My Maps. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains 170
5.7 Examples of location strategies of McDonald's versus Quick and McDonald's versus KFC in the central part of Paris in 2015 using Google My Maps. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains . . 171
5.8 Examples of location strategies of establishments of McDonald's, KFC, Quick, Paul, and Pomme de Pain in Paris, the closer suburbs, and the outer suburbs in 2015 using Google My Maps. The same scale has been kept through Fig. 5.6 - 5.8, so that the between store distances can be compared. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains 172
5.9 Locations of establishments of McDonald's, Quick, KFC, Paul, and Pomme de Pain in the Paris region (pictures on the left) and in 80 districts of Paris (pictures on the right) at the individual establishment level. Situation in 2015. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains . . 173

Listing of tables

2.1 Description of potential explanatory variables. 21
2.2 Market share of top 1, 10, 20, 50, 100 municipalities in 2007. 25
2.3 Ranking of top 10 municipalities based on the market share (#newly-created establishments to total #establishments in municipality) 27
2.4 Structure of newly-created establishments and existing establishments in each *département* in 2007; Comparison with 2003 and 2005 29
2.5 Structure of newly-created and pre-existing establishments by size category. 31
2.6 Hurdle model with distance matrix: explanatory variables represent density of pre-existing establishments in the respective sector, density of large pre-existing establishments, and various municipality characteristics. 49
2.7 Hurdle model with distance matrix: explanatory variables represent density of pre-existing establishments in the respective sector, density of large pre-existing establishments, and various municipality characteristics (continuation). 50

3.1 Description of explanatory variables used in the final models. 64
3.2 Pair correlations between Euclidean distance and transport distances. 72
3.3 Correlations between Euclidean distance and transport distances by increasing range intervals (in km). 73
3.4 Probabilistic mixture of hurdle-Poisson models for the construction sector that uses two latent classes, case #1: ED (class #1) and TtVhMph (class #2) versus two "mono-distance" models using first ED and then TtVhMph. . . 77
3.5 Estimated posterior probability of belonging to class #1 (`Pi`): Comparison across cases and sectors. 80
3.6 Probabilistic mixture of hurdle-Poisson models for spec., scien., tech. and real estate activities using two latent classes for case #1: ED (class #1) with TtVhMph (class #2). Hurdle and truncated-at-zero Poisson parts reported. . 89

4.1 Selected types of newly-created establishments 101
4.2 Structure of pre-existing establishments by activity sector in the beginning of 2007 103
4.3 Structure of pre-existing establishments by size category in the beginning of 2007 103

4.4 Pre-existing establishments' location across the Paris region in the beginning of 2007 104
4.5 Final list of variables used in the models (EGT zones characteristics dictionary) 106
4.6 Models estimates: Seven location choice models run simultaneously 120
4.7 Strategic interactions matrix: How the decision of one establishment type affects the decision of other establishment types. 123
4.8 Variance-covariance matrix of residuals. 124

5.1 The pairwise ANN index. Case #1: Paris 154
5.2 The pairwise ANN index. Case #2: Paris and its closer suburbs 155
5.3 The pairwise ANN index. Case #3: The Paris region 156
5.4 Location choices of McDonald's: Focus on market characteristics 165
5.5 Location choices of Quick: Focus on market characteristics 166
5.6 Location choices of KFC: Focus on market characteristics 166
5.7 Location choices of Paul: Focus on market characteristics 167
5.8 Location choices of Pomme: Focus on market characteristics 167
5.9 Number of newly-created and existing stores 167
5.10 Location choices of McDonald's: Focus on market structure (cLogit) 174
5.11 Location choices of all chains: Focus on new market and nearby markets variables (cLogit) 177
5.12 Location choices of all chain stores: Focus on same- and inter-chain effects (cLogit) 179

6.1 Explanation of French classification of activities (NAF nomenclature - Rev. 2) 208

1
General introduction

1.1 Location choice models of establishments

This thesis is breathing new life into the location choice models of establishments. The need for methodological advances in order to more realistically model the complexity of establishment decision-making processes, such as their optimal location choices is the key motivation of this thesis. A clear distinction between an establishment and a firm should be made in the first place. An establishment is defined as a distinct economic unit that produces goods or services at a single physical location. In contrast, a firm is a legal entity that consists of one or more establishments or plants under common ownership and control (van Wissen, 2000).

This thesis is comprised of six chapters and is written in a form of essays. The general introduction and the overall conclusion encompass four scientific papers entitled: (1) Location choices of newly-created establishments: Spatial patterns at the aggregate level, (2) Euclidean versus network distances: A mixture model of the intra-metropolitan location choice of establishments, (3) Location choices under strategic interactions: Interdependence of establishment types, and (4) Locational strategies of multi-store firms: A daytime population. The first paper has been published in the "Regional Science and Urban Economics Journal". The second paper is under the second review at the ”Networks and Spatial Economics Journal”. Matthieu de Lapparent contributed to Chapter 2, 3, and 4. Nicolas Coulombel contributed to Chapter 3 and 5.

The approach being proposed in this thesis is mostly computational. Techniques explored in this thesis embody the models which use aggregate data and account for the presence of excess zeros, namely count data models, such as zero-inflated or hurdle models, or Tobit model which controls for the excess of zeros problem by means of an additional sample selection equation. Discrete choice models have been applied to study the establishment location choices at the individual level.

1.2 Key questions

A number of questions will be addressed within this thesis. First, location choice models use geo-referenced data, for which choice sets have an explicit spatial component. It is thus critical to understand how to represent spatial aspects in location choice models. Second, there are insights to be gained by further reconsidering and measuring distance depending on the problem being analyzed. To account for the linkage between neighboring observations, we will examine which distance metric is the most likely to correctly capture spatial spillovers. Third, what makes these location choices particularly interesting and challenging to analyze is that decisions of a particular establishment are interrelated with choices of the others. These thorny problems posed by the interdependence of decisions generally cannot be assumed away, without altering the authenticity of the model of establishment decision making. The conventional approaches to location selection fail by providing only a set of systematic steps for problem-solving without considering strategic interactions between the establishments in the market. One of the goals of this thesis is to explore how to correctly adapt location choice models to study establishments' discrete choices when they are interrelated. Fourth, a firm can open a number of units and serve the market from multiple locations. Once again traditional theory and methods may not be suitable to situations wherein individual establishments instead of locating independently from each other, form a large organization, such as a chain facing a fierce competition from other chains. There is a necessity to incorporate interactions between units within the same and competing firms. Illustrative questions that can be answered are: What is the nature and degree of competition for each analyzed chain? How does a situation of a firm change if there are already rivals established in a particular market versus if a firm faces no local competition? How does a firm perceive a rival store located in a direct vicinity and how if it is located farther away? Can it happen that a chain concentrates its investment efforts in a particular geographical area and uses another type of strategy in other places? How to more appropriately proxy the demand in decisions of revenue-oriented sectors? An intensified research effort along the lines of location choices is still desirable to provide answers to many such questions.

1.3 Relevance of research

1.3.1 Chapter 2 (Paper 1)

In this thesis, we concentrate our research on the Paris region, which exemplifies an appealing terrain to address for several reasons. The Paris region is a vibrant and innovative site with over 5.6 million jobs, 37 percent of national executives, and 40 percent of national workforce in research and development. It is the leading research and development hub in Europe and the third worldwide. Over 19 percent of the country's population, 11.7 million people, reside in the area which occupies only 2.2 percent of the surface of France. The GDP of the region amounts to 29 percent of total French GDP (IAU IdF, 2014). The Paris region is the third world touristic destination (in 2013) (Global Destination Cities Index 2015) with 16 millions of visitors from abroad.

Yet, the Paris region's economy is spatially unbalanced (Combes et al., 2011). The region is highly heterogeneous, especially regarding economic activity. While a few municipalities host a large number of newly-formed establishments, others struggle to be chosen, and a large group of municipalities is left with no new creation of establishments. When the observed data display a higher fraction of zeros than would be typically explained by the standard count data models, two types of models can be suggested: the hurdle (Mullahy, 1986) or the zero-inflated model (Lambert, 1992). The hurdle model reflects a two-part decision making process. It relaxes the assumption that the zero observations and the positive observations come from the same data generating process.

Although much work has been done in the domain of location choice models, several issues arise when analyzing involved phenomena, which scholars have yet to fully appreciate: (1) addressing the excess of zeros problem in the location choice model in highly heterogeneous geographic areas and (2) determining an appropriate way to accommodate spatial effects in location decisions. In the second chapter of this thesis (Paper 1) we respond to the complaint voiced by Bhat et al. (2014) and Liviano-Solís and Arauzo-Carod (2013) that heretofore the hurdle model technique has not been well investigated when analyzing location patterns. These are the first challenges that we face in the second chapter of this thesis.

1.3.2 Chapter 3 (Paper 2)

The third chapter extends the research on the hurdle model and the study presented in the first part of this thesis. The final decision of an establishment seems to be related to the surrounding economic landscape. When accounting for the linkage between neighboring observations, the decision on the spatial weight matrix specification should be made. Yet, since there exists no solitary claims on the concept of space, the form of the weight matrix is largely debated. One of the problems hides in the definition of distance usually based on the straight-line segment connecting two locations. Euclidean distance is typically used in the empirical literature and has also been utilized in the second chapter of this thesis to account for spatial spillovers in location choice model. However, Euclidean distance is believed to be only one simplistic possibility out of an infinite number of shortest path relations. Other alternative distance metrics may be proposed when building the spatial distance weight matrices.

Geographic factors, such as terrain, land cover, infrastructure, and traffic congestion may cause agents not to adhere to pure Euclidean relations. The Euclidean distance might thus not always be the most relevant depending on the given problem. Interest in this question dates at least to the 1960s and research on network models in geography (Haggett, 1967). There are insights to be gained by mindfully reconsidering and measuring distance. The third chapter investigates establishment location decisions in the Paris region where high congestion, speed limits, or physical uncrossable barriers, such as rivers or industrial corridors can diminish or completely eliminate the linkage between neighboring areas. Rather than imposing a restrictive structure on the weight matrix, this research proposes a flexible toolkit to identify which distance metric is more appropriate to correctly account for the surrounding economic landscape. A probabilistic mixture of two ”mono-distance” hurdle-Poisson models is developed. Each model's latent class uses a different distance representation to incorporate spillover effects in location choices of establishments from several activity sectors. In addition to the standard Euclidean distance, six alternative metrics are considered: travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances. This methodology allows one to capture the diversity of agents' behavior, *i.e.*, to distinguish which establishments are more time- or more distance-oriented, given their location.

1.3.3 Chapter 4 (Paper 3)

In the fourth chapter we further enhance the literature on the location choices, this time incorporating strategic interactions among establishments. We shed light on strategic interactions, fundamental in establishments' location choices, which is still largely unheeded in the empirical literature. If establishments acted in isolation, it would be a relatively simple task to adapt existing discrete-choice models. Yet, being non-strategic means that an establishment ignores other players' decisions. Less is known about how to correctly adapt location choice models to study establishments' discrete choices when they are interrelated. In very sparse empirical applications, when locational choice models are developed for several activity sectors, each model is typically run independently.

What makes these discrete choices particularly interesting and challenging to analyze is that decisions of a particular establishment are interrelated with choices of the others because an establishment accounts for the actions of other agents when making its own decisions (Draganska et al., 2008). These thorny problems posed by the interdependence of decisions generally cannot be assumed away, without altering the authenticity of the model of establishment decision making (Berry and Reiss, 2007). The conventional approaches to location selection, *i.e.*, traditional theory and methods, fail (Thill, 1997) by providing only a set of systematic steps for problem-solving without considering strategic interactions between the establishments in the market. An appropriately specified model of simultaneous entry or location decisions needs to recognize this interdependence of profits (Berry and Reiss, 2007).

There is a need for more realistic studies of complex establishments' decision-making processes. Even though the computational burden imposed by these models considering strategic interactions is relatively high, it seems that the costs imposed are more than offset by the benefits being accumulated (Draganska et al., 2008).

Strategic interactions have been largely unsung in the empirical analyses since the year 1929 when Hotelling (1929) brought the discussion in the industrial organization literature. Most of the papers are less than a decade old (Bajari et al., 2013). This literature is in its infancy, in part, due to the complexity of expressions for the probabilities used in the models which increases along with the number of locations and establishment types (Draganska et al., 2008).

We estimate a static discrete game of incomplete information to obtain a Bayesian Nash equilibrium at the group level using data at the aggregate level. We permit asymmetries across establishment types in the impact of interaction effects and exogenous market characteristics. We develop one location choice model which embraces seven individual models for seven establishment types run simultaneously to account for interactions from all the types on each other.

1.3.4 Chapter 5 (Paper 4)

The motivation of the last chapter comes from the fact that most previous discussion on locational decisions has one common feature of making unrealistic and restrictive assumptions and perceives the industry in terms of an independent store. The analysis of multi-store competition has started already with the trailblazing work of Teitz (1968) who introduced the idea that a firm can open multiple facilities in the context of Hotelling's linear city model and serves the market from a number of locations. Yet, the subject of location of competing firms with multiple component units seems to have been largely unsung in the spatial location literature (Peng and Tabuchi, 2007). This gap is surprising for the systems which dominate in the market (Karamychev and van Reeven, 2009; Iida and Matsubayashi, 2011; Janssen et al., 2005; Pal and Sarkar, 2002; Peng and Tabuchi, 2007). The conventional single-store location theory may not apply to situations wherein individual stores are part of larger organizations under common strategy, intuition, and control, and where centralization is applied to reach global goals and consider the interest of a firm as a whole (Thill, 1997). Conceptually, a firm selects a distribution of locations instead of choosing a point location (Chu and Lu, 1998).

We do not attempt for the innovative solution or an intricate model. Instead, we focus on rectifying several misconceptions made in the existing papers. The proposed approach is computational. First, we consider five asymmetric fast-food chains. Second, we incorporate interactions between stores within the same firm and stores that belong to different chains. Third, since firms search for the large sources of demand trying to attract the greatest number of potential customers, we reconsider the list of variables usually used to capture the demand level in a market. It is particularly important to carefully define market segments and to select potential customer groups by observing their characteristics, their mobility patterns and the purpose of their activities during the day and the week. We signal a clear difference between a daytime and nighttime population. A combination of these elements can secure a more realistic model of locational strategies.

1.4 Literature

1.4.1 Chapter 2 (Paper 1)

Chapter 2 introduces the reader to the location choice models. The list of the key factors that potentially influence the locational decisions has been created based on the research of Maoh (2005), Strotmann (2007), Liviano-Solís and Arauzo-Carod (2014), Rocha (2008), Maoh and Kanaroglou (2005, 2007), Bondomi and Greenbaum (2007), Bodenmann (2011), Duvereux et al. (2007), Neumark and Kolko (2010), De Bok (2004), Bodenmann and Axhausen (2012, 2010), and the review of Arauzo-Carod et al. (2010). Chapter 2 also provides a discussion of the first attempts to incorporate spatial effects into location choice models starting with Bhat and Guo (2004) on modeling spatial dependence in residential locations using a mixed Logit. Sener et al. (2011) propose the generalized spatially correlated Logit and Miyamoto et al. (2004) the mixed Logit with the error autocorrelation and an autocorrelated deterministic component of utility to model the residential behavior. Garrido and Mahmassani (2000) discuss a multinomial probit with spatially and temporally correlated error structure to analyze and forecast the distribution of freight flows.

Nguyen et al. (2012) discuss a three-stage firm relocation model wherein spatial correlation between zones has been implemented in the error term and spatial interactions among firms in the deterministic part. Klier and McMillen (2008) provide a description of the generalized method of moments spatial Logit to model the clustering of the auto supplier establishments.

1.4.2 Chapter 3 (Paper 2)

Chapter 3 continues with a discussion on further attempts to incorporate spatial effects into establishments location choice models. It highlights the fact that, whenever a distance metric has been used to implement a spatial aspect into the establishments location models (see the research of Dubé et al., 2016; Bhat et al., 2014; Liviano-Solís and Arauzo-Carod, 2013; Lambert et al., 2010; and Klier and McMillen, 2008), the Euclidean distance is employed. The debate in this chapter demonstrates that there are insights to be gained by mindfully reconsidering and measuring distance depending on a given problem, such as a high congestion, speed limits, or physical uncrossable barriers which can diminish or totally eliminate the linkage between neighboring areas. Alternative distance metrices can be proposed and tested. Several articles, including the ones by Miller (2004, 2003) and the sudies by Combes and Lafourcade (2005), Graham (2007), Duran-Fernandez and Santos (2014), Weisbrod (2008), Faber (2014), Kwon (2002), Boscoe et al. (2012), Chalasani et al. (2005), or Rietveld et al. (1999) encourage the use of realistic distance metric based on the transport network over geographical distance metrics.

1.4.3 Chapter 4 (Paper 3)

Most of the analyses on establishments location choices presented in two previous chapters treat only one selected activity sector at a time, typically industrial or retail activities. Very few empirical studies, *e.g.*, Chatman et al. (2016) or Buczkowska and Lapparent (2014) develop models for a number of various sectors, still, these models are analyzed independently, without considering strategic interactions between the sectors. Literature revised in Chapter 4, especially works of Ellickson and Misra (2011, 2012), Aguirregabiria and Mira (2007), Zhu and Singh (2009), Jia (2008), Seim (2006), Berry (1992), Bresnahan and Reiss (1991a), Draganska et al. (2008) help understand why it is worth considering within- and inter-industry interactions. An excellent review of Draganska et al. (2008) gives a classification of possible modeling choices along plural dimensions, in the informational, temporal contexts, considering the timing of moves, and the discrete, continuous, or mixed decisions of establishments.

Our context is a static discrete game of incomplete information. An establishment's payoff from choosing a particular location depends on its expectation of the optimal location choices of its competitors and exogenous market characteristics (Zhu and Singh, 2009). Based on the expected distribution of other agents across market locations, each establishment selects the location that maximizes its payoff given its own type (Seim, 2006).

1.4.4 Chapter 5 (Paper 4)

The literature of Chapter 5 starts with the work of Teitz (1968) who introduces the idea that a firm can serve the market from multiple locations in the context of Hotelling's linear city model. Since then other researchers, such as Thill (1997), Chu and Lu (1998), Pal and Sarkar

(2002), Janssen et al. (2005), Karamychev and van Reeven (2009), Iida and Matsubayashi (2011), Takaki and Matsubayashi (2013), Neven (1987), Peng and Tabuchi (2007), Granot et al. (2010), or Nishida (2015) have been proposing multi-store firms location models.

The light is then shed on the research by: Yang (2015) and Igami and Yang (2015), who study Canada's fast-food hamburger industry comprised of five substitutable chains; Toivanen and Waterson (2005), who analyze fast-food restaurants in the UK; and Thomadsen (2005), who focuses on two fast-food chains in California. For comparison, three other studies are mentioned in Chapter 5: Nishida (2015) on convenience-store chains in Japan, Schiraldi et al. (2013) and Holmes (2011) on the UK supermarket industry.

Chapter 5 pronounces the fact that store-location models typically do not consider competing rival-chain stores and the strategic interactions between stores within the same chain. The profitability is a function of player *i's* entire network and the competitors' networks (Nishida, 2015). The literature on business stealing and learning effects, cannibalization effects and economies of scale described in Chapter 5 (see, *e.g.*, Toivanen and Waterson, 2005; Aquirregabiria and Suzuki, 2015) is not consistent in specifying which of these effects has a prevailing influence on location decisions and depends on the store type. In addition, Nishida (2015) and Toivanen and Waterson (2005) incorporate some form of spatial competition, demonstrating that the store's profitability is influenced by stores in the same location and by those in the adjacent locations, incorporating the interdependence of markets.

2

Location choices of newly-created establishments: Spatial patterns at the aggregate level

Abstract This paper explores the problems associated with the location choice of newly-created establishments at the aggregate level. Although much work has been done in this domain, several issues arise when analyzing involved phenomena, which scholars have yet to fully explore: (1) addressing the excess of zeros problem in the location choice model in highly heterogeneous geographic areas and (2) determining an appropriate way to accommodate spatial effects for location decisions. We tested models that include both stocks of pre-existing establishments and variables that represent measures of accessibility to the workforce and population, proximity to shops, services, transport infrastructure, availability of land, as well as prices and tax levels. We concluded that an establishment does not act in isolation during its decision-making processes and that it is likely to be influenced by other establishments located nearby. When selecting the appropriate location in which to set up in the market, an establishment may consider not only the characteristics of a particular area, but also the characteristics of neighboring zones. Having estimated 84 nested and non-nested count data models, we found that the hurdle models are preferred for taking into account the presence of excess zeros. Hurdle models offer greater flexibility in modeling zero outcomes and relax the assumption that the zero observations and the positive observations come from the same data generating process. In addition, the paper finds that the models tested with the distance matrix indicate that the incorporation of spatial spillovers leads to an enhancement in the models' performance.

Keywords: location choice model, count data model, hurdle model, spatial spillovers

JEL Classification: C35, D21, R12

2.1 Introduction

The present paper explores the problems associated with the location choice of newly-created establishments at the aggregate level within the Paris metropolitan area in the year 2007. Applications of disaggregate discrete choice modeling to analysis of birth, death, evolution, and location of establishments in the Paris region have been carried out by de Palma et al. (2008) and Motamedi (2008). We also refer the reader to the European SustainCity Project[1]. Much work has been done in this domain. However, several issues arise when analyzing involved phenomena, which scholars have yet to fully explore, most notably the excess of zeros problem in the location choice model in highly heterogeneous geographic areas and determining an appropriate way to incorporate spatial effects into the model. For example, spatial effects can be divided into two categories: the first considers the interactions between establishments that do not act in isolation and are thus likely to be influenced by the decisions of other players in the market, while the second looks at correlations across alternative choices. Inspired by spatial econometrics techniques, *e.g.*, Anselin (1988), Lambert et al. (2010), Klier and McMillen (2008), we seek solutions to these and related problems.

The proposed approach is descriptive, aggregate. For the purposes of the present research, all data are gathered at the municipality level. In the first step, we estimate standard non-spatial nested and non-nested count data models. In the second step, we run these count data models that accommodate spatial spillovers for location decisions. We compare all the obtained results and choose the best performing hurdle-Poisson model, after which a more

[1] See http://www.sustaincity.org/publications.

thorough interpretation of its results is undertaken.

The need to specify the model becomes apparent when dealing with various types of location choice models, which range from those focusing on employment location estimated at the job level to those which focus on establishment or firm location and the associated decision-making processes. Implementing the latter form of location choice models, we acknowledge that the decision to open a new unit is made at the establishment or firm level. We define an establishment as a distinct economic unit that produces goods or services at a single physical location. In contrast, a firm is a legal entity that consists of one or more establishments under common ownership and control (van Wissen, 2000). Thus, it is a newly-created establishment that offers job opportunities to members of a given population provided they possess the appropriate skills and desired characteristics. For this reason, the unit of analysis on which we focus is that of the establishment.

The rest of the paper is organized as follows. In Section 2.2, we review the literature on both location choice models and the initial attempts to incorporate spatial effects into such models. In Section 2.3, we characterize the area of study, which is the Paris region. In Section 2.4, we detail our data along with their statistical sources and their harmonization in order to carry out the empirical application. Descriptive statistics are presented in Section 2.5. We give a brief overview of the count data models and develop our parametric statistical model in Section 2.6, starting from the Logit-Poisson model, a type of hurdle-Poisson model, and discuss how to extend it in order to meet the needs of the present research. In Section 2.7, the results from our research are discussed. In the final section, we conclude and elaborate on extensions of the proposed approach.

2.2 Review of empirical literature

2.2.1 Location models

A number of empirical studies suggest variables that play a significant role in influencing the establishment's location choice. We briefly describe those which are most relevant. According to Arauzo-Carod (2005) and Maoh (2005), an establishment will decide to enter a market if it detects a potential business opportunity and if the capital and human resources are available. An establishment must choose a type of activity that reflects both its technological and organizational level and its minimum size. With these parameters in mind, the establishment makes a decision as to where to locate. Since choosing a particular area in which to locate is critical to the establishment's success or failure (Strotmann, 2007), the main task of the establishment is to identify the places which offer the highest possible potential profit. Local economic climate in terms of its ability to attract new entrants, as well as the structure of establishments at the local level are other factors that can influence the decision (Liviano-Solís and Arauzo-Carod, 2014). Locations with lower input costs are also more likely to experience a higher number of births of new establishments (Rocha, 2008).

Sutton (2007) notices that ranking sectors according to the degree of (entry-exit) turbulence appears to offer similar results across countries, suggesting that there are some sector-specific factors at work in molding this pattern. Rocha (2008) and Liviano-Solís and Arauzo-Carod (2014) concur, finding that local employment density attracts new entrants in related sectors and has a positive impact on the establishment's productivity. However, when this density is too high, the effect becomes negative due to congestion costs, including for example, high land prices and costly commuting. This results in an inverted U-shape profile between the effect of concentration of economic activity and site attractiveness. The relationship is initially positive, but becomes negative after a certain threshold is crossed. Maoh and Kanaroglou (2005, 2007) state that agglomeration effects tend to be more significant in particular activity sectors, such as retail and services.

Local authorities try to offer different incentives in a bid to attract establishments to less developed areas and accelerate the development of the whole region. These include, but are not limited to, supply of infrastructure, designation of new building zones, and tax reductions. Empirical results from various studies undertaken in different countries worldwide demonstrate that these policies may have positive, negative, or neutral effects on the location choice and activity distribution.

Special local fiscal arrangements are common practice for attracting establishments and have received considerable attention from economists (*e.g.*, Bondonio and Greenbaum, 2007). Offers of subsidies and tax reductions are popular yet controversial tools. Low taxes are another common instrument to attract establishments. Tax reductions show positive effects on activity development. Switzerland is a prime example (Bodenmann, 2011). Devereux et al. (2007) show that in the UK the impact of subsidies is rather small on the location choice but they are more effective in dense areas when a large number of establishments is already present. Bondonio and Greenbaum (2007) show that in the US enterprise zone programs

(EZ thereafter), which hope to convince establishments to locate in less developed zones by offering them subsidies, have a significant positive impact on the creation of new establishments. However, they also show that EZ policies tend to increase business closures. Neumark and Kolko (2010) show that the overall impact of the EZ program in California is ineffective and causes no positive impact on the number of jobs.

Providing special infrastructure is an effective but costly way to raise the attractiveness of a region. De Bok (2004) shows that accessibility has a positive effect on business services. Bodenmann and Axhausen (2012) claim that this also holds for other sectors and demonstrate that interventions regarding transport infrastructure have a larger impact in the densely populated areas of Switzerland. Devereux et al. (2007) show that the effects of transport infrastructure projects are unequally distributed spatially. All actions have larger impacts in denser regions. Siebert (2000) states that increasing accessibility has a relatively small effect on regional development.

Bodenmann and Axhausen (2010) summarize how the significance of location factors has changed over time. In the 1990s, the most critical location factors were human capital, agglomeration effects, and accessibility. Currently, a positive economic climate and direct costs (earning levels, tax burdens, etc.) play larger roles. More curious readers are encouraged to acquaint themselves with the referenced literature and the work of Arauzo-Carod et al. (2010). They provide a review of over fifty papers on location choice modeling with a focus on the analyses of location decisions of new industrial establishments or firms using appropriate econometric models. They describe the establishment/firm location determinants, the econometric methods used in these investigations, and their principal results[2].

[2]The most commonly used establishment/firm location determinants and the signs of their estimates used in both discrete choice and count data models according to the review of Arauzo-Carod et al. (2010) are: agglomeration economies (+,-: positive or negative effect), previous entries in the own sector (+), existing plants (+), own-industry employment (+), sectoral diversity (+,-), sectoral specialization (+,-), market size (+), establishment/firm size (+), productivity (+), unemployment (+,-), industrial employment share (+), services employment share (+), business services (+), share of employees in R&D (+), human capital (+,-), knowledge spillovers (+), skilled workforce (+), education (-), schooling (+), existence of high schools (+), overall R&D investment (+), R&D facilities (+), high-ranking hotels (+), population density (+,-), distance to urban areas (-), land area (+,-), land costs (-), entry costs (-), taxes (-), corporate tax rate (+,-), taxes on labor (-), labor costs (+,-), wages (+,-), income per capita (+), purchasing power per inhabitant (+), GDP (+), poverty (-), local demand (+), supplier accessibility (+), government spendings (+), promotional subsidies (+), labor and capital subsidies (+), economic promotion (+), investment climate (+), infrastructure (+), transportation infrastructure (+), road infrastructure (+), distance to highway (-), rail infrastructure (-), airports facilities (+), travel time to airport (-), energy costs (+,-), and environmental regulation (-).

2.2.2 First attempts to incorporate spatial effects in location choice models

As underscored by Nguyen et al. (2012), an establishment does not act in isolation during its decision-making process and is likely to be influenced by other establishments located nearby. When choosing an appropriate place in which to set up in the market, an establishment can take into account not only the characteristics of a particular area but also those of its surroundings. The reason for doing so is the spatial dependence of neighboring areas. In addition, the degree of spatial correlations is expected to be greater among choice alternatives (the options available to the decision-maker) that are close to one another. Jayet (2001) proves the existence of interactions among units located in space and demonstrates that their intensity decreases with distance. Thus, two types of spatial effects may be considered: (1) interactions between establishments and (2) dependencies among alternatives. Spatial effects can be incorporated in location choice models when modeling the observable explanatory variables and the unobservable components. However, most often, these spatial effects are either not properly treated or are completely ignored in the analysis of the establishment's location. An example of the former can be seen in Bhat and Guo (2004) who model the spatial dependence using the mixed Logit model, with an arbitrary spatial allocation instead of basing these spatial effects on a systematic specification. Furthermore, even if spatial effects are present they are not incorporated in traditional discrete choice models. For instance, multinomial Logit models (MNL) are commonly used in location choice processes, yet they are based on several simplifying assumptions, such as independent and identical Gumbel distribution (i.i.d.) of random components of the utilities and the absence of heterogeneity and autocorrelation in the model. These simplifying assumptions limit the ability of the model to represent the true structure of the choice process (Mohammadian et al., 2005).

Traditional discrete choice modeling methods are often based on the assumption of independence among choice alternatives, which according to Sener et al. (2011) is not appropriate. Furthermore, Sener et al. (2011) claim that the estimations of the parameters of the standard Logit models are biased and inconsistent. Instead, they suggest using the generalized spatially correlated Logit (GSCL) to model the residential location choice behavior and they incorporate the correlations between the alternatives. Their model can be enhanced to accommodate random taste variations across decision-makers. The authors compare the results obtained with their GSCL model against those obtained using the standard multinomial Logit and the spatially correlated Logit (for which only correlations between the adjacent alternatives are considered). Moreover, they do not limit themselves to a few choice alternatives since GSCL models eliminate the need for any kind of simulation and can be estimated by direct maximum likelihood techniques. They criticize prior studies that were carried out with a limited number of alternatives in the choice set. Miyamoto et al. (2004) estimate a mixed Logit model with the error autocorrelation and an autocorrelated deterministic component of utility using residential location choice data for only four zones in the city of Sendai in Japan. Garrido and Mahmassani (2000) recommend using a multinomial probit (MNP) model with spatially and temporally correlated error structure to model (analyze and forecast) the distribution of freight flows over space and time at the operational or tactical levels of planning. They use Monte-Carlo simulation with a choice set of 31-41 location alternatives.

It is worth mentioning here the 2010 paper by Smirnov (2010), which develops a new spatial random utility framework where individual decision-makers are spatially dependent in their preferences. In addition, pseudo maximum likelihood estimator is consistent and computationally feasible for large datasets. The author discusses recent developments of spatial discrete choice models and classifies them by the type of spatial effect and the way it is incorporated in the model.

The literature is scarce on previous attempts to incorporate spatial effects in an establishment's or firm's (re-)location decision process. Vichiensan et al. (2005) confirm that discrete choice models are applied mainly in the field of transport choice analysis and have been continuously developed for years. However, less attention has been put into the research and development of location choice models. Location choice models differ substantially from transport choice models given that they use geo-referenced data, for which choice sets have an explicit spatial component. For this reason, it is important to understand and represent spatial correlation in location choice models. Nguyen et al. (2012) offer a three-stage firm relocation process: a firm first takes a decision on whether to stay in its actual location or to move, it then chooses the region and, finally, chooses the zone in which to relocate. The authors incorporate spatial interactions among firms in the deterministic part and spatial correlation between zones in the error term (the generalized autoregressive term is used to explain the spatial correlation between the zones). Nevertheless, certain aspects of their research paper can be further improved: for instance, the accessibility of the region used in the analysis is estimated, following Allen et al. (1993), as a linear function whereas a non-linear function could have been proposed instead. Klier and McMillen (2008) use a GMM spatial Logit model that can be applied to the large samples to account for the clustering of auto supplier establishments in the US.

Thus, in spite of recognizing the importance of incorporating spatial interactions in location choice models (and establishment location choice models in particular), research is insufficient on this topic due to lack of data at the micro level, to numerous alternatives in location choice processes, and to difficulties in defining and representing the spatial effects and in measuring spatial dimensions (Sener et al., 2011). Inspired by spatial econometrics techniques (*e.g.*, Jayet, 2001 and Anselin, 1988, 2003) and drawing from a rich database, we try to find some solutions to the above-stated problems and present our ideas in this paper. We wish to combine the methods used to produce location choice models with spatial econometric techniques by examining the role of space in in these models. This is a new and challenging field of research.

2.3 Study area

In this paper, we concentrate our research on the Paris region, also known as Île-de-France. Île-de-France is a vibrant and innovative region with over 5,6 million jobs, 37% of French executives, and 40% of national workforce in research and development[3]. Yet, the Paris region's economy is also spatially unbalanced (Combes et al., 2011). Île-de-France is divided into eight *départements* - French administrative units consisting of 1300 municipalities that cover the city of Paris and its suburbs[4]. While the Paris region represents only 2.2% of the surface of France, over 19% of the country's population reside in this area (11.7 million)[5]. The GDP of the region amounts to 29% of total French GDP (IAU IdF, 2014). Large differences in population and employment densities are to be found between Paris and its outer periphery. As illustrated during the 2011 OECD Meeting, many issues continue to stand in the way of improving the situation in the suburbs (de Palma, 2011), such as poor access to public services and infrastructure.

[3]See http://www.iau-idf.fr/lile-de-france/un-portrait-par-les-chiffres/lile-de-france-capitale-de-rang-mondial.html.

[4]The Paris region consists of the Paris City, the inner ring with *départements* 92, 93, 94, and the outer ring with *départements* 77, 78, 91, and 95.

[5]See http://www.iau-idf.fr/lile-de-france/un-portrait-par-les-chiffres/population.html. Data are for the year 2009.

The Grand Paris is a development project for the whole of the Paris metropolitan area. It is designed to improve residents' quality of life, address regional inequalities, and build a sustainable city[6]. This project aims to link major territorial development contracts raising the attractiveness of the entire Paris metropolitan region. Estimates are that it will attract 1.5 million people, create or relocate 1 million jobs by 2030 and cost around 32.4 billion euros. Urban planners are likely to connect economically promising areas; yet, they could also use transport infrastructure to link the physically deteriorated, poorly planned areas in a bid to stimulate their development.

2.4 Data

Many different data sources were compiled for the present study, drawn primarily from the Census survey of establishments for the years 2003-2011[7]. Changes in nomenclature and definitions, as well as availability of data, meant that the year 2007 was used to run the models and generate the descriptive statistics. Data are pooled across activity sectors and additional data are drawn from other sources to build the final sample we used for estimation.

Local population and workforce are described by various characteristics, such as socio-professional class and education level. These data were gathered by the General Census and provided by the French National Institute of Statistics and Economic Studies (INSEE) for the years 2006-2009. French National Census of Population (RGP) provides us with information on the total number of trips made between work and home for the years 2006-2009. The Public Finances General Directorate provided information on income imputation and tax levels for 1998-2008. We also collected data on the rate of business and residence taxes for the years 2002-2010. Data on business taxes were only available for about half of the analyzed municipalities and the missing 563 out of 1300 values could not be implemented. For this reason, information on business taxes was not used to run the models. On the other hand, data on residence taxes paid per municipality were available for all the municipalities.

The Callon database enabled us to gather detailed information on the average price and rent levels per square meter for new, renovated, and pre-existing non-renovated offices, for first class shops (the most exclusive shops), for second and third class shops (the least exclusive

[6]See https://www.societedugrandparis.fr.

[7]The Census survey was carried out yearly until the year 2011 by the French National Institute of Statistics and and Economic Studies.

shops), as well as for individual industrial units and warehouses. Data on real estate prices were collected for the years 2000-2009. The hedonic pricing method was used to impute the missing values for some municipalities.

From the Sitadel2 database, managed by the Ministry of Ecology, Sustainable Development and Energy, we drew information on housing construction (net floor area and number of units for housing and professional offices) for the years 2002-2011. The numerical land use database (Corine Land Cover), also managed by this Ministry, was used to gather information on urban and agricultural zones for 2000 and 2006. The MOS database of the Urbanism and Development Institute enabled us to obtain information on the portion of the municipality areas dedicated to shops, services, industrial activities, universities and schools, hospitals and clinics, and various other amenities, for example, parkings and parks. The same database gave us the percentage of municipalities' surface covered by residential areas or vacant land. In addition, the BPE equipment database of DSDS/INSEE provided us with information on the location of individual services, education, health, social actions, transport services, sport and leisure zones, etc. Finally, information on the public transport accessibility level and the proximity to highways for 2005-2009 was obtained through the Regional and Interdepartmental Directorate of Public Works and Planning.

Table 2.1: Description of potential explanatory variables.

Variable and its expected sign	Description
Establishments from respective sector (+)[a]	Number of pre-existing establishments from the analyzed sector within a particular municipality divided by the surface of municipality (km^2)[b]
Large establishments from all sectors (+,-)	Number of large pre-existing establishments with fifty employees or more divided by the surface of municipality (km^2)
Employment density (+)	Size of active employment (x1000) divided by the surface of municipality (km^2)
White-collar employees (+)	Number of white-collar workers divided by the size of labor force
Blue-collar employees (+)	Number of blue-collar workers divided by the size of labor force
Trips home-work (nl)	Number of trips between home and work if municipality is both a place of residence and a workplace to the total number of trips home-work
Population density (+,-)	Size of population (x1000) divided by the size of residential area (km^2)
Income per person (+)	Average income level per capita (euros)
Offices (+)	Fraction of a municipality's surface dedicated to offices
Shops (+)	Fraction of a municipality's surface dedicated to shops
Parkings (nl)	Fraction of a municipality's surface dedicated to parkings
Universities and schools (+)	Fraction of a municipality's surface dedicated to universities and schools
Hospitals and clinics (nl)	Fraction of a municipality's surface dedicated to hospitals and clinics
Vacant land (+)	Fraction of a municipality's vacant land available for new investments
Distance to highway (-)	Distance to the nearest highway (km)
Public transport (+)	Number of subway, train stations, and bus stops in a municipality
Residence tax (-)	Average level of residence taxes
Price of offices (-)	Average price level of offices per square meter (euros)
Price of shops (-)	Average price level of shops per square meter (euros)

[a] (+) and (-) mean that the associated coefficient is expected to be positively or negatively statistically significant, respectively. (nl) means that no literature treats this problem or that no literature was reviewed on this issue.

[b] The data on stock of establishments are given for the 1st of January 2007. The range for the independent variables is 2005-2009.

Based on the reviewed literature, the paper of Arauzo-Carod et al. (2010) and the DREIF report of de Palma et al. (2008) in particular, the available data, and our prior knowledge, we select the variables that can be used in the models to explain the location choice of newly-created establishments. We run the models with variables that are expected to represent the structures of the population and employment area, the measure of accessibility to specific populations and employment characterized by certain relevant skills. We test whether the location choices of newly-created establishments are influenced by the proximity of retail, services, universities and schools, and other amenities. In addition, the proportion of a municipality's vacant land that can be available for new investments may play a role as well as the proximity to residential areas. The significance of price of shops and offices on the location choice of establishments is tested. We also analyze, whether an easy access to a particular municipality by either public or private transport influences the location choices, and for which sectors these accessibility measures are most relevant.

When modeling the location choice of establishments, care should be taken to select appropriate set of variables depending on the activity sector being analyzed. The expected signs of the estimated parameters may differ accordingly. As seen in the review by Arauzo-Carod et al. (2010) and the literature review presented in subsection 2.2.1, the focus has been, up to this point, on the manufacturing/industrial and high-tech or R&D sectors. We here consider several other sectors, such as construction, hotels and restaurants, and real estate activities. Thus, the existing literature can be treated as a guide to motivate factors that may be studied in the econometric models. One should be cautious, however, interpreting the models' results, as, for example, the signs of the obtained estimates depend on the sector analyzed.

The data used to estimate the models are gathered at the municipality level. We also used these raw data to compute additional variables for our application. When constructing the database, we paid special attention to the problem of definition changes, differences within the available periods and appropriate aggregation, interpolation, and extrapolation techniques. Selected variables that have been tested in the models along with their estimates expected signs are summarized in Table 2.1.

2.5 Descriptive statistics

To carry out the present study, we accessed the Census data of establishments at the aggregate level for the years 2003-2011. Due to changes in nomenclature, activities in 2003, 2005, 2007, 2009, 2010, and 2011 are based on the NAF nomenclature (the French classification of activities) while activities in 2006 and 2008 are based on the NES nomenclature (the summary economic classification). There is no one-to-one correspondence that would allow us for the comparison of all sectors or sub-sectors between these two classifications. Definitions of some sectors used in NES and NAF nomenclatures change across time. In addition, on August 4th 2008, the *Loi de Modernisation de l'Economie* (LME)[8] was introduced in an attempt to stimulate growth and employment creation. It gave small and medium enterprises the best opportunities to grow by reducing their payments, boosting competition (*e.g.*, simplifying the installation of supermarkets) and enhancing the attractiveness of the territory. On January 1st 2009, the definition of the auto-entrepreneur status was also modified[9]. Taking into account all these modifications and data accessibility, we built the models for the year 2007. Thus, 2007 is the year in which the dependent variable is measured.

Data are pooled across eleven main activity sectors: (1) industry (Industr), (2) construction (Constr), (3) commerce (Commer), (4) transport (Transp), (5) financial and insurance activities (Finan), (6) real estate activities (RealEst), (7) hotels and restaurants (HoteRes), (8) information and communication (InfoCom), (9) special, scientific and technical activities (SpecSci), (10) education (Educ), (11) health and social actions (HeaSoc), and the additional category "Others"[10].

---

[8] More information on LME can be found on: http://www.toute-la-franchise.com/vie-de-la-franchise-A3942-commerces-de-centre-ville-la-muta.html.

[9] It applies to the natural persons who set up or already possess a sole proprietorship, for the purpose of exercising a commercial or artisanal activity or one of the professions (with the exception of certain activities), as a main or complementary activity, and whose sole proprietorship fulfills the conditions of the micro-enterprise fiscal category, and who opt for VAT exemption. Source: http://www.insee.fr/en/methodes/default.asp?page=nomenclatures/liste-nomenclatures.htm.

[10] "Others" includes: administration services; public administration; arts, spectacles and recreation activities; other service activities: activities of membership organizations, reparation of computers and personal and household goods, and other personal services; activities of households as employers of domestic staff and activities of households as producers of goods and services for their own use; and extra-territorial activities.

2.5.1 Establishments' structure across activity sectors

In 2007, the largest number of newly-created establishments (over 22%) belonged to the commerce sector. The percentage of pre-existing commercial establishments among all establishments (also 22%) had been decreasing across time, when compared to 2003 and 2005. The second biggest sector was the special, scientific and technical activities sector, with 17.4% of newly-created units and 16.2% of pre-existing establishments. The percentage of establishments in this sector increased across time. The third biggest sector was construction, with 13.8% of newly-created units and a much lower 9.2% of pre-existing establishments. We also observed the smallest sectors: education (2% for newly-created establishments and 1.6% for pre-existing establishments) and transport (3.5% for newly-created establishments and 3.9% for pre-existing establishments).

2.5.2 Ranking of municipalities

Next, we analyze the numbers of newly-created establishments across all 1300 municipalities. We compute the market share of each municipality as the number of newly-created establishments in a given municipality from a given sector s over the total number of newly-created establishments from this sector s. This allows us to build the rankings of municipalities based on their share in the market. This method enables us to see which municipalities and *départements* are the most frequently chosen for a new location. From the rankings of municipalities, we select the municipality (one from each activity sector) which attracts the highest number of newly-created establishments. We also determine the number of new units that locate in the top 10, 20, 50, and 100 municipalities. This offers some insights on the concentration of activities in the Paris region (see Table 2.2). For instance, the most frequently chosen municipality, namely the 8th district of Paris, becomes home to over 14% of newly-created units in the real estate sector and it also accounts for 13.4% of all newly-created financial establishments. Approximately 9.3% of all new establishments in the special, scientific and technical activities sector are also located in just one of the 1300 potentially available municipalities (the 8th district of Paris). On the other hand, we observe relatively low fractions of transport and commerce newly-created establishments in the most frequently chosen municipality: only 2.3% and 2.7%, respectively. As a means of contrast, the average percentage of establishments that locate in the top municipality is 5.1%. The first 20 municipalities account for around half

Table 2.2: Market share of top 1, 10, 20, 50, 100 municipalities in 2007.

#Municip. [a]	Industr[b] % (#Est[c])	Constr %	Commer %	Transp %	Finan %	RealEst %
1	4.4 % (146)	4.2% (505)	2.7% (531)	2.3% (70)	13.4% (595)	14.4% (672)
10	25.6% (842)	21.8% (2644)	22.2% (4360)	17.6% (541)	39.9% (1772)	40.4% (1892)
20	37.9% (1249)	33.1% (4004)	34.7% (6815)	29.3% (900)	52.3% (2327)	52.9% (2478)
50	56.1% (1849)	51.7% (6266)	53.6% (10541)	52.1% (1600)	66.6% (2963)	68.2% (3195)
100	69.5% (2291)	67.8% (8211)	68.8% (13533)	70.8% (2175)	77.7% (3455)	78.0% (3654)
All municip.	3296	12115	19658	3072	4446	4683
Sector size	3.8%	13.8%	22.3%	3.5%	5.1%	5.3%

#Municip.	HoteRes %	InfoCom %	SpecSci %	Educ %	HeaSoc %	All %
1	3.6% (125)	5.4% (349)	9.3% (1427)	3.8% (66)	3.7% (149)	5.1% (4525)
10	25.4% (896)	31.7% (2048)	36.6% (5587)	26.9% (466)	24.7% (1008)	26.1% (23004)
20	38.6% (1359)	47.6% (3076)	52.4% (7984)	40.5% (702)	36.4% (1484)	38.6% (33996)
50	58.1% (2046)	67.4% (4349)	69.2% (10582)	58.3% (1011)	53.6% (2185)	56.9% (50026)
100	73.2% (2581)	79.3% (5120)	79.2% (12103)	72.2% (1251)	68.8% (2804)	70.2% (61780)
All municip.	3524	6457	15282	1733	4077	87974
Sector size	4.0%	7.3%	17.4%	1.97%	4.6%	100%

[a] Number of top 1, 10, 20, 50, 100 municipalities which gather the highest number of newly-created establishments from a particular activity sector in 2007.

[b] Market share of top municipalities. Number of newly-created establishments from a particular activity sector within these top municipalities given in the bracket.

[c] #Est: Total number of newly-created establishments in each sector in a particular municipality.

of all newly-created entities in the real estate, financial, scientific, special and technical activities, and information and communication sectors.

In addition, the top municipality in the case of the financial activities sector, namely the 8th district of Paris, gathers 2.6 times more newly-created establishments than the second top municipality, the 16th district of Paris. Similarly, the same top municipality (the 8th district of Paris) attracts 2.3 times more newly-created real estate units than the ”second best” municipality, the 16th district of Paris. A large difference in the number of newly-created establishments between the two first municipalities from the ranking lists can also be observed in the construction sector. In this sector, establishments tend to locate 1.7 more often in the 20th district of Paris than in the ”second best” municipality, the 10th district of Paris. The top municipalities (based on market share) are the ones that belong to Paris and to *départements* 92 and 93. However, some differences are observed across sectors. We summarize the results on the maps in Fig. 2.1.

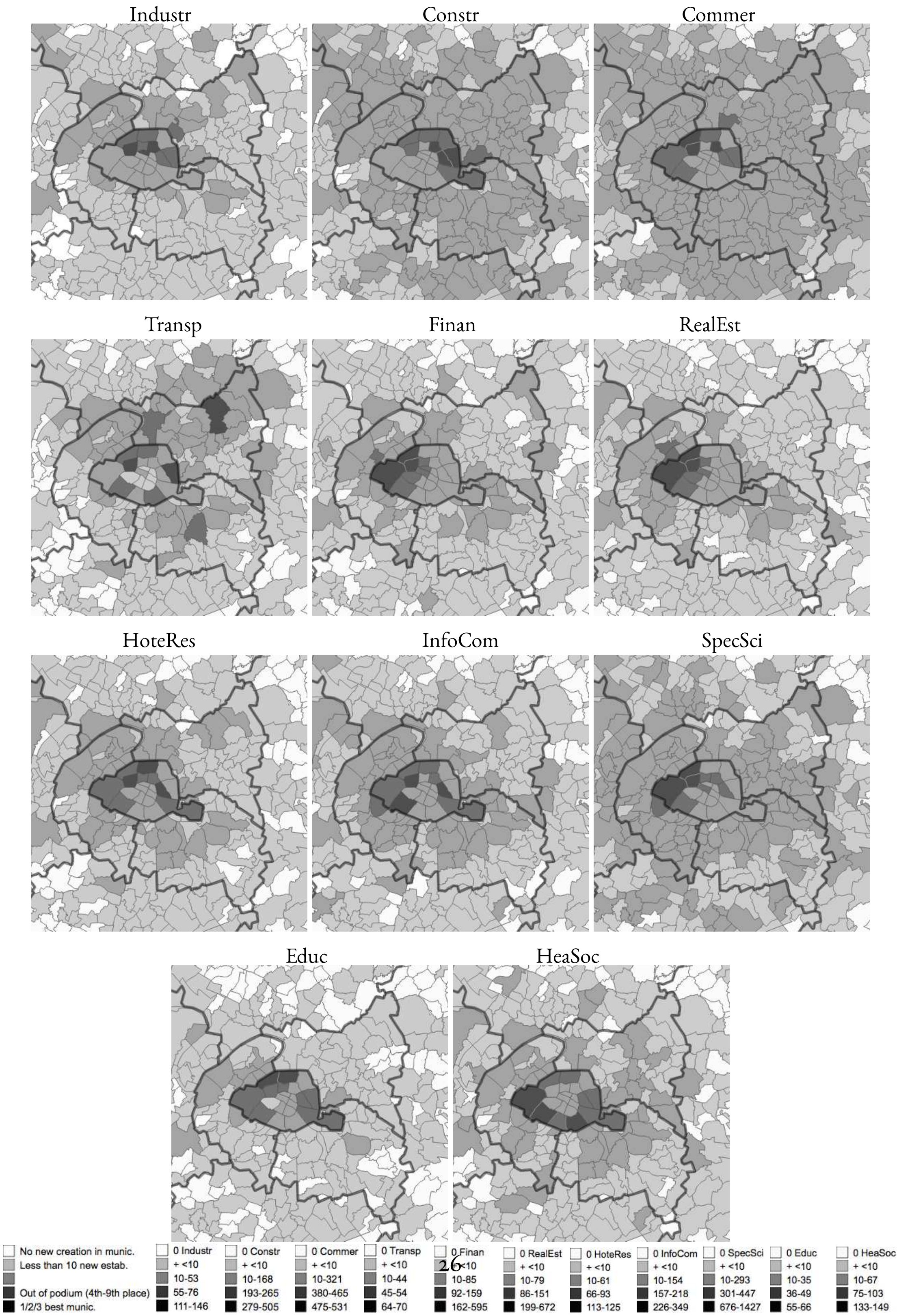


Figure 2.1: Newly-created establishments in 2007 - rankings based on the market share - focus on Paris and the inner ring.

Table 2.3: Ranking of top 10 municipalities based on the market share (#newly-created establishments to total #establishments in municipality)

	Industr Insee-Munic[a] (#Est[b])	Constr Insee-Munic (#Est)	Commer Insee-Munic (#Est)	Transp Insee-Munic (#Est)	Finan Insee-Munic (#Est)	RealEst Insee-Munic (#Est)
1	Paris2 (146)	Paris20 (505)	Paris8 (531)	Paris8 (70)	Paris8 (595)	Paris8 (672)
2	Paris8 (140)	Paris10 (302)	Paris17 (477)	Paris20 (68)	Paris16 (233)	Paris16 (293)
3	Paris10 (111)	Paris11 (279)	Paris10 (475)	93005-Aulnay (64)	Paris17 (162)	Paris17 (199)
4	Paris11 (76)	Paris18 (265)	Paris18 (465)	Paris18 (54)	Paris9 (159)	Paris9 (151)
5	Paris18 (73)	Paris19 (251)	Paris20 (455)	Paris15 (52)	Paris2 (128)	Paris15 (121)
6	Paris9 (66)	93048-Montr. (232)	Paris11 (442)	93066-SaintD. (50)	Paris15 (125)	92051-Neuilly (108)
7	Paris17 (61)	Paris17 (225)	Paris16 (416)	75117-Paris17 (47)	Paris1 (97)	92026-Courbe. (95)
8	93055-Pantin (61)	Paris8 (224)	93001-Auber. (398)	94028-Créteil (47)	Paris7 (96)	Paris7 (88)
9	Paris20 (55)	Paris12 (193)	Paris15 (380)	Paris13 (45)	92062-Puteaux (92)	Paris1 (86)
10	92026-Courbe. (53)	93001-Auber. (168)	Paris2 (321)	Paris19 (44)	92051-Neuilly (85)	92012-Boulog. (79)
Total[c]	3296	12115	19658	3072	4446	4683

	HoteRes Insee-Munic (#Est)	InfoCom Insee-Munic (#Est)	SpecSci Insee-Munic (#Est)	Educ Insee-Munic (#Est)	HeaSoc Insee-Munic (#Est)	All sectors Insee-Munic (Total)
1	Paris18 (125)	Paris8 (349)	Paris8 (1427)	Paris17 (66)	Paris16 (149)	Paris8 (4525)
2	Paris8 (118)	Paris11 (229)	Paris16 (960)	Paris18 (55)	Paris15 (146)	Paris16 (2721)
3	Paris11 (113)	Paris15 (226)	Paris17 (676)	Paris9 (49)	Paris13 (133)	Paris17 (2619)
4	Paris9 (93)	Paris17 (218)	Paris9 (447)	Paris20 (49)	Paris17 (103)	Paris11 (2055)
5	Paris10 (87)	Paris16 (195)	Paris15 (447)	Paris8 (48)	Paris18 (89)	Paris15 (2030)
6	Paris17 (87)	Paris10 (177)	Paris11 (414)	Paris16 (47)	Paris11 (85)	Paris20 (1988)
7	Paris15 (78)	92012-Boulog. (174)	92012-Boulog. (317)	Paris11 (44)	Paris14 (84)	Paris18 (1930)
8	Paris12 (68)	Paris18 (169)	Paris1 (305)	Paris15 (37)	Paris12 (77)	Paris10 (1891)
9	Paris16 (66)	Paris12 (157)	Paris10 (301)	Paris12 (36)	Paris20 (75)	Paris9 (1817)
10	Paris19 (61)	Paris9 (154)	Paris18 (293)	Paris14 (35)	Paris19 (67)	Paris12 (1428)
Total	3524	6457	15282	1733	4077	87974

[a] Insee-Munic: INSEE code and the municipality name (note that Courbe. stands for Courbevoie, Montr. = Montreuil, Auber. = Aubervilliers, Aulnay = Aulnay-sous-Bois, SaintD. = Saint-Danis, Neuilly = Neuilly-sur-Seine, Boulog. = Boulogne-Billancourt).

[b] #Est: Total number of newly-created establishments in each sector in a particular municipality.

[c] Total: Total number of newly-created establishments in the Paris region in a particular sector.

2.5.3 Composition of establishments in each *département*

Our next step is to analyze the composition of establishments in each *département* by looking at the ratio of establishments in sector s (where s=1, 2, ..., 11) in *département* j compared to the total number of establishments in this *département* j (where j=1, 2, ..., 8). We compute the number of establishments that belong to a particular sector in a *département* for both the newly-created establishments and the establishments in stock. Since the purpose of this paper is to analyze newly-created establishments, we describe in detail below how the structure of newly-created establishments differs between Paris, the *départements* belonging to the inner ring and the *départements* pertaining to the outer suburbs. Corresponding statistics for pre-existing establishments can be found in Table 2.4.

Approximately 37% (32 767 units) of all newly-created establishments locate in Paris, 13% (11 317 units) locate in dep 92 and 11% (9 825 units) locate in dep 93. *Départements* from the outer ring seem to be the least attractive for new establishments.

We also observe the following: (1) The lowest ratio of establishments in the construction sector (and equivalently the commerce sector) compared to the total number of establishments are registered in dep 92 and Paris. On the other hand, the highest percentage of establishments that belong to these two sectors are observed in dep 93. (2) The lowest ratio of establishments in the transport sector to the total number of establishments is registered in Paris and the highest ratio is registered in dep 93. (3) Establishments in the financial, real estate, and special, scientific and technical activities sectors gather the most frequently in Paris and dep 92. For each of these sectors, ratios of roughly a third of the size are registered in dep 93. (4) The highest fraction of establishments in the information and communication sector is located in dep 92 and in Paris, the lowest fraction in dep 77 and dep 93. (5) Similar percentages of establishments that belong to the industry, hotels and restaurants, health and social actions sectors are observed across almost all *départements*. (6) In the case of the education sector, the ratio of establishments within this sector to the total number of establishments is very low in each *département* and is particularly low in dep 93.

2.5.4 Size of newly-created establishments versus size of pre-existing units

We concentrate now on the size of newly-created units compared to the size of pre-existing entities (see Table 2.5). Around 60% of all pre-existing establishments are units with zero

Table 2.4: Structure of newly-created establishments and existing establishments in each *département* in 2007; Comparison with 2003 and 2005

Newly-created establishments	Dep[a]	Industr[b]	Constr	Commer	Transp	Finan	RealEst	HoteRes	InfoCom	SpecSci	Educ	HeaSoc
Paris	75	3.34%	9.49%	19.62%	1.73%	6.34%	7.01%	4.04%	8.89%	23.42%	2.03%	4.22%
Inner ring (dep 92, 93, 94)[c]	92	3.68%	9.00%	17.28%	2.92%	6.42%	6.49%	4.01%	10.14%	22.11%	2.27%	4.86%
	93	4.22%	23.47%	29.51%	7.03%	2.54%	2.04%	4.35%	4.97%	7.19%	1.06%	3.36%
	94	3.21%	15.94%	25.95%	5.41%	3.83%	3.94%	3.65%	6.66%	12.45%	1.81%	5.01%
Outer ring (dep 77, 78, 91, 95)[d]	77	5.01%	17.34%	25.30%	4.13%	4.14%	4.47%	4.50%	4.31%	10.06%	2.15%	5.42%
	78	3.75%	13.41%	21.98%	2.77%	4.60%	5.26%	3.84%	6.25%	17.99%	2.77%	6.11%
	91	3.79%	16.63%	24.33%	4.02%	4.29%	4.37%	3.64%	5.40%	12.99%	1.82%	5.56%
	95	4.50%	20.84%	25.47%	5.47%	3.55%	3.15%	3.73%	5.10%	10.43%	1.75%	4.44%
	Total 2007	3.75%	13.77%	22.35%	3.49%	5.05%	5.32%	4.01%	7.34%	17.37%	1.97%	4.63%
	Total 2005	3.59%	12.76%	23.51%	2.85%	4.58%	6.05%	4.39%	7.35%	18.13%	1.85%	4.36%
	Total 2003	4.39%	12.82%	25.17%	3.60%	4.22%	4.55%	4.95%	7.35%	17.57%	1.47%	3.83%
Establishments in stock	Dep	Industr	Constr	Commer	Transp	Finan	RealEst	HoteRes	InfoCom	SpecSci	Educ	HeaSoc
Paris	75	5.05%	6.02%	20.51%	1.98%	5.82%	6.66%	6.62%	7.63%	21.62%	1.57%	6.29%
Inner ring	92	5.40%	6.98%	18.51%	3.47%	6.56%	5.95%	5.83%	8.11%	18.35%	1.89%	8.42%
	93	7.15%	14.44%	28.64%	7.62%	3.12%	3.95%	6.65%	3.89%	7.45%	1.05%	7.03%
	94	5.92%	11.49%	24.01%	5.77%	4.02%	4.94%	6.02%	4.62%	11.70%	1.57%	9.39%
Outer ring	77	8.01%	13.64%	23.91%	6.01%	4.12%	4.86%	5.77%	2.83%	8.71%	1.68%	9.45%
	78	6.04%	10.34%	21.67%	3.89%	4.79%	5.16%	5.36%	4.71%	14.54%	2.17%	10.60%
	91	6.48%	12.95%	23.46%	5.13%	4.27%	4.65%	5.45%	4.11%	10.89%	1.61%	10.11%
	95	7.07%	13.55%	24.45%	6.92%	3.74%	4.01%	5.56%	3.29%	9.49%	1.61%	9.62%
	Total 2007	5.86%	9.16%	22.05%	3.93%	5.09%	5.64%	6.16%	6.03%	16.18%	1.63%	7.92%
	Total 2005	6.37%	8.91%	22.92%	4.22%	5.07%	4.31%	6.32%	5.88%	15.85%	1.58%	8.16%
	Total 2003	6.92%	8.69%	23.19%	4.26%	5.19%	4.09%	6.37%	5.66%	15.44%	1.53%	8.21%

[a] Dep: *Département* number.

[b] Sector names: industry; construction; commerce; transport; financial and insurance activities; real estate activities; hotels and restaurants; information and communication; special, scientific and technical activities; education; health and social actions.

[c] The inner ring of the Paris region consists of dep 92, dep 93, and dep 94.

[d] The outer ring of the Paris region consists of dep 77, dep 78, dep 91, and dep 95.

employees. An establishment with zero employees is an operating establishment with no employees other than the owner. Establishments with 0-9 employees (so called micro establishments) account for 93% of all establishments. In comparison, these figures stand respectively at 82% and 97% for newly-created establishments. Next, 7% of pre-existing units are registered as establishments with 10 or more employees, compared with only 2.8% of newly-created units from the same size category. Interestingly, 1.4% of existing units have at least 50 employees, whereas it is 7 times less frequent for newly-created establishment to be classified to this category. Thus, a newly-created establishment will most probably be one with zero employees and only very rarely it will have more than 10 employees. Stable units grow in size and hire more workers.

Large differences in the typical size of newly-created units can be observed across sectors. 96% of all the newly-created establishments in health and social actions are the ones with 0 employees (92% for education and special, scientific and technical activities), in comparison to only 57% for hotels and restaurants. Hotels and restaurants in 39% of cases, are units with 1-9 employees and this is the highest fraction across the sectors. Not surprisingly, the lowest number of new establishments from the size category with 1-9 employees, is registered in health and social actions (only 3%) and special, scientific and technical activities (7%). Relatively high fractions of bigger establishments with more than 10 employees are registered in construction and industry, 8.7% and 6.9%, respectively, comparing to rather low levels for other sectors. The case of the new creation of very large units (with 50 employees or more) is in general not common (from nearly zero percent for construction to 0.7% for industry), and will the most probably happen for establishments in industry, financial activities, and the transport sector.

Table 2.5: Structure of newly-created and pre-existing establishments by size category.

New	#Estab 0[a] (%)	#Estab 1-9[b] (%)	#Estab 10+[c] (%)	#Estab 50+[d] (%)	Total
Industr	2442 (74.1%)	628 (19.1 %)	226 (6.9%)	23 (0.70 %)	3296
Constr	8554 (70.6%)	2505 (20.7%)	1056 (8.7%)	3 (0.02 %)	12115
Commer	15472 (78.7 %)	3810 (19.4%)	376 (1.9%)	29 (0.15 %)	19658
Transp	2496 (81.3%)	491 (16.0 %)	85 (2.8%)	13 (0.42 %)	3072
Finan	3855 (86.7%)	503 (11.3 %)	88 (2.0%)	26 (0.58 %)	4446
RealEst	4049 (86.5%)	599 (12.8 %)	35 (0.7%)	7 (0.15 %)	4683
HoteRes	2013 (57.1%)	1384 (39.3%)	127 (3.6%)	8 (0.23 %)	3524
InfoCom	5812 (90.0%)	559 (8.7%)	86 (1.3%)	11 (0.17 %)	6457
SpecSci	14099 (92.3 %)	1040 (6.8%)	143 (0.9%)	23 (0.15 %)	15282
Educ	1589 (91.7%)	133 (7.7%)	11 (0.6%)	1 (0.06 %)	1733
HeaSoc	3916 (96.1%)	132 (3.2%)	29 (0.7%)	6 (0.15 %)	4077
Other	8028 (83.4%)	1380 (14.3%)	223 (2.3%)	31 (0.32 %)	9631
Total	72325 (82.2 %)	13164 (15.0 %)	2485 (2.8 %)	181 (0.21 %)	87974

Stock	#Estab 0 (%)	#Estab 1-9 (%)	#Estab 10+ (%)	#Estab 50+ (%)	Total
Industr	20872 (46.6 %)	16800 (37.5 %)	7072 (15.8%)	1694 (3.79%)	44744
Constr	36390 (52.1 %)	28622 (40.9 %)	4896 (7.0%)	589 (0.84%)	69908
Comm	92715 (55.1 %)	63047 (37.5 %)	12473 (7.4%)	1930 (1.15%)	168235
Transp	19624 (65.4 %)	6213 (20.7 %)	4183 (13.9%)	1071 (3.57%)	30020
Finan	23853 (61.4 %)	11404 (29.4 %)	3590 (9.2%)	892 (2.30%)	38847
RealEst	31325 (72.8 %)	10567 (24.6 %)	1139 (2.6%)	195 (0.45 %)	43031
HoteRes	16837 (35.8 %)	24621 (52.3 %)	5577 (11.9%)	521 (1.11 %)	47035
InfoCom	30536 (66.4 %)	11121 (24.2 %)	4346 (9.4%)	1139 (2.48%)	46003
SpecSci	88778 (71.9 %)	28419 (23.0 %)	6268 (5.1%)	1189 (0.96%)	123465
Educ	9079 (73.2%)	2583 (20.8%)	749 (6.0%)	94 (0.76 %)	12411
HeaSoc	50789 (84.1 %)	8365 (13.8%)	1266 (2.1%)	371 (0.61 %)	60420
Other	47052 (59.6 %)	26490 (33.5 %)	5470 (6.9%)	1202 (1.52%)	79012
Total	467850 (61.3 %)	238252 (31.2 %)	57029 (7.5%)	10887 (1.43 %)	763131

[a] #Estab 0: Number of establishments with 0 employees.
[b] #Estab 1-9: Number of establishments with 1-9 employees.
[c] #Estab 10+: Number of establishments with 10 employees or more.
[d] #Estab 50+: Number of establishments with 50 employees or more.

Large discrepancies in the typical size of newly-created units can be observed across sectors. 96% of all newly-created establishments in the health and social actions sector and 92% in the education and special, scientific and technical activities sectors have zero employees. These figures can be compared to the 57% of newly-created hotels and restaurants with zero employees. Relatively high figures of larger establishments with more than 10 employees are registered in the construction and industry sectors (8.7% and 6.9%, respectively) in comparison to rather low levels for other sectors. New creation of very large units (with 50 employees or more) is in general not common (from nearly zero percent for construction to 0.7% for industry) and is most likely to occur within the industry, financial activities, and transport sectors.

2.6 Parametric models

2.6.1 Count data models

In an economic context, the dependent variable is often an integer or count with the distribution that places probability mass at nonnegative integer values only (unlike a regression model). The Poisson regression is a very popular model and usually forms the starting point for count data analysis. Such a model captures the discrete and nonnegative nature of the data; it does, however, assume equidispersion (equality of the conditional variance and the conditional mean). This assumption is found to be too restrictive given that the conditional variance often exceeds the conditional mean. Large overdispersion leads to deflated standard errors and inflated t-statistics in the maximum likelihood output. Thus, a statistical test of the overdispersion should be performed after running the Poisson model. Overdispersion can be modeled either by introducing an additional parameter as in the negative binomial (NB) model (Consul and Jain, 1973), by allowing for an extra proportion of zeros using zero inflated models or by combining zero inflated models with overdisperse distribution.

In addition, when the observed data display a higher fraction of zeros than would be typically explained by the standard count data models, two types of models can be suggested: the hurdle model (Mullahy, 1986) or the zero-inflated model (Lambert, 1992). Where the Poisson model is used, the density function predicts the probability of zeros to be less than that is actually observed.

The hurdle model, also called the two-part model, reflects a two-part decision making process. It relaxes the assumption that the zero observations and the positive observations come from the same data generating process. Maximum likelihood estimation of the hurdle model involves separate maximization of the two terms in the likelihood, one related to the zeros and the other to the positive values. Typically, the distribution of the second part has a zero-truncated form of a standard discrete distribution such as Poisson or NB distribution, although other distributions defined directly on the positive integers (*e.g.*, the logarithmic) can be also considered. The two-stage decision-making process is reflected through the hurdle model interpretation. A zero-inflated model assumes that there are two sources of zero counts (not just one as in the hurdle model). For an extensive description of count data models and their comparison and applications, see, *e.g.*, Winkelmann (2003) or Cameron and Trivedi (1986).

In the current paper, we run several models, namely: Poisson, zero-inflated Poisson (ZIP), zero-inflated tau Poisson (ZIP τ[11]), negative binomial (NB), zero-inflated negative binomial (ZINB), and hurdle-Poisson models. All these models are run separately for selected activity sectors[12]: (1) not incorporating spatial effects and (2) using distance matrix. The results are then compared and the best performing hurdle model with incorporated spatial spillovers is described in the next section.

We agree that we may have used other modeling techniques by transforming the dependent variable into a rate instead of a count, using, for instance, an aggregate Logit model. We, thus, would have had to control for the excess of zeros problem (*i.e.*, rate equals to 0) by means of an additional sample selection equation (like a Tobit model applied to rates). We have no prior about which approach fits best the data. We leave this aside for future research. Yet, we have chosen to use an approach where it is easier to control for the excess zeros, as presented in the next section.

2.6.2 Hurdle-Poisson model

We statistically describe the location of newly-created establishments that belong to a given economic sector $\mathtt{s}$ in zone $\mathtt{l}$ at the aggregate level. Let $\mathtt{y}_{\mathtt{l},\mathtt{s}}$ models the count of newly-created establishments from sector $\mathtt{s}$ that locate at $\mathtt{l}$. By construction, $\mathtt{y}_{\mathtt{l},\mathtt{s}}$ takes nonnegative integer values. The rationale behind the choice of a hurdle-Poisson model is that observing a zero

[11] τ is a scalar parameter.

[12] These seven selected sectors are: industry, construction, commerce, financial activities, real estate activities, hotels and restaurants, and special, scientific and technical activities.

outcome means that it is not possible to capture additional net profit at a given location l for sector s provided set establishing conditions at the beginning of the year: there is no reason to locate where there is no profit to make. New establishments will locate in a zone where there remains profit to be captured. We agree that this is a debatable issue, as we do not account for behavioral dynamics (forward-looking behaviors) of establishments. However, we prefer to stick to the following: no current additional profit, no new establishment. We also understand that already installed establishments may increase in size and/or in production to capture any additional profit.

We describe observations of these counts by independent variables $x_{l,s}$. As our empirical approach is a two-part model, these variables may be used in either of the parts. They may also differ along with s. Our approach is parametric: let β_s (hurdle part) and α_s (conditional count part) be the weights of the independent variables.

The contribution of one observation to the sample likelihood function is written as (Cameron and Trivedi, 1986):

$$\begin{aligned}&\ell\left(\alpha_s, \beta_s | y_{l,s}; x_{l,s}\right) = \\ &\Pr\left(y_{l,s} = 0 | x_{l,s}; \beta_s\right)^{\mathbb{I}\left(y_{l,s}=0\right)} \times \\ &\left(\left(1 - \Pr\left(y_{l,s} = 0 | x_{l,s}; \beta_s\right)\right) g\left(y_{l,s} | y_{l,s} > 0; x_{l,s}; \alpha_s\right)\right)^{1-\mathbb{I}\left(y_{l,s}=0\right)}.\end{aligned} \tag{2.1}$$

We obtain an analytical formulation of the model by assuming that g is a truncated-at-zero Poisson distribution:

$$g\left(y_{l,s} | y_{l,s} > 0; x_{l,s}; \alpha_s\right) = \frac{\mu\left(x_{l,s}; \alpha_s\right)^{y_{l,s}} \exp\left(-\mu\left(x_{l,s}; \alpha_s\right)\right)}{y_{l,s}!\left(1 - \exp\left(-\mu\left(x_{l,s}; \alpha_s\right)\right)\right)}, \tag{2.2}$$

where the rate μ is defined as:

$$\mu\left(x_{l,s}; \alpha_s\right) = \exp\left(x'_{l,s}\alpha_s\right). \tag{2.3}$$

In addition, a logistic distribution is used to characterize the hurdle:

$$\Pr\left(y_{l,s}=0|x_{l,s};\beta_s\right)=\frac{\exp\left(-\frac{\Pi\left(x_{l,s};\beta_s\right)}{\sigma}\right)}{1+\exp\left(-\frac{\Pi\left(x_{l,s};\beta_s\right)}{\sigma}\right)}, \tag{2.4}$$

where:

$$\Pi\left(x_{l,s};\beta_s\right)=x'_{l,s}\beta_s. \tag{2.5}$$

Equation 2.5 models the deterministic part of a latent profit function. The probability that there is no establishment of type s which locates at l increases when the potential (latent) profit decreases.

It is straightforward to derive that the expectation of the count variable is given by:

$$\mathbb{E}\left(y_{l,s}|x_{l,s};\alpha_s,\beta_s\right)=\Phi\left(x_{l,s};\alpha_s,\beta_s\right)\mu\left(x_{l,s};\alpha_s\right) \tag{2.6}$$

and that the variance is equal to:

$$\mathbb{V}\left(y_{l,s}|x_{l,s};\alpha_s,\beta_s\right)=\Phi\left(x_{l,s};\alpha_s,\beta_s\right)\mu\left(x_{l,s};\alpha_s\right)\left(1+\left(1-\Phi\left(x_{l,s};\alpha_s,\beta_s\right)\right)\mu\left(x_{l,s};\alpha_s\right)\right), \tag{2.7}$$

where:

$$\Phi\left(x_{l,s};\alpha_s,\beta_s\right)=\frac{1-\Pr\left(y_{l,s}=0|x_{l,s};\beta_s\right)}{1-\exp\left(-\mu\left(x_{l,s};\alpha_s\right)\right)}. \tag{2.8}$$

Overdispersion occurs when $\Phi\left(x_{l,s};\alpha_s,\beta_s\right)<1$, underdispersion when $\Phi\left(x_{l,s};\alpha_s,\beta_s\right)>1$.

Another interesting feature of the model is that the total effect of an independent variable on the count process passes through two channels. Looking at the elasticity of the expectation of the count variable $y_{l,s}$ to any related variable $x_{l,s}$, we find that it is equal to the sum of the elasticity of the participation probability to this variable and the elasticity of the conditional expected count to this variable:

$$\epsilon_{\mathbb{E}(y_{l,s}|x_{l,s};\alpha_s,\beta_s)/x_{l,s}}=\epsilon_{1-\Pr\left(y_{l,s}=0|x_{l,s};\beta_s\right)/x_{l,s}}+\epsilon_{\mathbb{E}\left(y_{l,s}|y_{l,s}>0,x_{l,s};\alpha_s\right)/x_{l,s}}. \tag{2.9}$$

Given a sample of independently and identically distributed observations for l locations and S sectors, the log-likelihood function is written as:

$$\ell\ell\left(\alpha_{\mathsf{s}}, \beta_{\mathsf{s}} | \mathrm{y}; \mathrm{x}\right) = \sum\nolimits_{\mathsf{l}=1}^{\mathsf{L}} \sum\nolimits_{\mathsf{s}=1}^{\mathsf{S}} \ln \ell\left(\alpha_{\mathsf{s}}, \beta_{\mathsf{s}} | \mathsf{y}_{\mathsf{l},\mathsf{s}}; \mathrm{x}_{\mathsf{l},\mathsf{s}}\right). \tag{2.10}$$

Note that we here assume that there are no unobserved correlations between different sectors s. We also do not account for spatial errors and spatial lags. There is a need to fill this gap in the further studies. The paper of Lambert et al. (2010) and their spatial lag model of counts can give a hint.

2.6.3 Spatial spillovers

We discuss here the structure of the matrix of observed explanatory variables. $\mathrm{x}_{\mathsf{l},\mathsf{s}}$ is actually a generic notation used for the sake of conciseness in notations. Due to data availability and our prior knowledge regarding the various expected effects, this matrix has a specific structure: it contains implicitly the number of pre-existing establishments from a respective sector s located at l, the number of large pre-existing establishments from all sectors located at l, and other explanatory variables directly concerning either location l or sector s. It is also possible to make this structure more specific by accounting for spatial interdependence: the stocks of establishments from some sectors located in neighboring cities and locational-specific attributes of neighbor cities.

Without going too deeply into the discussion, we will simply state that we make use of the distance matrix to characterize spatial patterns[13]. Currently, there are two basic categories that define neighbors: contiguity (shared borders) and distance. Contiguity-based weights matrices include rook and queen matrices. A rook matrix defines a neighbor as an area with a shared border while a queen matrix defines a neighbor as an area with a shared border and a shared vertex (point) (*e.g.*, on a grid); in addition to the four cells included under a rook matrix, the four cells sharing a corner with the central location are also counted as neighbors. Distance-based weights matrices include distance bands and k nearest neighbors[14].

Spatial spillovers are simply modeled as:

$$\mathrm{x}_{\mathrm{l,s}} = \ln\left(\sum\nolimits_{\mathsf{j}=1}^{\mathsf{L}} \mathrm{e}^{-\mu \mathsf{d}_{\mathsf{l},\mathsf{j}}} \mathrm{z}_{\mathrm{j,s}}\right), \tag{2.11}$$

[13] Rook or queen contiguity matrices could have been proposed as alternatives.
[14] Source: https://geodacenter.asu.edu/node/390.

where $z_{j,s}$ is an attribute of the municipality that applies to sector s (the same attribute may apply to several sectors) or is the number of pre-existing establishments from this sector. μ is fixed to 1 for our application. There is further work to carry out about values that μ may take. As it models the range of spatial spillover effects, it may at least differ along with sectors s. This work is left aside for the moment. $d_{l,j}$ is the distance between the centroids of municipalities l and j.

No period-specific pattern is modeled. Nevertheless, we implicitly parameterize a dynamic structure in our cross-sectional data by assuming that the location of newly-created establishments during our period of observation is related to the stock of pre-existing establishments at the beginning of said period. The interdependence between activity sectors can be included by considering that newly-created establishments within one activity sector are related to the stock of pre-existing establishments from both said and other activity sectors. As our geographical structure is based on zones, we need to consider that there may exist some spatial patterns between them.

2.7 Results

2.7.1 Count data model comparison

We run Poisson, zero-inflated Poisson, zero-inflated tau Poisson, negative binomial, zero-inflated negative binomial, and hurdle models. Each model is run separately for each selected activity sector. Each model is specified along two dimensions: (1) without accommodating spatial effects and (2) using the distance matrix. We run the models with variables that are expected to represent the municipality characteristics and the numbers of pre-existing establishments.

Some of the models we run are nested (*e.g.*, Poisson and NB, Poisson and hurdle) while others are non-nested (*e.g.*, ZIP and Poisson, ZIP and NB, ZIP and hurdle). Two models are nested if one can be reduced to the other by imposing a set of linear restrictions on the parameter vector. Conversely, two models are non-nested (either partially or strictly) when this condition is not met (Clarke, 2003). In the statistics literature, two methods are generally used to compare non-nested models: the Vuong test proposed by Greene (1994) and the Clarke test (2003). Non-nested models can still be compared using information criteria, such as AIC (Akaike Information Criterion) and BIC (Schwartz' Bayesian Information Criterion) provided in most procedures using maximum likelihood estimation. Models with smaller values of these criteria are considered better models. However, no statistical test comparing criterion values is available.

The main conclusion we draw from our results is that incorporating spatial spillovers between location choice alternatives (municipalities) in the form of a distance matrix and for each type of model and for all the activity sectors[15] improves the estimates. Looking at the models for which the distance matrix was used and postulating that the justification of best model is solely based on statistical tests, it would be tempting to conclude that NB outperforms its counterparts for lower AIC and BIC. The NB model outperforms the Poisson model for each activity sector, as it is characterized by a lower level of the absolute value of log-likelihood (for nested models, the closer to zero the level of log-likelihood, the better the model). Yet, hurdle and ZIP regressions offer greater flexibility in modeling zero outcomes than do the other count data models that we run. The major difference between hurdle and ZIP models is that the Logit component of the hurdle model estimates the probability of a zero count, whereas the Logit component of the ZIP model describes the probability of a zero count from two groups of either always zero and not always zero. Thus, the interpretation of the hurdle model is more straightforward and more intuitive than that of the ZIP model. The first part of the estimates in the hurdle model explains why no new establishments are created in a given municipality, whereas the second part of the estimates describes the influence of various determinants on the positive creation of units within the municipality. Having no creation of establishments in a municipality happens when profit is exhausted (for instance by other pre-existing establishments within the municipality) or when there is no space for new units.

We observe that the ZIP model outperforms the ZIP τ model in all but one case[16]. The hurdle model outperforms the ZIP τ model in all of the analyzed cases, with no exceptions. A hurdle model usually gives better results than a ZIP model[17]. For these reasons, ZIP and ZIP

[15] We test six types of count data models: Poisson, ZIP, ZIP τ, NB, ZINB, and hurdle models for seven selected activity sectors with variables that represent the municipality's characteristics, the density of pre-existing establishments from the respective sector, and the density of large pre-existing establishments. Each model is then run with and without incorporating spatial spillover effects between the municipalities. In this way, we are able to obtain results from 84 models.

[16] This exception is the model run without the distance matrix for the industry sector.

[17] In the case of the ZIP model when no matrix is used, for the real estate and special, scientific and technical activities sectors, not all parameters were estimated.

τ models are not treated as rivals for the hurdle model. ZINB could have been suggested as an alternative to the hurdle model since each sector is characterized by very low values of the AIC and BIC criteria (lower than in the case of hurdle and ZIP models). However, when running the ZINB model, whether using the quasi-Newton optimization method or not in the case of many sectors, convergence could not be reached under the requirement of estimating all parameters. Taking into account all of these considerations, we therefore chose to favor the hurdle model.

2.7.2 Hurdle model: effect of various municipality characteristics on location choice

In the highly heterogeneous Paris region, some municipalities find themselves home to a large number of new establishments whereas others struggle to be chosen by any establishment. The first part of the hurdle estimates shows the reasons for no newly-created establishment in the municipality. Depending on the analyzed sector, the percentage of municipalities left with no new creation ranges from 34% up to 69%. The number of municipalities left with zero newly-created establishments in the industry sector equals to 734, in construction 439, commerce 440, transport 837, financial activities 799, real estate activities 738, hotels and restaurants 792, information and communication 771, special, scientific and technical activities 569, education 890, health and social activities 794 out of 1300 possible municipalities. In the subsection 2.7.2, we present an interpretation of the results from the first part of the hurdle models. In the following subsection (2.7.2), the reader can find an interpretation of the results from the second part of the hurdle models with a positive number of establishment creations occurring in the municipality.

Regarding the weight matrix, the spatial modeling accounts for the density of establishments from the respective sector in neighboring areas, the employment density of white- and blue-collar workers, and also the existence of large employers from all sectors with fifty workers or more. These large establishments might attract complementary upstream and downstream establishments (sometimes even within the same firm), leading to positive autocorrelation, or which might affect the labor market in such a way that it becomes costlier to hire employees and locate nearby, leading to negative autocorrelation. For this reason, it is necessary to capture the possible influence of large establishments on the location choice of newly-created entities.

No new creation of establishments in municipality

Large establishments with fifty employees or more seem not to add much in the explanation of why there are no new establishment openings in the area, beside that they have a weak negative effect on the real estate sector and a positive effect on the financial sector. A municipality is left with zero new openings of industrial establishments when there are no, or a few other establishments from the industry sector and/or when the average level of real estate prices is too high. Relatively few pre-existing construction establishments, low population density, high office prices, and/or difficulties in accessing public transport do not seem to encourage the creation of new establishments in the construction sector. There is a propensity for no new commercial establishments to be created in municipalities that lack other commercial units and/or when commercial rents are particularly high, both within the area and its surroundings. No new creation of hotels and restaurants can be observed in the areas characterized by steep price level, not enough vacant land where a new investment could be made, and/or few pre-existing units from this sector. No new financial establishments will be created when the number of pre-existing establishments within that same sector is too low. Steep office prices and limited access to white-collar workers, all beyond a certain threshold, may also lead to an absence of new openings for financial units. Difficulties in accessing the potential location by public transport may leave the municipality without any new openings of financial establishments either. There are also various reasons behind a municipality being left without any new real estate sector establishments. It is most likely to happen when there are too few, or indeed no pre-existing real estate units. It can also occur in case of high residential rate in the area, and/or when access to shops and services is hampered. Finally, a lack of other pre-existing establishments in special, scientific and technical activities are the most significant variables that cause zero new unit creation in this sector. In addition, difficulties for white-collar workers to access the premises, excessive office prices, and difficult access by public transport appear not to help to create a positive number of establishments in the special, scientific and technical activities sector.

Positive number of newly-created establishments in municipality

The models presented in this subsection, as well as the models explained in subsection 2.7.2, show the importance of interactions between establishments from a same activity sector. The number of new locations will be highly dependent on the quantity of competing establishments in the municipality and in its surroundings. Pre-existing establishments will encourage new units to locate nearby. This effect is found to be strongest in the case of construction, commerce, and special, scientific and technical activities. Overall, large establishments tend to attract industrial units, yet, have a negative effect on other analyzed sectors.

Drawing from the results from the model that describes the location choice of industrial establishments, we can observe that units tend to choose a location characterized by easy access to blue-collar workers, large areas of vacant land that can be made available for new investments, and/or a conveniently located highway. Newly-created establishments in the construction sector strongly prefer to locate in an area characterized by a high level of population density. In order to create this variable, we computed the ratio of population size to municipality surface zoned as residential area. We opted to consider only the surface zoned as residential area (instead of the entire municipality's surface) as a means to better express the population density measure. These establishments in the construction sector would rather choose areas with vacant land that could be used for new investments. Another important factor that should be taken into account is close proximity to public transport services. The distance to the highway does not seem to play a very significant role when the decision for a new location is taken. In addition, establishments in the construction sector pay close attention to office prices and residence tax rates. Access to the potential location by cars and close proximity to subway, train stations, or bus stops can be perceived as fundamental factors by a commercial establishment when choosing its location. It is important for future commercial establishments to be located in the vicinity of potential customers. Consequently, establishments will try to reach areas characterized by a high number of population who work and live nearby. Relatively steep shop prices will discourage the opening of new commercial units. Close proximity to subway, train stations, or bus stops seems to be very significant also in the location choice model of hotels and restaurants. Being located at a short distance from the highway will also play a positive, albeit weaker, role when deciding on the new location. Relatively easy access to white-collar workers who can profit from the restaurant or hotel services, availability of vacant land, and low real estate prices can also act as incentives in the location choice.

Establishments in the financial sector will look for municipalities easily accessible by cars. Easy public transport access is also relevant, yet slightly less significant. Financial units favor areas occupied by white-collar workers. It is worth mentioning that office prices appear to be weakly significant when financial activities choose a new location. Existence of shops and offices in and around a chosen municipality seems not to accelerate the new openings in the financial sector. The model for newly-created establishments in the real estate sector demonstrates that a potential location should be easily accessible by public transport. Public transit accessibility is the most significant explanatory variables in this model. A potential location should be visible and frequently used. Establishments that belong to the real estate sector favor zones with easily accessible shops and services. Being located in zones that offer ”a good image” is usually favored when the establishment chooses its new location. It could be argued that close proximity to wealthy clients attracts new real estate units. Low residence tax rates and low prices seem to strongly encourage new openings in the real estate sector.

Finally, ease of access, be it by public or private transport are significant variables in the model for special, scientific and technical activities. Access to the white-collar workforce within the close vicinity will accelerate the number of new openings in this sector. Yet, these establishments tend to avoid overpopulated areas. Office prices does not play a critical role. Neither the pre-existing offices attract the new establishments in the special, scientific and technical activities sector, indicating that they will rather locate close to other pre-existing units within the same sector. Tables 2.6, 2.7 summarize the results from these estimated hurdle models.

2.8 Conclusions

The following conclusions are based on both the descriptive statistics and the results from the count data models that we have estimated. Questions, such as: Where in the Paris region do new establishments decide to locate most often? and What is the level of activity concentration? are now answerable based on rankings of municipalities' market share. We found, that the top municipalities are the ones that belong to Paris and to *départements* 92 and 93. *Départements* from the outer ring seem to be the least attractive for new establishments. About 14% of all newly-created establishments in the real estate and financial and insurance sectors chose to locate in the 8th district of Paris. In contrast, there does not seem to be an ideal municipality, nor one that particularly appeals to the transport and commerce sectors so as to be very frequently chosen as a destination. After determining preference for municipality, our next step consisted in analyzing the size of newly-created establishments in comparison to the size of pre-existing ones. We observed a strong contrast between the typical size of these two types of establishments across various size classes. The most appealing conclusion rests on the fact that 82% of newly-created establishments have no employees beside the owner whatsoever, whereas this is the case for only 61% of pre-existing establishments. In addition, we detect 7.5% of pre-existing establishments having at least 10 employees, in comparison to the 2.8% of newly-created establishments of the same size. Just 0.2% of establishments that launch their activity in the market have at least 50 employees; this percentage is 7 times higher for establishments that have survived for some time in the market. Large differences in the size of newly-created establishments can also be observed across activity sectors.

The principal and most relevant conclusion we present is drawn from the results from the count data models that we have estimated. We have tested, for the purposes of the present research, six types of count data models: Poisson, zero-inflated Poisson, zero-inflated tau Poisson, negative binomial, zero-inflated negative binomial, and hurdle-Poisson models. Each type of model has been run with variables that represent: density of pre-existing establishments in the respective sector, density of large pre-existing establishments, and various municipality characteristics. Care was taken to distinguish outcomes across different activity sectors. We then checked if the results from each model could be improved by incorporating spatial spillovers between municipalities in the form of the distance matrix. In this way, we were able to obtain results from 84 models. In all the analyzed cases, the results from the models run with the distance matrix indicated that accommodating spatial spillovers significantly improves the model's performance. We suggest accounting for spatial spillovers when modeling establishment location choice.

Hurdle models turned out to be preferred. These models offer greater flexibility in modeling zero outcomes and relax the assumption that the zero observations and the positive observations come from the same data generating process. The major difference between hurdle and zero-inflated Poisson models is that the Logit component of the hurdle model estimates the probability of a zero count, whereas the Logit component of the zero-inflated Poisson model describes the probability of a zero count from two groups of either always zero and not always zero. Results also indicated that an establishment does not act in isolation and is influenced by other pre-existing establishments. Overall, the greater the presence of establishments from a particular sector, the greater the number of newly-created units observed locating nearby in the market. Large establishments tend to attract other industrial units, yet, have a negative effect on other sectors. In addition, when choosing a location in the market, an establishment as a decision-maker may take into consideration not only the characteristics of the potential municipality but also the characteristics of its surroundings due to spatial spillovers from neighboring areas. In addition, we found that new establishments tend to avoid areas characterized by high real estate prices. Price levels for shops or offices play a strong significant and negative role in location choice for units in construction, commerce, real estate, and for hotels and restaurants. High rates of residence tax appear to discourage the creation of units in the real estate and construction sectors. Establishments seek areas with high availability of vacant land. The new establishments in the construction sector are prone to concentrate on highly dense residential areas and commercial establishments tend to reach their potential customers. Access to the intellectual workforce is important in the special, scientific and technical sector, in the financial sector as well as for hotels and restaurants. Finally, convenient transport infrastructure seems to play a considerable role in the location choice decisions of establishments.

The contents of the paper may be useful for researchers in the field and those who are interested in policy implications. An approach taken in this paper is aggregate and descriptive and could be further extended in several ways: (1) Other measures beyond simple Euclidean distance (crow-fly distance) may be tested. Travel times by car (free-flow and for the peak period) and by public transit, and the corresponding network distances at equilibrium during peak hours for private vehicles and public transit may be proposed as alternative weight matrices (see Buczkowska et al., 2016). (2) The strength of spatial decay of neighborhood effects could be examined. (3) Care should be taken to control for spatial dependence. (4) A challenge would be to develop a disaggregated establishment location choice model incorporating both spatial and strategic interactions between establishments and spatial spillovers among choice alternatives. As discerned by Lambert et al. (2010), accommodation of spatial effects for location decisions may provide a richer, more explicit picture of the regional linkages supporting local growth, industry clustering, and economic development.

Acknowledgements

The authors would like to thank Alice Koegh and Tierra McMahon for their time and editorial support.

References

Allen, W.B., D. Liu, and S. Singer (1993). Accessibility measures of US metropolitan areas. *Transportation Research Part B*, 27(6), 439-449.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Dordrecht, Kluwer Academic Publishers.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review*, 26, 153-166.

Arauzo-Carod, J.-M. (2005). Determinants of Industrial Location: An Application for Catalan municipalities. *Papers in Regional Science*, 84(1), 105-120.

Arauzo-Carod, J.-M., D. Liviano-Solís, and M. Manjón-Antolín (2010). Empirical studies in industrial location choice: An assessment of their methods and results. *Journal of Regional Science*, 50(3), 685-711.

Bhat, C.R. and J.Y. Guo (2004). A Mixed Spatially Correlated Logit Model: Formulation and Application to Residential Choice Modeling, *Transportation Research Part B*, 38(2), 147-168.

Bodenmann, B.R. (2011). Modelling firm (re-)location choice in UrbanSim, Paper presented at the 51st ERSA Congress, Barcelona, September 2011.

Bodenmann, B.R. and K.W. Axhausen (2010). Synthesis report on the state of the art on firmographics. *SustainCity Working Paper 2.3*, Institute for Transport Planning and Systems (IVT), ETH Zurich.

Bodenmann, B.R. and K.W. Axhausen (2012). Destination choice of relocating firms - A discrete choice model for the region of St. Gallen, Switzerland. *Papers in Regional Science*, 91(2), 319-341.

Bondonio, D. and R.T. Greenbaum (2007). Do local tax incentives affect economic growth? What mean impacts miss in the analysis of enterprise zone policies. *Regional Science and Urban Economics*, 37(1), 121-136.

Buczkowska, S., N. Coulombel, and M. Lapparent (de) (2016). Euclidean distance versus network distance in business location: A probabilistic mixture of hurdle-Poisson models. *Working Paper*, Université Paris-Est, Marne la Vallée, France.

Cameron, A.C. and P.K. Trivedi (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, 1, 29-54.

Clarke, K.A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, 47, 72-93.

Combes, P.P., M. Lafourcade, J.-F. Thisse, and J.-C. Toutain (2011). The rise and fall of spatial inequalities in France: A long run perspective. *Explorations in Economic History*, 48, 243-271.

Consul, P.C. and G.C. Jain (1973). A generalization of Poisson distribution. *Technometrics*, 15, 791-799.

de Bok, M. (2004). Explaining the location decision of moving firms using their mobility profile and the accessibility of locations. *ERSA conference papers ersa04p338*, European Regional Science Association.

de Palma, A. (2011). The Grand Paris Project: Tools and Challenges. *International Transport Forum Discussion Papers 2011/28*, OECD Publishing.

de Palma, A., K. Motamedi, J. Moyano, and N. Picard (2008). Etude des décisions de localisation des emplois en Île-de-France. *Report for the Direction Régionale de l'Equipement Ile de France (DREIF)*, AdpC, Paris, France.

Devereux, M.P., R. Griffith, and H. Simpson (2007). Firm Location Decisions, Regional Grants and Agglomeration Economies. *Journal of Public Economics*, 91, 413-35.

Garrido, R.A. and H.S. Mahmassani (2000). Forecasting freight transportation demand with the space-time multinomial probit model. *Transportation Research Part B*, 34(5), 403-418.

Greene, W.H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. *Working paper*, Stern School of Business, NYU EC-94-10.

IAU IdF (2014). Emploi et crise, Île-de-France et 7 autres régions, January 2014.

Jayet, H. (2001). Econométrie des données spatiales. Une introduction à la pratique. *Cahiers d'Economie et de Sociologie Rurale*, 58-59, 105-129.

Klier, T. and D. McMillen (2008). Clustering of auto supplier plants in the United States. *Journal of Business & Economic Statistics*, American Statistical Association, 26(4), 460-471.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14.

Lambert, D., J.P. Brown, and R.J.G.M. Florax (2010). A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. *Regional Science and Urban Economics*, 40(4), 241-252.

Liviano-Solís, D. and J.-M. Arauzo-Carod (2014). Industrial Location and Spatial Dependence: An Empirical Application. *Regional Studies*, 48, 1-17.

Maoh, H.F. (2005). *Modeling firm demography in urban areas with an application to Hamilton, Ontario: towards an agent-based approach*, PhD Thesis, School of Geography and Earth Sciences, McMaster University, Hamilton, Canada.

Maoh, H.F. and P.S. Kanaroglou (2005). Agent-Based Firmographic Models: a Simulation Framework for the City of Hamilton, Proceedings of PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto.

Maoh, H.F. and P.S. Kanaroglou (2007). Business establishment mobility behavior in urban areas: a microanalytical model for the City of Hamilton in Ontario, Canada. *Journal of Geographical Systems*, 9(3), 229-252.

Miyamoto, K., V. Vichiensan, N. Shimomura, and A. Paez (2004). Discrete choice model with structuralized spatial effects for location analysis. In: *Transportation Research Record: Journal of the Transportation Research Board*, 312, TRB, National Research Council, Washington, DC.

Mohammadian, A.K., M. Haider, and P.S. Kanaroglou (2005). Incorporating spatial dependencies in random parameter discrete choice models. In: *TRB* (ed.) the 84th Annual Meeting of the Transportation Research Board, Washington, D.C.

Motamedi, K. (2008). *Contribution à la modélisation intégrée de la dynamique régionale et des transports: application à la région Parisienne*, PhD dissertation, University of Cergy-Pontoise, Cergy, France.

Mullahy, J. (1986). Specification and testing of some modified count models. *Journal of Econometrics*, 33, 341-365.

Neumark, D. and J. Kolko (2010). Do enterprise zones create jobs? Evidence from California's enterprise zone program. *Journal of Urban Economics*, 68(1), 1-19.

Nguyen, C.Y., K. Sano, T.V. Tran, and T.T. Doan (2012). Firm relocation patterns incorporating spatial interactions. *The Annals of Regional Science*, Published online: 11 August 2012.

Rocha, N. (2008). Firm Location Determinants: Empirical Evidence for France. *IHEID Working Papers 08-2008*, Economics Section, The Graduate Institute of International Studies.

Sener, I., R. Pendyala, and C. Bhat (2011). Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. *Journal of Transport Geography*, 19(2), 293-303.

Siebert, H. (2000). The paradigm of locational competition. *Kieler Diskussionsbeiträge*, 367, Institut für Weltwirtschaft, Kiel.

Smirnov, O. (2010). Modeling Spatial Discrete Choice. *Regional Science and Urban Economics*, 40(5), 292-298.

Strotmann, H. (2007). Entrepreneurial Survival. *Small Business Economics*, 28, 87-104.

Sutton, J. (2007). Market Structure: Theory and Evidence, Elsevier, vol. 3 of *Handbook of Industrial Organization*, chap. 35, 2301-2368.

van Wissen, L.J. (2000). A micro-simulation model of firms: Applications of concepts of the demography of the firm. *Papers in Regional Science*, 79, 111-134.

Vichiensan, V., Y. Tokunaga, and K. Miyamoto (2005). Mixed logit model framework with structuralized spatial effects: a test of applicability with area unit systems in location analysis. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 3789-3802

Winkelmann, R. (2003). *Econometric Analysis of Count Data*, 4th ed. Springer Verlag, Heidelberg.

Table 2.6: Hurdle model with distance matrix: explanatory variables represent density of pre-existing establishments in the respective sector, density of large pre-existing establishments, and various municipality characteristics.

Hurdle part I:	Industr[a]		Constr		Commer		Finan		RealEst		HoteRes		SpecSci	
zero counts	Est	t-st[b]	Est	t-st	Est	t-st	Est	t-st	Est	t-st	Est	t-st	Est	t-st
Constant	-18.330***	-6.51	-23.338***	-7.20	-23.420***	-6.76	-16.307***	-5.32	-9.949**	-2.06	-24.271***	-6.72	-20.789***	-5.37
Estab., resp. sec.[c]	-1.218***	-12.29	-0.663***	-5.31	-1.053***	-10.51	-0.580***	-6.27	-0.931***	-9.69	-1.374***	-12.59	-0.882***	-8.66
Large estab., all sec.[d]	0.053	1.10	0.049	1.16	0.003	0.06	-0.187**	-2.42	0.108*	1.83	0.066	1.15	0.046	0.86
White-collar workers							-0.283**	-2.34			-0.155	-1.45	-0.205**	-2.18
Blue-collar workers	0.070	0.62												
Trips home-work					-0.043	-0.20								
Population density			-1.043***	-4.13									-0.180	-0.79
Income per person									0.123	0.29				
Offices													-0.002	-0.12
Shops and offices							0.010	0.16	-0.098*	-1.76				
Vacant land	0.136	1.50	0.096	1.08	-0.032	-0.35	0.070	0.70			0.311***	2.93		
Distance to highway	0.114	1.28	0.117	1.36	0.006	0.06	-0.042	-0.41			-0.005	-0.05	0.057	0.62
Public transport			-0.052*	-1.84	-0.021	-0.73	-0.092**	-2.39	-0.006	-0.16	-0.010	-0.28	-0.091***	-2.88
Residence tax			0.005	0.02					-0.181	-0.73				
Price of offices	2.418***	6.65	3.016***	7.49			1.937***	5.10	1.212*	1.86			2.485***	5.22
Price of shops					2.814***	6.81					3.102***	7.09		

[a] Dependent variable: number of newly-created establishments in 2007.

[b] ***, **, * represent statical significance at the 1%, 5%, and 10% level, respectively.

[c] The data on stock of establishments used to construct the explanatory variables in the model are given for the 1st of January 2007. See Table 2.1 for description of potential explanatory variables.

[d] Large pre-existing establishments from all sectors with 50 employees or more.

Table 2.7: Hurdle model with distance matrix: explanatory variables represent density of pre-existing establishments in the respective sector, density of large pre-existing establishments, and various municipality characteristics (continuation).

Hurdle part II: positive counts	Industr[a] Est	t-st[b]	Constr Est	t-st	Commer Est	t-st	Finan Est	t-st	RealEst Est	t-st	HoteRes Est	t-st	SpecSci Est	t-st
Constant	0.269	0.44	12.596***	38.94	13.470***	42.49	3.255***	4.01	11.621***	15.78	5.857***	7.03	2.720***	6.94
Estab, resp. sec.[c]	0.593***	15.11	1.148***	54.85	0.527***	31.21	1.084***	20.21	0.812***	18.20	0.600***	12.10	1.001***	40.45
Large estab. all sec.[d]	0.087**	2.49	-0.128***	-9.37	-0.060***	-4.69	-0.760***	-16.18	-0.203***	-4.76	-0.228***	-4.89	-0.131***	-6.16
White-collar workers							0.771***	19.17			0.307***	6.14	0.154***	5.70
Blue-collar workers	0.083**	2.08												
Trips home-work					1.130***	35.10								
Population density			0.134***	4.09									-0.865***	-29.34
Income per person									0.377***	6.53				
Offices													-0.030***	-4.39
Shops and offices							-0.050	-1.31	0.240***	5.91				
Vacant land	0.064**	2.30	0.072***	5.01	0.049***	4.43	0.227***	10.07			0.146***	5.11		
Distance to highway	-0.215***	-6.19	-0.010	-0.58	-0.311***	-24.67	-0.478***	-16.18			-0.112***	-3.39	-0.562***	-38.46
Public transport			0.225***	18.84	0.260***	27.82	0.313***	12.25	0.251***	10.32	0.501***	16.47	0.430***	31.39
Residence tax			-1.312***	-30.47					-2.137***	-32.30				
Price of offices	0.013	0.16	-1.154***	-29.53			-0.188*	-1.80	-0.610***	-7.29			-0.072	-1.38
Price of shops					-1.220***	-31.81					-0.576***	-5.41		
Log-likelihood	-2071.0		-4472.4		-5194.5		-2214.4		-2449.1		-1803.5		-4424.4	
AIC	4169.9		8980.8		10421.0		4464.8		4930.2		3638.9		8884.7	
AICC	4170.3		8981.3		10422.0		4465.3		4930.7		3639.3		8885.3	
BIC	4242.3		9073.9		10504.0		4557.9		5013.0		3721.6		8977.8	
Convergence crit.	Yes		Yes		Yes		Yes		Yes		Yes		Yes	
#Parameters	2x7		2x9		2x8		2x9		2x8		2x8		2x9	

[a] Dependent variable: number of newly-created establishments in 2007.

[b] ***, **, * represent statical significance at the 1%, 5%, and 10% level, respectively.

[c] The data on stock of establishments used to construct the explanatory variables in the model are given for the 1st of January 2007. See Table 2.1 for description of potential explanatory variables.

[d] Large pre-existing establishments from all sectors with 50 employees or more.

Near is a more flexible and powerful concept than commonly appreciated.

- H.J. Miller, "Tobler's First Law and Spatial Analysis" (2004)

3

Euclidean versus network distances: A mixture model of the intra-metropolitan location choice of establishments

Abstract This article investigates the issue of the metric choice in spatial models. Although the Euclidean distance still prevails in the literature, more and more works advocate the use of alternate metrics, be it economic or transport distances, which may be more relevant to the problem considered. Geographic factors such as terrain, land cover, infrastructure, and traffic congestion may cause agents not to follow pure Euclidean relations. Instead of imposing a restrictive metric, this research proposes a flexible approach to point which alternative is more appropriate to correctly account for the surrounding economic landscape. A probabilistic mixture of two "mono-distance" hurdle-Poisson models is developed. Each model's latent class uses a different distance representation to incorporate spillover effects in location choices of establishments from several activity sectors. Seven metrics are considered: Euclidean distance, travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances. This methodology allows to capture the diversity of agents' behavior, i.e., to distinguish establishments which are more time- or more distance-oriented given location. The distance measure being most likely to capture spatial spillovers varies de-

pending on the economic sector. This supports the choice of a "meaningful" over "more abstract" metric in spatial econometric models. The peak travel time strongly predominates for establishments from the construction and special, scientific and technical activities sectors, while the off-peak travel time or the Euclidean distance (depending on the pair of metrics used) tend to prevail for the real estate sector, reflecting differences between sectors in operations and location choice criteria.

Keywords: Euclidean distance, length-based distance, time-based distance, network distance, congestion, physical barriers, intra-region location, latent classes

JEL Classification: C31, C35, L22, R41

3.1 Introduction

Numerous studies have attempted to determine which specification of the spatial weight matrix (W) best fits their data and investigated the robustness of their results to different W specifications (*e.g.*, Bell and Bockstell, 2000; Kostov, 2010). Examinations cover the neighborhood definition (rook or queen matrix, k nearest neighbors, etc.), the specification of the distance decay function, or the bandwidth size. Getis and Altstadt (2004) summarize the typical well-known schemes that researchers follow to find an appropriate W matrix. These schemes are: (1) spatially contiguous neighbors, (2) inverse distances raised to some power, (3) lengths of shared borders divided by the perimeter, (4) bandwidth as the k-th nearest neighbor distance, (5) ranked distances, (6) constrained weights for an observation equal to some constant, (7) all centroids within distance d, (8) k nearest neighbors, and so on. Some of the newer propositions include: (1) bandwidth distance decay (Fotheringham et al., 1996), (2) Gaussian distance decline (LeSage, 2003), and (3) "tricube" distance decline function (McMillen and McDonald, 2004). As yet, the definition of distance has attracted less attention. When the spatial weight matrix is based on distance, the choice of the metric is often glossed over, with ultimately a preponderating use of the Euclidean metric (exceptions include Nguyen et al., 2012). The requirement that the weights should be exogenous to the model is usually called upon to motivate the choice of the Euclidean distance, because the underlying geographical structure is arguably exogenous in most applications[1]. As noticed by Miller and Wentz (2003),

[1]See Drukker et al. (2013) for a joint test of zero spatial interactions in the dependent variable, the exogenous variables and the disturbances; Kelejian and Prucha (1998) for a feasible generalized spatial two-stage least squares

many researchers adopt the Euclidean metric without realizing its underlying assumptions or its alternatives. This representation has been originally proposed due to scarce data and low computational power at the time, rather than because of its alleged universality. Yet, it remains overwhelmingly applied to this day.

Corrado and Fingleton (2012) argue that the specification of the spatial weight matrix, including the definition of distance, should be supported by economic theory. Let us imagine two neighborhoods that are contiguous yet separated by some uncrossable physical barrier (a transport axis, a river, etc.). One would indeed expect that spatial spillovers would be smaller, if any at all, than that if the barrier was not there. Tobler's (1970) first law of geography ("everything is related to everything else, but near things are more related than distant things") proposes nearness as a main determinant of interactions among objects. As suggested by Miller (2004), the problem lies in an appropriate definition of nearness, usually based in the empirical literature on the straight-line segment connecting two locations. Miller (2004) points out that this is only one possibility out of an infinite number of shortest path relations. Natural (*e.g.*, rivers, lakes, mountains, steep hills) or artificial (*e.g.*, parks, cemeteries, golf courses, landmarks, transport infrastructures, industrial corridors) obstacles, in particular, require to make detours and expend additional energy and resources (see Fig. 3.2 and 3.3), something that is usually minimized (Miller 2004; Boscoe et al., 2012). This tends to limit interactions between the sundered territories (Jacobs, 1961), reducing the mobility of population nearby (Héran, 2011), restricting access to employment and population (Motte-Baumvol et al., 2015), and mitigating neighborhood externalities (Noonan, 2005). Physical features may even in some cases imply a total non-contiguity (Chakravorty, 1996). Accordingly, there is a good reason to expect geographic factors such as terrain, land cover, infrastructures, and traffic congestion to matter in human interactions.

The Euclidean metric might therefore not always be the most relevant one depending on the problem considered. Interest in this question dates at least to the 1960s and the research on network models in geography (Haggett, 1967). Talen and Anselin (1998) suggest that the distance can be computed in a variety of ways, and that when incorporating spatial externalities

procedure for estimating cross-sectional linear regression models that contain a spatial lag of dependent variable as a regressor or a disturbance term that is spatially autoregressive where the neighboring units can be those that are close in some dimension, such as geographical or technological. This procedure accounts for the endogeneity of y_{-i} by instrumental variables constructed as spatial lags of the exogenous characteristics x_i as well as spatial error correlation; Rincke (2010) for an application of Kelejian and Prucha' (1998) procedure.

the measure of access should be correlated with the socio-economic characteristics. Fingleton and Le Gallo (2008) state, for instance, that spillovers between areas are not a function of spatial proximity and that "it is more realistic to base it on relative economic distance". Parent and LeSage (2008) measure the level of regional technological proximity applying a number of spatial weight matrices based on geographical, transportation, and technological proximity. Another very stimulating approach is taken by Fingleton (2008) who proposes to use a matrix $C_{ij} = -\exp(\delta_i d_{ij})$ which is a function of the straight-line distance d_{ij} between areas i and j and an area specific coefficient δ_i which allows to account for varying levels of transport infrastructures and commuting patterns within the study area. More generally, several studies have considered alternative metrics that are not purely based on topography (*e.g.*, Bodson and Peeters, 1975; Aten, 1997; Duran-Fernandez and Santos, 2014; Conley and Ligon, 2002; Slade, 2005; Le Gallo and Dall'erba, 2008), including network distances and transport costs. All in all, when considering alternative metrics to the Euclidean one, a comparison based on the relative performances of two models, one using the Euclidean distance, the other the alternative distance, still largely prevails in the literature. Rather than locking researchers into a restrictive metric choice, this research proposes a flexible approach, wherein several metrics may be used within the same model, to point which alternative is more appropriate to correctly account for the surrounding economic landscape. A probabilistic mixture of two "mono-distance" hurdle-Poisson models is developed. Each model's latent class uses a different distance representation to incorporate spillover effects in location choices of establishments from several activity sectors. This methodology allows to capture the diversity of agents' behavior, i.e., to distinguish establishments which are more time- or more distance-oriented given location. We address the criticism of Rincke (2010), Vega and Elhorst (2013), and other authors that the choice of the matrix used in the spatial models is usually quite arbitrary, while it refers to the choice of the distance measure.

The methodology is applied to the establishment location choice problem. Because establishments are subject to interactions that are extremely varied by nature, spatial effects in location choice models might not be correctly captured using a single metric. Accordingly, this paper investigates the complementarity of the Euclidean distance and transport distances in the case of the location choice of newly-created establishments in the Paris region. The Paris region features high congestion levels as well as many physical barriers (including the Seine river), so that the Euclidean distance might indeed not always well reflect nearness between the zones. To test, compare, and mix various metrics, this paper enhances the location choice model of Buczkowska and Lapparent (2014) by estimating a probabilistic mixture of two "mono-distance" hurdle-Poisson models. The hurdle-Poisson model reflects a two-step decision making process. The hurdle model represents the establishments' decision whether to settle within a given area or not, based on the expected profitability. If an area is deemed profitable by establishments of a given sector as a whole, then and only then the Poisson model aims to predict the (strictly positive) number of establishments that do so. The proposed mixture of models uses two latent classes, each representing a different metric to capture spatial spillovers in the location choice of establishments. This provides a direct and integrated way of testing several metrics at a time. In addition to the standard Euclidean metric used in Buczkowska and Lapparent (2014), six alternative metrics are considered: travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances[2]. To the best of our knowledge, this work is one of the first formulations and applications of spatial count data models of location choice in regional science, wherein various metrics other than the Euclidean one are tested and associated to build the spatial distance weight matrices[3]. The methodology allows to capture the heterogeneity of agents' behavior, and to distinguish establishments which are more time- or distance-oriented given a particular location. Once the model is estimated, posterior class assignment probabilities may be computed based on observed location choices, updating for each location the knowledge about which spatial representation is most likely to correctly account for spatial spillovers.

The mixture of hurdle-Poisson models based on two metrics is found to perform markedly better than "pure" hurdle-Poisson models based on a single metric (according to the Bayesian Information Criterion). Parameter estimates for the mixture model are broadly consistent with the literature and not overly sensitive to the model specification. Second, the combination of peak and off-peak road travel times (slightly) outperforms other pairs of metrics that include the Euclidean distance. This supports the choice of economically meaningful metrics over more neutral ones in spatial econometric models. As a corollary, considering both the peak and off-peak travel conditions (rather than just one out of the two) proves worthwhile to better understand establishment location choices. Last, the distance measure most likely

[2]See Nguyen et al. (2012) for a relocation choice model where an average travel distance is used to proxy the distance among zones and firms.

[3]In the case of continuous data, spatial cross-regressive models have already been used to combine several metrics (*e.g.*, Kang and Dall'erba, 2016).

to capture spatial spillovers varies depending on the economic sector. The peak travel time strongly predominates for establishments from the construction and special, scientific and technical activities sectors, while the off-peak travel time or the Euclidean distance (depending on the pair of metrics used) tend to prevail for the real estate sector, reflecting differences between sectors in operations and location choice criteria.

The remainder of the paper is organized as follows. Section 3.2 reviews the literature first on establishment locational decisions and spatial effects, then on the metrics that can be used in these types of analyses. Section 3.3 describes the data. The econometric model is developed in Section 3.4. Section 3.5 elaborates on the results. Section 3.6 concludes.

3.2 Literature review

3.2.1 Spatial effects in the context of establishment location choice

Establishment location choice has attracted considerable attention in the past decades (see, *e.g.*, a survey of Arauzo-Carod et al., 2010). The objective of the studies on location of establishments is the analysis of the determinants of how the choices are being made. Distance plays usually only a secondary role in the estimation. Only recently has the importance of spatial effects been emphasized in this context. Despite the existence of spatial effects, there is a little mention in the literature of previous attempts to incorporate spatial effects in establishment or firm location decision-making processes (Bhat et al., 2014; Buczkowska and Lapparent, 2014; Liviano-Solís and Arauzo-Carod, 2013; Liesenfeld et al., 2015; Lambert et al., 2010; Klier and McMillen, 2008).

Klier and McMillen (2008) build a model with a spatially weighted dependent variable to analyze location decisions of auto supplier plants in the US (discrete choice framework). They account for the clustering tendency assuming that the location of a plant in a particular county depends on the location of plants in contiguous counties. Lambert et al. (2010) develop the Spatial Autoregressive Poisson model and assess the use of a two-step limited information maximum likelihood approach. This model includes a spatially lagged dependent variable as a covariate. The proposed estimator models the location events of start-up firms in the manufacturing industry as a function of neighboring counts. Effects of location determinants can be divided into direct, indirect, and induced effects thus providing information to better understand regional patterns. Liesenfeld et al. (2015) propose a maximum likelihood

(ML) approach based on the spatial efficient importance sampling applied to the spatial Poisson and negative-binomial models for manufacturing establishment location choices. ML estimation of parameter-driven count data models requires high-dimensional numerical integration. Bhat et al. (2014) formulate a spatial multivariate model to predict the count of new industrial businesses at a county level in the state of Texas. It allows for a better recognition of the industry specific determinants. The authors accommodate overdispersion and excess zero problems. They account for the unobserved factors that simultaneously affect the county-level count of new businesses in different sectors and spatial dependence effects across counties. Dubé et al. (2016) use a discrete choice model to describe the individual business location choices in the nonmetropolitan area in Canada demonstrating that the final decision of an establishment should be related to the surrounding economic landscape.

However, whenever a distance metric was used in the weight matrix to implement the spatial effects in the location choice model, beside the work of Nguyen et al. (2012)[4], no discussion was provided on the choice of the distance definition itself. Starting with the research on which this project has been based, Buczkowska and Lapparent (2014) test various nested and non-nested count data models implementing spillover effects as: $x_{i,s} = \ln\left(\sum_{j=1}^{I} e^{-\mu d_{i,j}} z_{j,s}\right)$, where $z_{j,s}$ is a municipality attribute that applies to the activity sector s or the number of pre-existing establishments from this sector, and $d_{i,j}$ is the Euclidean distance between the centroids of municipalities i and j. Liviano-Solís and Arauzo-Carod (2013) consider a distance matrix such that $w_{ij} = 1/d_{ij}$, where d_{ij} is the Euclidean distance between the municipalities i and j when implementing spatial spillovers in a number of distinct count data models. Bhat et al. (2014) test different specifications of the weight matrix in their spatial multivariate model, including inverse distance, inverse of the square of the distance, and inverse of the cube of the distance between counties. Yet, they do not concentrate on the distance definition. Lambert et al. (2010) propose, among others, a row-standardized inverse distance matrix based on the Euclidian distance between the nearest neighbors. Dubé et al. (2016) calculate the Euclidean distance separating all the establishments introducing a cut-off criterion, *i.e.*, a critical radius of influence. A number of models are estimated, each increasing the radius of influence by 250 meters (from 0 to 2250 meters) to calculate the number of establishments and total employment located in the direct vicinity of each establishment using a connectivity

[4]Nguyen et al. (2012) develops a relocation choice model using an average travel distance as a proxy for the distance among zones and firms.

matrix.

In light of these arguments, this paper seeks to contribute to the literature on location choice modeling, entreating a discussion on new directions for empirical explorations using more appropriate econometric tools and putting more consideration on the definition of distance itself.

3.2.2 Distance measures

The use of the Euclidean distance is widespread in economics (*e.g.*, Duranton and Overman, 2005; Partridge et al., 2008). This metric is known to all and experienced by all in everyday life, hence a prime candidate to capture economic interactions. According to Guy (1983), the use of straight-line distance to represent a travel function is unsatisfactory, however, though it does simplify computations. In the same line, Combes and Lafourcade (2003, 2005) claim that the Euclidean distance can only be regarded as a simplistic measure of the actual physical distance. The curvature of the Earth is the first source of systematic error. When calculating the straight-line (crow fly) distance between two remote points, the Euclidean distance may be replaced by the great circle distance, which takes the Earth's spherical shape into account (Axhausen, 2003). The second source of systematic error comes from the fact that in practice, people (or goods) rarely move from point A to point B along a straight line as assumed in the Fetter's "Law of Markets" (1924). For instance, car users may only drive on the existing road network, hence the well-known example of the Manhattan distance (Eaton and Lipsey, 1980). As the road network plays a central role in most urban transportation systems around the world, the Euclidean metric would therefore be inappropriate for the study of intra-urban location (Eaton and Lipsey, 1980; Perreur and Thisse, 1974).

Talen and Anselin (1998) suggest that distance can be computed in a variety of ways and use the actual street network in their empirical study of public facilities locations. Distance is measured by the means of a shortest path algorithm between the centroids of census tracts and the coordinates of public facilities of interest. They consider the shortest path network distance to be more relevant than the straight-line distance regarding the analyzed issue. Actual driving distances and travel times over a road network are perceived also by Boscoe et al. (2012) to be superior and substantially more precise than the straight-line distance. Previously, these metrics were considered as an expensive and labor-intensive to obtain. Nowadays, the commercial websites, such as Google, Yahoo!, Mapquest, Bing, Rand McNally, Michelin offer precise driving directions between nearly all locations in the developed world (Boscoe et al., 2012; Le Gallo and Dall'erba, 2008).

Given these considerations, an increasing number of works advocate the use of realistic distances based on a transport network over Euclidean and great circle alike (Combes and Lafourcade, 2005; Graham, 2007; Duran-Fernandez and Santos, 2014; Weisbrod, 2008; Faber, 2014; Kwon, 2002). This point is especially cogent when it comes to the location choice of economic establishments, for which the role of transportation infrastructures is well-known (Arauzo-Carod et al., 2010). In practice, several studies have compared whether and to what extent Euclidean and great circle distances differ from realistic transport distances. Boscoe et al. (2012) assess the extent to which travel time or network distance confers a genuine advantage over straight-line distance and identify locations where differences between the two are most pronounced. Chalasani et al. (2005) look at the differences between crow fly, shortest distance path, shortest time path, mean user equilibrium path distances, and the distance reported by the respondent, using data from three large-scale surveys carried in Norway and Switzerland. Rietveld et al. (1999) carry out a similar analysis in the Netherlands. In the case of France, Combes and Lafourcade (2005) compare the great circle distance, the network distance and the travel time, as well as an "economic distance". All works find strong correlations between transport distances and geographical distances for cross-sectional data (*i.e.*, at a given time). Yet, it is not a perfect correlation. This point will be carefully considered and will ultimately lead us to restrict the set of metrics used in our analysis.

3.3 Data

3.3.1 Distance matrices

This research considers several metrics, namely the Euclidean distance and various transport distances, to model establishment location choices. Distances are computed between the 1300 municipalities of the Paris region[5], hence matrices of size 1 690 000 (1300 by 1300).

[5]The Paris region, also called Île-de-France, is a vibrant and innovative region with over 5,6 million jobs, 37 percent of national executives, and 40 percent of the national workforce in research and development. It is the 1st R&D hub in Europe and the third worldwide. 11.7 million people representing over 19 percent of the country population live in an area which covers only 2.2 percent of national grounds. The GDP of the region amounts to 31 percent of total French GDP (http://www.grand-paris.jll.fr/fr/paris/chiffres-cles/) that is 612 billion euros (2012). The Paris region is the 1st European region considering the number of firms classified in Fortune 500 (July

The Euclidean distance matrix is computed between the centroids of each municipality using QGIS.

Transport matrices include the network distance and travel time matrices for the road network and the public transit network. These matrices are computed by means of the assignment models, one for each transport mode. Assignment models simulate the route choice behavior of users on a transport network. Congestion plays a major role in road models. As more drivers use the same road, it becomes more congested and travel time increases. Eventually, the travel time becomes so long that some drivers turn to alternative routes, which increases the traffic flow on the corresponding roads. This phenomenon develops until a traffic equilibrium - called Wardrop's equilibrium - is reached (Ortuzar and Willumsen, 2011). As far as public transit models are concerned, travel conditions typically include access and egress time, fare, waiting time, in-vehicle travel time, and transfer costs, which depend on the characteristics of the services that are used, such as frequency or speed. For some transit lines, congestion (Lapparent and Koning, 2015) and service reliability (Benezech and Coulombel, 2013) may also represent a non-negligible share of the generalized cost of travel. These are seldom considered in standard transit assignment models however, as introducing either of these items drastically increases the model complexity. Assignment models are primarily applied to determine the usage of road infrastructure or of transit routes for a given time period, typically the morning or evening peak periods. They can also serve to derive the shortest path between any origin-destination (O-D) pair to minimize the travel time, distance, and speed (Coulombel and Leurent, 2013). The variable to minimize when computing the shortest path is defined by the user. Unlike the shortest distance path, which only depends on the network geometry, the shortest time path - the most frequently used and the one included here - also depends on the network characteristics, firstly free-flow travel times and link capacities (Chalasani et al., 2005).

The road traffic and public transit assignment models are based on original models developed by the DRIEA Île-de-France (DRIEA Île-de-France 2008) and were adapted to run with the TransCAD software. Due to data availability issues, the two original assignment models were calibrated for different years, 2008 for the road model and 2009 for the transit model[6].

2014). The Paris region consists of 1300 municipalities encompassing the inner city of Paris and its suburbs. Yet, very large differences in population and employment densities are to be found between the Paris city and its outer periphery.

[6] Transport matrices being relatively stable over time at a regional scale, especially in the Paris region where

The road network, which comprises 65,692 links, includes all the main roads of the Paris region. The public transit network is even more radial than the road network, as most heavy transit lines pass through Paris, and is relatively undeveloped in the most remote areas of the Paris regions (Fig. 3.1), corroborating the observation of Chatman et al. (2016) that public transit networks tend to be sparser than road networks.

For each transport mode, two matrices are derived: the shortest travel time matrix, *i.e.*, the minimum travel time between each O-D pair, and the associated network distance matrix. Unlike Euclidean matrices, shortest travel time matrices are not symmetric. The time needed to go from A to B may differ from that needed to go from B to A because of, *e.g.*, one-way roads, asymmetric congestion patterns, or asymmetric travel conditions (*e.g.*, frequency) depending on the line direction in the case of transit (Fransen et al., 2015). Since the fastest path is not necessarily the shortest, the network distance computed is greater than the shortest path distance.

The transport matrices are computed between the centroids of the 1300 municipalities for the morning peak period, which is defined differently depending on the transport mode: from 6.50 a.m. to 9.10 a.m. for private vehicles and from 8 a.m. to 9 a.m. for public transit (DRIEA Île-de-France 2008). In the case of the road network, we also compute the travel time and network distance matrices under free-flow conditions, *i.e.*, when there is no congestion at all. The free-flow situation is used as a proxy for the off-peak period. We choose to consider it based on the hypothesis that travel conditions might be more relevant to some industries during the off-peak period if most of their deliveries and/or shippings are concentrated during this period.

3.3.2 Other data sources

Various data sources were compiled for the present study, drawn primarily from the Census survey of establishments carried out by the French National Institute of Statistics and Economic Studies. Data on the stock of establishments are measured at the 1st of January 2007. In our sample, 763 131 establishments are registered on the market at this date. The number of newly-created units amounts to 87 974 for year 2007. Data are pooled across activity sectors. For clarity and concision, this paper focuses on three sectors: special, scientific,

the transport networks are already well developed, this one year difference should have a very limited impact on our results.

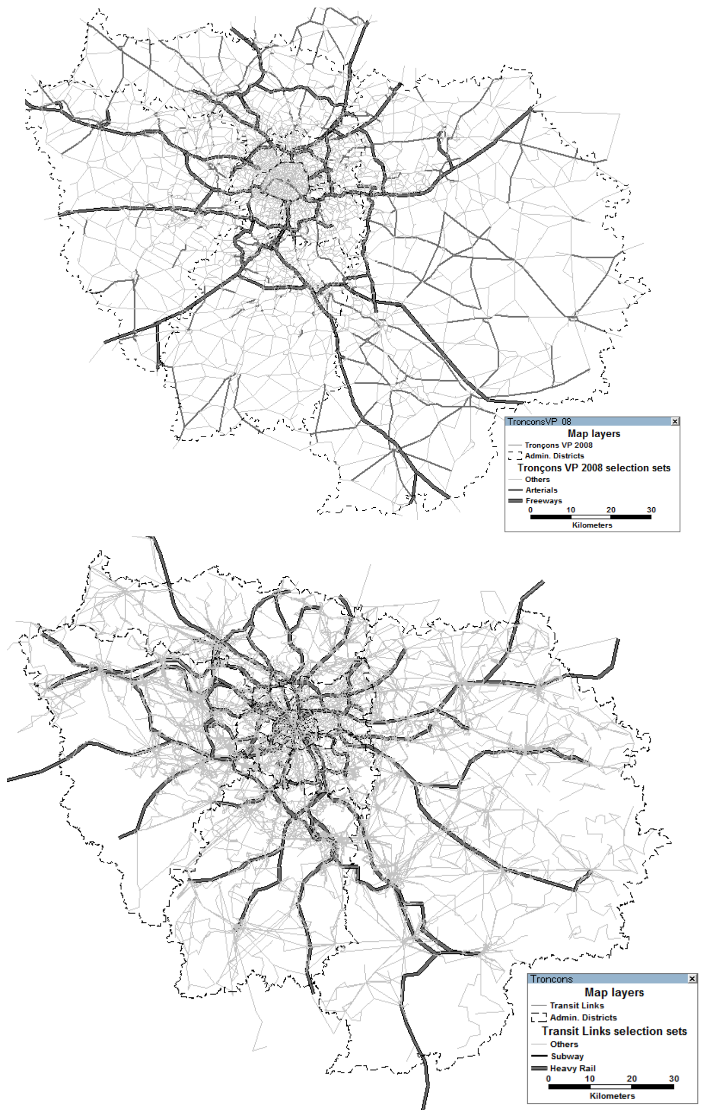


Figure 3.1: Road network (upper figure) and public transit network (lower figure) in the Paris region

and technical activities (SpecSci), construction (Constr), and real estate (RealEst). Among the newly-created establishments for year 2007, 17.4% (15 282 units) belong to the special, scientific and technical activities sector, 13.8% (12 115 units) to the construction sector, and 5.3% (4 683 units) to the real estate sector.

Detailed specification of other data used in the models that represent, among others, population and employment structure, proximity to retail, services, universities and schools, public transport and highways, and the levels of prices and taxes, along with their sources can be found in Buczkowska and Lapparent (2014). For concision, their presentation has been limited here to the summary Table 4.5.

Table 3.1: Description of explanatory variables used in the final models.

Variable and expected sign	Description
Establishments from respective sector (+)[a]	Number of pre-existing establishments from the analyzed sector within a particular municipality divided by the surface of municipality (km^2)[b]
Large establishments from all sectors (-)	Number of large pre-existing establishments with fifty employees or more divided by the surface of municipality (km^2)
Trips home-work (+)	Number of trips between home and work if a municipality is both a place of residence and a workplace to the total number of trips home-work
Trips home-work, intelectual workers (+)	Number of trips between home and work if a municipality is both a place of residence and a workplace to the total number of trips home-work made by white-collar workers
Offices (+)	Proportion of a municipality's surface dedicated to offices
Shops (+)	Proportion of a municipality's surface dedicated to shops
Vacant land (+)	Proportion of a municipality's vacant land available for new investments
Residential area (+)	Proportion of a municipality's land dedicated to the residential area
Universities and schools (+)	Proportion of a municipality's surface dedicated to universities and schools
Hospitals and clinics (nl)	Proportion of a municipality's surface dedicated to hospitals and clinics
Distance to highway (-)	Distance to the nearest highway (km)
Public transport (+)	Number of subway, train stations, and bus stops in a municipality
Residence tax (-)	Average level of residence taxes
Income per person (+)	Log value of the average income level per capita (euros)
Price of offices (-)	Log value of the average price level of offices per square meter (euros)
Price of shops (-)	Log value of the average price level of shops per square meter (euros)

[a](+) and (-) mean that the associated coefficient is expected to be positively or negatively statistically significant, respectively. (nl) means that no literature treats this problem or that no literature was reviewed on this issue. See, *e.g.*, a survey of Arauzo-Carod et al. (2010) which summarizes location determinants and main fundings of over fifty papers on establishment or firm location choice modeling.

[b]Data on the stock of establishments are given for the 1st of January 2007. The range for the independent variables is 2005-2009.

3.4 Model

3.4.1 Motivation of model choice

The Paris region is highly heterogeneous, especially regarding economic activity. While few municipalities host a large number of new establishments, others struggle to be chosen by any, and a large group of municipalities is left with no new entries. Based on the data for the Paris region aggregated at the municipality level, depending on the analyzed sector, the percentage of municipalities left with no new creation ranges from 34% up to 61%. The number of municipalities left with zero new entries in the construction sector equals to 439, in the real estate activities sector to 738, and in the special, scientific, technical activities sector to 569 out of 1300 possible municipalities that met no new establishments. These findings are similar to the remark made by Liviano-Solís and Arauzo-Carod (2014) based on the analysis of the Catalan data. The authors state that the distribution of entries is heavily skewed: a small group of municipalities meet the largest number of entries, while more than a half receive no entries at all. Municipalities range from small isolated villages in rural areas to huge and densely populated cities.

When the observed data display a higher fraction of zeros than would be typically explained by the standard count data models, the zero inflated or hurdle models can be suggested. This paper seeks to address the complaint voiced by Liviano-Solís and Arauzo-Carod (2013) and Bhat et al. (2014) who notice that heretofore scholars have not fully explored the hurdle model technique when analyzing location phenomena. It is thereby in line with Liviano-Solís and Arauzo-Carod (2013) and Buczkowska and Lapparent (2014). In contrast to the two previous works, which focus on determining which type of model (Poisson, negative binomial, zero-inflated versions of these models, hurdle-Poisson, and hurdle negative-binomial) best fits their data, this paper focuses instead on the issue of which type(s) of spatial representation(s) best characterizes spatial spillovers. Considering all these points, a discrete mixture of hurdle-Poisson models is developed, wrapping the spatial metrics in a common statistical framework of analysis.

3.4.2 Model specification

Contingently on a type m of spatial metric, the likelihood function is built up on a hurdle-Poisson count data model:

$$\begin{aligned}
&\ell\left(\delta_l, y_l | x_{l,m}; \boldsymbol{\theta}_{1,m}, \boldsymbol{\theta}_{2,m}\right) = \\
&\quad \left(1 - p\left(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m}\right)\right)^{1-\delta_l} \times \\
&\quad \left(p\left(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m}\right) h\left(y_l | y_l > 0; x_{l,m}; \boldsymbol{\theta}_{2,m}\right)\right)^{\delta_l},
\end{aligned} \tag{3.1}$$

where

$$\delta_l = \begin{cases} 0 \text{ if } y_l = 0 \\ 1 \text{ otherwise} \end{cases}. \tag{3.2}$$

$\forall l, y_l \in \mathbb{N}$ is the number of new establishments that locate at l. $x_{l,m}$ is a vector of independent variables that characterize location l using spatial metric m. p and h are function that will be defined below. $\boldsymbol{\theta}_m := \left[\boldsymbol{\theta}'_{1,m}, \boldsymbol{\theta}'_{2,m}\right]'$ is a vector of parameters to estimate when the spatial metric type is m.

The probability that at least one establishment settles in location l is based on a latent profit variable: establishments locate at l as long as local profit is not exhausted. The local profit function is defined as a linear combination of observed and unobserved variables:

$$\Pi\left(x_{l,m}; \boldsymbol{\theta}_{1,m}\right) = x'_{l,m}\boldsymbol{\theta}_{1,m} + \varepsilon_{l,m}. \tag{3.3}$$

The error terms $\varepsilon_{l,m}$ are assumed to be *iid* according to a logistic distribution with a location parameter equal to 0 and a scale parameter equal to 1. The probability to observe one or more new establishments locating at l is then given by:

$$p\left(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m}\right) = \frac{1}{1 + \exp\left(-x'_{l,m}\boldsymbol{\theta}_{1,m}\right)}. \tag{3.4}$$

When the number of new establishments that locate at l is strictly positive, the probability to observe a number $y_l > 0$ of establishments at l is defined as a truncated-at-zero Poisson distribution:

$$h\left(y_l | y_l > 0; x_{l,m}; \boldsymbol{\theta}_{2,m}\right) = \frac{\lambda\left(x_{l,m}; \boldsymbol{\theta}_{2,m}\right)^{y_l} \exp\left(-\lambda\left(x_{l,m}; \boldsymbol{\theta}_{2,m}\right)\right)}{y_l!\left(1 - \exp\left(-\lambda\left(x_{l,m}; \boldsymbol{\theta}_{2,m}\right)\right)\right)}, \tag{3.5}$$

where the rate of occurrence is parametrically defined as:

$$\lambda\left(x_{l,m};\boldsymbol{\theta}_{2,m}\right)=\exp\left(x'_{l,m}\boldsymbol{\theta}_{2,m}\right). \tag{3.6}$$

3.4.3 Spatial spillovers

The observed explanatory variables relate to the location l and/or the sector s, including the stock of (*i.e.*, number of pre-existing) establishments in the sector s. To reduce notation clutter, the argument s will be omitted in the remainder of the text. According to Dubé et al. (2016), the final locational decision of an establishment should be related to the surrounding economic landscape. Following this remark, the characteristics of the surrounding areas and the stock of establishments located nearby when modeling the spatial spillovers are accounted for as follows:

$$x_{l,m}=\ln\left(\sum_{j=1}^{L}e^{-\mu d^{m}_{l,j}}z_j\right), \tag{3.7}$$

where z_j is a vector of attributes characterizing the municipality j (including the stock of establishments in the considered sector s, as well as in the other sectors). The parameter μ is fixed to one[7] while $d^{m}_{l,j}$ is the distance between the centroids of municipalities l and j according to the metric m.

The metric m influences how the surrounding economic landscape is considered through the exponential decay function. The relative influence of a given zone j on the vector $x_{l,m}$ varies depending on how close zone j is to zone l according to the metric m. For instance, if access to zone j is made difficult during rush hour because of congestion, zone j will have less influence on the vector $x_{l,m}$ if m is the congested travel time metric than if m is the Euclidean metric or the free-flow travel time metric.

3.4.4 Full information maximum (log-)likelihood function

Considering the M types of metrics together, the vector $\pi=(\pi_1,\cdots,\pi_M)$ defines the probabilities to belong to each class associated to the corresponding metric. These probabilites sum up to 1, $\sum_{m=1}^{M}\pi_m=1$. The full information maximum likelihood estimator (FIMLE) is based on maximizing the following marginal log-likelihood function with respect

[7] One can play with the parameter μ setting its value at the level smaller or larger than 1 for more global or more local spillover effects. This is left aside for the further research.

to unknown parameters π and $\boldsymbol{\theta}$ conditionally to observed data $x_{\cdot,l} = (x_{1,l}, \cdots, x_{M,l})$:

$$\ell(\boldsymbol{\theta}, \pi | y_{\cdot}, x_{\cdot,\cdot}) = \sum_{l=1}^{L} \ln \left(\sum_{m=1}^{M} \pi_m \ell(\delta_l, y_l | x_{m,l}; \boldsymbol{\theta}_{1,m}, \boldsymbol{\theta}_{2,m}) \right). \tag{3.8}$$

3.4.5 Partial effects

For a discrete mixture of hurdle-Poisson models, partial effects are simply defined as the discrete mixture of the conditional hurdle-Poisson partial effects. For instance, the expected number of establishments that locate at l is a discrete mixture of the expectations of different hurdle-Poisson models:

$$\mathbb{E}(y_l | x_{l,\cdot}; \boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m \frac{p(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m})}{1 - \exp(-\lambda(x_{l,m}; \boldsymbol{\theta}_{2,m}))} \lambda(x_{l,m}; \boldsymbol{\theta}_{2,m}). \tag{3.9}$$

As in the standard hurdle-Poisson model, this allows for a straightforward decomposition of the overall effect into an effect at the extensive margin and an effect at the intensive margin. Consider a variable z_l that characterize l and that is then transformed using a metric m. The effect on the expected number of new establishments that locate at l with respect to a variation of it is defined as:

$$\begin{aligned} &\frac{\partial \mathbb{E}(y_l | x_{l,\cdot})}{\partial z_l} = \\ &\sum_{m=1}^{M} \pi_m \frac{\partial p(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m})}{\partial z_{l,m}} \mathbb{E}(y_l | y_l > 0, x_{l,m}; \boldsymbol{\theta}_{2,m}) + \\ &\sum_{m=1}^{M} \pi_m p(y_l > 0 | x_{l,m}; \boldsymbol{\theta}_{1,m}) \frac{\partial \mathbb{E}(y_l | y_l > 0, x_{l,m}; \boldsymbol{\theta}_{2,m})}{\partial z_{l,m}}. \end{aligned} \tag{3.10}$$

Derivation of direct and cross elasticities and other partial effects are in the same vein: they are defined as discrete mixtures of the associated conditional elasticities and partial effects.

3.4.6 Posterior class assignment probabilities

Once the model estimated, one may compute posterior class assignment probabilities, *i.e.*, the probability of metric m contingently on location l:

$$\psi_{j|l} = \frac{\pi_j \ell(\delta_l, y_l | x_{j,l}; \boldsymbol{\theta}_{1,j}, \boldsymbol{\theta}_{2,j})}{\sum_{m=1}^{M} \pi_m \ell(\delta_l, y_l | x_{m,l}; \boldsymbol{\theta}_{1,m}, \boldsymbol{\theta}_{2,m})}. \tag{3.11}$$

By doing so, one may update the knowledge about which spatial representation is appropriate (in the sense of correctly accounting for spatial spillovers) for each location l using observed aggregate choices of establishments .

3.5 Results

3.5.1 Preselection of metrics based on correlation analysis

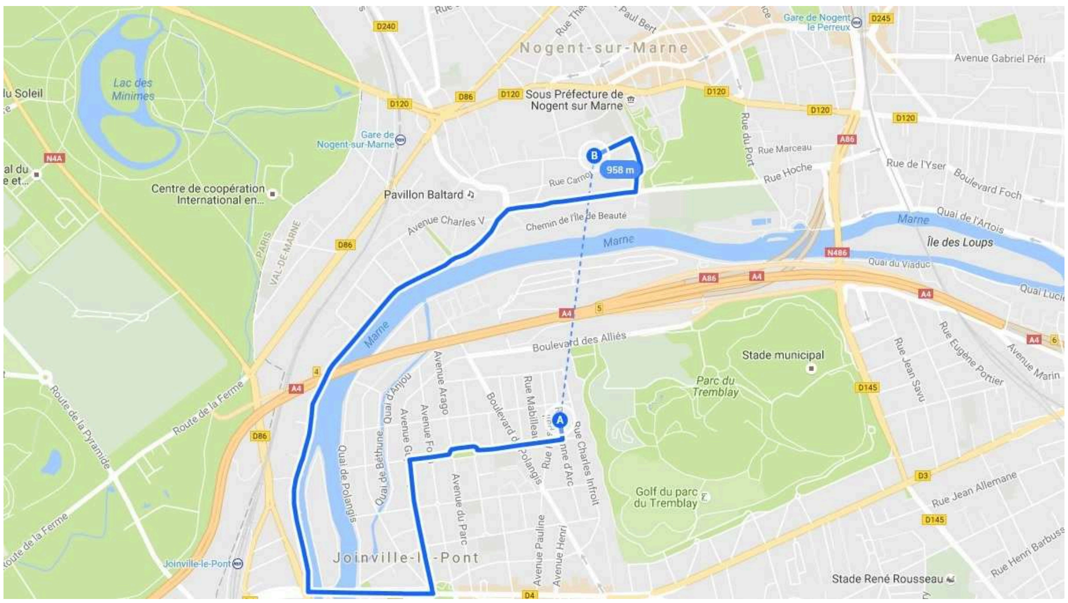


Figure 3.2: Example 1. Google map image: Possible barriers which require to make detours: a river, islands, a lake, parks, a golf course, stadiums, highways, a national road, departmental roads, train routes. Euclidean (958 m) versus real distance (4.0 km and 12 min) from point A to B in the area of Boullereaux-Champigny, the Paris region. Source: Google My Maps

A preliminary analysis of correlation between the various distance measures (Table 3.2) confirms in the case of the Paris area the main findings from the literature review (subsection 3.2.2). The Euclidean distance (ED) is strongly correlated with road network distances, with a correlation coefficient of 0.987 for the morning peak period (DistVhMph) and 0.988 under free-flow conditions (DistVhFlow). By comparison, Combes and Lafourcade (2005) find a correlation coefficient of 0.990 between the great circle distance and the road network distance, Rietveld et al. (1999) a value of 0.966 between the Euclidean distance and the road net-

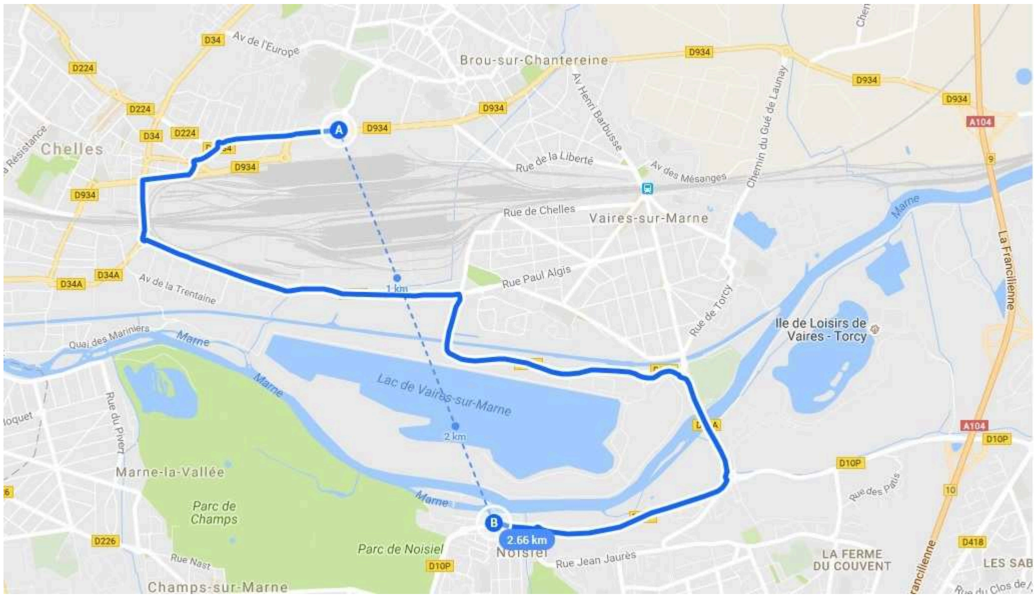


Figure 3.3: Example 2. Google map image: Possible barriers which require to make detours: a river, lakes, an attraction island theme park, parks, a highway, departmental roads, train routes, electric and thermic centers, cemetery. Euclidean (2.66 km) versus real distance (8.2 km and 13 min) from point A to B in the area of Champs-sur-Marne, the Paris region. Source: Google My Maps

work distance, and Duran-Fernandez and Santos (2014) find correlation levels between 0.838 and 0.977 depending on the considered metrics. Duran-Fernandez and Santos (2014) consider two metrics based on the great circle formula and two other metrics based on a restricted and unrestricted road network distance (transit). The former two metrics are the average distance in kilometers between metropolitan area in a given state and between the geographic centroids of each state. The latter two metrics are the transit network times along the route with and without mandatory rest after 11 continuous hours of service, and the network length associated to the route that minimizes travel time.

The level of correlation falls markedly with distance, however. For municipalities distant by less than 10 km (according to the ED), the correlation between ED and DistVhMph and between ED and DistVhOph are equal to 0.869 and 0.868, respectively (Table 3.3). But the same values drop to 0.541 and 0.586 for municipalities which are 40 to 50 km distant. The mean detour factor - firstly defined by Cole and King (1968) as the ratio between the network distance and the straight-line distance, with computations done by students tracing road ways on paper maps[8] - is in our analysis equal to 1.287 when there is no congestion. It is slightly higher for the morning peak period (1.294), reflecting the fact that individuals make additional detours to avoid congestion. Results are quite different regarding the public transit network. The overall level of correlation with ED falls to 0.635, with a mean detour factor of 1.624. This stresses the fact that the public transit network is both sparser than the road network, especially in suburban and rural areas, and more radial (hence a higher detour factor).

The comparison of ED with travel times leads to similar conclusions. For the road network, the correlation between the ED and travel times equals to 0.948 for the morning peak period (TtVhMph) and 0.973 for free-flow conditions (TtVhOph), against 0.974 for Combes and Lafourcade (2005) and 0.947 for Rietveld et al. (1999). For the public transit network, the correlation coefficient is lower and equal to 0.708. Last, correlation coefficients are again significantly lower when computed by increasing distance interval, and sharply decrease with distance.

All in all, road transport distance measures (travel time and network distance alike) turn out to be strongly correlated with the Euclidean distance at first glance, but less so when disaggregating O-D pairs by distance range. While the corresponding values are not reported here, for brevity, road network distances (with and without congestion) are strongly correlated with each other, even for a given distance interval, and are also relatively strongly correlated with their associated travel times. On the contrary, congested and free-flow travel times are less correlated between each other, especially so when considering distance intervals. Road network distances are therefore discarded in the subsequent analysis. Similarly, public transport distance measures (again, network distance and travel time alike) have been tested in several model specifications. Cases involving public transit network for the sectors of construction, real estate, and special, scientific, technical activities mix Euclidean distance with travel time for public transit and travel time for private vehicles under road congestion with travel time for public transit. These cases do not assure the convergence of the model or yield poor results. In all the cases, the optimization could not be completed or could not improve the function value. Cases related to the public transit network have been hence discarded. Two facts might account for the poor performance of the public transit network metrics. First, establishments might focus on road travel conditions because it is the predominant transport

[8] Cole and King (1968) report typical values of 1.2 to 1.6 for rural areas in various parts of Britain, while Boscoe et al. (2012) find a nationwide detour index of 1.417.

Table 3.2: Pair correlations between Euclidean distance and transport distances.

Basic statistics	ED[a]	DistVhMph	DistVhOph	DistTcMph	TtVhMph	TtVhOph	TtTcMph
Mean	55.44	70.72	70.27	105.97	64.91	51.25	195.17
Std. dev.	28.28	34.92	34.55	68.16	28.10	21.85	145.28
#Observations	1690 000	1690 000	1690 000	1690 000	1690 000	1690 000	1690 000
Min	0	0	0	0	0	0	0
Max	157.56	195.29	195.96	239.28	156.05	126.03	509.77

Distance	ED	DistVhMph	DistVhOph	DistTcMph	TtVhMph	TtVhOph	TtTcMph
ED	1						
DistVhMph	0.987 (0.00002)[b]	1					
DistVhOph	0.988 (0.00002)	0.992 (0.00001)	1				
DistTcMph	0.907 (0.00015)	0.895 (0.00017)	0.905 (0.00015)	1			
TtVhMph	0.948 (0.00008)	0.959 (0.00007)	0.953 (0.00008)	0.829 (0.00026)	1		
TtVhOph	0.973 (0.00004)	0.975 (0.00004)	0.978 (0.00004)	0.908 (0.00015)	0.960 (0.00007)	1	
TtTcMph	0.708 (0.00042)	0.695 (0.00043)	0.706 (0.00042)	0.834 (0.00026)	0.638 (0.00050)	0.740 (0.00038)	1

[a] Euclidean distance (ED), travel times by car for the peak (TtVhMph) and off-peak periods (TtVhOph), travel time by public transit during the morning peak period (TtTcMph), and the corresponding distances (DistVhMph, DistVhOph, and DistTcMph, respectively).

[b] Standard errors reported in the parentheses.

mode for freight or for intrametropolitan business trips. Second, the strong spatial irregularities of public transport metrics, in particular in the most distant parts of the metropolitan area, might make them unsuitable to model spatial spillovers.

3.5.2 Parameter estimates and model performance

In order to demonstrate the relevance and usefulness of our approach, we proceed in two main steps. First, we examine the parameter estimates in order to verify that the mixture model does provide sensible results in light of the literature, and that said estimates are not overly sensitive to the model specification. Second, we confirm that the mixture model does perform better than ”pure”, mono-distance models.

Table 3.3: Correlations between Euclidean distance and transport distances by increasing range intervals (in km).

Distance	0-10	10-20	20-30	30-40	40-50	50-60	60-70
DistVhMph	0.869	0.776	0.676	0.599	0.541	0.510	0.486
	(0.00116)[a]	(0.00118)	(0.00134)	(0.00146)	(0.00158)	(0.00170)	(0.00190)
DistVhOph	0.868	0.785	0.696	0.633	0.586	0.555	0.517
	(0.00118)	(0.00113)	(0.00127)	(0.00137)	(0.00147)	(0.00159)	(0.00182)
DistTcMph	0.355	0.323	0.311	0.283	0.244	0.207	0.193
	(0.00416)	(0.00264)	(0.00223)	(0.00210)	(0.00211)	(0.00220)	(0.00240)
TtVhMph	0.701	0.488	0.382	0.320	0.290	0.278	0.279
	(0.00243)	(0.00225)	(0.00211)	(0.00204)	(0.00205)	(0.00212)	(0.00229)
TtVhOph	0.763	0.634	0.541	0.469	0.428	0.401	0.377
	(0.00199)	(0.00176)	(0.00175)	(0.00178)	(0.00183)	(0.00193)	(0.00213)
TtTcMph	0.354	0.215	0.179	0.146	0.117	0.101	0.088
	(0.00417)	(0.00281)	(0.00239)	(0.00223)	(0.00221)	(0.00227)	(0.00247)

[a] Standard errors reported in the parentheses.

The hurdle-Poisson mixture model with two latent classes is estimated for the three selected sectors: (1) special, scientific, and technical activities, (2) construction, and (3) real estate. Three various combinations of metrics are tested. The first mixture (case #1) uses the Euclidean distance (ED) and the peak (road) travel time (TtVhMph) to capture spatial spillovers. In the second mixture (case #2), the first class is based on the off-peak travel time (TtVhOph) and the second class is again based on the peak travel time. In the third and last mixture (case #3), the Euclidean distance (ED) is combined with the off-peak travel time (TtVhOph). For concision, not all cases are discussed to focus on the most illustrative ones instead. Parameter estimates of the probabilistic mixture of hurdle-Poisson models for the construction sector are presented in Table 3.4. For the two other sectors (special, scientific, technical activities and real estate) the parameter estimates for case #1 (ED with TtVhMph), are reported in Table 3.6.

We first present the results for the truncated-at-zero Poisson parts of the mixture models. There is a perpetual debate in the literature about Jacobs externalities (Jacobs, 1969) versus Marshall-Arrow-Romer externalities (Marshall, 1920; Arrow, 1962; Romer, 1986), *i.e.*, diversity versus specialization matter the most for the location choices of new establishments. MAR externalities describe a set of positive economic externalities that arise from the concentration of activities belonging to the same industry. Certain industries benefit from economic specialization (co-location within an industry). Specialization allows establishments to lower transportation time and costs from suppliers, sharing infrastructures and equipment, to benefit from knowledge spillovers, and to facilitate communication between them (Marshall, 1920; Arrow, 1962; Romer, 1986; Dubé et al., 2016). For the present case study, marked localization patterns are observed for all considered sectors. The greater the number of establishments from a given sector, the greater the number of newly-created units of this sector locating within the same area, whatever the metric considered (Table 3.4, Table 3.6). As stated by Jacobs (1969), some advantages can also arise through the clustering of different sectors and economic activities. To limit the model complexity, the issue of diversity is not considered here, however[9].

The presence of large establishments tends to repel new establishments. In the case of establishments that belong to special, scientific and technical activities sector, this result corresponds to the findings of Kang and Dall'erba (2016), that an exceeding presence of large establishments lessens knowledge creation. This suggests that smaller establishments might be more efficient at innovating (Acs et al., 1994).

Price levels for shops or offices can be treated as a proxy for the average price an establishment needs to pay to set up on the market. High real estate prices (of either shops or offices depending on the sector considered) also deter new establishments from settling in the area, which is conform to economic intuition.

Transport accessibility seems to play a role in the location choices of newly-created establishments. Establishments from the construction and special, scientific and technical activities sectors seek proximity to the highway network as well as to public transit stations. As far as real estate establishments are concerned, close proximity to public transport is an important criterion whereas proximity to the highway network did not turn out to be significant. One possible interpretation is that the real estate establishments act more locally and settle preferentially in dense areas with good access to public transit, with customers more prone to come by foot or by public transit than using the highway.

Looking at sector specific effects, high rates of residence tax appear to discourage the creation of units in the construction and real estate sectors. These also seek to locate nearby shops and offices. Establishments from the construction sector favor proximity to public amenities, such as schools, universities, hospitals and clinics. They also prefer municipalities where people both live and work assuring a fine level of vehicular and pedestrian traffic passing by the

[9]Readers interested in this issue may refer to Buczkowska and Lapparent (2017).

site.

The role of intraregional knowledge spillovers is rarely explicitly explored in the literature. The level of knowledge spillovers depends, among others, on close links between business, academic, and government sectors (Kang and Dall'erba, 2016). Private and university research laboratories tend to be the main institutions of innovation and have a positive influence on knowledge creation. Jaffe (1989) measures the coincidence between the location of a university and of industrial research activities and uses his measurements as a proxy for knowledge spillovers due to academic research. In this paper, private and academic knowledge stocks are proxied by home-work flows of intellectual workers, universities, and offices. Good access to the intellectual workforce, the presence of academic establishments and offices increases the probability of new creation of establishments dedicated to special, scientific and technical activities. This corroborates the findings of Kang and Dall'erba (2016).

Last, the presence of high-income households increases the probability of new establishments from the real estate sector setting in the area, thus taking into account some of the population features, such as the purchasing power. It may be worth having in mind that a particular site may produce more total traffic, but another location which produces more of the desired traffic will be chosen.

Next, the parameter estimates of the hurdle parts of the mixture models are reported. They describe which factors influence the fact that at least one new establishment settles in a given municipality. As such, only sufficient number of nearby public amenities, a good enough access to the public transport, and level of flows of people seem to assure an establishment engaged in construction to locate in a particular municipality. The presence of establishments from the same sector in the vicinity increases the probability that at least one establishment settles in a municipality. Conversely, large establishments or high real estate prices act as deterrents to the implantation of new establishments. One may expect that the amount of vacant land increases the probability that at least one establishment in the construction sector locates in the municipality. The sign of the associated parameter should consequently be positive, which is the case for class #2 (TtVhMph) but not for class #1 (ED).

In case of the special, scientific and technical activities sector, economic specialization, a good accessibility to the intellectual workforce, the presence of offices and universities in the vicinity, and an easy access to the public transport turn out to increases the probability that at least one establishment locates in a particular municipality. The estimates of the hurdle parts of the location choice model for units in the real estate sector confirm that co-location, the mixture of shops and offices in a particular municipality and its surroundings, land dedicated to the residential use play an important and positive effect and high real estate prices have a negative influence on implantation of new establishments.

To check the robustness of our model, we investigate whether the parameter estimates are sensitive to: (1) the choice of the other metric for a given metric and (2) to the fact of using a mixture model, as opposed to a "pure, mono-distance" model. To investigate point (1), the parameter estimates of the positive count (or truncated-at-zero Poisson) parts for the construction sector are compared for two mixture models: cases #1 and #2. Both share the same metric - the congested travel time (TtVhMph) - for the second latent class. The first class is based on the Euclidean distance in case #I, and on the off-peak travel time in case #II. The parameter estimates tend to behave in relatively similar ways for most model variables (Fig. 3.4). This result is critical; otherwise, sensitivity of estimates could be a good reason to consider spatial models as ill-conditioned. In particular, the parameter estimates for the second latent class (TtVhMph) are very little sensitive to the metric choice for the other latent class (ED or TtVhOph), again corroborating the robustness of the results.

To examine (2), two pure mono-distance hurdle-Poisson models are run independently from each other. The first one uses the Euclidean distance to account for spillover effects, while the second one uses the peak travel time. By comparing the results line by line in Table 3.4, the parameter estimates are relatively similar for the mixture model and for the pure models. Again, this tends to confirm the robustness of our results, and that using a mixture model does not introduce any major misspecification of any sort.

Last, the Bayesian Information Criterion (BIC) is used to compare the performance of the mixture model with that of the pure models. All models use the same set of variables and the same number of observations (1300). However, the number of parameters does double when using the mixture of hurdle-Poisson models (48 parameters) in comparison to the pure mono-distance HP models (24 parameters), hence the choice of the BIC to compare the relative performances of the various models. Whatever the sector considered, the mixture model performs significantly better than the pure hurdle-Poisson models based on a single metric. For the construction sector, the BIC is equal to 5692.37 for the mixture model, against 7870.60 and 7856.60 for the pure models based on ED and TtVhMph, respectively (Table 3.4). For the real estate sector, the BIC are respectively equal to 3628.71 for the mixture model, and to

Table 3.4: Probabilistic mixture of hurdle-Poisson models for the construction sector that uses two latent classes, case #1: ED (class #1) and TtVhMph (class #2) versus two "mono-distance" models using first ED and then TtVhMph.

	Mixture of two HP models Estimate Class #1, Matrix used: ED	(t-st)	1st mono-distance HP model Estimate Matrix used: ED	(t-st)
	Hurdle: Zero counts		Hurdle: Zero counts	
Constant	-4.267	(-0.48[a])	13.792	(7.19) ***
Same sector estab.[b]	1.341	(1.49)	0.287	(2.07) **
Large estab., all sectors	0.341	(1.05)	-0.031	(-0.80)
Trips home-work	-1.223	(-0.92)	0.984	(4.70) ***
Shops and offices	-0.211	(-0.94)	0.066	(1.60)
Vacant land	-2.895	(-1.74) *	0.023	(0.27)
Universities and schools	0.609	(1.06)	0.402	(4.88) ***
Hospitals and clinics	-0.677	(-1.38)	-0.029	(-0.98)
Distance to highway	-2.77	(-1.76) *	-0.21	(-2.31) **
Public transport	0.111	(0.48)	0.092	(3.01) ***
Residence tax	6.973	(2.06) **	1.135	(4.53) ***
Price of shops (log)	-7.331	(-1.72) *	-4.266	(-6.98) ***
	Poisson: Positive counts		Poisson: Positive counts	
Constant	23.194	(40.7) ***	22.375	(68.41) ***
Same sector estab.	1.296	(24.28) ***	1.553	(45.47) ***
Large estab., all sectors	-0.114	(-4.00) ***	-0.256	(-13.99) ***
Trips home-work	2.022	(30.07) ***	1.689	(46.49) ***
Shops and offices	0.115	(4.59) ***	0.355	(21.49) ***
Vacant land	0.077	(3.00) ***	-0.04	(-2.61) ***
Universities and schools	0.559	(12.96) ***	0.45	(18.28) ***
Hospitals and clinics	0.055	(3.13) ***	0.024	(2.44) **
Distance to highway	-0.136	(-4.96) ***	-0.159	(-10.06) ***
Public transport	0.098	(6.16) ***	0.222	(19.55) ***
Residence tax	-0.381	(-4.07) ***	-0.185	(-3.46) ***
Price of shops (log)	-4.342	(-26.63) ***	-4.633	(-53.34) ***
			Convergence criterion	Satisfied
			#Parameters	2 x 12
			#Observations	1300
	... continuation of mixture of HP models below		Log-likelihood	-3849.25
			AIC	7746.5
			AICC	7747.4
			BIC	7870.6
	Class #2. Matrix: TtVhMph		2nd mono-distance HP model Estimate Matrix: TtVhMph	(t-st)
	Hurdle: Zero counts		Hurdle: Zero counts	
Constant	29.572	(4.28) ***	15.779	(3.30) ***
Same sector estab.	0.344	(2.18) **	0.174	(2.63) ***
Large estab., all sectors	-0.096	(-2.03) *	0.009	(0.35)
Trips home-work	1.53	(4.71) ***	0.778	(4.82) ***
Shops and offices	0.166	(3.59) ***	0.08	(3.10) ***
Vacant land	0.424	(4.05) ***	0.046	(0.74)
Universities and schools	0.275	(3.14) ***	0.062	(1.54)
Hospitals and clinics	0.103	(2.46) **	0.008	(0.37)
Distance to highway	0.046	(0.37)	-0.167	(-2.13) **
Public transport	0.135	(3.43) ***	0.072	(3.35) ***
Residence tax	0.135	(0.40)	1.124	(5.52) ***
Price of shops (log)	-8.892	(-2.89) ***	-6.222	(-2.84) ***
	Poisson: Positive counts		Poisson: Positive counts	
Constant	19.842	(8.21) ***	16.471	(33.97) ***
Same sector estab.	1.14	(16.34) ***	1.147	(48.29) ***
Large estab., all sectors	-0.092	(-4.58) ***	-0.087	(-8.84) ***
Trips home-work	1.385	(24.88) ***	1.47	(53.28) ***
Shops and offices	0.245	(12.21) ***	0.166	(18.41) ***
Vacant land	0.002	(0.12)	-0.001	(-0.15)
Universities and schools	0.476	(12.83) ***	0.432	(27.69) ***
Hospitals and clinics	0.056	(5.39) ***	0.039	(7.08) ***
Distance to highway	-0.083	(-4.94) ***	-0.089	(-9.73) ***
Public transport	0.08	(7.22) ***	0.101	(20.67) ***
Residence tax	-0.213	(-2.68) **	-0.229	(-6.11) ***
Price of shops (log)	-4.107	(-3.96) ***	-2.494	(-11.79) ***
	Pi = 0.322	(13.70) ***	Convergence crit.	Satisfied
	Convergence crit.	Satisfied	#Parameters	2 x 12
	#Parameters	2 x 12 x 2	#Observations	1300
	#Observations	1300	Log-likelihood	-3842.30
	Log-likelihood	-2790.61	AIC	7732.6
	BIC	5692.4	AICC	7733.5
			BIC	7856.6

[a] ***, **, * represent statical significance at the 1%, 5%, and 10% level, respectively.

[b] See Table 4.5 for the variables description.

4497.30 and 5079.90 for the pure models. Last, for the special, scientific, and technical activities sector, the BIC of the mixture model is 5270.61, which is markedly lower than the BIC for the pure models, being equal to 8839.90 and 9090.10, respectively. To conclude, for all three considered sectors the mixture of hurdle-Poisson models performs significantly better than pure hurdle-Poisson models that use a single distance matrix to incorporate spatial spillovers. The significance and sign of the parameter estimates are generally conform to the literature, corroborating that the methodology is sound and that the results are robust. The consideration of alternative metrics to the Euclidean distance even improves the results of the hurdle part in the case of the construction sector and of the special, scientific and technical activities sector. On the other hand, for two variables, the availability of vacant land and the proximity to residential areas, the signs turned out to be not always positive (if still significant) - as one might expect them to be based on the literature review - depending on the metric used. These variables were found to have little effect on the objective function, however, and consequently on the location choices of the establishments. Still, this point should be further investigated in the future. Last, two sensitivity tests - relative to the choice of the other metric, and to the mixture per se - have further corroborated the robustness of the results.

3.5.3 Which metric(s) to choose?: Comparison of metric mixtures

The relevance and soundness of the methodology being established, we may now tackle our main research question, that is which distance measure(s) is/are the most appropriate to capture spatial spillovers in the location choice of new establishments. Again, the focus is on the same three sectors and the same three mixtures as in subsection 3.5.2.

- ⋆ Case #1. Probabilistic mixture of hurdle-Poisson models with two latent classes: Euclidean distance (class #1, ED) and peak period road travel time (class #2, TtVhMph).

- ⋆ Case #2. Probabilistic mixture of hurdle-Poisson models with two latent classes: Off-peak period road travel time (class #1, TtVhOph) and peak period road travel time (class #2, TtVhMph).

- ⋆ Case #3. Probabilistic mixture of hurdle-Poisson models with two latent classes: Euclidean distance (class #1, ED) and off-peak period road travel time (class #2, TtVhOph).

For the construction sector, the estimated class assignment probability π is equal to 0.322 in case #1 (ED with TtVhMph) and 0.319 in case #2 (TtVhOph with TtVhMph). In case #3 (ED with TtVhOph), optimization could not be completed. For this sector, the peak travel time is the most likely to adequately capture spatial spillovers. A similar conclusion can be reached for establishments dedicated to special, scientific and technical activities, with π equal to 0.283 and 0.222 in case #1 and #2, respectively. In case #3, π is equal to 0.278 which indicates that the off-peak travel time prevails over the Euclidean distance. For the real estate sector, the value of π is equal to 0.522 when the spatial weight matrices are based on ED and TtVhMph, 0.639 with TtVhOph and TtVhMph, and 0.525 with ED and TtVhOph.

Looking now at log-likelihoods, these are always minimized in case #2. The combination of the two road travel time measures (peak and off-peak) outperforms the two other combinations for all three sectors. Stated differently, the Euclidean distance would be less suited to capture spatial spillovers than transport distances (namely travel times) in our case study. As a corollary, our results tend to show that considering both the peak and off-peak travel conditions proves worthwhile to better understand establishments' location choices. Because the differences in log-likelihood remain relatively minor, these findings should be considered with caution, however, and ought to be confirmed with other case studies.

For the first two sectors, the predominance of peak travel time (as opposed to the off-peak travel time or the Euclidean distance) most likely underlines the importance of road travel conditions either for work operations, or to ensure a smooth commute to workers. On the other hand, the slight predominance of the Euclidean distance and of the off-peak travel time in real estate tends to emphasize that spatial spillovers are channeled not only through the road mode but also through other modes, such as walk, public transit, or some other communication modes.

3.5.4 Posterior class probabilities

As indicated in the econometric model (Section 3.4), it is useful to compute the posterior probabilities for each of the 1300 municipalities of the Paris region. For the sake of concision, case #1 (ED and TtVhMph) for all the three sectors, special, scientific and technical activities, construction, and real estate is reported. The results are presented in Fig. 3.5. Overall, it is clear that the peak road travel time prevails in more municipalities for the construction and special, scientific and technical activities sectors, while the situation is more mixed for the

Table 3.5: Estimated posterior probability of belonging to class #1 (`Pi`): Comparison across cases and sectors.

Sector[b]	Case #1[a] Mixture of two HP models Class #1, Matrix used: ED Class #2, Matrix used: TtVhMph Convergence crit.	π[c]	Indication	#Param.	#Obs.	Log-lik.
Constr	Satisfied	0.322***[d]	TtVhMph	2x12x2+1=49	1300	-2790.61
SpecSci	Satisfied	0.283***	TtVhMph	2x9x2+1=37	1300	-2593.62
RealEst	Satisfied	0.524***	ED	2x9x2+1=37	1300	-1772.71

Sector	Case #2 Mixture of two HP models Class #1, Matrix used: TtVhOph Class #2, Matrix used: TtVhMph Convergence crit.	π	Indication	#Param.	#Obs.	Log-lik.
Constr	Satisfied	0.319***	TtVhMph	2x12x2+1=49	1300	-2770.42
SpecSci	Satisfied	0.222***	TtVhMph	2x9x2+1=37	1300	-2560.34
RealEst	Satisfied	0.639***	TtVhOph	2x9x2+1=37	1300	-1754.76

Sector	Case #3 Mixture of two HP models Class #1, Matrix used: ED Class #2, Matrix used: TtVhOph Convergence crit.	π	Indication	#Param.	#Obs.	Log-lik.
Constr	Optimization cannot be completed	-	-	2x12x2+1=49	1300	-
SpecSci	Satisfied	0.278***	TtVhOph	2x9x2+1=37	1300	-2582.04
RealEst	Satisfied	0.525***	ED	2x9x2+1=37	1300	-1755.75

[a] Three other cases have been tested: ED with TtTcMph,TtVhMph mixed with TtTcMph, and ED with DistTcMph. Yet, cases involving public transit network do not assure the convergence of the model or yield poor results. In all the cases, the optimization could not be completed or could not improve the function value.

[b] Constr stands for the construction sector; SpecSci: special, scientific, technical activities; RealEst: real estate activities.

[c] The level of estimated probability `Pi` inferior to 0.5 indicates that the distance metric used in the second class has a larger probability to be the appropriate representation to account for spatial spillovers as compared to the distance metric of the first class.

[d] *** represents statical significance at the 1% level.

real estate sector. This being said, no clear spatial patterns appear at this stage. Proximity to highway tends to be associated with the predominance of peak travel time. There are counter-examples, especially in the vicinity of Paris. Density might also play some role: the most dense areas are usually associated with the Euclidean distance, while the least dense ones are more often associated with the peak travel time. One possible interpretation would be that if density is high enough, the market size allows establishments to operate at a local scale, whereas in low-density areas establishments must increase their market area and thus rely more heavily on car use. A minimum Euclidean distance approach may be most warranted when the establishments service area tends to be localized. Spatial spillovers would be more circumscribed since the degree to which establishments in one location serve the needs of users in other locations is fairly limited (Talen and Anselin, 1998). These points call for further investigation.

3.6 Conclusions

This article has exposed the need to further investigate the role of distance measures in spatial models. We contribute to the literature on location choice modeling by proposing one of the first formulations and applications of spatial count data models wherein various metrics other than Euclidean distance are considered jointly - rather than separately - to account for the linkage between neighboring observations. The methodology was applied to the location choice of new establishments in the Paris region. We show that the combination of peak and off-peak road travel times were found to (slightly) outperform other combinations including the Euclidean distance. This suggests that in areas where high congestion or uncrossable physical barriers problems clearly arise, distances purely based on topography may not be the most appropriate to capture interactions between neighboring observations. More generally, this supports the use of economically meaningful metrics over more neutral ones in spatial econometric models. The model specification proposed in the article, based on a mixture of spatial count models, can easily be applied to other fields as well as to other forms of spatial dependence that are not necessarily proximity- or transportation time-dependent, whenever the phenomenon studied involves spatial interactions that are heterogeneous by nature. To cite just a few, instances involving different forms of spatial dependence to which the methodology could be applied include interregional flows based on economic distance (Fingleton, 2008; LeSage and Pace, 2008), trade (Eliste and Fredriksson, 2004; Chen and Haynes, 2014), or innovation (Kang and Dall'erba, 2016).

To provide correct results, the various metrics used in the model must be both sufficiently differentiated (*i.e.*, not overly correlated) and regular enough. Public transport distances (travel time and network distance alike) were found to yield poor results, which is most likely related to the fact that they are quite irregular in suburban and rural areas. More generally, these conditions on the metrics raise two questions. The first one is the role of scale in the results. At the metropolitan level, variations between the Euclidean distance and transport distances seem to be pronounced enough for the model to produce good results. Yet, this might not be the case at a larger geographic scale. For instance, the studies of Le Gallo and Dall'erba (2008) or Duran-Fernandez and Santos (2014), carried out at the level of European regions and of Mexican states, respectively, find the difference between transport distances and geographical distances to be modest. On the other hand, looking at the anamorphic maps for the travel times by train for France and the UK, or by train or plane for the whole Europe[10], the map profiles are strongly deformed even in the case of long distances between cities or countries, suggesting the opposite (Denain and Langlois, 1998). This issue calls for further investigation. The second question is the role of network structure. The Paris network is strongly radial, which leads to significant variations in the ratio between travel time and Euclidean distance depending on the O-D configuration. Again, this might not be the case for other network structures. For the Euclidean and transport distances not to be excessively correlated, there must be sufficient spatial heterogeneities in the ratio of travel time (or network distance) to Euclidean distance. The network structure is not the only relevant factor in this matter; one must also consider speed limits, and congestion for the peak-period. This being said, all other things being equal, the correlation between the Euclidean and transport distances is likely to be greater for regular and dense transport networks (*e.g.*, mesh, and to a lesser extent hub and spoke, see Le Gallo and Dall'erba, 2007) than for sparser and unbalanced networks (*e.g.*, linear, radial, circular). Accordingly, the methodology - if applied to the Euclidean and transport distances - is likely to produce better results in cities of the latter category.

The Euclidean model remains useful in a number of contexts, such as cartography or navigation to mention just a few (Miller and Wentz, 2003). Yet, its explanatory power may be unable to fully capture many physical or human geographic phenomena and the impact of geographic variations on their patterns. With the advent of powerful computers and of "big

[10] See, *e.g.*, http://www.mgm.fr/PUB/Mappemonde/M198/LangloisDenain.pdf.

data", these issues await to be explored further.

References

Acs, Z.J., D. Audretsch, and M.P. Feldman (1994). R&D spillovers and recipient firm size. *Review of Economics and Statistics*, 100(1), 336-367.

Arauzo-Carod, J.M., D. Liviano-Solís, and M. Manjón-Antolín (2010). Empirical studies in industrial location choice: An assessment of their methods and results. *Journal of Regional Science*, 50(3), 685-711.

Arrow, K. (1962). Economic Welfare and the Allocation of Resources for Invention, In *The Rate and Direction of Inventive Activity. Economic and Social Factors*. Princeton University Press, 609□626.

Aten, B. (1997). Does Space Matter? International Comparisons of the Prices of Tradables and Nontradables, *International Regional Science Review*, , 20, 35-52.

Axhausen, K.W. (2003). Definitions and Measurement Problems. *Capturing Long Distance Travel*. Edited by Axhausen K.W., J.-L. Madre, J.W. Polak, and P. Toint. Baldock, Herdfordshire, England: Research Science Press.

Bell, K.P. and N.E. Bockstael (2000). Applying the Generalized-Moments Estimation Approach to Spatial Problems Involving Microlevel Data. *Review of Economics and Statistics*, 87(1), 72-82.

Benezech, V. and N. Coulombel (2013). The value of service reliability. *Transportation Research Part B: Methodological*, 58, 1-15.

Bhat, Ch.R., R. Paleti R., and P. Singh (2014). A Spatial Multivariate Count Model for Firm Location Decisions. *Journal of Regional Science*, 54(3), 462-502.

Bodson, P. and D. Peeters (1975). Estimations of the coefficients in a linear regression in the presence of spatial autocorrelation: an application to a Belgian labour-demand function, *Environment and Plannings A*, 7, 455-72.

Boscoe, F.P., K.A. Henry, and M.S. Zdeb (2012). A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals. *Professional Geographer*, 64(2), 188-196.

Buczkowska, S. and M. Lapparent (de) (2014). Location choices of newly-created establishments: Spatial patterns at the aggregate levels. *Regional Science and Urban Economics*, 48, 68-81.

Buczkowska, S. and M. Lapparent (de) (2017). Location choices under strategic interactions: Interdependence of establishment types. *Working paper*.

Chakranorty, S. (1996). A measurement spatial disparity: the case of income inequality. *Urban Studies*, 33, 1671-1686.

Chalasani, V.S., J.M. Denstadli, Ø. Engebretsen, and K.W. Axhausen (2005). Precision of Geocoded Locations and Network Distance Estimates. *Journal of Transportation and Statistics*, 8(2), 1-15.

Chatman, D.G., R.B. Noland, and N.J. Klein (2016). *Journal of the Transportation Research Board*, 2598, 1-10.

Chen, Z. and K. Haynes (2014). Spatial Impact of Transportation Infrastructure: A Spatial Econometric CGE Approach, in: *Regional Science Matters? Studies dedicated to Walter Isard*, Adam Rose and Peter Nijkamp (Eds.), Springer-Verlag.

Cole, J.P. and C.A.M. King (1968). *Quantitative Geography*. London: Willey.

Combes, P.P. and M. Lafourcade (2003). Core-Periphery Patterns of Generalized Transport Costs: France, 1978-98. *C.E.P.R. Discussion Papers 3958*.

Combes, P.P. and M. Lafourcade (2005). Transport costs: measures, determinants, and regional policy implications for France. *Journal of Economic Geography*, Oxford University Press, 5(3), 319-349.

Conley, T.G. and E. Ligon (2002). Economic distance and cross-country spillovers. *Journal of Economic Growth*, 7(2), 157-187.

Corrado, L. and B. Fingleton (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, 52(2), 210-239.

Coulombel, N. and F. Leurent (2013). Les ménages arbitrent-ils entre coût du logement et coût du transport: une réponse dans le cas francilien. *Economie & Statistique*, 457-458, 57-75.

Denain, J.C. and P. Langlois (1998). Cartographie en anamorphose. *Mappemonde*, 49(1), 16-19.

DRIEA, Île-de-France (2008). MODUS v2.1 Documentation détaillée du modéle de déplacements de la DREIF. *Technical report*.

Drukker, D.M., P. Egger, and I.R. Prucha (2013). On two-step estimation procedure of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econometric Reviews*, 32(5-6), 686-733.

Dubé, J., C. Brunelle, and D. Legros (2016). The Review of Regional Studies Location Theories and Business Location Decision: A Micro-Spatial Investigation of a Nonmetropolitan

Area in Canada. *The Review of Regional Studies*, 46, 143-170.

Duran-Fernandez, R. and G. Santos (2014). Gravity, distance, and traffic flows in Mexico. *Research in Transportation Economics*, 46, 30-35.

Duranton, G. and H.G. Overman (2005). Testing for Localization Using Micro-Geographic Data. *Review of Economic Studies*, 72(4), 1077-1106.

Eaton, B.C. and R.G. Lipsey (1980). The Block Metric and the Law of Markets. *Journal of Urban Economics*, 7, 337-347.

Eliste, P. and P.G. Federiksson (2004). Does trade liberalization cause a race to the bottom in environment policies? A spatial econometric analysis, in: L. Anselin, R. Florax (Eds.), *Advances in Spatial Econometrics*, Springer-Verlag, New York.

Faber, B. (2014). Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System. *Review of Economic Studies*, 81(3), 1046-1070.

Fetter, F.A. (1924). The Economic Law of Market Areas. *Quarterly Journal of Economics*, 39, 520-529.

Fingleton, B. (2008). A Generalized Method of Moments Estimator for a Spatial Model with Moving Average Errors, with Application to Real Estate Prices. *Empirical Economics*, 34, 35-57

Fingleton, B. and J. Le Gallo (2008). Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties. *Papers in Regional Science*, 87(3), 319-339.

Fotheringham, A.S., M.E. Charlton, and C. Brunsdon (1996). The Geography of Parameter Space: An Investigation into Spatial Nonstationarity. *International Journal of GIS*, 10(5), 605-627.

Fransen, K., T. Neutens, P. De Mayer, and G. Deruyter (2015). A commuter-based two-step floating catchment area method for measuring spatial accessibility of daycare centers. *Health & Place*, 32, 65-73.

Getis, A. and J. Aldstadt (2004). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis*, 36(2), 90-104.

Graham, D.J. (2007). Variable Returns to Urbanization and the Effect of Road Traffic Congestion, *Journal of Urban Economics*, 62(1), 103-120.

Guy, C.M. (1980). *Retail Location and Retail Planning in Britain* (Gower, Aldershot, Hants).

Guy, C.M. (1983). The assessment of access to local shopping opportunities: A comparison of accessibility measures. *Environment and Planning B*, 10, 219-238.

Haggett, P. (1967). Network models in geography. In: Chorley, RJ and Haggett P, editors. *Integrated models in geography*. London: Methuen; 609-668.

Héran, F. (2011). *La ville morcelée. Effets de coupure en milieu urbain*. *Economica*, Collection Méthodes et Approches, Paris.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. New York: Random House.

Jacobs, J. (1969). *The Economy of Cities*. New York: Vintage Books.

Jaffe, A.B. (1989). Real effects of academic research. *American Economic Review*, 79(5), 957-970.

Kang, D. and S. Dall'erba (2016). An Examination of the Role of Local and Distant Knowledge Spillovers on the US Regional Knowledge Creation. *International Regional Science Review*, 39(4), 355-385.

Kelejian, H.H. and I.R Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17(1), 99-121.

Klier, T. and D. McMillen (2008). Clustering of auto supplier plants in the United States. *J. Bus. Econ. Stat*, Am. Stat. Assoc., 26(4), 460-471.

Kostov, P. (2010). Model boosting for spatial weighting matrix selection in spatial lag models. *Environment and Planning B: Planning and Design*, 37 (3), 533-549.

Kwon, Y. (2002). Rent-commuting cost function versus rent-distance function. *Journal of Regional Science*, 42(2), 773-791.

Lambert, D.M., J.P. Brown, and R.J.G.M. Florax (2010). A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. *Regional Science and Urban Economics*, 40(4), 241-252.

Lapparent, M. (de) and M. Koning (2015). Travel discomfort-time tradeoffs in Paris subway: an empirical analysis using interval regression models. *EPFL-REPORT-204605*.

Le Gallo, J. and S. Dall'erba (2008). Spatial and Sectoral Productivity Convergence between European Regions, 1975-2000. *Papers in Regional Science*, 87(4), 505-525.

LeSage, J.P. (2003). A Family of Geograhically Weighted Regression Models. In *Advances in Spatial Econometrics: Methodology, Tools and Aplications*, edited by L. Anselin, R.J.G.M. Florax, and S.J. Rey. Heidelberg: Springer.

LeSage, J.P. and R.K. Pace (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48(5), 941-967.

Liesenfeld, R., J.F. Richard, and J. Vogler (2015). Likelihood Evaluation of High-Dimensional Spatial Latent Gaussian Models with Non-Gaussian Response Variables. Available at SSRN: SSRN-id2196041 2.

Liviano-Solís, D. and J.M. Arauzo-Carod (2013). Industrial Location and Interpretation of Zero Counts. *Annals of Regional Science*, 50, 515-534.

Liviano-Solís, D. and J.M. Arauzo-Carod (2014). Industrial Location and Spatial Dependence: An Empirical Application. *Regional Studies*, 48(4), 1-17.

McMillen, D.P. and J.F. McDonald (2004). Locally Weighted Maximum-Likelihood Estimation: Monte Carlo Evidence and an Application. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, edited by L. Anselin, R.J.G.M. Florax, and S.J. Rey. Heidelberg: Springer.

Miller, H.J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2), 284-289.

Miller, H.J. and E.A. Wentz (2003). Representation and Spatial Analysis in Geographic Information Systems, *Annals of the Association of American Geographers*, 93(3), 574-594.

Motte-Baumvol, B., O. Bonin, C.D. Nassi, and L. Belton-Chevallier (2015). Barriers and im(mobility) in Rio de Janeiro. *Urban Studies*, 1-17.

Nguyen, C.Y., K. Sano, T.V. Tran, and T.T. Doan (2012). Firm relocation patterns incorporating spatial interactions. *The Annals of Regional Science*, 50(3), 685-703.

Noonan, D.S. (2005). Neighbours, Barriers and Urban Environments: Are Things "Different on the other Side of the Tracks"? *Urban Studies*, 42(10), 1817-1835.

Ortuzar, J. de D. and L.G. Willumsen (2011). *Modeling Transport* 4th Edition. Chester, England: Wiley.

Parent, O. and J.P. LeSage (2008). Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. *Journal of Applied Econometrics*, John Wiley & Sons, Ltd., 23(2), 235-256.

Partridge, M.D., D.S. Rickman, K. Ali, and M.R. Olfert (2008). Lost in space: population growth in the American hinterlands and small cities. *Journal of Economic Geography*, 8(6), 727-757.

Perreur, J. and J. Thisse (1974). Central metrics and optimal location. *Journal of Regional Science*, 14, 411-421.

Rietveld, P., B. Zwart, B. van Wee, and T. van den Hoorn (1999). On the relationship between travel time and travel distance of commuters. *The Annals of Regional Science*, Springer, 33(3), 269-287.

Rincke, J. (2010). A commuting-based refinement of the contiguity matrix for spatial models, and an application to local police expenditures. *Regional Science and Urban Economics*, 40, 324-330.

Slade, M.E. (2005). The Role of Economic Space in Decision Making, ADRES Lecture. *Annales D'Economie et de Statistique*, 77, 1-20.

Talen, E. and L. Anselin (1998). Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds. *Environmental and Planning A*, 30, 595-613.

Tobler, W. (1970). A computer movie stimulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.

Vega, S.H. and J.P. Elhorst (2013). On spatial econometrics models, spillovers effects, and W. *ERSA conference papers ersa13p222*, European Regional Science Association.

Weisbrod, G. (2008). Models to predict the economic development impact of transportation projects: historical experience and new applications. *The Annals of Regional Science*, 42(3), 519-543.

Table 3.6: Probabilistic mixture of hurdle-Poisson models for spec., scien., tech. and real estate activities using two latent classes for case #1: ED (class #1) with TtVhMph (class #2). Hurdle and truncated-at-zero Poisson parts reported.

SpecSci Class #1, Matrix:	Estimate ED	(t-st)
	Hurdle: Zero counts	
Constant	18.417	(3.20) ***[a]
Same sector estab.[b]	0.139	(0.53)
Large estab.	0.064	(0.44)
Trips HW, intel.	-0.320	(-1.27)
Offices	-0.051	(-0.76)
Univ., schools	1.126	(3.00) ***
Dist. highway	-0.281	(-0.81)
Public transp.	-0.017	(-0.14)
Price of offices (log)	-4.538	(-2.52) **
	Poisson: Positive counts	
Constant	14.043	(23.17) ***
Same sector estab.	0.490	(8.17) ***
Large estab.	-0.185	(-5.92) ***
Trips HW, intel.	0.932	(22.98) ***
Offices	0.039	(3.10) ***
Univ., schools	0.323	(7.28) ***
Dist. highway	-0.318	(-11.14) ***
Public transp.	0.222	(10.31) ***
Price of offices (log)	-2.429	(-15.74) ***
Class #2. Matrix:	TtVhMph	
	Hurdle: Zero counts	
Constant	20.338	(2.60) **
Same sector estab.	0.259	(1.98) *
Large estab.	-0.072	(-1.40)
Trips HW, intel.	0.734	(5.15) ***
Offices	0.103	(3.09) ***
Univ., schools	0.517	(4.37) ***
Dist. highway	0.010	(0.07)
Public transp.	0.162	(3.87) ***
Price of offices (log)	-5.036	(-1.39)
	Poisson: Positive counts	
Constant	12.419	(9.19) ***
Same sector estab.	0.118	(1.74) *
Large estab.	0.0004	(0.02)
Trips HW, intel.	0.638	(18.77) ***
Offices	0.080	(9.42) ***
Univ., schools	0.388	(14.71) ***
Dist. highway	-0.00004	(0.00)
Public transp.	0.078	(6.92) ***
Price of offices (log)	-2.423	(-4.16) ***
	Pi =0.283	(9.12) ***
	Convergence crit.	Satisfied
	#Parameters	2 x 9 x2
	Log-likelihood	-2593.621

RealEst Class #1, Matrix:	Estimate ED	(t-st)
	Hurdle: Zero counts	
Constant	30.167	(4.89) ***
Same sector estab.	1.142	(2.51) **
Large estab.	0.270	(0.95)
Shops. offices	0.652	(2.33) **
Resid. area	2.065	(2.80) ***
Public transp.	-0.236	(-1.28)
Residence tax	1.271	(1.57)
Income (log)	-0.609	(-0.64)
Price of shops (log)	-9.107	(-4.56) ***
	Poisson: Positive counts	
Constant	16.327	(20.49) ***
Same sector estab.	1.729	(11.77) ***
Large estab.	-0.453	(-7.41) ***
Shops. offices	0.598	(13.07) ***
Resid. area	-0.173	(-2.40) **
Public transp.	0.425	(17.24) ***
Residence tax	-0.573	(-3.53) ***
Income (log)	0.688	(5.12) ***
Price of shops (log)	-3.085	(-11.82) ***
Class #2. Matrix:	TtVhMph	
	Hurdle: Zero counts	
Constant	4.150	(0.40)
Same sector estab.	0.148	(2.90) ***
Large estab.	-0.057	(-0.86)
Shops. offices	0.055	(1.10)
Resid. area	0.762	(3.30) ***
Public transp.	0.070	(1.28)
Residence tax	0.401	(1.05)
Income (log)	0.730	(1.18)
Price of shops (log)	-0.915	(-0.19)
	Poisson: Positive counts	
Constant	10.481	(8.05) ***
Same sector estab.	0.710	(3.02) ***
Large estab.	-0.094	(-2.09) **
Shops. offices	0.418	(6.65) ***
Resid. area	0.239	(2.67) **
Public transp.	0.366	(5.55) ***
Residence tax	-0.850	(-3.47) ***
Income (log)	1.099	(5.61) ***
Price of shops (log)	-1.513	(-12.58) ***
	Pi =0.522	(11.12) ***
	Convergence crit.	Satisfied
	#Parameters	2 x 9 x2
	Log-likelihood	-1772.672

[a] ***, **, * represent statical significance at the 1%, 5% and 10% level, respectively.

[b] See Table 4.5 for the description of variables.

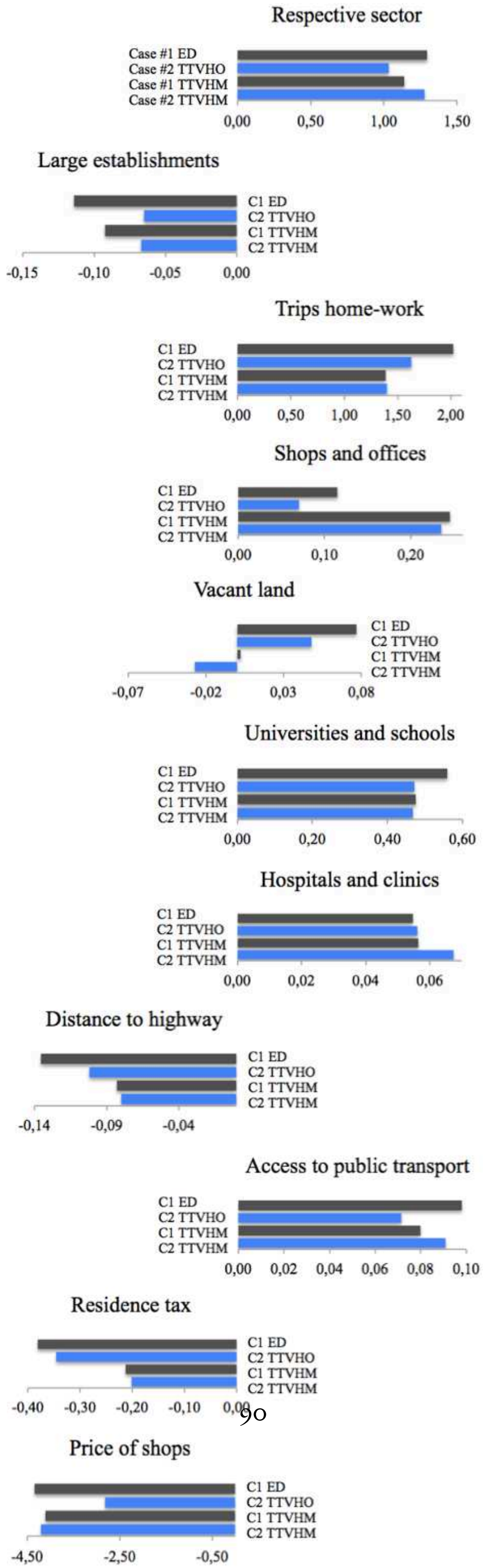


Figure 3.4: Selected example of the estimates of the truncated-at-zero Poisson parts of the mixture models for the construction sector. Comparison of two cases for models that use two classes.

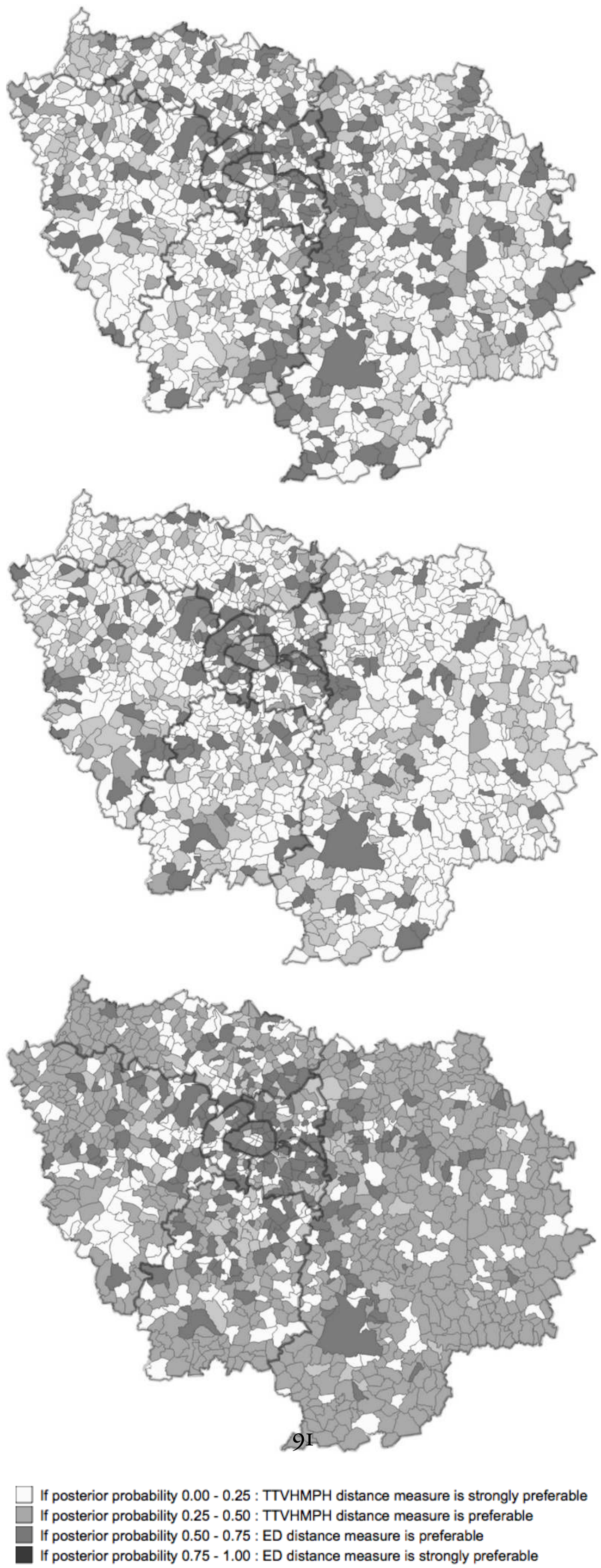


Figure 3.5: Posterior probabilities of belonging to class #1 (ED) as opposed to class #2 (TtVhMph) at the municipality level (each municipality can be treaten as an alternative in the decision-making process of an establishment): construction sector (upper figure), special, scientific and technical activities (central figure), and real estate sector (lower figure)

The complexity of the expressions for the probabilities increases in the number of locations and firm types.

- K. Seim, "An emprirical model of firm entry with endogenous product-type choices" (2006)

4

Location choices under strategic interactions: Interdependence of establishment types

ABSTRACT The need for methodological advances in order to model more realistically the complexity of establishment optimal location choices is the key motivation of our present paper. We shed light on strategic interactions, fundamental in establishment location choices, yet largely unheeded in the empirical literature. Being non-strategic means that an establishment ignores other players' decisions. Less is known about how to correctly adapt location choice models to study establishments' discrete choices when they are interrelated. When locational choice models are developed for several activity sectors, each model is typically run independently. We estimate a static discrete game of incomplete information to obtain a Bayesian Nash Equilibrium. Our modeling approach includes several nice features. We account for the spatial spillover effects from the neighboring zones using accessibility measures. We deal with the price endogeneity issue. Location choice models are developed simultaneously for seven various establishment types allowing us to explore strategic interactions between these types.

Establishments are prone to account for the actual and potential presence of other players. We find strong negative effects among inter-industry pre-existing establishments. Still, new establishments are likely to enter the market where other new establishments of the same type are expected to locate. The influence of variables that capture market conditions, within- and inter-industry economies at each location is found to be critical in establishment location decisions, yet quite different depending on the establishment type being considered.

Keywords: location choice, strategic interactions, Bayesian Nash, discrete game, interdependence of sectors, spatial spillovers

JEL Classification: D22, C57, C31, R15, C36

4.1 Introduction

Many key strategic decisions establishments make, such as where to set up in the market, involve discrete choices (Draganska et al., 2008). These decisions are fairly complex, yet particularly important since, unlike other marketing mix elements, they are less adjustable in the short-run without incurring significant costs (Zhu and Singh, 2009). In the literature on establishment location choices, when locational choice models are developed for several activity sectors, each model is typically run independently (see, *e.g.*, Chatman et al., 2016; Buczkowska and Lapparent, 2014). An establishment chooses its optimal location that maximizes its payoffs given its own type. If establishments were making location decisions in isolation, existing discrete choice models can be of service to model their decisions. In concentrated markets, however, establishment decisions are interdependent, which considerably complicates the formulation and estimation of market structure models (Berry and Reiss, 2007). These thorny problems posed by interrelated decisions generally cannot be assumed away, without altering the realism of the model of establishment decision making (Berry and Reiss, 2007). The conventional approaches to location selection, *i.e.*, traditional theory and methods, fail (Thill, 1997) by providing only a set of systematic steps for problem-solving without considering strategic interactions between the establishments in the market. Being non-strategic would mean that an establishment ignores other players' decisions (Toivanen and Waterson, 2005).

However, less is known about how to correctly adapt location choice models to study establishment discrete choices when they are interrelated (Draganska et al., 2008). Strategic motives are widespread phenomena and they are fundamental in establishments' location choices, in particular. Yet, the small amount of attention the topic has received is surprising. Bajari et al. (2010a), Anselin (2010), and Blume (1993) believe that it is as an important oversight. Strategic interactions have been largely unsung in the empirical analyses since the year 1929 when Hotelling (1929) brought the discussion in the industrial organization literature. Most of the papers are less than a decade old (Bajari et al., 2013). This literature is in its infancy, in part, due to the complexity of expressions for the probabilities used in the models which increases along with the number of locations and establishment types (Draganska et al., 2008). The study of interrelated discrete decisions poses several methodological challenges, such as large state spaces. Often, there is a very large number of location configurations and hence significant computational challenges can be involved in modeling these decisions (Draganska et al., 2008).

What makes these discrete choices particularly interesting and challenging to analyze is that decisions of a particular establishment are interrelated with choices of the others because an establishment accounts for the actions of other agents when making its own decisions (Draganska et al., 2008)[1]. A discrete game strictly generalizes a standard random utility model, but does not impose the often strong assumption that agents act in isolation. A static game is a generalization of a discrete choice model which allows the actions of a group of agents to be interdependent (Bajari et al., 2010a). Moreover, unlike in a standard discrete choice model, the actions of other players in the game enter the utility of a particular establishment. Thus, the utility depends on the possible actions of other agents (Bajari et al., 2010a).

In our application, we therefore develop a framework which allows payoffs to be interdependent. We insist on the fact that an appropriately specified model of simultaneous location choices needs to recognize the interdependence of actions and profits (Berry and Reiss, 2007). We focus on incorporating strategic interactions between several establishment types at once. We distinguish seven establishment types based on their activity sector[2]. These are: commercial establishments in the retail sector and units in wholesale, hotels and restaurants, establishments involved in transport and storage, manufacturing establishments, financial establishments, and units dedicated to professional, scientific and technical activities. Decisions

[1]The goal of the survey by Draganska et al. (2008) is to illustrate how a game-theoric framework can aid in construction and estimation of interrelated choice models.

[2]Similarly, establishment types can be based on their sector and size.

of establishments of various types are determined simultaneously, thus not as it has been postulated in a standard location model. Establishments are prone to account for the actual and potential presence of other players. Thus, we choose to include both, pre-existing establishments and units expected to set up in the market, since they seem to play a critical role in location choices. In addition, an establishment's profit seems to depend on the actual type rather than just on the total number of other establishments in the market.

We estimate a static discrete game of incomplete information to obtain a Bayesian Nash Equilibrium. We develop a baseline model at a group level. Our modeling approach includes several nice features: (1) We incorporate strategic interactions between several establishment types and represent them in a form of a matrix to show how dependent each establishment type is on every other type at the level of regional economy of the Paris region. (2) We account for the spatial spillover effects from the neighboring zones using accessibility measures. (3) We deal with the price endogeneity issue.

Our objective is to answer questions, such as: (1) What are the push and pull factors that play a critical role in location processes and how does the significance of these factors vary depending on the establishment type? (2) Which types of establishments attract the others and which push them out from locating nearby? (3) How strong are these effects? (4) How important are competitive effects across establishments? (5) How do competitive effects vary depending on the establishment type?

We report results for location choice models run simultaneously for seven establishment types. We notice that establishment locational choices differ by type. Overall, all establishments exert a strong negative impact on their pre-existing competitors of the same type. Yet, new establishments prefer to enter the market where other newly-formed establishments of the same type are expected to locate. We observe impact of asymmetries in strategic effects for establishments of different types. The influence of variables that capture market conditions at each location, such as demand and cost factors, availability of dedicated space, access to public transit networks or road networks, access to amenities, physical qualities of facilities, within-industry and inter-industry economies is found to be critical in establishment location decisions, yet quite different depending on the establishment type being analyzed.

The rest of the paper is organized as follows. We explain why it is critical to incorporate strategic interrelations in location choice models and how these interactions are formed. We give a brief description of possible push and pull factors that can be considered in location choice models and present our data. Then we explain the concept of matrix of strategic interactions. We shed light on possible modeling approaches and develop our parametric statistical model. We discuss the results. Conclusions are drawn in the final section.

4.2 Within-industry and inter-industry interactions

An establishment does not seem to act in isolation during its decision-making processes (Nguyen et al., 2012). Jayet (2001) proves the existence of interactions among units located in space. Interactions of agents may occur in various forms and with various levels of magnitude. A long discussion on within-industry economies (own-industry competition effects versus Marshallian agglomeration economies, *i.e.*, own-sector collocation), and inter-industry economies (Jacobsian agglomeration economies, *i.e.*, economic diversity) can be found in the existing literature. A number of empirical methods to quantify and characterize the tendency of individual establishments or industries to cluster or disperse in space have been proposed. These measures can be divided into three groups: (1) empirical measures, such as Herfindahl or Gini indices which do not take space into account (see, *e.g.*, Krugman, 1991); (2) indicators, such as Ellison and Glaeser index (Ellison and Glaeser, 1997) which uses space as being discrete (administrative units); and (3) measures that treat space as being continuous (*e.g.*, Duranton and Overman, 2005; Casanova et al., 2017).

Yet, instead of analyzing spatial distribution of economic activities, *e.g.*, concentration or dispersion tendencies of pre-existing establishments that belong to two industries, horizontally or vertically linked, we rather opt to explain how the actual and potential presence of other establishments from both, the same and other sectors, influence the location decisions of new units. The interrelated nature of locational decisions suggests modeling them as strategic games to recognize the interdependence of payoffs and actions.

The impact on payoffs due to the behavior of other establishments is a vector that contains the number of establishments in each location in the market (Seim, 2006). In addition, since establishments have different costs, sell different products, they seem to care about the actual identities of other players rather than simply the total number of establishments. It is therefore important to explore how these potential heterogeneities might affect estimation (Berry and Reiss, 2007). In the paper by Seim (2006), profits depend only on the number of entrants at every location and not on the entrants' identities. Seim (2006) maintains symmetry for computational reasons suggesting that more heterogenous payoff functions can be proposed instead. To cope with this problem, Zhu and Singh (2009) propose a modeling approach that explicitly allows for asymmetries related to identity of establishments. Clearly, allowing for such establishment asymmetries is not without cost since it leads to a significantly larger number of parameters to be estimated (Zhu and Singh, 2009).

We estimate a static discrete game of incomplete information to obtain a Bayesian Nash Equilibrium at the group level using data at the aggregate level. We permit asymmetries across establishment types in the impact of interaction effects and exogenous market characteristics.

4.3 Why bother about strategic interactions?

There are several relevant reasons why incorporating strategic interactions may turn out to be necessary. If strategic effects matter and are ignored, other factors included in the payoffs will be estimated with a bias (Draganska et al., 2008). The magnitude of such a bias will depend on the degree to which strategic effects matter. In some cases, this bias can be severe and even go as far as changing the sign of the effects of interest. Without strategic effects, counterfactuals and policy experiments become difficult to evaluate. Reduced form models can give us only some insights into the effects of market and establishments' characteristics on decisions, yet nothing beyond that. In addition, incorporation of strategic effects often allows for the inclusion of other players' characteristics (such as sector or size) via excluded variables and fixed effects into the payoff function. Without strategic effects, there would be no particular reason why the characteristics of other establishments would enter the payoff function of a particular agent. The main purpose of such exclusions is to aid identification. These strategic interactions also improve the fit to the data and the predictions.

There is a need for more realistic studies of complex establishment's decision-making processes. Even though the computational burden imposed by these models considering strategic interactions is relatively high, it seems that the costs imposed are more than offset by the benefits that accumulate (Draganska et al., 2008).

4.4 Push and pull factors

When deciding where to locate in the market, establishments account for a variety of factors including demand, proximity to and size of their consumer or user base, cost conditions, and logistical issues (Zhu and Singh, 2009). Yet, the influence of variables might be quite different depending on the establishment type considered. As highlighted by Min (1987), the retail, service sector, and professional location analyses are likely to be highly sensitive to the demand and revenue-generating factors. Whereas cost factors should play a critical role rather in the locational strategies of, *e.g.*, warehouses or goods-producing establishments.

In case of revenue-demand oriented establishments, location is a major determinant of revenue. High customer- or user-contact issues or higher traffic counts seem to be critical. Costs are relatively constant for a given area, therefore the revenue function is important. Revenue-demand oriented establishments tend to carefully consider the purchasing power of customer-drawing area and to make sure that their proposed services and image are compatible with demographics of potential clients. Physical qualities of facilities, such as parking places, access, security and lighting may also be tested whether they play a role[3].

Considering that, in case of cost focused establishments, such as manufacturing establishments, low customer contact allows to concentrate on the identifiable costs and most major costs can be identified explicitly for each site. These establishments tend to pay a particular attention to transport cost of raw materials, shipment cost of finished goods, energy and utility cost, accessibility to labor and materials, or the level of taxes[4].

While favorable market conditions encourage establishments' choices, such markets are also likely to attract competitors. In the long-run, the interaction of described different push and pull factors determines the observed location patterns by establishments (Zhu and Sing, 2009). Opposing forces drive profits. The trade-off between the proximity to competitors and the desirability of certain location characteristics should be pondered (Zhu and Singh, 2009; Seim, 2006). Yet, establishments' behavior also depend on the choices of others because increased competition at a given location adversely affects profits (Seim, 2006). Therefore, establishments should consider competitive pressure (Draganska et al., 2008) and the actual and potential presence of competitors (Zhu and Singh, 2009; Seim, 2006).

Site-specific operating costs are often neglected in the location choices of establishments (Nwogugu, 2006). Costs consist of a rent and a fixed set-up cost (Chu and Lu, 1998). Location-specific costs of running an establishment mainly take a form of property costs and lease payments (Seim, 2006). Set-up costs do not depend on the location site and are exogenously

[3]See http://fr.slideshare.net/Joanmaines/location-strategy.
[4]See http://fr.slideshare.net/Joanmaines/location-strategy.

given. A fixed set-up cost is paid by every player entering the market (Granot et al., 2010; Peng and Tabuchi, 2007). Cost determinants are typically proxied by rent and operational policies, such as wage rates or the number of hours worked, and the management caliber. According to Chang and Hsieh (2014), a proportion of rent expenses against sales should be carefully calculated. As the authors claim, "value for money" locations are usually hard to find due to the fact that most of this kind of sites are already being rented or that rent expenses of good locations, such as in the downtown or newly developed areas, are extremely high eliminating the profit. Aquirregabiria and Suzuki (2016) highlight the fact that attractive locations are typically expensive and can be associated with stronger competition. To illustrate this logic, let us use the example of Seim (2006). The placement out of the edge of the city may have a small set of adjacent locations and consequently a lower demand, yet be characterized by smaller number of competitors. Placements in the central parts of the city with many close-by competitors are exposed to a stronger competition. Yet, from a demand perspective, they will be more attractive if it grants an easy accessibility and convenience to consumers or users living in the neighboring locations (Seim, 2006). Establishments should thus consider the trade-off between being accessible to many potential consumers or users, higher land prices, and ferocious competition when deciding where to open a new store.

4.5 Data

In the present paper, we implement strategic interactions in establishment location choices. Seim (2006) believes that the insufficient empirical literature in the strategic interactions field is due to the complexity of the expressions for the probabilities used in the models which increases along with the number of potential locations and establishments (in our application, these would be the number of locations and establishment types). Empirical implemention of strategic interactions in the models is computationally complicated and difficult to estimate. As explained by Bajari et al. (2010a, b), the set of possible alternatives that the decision-maker has at its disposal plays a critical role in the location decisions. In addition, when making a choice, individuals try to restrict the number of alternatives by eliminating the unsatisfactory alternatives (Tversky, 1972). The reduction of the size and complexity of entropy of the choice set can be done by following some hierarchical process of simplification: by elimination or grouping (Genesh and Svestka, 1979; Pooler, 1994), or by comparing a weighted average of

Table 4.1: Selected types of newly-created establishments

Establishment type	#Newly-created establishments
Type 1: Manufacturing	3 296
Type 2: Retail	10 899
Type 3: Wholesale	8 572
Type 4: Transport, storage	3 072
Type 5: Financial activities	4 446
Type 6: Hotels, restaurants	3 524
Type 7: Professional, scientific and technical activities	15 282
Other newly-created establishments	38 883
Total	87 974

payoffs across locations (Seim, 2006). Similarly, by a suitable augmentation of extension of the notion of state, it is possible to create several types of agents by classifying players into distinct groups (Aoki, 1995). In the following subsections 4.5.1 and 4.5.2, we explain how we group the establishments and the alternative choices these establishments are choosing from.

4.5.1 Primary data sources

Different data sources were compiled for the present study, drawn primarily from the Census survey of establishments carried by the French National Institute of Statistics and Economic Studies (INSEE) between the years 2003 and 2011. The survey includes information on both, newly-created establishments and stocks of pre-existing establishments.

We distinguish seven groups of newly-created establishments that set up in the market in 2007, based on information containing their activity sector (see Table 4.1). These are: (1) retail and (2) wholesale establishments, (3) hotels and restaurants, (4) establishments involved in transport and storage, (5) manufacturing, (6) financial establishments, and (7) entities dedicated to professional, scientific and technical activities. We assume that within a group, these establishments have similar multivariate profiles and can therefore be grouped together[5].

We analyze the behavior of newly-created establishments created in the year 2007, these are 87 974 units in total and 49 091 establishments that belong to the selected categories. A newly-created establishment chooses one zone from the set of 109 alternatives. The Paris region has

[5]It can be suggested to introduce some level of heterogeneity. *E.g.*, using data at the individual establishment level, establishments could be distinguished based on their activity sector, size, age, number of establishments per firm, location of its headquarter, distance to the distribution center.

been divided into 109 zones following the way it was done by STIF-OMNIL-DRIEA for the purpose of the Global Survey of Transport EGT 2010 of the Paris region (see Fig. 4.1).

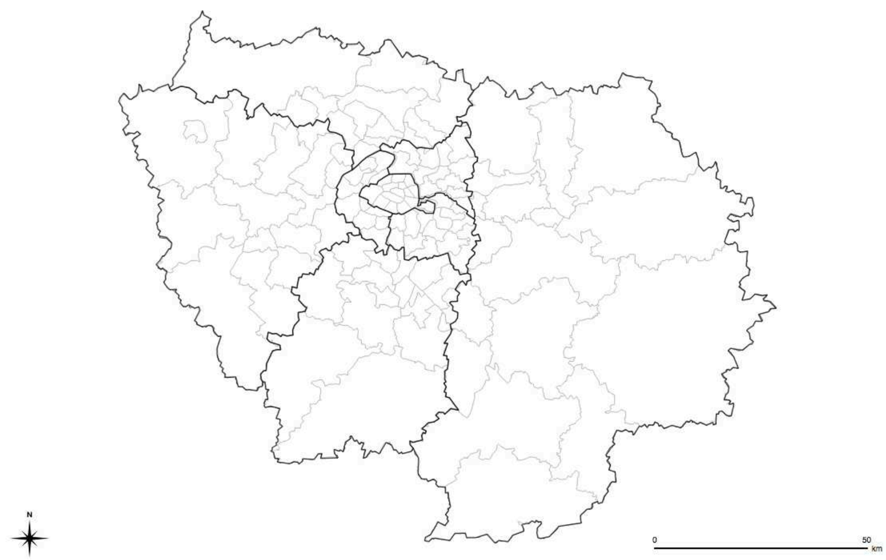


Figure 4.1: 109 EGT zones of the Paris region

We also observe 763 131 pre-existing establishments in the market in the beginning of 2007. Pre-existing establishments are pooled across eleven sectors: industry; construction; commerce; transport; financial and insurance activities; real estate activities; hotels and restaurants; information and communication; special, scientific and technical activities; education; health and social actions, and the additional category "Others" and several size classes: establishments with zero employees[6]; establishments with 0-9 employees (so called micro establishments); establishments with 10 or more employees; and entities with at least 50 employees.

The percentage of pre-existing commercial establishments among all the establishments is as high as 22%. The second biggest sector is the special, scientific and technical activities sector, with 16.2% of pre-existing establishments. The third most numerous sector is construction, with 9.2% of entities already set up in the market for some time. We also observe the smallest sectors, namely education and transport with 1.6% and 3.9% of all pre-existing units, respectively. Around 60% of all pre-existing establishments are entities with zero employees. Establishments with 0-9 employees account for 93% of all establishments. Next, 7% of pre-existing

[6]An establishment with zero employees is an operating entity with no employees other than the owner.

Table 4.2: Structure of pre-existing establishments by activity sector in the beginning of 2007

Category	Stock: #Pre-existing establishments	(%)
Main sectors:		
Manufacturing	40 258	5.275
Construction	69 908	9.161
Commerce (wholesale, retail)	16 8235	22.045
Transport, storage	30 020	3.934
Accommodation, food	47 035	6.163
Information, communication	46 003	6.028
Financial, insurance activities	38 847	5.090
Real estate activities	43 031	5.639
Professional, scientific, tech.	123 565	16.192
Administrative, support service	39 385	5.161
Arts, entertainment, recreation	12 701	1.664
Other service activities	26 898	3.525
Other sectors:		
Agriculture; Mining	4 486	0.588
Public admin., defence	72 859	9.547
Total	763 131	100

Table 4.3: Structure of pre-existing establishments by size category in the beginning of 2007

Category	Stock: #Pre-existing establishments	(%)
Est with 0-9 employees	706 102	92.527
Establishments with 10-49 employees	46 142	6.046
Establishments with 50 or more employees	10 887	1.427
Total	763 131	100

units are registered as establishments with 10 or more employees and only 1.4% with at least 50 employees.

4.5.2 Zones' characteristics

An establishment chooses one zone from the set of 109 alternatives - 109 EGT zones of the Paris region. These zones are described by variables that represent measures of demand, cost, availability of vacant land and of floor space, accessibility to shops, services, universities and schools, transport infrastructure, as well as real estate prices, wages, income, tax levels, and competitive factors. The complementary data sources cover several categories described below.

Various proxies can be suggested to capture the level of demand in a particular market and

Table 4.4: Pre-existing establishments' location across the Paris region in the beginning of 2007

Category		Stock: #Pre-existing establishments	(%)
Paris	department 75	289 416	37.924
Inner ring	department 92	85 133	11.156
	department 93	62 538	8.195
	department 94	53 157	6.966
Outer ring	department 77	48 186	6.314
	department 78	58 396	7.652
	department 91	45 215	5.925
	department 95	42 078	5.514
	Total	763 131	100

in its surroundings. In the current paper we test: (1) The measure of accessibility to population with some specific characteristics. Population living within a particular market area can be described by income, family status, age, education level (Seim, 2006; Chatman et al., 2016) as well as a socio-professional class. (2) The number of trips including its origin and destination and their purpose, such as shopping, restaurant, leisure, or social activities. Depending on the establishment types being analyzed, these variables can be treated as accessibility to the potential customers or users[7].

Data on characteristics of local population were gathered through the General Census of INSEE. Information on income imputation was prepared by the Public Finances General Directorate (DGFiP) and available through INSEE. Information on the total number of trips came from the Global Survey of Transport EGT 2010 of the Paris region (STIF-OMNIL-DRIEA).

Next, cost can be proxied using: (1) detailed information on the average value and rent levels per square meter for new, renovated, and non-renovated offices, for the most exclusive shops, the second class shops, and the least exclusive shops, as well as for the individual industrial units and warehouses, (2) the level of wages of white-collar workers, managers, technicians, and blue-collars, (3) professional and housing taxes. The Callon database enables us to gather

[7]In our current paper, we propose limited number of proxies to capture the demand level since our analyses treat seven different establishment types and we do not want to go too much in details. Yet, we extend this interesting topic in the fourth paper of this thesis. For the detailed discussion on the proxies of demand for the establishments highly sensitive to the demand and revenue-generating factors, such as retails and services, see the paper by Buczkowska and Coulombel (2017).

detailed information on the real estate prices. Wages were provided through the Annual Declaration of Social Data of INSEE. The tax levels were gathered by DGFiP.

We also measure: (1) net floor area and number of units for housing and professional offices, (2) availability of vacant land that can meet new projects, (3) accessibility to skilled workforce (white-collar workers and managers), blue-collar workers, employment in the industrial sector, commerce, services, construction sector, health, eduction, public administration, etc., (4) number of trips with the purpose of a professional meeting or a meeting in academia. Data on housing and offices construction are drawn from the Sitadel2 database managed by the Ministry of Ecology, Sustainable Development and Energy (MEDDE). Information on availability of vacant land can be found in the MOS database of IAU. The characteristics of local workforce were gathered through the General Census and supplied by INSEE.

Three sources of the numerical land use databases are considered. These are MOS (IAU) and Corine Land Cover (MEDDE), and the BPE equipment database provided by INSEE through the Demographic and Social Statistics Directorate. These databases allow us to account for the proximity to shops, services, industrial and commercial zones, zones dedicated to material extraction, construction sites, universities and schools, hospitals and clinics, zones dedicated to sport and leisure activities, and various other amenities, for example, parks, green spaces, and parking places, road network, rail terminals, and airports.

Information on transport infrastructures, *e.g.*, access to the train, subway, tramway stations, buss stops, airports, and distance to highways, was obtained through the Regional and Interdepartmental Directorate of Public Works and Development (DRIEA) of the French National Institute of Geographic and Forest Information (IGN).

When running the models, we test all of the mentioned variables. Different sets of variables are used depending on the establishment type being analyzed. In Table 4.5, we present the list of variables selected in the final stage of models construction.

Table 4.5: Final list of variables used in the models (EGT zones characteristics dictionary)

Variable	Description
Geographical position:	
Paris/La Défense/New Cities/Outer suburbs	1 if a zone is located in Paris/La Défense/New Cities/the outer suburbs excl. New Cities
Stocks of pre-existing establishments:	
Entities that belong to the same estab. type (INSEE)	Stock of pre-existing establishments of the same type (within-industry) to the zone surface (km^2)[a]
Estab. from other sectors (INSEE)	Stock of pre-existing establishments that belong to the commerce, hotels, restaurants, financial, insurance activities, professional, scientific, technical activities, real estate activities, transport and storage, and manufacturing sectors (inter-industry) divided by the zone surface (km^2)
Access to workforce:	
Labor force density (/1000) (INSEE)	Total active population divided by the zone surface (km^2) (accessibility measure)
White-collar workers ratio (INSEE)	Number of white-collar workers divided by the size of labor force (accessibility)
Access to population (demand):	
Population density (/1000) (INSEE)	Total population divided by the zone surface (km^2) (accessibility)
Trips (EGT/STIF-OMNIL-DRIEA)	Total number of trips with the purpose of doing shopping, restaurant visit, professional or academic meeting (accessibility)
Land availability:	
Dedicated land (MOS/IAU)	Fraction of a zone's surface dedicated to commerce/universities or schools (%) (accessibility)
Dedicated land (Corine Land Cover/MEDDE)	Fraction of a zone's surface (/100) dedicated to industrial and economic activities, extraction of materials (%) (accessibility)
Surface	Log value of total zone surface (km^2)
Transport infrastructures:	
Distance to highway (DRIEA/IGN)	Distance to the nearest highway (km) (accessibility)
Public transport (DRIEA/IGN)	Number of subway, train stations, and bus stops in a zone (log)
Roads, rail terminals (Corine Land Cover/MEDDE)	Fraction of a zone's surface (/100) dedicated to the road and rail networks in 2006 (%) (accessibility)
Airport (Corine Land Cover/MEDDE)	Fraction of a zone's surface (/100) dedicated to airports in 2006 (%) (accessibility)
Rents and wages:	
Rent (or value) of offices or shops (Callon database)	Log value of the average rent (price) level of offices or shops per square meter (euros)
Wage of blue-collar workers (INSEE)	Average wage of blue-collar workers (euros)
Instruments used to predict rents[b]:	
Pop/Emp (INSEE)	Population/Employment level in 1999 (log)
Rev (DGFiP/INSEE)	Average net revenue per household in 1990 (log)
In,Co/Le,Sp (Corine Land Cover/MEDDE)	Fraction of a zone's surface (/100) dedicated to industry, commerce/leisure, sport facilities in 1990 (%)

[a] Stock of establishments is given for the 1st of January 2007. The range of the explanatory var. is 2005-2010.
[b] Proposed instruments present the situation in the years 1990 and 1999.

4.6 Model

4.6.1 Modeling choices

The interrelated nature of locational decisions advocates modeling them as strategic games. The specific structure of the game will clearly depend on the particular application. Draganska et al. (2008) categorize the modeling choices that affect the econometric specification, identification, and estimation along plural dimensions: (1) In the informational context, information can be complete or private. (2) In the temporal context, games can be static or dynamic. (3) Considering the timing of moves, games can involve simultaneous or sequential moves. (4) In addition, the choice may be discrete, continuous, or mixed.

Informational context: Complete vs private information

Before estimation proceeds, the researcher must specify how variable profit depends on what is observable to the establishments and to the researcher (Berry and Reiss, 2007). In general, the less the economist knows about potential entrants, establishments' expectations, agents' profits, etc., the less the researcher is able to infer from the data on market structures and the more he has to rely on untestable modeling assumptions (Berry and Reiss, 2007). The discussion can examine how empirical predictions about market structure depend on observable and unobservable economic variables.

In case of receiving imperfect information about other players' profitabilities or when the errors constitute private information for each player, each agent must form beliefs about the choices of other players. Establishments base their decisions on what they expect other players will do, where they will locate, and vice versa (Draganska et al., 2008). Due to imperfect information about other players' profitability, an establishment can only form expectations of the others' optimal location choices. Overall, an establishment's payoff from choosing a particular location depends on its expectation of the optimal location choices of its competitors and exogenous market characteristics (Zhu and Singh, 2009). Modeling establishments' discrete decisions requires describing establishments' beliefs. The indicators of other players choices are then replaced by their expectations - the probabilities of such choices (Draganska et al., 2008).

Based on the expected distribution of other agents across market locations, each establishment selects the location that maximizes its payoff given its own type (Seim, 2006). Seim (2006) solves the large choice set problem (dimensionality problem) by allowing establishments to possess private information about their own profitability. Seim (2006) sets up the model by using an imperfect information (incomplete information game) framework where players privately observe their own location-specific potential profits, but observe only the distribution of the other players' profitability.

There is often information that is a common knowledge to all the players but unobserved to the researcher. If players have common information about the discrete choices, *e.g.*, a location, then ignoring such common unobservables may create a simultaneity bias in estimating the effect of strategic interactions among these players on their payoffs from the revealed choice data. For example, establishments may all choose to enter a given market because they know of a favorable (but unobserved to the researcher) demand condition.

Temporal context: Static vs dynamic

Most researchers will agree that strategic interactions between establishments occur over time. Yet, while dynamic games are more realistic, static games are usually econometrically less complicated to estimate. It is an open question, how well static games approximate the outcomes of dynamic games (Draganska et al., 2008). The static model is assumed to approximate the repeated establishment interaction that characterizes the evolution of an industry, in part because dynamic models of industry evolution are computationally complex and difficult to estimate for markets with many product types and many establishments (Seim, 2006) or as in our case, many establishment types.

To date, there have been relatively few attempts at estimating dynamic models (Berry and Reiss, 2007; Bajari et al., 2010a). One reason for this is the "curse of dimensionality" that makes computation take a long time. To compare static and dynamic approaches, we suggest to follow several excellent surveys: (1) the review on emerging literature on econometric analysis of static and dynamic models of strategic interactions by Bajari et al. (2013), (2) the survey on dynamic discrete choice structural models by Aguirregabiria and Mira (2010), (3) the survey on dynamic models on strategic interactions by Doraszelski and Pakes (2007), or (4) on static models by Berry and Reiss (2007).

Dynamic approach would require to model the behavior of players as forward-looking establishments that maximize some form of discounted profits. In general, establishments either know the choices that other players will make or they have to form conjectures about the other players' actions.

Timing of moves: Games with simultaneous vs sequential moves

Each player needs to form conjectures about the reaction of the others. If players move sequentially and the sequence of moves is observed, then that information can be used to inform the parameters. Otherwise, the researcher needs to integrate over all possible sequence combinations causing the problem to be more pronounced (Draganska et al., 2008).

Possible approaches

A number of methods can be applied to model strategic interactions. In the literature, structural models are often estimated using the nested fixed-point algorithm. While it is a useful approach, it often requires considerable central processing unit time and numerical sophistication on the part of the econometrician (Bajari et al., 2013). In the review of Draganska et al. (2008), we read that the computational cost of the nested fixed-point algorithm has prompted the development of alternative methods, such as the two-step approach of Bajari et al. (2007) and the nested pseudo likelihood by Aguirregabiria and Mira (2007). Alternatives to the likelihood-based approaches are generalized method of moments estimators. These estimators rely on the specification of moment conditions based on a structural game.

The approach proposed by Bajari et al. (2007) or Bajari et al. (2013), the two-step approach, has the disadvantage of being inefficient as compared to the nested fixed-point. However, it has a practical advantage of being much simpler and more accurate numerically. In addition, the two-step approach solves the multiplicity problem by letting the data tell us which equilibrium is played if the model generates multiple outcomes.

In our current application, we set up the model using an imperfect-information (incomplete information game) framework in a static profit maximizing context (see Seim, 2006) where players privately observe their own location-specific potential profits, but observe only the distribution of the other players' profitability. We develop a baseline model at the group level estimating a Bayesian Nash Equilibrium. We will describe this approach in details in the model subsection 4.6.3.

4.6.2 Unique features of our modeling approach

Our modeling approach is a combination of ideas proposed by Ellickson and Misra (2011, 2012), Aguirregabiria and Mira (2007), Zhu and Singh (2009), Jia (2008), and Leontief and includes several nice features: (1) We incorporate strategic interactions between several establishment types and present them in a form of a matrix. (2) We account for the spatial spillover effects from the neighboring zones. (3) We correct for the price endogeneity.

Matrix of strategic interactions

In the present paper, we draw from the concept of a matrix à la Leontief[8]. A great deal of empirical work has been done to identify coefficients at the level of regional and national economies. Leontief's input-output model is a quantitative economic technique that depicts inter-industry relationships within an economy, showing how output from one sector may become an input to another sector. In the inter-industry matrix, column entries typically represent inputs to an industrial sector, while row entries represent outputs from a given sector. This format therefore shows how dependent each sector is on every other sector, both as a customer of outputs from other sectors and as a supplier of inputs. Each column of the input-output matrix shows the monetary value of inputs to each sector and each row represents the value of each sector's outputs.

In our current application, we draw from the concept of Leontief in a way, that we represent interdependencies between various establishment types in a form of a matrix. Yet, due to the lack of information on sales values, purchases, inventories, or production levels, we do not represent them in a form of monetary relations. Our focus is to rather understand the linear relationship between two different establishment types, thus how these two types behave as a pair, whether the relation is positive, negative, or whether there is no linear relationship between a particular pair of establishment types. We provide the direction of influence between the establishment types, the level of magnitude of these strategic interrelations, and establish possible asymmetries in establishments' decisions. These effects are analyzed at the regional level for the Paris region economy.

[8]Wassily Leontief earned the Nobel Prize in Economics for his development of the input-output model. An understanding of the economy as consisting of linked sectors goes back to the French economist François Quesnay and was fully developed by Walras in 1874 (Walras, 1874).

The structure of a matrix of observed explanatory variables summarized in Table 4.5 is described next. Spatial spillovers are modeled using a definition of accessibility measure. Potential accessibility measures (called gravity-based measures) have been widely used in urban and geographical studies since late 1940s (see, *e.g.*, Stewert, 1947; Hansen, 1959; Ingram, 1971; Vickerman, 1974). The potential accessibility measure estimates the accessibility of opportunities in zone $\mathtt{i}$ to all other zones in which more distant opportunities provide diminishing influences. The negative exponential distance has been the most often used (Geurs and van Wee, 2004), yet, various other impedance functions, such as power, Gaussian, or logistic functions can be also proposed. Publications advocating the use of the logsum have also appeared (see, *e.g.*, Cochrane, 1975; Williams, 1977).

We represent spatial spillovers in the following form:

$$x_{l,s} = \ln\left(\sum_{j=1}^{L} e^{-\mu d_{l,j}} z_{j,s}\right) \tag{4.1}$$

assuming a negative exponential distance function to indicate that some amount of interaction intensity falls with each unit of distance. This matrix contains variables $z_{j,s}$ directly concerning either location l or establishment type s and the number of pre-existing establishments. In the current application, we fix μ to one. The value of $\mu < 1$ would denote more global spillovers and the value of $\mu > 1$ more local spillovers[9]. In addition, as it models the range of spatial spillover effects, the value of μ may differ along with establishment types. The values that the parameter μ may take is left aside for further investigation. $d_{l,j}$ is the Euclidean distance between the centroids of the EGT zones l and j. Other distance measures, such as travel time, could be proposed (see Buczkowska et al., 2016)[10].

[9]See the example of distance decay coefficient of minus 0.1 used in the application by Chatman et al. (2016).

[10]Buczkowska et al. (2016) investigate this issue in the case of establishment locational choices. The authors suggest that high congestion, speed limits, or physical uncrossable barriers can diminish or totally eliminate the linkage between neighboring areas challenging the choice of the Euclidean distance in representing spatial spillovers. They develop a probabilistic mixture of hurdle-Poisson models for several activity sectors. Seven metrics are considered: Euclidean distance, the travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances.

Price endogeneity

We extend the baseline model by accounting for the endogeneity of business real estate prices. To predict urban prices, studies, such as the ones of Chakir and Parent (2009) or Chakir and Le Gallo (2013) propose to use, *e.g.*, the population density, per capita income, or the distance to the closest major urban centers.

We use Feasible Generalized Least Square method considering the following variables as instruments for the urban prices of offices and shops: (1) average net revenue per household from the year 1990 (DGFiP/INSEE), (2) population and employment levels from 1999 (INSEE), (3) land use proxies using the Corine Land Cover database of MEDDE from the year 1990 which include the fraction of a zone dedicated to the industrial and commercial activities, road and railway network, airports, materials extraction sites, construction sites, urban green areas, and leisure and sport facilities.

In case of revenue oriented establishments, such as retail, service, or entities dedicated to professional activities, we predict the urban prices and rents. Being more precise, for the professional, scientific, technical activities, financial activities, and hotels and restaurants, office rents are predicted and further used in the location choice models. In case of the commerce and transport and storage industries, we predict the shop rents. In case of goods-producing establishments, such as entities involved in the manufacturing industry, the focus is put on cost. Thus, a labor cost is supposed to be generally more significant than the level of rent. For this reason, we decide to predict the wage of blue-collar workers to be included in the location models of manufacturing establishments.

To correctly construct a model for each sector, we test various possible combinations of instruments among the ones just described. Variables indicating the level of population and employment in a particular EGT zone in 1999, average income per person from 1990, land dedicated to the industrial and commercial use, and land purposed for the leisure and sport activities in the year 1990 turn out to be the most appropriately working instruments in our application.

An indirect expected profit function now writes as:

$$\begin{aligned}\mathbb{E}\left(\pi_{i_s,l}\right) &= V_{s,l}\left(x_{s,l},\boldsymbol{\theta}\right) - \gamma_{s,l}p_l \\ &\quad + \textstyle\sum_{j=1}^{S}\tilde{n}_j\sum_{m=1}^{L}\alpha_{s,j,l,m}\mathbb{E}\left(y_{i_j,m}\right) + \eta_l + \varepsilon_{i_s,l}.\end{aligned} \tag{4.2}$$

A long-run pure and perfect competition in price between locations is represented as:

$$p_l = mc_l + \nu_l, \tag{4.3}$$

$$\mathbb{E}\left(\eta_l \nu_l\right) = \sigma_{\eta,\nu}. \tag{4.4}$$

4.6.3 Model specification

We assume that there are S establishment types. In our application, an establishment type is equivalent to a particular sector or a part of it. The stock of establishments in sector s is known and defined as n_s. We also assume a static and simultaneous move setting at the disaggregate level. We refer to Ackerberg et al. (2005) and Aguirregabiria and Mira (2010) for a detailed review of how type of strategic games are affecting choices and resulting equilibria. An establishment i from sector s locating at l is endowed with the following profit function:

$$\pi_{s,i_s,l} = V_{s,l} + \sum_k \sum_{i_k \neq i_s} \alpha_{s,k} \mathbb{E}\left(y_{i_k,l} = 1\right) + \xi_{s,l} + \varepsilon_{s,i_s,l}. \tag{4.5}$$

This paper discusses a strategic game with incomplete information, *i.e.*, a Bayesian Nash game, assuming within-group homogeneity and heterogeneity between groups (Ellickson and Misra, 2011). The specification of the profit function considers that all establishments from one sector have, up to $\varepsilon_{s,i_s,l}$, the same profit function. As stated by Aguirregabiria and Mira (2007), they are defined as "global players", *i.e.*, they are assumed to be homogeneously defined within an industrial sector (or a part of it) s. We recognize that it is a stringent assumption as it constrains to define homogenous groups, which is not here really the case. We are limited, however, by the available data and our application might be considered as a first step towards a more realistic representation of location choices while accounting for strategic location choice behaviors. Establishments are assumed to have different behaviors along with sectors s. We also might extend our framework to account for intra-group heterogeneity by better modeling the taste heterogeneity. Yet, again, we do not have available sub-group a priori known probabilistic descriptions of such behaviors.

$V_{s,l}$ is the systematic part of the profit function. $V_{s,l}$ is modeled as a linear in parameters function of covariates x that only depend on s and l, *e.g.*, rents of offices, shops, market potential (client attraction), land uses:

$$V_{s,l} = x'_{s,l}\boldsymbol{\theta}. \tag{4.6}$$

$\sum_k \sum_{i_k \neq i_s} \alpha_{s,k} \mathbb{E}\left(y_{i_k,l} = 1\right)$ is modeling how an establishment anticipates how its profit is likely to be affected by location choices of other establishments. Since each establishment has no perfect information about what the decisions of others will be, it is a function of their expected choices. Note that in case of perfect information, there is no need to anticipate location choices and cumulated sums just recognize the actual number of installed establishments by sector. As it will be explained in the application, we consider focusing on behaviors of newcomers conditional to already existing establishments.

$\xi_{s,l}$ is an unobserved (by the modeler) component that captures sector and location effects. It is not establishment specific. It can be correlated to real estate location-specific rents or prices. It might be the case that prices/rents (or any related measures) are not the only endogenous variables of the main estimation problem. We here assume that real estate prices/rents are the only endogenous variables. We first estimate a set of auxiliary price regressions from which we compute estimated prices that we further use as exogenous variables in the location choice model.

In addition, it is assumed that these unobservables are correlated across sectors but not across locations. It is a specific assumption since there is no spatial correlation arising from sector-location specific unobservables. We recognize that it is an important topic which is left aside for future research. Unobservables are assumed to be normally distributed as follows:

$$\forall s, l, \xi_{s,l} \overset{iid}{\longrightarrow} \mathcal{N}\left(0, \sigma_s^2\right) \tag{4.7}$$

and

$$\forall s, m, l, k, \mathbb{C}\mathrm{ov}\left(\xi_{s,l}, \xi_{m,k}\right) = \begin{cases} \sigma_{s,m} \text{ if } l = k \\ 0 \text{ otherwise.} \end{cases} \tag{4.8}$$

Finally, $\varepsilon_{s,i_s,l}$ is an establishment specific error term independent from V and ξ. Assuming that $\varepsilon_{s,i_s,l}$ are iid Gumbel distributed and that establishments choose locations to maximize profits at the same time, the (unconditional to $\boldsymbol{\xi}$ unobservables) probability that an establishment of sector s locates at l is defined as:

$$\begin{aligned}&\Pr_{s,l}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|n,x,\Pr)=\\&\int_{\mathbb{D}(\boldsymbol{\xi}_s)}\frac{\exp\left(x'_{s,l}\boldsymbol{\theta}+(n_s-1)\alpha_{s,s}\Pr_{s,l}+\sum_{k\neq s}n_k\alpha_{s,k}\Pr_{k,l}+\xi_{s,l}\right)}{\sum_{m=1}^{L}\exp\left(x'_{s,m}\boldsymbol{\theta}+(n_s-1)\alpha_{s,s}\Pr_{s,m}+\sum_{k\neq s}n_k\alpha_{s,k}\Pr_{k,m}+\xi_{s,m}\right)}f(\boldsymbol{\xi}_s|\boldsymbol{\sigma})\,d\boldsymbol{\xi}_s,\end{aligned}\tag{4.9}$$

where $\Pr_{s,l}\equiv\Pr_{s,l}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|n,x,\Pr)$. Such expressions also define ”market shares” of locations. Note that these are best response probability functions which map expected profits (conditional on beliefs characterized by Pr) into (ex ante) choice probabilities (Ellickson and Misra, 2011). A Nash equilibrium (Nash, 1951) is a strategy profile in which each player's strategy is a best reply to the others' strategies (see Bicchieri, 1995).

We observe vectors of locations of establishments $d_{s,1},\cdots,d_{s,L},\prod_l d_{s,l}=n_s,\forall s=1,\cdots,S$. The likelihood function is defined as:

$$\ell(d|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma},n,x,\Pr)=\prod_s\prod_l\Pr_{s,l}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|n,x,\Pr)^{d_{s,l}}.\tag{4.10}$$

We maximize the log-likelihood function subject to the fixed point problem:

$$\begin{aligned}&\max_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}}\ln\ell(d|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma},n,x,\Pr)\\&\text{s.t. }\Pr_{s,l}(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|n,x,\Pr)=\\&\int_{\mathbb{D}(\boldsymbol{\xi}_s)}\frac{\exp\left(x'_{s,l}\boldsymbol{\theta}+(n_s-1)\alpha_{s,s}\Pr_{s,l}+\sum_{k\neq s}n_k\alpha_{s,k}\Pr_{k,l}+\xi_{s,l}\right)}{\sum_{m=1}^{L}\exp\left(x'_{s,m}\boldsymbol{\theta}+(n_s-1)\alpha_{s,s}\Pr_{s,m}+\sum_{k\neq s}n_k\alpha_{s,k}\Pr_{k,m}+\xi_{s,m}\right)}f(\boldsymbol{\xi}_s|\boldsymbol{\sigma})\,d\boldsymbol{\xi}_s.\end{aligned}\tag{4.11}$$

As stated by Ellickson and Misra (2011):

> ”*[...] this system [...] of nonlinear equations is guaranteed to have a solution by Brouwers fixed point theorem. Moreover, this fixed point representation provides a direct method of solving for equilibria: the method of successive approximations (i.e. fixed point iteration). Thus, one possible estimation strategy (originally proposed by Rust (1994)) is another full solution approach that first solves this system of equations (for a given set of parameters) and then matches the predicted conditional choice probabilities to the choices observed in the data. This is essentially a static games version of Rust's nested fixed point (NFXP) algorithm (Rust, 1987)*”.

We implement a nested fixed point maximum likelihood estimator. The estimation routine requires solving the fixed point problem. It is carried out by successive approximation. We keep in line with most of the existing literature in that we are not searching for all possible fixed points. We presume that only one equilibrium is played in the data. As we use estimated prices as explanatory variables, we have to adjust standard errors. We do so by bootstrap.

4.7 Results

This section reports estimation results for seven location choice models run simultaneously for seven establishment types. Establishments are divided based on their activity sector into the following groups: (1) retail establishments of all size classes, (2) wholesale establishments, (3) hotels and restaurants, (4) manufacturing establishments, (5) entities involved in transport and storage, (6) financial establishments, and (7) establishments dedicated to professional, scientific and technical activities.

4.7.1 Results: Estimates

Strategic interaction models can become a powerful tool for location analysis. The right location choice can make a big difference for the newly-created business. To test which factors are the most critical in the location choices of establishments, we include a variety of variables that capture demand, cost, local agglomeration measures, such as access to the workforce, demographic characteristics, physical quality, availability of vacant land, availability of dedicated floor space, accessibility to transport infrastructures (access to public transit networks, road networks, or proximity to airports), access to other amenities, such as universities, services and commerce, within-industry economies (own-industry competition effects versus Marshallian agglomeration economies, *i.e.*, own-sector collocation), and inter-industry economies (Jacobsian agglomeration economies, *i.e.*, economic diversity).

Depending on the analyzed establishment type, location can be a major determinant of revenue or cost. In the case of services or commerce, location is a main determinant of revenue and costs will be relatively constant for a given area. To boost the revenue, a focus should be laid on the high customer- or high user-contact issues, high traffic flows, the level of purchasing power, and finally the compatibility of products or services with demographics of a drawing area. In our empirical application, a proximity to customers or users is reflected by the level of high population density which is likely to boost the revenue of commercial establishments, hotels and restaurants. In addition, high traffic flows have been proxied by the number of trips with the purpose of both everyday and occasional shopping. Areas where their potential customers do their shopping turn out to be critical for the entities in the retail industry. The areas with a high number of population traveling with the purpose of visiting a restaurant seems to positively influence the successful prospering of restaurants and the hotel industry.

Next, the physical quality, such as facilities or access also play a role. Convenient transport infrastructures seem to play a considerable role in the location decision of establishments in general. In our application, we proxy the access to the transport infrastructure. For the units in retail, hotels and restaurants, the professional, scientific, technical activities sector, and the financial industry, a good access from and to the subway stations and bus stops are of a very high value. Access to public transport plays also a positive role in case of the manufacturing and wholesale sectors. In the main, a good road network (car accessibility), accessibility to rail stations, and airports seem to have a strong positive influence on the location choices of establishments in the transport and storage sector. For the more detailed discussion on the relation between access to transport infrastructure (transit accessibility, road transport infrastructure, etc.) and the new establishments' formation and location in the US, Spain, and Portugal, see works of Chatman et al. (2016), Holl (2004a, b), Melo et al. (2010), and Mejia-Dorantes et al. (2012).

Other physical qualities, such as facilities and neighboring businesses, security, or image are generally relevant in the location choices of establishments. For example, availability of space dedicated to the commercial use is important in the retail sector. Manufacturing, wholesale, and transport, storage opt for the proximity to the areas dedicated to extraction of materials. Units involved in manufacturing appear to avoid zones highly dedicated to economic or industrial activities. Yet, we acknowledge that two separate variables could be proposed instead which would better proxy the economic and industrial activities.

Paris tends to give incentives to establishments from most of the analyzed sectors. Paris appears to have a negative influence on the establishments involved in manufacturing and have no particular effect on retail and hotels and restaurants. La Défense attracts establishments from all the sectors, beside wholesale. In addition, retail establishments seem not to have a preference over a particular zone of the Paris region. New Cities are prone to faze a negative effect on the entities involved in professional, scientific and technical activities, and hotels and restaurants, whereas New Cities have a positive effect on other sectors. Establishments in the manufacturing, transport, storage, and retail sectors seem not to mind locating in the outer suburbs of the Paris region. This is not the case for financial units, establishments in professional, scientific and technical activities, wholesale, and hotels, restaurants.

The cost determinants themselves are rents, management quality, and operation policies, such as wage levels. We observe that overall, all new establishments tend to evade paying high real estate prices. For all the sectors being analyzed, this effect is significant and negative.

In the further analysis of establishments whose strategy is based on innovation or specialized knowledge, the key attributes that could be tested are mixture of critical mass of information, ready supply of skilled, educated workforce with large pool of engineers and managers, high-quality and specialized inputs, an environment that encourages investment and local rivalry, a sophisticated local market, and a local presence of related and supporting industries. In our application, financial establishments and units dedicated to the professional, scientific, technical activities seem to strongly prefer sites that provide an easy access to white-collar workers and managers. Also the zones and their surroundings where professional meetings take place relatively often tend to strongly attract the financial establishments. Similarly, higher number of professionals and persons involved in the academic activities positively influence the decisions of units in the professional, scientific, technical activities sector.

Labor availability, *i.e.*, labor force sources, proxied by a high density of total active population can also raise the profit of manufacturing establishments, which is in line with the results obtained by, *e.g.*, Chatman et al. (2016). Labor availability seems to increase the efficient functioning and to have a positive effect on the prosperity of establishments in the transport and storage sector.

As Zhu and Singh (2009) highlight, significant asymmetries across players in their response to market conditions and interactions are recognized. We control thus for within-industry establishments. We observe that an establishment does not seem to act in isolation. We acknowledge that all establishment types exert a strong negative impact on the pre-existing establishments of exactly the same type. In our analysis, in case of the transport and storage, and hotels and restaurants sectors this effect is insignificant. We thus validate the findings of Chatman et al. (2016), who demonstrate that the presence of competitive establishments in a particular industry (establishments in the same industry category) seems to negatively predict the number of new establishments' formation in that industry. According to Chatman et al. (2016), the fact that the number of establishments in the same sector category tends to negatively predict the number of establishment births in that sector suggests that the own-industry competition effect overrides any localized Marshallian agglomeration economies. We can ratify the result of Seim (2006) and Zhu and Singh (2009) that the competitive effects have a negative influence on the agent's situation.

We also observe the impact of asymmetries in competitive effects across locations when competing establishments that belong to different types collocate. The result of Zhu and Singh (2009) that establishments care about the actual identities of their competitors rather than simply the total number of competitors can be confirmed. We notice the effects of negative inter-industry competition of the professional, scientific, technical activities sector on the financial sector. Similarly the pre-existing establishments in the commercial sector tend to push out entities dedicated to transport, storage and manufacturing to locate their new units nearby. Whereas, the positive inter-industry economies are identified in case of pre-existing manufacturing entities attracting new commercial units (both retail and wholesale) and newly-created establishments in the transport and storage industry. Units dedicated to real estate tend to attract the financial establishments to locate nearby. Zones occupied by higher number of pre-existing hotels and restaurants are often chosen also by new entities in commerce and manufacturing.

To control for the size of a zone, we include a variable indicating its surface which turns out to be significant and positive for all the analyzed types. In addition, it seems that establishments account for the spatial spillover effects from the surrounding zones when looking for the place to locate in the market.

One clarification should be made at this point. In some industries, such as the call center industry, neither face-to-face contact nor movement of materials are required. Thus, this type of establishments may have very broad location options. Traditional variables would be probably no longer relevant and cost and availability of labor may be driving in this case location decisions. We do not include this kind of establishment types in our analysis. Nor the units, such as schools, universities, hospitals, or public administration are considered. We prefer instead to focus on a large variety of establishments whose decisions will be relatively strongly steered by the market forces.

4.7.2 Results: Matrix of strategic interactions

We incorporate strategic interactions between several establishment types and present them in a form of a matrix to show how dependent each establishment type is on every other type.

Table 4.6: Models estimates: Seven location choice models run simultaneously

	Retail		HoRes		Finan		ProSci		Whole		TranSt		Manuf	
	Est	t-st	Est	t-st	Est	t-st	Est	t-st	Est	t-st	Est	t-st	Est	t-st
Paris	0.0012	0.36	-0.0003	-0.04	0.3662	90.19	0.1673	36.06	0.1369	293.58	0.3098	50.50	-0.0680	-13.39
La Défense	0.0904	11.68	0.1160	86.65	0.3814	432.02	0.0284	8.14	-0.0135	-3.77	0.0908	34.37	0.2694	288.35
New Cities	0.0592	11.02	-0.0658	-28.05	0.0433	18.57	-0.1664	-40.99	0.0747	34.56	0.2056	82.67	0.1246	160.80
Outer suburbs (excl. New Cities)	0.0370	5.55	-0.0393	-30.88	-0.0456	-13.52	-0.0353	-9.24	-0.0120	-3.33	0.0396	8.49	0.1572	48.41
Zone's surface (log)	0.1659	22.37	0.4644	895.97	0.1592	18.29	0.1554	81.79	0.3782	81.02	0.2797	58.36	0.4943	140.15
Stock: The same estab. type	-0.0004	-8.66	-0.0001	-1.28	-0.0032	-24.13	-0.0005	-41.31	-0.0006	-11.25	<.0001	0.13	-0.0005	-2.87
Stock: Commerce (G)			-0.0001	-1.23							-0.0003	-4.02	-0.0003	-4.17
Stock: Hotels, restaurants (I)	0.0002	1.71							0.0008	8.15	<-.0001	-0.08	0.0006	9.62
Stock: Finance, insurance (K)							-0.0001	-0.84						
Stock: Pro., scien., tech. (M)					-0.0007	-3.66								
Stock: Real estate (L)					0.0018	4.75								
Stock: Manufacturing (C)	0.0005	6.54							0.0008	8.05	0.0007	4.69		
Active density (labor force)/1000											0.2929	41.46	0.4722	262.70
White-collar/manager (%tot emp)					0.5029	90.90	0.5967	284.72						
Trips: Prof. meeting					0.1549	33.71								
Trips: Prof. meeting/university							0.0825	17.17						
Total pop. density/1000	0.1241	17.10	0.3855	246.60					0.3191	298.98				
Trips: Shopping purpose	0.2557	187.95												
Trips: Restaurant visit			0.0712	14.26										
Land: Shops (%zone)	0.0426	16.82							-0.0004	-0.18				
Land: Industrial. economic act.											-0.0002	-0.03	-0.0184	-2.78
Land: Extraction of materials									0.0099	5.85	0.0177	4.37	0.0098	23.50
Road. rail terminal (%zone)											0.0109	5.28		
Airport (%zone)											0.0073	9.94		
Access to public transport (log)	0.0371	16.73	0.0859	15.95	0.1410	21.16	0.1655	28.32	0.0136	3.63			0.0370	10.37
Predicted rent (offices or shops)	-0.0387	-18.10	-0.1113	-63.76	-0.2468	-117.43	-0.2753	-454.27	-0.1560	-64.55	-0.0733	-19.15	-0.0322	-6.07
Instruments to predict rents[a]	Pop In,Co	Rev Le,Sp	Emp In,Co	Rev Le,Sp	Emp In,Co	Rev Le,Sp	Emp	Rev	Pop In,Co	Rev	Emp In,Co	Rev Le,Sp	Emp In,Co	Rev

[a] Proposed instruments: Pop/Emp: Population/Employment level in 1999 (log); Rev: Average net revenue per household in 1990 (log); In,Co/Le,Sp: Fraction of a zone's surface/100 dedicated to industry, commerce/leisure, sport facilities in 1990 (%).

Three results merit being highlighted: (1) Location choices of newly-formed establishments seem to be interrelated with the choices of other new establishments that are expected to install their activities in the market. (2) As shown in the subsection 4.7.1, all the newly-formed establishments of one type exert a strong negative impact on their pre-existing competitors of the same type beside the transport and storage sector. Yet, we also observe that all the new establishments prefer to enter the market where other new establishments of the same type are expected to locate. Values on the diagonal of matrix 4.7 indicate the positive intra-sectoral reactions between the newly-created establishments. (3) Once again, when analyzing the interactions between the newly-created establishments of different types, an impact of asymmetries in establishments' decisions can be acknowledged which allows us to validate the results of Zhu and Singh (2009) or Jia (2008). Detailed relations can be read through the values off diagonal of matrix 4.7. In addition, the results of the variance-covariance matrix of residuals are reported in Table 4.8. Variances are given along the main diagonal and covariances are the off-diagonal elements of the matrix.

When looking closer at the reaction of each type on every other type, we notice that intensified openings of transport and storage units seem to be positively influenced by decisions of retail and manufacturing establishments, and negatively by new creation of establishments involved in financial and professional, scientific and technical activities. A decision of setting up a new retail store seems to be positively influenced by potential openings of both establishments involved in transport and storage and in professional, scientific and technical activities. Yet, retails tend to keep at a respectful distance from the sites that are about to be selected by manufacturing entities. Hotels and restaurants seem to follow the ideas of establishments involved in commerce (retail and wholesale alike) and tend to choose locations where new manufacturing units will potentially set up their units. Hotels and restaurants tend to get discouraged from choosing the sites which are about to be occupied by the financial units and establishments in professional, scientific and technical activities, and transport and storage. Overall tendencies of entities dedicated to the financial activities indicate that these units are prone to be discouraged by most of the sectors beside transport and storage. On the other hand, units in the wholesale or in the professional, scientific and technical activities sector tend not to exhibit strong concerns over the type of establishment which are expected to locate in a zone. Finally, newly established commerce seems to have a negative influence on the locational decisions of manufacturing entities, so do the professional, scientific and technical services. Zones to be chosen by hotels and restaurants, as well as new financial establishments tend to send a positive signal to enforce the decision of manufacturing units to locate nearby.

Table 4.7: Strategic interactions matrix: How the decision of one establishment type affects the decision of other establishment types.

	Retail	HoteRes	Finan	ProSci	Whole	TranSto	Manuf
Retail	0.0067[a] (15.22)	0.0022[b] (3.54)	-0.0004 (-1.92)	0.0006 (5.42)	0.0017 (5.38)	0.0015 (3.15)	-0.0036 (-5.62)
HoteRes	-0.0006 (-0.61)	0.0182 (36.62)	-0.0025 (-4.49)	0.0006 (3.59)	0.0005 (0.51)	0.0003 (0.11)	0.0024 (2.21)
Finan	0.0005 (0.63)	-0.0114 (-5.96)	0.0225 (29.03)	0.0051 (4.69)	0.0010 (2.04)	-0.0038 (-3.28)	0.0031 (3.27)
ProSci	0.0019 (6.61)	-0.0042 (-6.63)	-0.0057 (-10.58)	0.0085 (107.56)	0.0005 (2.84)	-0.0056 (-10.00)	-0.0019 (-3.66)
Whole	-0.0004 (-0.73)	0.0013 (1.99)	-0.0009 (-2.10)	0.0005 (3.48)	0.0102 (26.49)	0.00003 (0.06)	-0.0029 (-3.11)
TranSto	0.0035 (5.93)	-0.0059 (-13.47)	0.0023 (6.82)	0.00002 (0.21)	-0.0021 (-6.59)	0.0253 (18.54)	0.0002 (0.15)
Manuf	-0.0021 (-2.79)	0.0048 (5.04)	-0.0058 (-7.92)	0.0015 (5.38)	0.0029 (4.66)	0.0020 (1.73)	0.0206 (13.17)

[a] Influence from type 1 on type 1. Upper value stands for the estimate and the lower value for t-ratio.
[b] Influence from type 1 on type 2. etc.

Table 4.8: Variance-covariance matrix of residuals.

	Retail	HoteRes	Finan	ProSci	Whole	TranSto	Manuf
Retail	0.0079[a] (0.84)	0.1027 (75.97)	-0.2133 (-54.04)	0.3612 (51.69)	0.0426 (22.09)	-0.2226 (-84.28)	-0.0436 (-7.31)
HoteRes		0.3843 (69.72)	-0.6597 (-298.43)	-0.0166 (-4.82)	0.4869 (331.50)	0.3828 (104.04)	0.2404 (180.10)
Finan			-0.0930 (-13.55)	0.3484 (370.16)	-0.0707 (-30.80)	-0.5531 (-283.17)	0.6444 (458.93)
ProSci				0.1504 (132.77)	0.2502 (198.17)	-0.0871 (-44.43)	-0.2066 (-42.76)
Whole					-0.3201 (-164.49)	-0.3220 (-619.88)	0.1287 (20.10)
TranSto						-0.0471 (-87.42)	0.4491 (82.98)
Manuf							-0.3043 (-109.98)

[a] Upper value stands for the estimate and the lower value for t-ratio. Variances are given along the main diagonal. The off-diagonal elements of the matrix provide the covariances.

4.8 Conclusions and possible applications

We found it necessary to vitalize the still deficient empirical literature on strategic interactions in establishment location choice models. By observing real-world-like examples, we focused on finding characteristic patterns in several establishment types' placements.

The main objective of this paper was to implement the strategic interactions between the newly-created establishments regarding their location choice relative to their activity sector. We estimated a simultaneous discrete game between seven types of establishments. Development of a simple baseline model at the group level was our starting point. We accounted for the spatial spillovers and dealt with the endogeneity of business real estate prices issue.

We reported that establishment location choices differ by type and that all establishments exert a strong negative impact on their pre-existing competitors of the same type. Yet, all the new establishments prefer to enter the market where other new establishments of the same type are about to locate. This is a new and stimulating finding. We observe an impact of asymmetries in strategic interactions between establishment types across locations. Variables that capture demand and cost factors, local agglomeration measures, such as access to the workforce, demographic characteristics, availability of dedicated space, access to amenities, and access to transport infrastructure turn out to be relevant in establishment location decisions.

The level of analysis has not reached the high methodological level. There is a need for methodological advances in order to be able to model realistically the complexity of establishment decision-making processes. Draganska et al. (2008) provide a broad discussion of the problems and opportunities that exist in this research domain and hope that their review will spur new developments in the years to come. There is still a scope for further research and many fertile fields yet to be cultivated in a number of areas, such as dynamic aspects of strategic interactions.

Interaction models and strategic interaction models, in particular, provide a powerful tool for location analysis (Ghosh and McLafferty, 1987). Interaction modeling can be potentially used in a wide variety of geographical problems and situations (Pooler, 1994). Also game theory has strongly influenced some fields, such as industrial organization, where it helps to transform the analysis of market interactions (Bajari et al., 2010a).

In theoretical models in industrial organization, game theory is by far the most common tool used to model industries. Game theory can be used to model a very broad set of economic problems (Bajari et al., 2013). Yet, the domain of applicability is not limited to industrial organization. These models could be applied to estimate seemingly complicated structural models in other fields, such as macroaconomics, labor, public finance, or marketing (Bajari et al., 2013). The continued use of spatial interactions and relative location models by city planners, transport analysts, retail location firms, shopping center investors, plan developers, and urban social theorists is without precedent (Haynes and Fotheringham, 1984). Our method, once developed to account for the strategic interactions in location choice model, can be furthered applied to, *e.g.*, the housing choices or driving behavior.

References

Ackerberg, D., Benkard, C., Berry, S., and A. Pakes (2005). Econometric Tools for Analyzing Market Outcomes, in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, vol. 6A, chap. 63, 4171-4279. North-Holland, Amsterdam.

Aguirregabiria, V. and P. Mira (2007). Sequential Estimation of Dynamic Discrete Games. *Econometrica*, 75, 1-54.

Aguirregabiria, V. and P. Mira (2010). Dynamic Discrete Choice Structural Models: A Survey. *Journal of Econometrics*, 156, 38-67.

Aguirregabiria, V. and J. Suzuki (2016). Empirical Games of Market Entry and Spatial Competition in Retail Industries. In *Handbook on the Economics of Retailing and Distribution*, ed. by E. Basker, Edward Elgar, University of Missouri-Columbia, US.

Anselin, L. (2010). Thirty Years of Spatial Econometrics. *Papers in Regional Science*, 89(1), 3-25.

Aoki, M. (1995). Economic fluctuations with interactive agents: Dynamic and stochastic externalities. *The Japanese Economic Review*, 46(2), 148-165.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010a). Estimating static models of strategic interactions. *Journal of Business & Economic Statistics*, 28(4), 469-482.

Bajari, P., H. Hong, and D. Nekipelov (2013). Game Theory and Econometrics: A Survey of Some Recent Research. In *Advances in Economics and Econometrics*, vol. 3, Econometrics, ed. by D. Acemoglu, M. Arellano, and E. Dekel. Publisher: Cambridge University Press.

Bajari, P., H. Hong, and S.P. Ryan (2010b). Identification and Estimation of a Discrete Game of Complete Information. *Econometrica*, 78(5), 1529-1568.

Berry, S. and P. Reiss (2007). Empirical Models of Entry and Market Structure. In *Handbook of Industrial Organization*, ed. by M. Armstrong and R. Porter, Chapter 29, 3, 2007, 1845-1886.

Bicchieri, C. (1995). The Epistemic Foundations of Nash Equilibrium. In *On the Reliability of Economic Models: Essays in the Philosophy of Economics*, ed. by D. Little, Kluwer 1995, 42, 91-146.

Blume, L. (1993). The statistical mechanics of strategic interaction. *Games and Economics Behavior*, 5, 387-424.

Buczkowska, S. and M. Lapparent (de) (2014). Location choices of newly-created establishments: Spatial patterns at the aggregate levels. *Regional Science and Urban Economics*, 48, 68-81.

Buczkowska, S., N. Coulombel, and M. Lapparent (de) (2016). Euclidean distance versus network distance in business location: A probabilistic mixture of hurdle-Poisson models. *Working Paper*, Université Paris-Est, Marne la Vallée, France.

Buczkowska, S. and N. Coulombel (2017). Locational strategies of multi-store firms. *Working paper*.

Casanova, M.R., V. Orts, and J.M. Albert (2017). Sectoral scope and colocalisation of Spanish manufacturing industries. *Journal of Geographical Systems*, 19, 65-92.

Chakir, R. and O. Parent (2009). Determinants of land use changes: A spatial multinomial probit approach. *Papers in Regional Science*, 88(2), 327-344.

Chakir, R. and J. Le Gallo (2013). Predicting land use allocation in France: A spatial panel data analysis. *Ecological Economics*, 92, 114-125.

Chang, H.-J. and C.-M. Hsieh (2014). A TOPSIS Model for Chain Store Location Selection. *Review of Integrative Business and Economics*, 4(1), 410-416.

Chatman, D.G., R. Noland, and N.J. Klein (2016). Firm Births, Access to Transit, and Agglomeration in Portland, Oregon, and Dallas, Texas. *Transportation Research Record*, 2598, 1-10.

Cochrane, R.A. (1975). A possible economic basis for the gravity model. *J. Transport Economics and Policy*.

Chu, C.Y. Cyrus and H.-ch. Lu (1998). The multi-store location and pricing decisions of a spatial monopoly. *Regional Science and Urban Economics*, 28(3), 255-281.

Doraszelski, U. and A. Pakes (2007). A framework for applied dynamic analysis in IO. In *Handbook of industrial organization*, ed. by M. Armstrong and R. Porter, 3, 1887-1966. Amsterdam:

North-Holland.

Draganska, M., S. Misra, V. Aguirregabiria, P. Bajari, L. Einav, P. Ellickson, D. Horsky, S. Narayanan, Y. Orhun, R. Reiss, K. Seim, V. Singh, R. Thomadsen, and T. Zhu (2008). Discrete Choice Models of Firms' Strategic Decisions. *Marketing Letters*, 19(3-4), 399-416.

Duranton, G. and H.G. Overman (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, 72, 1077-1106.

Ellickson, P. and S. Misra (2011). Estimating Discrete Games. *Marketing Science*, 30(6), 997-1010.

Ellickson, P. and S. Misra (2012). Enriching Interactions: Incorporating Outcome Data into Static Discrete Games. *Quantitative Marketing and Economics*, 10(1), 1-26.

Ellison, G. and E. Gleaser (1997). Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5), 889-927.

Genesh, D.H. and J.A. Svestka (1979). An exact hierarchical algorithm for determining aggregate statistics from individual choice data. *Management Science*, 10, 939-952.

Geurs, K.T., and B. van Wee (2004). Accessibility evaluation of land-use and transport strategies: Review and research directions. *Journal of Transport Geography*, 12, 127-140.

Ghosh, A. and S.L. McLafferty (1987). Location strategies for retail and service firms. Lexinton, MA: D.C. Heath.

Granot, D., F. Granot, and T. Raviv (2010). On competitive sequential location in a network with a decreasing demand intensity. *European Journal of Operational Research*, 205(2), 301-312.

Hansen, W.G. (1959). How accessibility shapes land use. *Journal of American Institute of Planners*, 25 (1), 73-76.

Haynes, K. E. and A.S. Fotheringham (1984). *Gravity and spatial interaction models*, Beverly Hills, CA: SAGE publications, 9-13.

Holl, A. (2004a). Manufacturing Location and Impacts of Road Transport Infrastructure: Empirical Evidence from Spain. *Regional Science and Urban Economics*, 34(3), 341-363.

Holl, A. (2004b). Transport Infrastructure, Agglomeration Economies, and Firm Birth: Empirical Evidence from Portugal. *Journal of Regional Science*, 44(4), 693-712.

Hotelling, H. (1929). *The Economic Journal*, 39(153), 41-57.

Ingram, D.R. (1971). The Concept of Accessibility: A Search for an Operational Form. *Regional Studies*, 5, 101-107.

Jayet, H. (2001). Econométrie des données spatiales. Une introduction á la pratique. *Cah. Econ. Social. Rural*, 58-59, 105-129.

Jia, P. (2008). What happens when Walmart comes to town: An empirical analysis of the discount retailing industry. *Econometrica*, 76(6), 1263-1316.

Krugman, P. (1991). *Geography and trade*. MIT Press, Cambridge.

Mejia-Dorantes, L., A. Paez, and J.M. Vassallo (2012). Transportation Infrastructure Impacts on Firm Location: The Effect of a New Metro Line in the Suburbs of Madrid. *Journal of Transport Geography*, 22, 236-250.

Melo, P.C., D.J. Graham, and R.B. Noland (2010). Impact of Transport Infrastructure on Firm Formation: Evidence from Portuguese Municipalities. In Transportation Research Record: *Journal of the Transportation Research Board*, 2163, Transportation Research Board of the National Academies, Washington, D.C., 133-143.

Min, H. (1987). A multiobjective retail service location model for fast food restaurants. *Omega*, 15(5), 429-441.

Nash, J. (1951). Non-Cooperative Games. *The Annals of Mathematics*, 54(2), 286-295.

Nguyen, C.Y., K. Sano, T.V. Tran, and T.T. Doan (2012). Firm relocation patterns incorporating spatial interactions. *The Annals of Regional Science*, 50(3), 685-703.

Nwogugu, M. (2006). Site selection in the US retailing industry. *Applied Mathematics and Computation*, 182(2), 15, 1725-1734.

Peng, S.K. and T. Tabuchi (2007). Spatial Competition in Variety and Number of Stores. *Journal of Economics and Management Strategy*, 16(1), 227-250.

Pooler, J. (1994). An extended family of spatial interaction models. *Progress in Human Geography*, 18(1), 17-39.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical of Harold Zurcher. *Econometrica*, 55, 999-1033.

Rust, J. (1994). Estimation of dynamic structural models, problems and prospects: Discrete decision processes. In *Advances in econometrics: Sixth World Congress*, ed. by C. Sims, vol. 2, chap. 4, 119-170. Cambridge: Cambridge University Press.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The Rand Journal of Economics*, 37(3), 619-640.

Stewert, J.Q. (1947). Empirical mathematical rules concerning the distribution and equilibrium of population. *Geography Review*, 37, 461-485.

Thill J.-C. (1997). Multi-outlet firms, competition and market segmentation strategies. *Regional Science and Urban Economics*, 27, 67-86.

Toivanen, O. and M. Waterson (2005). Market Structure and Entry: Where's the Beef? *Round Journal of Economics*, 36, 680-699.

Tversky, A. (1972). Elimination by aspects: a theory of consumer choice. *Psychological Review*, 79, 281-298.

Vickerman, R.W. (1974). Accessibility, attraction, and potential: A review of some concepts and their use in determining mobility. *Environment and Planning A*, 6, 675-691.

Walras, L. (1874). Elements D'Economie Politique Pure: Ou Theorie de La Richesse Sociale.

Williams, H.C.W.L. (1977). On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 9 (3), 285-344.

Zhu, T. and V. Singh (2009). Spatial competition with endogenous location choices: An application to discount retailing. *Quantitative Marketing and Economics*, 7, 1-35.

France - the land of haute cuisine, fine wine and cheese - would be the last place you would expect to find a thriving fast-food market.

- L. Fancourt, B. Lewis, and N. Majka, "Born in the USA, Made in France: How McDonald's Succeeds in the Land of Michelin Stars" (2012)

5

Locational strategies of multi-store firms: A daytime population

Abstract This paper examines competing firms with multiple component units. Most previous discussion on locational decisions has one common feature of making unrealistic assumptions and perceives the industry in terms of independent store. Yet, the conventional single-establishment location theory, which assumes that a firm can choose only one store site, may not apply to situations wherein individual outlets are part of larger organizations under common strategy and intuition. The approach used in this paper is computational. Rather than proposing an innovative solution or striving for an intricate model, we focus on rectifying several misconceptions made in other papers. First, we consider five asymmetric fast-food chains. Further, we insist on including intra- and inter-chain pre-existing stores. Finally, we propose to reconsider the list of variables commonly used to capture the demand level in a market. Firms, such as fast-food chains, search for major sources of demand trying to attract the greatest number of potential customers. Yet, some restrictive assumptions about customer behavior are still present in the existing models. Commonly applied population and income might inadequately represent the demand. Population statistics usually reflect the number of residents but not necessarily the real daytime population. It is particularly important to carefully define market segments and to select potential customer groups by observing their characteristics, their mobility patterns and the purpose of their activities during the day and the week. Proxies used in the models presented in this paper, include flows of people doing shopping, going out to eat in restaurants, or making trips for a leisure purpose; flows of middle- and low-income persons; crowds of potential clients passing through the commercial centers, train stations, airports, or touristic sites.

Our main contributions are to demonstrate that the observed location patterns of analyzed fast-food chains are prone to diverge. Fast-food chains react distinctively to market characteristics and market structure making existing papers rethink their assumption on homogenous firms. Imposing symmetry upon firms' profit functions might be incorrect, even if they seem similar upon first inspection. Contrary to what some researchers found, demand, when carefully accounted for, turns out to be very significant for all chains being studied. In addition, different demand characteristics may be proposed to proxy demand depending on the chain. Our results caution against the simplistic representation of demand. Our method adds behavioral realism to the existing models that are typically based on static, nighttime representations of population.

Keywords: location patterns, multi-store firm, asymmetric players, inter- and intra-chain competition, market structure, flows, daytime population

JEL Classification: D22, C57, C33, C31, L22, L14

5.1 Introduction

The subject of location of competing firms with multiple component units seems to have been largely ignored in the spatial location literature (Peng and Tabuchi, 2007). This gap is inquisitive for the systems which dominate in the market (Karamychev and van Reeven, 2009; Iida and Matsubayashi, 2011; Janssen et al., 2005; Pal and Sarkar, 2002; Peng and Tabuchi, 2007). Chu and Lu (1998) note that most previous discussion on locational decisions has one common feature of making unrealistic and restrictive assumptions and perceives the industry in terms of independent store. The conventional approaches to location selection, *i.e.*, traditional theory and methods, fail (Thill, 1997) by providing only a set of systematic steps for problem-solving without considering strategic interactions between the stores in the market.

Being non-strategic means that a firm ignores other players' decisions (Toivanen and Waterson, 2005). Pal and Sarkar (2002) show that when a firm has several stores, each store's behavior influences the choices of other stores in the market. The correct identification of real competitors provides critical information for the managing team. From the practical point of view, stores draw a bigger attractiveness among potential customers when more competitive facilities locate in their close surrounding. For the customer, it means a wider range of homogeneous products, plenty of opportunities, and a reduced search cost (Min, 1987; Watson, 2009). Yet, a new store will not only compete with pre-existing, previously opened rival firms but may also cannibalize demand from stores that belong to its own chain (Igami and Yang, 2016; Holmes, 2011; Iida and Matsubayashi, 2011; Granot et al., 2010; Schiraldi et al., 2012; Toivanen and Waterson, 2005)[1]. As a consequence, the results for the independent single-store firms and the multi-store ones are prone to diverge. Conceptually, a firm selects a distribution of locations instead of choosing a point location (Chu and Lu, 1998).

Major modern international brands, particularly those in the convenience stores, supermarkets, coffee shops, fast-food restaurants, and fast fashion retailing industries sell their goods through chain stores (Takaki and Matsubayashi, 2013; Peng and Tabuchi, 2007). Examples include Starbucks, McDonald's, Zara, or H&M. These chains are marked by their well-designed networks of stores. Expansion of the chain stores allows players to boost recognition among potential consumers and ensure greater visibility, increase bargaining power, lower costs in distribution, and reduce wastage of resources (Takaki and Matsubayashi, 2013; Janssen et al., 2005; Chang and Hsieh, 2014). All that to maximize their market power or their profits, or to secure sustainable markets.

Typically, stores within a chain reflect a single concept of identity or image (Thill, 1997; Takaki and Matsubayashi, 2013; Toivanen and Waterson, 2005) and thus typically share the same facade, shop format, employ similar design and product offering. The conventional single-store location theory may not apply to situations wherein individual stores are part of larger organizations under common strategy, intuition, and control, where a centralization is applied to reach global goals and consider the interest of a firm as a whole (Thill, 1997).

Chains are usually globally active and well-known by almost every potential client who nowadays is becoming more mobile and consumes repeatedly. The standardization of ap-

[1]For examples of clustering of stores within the same firm, see also Iida and Matsubayashi, 2011; Pal and Sakar (2002); de Palma et al. (1985).

pearance or operations of the chain stores allows the consumers to expect the same type of products, quality, and service within a chain. The created reputation can be relatively easily spread over various markets (Loertscher and Schneider, 2011). Thus, when moving from one location to the other, buying from a chain store decreases the risk of purchasing unknown products and using unfamiliar service.

Interestingly, albeit the products offered by various chains may be technically considered as close substitutes, from the customers' perspective it may not be the case and their psychological attachment to specific brands or chains, their individual taste will rather play a role. The store format and its strategy is supposed to fit well the targeted customer groups identified by their desires, interests, buying requirements, and their expectations (Thill, 1997). Thill (1997) believes that appropriate stores' locations along with relevant brand characteristics create a powerful combination which allows a chain to succeed and to minimize the competitive threat. Very similar firms may have to compete for the same customers (Thill, 1997).

This paper examines location patterns of five competing fast-food firms with multiple component units. The approach used in this paper is computational. Rather than proposing an innovative solution or striving for an intricate model, we focus on rectifying several misconceptions made in other papers. First, we consider a number of asymmetric chains. Second, we insist on including intra- and inter-chain pre-existing stores which are potentially critical, yet have been largely unsung in location analysis literature. Finally, since firms search for large sources of demand trying to attract the greatest number of potential customers, we propose to reconsider the list of variables usually use to capture the demand level in a market. It is particularly important to appropriately define market segments and to select potential customer groups by observing their characteristics, their mobility patterns, and activities' purposes during the day and the week. We emphasize the need to state a clear difference between the daytime and nighttime population. Igami and Yang (2016) claim that typically used population and income might not adequately proxy demand. Population statistics usually reflect the number of residents but not necessarily the real daytime population, which might be much more critical in the fast-food industry. Instead, we test a number of proxies to more appropriately represent demand in a market, namely car flows, pedestrian or all active modes flows (walk, bike and other active modes aggregated), and flows of people using public transport by income level (low, medium, high income classes), age (18-25, 25-34, and plus), or purpose of the trip (restaurant, shopping, leisure, professional meeting) along the day and week;

flows of tourists visiting the most attractive sites of the Paris region; proportion of a zone dedicated to commerce, offices, universities and high schools, hospitals, cultural and leisure equipment, congress and exposition centers. Beside the level of flows of people estimated through the Global Survey of Transport EGT 2010 of the Paris region, we add variables which proxy the crowds of potential clients passing through the commercial centers, train, express underground and subway stations, and airports.

Our method adds behavioral realism to the existing models that are typically based on static, nighttime representations of population. We believe that the incorporation of these elements can secure a more realistic model of locational strategies. Illustrative questions that can be answered using the proposed approach are: What is the nature and degree of competition for each analyzed chain? How strong are cannibalization effects within a chain and business stealing or learning effects among the rivals? How does a firm perceive a rival store located in a direct vicinity and how if it is located farther away? How does a situation of a firm change if there are already rivals established in a particular market versus if a firm faces no local competition? Can it happen that a chain concentrates its investment efforts in a particular geographical area and uses another type of strategy in other places? How to more appropriately proxy the demand in decisions of revenue-oriented sectors? An intensified research effort along the lines of multi-store locational choices is still desirable to bring the answers to many questions of this type.

To open the door to possible answers, we review the literature on the topic, we introduce the reader to our database and we give a brief synopsis of the fast-food market focusing on the Paris region. We show some basic statistics presenting the multiple-store firms. We next compute the Average Nearest Neighbor Distance index to determine if the chains' point patterns are random, disperse, or clustered. We present the reduce-form model and describe the empirical results. Finally, we draw to the concluding section and possible extensions of this paper.

5.2 Literature review

5.2.1 Multiple-store firms

Chu and Lu (1998) note that most previous discussion on location decisions has one common feature of making unrealistic and restrictive assumptions and perceives the industry in terms of independent stores. The conventional approaches to location selection, *i.e.*, traditional theory and methods, fail (Thill, 1997) by providing only a set of systematic steps for problem-solving without considering strategic interactions between the stores within the same firm and stores that belong to different chains. Being non-strategic means that a firm ignores the rivals' decisions (Toivanen and Waterson, 2005) and decisions of other establishments within a firm. In reality, the location decisions of multi-store firms are dependent on each other (Peng and Tabuchi, 2007).

Scant academic research has been conducted to develop empirically tractable models of multiple-store firms' choices, as summarized by Nishida (2015), it is due to the fact that solving and estimating an equilibrium model of store-network choice pose substantial computational challenge. Pal and Sarkar (2002) show that when a firm has several stores, each store's behavior influences the decisions of other stores in the market. The results for the independent single-store firms and the multi-store ones are prone to diverge. Conceptually, a firm selects a distribution of locations instead of choosing a point location (Chu and Lu, 1998). The first work that considers the multi-store firms dates back to 1968 and the work of Teitz who introduced the idea that a firm can open several facilities in the context of Hotelling's linear city model and serves the market from multiple locations.

5.2.2 Rival- and same-chain stores

Existing location-store models typically do not consider pre-existing competing stores[2], which may indicate the level of competitive threat in a particular area, nor the potential reactions of these direct-competitor stores, nor the strategic interactions between stores within the same chain are clearly analyzed. The entry of new stores seem to harm the profitability of pre-existing outlets. Thus, there is a necessity to incorporate strategic interactions between stores within the same firm which may capture the cannibalization effects (see, *e.g.*, Igami and

[2]Possible business categories mentioned by Chang et al. (2015) are: same chain stores; rival stores; famous stores, such as Apple or Nike, which can provide a crowd flow and certain customers because of their customer attraction; complementary stores which sell complementary goods or services; mutually exclusive stores which might resent or block potential customers from frequenting a store. Schiraldi et al. (2012) differentiate stores depending on their size/format into small and big. In addition, local business activity can be proxied by using the number of entertainment establishments (*e.g.*, theaters, swimming pools), speciality establishments (*e.g.*, pawn shops, pet shops), amenities and services (car rental, hair salons, laundry, or parking), and malls as was proposed in Yang (2016).

Yang, 2016; Holmes, 2011; Iida and Matsubayashi, 2011; Granot et al., 2010; Aquirregabiria and Suzuki, 2016) and among stores that belong to different chains and the accompanying business-stealing effects between them (see, *e.g.*, Watson, 2009).

Having more rivals in a market may decrease the probability of sale (Watson, 2009). The profitability is a function of player *i*'s entire network and the competitors' networks (Nishida, 2015). Firm *i*'s total profit is the sum of the market-level profits across all the markets (Nishida, 2015). Igami and Yang (2016) find that the presence of both same-chain stores and rivals lessens an outlet's profitability. Based on the interviews with store-development officers of several fast-food chains in Canada, Igami and Yang (2016) discover that cannibalization is one of their most prominent concerns. A higher number of same-chain stores tend to dampen the chance of further entry, whereas the influence of the rival-chain shops seems to be rather subtle and highly nonmonotonic (Igami and Yang, 2016).

Firms face business stealing effects but also incentives to collocate so that the relative magnitude of these two effects varies substantially across store types (Aquirregabiria and Suzuki, 2016). Yang (2016) who studies Canada's hamburger fast-food industry of five major substitutable with one another chains, A&W, Burger King, Harvey's, McDonald's, and Wendy's, explains that clustering (when stores systematically choose the same locations) can be provoked by several drivers, such as learning, demand externalities[3], or unobserved heterogeneity[4]. Toivanen and Waterson (2005) emphasize that while traditional industrial organization theories cannot explain the positive effect of rival presence on own entry, learning models can. They are contrary to what would be observed from an entry prevention strategy. They indicate positive spillovers to both firms from the presence of the rival in a given market. These

[3]An expanding cluster of rival chains may boost demand of other active firms through the increased customer traffic or higher goodwill towards similar products (Yang, 2016). Yet, Yang (2016) demonstrates that demand externalities do not appear to be significant in retail clustering.

[4]Retailers can make decisions based on information unobservable to researchers which therefore will not be included in the models (unobserved heterogeneity) and may be leading to biased estimates and subsequent counterfactual analysis (Yang, 2016; Igami and Yang, 2016; Orhun, 2013). Unobserved heterogeneity, if ignored, may also attenue competition effects (Arcidiacono and Miller, 2011; Igami and Yang, 2016). Igami and Yang (2016) discover that their data on hamburger fast-food chains in Canada feature puzzling patterns of firms favoring the presence of more rivals caused unobservably by more lucrative markets. Toivanen and Waterson (2005) state that if unobserved market specific factors are positively correlated between firms, an estimated positive effect of rival presence on the probability of entry could be counterfeit. We suggest readers particularly interested in issues of unobserved heterogeneity, to acquaint themselves with the work of Igami and Yang (2016) whose purpose was to explicitly assess and address the endogeneity problem caused by unobserved market heterogeneity in the context of a dynamic entry game among fast-food multi-store firms.

spillovers are best explained as the product of learning. Toivanen and Waterson (2005) show that the rival presence increases the probability of entry due to the firm learning, yet, profits are decreasing in the number of rival stores and are increasing in the number of own outlets. The Toivanen and Waterson' (2005) results further suggest that learning effects are strong enough to dominate any negative effects that competition between firms may have on entry decisions.

In addition, learning turns out to be the most prominent and plays the largest role in retail clustering according to Yang (2016). Observation of other stores in the market could send positive or negative signals to update beliefs about the profitability of the market. More attractive markets could be considered the ones which host well prospering firms in comparison with the markets where others have failed. In some past studies, several authors (*e.g.*, Ackerberg, 2001) postulate that learning should matter only to firms which lack information or are inexperienced. Yang (2016) claims that learning helps resolve the uncertainty firms face in the market by observing the past decisions of their informed rivals. He finds that Burger King and McDonald's seem to take advantage of information externalities via learning from each other. Interestingly, two other analyzed chains, namely a Canadian Harvey's and A&W which entered the Canadian market long before Burger King and McDonald's seem to already have a necessary knowledge about market profitability so that they exhibit the smallest difference between being uninformed and informed. Toivanen and Waterson (2005) also suggest that a substantial amount of clustering of fast-food restaurants (McDonald's and Burger King) in the UK can be explained by informational spillovers from the pre-existing firms to the potential entrants.

Aquirregabiria and Suzuki (2016) highlight that cannibalization and the economies of scope between same-chain stores are important factors in location choices of multi-store firms. Economies of scope may lead to the reduction of some operating costs, such as advertising, inventory, personnel, or distribution costs, which will be shared between the stores of the same chain in case of the chain expansion into new markets.

5.2.3 Assumptions made in other papers

The analysis of multi-store competition has started with the trailblazing work of Teitz (1968) who introduces the idea that a firm can open several facilities in the context of Hotelling's linear city model and serves the market from multiple locations. Since then infrequent papers tried to treat complex decisions of multi-store firms proposing two- or three-stage models wherein firms choose, *e.g.*, stores' locations and their qualities or value positions (Thill, 1997); the number and locations of stores along the geographic line, and optimal mill prices charged by these outlets (Chu and Lu, 1998); locations of stores and quantities in Hotelling's linear city using the Cournot competition (Pal and Sarkar, 2002); simultaneously or sequentially the number and locations of their stores, and prices accounting for consumers preferences, consumers identity, the identity of two analyzed firms, and the identity of the stores within these chains (Janssen et al., 2005); the number, locations of their facilities, and then the level of prices at every store where heterogeneous products are offered in the context of the Salop's circular product space (Karamychev and van Reeven, 2009); the number of their stores and their locations, and the number of varieties by two retail firms showing that the location-then-variety model can better describe the retail sector in the real world, such as market segmentation, interlacing, sandwich, or enclosure. The work of Iida and Matsubayashi (2011) examines a number of store openings of two firms offering identical products and their locations on a linear street under financial constraints. An exogenously given brand power ascribed to each store and the distance necessary to reach its location site allows consumers to differentiate between stores. Takaki and Matsubayashi (2013) analyze a sequential multi-store Stackelberg competition between two symmetric firms that choose a number and locations of their outlets given current financial constraints, a fixed opening cost, and uniformly distributed consumers. Neven's (1987) model allows for a sequential entry of up to five firms in the retail industry, claiming that the problem solved for five firms covers a sufficient variety of strategies and that a more extensive simulation would not yield many more insights, yet it also assumes the uniform distribution of consumers' tastes. In the paper by Peng and Tabuchi' (2007), consumers are assumed to be uniformly distributed. To deal with the problem of the uniform distribution of the clients, Granot et al. (2010) propose instead a sequential locations in Hotelling's linear city and a graph city at a fixed cost assuming the demand intensity to be strictly decreasing in distance. Nishida (2015) incorporates some form of spatial competition, showing that a store's revenue is influenced not only by other stores in the same location but also by those in adjacent locations. Yet, he builds his model in a static framework where two firms compete. Seim's (2006) model features several distance bands, yet it is static and assumes the independence of locations, and the symmetry across the analyzed chains.

What emerges from the cited examples is that even the most recent papers still keep on making some unrealistic assumptions, such as: a linear market, no more than two firms competing in a market[5], zero level costs, a constant density and elasticity of demand, uniformly distributed consumers, a static approach, or an independence between the markets. Several papers examine consequences of relaxing some of these assumptions individually. In addition, most of the papers considering the multi-store firm location are theoretical.

5.2.4 Distance

Teitz (1968) claims that distance influences a consumer's selection of a firm and branch assuming that this distance is measured from the consumer to the nearest branch of either firm and that the cost of travel is linear with distance and borne by the consumer who behaves as a distance minimizer. Yet, Nwogugu (2006) finds the issue with distance measure particularly disturbing. The author claims that existing store-location models do not incorporate the distance element appropriately, erroneously assuming that all residents of each community travel the full length of the distance between their community and the store location each time they visit a store. However, in reality most potential buyers do not work in their immediate communities and tend to stop over a store on their way to other destinations and so will prefer locations that are on the route to destinations that they regularly visit. Most people drop in on the store before going to work, during lunch, after work, or on weekends. Thus, there is a need to use data about potential customers activities performed in a given area and to well understand the process of chaining behavior. Fransen et al. (2015) demonstrate that the introduction of trip chaining behavior into the accessibility analysis has a significant impact on the resulting accessibility pattern. Considering more complex travel behavior by accounting for daily mobility patterns results in significantly distinct spatial patterns of accessibility. Fransen et al.' (2015) method adds a behavioral realism to the existing metrics that are typically based on static, nighttime representations of population. For these reasons, we keep in mind the fact that visiting a fast-food restaurant can often be combined with other activities.

In addition, the limitations of the "isolated small towns" approach which assumes no coordinated decisions across markets, have motivated the development of more appropriate empirical models. Determinants of the multiple-store firms' network choice make location decisions across markets interdependent, unlike a standard location model (Nishida, 2015; Jia,

[5] See, *e.g.*, Granot et al. (2010) who propose characterization that can deliver robust results for an arbitrarily large number of players.

2008). The work of Seim (2006) is the first study which endogenizes store locations and introduces the concept of spatial competition in an entry game framework. Seim (2006) shows how the market power and profits change depending on stores' locations of one firm in relation to competitors' locations. Her idea was to capture stronger effects of closer stores and a decreasing profit in case of decreasing distances from competing stores. A particular surface can be divided into dozens, hundreds, or thousands of contiguous blocks or locations, yet, these locations should not be isolated, allowing the model, for example, for competition effects between the stores at different locations. Nishida (2015) and Toivanen and Waterson (2005) incorporate some form of spatial competition, showing that a store's revenue is influenced not only by other stores in the same location but also by those in adjacent locations. We notice that in the papers by Igami and Yang (2016), Ellickson et al. (2013), Beresteanu et al. (2010), the authors rule out the interdependence between the markets.

In our work, we insist on including interdependent markets (Holmes, 2011). Within-market and across-markets effects should be analyzed (Nishida, 2015). The neighboring characteristics of each site-location and surrounding businesses will be included in our model (see *e.g.*, Seim, 2006; Chang et al., 2015; Schiraldi et al., 2013). For any given location *l*, Schiraldi et al. (2013) define a set of neighboring locations within some distance, whose payoffs affect firm's profits at this site *l*. Thus, variables the total number of same-chain and rival stores could all affect the profit of a store at *l*. This way, for example, Schiraldi et al. (2013) allow business stealing and cannibalization effects to operate not only within the boundary of a particular location, but also across locations which may be the source of the spatial competition. Igami and Yang (2016) notice the future need to include the shops' distances within geographical market stating that closer shops seem to compete more fiercely. Thus, the Seim's (2006) model which tests several distance bands should be further extended to enable exploration of spatial patterns of, *e.g.*, the fast-food competition.

Also Aquirregabiria and Suzuki (2016) highlight the importance of distance from a store to potential customers, wholesales, and competitors which appear to have a substantial effects on demand and costs and consequently on profits or consumer welfare. Firms are expected to carefully choose locations to ease the access to the highest number of spatially dispersed potential customers.

Interestingly, Igami and Yang (2016) observe that fast-food stores seem to compete within relatively small markets concentrating their efforts on a micro-level location. Their results indicate a highly localized nature of competition among fast-food stores. The authors state that a distance criterion equal or greater than a mile (1,61 kilometer) would appear to be useless for an empirical analysis of competition among fast-food stores. According to calculations of Thomadsen (2005) for the fast-food market (McDonald's and Burger King) in Santa Clara County, only outlets within 0.5 mile (800 meters) will compete as close substitutes, even in car-obsessed California. Toivanen and Waterson (2005) emphasize that people in the UK, for example, are not willing to travel far to satisfy their fast-food needs, seldomly traveling outside their districts in search of fast-food outlets. In the case of the convenience-store chains, Nishida (2015) notices that convenience store demand is more localized in Japan than other types of service industries, such as supermarkets or gas stations, with 70% of customers' visits on foot or bicycle and 30% by car. Nishida's (2015) findings indicate that a trade area for a typical convenience store has a radius of about 500-700 metres and the consumer's average travel time to a convenience store is about 10 to 20 minutes by walking. The author states that the convenience-store industry invests heavily in sophisticated distribution networks because stores usually do not have much inventory space and because 70% of the sales are perishable, so that a store needs to preserve the freshness of food, such as sandwiches and lunchboxes, which delivery tracks replace two or three times a day. On the other hand, Walmart and other supercenters seem to compete in the efficiency of purchasing and distribution logistics on a relatively larger geographical scale. Schiraldi et al. (2013) indicate that the radius of 10 kilometers needs to be taken to define the neighboring locations of a particular site l in the UK supermarket industry. Moreover, Holmes (2011) shows a systematic geographical pattern of Walmart's entry that radiates from the headquarters.

5.2.5 Demand

As highlighted by Min (1987), the retail and service sector location analyses, such as investigation of the fast-food stores' locational decisions, are likely to be highly sensitive to the demand and revenue-generating factors, whereas cost factors should play a critical role rather in the location choices of, *e.g.*, warehouse locational strategies. These remarks line up with the work of Fransen et al. (2015) and Wang (2012). Fransen et al. (2015) notice that the demand for services needs to be differentiated. Yet, some unrealistic assumptions about customer behavior are still present in existing models. Interestingly, Fransen et al. (2015) suggest that not the entire working population should be treated as potential customers for a simple reason, that not every person commuting to work make use of a particular facility (in our application, service or store). This can lead to an overestimation of the facility's (service's or store's) demand. Integrating non-spatial factors such as demographic and socio-economic variables that impact the access to a particular facility/store is an important challenge (Wang, 2012). The accessibility for specific population groups, such as those with a higher income or without a privately owned vehicle could be examined. As well stated by Jerry W. Thomas (2007), the CEO and the founder of Decision Analyst, "Segmentation studies tend to be large and complicated, so it is easy for errors to be made. Thus, one of the most common mistakes is targeting people instead of targeting money. A market segment may represent a large percentage of population, but a small part of the market. There is always a need to look at the money potential of market segments, not just the number of people in the segments". Thus, one of the results of Chu and Lu (1998) seems not to tell the full story when indicating that more populated areas will encourage the player to establish more stores.

Demographic and socio-economic characteristics of potential customers who reside in a particular area and its surrounding along with their traveling and shopping boundaries should be determined and implemented into the model (Min, 1987). The demand faced by a particular outlet according to Karamychev and van Reeven (2009) and Thomadsen (2007) consists of only consumers who fix on buying at that store, of consumers who patronize a particular outlet.

Communities form and disperse at various times of the day within a given trade area. People travel to other communities for work, leisure, and for other purposes. For any given trade area, there are several time-dependent, transport-dependent communities: communities of people before they leave to work, communities during the peak-hours period when people travel to work[6], communities of people at work before the lunch time, communities of people having their lunch, communities when children and teens leave schools for their homes, communities during the evening peak-hours[7] when people travel from work to their homes and for recreation, and communities after the evening peak-hours period when people are resting at home. These activities and times change during the weekend.

Gira Conseil company specialized in analyzing the Parisian consumption market explains

[6] That is from 6.50 a.m. to 9.10 a.m. for private vehicles and from 8 a.m. to 9 a.m. for public transit for the Paris region (DRIEA Île-de-France, 2008).

[7] That is from 5 p.m. to 7 p.m. for private vehicles and from 5 p.m. to 6 p.m. for public transit for the Paris region (DRIEA Île-de-France, 2008).

how important it is for a firm to understand and well define the customer catchment area, its clients' stratums, consumption moments, such as an "empty fridge". In addition, Gira Conseil evaluates that a potential customer is willing to walk 10-12 minutes or drive a car for 5-7 minutes to reach a particular restaurant.

Thomadsen (2007) shows that firms will choose to locate near large sources of demand. Thus, it is particularly critical to correctly capture the demand level, to define market segments and to select potential customer groups by observing their characteristics, such as age or occupation (Chang et al., 2015), their mobility patterns, their chaining behavior, and activities' purposes during the day and during the week. Influx of cars and pedestrians will be also analyzed in our model (see Chang et al., 2015). Consumers are heterogeneous considering their preferences over the brands proposed by various sellers. Firms tend to set more locations where consumers are more prone to buy from that firm (Janssen et al., 2005). Janssen et al. (2005) claim that firms open more stores where more consumers agglomerate. Yet, Igami and Yang (2016) show that typically used population and income might not adequately capture the demand. Population statistics usually reflect the number of residents but not necessarily the real daytime population, which might be much more critical in the fast-food industry. Therefore, other alternative measures should be tested. Igami and Yang (2016) propose they use the size of office or commercial floor space, among others. Yet, these measures were unfortunately not recorded in the publicly available statistics and were not available to Igami and Yang. Another solution that they suggest is to test instead the number of traffic lights. Following the remark of Nishida (2015), a clear distinction between a daytime versus nighttime population present in a particular area is to be made.

In addition, people tend to have psychological attachments to specific brands or chains and can drive or walk a couple of additional minutes to reach their preferred store (Kuo et al., 2002). According to Karamychev and van Reeven (2009) consumer's utility depends on both real transport costs and on their preferences or tastes. Another type of segmentation distinguished by Thomas (2007), which can also be significant in defining a potential customer group, is a psychographic or lifestyle segmentation. It is based upon multivariate analyses of consumers' attitudes, values, behaviors, emotions, perceptions, beliefs, and their interests[8].

[8] A definition of a market segmentation given by Thomas (2007): "The purpose of segmentation is the concentration of marketing energy and force on the subdivision or the market segment to gain a competitive advantage within that segment. It is analogous to the military principle of "concentration of force" to overwhelm an

An identification of the store market area boundaries is critical in determining the aggregate demand and potential market share (Min, 1987).

5.3 Data

Min (1987) proposes to divide factors possibly considered in locational strategies analyses into three broader categories: (1) locations of the nearest pre-existing competitor stores, (2) an accessibility to public amenities. (3) an average profitability of a particular site. We add to this list some other factors that will be described in the following subsections.

5.3.1 Stores' characteristics and competitors' locations

The consolidation of businesses into a few large multiple-store corporations has been a documented evolution of trade in Western countries after the World War II (Hollander and Omura, 1989). In the fast-food industry, there are several brands, such as McDonald's and Quick which according to the French daily newspaper "Le Figaro" seem to dominate and set the rules in the French market[9]. For the purpose of the current project, we construct a database for the following chains: McDonald's, KFC, Quick, Paul, Pomme de Pain. We have chosen to investigate the fast-food industry considered as an archetypical chain business for which the provision of homogenous goods and services is one of the main purposes (Igami and Yang, 2016). The scope of the product differentiation is limited among stores. The production of identical products have been so effectual that the Big Mac index, based on the premise of purchasing-power parity, is being used across different countries to analyze foreign-exchange rates[10]. Beside the three best performing fast-food chains in the French market, namely McDonald's, Quick, and KFC, we examine Paul and Pomme de Pain. Our motivation was based on the fact that France is the country where people consume more traditional sandwiches than burgers. A 2009 study carried by the French restaurant industry consulting firm Gira Conseil revealed that the French population consumes nine times more traditional sandwiches

enemy. Market segmentation is the conceptual tool to help achieve this focus". Source: "Market Segmentation", Decision Analyst.

[9]Source: "Le Figaro", Geraldine Russell, September 28, 2015, "Burger King-Quick, l'alliance qui rebat les cartes de la restauration rapide".

[10]See "The Big Mac index: Bunfight", "The Economist", February 2, 2013.

than hamburgers and that more than 70% of all sandwiches consumed in France are made on baguettes[11].

For the active stores, information is based on the SIRENE/INSEE database and includes the year of the store's creation, its size class, its exact address, and the address of its headquarter. For the previously closed stores, not active any more in the market, the exact date of the store's closure is also available. The BAN database - a collaborative national database of geo-referenced addresses in France - created in the cooperation of the French National Institute of Geographic and Forest Information (IGN), the mail service of France - La Poste, and the Public Finances General Directorate (DGFiP) was used to geo-code all the addresses, to obtain the longitude-latitude coordinates of each store. Based on the X-Y coordinates, we further compute the Euclidean distance between the stores and between each store and its headquarter[12]. Nowadays, also other commercial websites, such as Google, Yahoo!, Mapquest, Bing, or Rand McNally offer precise driving directions between nearly all locations in the developed world (Boscoe et al., 2012)[13].

We have further verified the consistency of the SIRENE/INSEE database comparing it with the official websites of each fast-food chain[14], so that some small number of missing addresses were added to our database. In this way a unique database has been created, which has been never used in any other application.

5.3.2 Area characteristics

The considered area is the Paris region. When choosing a location for a new fast-food restaurant, one needs to identify feasible sites with respect to various factors. Following the remarks of Nwogugu (2006) who identifies several major mistakes of existing store-location

[11] See, *e.g.*, "Born in the USA, Made in France: How McDonald's Succeeds in the Land of Michelin Stars", Jan 03, 2012, http://knowledge.wharton.upenn.edu/article/born-in-the-usa-made-in-france-how-mcdonalds-succeeds-in-the-land-of-michelin-stars/.

[12] See Toivanen and Waterson (2005) for the use of distance from the market to the headquarters of two analyzed firms, namely McDonald's and Burger King with the application to the UK market.

[13] See the following website: http://adresse.data.gouv.fr/csv/;
http://batchgeocodeur.mapjmz.com/; http://www.doogal.co.uk/DrivingDistances.php;
http://mondeca.com/index.php/en/any-place-en.

[14] See https://www.restaurants.mcdonalds.fr; http://www.kfc.fr/#/restaurants_storelocator;
https://www.quick.fr/fr/localisation; http://www.paul.fr/fr/nos-boutiques;
http://pommedepain.fr/accueil/nos-adresses/paris-2.

models, we focus first on an appropriate set of explanatory variables. Primo, we look for the measures that will correctly proxy consumers' demand including their demographics (income, age), customers' spatial behavioral patterns characterized by their buying patterns, their activities, the purpose of their trips and the time of the day when these trips occur, and traffic counts of cars and pedestrians.

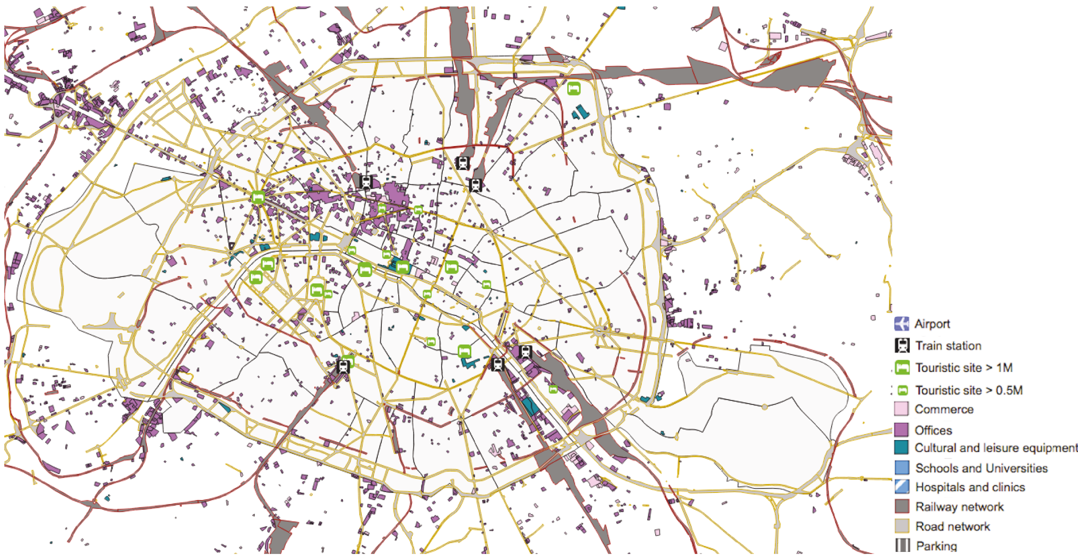


Figure 5.1: Example of data representation: Zones' characteristics. Various data sources; The most touristic sites in the Paris region with over 1 million and 0.5 million tourists each year in 2012; Locations of train stations. Source: Ministry of Crafts, Commerce and Tourism, "Memento du tourisme 2013", General Directorate for Competitiveness, Industry and Services

Cliquet (2006) agrees on using more complete databases concerning attitudes and real behaviors of consumers. Cliquet (2006) makes a stand that demand can be measured in terms of revenue, number of households, budgetary structures of spending, and lifestyles. Schiraldi et al. (2013) exploit a rich data of consumer choices in the UK supermarket industry and consumers' locations[15]. Yet, the population is proxied by the total number of households at l and in its neighborhood, which might not be accurate in the case of the fast-food industry. There should be a clear distinction made between the daytime and nighttime population present at a particular site (Nishida, 2015). Chang et al. (2015) propose to consider instead the stream

[15]The consumer survey contains information on supermarkets or at least a firm visited by each consumer, consumer's location, his social class, and the household size.

of people, crowd flow rate in front of a store, target customer number within a particular area[16], customers preferences, their lifestyles, their demographics, the shopping habits of the crowd, customer purchasing characteristics, and their purchasing frequency. Moreover, their commute tools, traffic patterns in the area should be rather considered.

Other explanatory factors that will be included in our analysis are: land-use characteristics, the transport convenience, the proximity to public transport system, roads and highways traits, and parking spots (see works of Chatman et al., 2016; Holl, 2004a, b; Melo et al., 2010; and Mejia-Dorantes et al., 2012; Chang et al., 2015), proximity to schools, universities, offices, shops, etc. (Min, 1987). All this to assure that location is convenient and accessible for potential customers (Chang and Hsieh, 2014).

5.3.3 Data sources of zones' characteristics

We include here a short description of data used in the present project along with their sources. Our data set covers variables that represent measures of demand, availability of floor space, and accessibility measures.

Data about characteristics of local population were gathered through the General Census of INSEE. Information on income imputation was prepared by DGFiP and available through INSEE. Information on the total number of trips were available through the Global Survey of Transport EGT 2010 of the Paris region. The movements and activities' purposes, such as shopping, restaurant, leisure or social activities of groups of potential customers characterized by various socio-economics features were being observed during the day from Monday to Sunday. The origin and destination points of each trip, the day of the week and the exact time when the activity was performed, and the mode of transport used have been registered. Data about the potential customers' activities performed in a given area including their mobility patterns can be extracted from the EGT database, so as the car and pedestrian flows.

To account for the proximity to various amenities, for example, schools, hospitals, parks and parking places we considered two sources of numerical land use databases, these are MOS (IAU-IDF). Information on the transport infrastructure, *e.g.*, access to the train, subway, tramway stations, buss stops, airports, and distance to highways, was obtained through the Regional and Interdepartmental Directorate of Public Works and Planning of IGN.

[16] According to Chang and Hsieh (2014), generation of consumption power is caused by crowd flow, population, households, and workforce.

5.4 Market description

We epitomize here the fast-food restaurants market in France and the Paris region, in particular, and its most important players. The commercial restaurant food service market is composed of traditional and fast-food restaurants[17]. The amount of money spent on the commercial restaurant food services in France in the year 2012 was as high as 35.9 milliard euros (Roux, 2014)[18]. A very dynamic evolution of the fast-food market has been observed since the year 2000, with a slight slow down in 2012. According to the analysis performed by the French Ministry of Economy, Finance and Industry[19], the overall turnover value of the fast-food restaurants market was for the first time higher than the value of the traditional restaurant market in 2010, while the discrepancy between these two markets were significantly large back to the year 1995 (see Fig. 5.2). Along with the fast-food market turnover growth, a significant number of new fast-food store openings was registered.

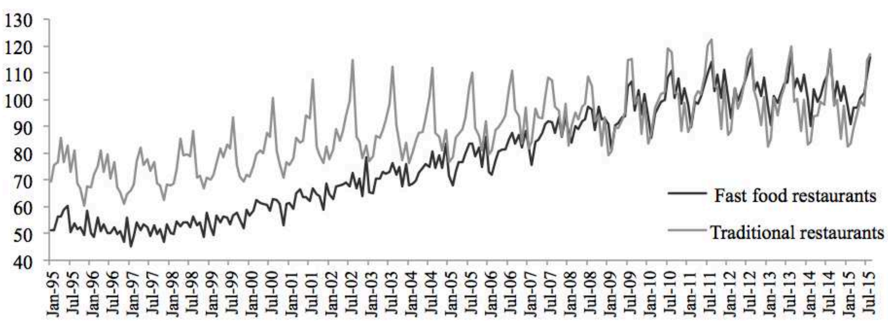


Figure 5.2: Monthly turnover index value in million euros for the fast-food and traditional restaurants over time. Situation for the metropolitan France from January 1995 until August 2015. Reference period: 2010. Source: INSEE, DGFiP

McDonald's has unquestionably conquered France (Wile, 2014)[20] and France became to this chain the most profitable country outside the US. The McDonald's company has stayed

[17] Toivanen and Waterson (2005) emphasize also the dissimilarities between the two everyday markets services, namely the local counter-service market and table service market which are significantly different in view of time element, for example, and the transit market. In the examination of the transit market, which assumes that outlets operate from airports or motorways, the passenger flows are most likely to be involved.

[18] Source: DGCCRF ECO.

[19] Source: DGCCRF ECO.

[20] Source: Rob Wile, 2014, Business Insider, "The True Story of How McDonald's Conquered France".

the leader in the fast-food industry for decades, dominating fast-food culture on the world stage (Carlsen, 2010)[21]. The company hires now 3 000 employees per year, has more than 69 000 workers in France (Wile, 2014), and more than 1 200 locations across France, including locations such as Louvre, Sorbonne, or Champs-Elysées. In 2013, McDonald's announced that it was going to invest 200 million euros in the further expansion (Wile, 2014). However, the beginning was rather tough for the McDonald's chain in France. The company installed its first French store in Créteil in the suburbs of Paris, back in 1972, yet, at this time McDonald's did not experience success as the French population was opposed to the Americanization of their country (Carlsen, 2010). McDonald's was obliged to close its franchise. The company official history dates to 1979 when it was trying again slowly to synthesize its American DNA with the French culture.

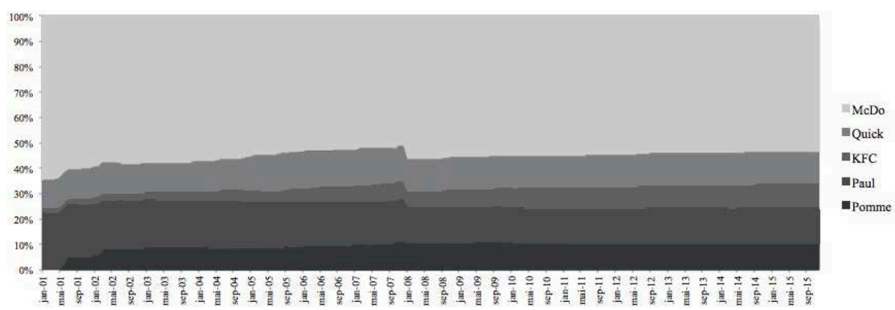


Figure 5.3: How the best performing fast-food chains in the Paris region share the market between each other over time. Accumulated monthly number of newly opened stores in the Paris region between 2001 and 2015, including the stock of stores registered before January 2001. Source: Calculations based on the SIRENE/INSEE database

McDonald's company gained 80% of the ground in France, yet, it is not immune to other competitors which do not stay unresponsive according to information provided by the McDonald's website considering its commercial strategies[22]. Many companies try to integrate into the market constantly challenging McDonald's and provoking its managing board to rethink its strategy and renew its ideas. The biggest McDonald's competitor is the Belgian chain, Quick who entered the French market in the year 1980. It proposes similar style and

[21]Source: https://uramericansinparis.wordpress.com, Lindsey Carlsen (2010), "Americans in Paris, McDonald's in France".

[22]Source: http://mcdofrance.e-monsite.com.

alike products, yet remains far behind. The third player is Kentucky Fried Chicken (KFC), an American restaurant specializing in products based on the chicken[23]. Interestingly, a burger operator Quick which is successful in France and other countries has failed to set up in the UK market (Toivanen and Waterson, 2005). According to "Le Figaro", the sales of McDonald's reached 4.097 billion euros[24] and were about 826 million euros for Quick. KFC was not willing to provide this kind of information[25].

For the purpose of the current project, we construct a database for the following asymmetric chains: McDonald's, Quick, KFC, Paul, and Pomme de Pain. Beside the three best performing fast-food chains, we also add two more players, namely, Paul[26] and Pomme de Pain - the pioneers of sandwiches "à la française". Pomme de Pain proposes to French people to discover "the good taste of sandwich"[27]. It opened its first bakery in 1980 at rue de Rivoli in Paris. Paul was founded as far as in 1889, and it opened its first tea salon in 1987, it expands internationally and is nowadays present in 30 countries around the world, opening its stores at the train stations and airports. Paul is known as the most famous bakery in the world. Our motivation to choose Paul and Pomme de Pain along with McDonald's, Quick, and KFC was the fact, that France is the country where people consume more traditional sandwiches than burgers. For this reason, we want to examine how also these two players compete with each other and other fast-food best performing players. Moreover, these firms may attract different customer types.

We investigate further the locational tendencies across the Paris region. We observe where across the region each fast-food chain concentrates its efforts. We examine the location choices of each chain year by year. We accumulate over time the number of newly-formed stores opened in the Paris city, Paris region closer and outer suburbs and show the locational tendencies of store openings between 2001 and the year 2015 (see Fig. 5.5). We notice that recently,

---

[23]Notes: According to the analysis of the Xerfi Groupe which provides the market and firms' analyses, major groups that dominate the fast-food market in France include: McDonald's, Quick, Duff, Subway, Domino's Pizza, Yum! Brands (KFC, Pizza Hut), Speed Rapid Pizza.

[24]Source: Le Figaro, Geraldine Russell, September 28, 2015, "Burger King-Quick, l'alliance qui rebat les cartes de la restauration rapide". According to Wile (2014), the McDonald's sales reached 4.46 billion euros in 2013.

[25]McDonald's sales share is estimated at 40% of the burger fast-food market in the UK in 1994. The UK market is dominated by three players, namely McDonald's, Burger King, and Wimpy. These three chains were estimated to have 45% of fast-food outlets in the UK (Toivanen and Waterson, 2005).

[26]Source: http://www.paul125ans.fr/saga125ans/index.php.

[27]Source: http://pommedepain.fr/en/home/the-brand/our-history/.

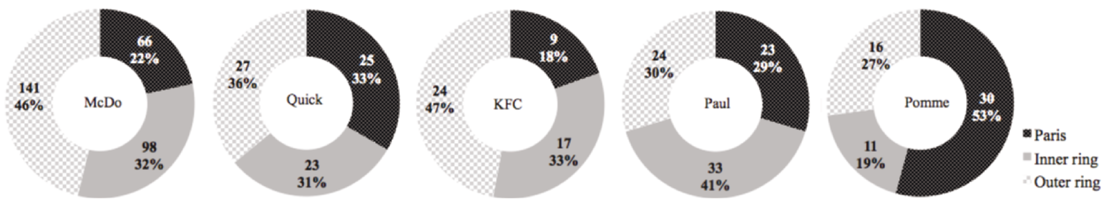


Figure 5.4: Attractiveness of the Paris region for the best performing chains: The global situation which considers the total accumulated number of all pre-existing stores of these chains in 2015. Source: Calculations based on the SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

KFC and McDonald's chains opt for a similar strategy, concentrating their efforts more and more in the outer parts of the Paris region and choosing to invest relatively the smallest efforts in new store openings in Paris itself. The proportion of new stores opened in the outer suburbs come close to the level of 50% in comparison with the number of stores opened in Paris (18% and 20% for KFC and McDonald's, respectively) and the closer suburbs of the Paris region (34% and 33%). Thus, every second new store is being opened nowadays in the outer suburbs and only every fifth in Paris. Further, we discern that Quick demonstrates the tendency to share its efforts relatively evenly across the region with a slight privilege for the outer suburbs and that Paul opts for the locations in the closer suburbs slightly leaving behind Paris and the outer suburbs. Last, quite opposite to the strategies of McDonald's and KFC, Pomme de Pain still chooses to concentrate its actions in Paris, leaving the suburbs much more unserved. A static picture of the situation in the year 2015 that explains which part of the Paris region seems to be the most attractive for each analyzed chain can be read from Fig. 5.4.

5.5 Nearest neighbor distance

We investigate first the point pattern between same-chain and inter-chain stores. For the purpose of this paper, store addresses have been geo-coded to obtain the longitude-latitude coordinates of each establishment. Based on the X-Y coordinates, we further compute the Euclidean distances between all the possible pairs of stores that belong to selected five fast-food chains[28]. Given the exact location of 571 stores, we build a matrix of size 325 470 (without the zeros on the diagonal). Other alternatives are, *e.g.*, the walking and driving distances. We

[28] www.doogal.co.uk website has been used to geo-code the data and compute the distances.

next compute the Average Nearest Neighbor Distance index, the ANN statistic. The Average Nearest Neighbor Distance index uses the distance between each store (each point) and its nearest neighbor's location, *i.e.*, its closest neighboring point in a layer (Ebdon, 1985). The Nearest Neighbor Analysis determines if the point pattern is random, disperse, or clustered.

The ANN tool averages the nearest neighbor distances. If the average distance, computed this way, is less than the average distance that would be obtained in case of a hypothetical random distribution, the distribution of the stores is considered clustered. In case of average distance being greater than a hypothetical random distribution, the stores are considered dispersed. The index is expressed as the ratio of the observed distance divided by the expected distance:

$$\mathsf{ANN} = \frac{\bar{D}_O}{\bar{D}_E}, \tag{5.1}$$

where $\bar{D}_O$ is the observed mean distance between each studied point and its closest neighbor:

$$\bar{D}_O = \frac{\sum_{i=1}^{N} d_i}{N}, \tag{5.2}$$

and $\bar{D}_E$ is the expected mean distance between the points given the random pattern:

$$\bar{D}_E = \frac{0.5}{\sqrt{\frac{N}{A}}}, \tag{5.3}$$

where d_i equals the distance between store i and its nearest neighboring outlet, N is the number of studied points and A is the studied area.

Expected distance is based on a hypothetical random distribution with the exactly same number of stores covering a particular area. This ratio permits to valuate competition of stores locating over some selected terrain. This tool is most appropriate when the study area is fixed.

We analyze in detail three cases, concretely, the area of Paris, Paris along its closer suburbs, and the Paris region. The surface of these areas equals to 105 km^2, 762 km^2, and 12 012 km^2, respectively. We investigate whether stores which belong to the same chain cluster with each other and what are the pairwise tendencies of stores which belong to different chains. Fig. 5.9, 5.6, and 5.8 visualize the location choices of stores that belong to each of the five chains being

Table 5.1: The pairwise ANN index. Case #1: Paris

Case #1: Paris					
	McDo - McDo	McDo - Quick	McDo - KFC	McDo - Paul	McDo - Pomme
SumDist	42.75	61.90	85.16	78.97	92.14
N	66	91	76	90	98
ANN	1.03	0.47	0.71	0.60	0.67
		Quick - Quick	Quick - KFC	Quick - Paul	Quick - Pomme
SumDist		23.99	38.59	38.18	44.01
N		25	35	49	57
ANN		0.94	0.47	0.40	0.42
			KFC - KFC	KFC - Paul	KFC - Pomme
SumDist			16.02	45.77	78.61
N			10	34	53
ANN			0.99	0.57	0.78
				Paul - Paul	Paul - Pomme
SumDist				17.53	30.90
N				24	56
ANN				0.70	0.30
					Pomme - Pomme
SumDist					18.74
N					32
ANN					0.65

analyzed.

Table 5.3 presents the level of ANN pairwise index. The ANN ratio allows to quantify and compare patterns in distributions for stores within- and inter- different fast-food chains. If the index is less than 1, the pattern exhibits clustering. If the index is greater than 1, the trend is toward dispersion. The index of around 1 suggests that the pattern is random.

⋆ Case #1. Paris

We first focus on inter-chain stores' location tendencies in the case of Paris. The stores of all chains considered pair-wisely, appear to cluster together in Paris, yet, this is happening with different levels of magnitude. In particular, the lowest ANN ratio of 0.2992 has been obtained in case of pairwise relations between the stores of Paul and Pomme indicating the strongest

Table 5.2: The pairwise ANN index. Case #2: Paris and its closer suburbs

Case #2: Paris and its closer suburbs					
	McDo - McDo	McDo - Quick	McDo - KFC	McDo - Paul	McDo - Pomme
SumDist	175.74	309.11	318.59	306.06	468.73
N	164	212	191	221	207
ANN	0.99	1.54	1.67	1.49	2.36
		Quick - Quick	Quick - KFC	Quick - Paul	Quick - Pomme
SumDist		84.09	124.45	158.20	179.91
N		48	75	105	91
ANN		0.88	1.04	1.12	1.37
			KFC - KFC	KFC - Paul	KFC - Pomme
SumDist			74.90	140.02	149.32
N			27	84	70
ANN			1.04	1.11	1.29
				Paul - Paul	Paul - Pomme
SumDist				92.42	168.90
N				57	100
ANN				0.89	1.22
					Pomme - Pomme
SumDist					47.06
N					43
ANN					0.52

Table 5.3: The pairwise ANN index. Case #3: The Paris region

Case #3: The Paris region					
	McDo - McDo	McDo - Quick	McDo - KFC	McDo - Paul	McDo - Pomme
SumDist	672.57	1251.77	1260.17	1518.48	1709.62
N	305	380	356	386	364
ANN	0.70	1.17	1.22	1.41	1.64
		Quick - Quick	Quick - KFC	Quick - Paul	Quick - Pomme
SumDist		279.16	390.72	407.89	541.51
N		75	126	156	134
ANN		0.59	0.64	0.60	0.85
			KFC - KFC	KFC - Paul	KFC - Pomme
SumDist			258.88	491.21	440.55
N			51	132	110
ANN			0.66	0.78	0.77
				Paul - Paul	Paul - Pomme
SumDist				258.43	439.58
N				81	140
ANN				0.52	2.69
					Pomme - Pomme
SumDist					142.26
N					59
ANN					0.34

clustering effects between these two firms. In addition, stores of Quick tend to herd quite intensively with all other outlets (Paul, Pomme, McDonald's, and KFC) setting the index levels between 0.3952 and 0.4726. The ratio for the KFC-Paul pair is at the level of 0.5687 indicating some level of clustering. The McDonald's-Pomme and McDonald's-Paul ratios equal to 0.6744 and 0.6031, respectively, suggesting also some forms of herding. In addition, outlets of McDonald's and KFC seem to cluster together in Paris, yet, the index of 0.7078 means that this process is not very intense. The KFC and Pomme chains do not express strong willingness to accrue (index of 0.7823).

⋆ Case #2: Paris and its closer suburbs

Contrary to what is being observed in Paris, stores of McDonald's and outlets that belong to all other chains analyzed pair-wisely seem to disperse when we enlarge the considered area to Paris along with its closer suburbs. This effect is particularly strong for McDonald's and Pomme, and McDonald's and KFC, resulting in the index at the high level of 2.3604 and 1.6702, respectively. Stores which belong to Quick and Pomme tend to disperse as well (the ANN index is at the level of 1.3664). For various pairwise relationships the index is close to 1 indicating random mutual locations between the following pairs of chains: Quick-KFC and Quick-Paul. Interestingly, note that this is not the case when only the area of Paris alone is being analyzed. In addition, the index for KFC-Paul in case of establishments locating in Paris or its surrounding territory indicates a uniform distribution of the stores (1.1069). Stores of KFC and Pomme, and Pomme and Paul seem to rather avoid each other (1.2931 and 1.2237, respectively) in this enlarged territory.

⋆ Case #3: The Paris region

Only two out of five analyzed chains, namely McDonald's and KFC, opt to locate almost half of their units in the outer suburbs. For other chains, this fraction is at the level of 16%-27%. McDonald's and other stores seem to disperse from each other when the area of the Paris region is taken into account. It is especially pronounced in case of McDonald's compared to Pomme and McDonald's compared to Paul. The stores of Quick still appear to locate in close proximity to other rival chains (ratios are between 0.5959 - 0.8536). When the Paris region area is taken into consideration, KFC outlets do not display strong tendencies. The highest levels of ANN index, 2.6917 and 1.6352, are obtained in case of pairwise relationships between Pomme and Paul, and McDonald's and Pomme. Again, it should be highlighted that the Pomme chain concentrates its efforts mainly in the city of Paris.

We look next at the behavior of same-chain stores. Interestingly, the ANN index for most of the same-chain outlets located in Paris or in the area of Paris extended to its closer suburbs is nearly one indicating that these chains try to uniformly distribute their stores over these two markets. The strategy of opening stores at equidistant locations throughout the market has been used by McDonald's, Quick, and KFC in case of the city of Paris and McDonald's and KFC in case of the area of Paris along with its inner ring. Only stores of Paul and Pomme express tendencies to cluster their stores within Paris focusing their efforts on some specific zones, and does so Pomme in the "Paris and its closer suburbs" case.

Two brands, namely, KFC and McDonald's seem to master the equidistant location strategy over the Paris market and the area of Paris together with its closer suburbs, where every store is opened at equidistant locations. The most pronounced difference is in the number of stores established by McDonald's and KFC. Interestingly, McDonald's opens 6 times more outlets than KFC in the area of Paris along with its closer suburbs (164 versus 27) or 5 times more stores in case of the Paris region (305 versus 51 in total). Overall, KFC is the weakest player considering the number of openings compared to all other chains. Finally, when the area of the Paris region is studied, it can be observed that stores of Paul strongly cluster (0.3380) and do so the outlets of other chains, yet, the magnitude is relatively lower (index takes the value between 0.5240-0.7028).

Overall, visualization of exact locations of stores and the ANN analysis permit to discover that some chains tend to asymmetrically segment the market and use different location patterns in different zones of the studied market which confirms the observations made by Takaki and Matsubayashi (2013) on differentiations of strategies depending on the area considered. To summarize, the strongest tendencies to cluster in Paris are observed between the stores of Pomme and Paul, as well as stores of Quick and all other chains taken pair-wisely. This suggests that these pairs of chains opt for the head-to-head locations keeping only very small distances from each other and exhibiting a special type of interlacing (Takaki and Matsubayashi, 2013). In the extended area of Paris along with its closer suburbs, all of the chain pairs of both same-chain and inter-chain stores, beside the Pomme company, disperse or display a perfectly uniform distribution. In addition, the obtained results suggest that Pomme is prone to concentrate its efforts on a particular confined area by opening many stores covering relatively small surface, as compared with its competitors, whereas this firm seems to opt for another strategy in the outer suburbs of the Paris region. Finally, when the whole Paris region area is investigated, two chains, concretely McDonald's and KFC seem to keep relatively the longest distances between their establishments. Another very interesting example of strategy diversification is presented by Quick. This chain locates its stores in very close proximity to outlets of other chains in Paris, yet, it does not appear to be the case when the area of Paris along with its closer suburbs is analyzed.

To visualize different locational strategies of all chains being analyzed, interactive maps have been created using Google My Maps[29]. We investigate next the possible incentives which attract establishments to locate in some particular areas in the subsection 5.6 introducing a location choice model for multi-store firms.

5.6 Model: Reduced-form profit function

In the literature, structural models are often estimated using the nested fixed point algorithm. While this is a useful approach, it often requires considerable central processing unit time and numerical sophistication on the part of the econometrician (Bajari et al., 2013). In the review of Draganska et al. (2008), we read that the computational cost of the nested fixed-point algorithm has prompted the development of alternative methods, such as the two-step approach of Bajari et al. (2007) and the nested pseudo likelihood by Aguirregabiria and Mira (2007). Alternatives to the likelihood-based approaches are method of moment (generalized method of moments) estimators. These estimators rely on the specification of moment conditions based on a structural game.

As highlighted by Seim (2006), the complexity of the expressions for the probabilities increases in the number of locations and firm types. Simulation evidence suggests, however, that the result from simpler model carries over. Scant academic research has been conducted to develop empirically tractable models of multiple-store firms' choices, as summarized by Nishida (2015), it is due to the fact that solving and estimating an equilibrium model of store-network choice pose substantial computational challenge.

[29]Visit https://www.google.com/maps/d/edit?mid=1PZpVCQB6SR7-hqhecsEjzm93k90 to see examples of interactive maps.

We empirically analyze the behavior of five, potentially asymmetric, chains in the Paris region, which is economically a highly heterogenous market. As mentioned in the introductory Section 5.1, even the most recent papers still keep on making some unrealistic assumptions, such as: no more than two (symmetric) firms competing in a market, a linear market, or uniformly distributed consumers. Moreover, most of the papers considering the multi-store firm location are theoretical. In this project we do not strive for the most elaborate or econometrically sophisticated model. Instead, we try to rectify several misconceptions made in other papers considering the problems of, *e.g.*, demand representation, daytime versus nighttime population, or the distance to the store. We apply the reduced-form model suggested first by Bresnahan and Reiss (1991) and further explored by Mazzeo (2002), Toivanen and Waterson (2005), Reiss (1996), and Berry (1992), Bresnahan and Reiss (1991).

$$\pi = X_m \beta + g(\theta; \overrightarrow{N}) + \epsilon_m. \tag{5.4}$$

The first term stands for the market demand characteristics that influence firm payoffs. It should be mentioned that the effect of X_m varies by chain, so it is firm specific. The term $g(\theta; \overrightarrow{N})$ represents the part of the payoffs which captures the competitive effects. The vector $\overrightarrow{N}$ stands for the stocks of within- and inter-chain competitors. The competitive effects may differ by type. The error term is assumed to vary with type in a given market and represents unobserved payoffs from operating as a particular brand in a given market. The error term is assumed additively separable, independent of the observables, including the number of market competition, and identical for each chain of the same type in a given market. This profit function specification makes the estimation tractable. We leave the extension of this model for the future research.

To extend this model, for instance, Bajari et al., (2013) apply the two-step estimator. In the first step, the econometrician estimates the reduced-form of the model and in the second step, the structural parameters take the reduced form. The reduced form is the distribution of the dependent variable given the exogenous variables in the model. The authors find that with large sample sizes, results are reasonably robust to the specification of the first stage so long as it is sensibly and flexibly specified.

We use the conditional Logit technique to model the location decisions of newly-formed stores, *i.e.*, opened in or after the year 2001. In addition, since the distributions of our dependent variables are relatively skewed, we test also the Tobit - equivalent to the left-censored - model. The skewness of fast-food chains data has been noticed also by Toivanen and Waterson (2005), who analyze the McDonald's and Burger King data for London. Toivanen and Waterson (2005) admit that there is an issue of sample for estimation since a substantial proportion of markets has no outlets of either firm.

To remind, in our application, fast-food chains can choose between 175 alternative markets to locate their stores. These are 80 districts of Paris and 95 EGT zones. We observe no new outlets of McDonald's in 103 out of 175 of markets of the Paris region. It means that 125 newly-created outlets of McDonald's can be found in 72 zones. Moreover, 139 out of 175 alternative markets are left with no newly-formed stores of Quick, 133 with no KFC outlets, 158 with no points of Paul, and finally, POM leaves 134 out of 175 proposed zones with none of its newly-created stores. The fact that a substantial proportion of markets has no outlets of any firm can be also observed through the location pattern maps presented in Fig. 5.9.

5.7 Empirical results

5.7.1 Market characteristics

Demand, cost, and competitive factors contribute to profits of firms. The profit function includes both, market characteristics and market structure. We first study how market characteristics influence the behavior of five chains analyzed. We present the results of the conditional Logit and compare it with the Tobit model's recommendations. A number of papers analyze the fast-food chains' behavior (Toivanen and Waterson, 2005; Igami and Yang, 2016; Thomadsen, 2007) or the coffee shops locations (Iida and Matsubayashi, 2011). Yet, the focus in this paper is to reconsider the list of variables taken to proxy the demand level in the markets. Our motivation has been clearly put over in subsection 5.2.5. In the survey by Draganska et al. (2008), the authors call for a more realistic representation of demand in location choice models. Later on, other authors including Fransen et al. (2015), Wang (2012), Nishida (2015), Igami and Yang (2016), Chang and Hsieh (2014), and the founder of Decision Analyst, Jerry W. Thomas (2007), still express this burning need. These works have substantailly stimulated the choice of our approach.

Some restrictive assumptions about customer behavior are still present in the existing models. One of the most common mistakes is targeting people instead of targeting money. Not all residents of each community travel the full length of the distance between their community and the store location each time they visit a store. Not the entire population should be treated as potential customers. Igami and Yang (2016) state that usually applied population and income might not adequately capture the demand. Population statistics usually reflect the number of residents but not necessarily the real daytime population, potentially critical in the fast-food industry. For instance, Toivanen and Waterson (2005) use the absolute number of population, proportion of young population under 16-year old, the proportion of pensioners, and unemployed rate. The authors state that population is likely the leading determinant of market size. They add the age group controls since different age groups may display different tastes for eating hamburgers. Yet, in the reduced model of Toivanen and Waterson (2005), population turns out to be insignificant in case of Burger King and significant positive for McDonald's. A clear distinction between a daytime versus nighttime population present in a particular area is to be made. To better represent demand, for instance, Igami and Yang (2016) propose other alternative measures, such as the size of office or commercial floor space. However, these measures turn out to be out of reach to Igami and Yang. Another solution they suggest is to test instead the number of traffic lights.

Taking into account all these remarks, we test the following demand proxies:

1. Car flows, pedestrian or all active modes flows (walk, bike and other active modes), and flows of people using public transport by income level (low, medium, high income classes), age (18-25, 25-34, and plus), or purpose of the trip (restaurant, shopping, leisure, professional meeting), hour by hour along the day, and day by day along the week.

2. Flows of tourists visiting the most attractive sites of the Paris region, such as Disneyland Paris or Musée du Louvre. These are the places which host over 0.5 million and 1 million tourists per year.

3. Proportion of a zone dedicated to commercial centers, commerce, offices, universities and high schools, hospitals, cultural and leisure equipment, congress and exposition centers. Thus, the proxies suggested by Igami and Yang (2016), *i.e.*, the size of office or commercial floor space, have been successfully obtained and tested in the models.

4. Number of train, express underground and subway stations.

5. Train stations: Gare St-Lazare, Gare du Nord, Gare de l'Est, Gare de Lyon, Gare d'Austerlitz, Gare Montparnasse.

6. Airports of Charles de Gaulle and Paris Orly.

Firms search for large sources of demand (Thomadsen, 2007). Beside the flows of people computed through the EGT Survey, we add variables which proxy the mass of potential clients passing through the commercial centers, train, express underground and subway stations, and airports. These are also the places where the fast-food chains declare to locate their stores (see, *e.g.*, the official website of Paul for details). Fig. 5.1 gives one example of data representation. This figure identifies, among others, the location points of the most famous touristic sites and all the train stations in the Paris region. All these proxies have been examined and the followingones have been eventually selected: flows of people doing shopping, going to restaurants, or making trips for a leisure purpose; flows of people characterized by low and medium income; and others which represent the crowds of potential clients passing through the commercial centers, train, express underground and subway stations, airports, touristic sites, highways, and nearby high schools, universities, and offices.

When comparing the first part of the results, *i.e.*, how market characteristics influence the behavior of fast-food chains, especially with the fundings of Toivanen and Waterson (2005) for McDonald's and Burger King locating in London, we notice that overall, the variables used to proxy the demand are highly significant in the fast-food chain location choice models.

We testify that the results of conditional Logit and Tobit models provide very similar results. For the comparison purposes, we present both model estimations through Tables 5.4 - 5.8. Later on, we focus our attention on the estimation of discrete choice model's recommendations.

Results: All fast-food chains

We first look at the behavior of the market leader. McDonald's aims at catching a large number of people in airports, commercial centers, or in the proximity to universities or high schools. Moreover, flows of people doing their shopping, spending their time on some leisure activities or going out with a purpose of eating in a restaurant seem to significantly attract McDonald's to establish its businesses nearby their sources.

We further notice that Quick opts for locating inside the commercial centers, on highways, or nearby sites where it can catch a large number of tourists and middle-income persons. The variable "Tourists" becomes much less significant in the Tobit model. KFC seems to be a quite particular fast-food chain for whom sufficiently high flows of middle-income people send the positive sign to locate its stores nearby. In case of this firm, other variables tested do not turn out to significantly influence its behavior. In case of Paul, a variable that represents flows of people from lower-income categories carries a negative sign. On the contrary, rather middle-income people display a tendency to buy in the outlets of Paul. This chain tends to locate nearby large aggregations of people passing through the train, express underground and subway stations. In addition, a variable "Airports" shows a positive, yet, not very significant effect on the locational decisions of Paul (probability is at the level of 0,2723) according to the results of the conditional Logit model. This positive effect becomes slightly more pronounced in the Tobit model (probability equals to 0,1530). We finally observe, that Pomme tends to locate its outlets in the commercial centers and nearby offices and famous touristic sites, and other places where flows of medium income people is elevated.

In addition, contrary to the results of Toivanen and Waterson (2005), we obtain the significant and positive effect of variable area in all chains being analyzed beside Paul. Toivanen and Waterson (2005) notice that in both, reduced and structural, models, the variable area carries insignificant coefficient for fast-food chains.

Finally, by looking at the results from all the models presented in this paper, it seems true that firms open more stores where more potential consumer agglomerate, which is in line with the paper by Janssen et al. (2005). We also notice that consumers seem to be heterogeneous considering their preferences over various firms (Janssen et al., 2005). For instance, Paul tends to avoid flows of people from lower-income class. Another example shows that Pomme tries to locate nearby offices, whereas McDonald's goes, among others, after younger generation, establishing its outlets in close proximity to high schools and universities. Overall, firms are likely to set more locations where their potential consumers are prone to buy from them, as it was shown in Janssen et al. (2005).

Our main contributions, after analyzing the first part of the model results, are to demonstrate that, contrary to the paper of Toivanen and Waterson (2005), demand, when properly accounted for, turns out to be very significant for all the chains studied. Yet, different demand characteristics shall be used depending on the chain being analyzed. In addition, it should be

Table 5.4: Location choices of McDonald's: Focus on market characteristics

	MCD: cLogit Est	Pr > ChiSq	MCD: Tobit Est	Pr > |t|
Constant	-16.7196	<.0001	-16.2145	<.0001
Area (log)	1.3227	<.0001	1.8530	<.0001
Airports (dummy)	0.9486	0.0252	1.9920	0.0981
Commercial centers (% of zone)	22.9299	0.0004	46.6546	0.0171
High schools, universities (% of zone)	7.2468	0.0348	14.4959	0.0251
Flows: Shopping, restaurants, leisure (log)	1.0444	<.0001	1.4404	<.0001
Sigma			2.0467	<.0001
	Log-likelihood	-729.01	Log-likelihood	-205.64
	Full log-likelihood	-729.01	AIC	425.27
	AIC	1470.01	Schwarz crit.	447.42
	AICC	1470.02		
	BIC	1517.97		

clarified that fast-food firms react differently to market characteristics. Thus, different sets of variables are finally selected to build the models of five fast-food chain location decisions, confirming the results of, *e.g.*, Toivanen and Waterson (2005). Assumptions that firms are symmetric made in several papers (*e.g.*, Takaki and Matsubayashi, 2013; Seim, 2006) seem not to be valid. Our results caution against the use of simpler representation of demand. Our method adds behavioral realism to the existing models.

5.7.2 Market structure

Other factors of interest in this paper influencing the location decisions are the market structure variables. We include the same- and inter-chain stocks of outlets. We test a few different forms of variable which expresses the pre-existing stocks of establishments, concretely, number of establishments of a chain i; logarithmic values of stocks; densities of stocks (number of units over the zone's surface); and we introduce the market structure dummies. In the latter case, dummies for a particular chain i take a value of one (1) if a particular chain i locates more of its stores in a market than there are stores of another chain j, where i, j are McDonald's, Quick, KFC, Paul, or Pomme and $i \neq j$ or (2) if a market is dominated by the leader. Based on the descriptive statistics presented in subsection 5.4, we assume McDonald's to be the leader. Only one pre-existing establishment of Pomme was registered before the

Table 5.5: Location choices of Quick: Focus on market characteristics

	Quick: cLogit		Quick: Tobit	
	Est	Pr > ChiSq	Est	Pr > |t|
Constant	-17.0206	<.0001	-11.7786	0.0009
Area (log)	0.6870	0.0123	0.5847	0.0215
Commercial centers (% of zone)	23.0050	0.0415	29.4570	0.0761
Flows of medium income people (log)	0.8742	0.0010	0.7920	0.0039
Tourists (log)	0.8660	0.0408	0.6841	0.1983
Highways (log)	0.3113	0.1057	0.3165	0.1129
Sigma			1.5718	<.0001
	Log-likelihood	-249.92	Log-likelihood	-114.84
	Full log-likelihood	-249.92	AIC	243.68
	AIC	511.85	Schwarz crit.	265.83
	AICC	511.86		
	BIC	553.40		

Table 5.6: Location choices of KFC: Focus on market characteristics

	KFC: cLogit		KFC: Tobit	
	Est	Pr > ChiSq	Est	Pr > |t|
Constant	-19.5181	<.0001	-13.7269	<.0001
Area (log)	1.1000	<.0001	0.9690	<.0001
Flows of medium income people (log)	1.0700	<.0001	0.9772	0.0001
Sigma			1.3676	<.0001
	Log-likelihood	-273.58	Log-likelihood	-120.76
	Full log-likelihood	-273.58	AIC	249.53
	AIC	553.15	Schwarz crit.	262.19
	AICC	553.16		
	BIC	574.20		

Table 5.7: Location choices of Paul: Focus on market characteristics

	Pau: cLogit		Pau: Tobit	
	Est	Pr > ChiSq	Est	Pr > |t|
Constant	-19.8548	0.0018	-15.5490	0.0246
Area (log)	0.1715	0,6543	0.0958	0.8095
Airports (dummy)	1.1941	0.2723	1.9193	0.1530
Flows of low income people (log)	-1.2455	0.0348	-1.1765	0.0571
Flows of medium income people (log)	2.2547	0.0063	2.0741	0.0199
Train, subway stations (log)	0.8802	0.0872	0.9723	0.0912
Sigma			1.7637	<.0001
	Log-likelihood	-107.03	Log-likelihood	-63.06
	Full log-likelihood	-107.03	AIC	140.12
	AIC	226.07	Schwarz crit.	162.27
	AICC	226.09		
	BIC	262.72		

Table 5.8: Location choices of Pomme: Focus on market characteristics

	Pomme: cLogit		Pomme: Tobit	
	Est	Pr > ChiSq	Est	Pr > |t|
Constant	-11.7071	<.0001	-7.7669	0.0266
Area (log)	0.3651	0.0226	0.3726	0.0768
Commercial centers (% of zone)	20.0631	0.0015	43.4724	0.0116
Offices (% of zone)	3.2602	0.0103	5.5614	0.0187
Flows of medium income people (log)	0.4885	0.0197	0.4831	0.0828
Tourists (log)	0.9377	0.0004	1.1758	0.0290
Sigma			1.7841	<.0001
	Log-likelihood	-337.34	Log-likelihood	-134.38
	Full log-likelihood	-337.34	AIC	282.76
	AIC	686.69	Schwarz Crit.	304.92
	AICC	686.70		
	BIC	730.04		

Table 5.9: Number of newly-created and existing stores

	McDonald's	Quick	KFC	Paul	Pomme	Total
New (created in or after 2001)	125	43	47	19	58	292
Stock	180	32	4	62	1	279
Total	305	75	51	81[a]	59	571

[a]The headquarters of Paul is located in Belgium.

year 2001. We therefore remove the variable stock of Pomme from the analysis. One should be careful about the stock of KFC, since only four stores of KFC were registered before 2001 (see Table 5.9).

Following the idea of Toivanen and Waterson (2005), we add dummies to control for the omitted market structure, *i.e.*, markets where there are no pre-existing stores established in a market. We test two variables. The first dummy being considered, builds on the stocks of all the analyzed chains taken together. This variable demonstrates whether chains locate their outlets in the new, unfrequented markets, where none of the firms has yet arrived. The second dummy takes one if there is no stock of a particular chain.

In addition, to control for market specific unobservables, Toivanen and Waterson (2005) suggest that the number of own and rival outlets in neighboring markets be used. Since some characteristics of potential customers are correlated among neighboring markets, firms may use information they learn in adjacent markets to update their knowledge of market size in a particular market. We account for the outlets located in the proximity of the analyzed markets using the accessibility measure.

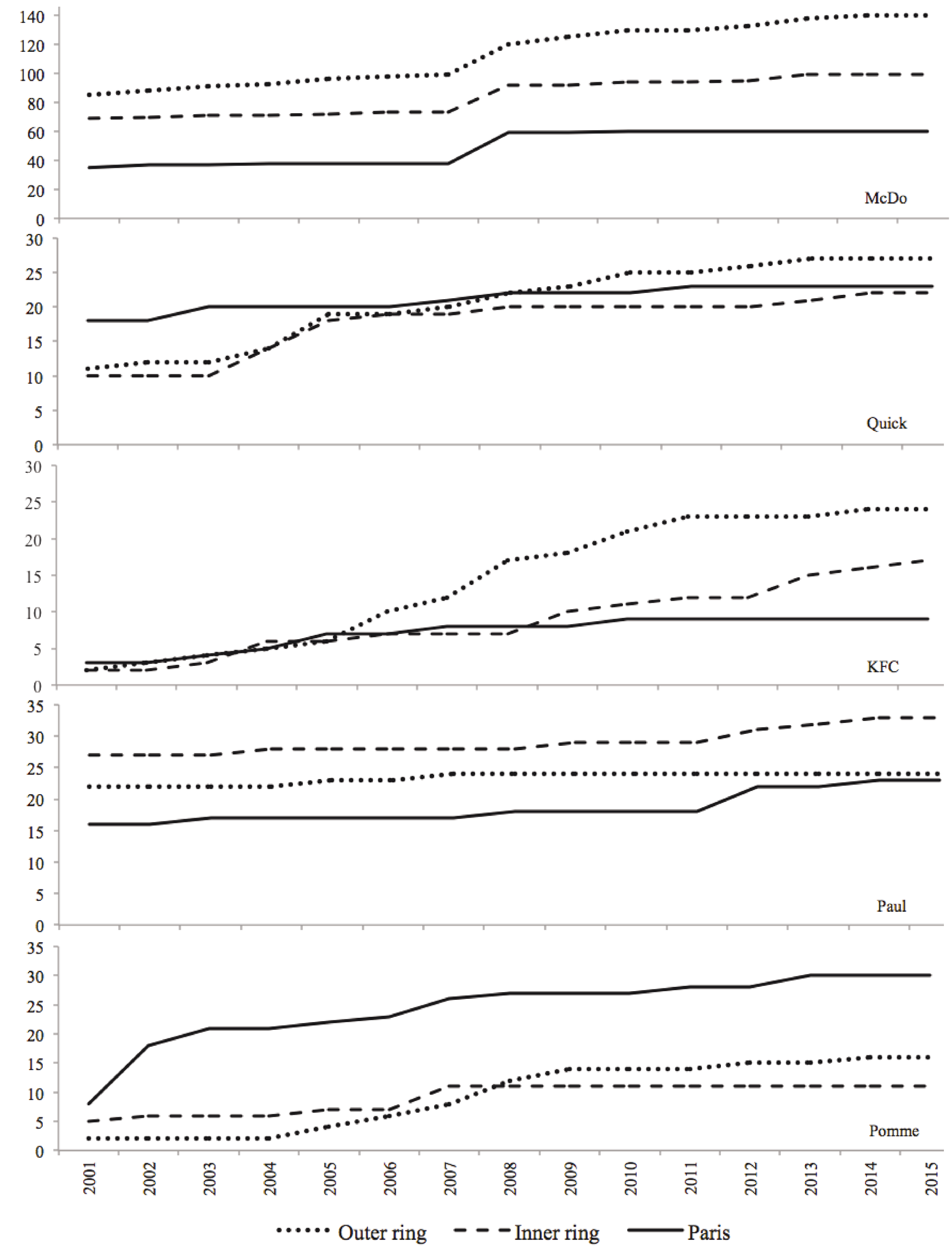
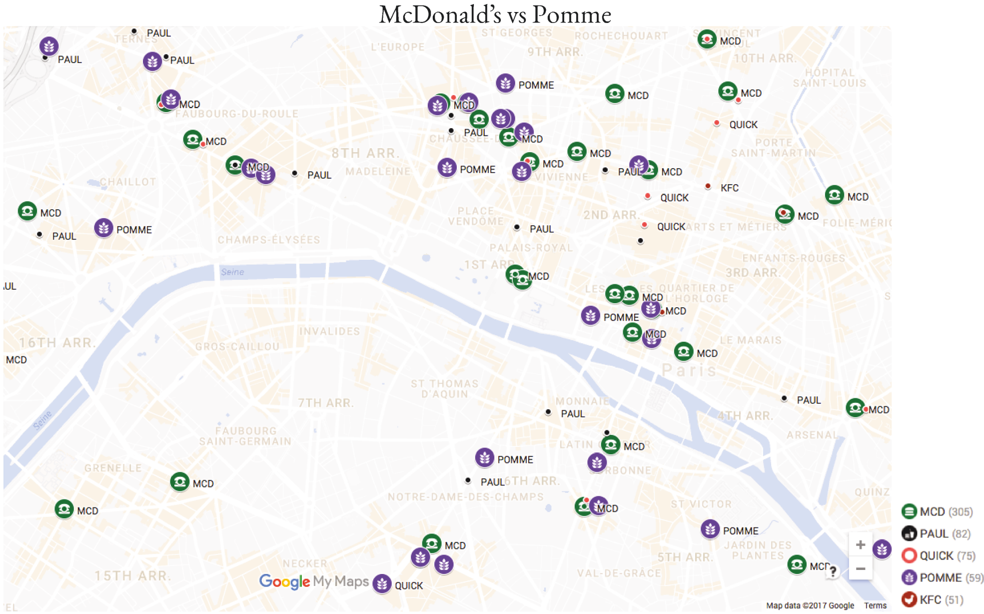


Figure 5.5: The change of location strategy over time: The change of the attractiveness level of Paris, the closer and outer suburbs for the best performing chains. Focus on the situation between 2001 and 2015. Source: Calculations based on the SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

McDonald's vs Pomme



McDonald's vs Pomme and Paul

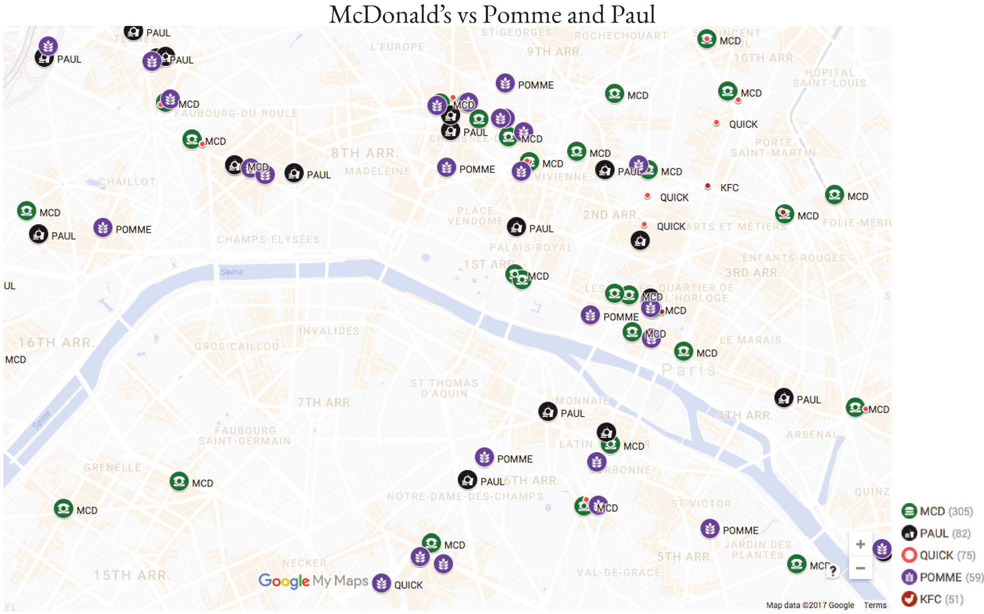
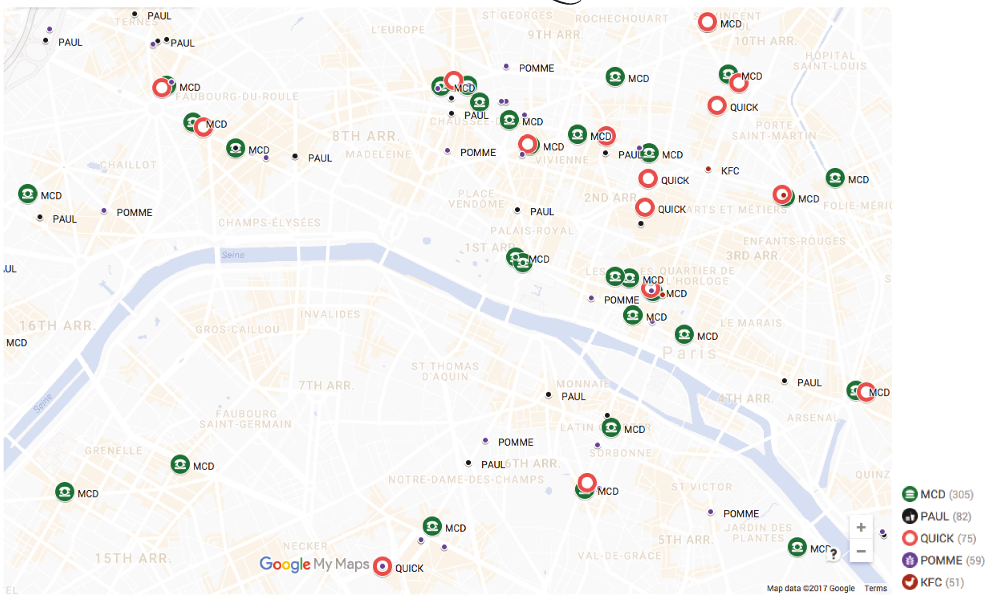


Figure 5.6: Examples of location strategies of McDonald's versus Pomme and McDonald's versus Pomme and Paul in the central part of Paris in 2015 using Google My Maps. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

McDonald’s vs Quick



McDonald’s vs KFC

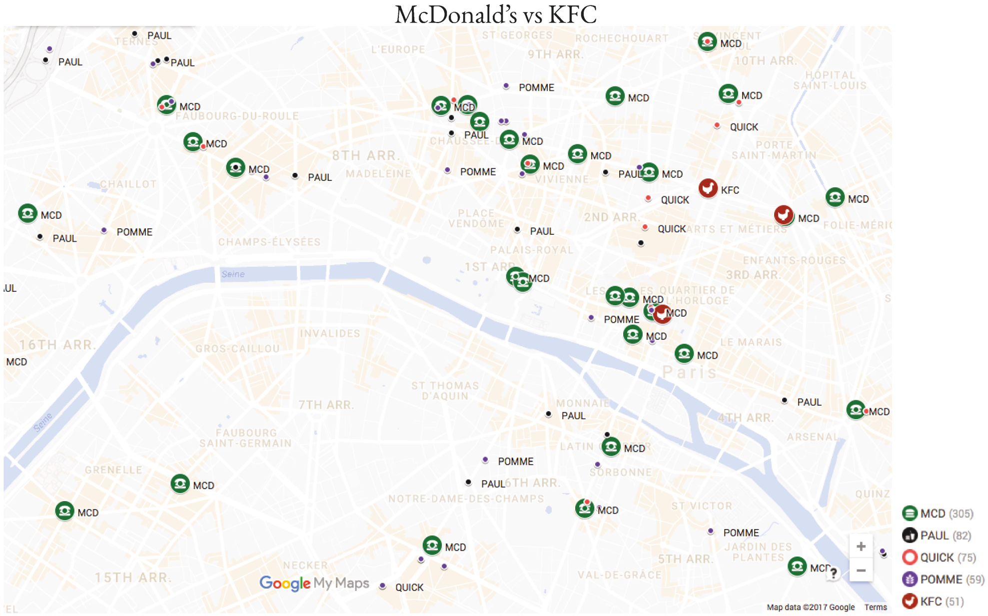


Figure 5.7: Examples of location strategies of McDonald’s versus Quick and McDonald’s versus KFC in the central part of Paris in 2015 using Google My Maps. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

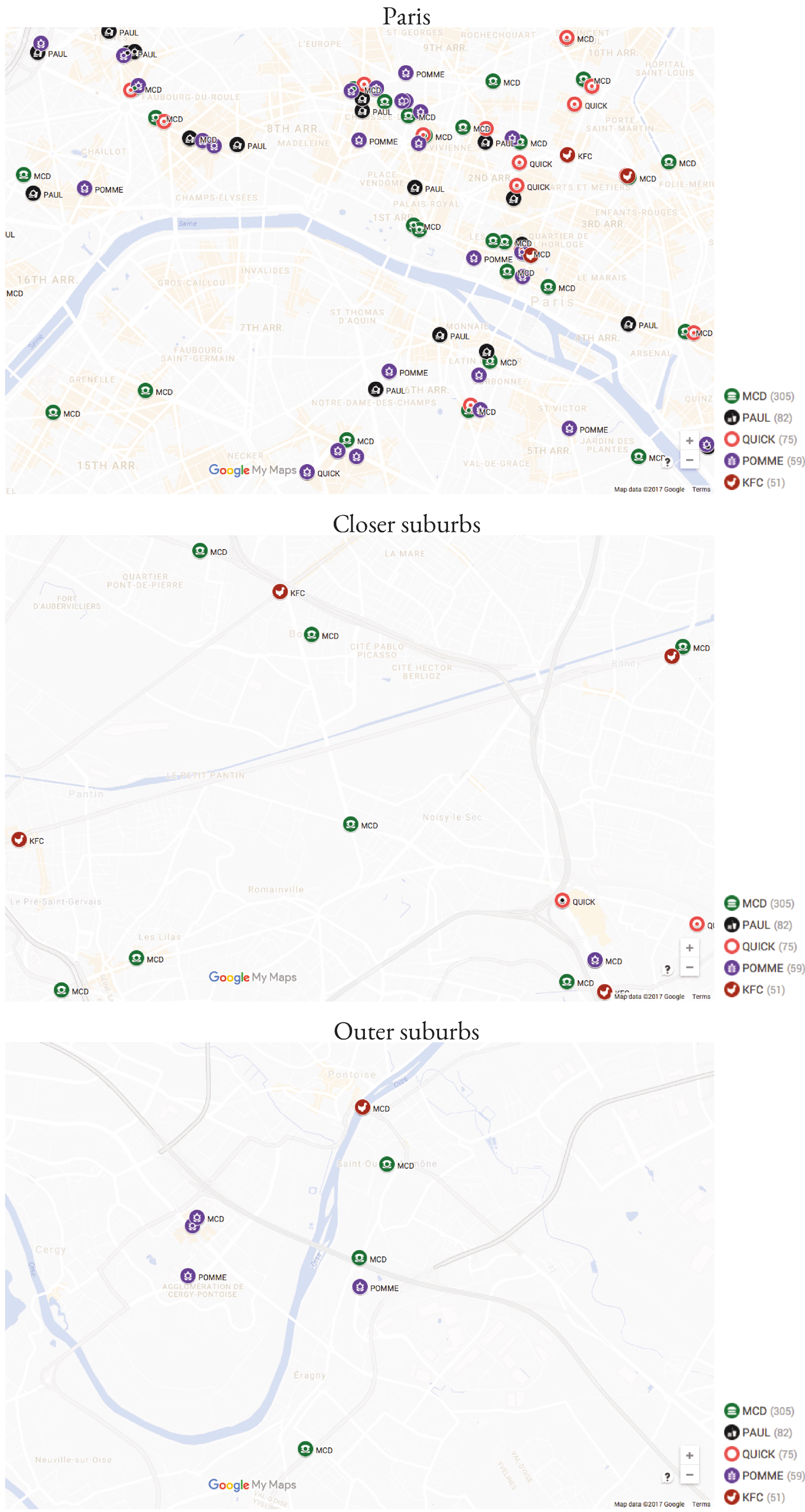


Figure 5.8: Examples of location strategies of establishments of McDonald's, KFC, Quick, Paul, and Pomme de Pain in Paris, the closer suburbs, and the outer suburbs in 2015 using Google My Maps. The same scale has been kept through Fig. 5.6 - 5.8, so that the between store distances can be compared. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

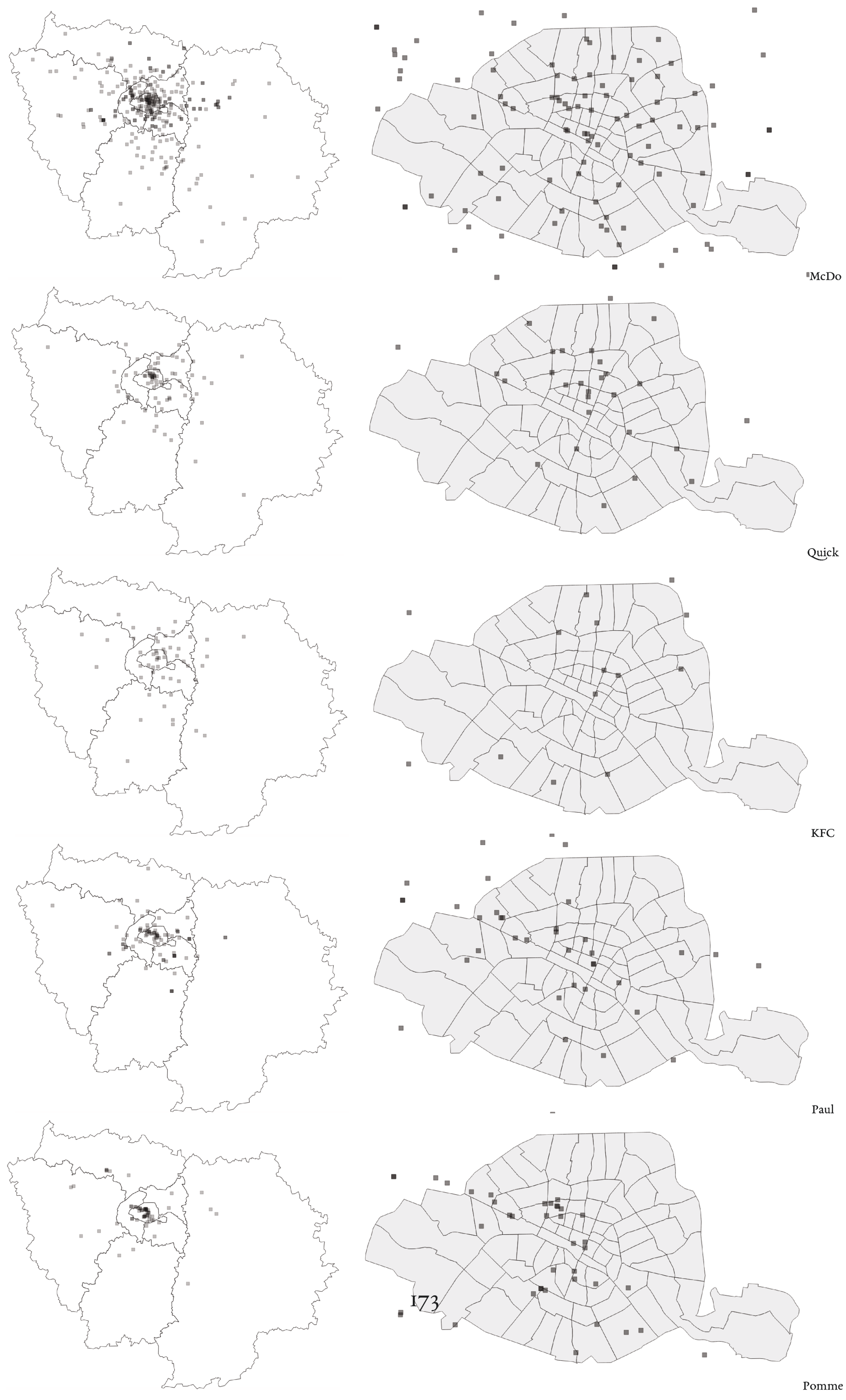


Figure 5.9: Locations of establishments of McDonald's, Quick, KFC, Paul, and Pomme de Pain in the Paris region (pictures on the left) and in 80 districts of Paris (pictures on the right) at the individual establishment level. Situation in 2015. Source: SIRENE/INSEE database verified and extended using the addresses provided through the official websites of all the analyzed fast-food chains

Table 5.10: Location choices of McDonald's: Focus on market structure (cLogit)

	MCD: Example #1 Est	MCD: Example #1 Pr > ChiSq	MCD: Example #2 Est	MCD: Example #2 Pr > ChiSq	MCD: Example #3 Est	MCD: Example #3 Pr > ChiSq
Constant	-16.3160	<.0001	-16.1744	<.0001	-17.5760	<.0001
Area (log)	1.3083	<.0001	1.3189	<.0001	1.4424	<.0001
Airports (dummy)	0.9666	0.0235	1.0043	0.0219	1.0760	0.0129
Commercial centers (% of zone)	23.7378	0.0004	21.2073	0.0019	20.2887	0.0020
High schools, universities (% of zone)	7.4543	0.0317	6.2167	0.0844	6.3634	0.0660
Flows: Shopping, restaurants, leisure (log)	1.0012	<.0001	1.0180	<.0001	1.1473	<.0001
Same-chain stores (spillovers)	-0.0299	0.0731				
Inter-chain stores (spillovers)	0.0179	0.1467				
At least 1 McDonald's in a market (dummy)			- 0.5285	0.0223		
At least 1 Quick in a market (dummy)			- 0.0266	0.9143		
At least 1 KFC in a market (dummy)			0.1754	0.6974		
At least 1 Paul in a market (dummy)			0.5071	0.0175		
McDonald's dominates Quick (dummy)					-0.3453	0.1094
Log-likelihood		-726.90		-724.32		-727.76
Full log-likelihood		-726.90		-724.33		-727.76
AIC		1469.81		1468.65		1469.52
AICC		1469.81		1468.66		1469.53
BIC		1533.75		1548.58		1525.48

Results: The leader

First, we present the results from a number of models describing the behavior of the leader, McDonald's. These models test different combinations of market structure variables. For instance, the first model presented in Table 5.10 (example #1) includes information on previous location choices of its own stores (within-chain effects) and past decisions of all its competitors (inter-chain effects). In this model, we apply an accessibility measure to compute the within-chain and inter-chain effects to account for spatial spillovers from the nearby markets to control for market specific unobservables. We observe that McDonald's avoids to locate its new stores in the markets where it has already entered. This can indicate the negative cannibalization effects. On the other hand, McDonald's is willing to open its stores where the other fast-food firms are already active. This effect is positive, yet rather weak.

The second model presented in Table 5.10 (example #2) builds on information on previous location choices of within-chain stores and past decisions of all its competitors (inter-chain effects) considered individually. We introduce dummies to control for the omitted market structure, *i.e.*, markets where there are no pre-existing stores of a particular chain. We notice that McDonald's tends to avoid markets where its stores have been already established, possibly keeping in mind the negative effects from cannibalization. McDonald's is likely to search for the markets where at least one store of Paul can be found nearby.

The third selected model shown in Table 5.10 (example #3) indicates how the leader reacts to its potentially main rival, Quick. This model introduces a dummy, which takes one if McDonald's dominates over Quick in a particular market. The results of model in Table 5.10 suggest that McDonald's tries to stay clear from the sites, where the domination of McDonald's over Quick has been already established.

It is also worth highligting that the market characteristics variables stay quite insensitive to further extensions using the market structure variables.

Results: All fast-food chains

We look now into the behavior and location tendencies of all five chains being studied. Based on the model estimations, the statistics presented in subsection 5.4, and the levels of the ANN index computed in subsection 5.5, it becomes visible, that the observed location patterns of analyzed fast-food chains vary significantly confirming the observations of Toivanen and

Waterson (2005). The firms' profit functions seem to be heterogenous. Fast-food chains are prone to react differently to market characteristics and market structure. The set of variables ultimately selected to build a location choice model for each fast-food chain diverges between the firms. In this case, imposing symmetry upon firms' profit functions might be incorrect, even if they seem similar upon first inspection.

The leader, McDonald's, enters into the market more frequently than its rivals and it operates over half (305 out of 571) of all shops which belong to five biggest players in the Paris region. In addition, around 43% of the newly-created stores, *i.e.*, stores opened in or after the year 2001, belong to McDonald's.

We summarize next the results from the models which try to answer two following questions: Are fast-food firms prone to locate in the new markets? Does the situation from the nearby markets is being considered when choosing a site? We verify the hypothesis made by Toivanen and Waterson (2005), that location in new markets is less likely than location in markets where at least one of the analyzed players is already present. We use a variable which represents new markets and a variable which proxies the overall situation in the nearby markets. A control variable for the surface of a zone has been equally added.

It turns out (see Table 5.11), that the observation made by Toivanen and Waterson (2005) seems to be only partially correct in the application to the Paris region since for two out of five chains, a variable which demonstrates whether chains locate their outlets in the new markets, carries a negative coefficient and for the rest the effect is insignificant. KFC tends to strongly avoid locating in new markets. Paul's reaction is similar.

The included control variable which proxies the overall situation in the nearby markets plays a role in the decision processes of some fast-food chains. In case of all the chains being studied, the fact that nearby markets are quite busy already hosting some stores, sends a positive sign to locate in their proximity. Yet, this effect is particularly strong only for Quick meaning that it carefully considers the situation in the nearby markets when making its decisions. Pomme is also prone to do so, yet, this effect is almost inaudible.

Table 5.11: Location choices of all chains: Focus on new market and nearby markets variables (cLogit)

	MCD		Quick		KFC		Paul		Pomme	
	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq
Constant	-5.5617	<.0001	-6.0216	<.0001	-5.3962	<.0001	-5.0817	<.0001	-5.3197	<.0001
Area (log)	0.1734	0.0007	0.2692	0.0053	0.1811	0.0240	-0.1533	0.2376	-0.0084	0.9186
All chain stocks in nearby markets	0.0012	0.9529	0.2656	0.0894	-0.0198	0.6352	0.5058	0.2027	0.1276	0.1542
New market (dummy)	-0.1368	0.6375	0.4415	0.4317	-1.3596	0.0437	-1.3061	0.1149	0.4336	0.2453
Log-likelihood	-759.88		-257.80		-279.59		-111.49		-353.95	
Full log-likelihood	-759.88		-257.80		-279.59		-111.49		-353.95	
AIC	1527.76		523.60		567.19		230.98		715.89	
AICC	1527.76		523.60		567.19		230.99		715.90	
BIC	1559.73		551.30		595.25		255.42		744.79	

We study now in more details the variables that proxy the influence from the same- and inter-chain stores (see Table 5.12). Existing same-chain stores tend to decrease the probability of location choice in a particular market, indicating possible negative cannibalization effects in case of MDonald's. This effect is also negative for Quick and KFC, yet, very weak or not significant. On the other hand, we further observe very weak positive spillovers of Pomme on its stores and Paul on its outlets indicating some level of clustering.

We observe positive learning effect, *e.g.*, of Paul on McDonald's and negative stealing effect in case of other pairs of fast-food firms, *e.g.*, Quick seems to avoid locating in the direct proximity to the leader, McDonald's, indicating a potential fear of the stealing effects, confirming the fundings of Neven (1987) that profits tend to decrease along with a more intense competition. KFC tends to collocate with Quick and seem to select different markets than Paul. Quick has a tendency to follow KFC and Paul, yet, this effect is rather negligible (significance is at the level of 0.2241 and 0.2845). Pomme is also prone to collocate with Paul and KFC.

According to Toivanen and Waterson (2005), traditional industrial organization theories cannot explain the positive effect of rival presence on own entry, learning models can. The most plausible explanation of the results is therefore firm learning from each other. Their results suggest that learning effects are strong enough to dominate any negative effects that competition between firms may have on entry decisions.

In addition, the significance of dummies for Paris and the outer suburbs (keeping the closer suburbs as a reference) has been tested. When extending the models using the dummies for markets within Paris and in the far away suburbs, these variables turn out insignificant for McDonald's, Quick, Paul, and Pomme, whereas in case of KFC, its decisions are towards the faraway suburbs. This effect, however, is relatively weak.

It has been demonstrated again that the market characteristics variables stay quite insensitive to extending the models using the market structure variables.

The charts presented in Fig. 5.5 only partly confirm the fundings of Neven (1987) that early entrants tend to locate around the center and with more numerous firms, successive entrants locate in more distant slots of the market. It seems to be true in case of McDonald's and KFC. Yet, *e.g.*, Pomme concentrates its efforts in the very central parts of the Paris region. Based on the ANN ratio in case of Pomme, we confirm the observations made by Thomas (2007) that numerous fast-food restaurants chains typically focus on a limited geographic area to achieve consolidation/concentration of force. This is called a geographic segmentation. Firms

Table 5.12: Location choices of all chain stores: Focus on same- and inter-chain effects (cLogit)

	MCD		Quick		KFC		Paul		Pomme	
	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq	Est	Pr > ChiSq
Constant	-16.1350	<.0001	-16.5392	0.0002	-19.8657	<.0001	-21.2135	0.0050	-10.5210	0.0007
Area (log)	1.3046	<.0001	0.8004	0.0105	1.0614	0.0004	0.1556	0.7134	0.4034	0.0329
Commercial centers (% of zone)	21.3492	0.0018	19.3344	0.0960					16.3810	0.0168
Offices (% of zone)									4.0846	0.0035
High schools, universities (% of zone)	6.2851	0.0816								
Flows of low income people (log)							-1.2949	0.0311		
Flows of medium income people (log)			0.8348	0.0083	1.0609	0.0010	2.3195	0.0108	0.4405	0.0562
Flows: Shopping, restaurants, leisure (log)	1.0193	<.0001								
Tourists (log)			0.8442	0.0519					1.0201	0.0001
Airports (dummy)	0.9996	0.0226					1.6257	0.1465		
Train, subway stations (log)							0.9526	0.0716		
Highways (log)			0.2968	0.1416						
At least 1 McDonald's in a market (dummy)	-0.4544	0.1219	-0.8055	0.1544	0.4743	0.4236	-0.2471	0.7504	-0.2958	0.4829
At least 1 Quick in a market (dummy)	-0.0189	0.9396	-0.5978	0.1814	0.5190	0.1542	-0.3009	0.6488	-0.4777	0.2376
At least 1 KFC in a market (dummy)	0.1556	0.7324	0.8372	0.2241	-0.6398	0.5433	-0.5878	0.6183	0.9619	0.1031
At least 1 Paul in a market (dummy)	0.5177	0.0162	0.4035	0.2845	-0.5547	0.1571	0.6190	0.3242	0.4935	0.1418
New/Unfrequented market (dummy)										
Stocks in nearby markets (all chains)	-0.0086	0.6759	0.1737	0.1704	-0.0204	0.6235	0.5655	0.2484	0.0612	0.3908
Log-likelihood	-724.24		-245.92		-271.47		-104.19		-334.23	
Full log-likelihood	-724.24		-245.92		-271.47		-104.19		-334.23	
AIC	1470.48		513.84		558.94		230.38		690.47	
AICC	1470.49		513.88		558.96		230.46		690.49	
BIC	1558.40		590.03		615.06		297.58		769.94	

segment the market by attacking a restricted geographic area.

The outcomes of studying maps in Fig. 5.8 and Fig. 5.9 are in line with the observations made by Takaki and Matsubayashi (2013), that a leader may be using two opposing locational strategies: (1) a segmentation strategy where the leader opens its outlets at small and equal distances to prevent the entry of its rivals. Yet, at the same time, the cannibalization of sales between stores is probable. Thus, the segmentation strategy is likely to decrease the surplus of firms, (2) an equidistant location strategy which consists of opening stores equidistantly throughout the market.

Peng and Tabuchi (2007) believe that firms hardly establish stores at edges of our outside consumer distributions. Interestingly, when observing the decisions of KFC, this remark may not be telling the true story described by the maps Fig. 5.8 and Fig. 5.9.

5.8 Conclusions

This paper shed light on competing firms with multiple component units. Rather than striving for a very sophisticated model, we attempted to remedy certain misconceptions made in other papers. First, we considered five asymmetric fast-food chains. Further, we insisted on including intra- and inter-chain pre-existing stores, which are usually unsung in the literature. Finally, we proposed to reconsider the list of variables commonly used to capture the demand level in a market.

Firms, such as fast-food chains, search for major sources of demand trying to attract the greatest number of potential customers. Yet, some unrealistic assumptions about customer behavior are still present in the existing models. Usually applied population and income might not adequately capture the demand. One of the most common mistakes is targeting people instead of targeting money. Not all residents of each community travel the full length of the distance between their community and the store location each time they visit an outlet. Not the entire population should be treated as potential customers. Motivated by works by Draganska et al. (2008), Fransen et al. (2015), Wang (2012), Nishida (2015), Igami and Yang (2016), Chang and Hsieh (2014), and suggestions made by the founder of Decision Analyst, Jerry W. Thomas (2007), among others, we found it particularly relevant to carefully define market segments and to select potential customer groups by observing their characteristics, their mobility patterns, and purpose of their activities during the day and week. We emphasized the need to state a clear difference between the daytime and nighttime population. We tested a number of proxies to capture more appropriately the demand level, to eventually select the following ones: flows of people doing shopping, going to restaurants, or making trips for a leisure purpose; flows of people characterized by low and medium income; and others which represent the crowds of potential clients passing through the commercial centers, train, express underground and subway stations, airports, touristic sites, highways, and nearby high schools, universities, and offices.

Our main contributions were to demonstrate that the observed location patterns of analyzed fast-food chains vary significantly. Fast-food chains react differently to market characteristics and market structure. Thus many papers which assume that firms are homogenous may be incorrect. Contrary to some existing literature, demand, when properly accounted for, turned out to be very significant for all chains being studied. In addition, different demand characteristics shall be used to capture demand depending on the chain being analyzed.

For the purpose of this paper, a unique database has been created. Detailed information on five Parisian strongest players in the fast-food industry has been manually drawn (using the SIREN code of each firm) from the SIRENE/INSEE websites and its consistency has been verified with the official websites of each fast-food chain, so that some missing addresses were added to our database, not existing any more establishments deleted, or some coordinates of several entities adjusted. In this way a unique database has been built, which has been never used in any other application.

We acknowledge that the reduce-form model proposed in this paper is not a very advanced solution from the methodological and econometrical point of view. We wish to extend it further in a close future, for instance, treating it as a base for the two-stage structural model. Yet, we believe, that a necessary call for the more realistic representation of the demand and a clear difference between the daytime and nighttime population have been made and certain misconceptions made in other papers have been rectified. Our results caution in particular, against the use of simpler representation of demand. Our method adds behavioral realism to the existing models that are typically based on static, nighttime representations of population.

References

Ackerberg, D.A. (2001). Empirically Distinguishing Informative and Prestige Effects of Advertising, *RAND Journal of Economics*, The RAND Corporation, 32(2), 316-333.

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica*, 75(1), 1-53.

Aguirregabiria, V. and J. Suzuki (2016). Empirical Games of Market Entry and Spatial Competition in Retail Industries. In *Handbook on the Economics of Retailing and Distribution*, ed. by E. Basker, Edward Elgar, University of Missouri-Columbia, US.

Arcidiacono, P. and R.A. Miller (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6), 1823-1867.

Bajari, P., L. Benkard, and J. Levin (2007). Estimating dynamic models of imperfect competition. *Econometrica*, 75(5), 1331-1370.

Bajari, P., H. Hong, and D. Nekipelov (2013). Game theory and econometrics: A survey of some recent research. *Advances in Economics and Econometrics*, 10th World Congress, 3, 3-52.

Beresteanu, A., P. Ellickson, and S. Misra (2010). The Dynamics of Retail Oligopoly. *Working Paper*.

Berry, S. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4), 889-917.

Boscoe, F.P., K.A. Henry, and M.S. Zdeb (2012). A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals. *Professional Geographer*, 64(2), 188-196.

Bresnahan, T. and P. Reiss (1991). Empirical models of discrete games. *Journal of Econometrics*, 48(1), 57-81.

Carlsen, L. (2010). Americans in Paris, McDonald's in France. *Wordpress*.

Chang, H.-J. and C.-M. Hsieh (2014). A TOPSIS Model for Chain Store Location Selection. *Review of Integrative Business and Economics*, 4(1), 410-416.

Chang, H.-J., C.-M. Hsieh, and F.-M. Yang (2015). Acquiring an Optimal Retail Chain Location in China. In *Proceedings of the 2nd International Conference on Information Science and Control Engineering (ICISCE '15)*, 96-99.

Chatman, D.G., R. Noland, and N.J. Klein (2016). Firm Births, Access to Transit, and Agglomeration in Portland, Oregon, and Dallas, Texas. *Transportation Research Record*, 2598, 1-10.

Chu, C.Y. Cyrus and H.-Ch. Lu (1998). The multi-store location and pricing decisions of a spatial monopoly. *Regional Science and Urban Economics*, 28(3), 255-281.

Cliquet, G. (2006). *Geomarketing: Methods and Strategies in Spatial Marketing*. John Wiley & Sons.

de Palma, A., V. Ginsburgh, Y. Papageorgiou, and J.-F. Thisse (1985). The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica*, 53, 767-782.

Draganska, M., S. Misra, V. Aguirregabiria, P. Bajari, L. Einav, P. Ellickson, D. Horsky, S. Narayanan, Y. Orhun, R. Reiss, K. Seim, V. Singh, R. Thomadsen, and T. Zhu (2008). Discrete Choice Models of Firms' Strategic Decisions. *Marketing Letters*, 19(3-4), 399-416.

Ellickson, P.B., S. Houghton, and C. Timmins (2013). Estimating network economies in retail chains: A revealed preference approach. *The RAND Journal of Economics*, 44(2), 169-193.

Fransen, K., T. Neutens, P. De Mayer, and G. Deruyter (2015). A commuter-based two-step floating catchment area method for measuring spatial accessibility of daycare centers. *Health & Place*, 32, 65-73.

Granot, D., F. Granpot, and T. Raviv (2010). On competitive sequential location in a network with a decreasing demand intensity. *European Journal of Operational Research*, 205(2), 301-312.

Holl, A. (2004a). Manufacturing Location and Impacts of Road Transport Infrastructure: Empirical Evidence from Spain. *Regional Science and Urban Economics*, 34(3), 341-363.

Holl, A. (2004b). Transport Infrastructure, Agglomeration Economies, and Firm Birth: Empirical Evidence from Portugal. *Journal of Regional Science*, 44(4), 693-712.

Hollander, S.C. and G.S. Omura (1989). Chian Store Developments and Their Political, Strategic, and Social Interdependencies. *Journal of Retailing*, 65(3), 299.

Holmes, T.J. (2011). The Diffusion of Wal-Mart and Economics of Density. *Econometrica*, 79, 253-302.

Igami, M. and N. Yang (2016). Unobserved Heterogeneity in Dynamic Games: Cannibalization and Preemptive Entry of Hamburger Chains in Canada.*Quantitative Economics*, 7(2), 483-521.

Iida, T. and N. Matsubayashi (2011). Strategic multi-store opening under financial constraint. *European Journal of Operational Research*, 210, 379-389.

Janssen, M., V. Karamychev, and P. Van Reeven (2005). Multi-store competition: market segmentation or interlacing. *Regional Science and Urban Economics*, 35(6), 700-714.

Jia, P. (2006). What happens when Walmart comes to town: An empirical analysis of the discount retailing industry. *Technical Report*, MIT.

Karamychev, V. and P. van Reeven (2009). Retail sprawl and multi-store firms: An analysis of location choice by retail chains. *Regional Science and Urban Economics*, 39, 277-286.

Kuo, R.J., S.C. Chi, and S.S. Kao (2002). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neutral network. *Computers in Industry*, 47, 199-214.

Loertscher, S. and Y. Schneider (2011). Chain Stores, Consumer Mobility, and Market Structure, *Journal of Institutional and Theoretical Economic*, 167(2), 236-246.

Mazzeo, M. (2002). Product choice and oligopoly market structure. *RAND Journal of Economics*, 33, 221-242.

Mejia-Dorantes, L., A. Paez, and J.M. Vassallo (2012). Transportation Infrastructure Impacts on Firm Location: The Effect of a New Metro Line in the Suburbs of Madrid. *Journal of Transport Geography*, 22, 236-250.

Melo, P.C., D.J. Graham, and R.B. Noland (2010). Impact of Transport Infrastructure on Firm Formation: Evidence from Portuguese Municipalities. In Transportation Research Record: *Journal of the Transportation Research Board*, 2163, Transportation Research Board of the National Academies, Washington, D.C., 133-143.

Min, H. (1987). A multiobjective retail service location model for fastfood restaurants. *Omega*, 15(5), 429-441.

Neven, D.J. (1987). Endogenous sequential entry in a spatial model. *International Journal of Industrial Organization*, 5, 419-434.

Nishida, M. (2015). Estimating a Model of Strategic Network Choice: The Convenience-Store Industry in Okinawa. *Marketing Science*, 34 (1), 20-38.

Nwogugu, M. (2006). Site selection in the US retailing industry. *Applied Mathematics and Computation*, 182(2), 15, 1725-1734.

Orhun, A.Y. (2013). Spatial differentiation in the supermarket industry: The role of common information. *Quantitative Marketing and Economics*, 11(1), 3-37.

Pal, D. and J. Sarkar (2002). Spatial competition among multi-store firms. *International Journal of Industrial Organization*, 20, 163-190.

Peng, S.K. and T. Tabuchi (2007). Spatial Competition in Variety and Number of Stores. *Journal of Economics and Management Strategy*, 16(1), 227-250.

Reiss, P.C. (1996). Empirical models of discrete strategic choices. *American Economic Review*, 86, 421-426.

Roux, N. (2014). Le secteur de la restauration commerciale : données économiques, évolution des prix et du nombre de plaintes du secteur. DGCCRF-éco, July 2014, 28.

Schiraldi, P., S. Seiler, and H. Smith (2012). Supermarket Retail Competition with Multi-stop Shopping: Who Competes with a Big Box Store? *Working Paper*.

Schiraldi, P., H. Smith, S. Seiler, and Y. Takahashi (2012). Estimating a Dynamic Game of Spatial Competition: The Case of the U.K. Supermarket Industry. *Working Paper*.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The Rand Journal of Economics*, 37(3), 619-640.

Takaki, M. and N. Matsubayashi (2013). Sequential multi-store location in a duopoly. *Regional Science and Urban Economics*, 43(3), 491-506.

Teitz, M.B. (1968). Location strategies for competitive systems. *Journal of Regional Science*, 8, 135-148.

Thill, J.-C. (1997). Multi-outlet firms, competition and market segmentation strategies. *Regional Science and Urban Economics*, 27, 67-86.

Thomadsen, R. (2005). The effect of ownership structure on prices in geographically differentiated industries. *RAND Journal of Economics*, 36, 908□929.

Thomadsen, R. (2007). Product positioning and competition: The role of location in the fast food industry. *Marketing Science*, 26, 792-804.

Thomas, J.W. (2007). Market Segmentation, Decision Analyst. *Wordpress*.

Toivanen, O. and M. Waterson (2005). Market Structure and Entry: Where's the Beef? *Round Journal of Economics*, 36, 680-699.

Wang, F.H. (2012). Measurement, optimization, and impact of healthcare accessibility: A methodological review. *Annals of the Association of American Geographers*, 192, 1104-1112.

Watson, R. (2009). Product Variety and Competition in the Retail Market for Eyeglasses. *The Journal of Industrial Economics*, 57(2), 217-251.

Wile, R. (2014). The True Story of How McDonald's Conquered France. *Business Insider*.

Yang, N. (2016). Learning in Retail Entry. *NET Institute Working Paper*.

6
Concluding remarks

The level of establishment location analysis in the existing literature has not yet reached the highest methodological standard. In this thesis, we focused on accommodation of spatial aspects and strategic interactions for location choices to provide a more realistic, more explicit, and richer picture of the regional linkages reinforcing local growth and economic development. In addition, we tried to rectify several false assumptions and misconceptions made in the existing literature, such as the use of Euclidean distance to capture interactions between neighboring observations no matter the terrain analyzed, establishments acting in isolation and ignoring others players' decisions, perception of the industry in terms of independent store, or finally a static, nighttime representation of demand in decisions of revenue-oriented sectors for which high customer- or user-contact seems to be critical. We believe that this thesis delivers a number of novel solutions and demonstrates the need to reconsider some critical aspects of location choices.

This thesis has concentrated on the Paris region, which typifies an interesting site for analyzing business location decisions in an intra-metropolitan context, as it is characterized by a sufficient level of spatial and industrial heterogeneity. First, the Paris region is a vibrant and innovative site with over 11.7 million residents and 5.6 million jobs. It is the leading research and development hub in Europe. Yet, its economy is spatially unbalanced, especially regarding economic activity. Some of its markets continually accommodate newly-formed establishments whereas a large group of municipalities is left with no new creation of establishments. Second, the Paris region is an appealing territory to address, where the problems of high congestion, speed limits, or uncrossable physical barriers (such as rivers or industrial corridors, that can diminish or completely eliminate the linkage between neighboring areas) clearly apply. These arguments strongly suggest the need to derive advantages from several disciplines, namely, spatial econometrics, industrial organization, transport economics, along with land use planning or real estate studies to secure more realistic models of establishment locational choices.

Four papers of which this thesis is comprised, sought to open a dialogue with numerous researchers, who voiced in their works some complaints and burning needs to be accomplished. For instance, the first paper (Chapter 2) is the return to Bhat et al. (2014) and Liviano-Solís and Arauzo-Carod (2013), who acknowledge that heretofore scholars have not fully inquired into the specifics of the hurdle model method when analyzing location phenomena. In addition, the first paper responds to the remark made by Dubé et al. (2016) that the existing literature of location choices relies on aggregate geographic entities (*e.g.*, municipalities) with the implicit consequence of limiting patterns to choices occurring only between these administrative areas. Our second paper (Chapter 3) is a reaction to the discontent expressed by Miller and Wentz (2003) that with the advent of powerful computers and "big data" many researchers still overwhelmingly apply the Euclidean metric without realizing its underlying assumptions or its alternatives. The third paper (Chapter 4) takes into consideration the suggestion made in the survey by Draganska et al. (2008) on how a game-theoric framework can aid in construction and estimation of interrelated choice models. In the literature on location choices of establishments, one can find industry-specific regressions, *i.e.*, where only one activity sector, typically an industrial or a retail sector, is considered at a time. In very sparse empirical applications, when locational choice models are developed for several activity sectors, each model is typically run independently. The last paper (Chapter 5) is the answer to Draganska et al. (2008), Fransen et al. (2015), Wang (2012), Nishida (2015), Igami and Yang (2015), Chang and Hsieh (2014), and the founder of Decision Analyst, Jerry W. Thomas (2007), who identify that some unrealistic assumptions about customer behavior are still present in existing models and that usually applied static, nighttime representations of population and income might inadequately capture the demand. In addition, when writing this this we kept in mind a request made by Chatman et al. (2016), to further explore the relationship of establishment and firm births to transportation services.

6.1 Main findings

All research objectives outlined in the introduction have been fulfilled. We inscribe the individual contributions of each chapter in the following.

Chapter 2 (Paper 1) addressed the excess of zeros problem in the location choice model in highly heterogeneous geographic areas. In addition, since the final decision of an establishment seems to be related to the surrounding economic landscape, it proposed a way to accommodate spatial spillovers in location decisions. The approach was twofold. In the first step, we estimated standard non-spatial nested and non-nested count data models. Six techniques of count data models, namely Poisson, ZIP, ZIP tau, NB, ZINB, and hurdle-Poisson model, have been tested. Each model type has been run with variables that represent: access to the workforce, demographic characteristics, physical quality, availability of vacant land, availability of dedicated floor space, accessibility to transport infrastructure (access to public transit networks, road networks), access to other amenities such as universities, services and commerce, pre-existing large establishments and within-industry entities, and prices and tax levels. Particular attention was paid to distinguish outcomes across different activity sectors. In the second step, we ran these count data models accommodating spatial spillovers for location decisions. In this way, we obtained results from 84 models. We compared all the obtained results and chose the best performing hurdle-Poisson model. In all the analyzed cases, the results from the models run with the distance matrix indicated that accommodating spatial spillovers significantly improves the model's performance. We thus suggest accounting for spatial spillovers when modeling establishment location choice. In addition, when the observed data display a higher fraction of zeros than would be typically explained by the standard count data models, the use of hurdle models can be advocated.

Chapter 3 (Paper 2) explored further the hurdle model technique and establishment location decisions in relation to both the characteristics of their immediate economic environment as well as their relative position to surrounding economic landscape. This chapter contributed to the literature on location modeling entreating a discussion on new directions for empirical explorations using more appropriate econometric tools and putting more consideration on the definition of distance itself. Chapter 3 proposed one of the first formulations and applications of spatial count data models of location choice in regional science, wherein various metrics other than Euclidean distance are investigated to account for the linkage between neighboring observations. Rather than locking researchers into a restrictive structure of the weight matrix, this study proposed a flexible approach to intimate which distance metric is more appropriate to correctly account for the nearby markets. Seven metrics have been considered: Euclidean distance, travel times by car (for the peak and off-peak periods) and by public transit, and the corresponding network distances. A probabilistic mixture of two "mono-distance" hurdle-Poisson models has been developed by means of this chapter. Each model's latent class used a different distance representation to incorporate spillover effects in location choices of establishments from several activity sectors.

This chapter suggested that in areas where high congestion or uncrossable physical barriers problems clearly arise, distances purely based on topography may not be the most appropriate for the study of intra-urban location to capture interactions between neighboring observations. For all economic sectors considered, the mixture of hurdle-Poisson models performed markedly better than the "pure" mono-distance models, emphasizing the usefulness of our approach. This corroborated that spatial spillovers are indeed channeled by different means, hence best represented using several metrics. Second, the combination of peak and off-peak road times (slightly) outperformed other combinations including the Euclidean distance. This supported the choice of meaningful over more abstract metrics in spatial econometric models. The distance measure being most likely to capture spatial spillovers varied depending on the economic sector. The peak travel time strongly predominated for establishments from the construction and special, scientific and technical activities sectors, while the off-peak travel time or the Euclidean distance (depending on the pair of metrics used) seemed to prevail for the real estate sector, reflecting differences between sectors in operations and location choice criteria.

The model specification proposed in the third chapter, based on a mixture of spatial count models, can easily be applied to other fields as well as to other forms of spatial dependence including interregional flows based on economic distance, trade, or innovation.

Chapter 4 (Paper 3) had for the main objective to further improve the location choice models by implementing the strategic interactions between the newly-created establishments regarding their location choice relative to their activity sector. It discussed a strategic game with imperfect information, *i.e.*, it employed a Bayesian Nash game, assuming within-group homogeneity and heterogeneity between groups. In the literature on location choices of establishments, one can find industry-specific regressions, *i.e.*, only one activity sector, typically an industrial or a retail sector, is considered at a time. In very sparse empirical applications, when locational choice models are developed for several activity sectors, each model is typically executed independently. The novelty of this paper was to simultaneously run location choice models for a number of establishment types, so that decisions of newly-created establishments on other establishments' births could have been examined.

We found the impact of asymmetries in strategic effects for establishments of different types. The influence of variables that captured market conditions at each location, such as demand and cost factors, local agglomeration measures, or within-industry economies (own-industry competition effects versus Marshallian agglomeration economies), and inter-industry economies (Jacobsian agglomeration economies) were found to be critical in establishments' location decisions, yet quite different depending on the establishment type. Strategic interactions between seven various establishment types have been represented in a form of a matrix. We found strong negative effects among pre-existing establishments of the same type confirming the results of, *e.g.*, Chatman et al. (2016), who suggested that the fact that number of establishments in the same industry category tends to negatively predict the number of establishments in that industry may indicate that the own-industry competition effect overrides any localized Marshallian agglomeration economies. What has not been yet demonstrated in the literature, is that new establishments were likely to enter the market where other newly-created establishments of the same type have just located.

Chapter 5 (Paper 4) aimed to carefully reconsider the list of variables usually used to capture the demand level in a market in the case of services or commerce for which location is a main determinant of revenue and the high customer- or user-contact is indispensable. Stores, such as fast-food restaurants, which are the focus of this chapter, as well as coffee shops, supermarkets, bank branches, do have a direct impact on demand and can be cited as examples.

For the purpose of this chapter, a unique database has been created. Detailed information on five Parisian strongest players in the fast-food industry has been manually drawn (using the SIREN code of each firm) from the SIRENE/INSEE source and its consistency has been compared with the official websites of each fast-food chain, so that some missing addresses were added to our database or some other coordinates modified. In this way a unique database has been created, which has been never used in any other application.

The need to state a clear difference between the daytime and nighttime population has been emphasized. Demand proxies used in the models included flows of people doing shopping, going out to eat in restaurants, or making trips for a leisure purpose; flows of middle- and low-income persons; and other variables which represent the crowd of potential clients passing through the commercial centers, train, express underground and subways stations, airports, highly touristic sites, highways, and the mass of potential clients coming from the nearby high schools, universities, and offices.

Conditional Logit and Tobit models have been applied to model the locational choices of fast-food chains. We acknowledged that the reduce-form model to explain the location strategies of multi-store fast-food firms, proposed in this chapter, was not the most sophisticated solution from the methodological and econometrical point of view and can be treated as a base for the two-stage structural model to be extended in the future. Yet, the outcomes from this chapter demonstrated that the observed locational patterns of analyzed fast-food chains were prone to diverge. Fast-food chains reacted distinctively to market characteristics and market structure making existing papers rethink their assumption on homogenous firms. Contrary to some existing literature, demand, when properly expressed, turned out to be very significant for all chains investigated. In addition, different demand characteristics were used to capture demand depending on the chain being examined. Our results cautioned against the simplistic representation of demand. Our method added behavioral realism to the existing models that are typically based on static, nighttime representations of population.

6.2 Practical recommendations

More realistically designed location choice models accounting for spatial spillovers, strategic interaction, price endogeneity, and with a more appropriate definition of distance and demand can become a powerful and flexible tool. This tool assists in finding a befitting site, which in turn can make an implicative difference for the newly-created business. The contents of this thesis provide some useful recommendations for city planners, transport analysts, real estate agents, plan developers, urban social theorists, for entrepreneurs, business owners, and shopping center investors.

6.3 Future research directions

We asked a number of burning questions through this thesis and attempted to open the door to possible answers or at least to stimulate further reflection on location choice models. We recognize that substantial work remains in dealing with important limitations of this thesis itself. Specifically, we call for the robustness check of results produced by the means of the hurdle, zero-inflated, or Tobit models to account for the excess of zeros problem. A number of papers cited in this thesis, demonstrated that in the similar cases to ours, the distribution of newly-formed establishments is heavily skewed. Examples include the paper of Liviano-Solís and Arauzo-Carod (2014) who studied the Catalan establishment entries and Toivanen and Waterson (2005) who carried fast-food chain data analysis for London. When the terrain being analyzed is highly heterogeneous, few sites host a large number of newly-formed establishments, whereas a large group of sites are left with no new entries.

First, using the data for the Paris region, it has been demonstrated in the first paper of this thesis, that a proportion of municipalities left with no new creation ranges from 34 % up to 61 % depending on the sector being analyzed. Second, the skewness of newly-formed fast-food stores data was acknowledged in the fourth paper. In our application of fast-food data in the case of the Paris region, each outlet of five fast-food chains was choosing one out of 175 alternative markets, these were 80 districts of Paris and 95 EGT zones for the rest of the Paris region. We observed no newly-formed outlets of McDonald's in 103 out of 175 of sites. Moreover, 139 out of 175 alternative markets were left with no new stores of Quick, 133 with no KFC outlets, 158 with no points of Paul, and finally, POM left 134 out of 175 proposed zones with none of its newly-created stores.

We spur the researchers to further investigate how to distinguish the zeros in the data to be the "true" zeros or the "false" zeros. For instance, "true" or "structural" zeros occur if establishments have a possibility, yet, have no incentives to locate in some particular sites of the terrain being studied, whereas the "false" zeros are due to design, survey, or observer errors. Additional tests should be proposed. The number (or proportion) of zeros depends on the length of the observed period, and that the robustness of the results could be tested, for example, by reducing this period.

Other fertile fields yet to be cultivated include: (1) more investigations using continuous or spatial micro-data on individual establishments (see Dubé et al., 2016), (2) vertical in addition to the traditional horizontal location choices (see Liu et al., 2016), (3) further analysis on strategic games and complex questions about dynamics in location strategies of multi-store heterogenous firms, in particular (see various works of Aguirregabiria), (4) and finally incorporation of spatial dependence in discrete choice and count data models (see Liesenfeld et al., 2016[1]). We hope to respond to some of these issues in the near future.

[1]Two papers that propose some solutions to these problems and have not been listed in the references of this thesis are: (1) Liu, C.H., S.S Rosenthal, and W.M. Strange (2016). The Vertical City: Rent Gradients, Spatial Structure, and Agglomeration Economies. *Working Paper* and (2) Liesenfeld, R., J.-F. Richard, and J. Vogler (2016). Likelihood-Based Inference and Prediction in Spatio-Temporal Panel Count Models for Urban Crimes. *Journal of Applied Econometrics*, http://dx.doi.org/10.1002/jae.2534.

Appendix

Code	Activity sector, NAF nomenclature - Rev. 2
Section A	AGRICULTURE, FORESTRY AND FISHING
01	Crop and animal production, hunting and related service activities
01.1	Growing of non-perennial crops
01.2	Growing of perennial crops
01.3	Plant propagation
01.4	Animal production
01.5	Mixed farming
01.6	Support activities to agriculture and post-harvest crop activities
01.7	Hunting, trapping and related service activities
02	Forestry and logging
02.1	Silviculture and other forestry activities
02.2	Logging
02.3	Gathering of wild growing non-wood products
02.4	Support services to forestry
03	Fishing and aquaculture
03.1	Fishing
03.2	Aquaculture
Section B	MINING AND QUARRYING
05	Mining of coal and lignite
05.1	Mining of hard coal
05.2	Mining of lignite
06	Extraction of crude petroleum and natural gas
06.1	Extraction of crude petroleum
06.2	Extraction of natural gas
07	Mining of metal ores

Code	Activity sector, NAF nomenclature - Rev. 2
07.1	Mining of iron ores
07.2	Mining of non-ferrous metal ores
08	Other mining and quarrying
08.1	Quarrying of stone, sand and clay
08.9	Mining and quarrying n.e.c.
09	Mining support service activities
09.1	Support activities for petroleum and natural gas extraction
09.9	Support activities for other mining and quarrying
Section C	MANUFACTURING
10	Manufacture of food products
10.1	Processing and preserving of meat and production of meat products
10.2	Processing and preserving of fish, crustaceans and molluscs
10.3	Processing and preserving of fruit and vegetables
10.4	Manufacture of vegetable and animal oils and fats
10.5	Manufacture of dairy products
10.6	Manufacture of grain mill products, starches and starch products
10.7	Manufacture of bakery and farinaceous products
10.8	Manufacture of other food products
10.9	Manufacture of prepared animal feeds
11	Manufacture of beverages
11.0	Manufacture of beverages
12	Manufacture of tobacco products
12.0	Manufacture of tobacco products
13	Manufacture of textiles
13.1	Preparation and spinning of textile fibers
13.2	Weaving of textiles
13.3	Finishing of textiles
13.9	Manufacture of other textiles
14	Manufacture of wearing apparel
14.1	Manufacture of wearing apparel, except fur apparel

Code	Activity sector, NAF nomenclature - Rev. 2
14.2	Manufacture of articles of fur
14.3	Manufacture of knitted and crocheted apparel
15	Manufacture of leather and related products
15.1	Tanning and dressing of leather; manufacture of luggage, handbags, saddlery and harness; dressing and dyeing of fur
15.2	Manufacture of footwear
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
16.1	Sawmilling and planing of wood
16.2	Manufacture of products of wood, cork, straw and plaiting materials
17	Manufacture of paper and paper products
17.1	Manufacture of pulp, paper and paperboard
17.2	Manufacture of articles of paper and paperboard
18	Printing and reproduction of recorded media
18.1	Printing and service activities related to printing
18.2	Reproduction of recorded media
19	Manufacture of coke and refined petroleum products
19.1	Manufacture of coke oven products
19.2	Manufacture of refined petroleum products
20	Manufacture of chemicals and chemical products
20.1	Manufacture of basic chemicals, fertilizers and nitrogen compounds, plastics and synthetic rubber in primary forms
20.2	Manufacture of pesticides and other agrochemical products
20.3	Manufacture of paints, varnishes and similar coatings, printing ink and mastics
20.4	Manufacture of soap and detergents, cleaning and polishing preparations, perfumes and toilet preparations
20.5	Manufacture of other chemical products
20.6	Manufacture of man-made fibres
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations
21.1	Manufacture of basic pharmaceutical products

Code	Activity sector, NAF nomenclature - Rev. 2
21.2	Manufacture of pharmaceutical preparations
22	Manufacture of rubber and plastic products
22.1	Manufacture of rubber products
22.2	Manufacture of plastics products
23	Manufacture of other non-metallic mineral products
23.1	Manufacture of glass and glass products
23.2	Manufacture of refractory products
23.3	Manufacture of clay building materials
23.4	Manufacture of other porcelain and ceramic products
23.5	Manufacture of cement, lime and plaster
23.6	Manufacture of articles of concrete, cement and plaster
23.7	Cutting, shaping and finishing of stone
23.9	Manufacture of abrasive products and non-metallic mineral products n.e.c.
24	Manufacture of basic metals
24.1	Manufacture of basic iron and steel and of ferro-alloys
24.2	Manufacture of tubes, pipes, hollow profiles and related fittings, of steel
24.3	Manufacture of other products of first processing of steel
24.4	Manufacture of basic precious and other non-ferrous metals
24.5	Casting of metals
25	Manufacture of fabricated metal products, except machinery and equipment
25.1	Manufacture of structural metal products
25.2	Manufacture of tanks, reservoirs and containers of metal
25.3	Manufacture of steam generators, except central heating hot water boilers
25.4	Manufacture of weapons and ammunition
25.5	Forging, pressing, stamping and roll-forming of metal; powder metallurgy
25.6	Treatment and coating of metals; machining
25.7	Manufacture of cutlery, tools and general hardware
25.9	Manufacture of other fabricated metal products
26	Manufacture of computer, electronic and optical products
26.1	Manufacture of electronic components and boards

Code	Activity sector, NAF nomenclature - Rev. 2
26.2	Manufacture of computers and peripheral equipment
26.3	Manufacture of communication equipment
26.4	Manufacture of consumer electronics
26.5	Manufacture of instruments and appliances for measuring, testing and navigation; watches and clocks
26.6	Manufacture of irradiation, electromedical and electrotherapeutic equipment
26.7	Manufacture of optical instruments and photographic equipment
26.8	Manufacture of magnetic and optical media
27	Manufacture of electrical equipment
27.1	Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus
27.2	Manufacture of batteries and accumulators
27.3	Manufacture of wiring and wiring devices
27.4	Manufacture of electric lighting equipment
27.5	Manufacture of domestic appliances
27.9	Manufacture of other electrical equipment
28	Manufacture of machinery and equipment n.e.c.
28.1	Manufacture of general-purpose machinery
28.2	Manufacture of other general-purpose machinery
28.3	Manufacture of agricultural and forestry machinery
28.4	Manufacture of metal forming machinery and machine tools
28.9	Manufacture of other special-purpose machinery
29	Manufacture of motor vehicles, trailers and semi-trailers
29.1	Manufacture of motor vehicles
29.2	Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semi-trailers
29.3	Manufacture of parts and accessories for motor vehicles
30	Manufacture of other transport equipment
30.1	Building of ships and boats
30.2	Manufacture of railway locomotives and rolling stock

Code	Activity sector, NAF nomenclature - Rev. 2
30.3	Manufacture of air and spacecraft and related machinery
30.4	Manufacture of military fighting vehicles
30.9	Manufacture of transport equipment n.e.c.
31	Manufacture of furniture
31.0	Manufacture of furniture
32	Other manufacturing
32.1	Manufacture of jewelry, bijouterie and related articles
32.2	Manufacture of musical instruments
32.3	Manufacture of sports goods
32.4	Manufacture of games and toys
32.5	Manufacture of medical and dental instruments and supplies
32.9	Manufacturing n.e.c.
33	Repair and installation of machinery and equipment
33.1	Repair of fabricated metal products, machinery and equipment
33.2	Installation of industrial machinery and equipment
Section D	ELECTRICITY, GAS, STEAM AND AIR CONDITIONING SUPPLY
35	Electricity, gas, steam and air conditioning supply
35.1	Electric power generation, transmission and distribution
35.2	Manufacture of gas; distribution of gaseous fuels through mains
35.3	Steam and air conditioning supply
Section E	WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES
36	Water collection, treatment and supply
36.0	Water collection, treatment and supply
37	Sewerage
37.0	Sewerage
38	Waste collection, treatment and disposal activities; materials recovery
38.1	Waste collection
38.2	Waste treatment and disposal
38.3	Materials recovery

Code	Activity sector, NAF nomenclature - Rev. 2
39	Remediation activities and other waste management services
39.0	Remediation activities and other waste management services
Section F	CONSTRUCTION
41	Construction of buildings
41.1	Development of building projects
41.2	Construction of residential and non-residential buildings
42	Civil engineering
42.1	Construction of roads and railways
42.2	Construction of utility projects
43	Specialised construction activities
43.1	Demolition and site preparation
43.2	Electrical, plumbing and other construction installation activities
43.3	Building completion and finishing
43.9	Other specialised construction activities
Section G	WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES
45	Wholesale and retail trade and repair of motor vehicles and motorcycles
45.1	Sale of motor vehicles
45.2	Maintenance and repair of motor vehicles
45.3	Sale of motor vehicle parts and accessories
45.4	Sale, maintenance and repair of motorcycles and related parts and accessories
46	Wholesale trade, except of motor vehicles and motorcycles
46.1	Wholesale on a fee or contract basis
46.2	Wholesale of agricultural raw materials and live animals
46.3	Wholesale of food, beverages and tobacco
46.4	Wholesale of household goods
46.5	Wholesale of information and communication equipment
46.6	Wholesale of other machinery, equipment and supplies
46.7	Other specialised wholesale
46.9	Non-specialised wholesale trade

Code	Activity sector, NAF nomenclature - Rev. 2
47	Retail trade, except of motor vehicles and motorcycles
47.1	Retail sale in non-specialised stores
47.2	Retail sale of food, beverages and tobacco in specialised stores
47.3	Retail sale of automotive fuel in specialised stores
47.4	Retail sale of information and communication equipment in specialised stores
47.5	Retail sale of other household equipment in specialised stores
47.6	Retail sale of cultural and recreation goods in specialised stores
47.7	Retail sale of other goods in specialised stores
47.8	Retail sale via stalls and markets
47.9	Retail trade not in stores, stalls or markets
Section H	TRANSPORTATION AND STORAGE
49	Land transport and transport via pipelines
49.1	Passenger rail transport, interurban
49.2	Freight rail transport
49.3	Other passenger land transport
49.4	Freight transport by road and removal services
49.5	Transport via pipeline
50	Water transport
50.1	Sea and coastal passenger water transport
50.2	Sea and coastal freight water transport
50.3	Inland passenger water transport
50.4	Inland freight water transport
51	Air transport
51.1	Passenger air transport
51.2	Freight air transport and space transport
52	Warehousing and support activities for transportation
52.1	Warehousing and storage
52.2	Support activities for transportation
53	Postal and courier activities
53.1	Postal activities under universal service obligation

Code	Activity sector, NAF nomenclature - Rev. 2
53.2	Other postal and courier activities
Section I	ACCOMMODATION AND FOOD SERVICE ACTIVITIES
55	Accommodation
55.1	Hotels and similar accommodation
55.2	Holiday and other short-stay accommodation
55.3	Camping grounds, recreational vehicle parks and trailer parks
56	Food and beverage service activities
56.1	Restaurants and mobile food service activities
56.2	Event catering and other food service activities
56.3	Beverage serving activities
Section J	INFORMATION AND COMMUNICATION
58	Publishing activities
58.1	Publishing of books, periodicals and other publishing activities
58.2	Software publishing
59	Motion picture, video and television programme production, sound recording and music publishing activities
59.1	Motion picture, video and television programme activities
59.2	Sound recording and music publishing activities
60	Programming and broadcasting activities
60.1	Radio broadcasting
60.2	Television programming and broadcasting activities
61	Telecommunications
61.1	Wired telecommunications activities
61.2	Wireless telecommunications activities
61.3	Satellite telecommunications activities
61.9	Other telecommunications activities
62	Computer programming, consultancy and related activities
62.0	Computer programming, consultancy and related activities
63	Information service activities
63.1	Data processing, hosting and related activities; web portals

Code	Activity sector, NAF nomenclature - Rev. 2
63.9	Other information service activities
Section K	FINANCIAL AND INSURANCE ACTIVITIES
64	Financial service activities, except insurance and pension funding
64.1	Monetary intermediation
64.2	Activities of holding companies
64.3	Trusts, funds and similar financial entities
64.9	Other financial service activities, except insurance and pension funding
65	Insurance, reinsurance and pension funding, except compulsory social security
65.1	Insurance
65.2	Reinsurance
65.3	Pension funding
66	Activities auxiliary to financial services and insurance activities
66.1	Activities auxiliary to financial services, except insurance and pension funding
66.2	Activities auxiliary to insurance and pension funding
66.3	Fund management activities
Section L	REAL ESTATE ACTIVITIES
68	Real estate activities
68.1	Buying and selling of own real estate
68.2	Renting and operating of own or leased real estate
68.3	Real estate activities on a fee or contract basis
Section M	PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES
69	Legal and accounting activities
69.1	Legal activities
69.2	Accounting, bookkeeping and auditing activities; tax consultancy
70	Activities of head offices; management consultancy activities
70.1	Activities of head offices
70.2	Management consultancy activities
71	Architectural and engineering activities; technical testing and analysis
71.1	Architectural and engineering activities and related technical consultancy

Code	Activity sector, NAF nomenclature - Rev. 2
71.2	Technical testing and analysis
72	Scientific research and development
72.1	Research and experimental development on natural sciences and engineering
72.2	Research and experimental development on social sciences and humanities
73	Advertising and market research
73.1	Advertising
73.2	Market research and public opinion polling
74	Other professional, scientific and technical activities
74.1	Specialised design activities
74.2	Photographic activities
74.3	Translation and interpretation activities
74.9	Other professional, scientific and technical activities n.e.c.
75	Veterinary activities
75.0	Veterinary activities
Section N	ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES
77	Rental and leasing activities
77.1	Renting and leasing of motor vehicles
77.2	Renting and leasing of personal and household goods
77.3	Renting and leasing of other machinery, equipment and tangible goods
77.4	Leasing of intellectual property and similar products, except copyrighted works
78	Employment activities
78.1	Activities of employment placement agencies
78.2	Temporary employment agency activities
78.3	Other human resources provision
79	Travel agency, tour operator and other reservation service and related activities
79.1	Travel agency and tour operator activities
79.9	Other reservation service and related activities
80	Security and investigation activities
80.1	Private security activities
80.2	Security systems service activities

Code	Activity sector, NAF nomenclature - Rev. 2
80.3	Investigation activities
81	Services to buildings and landscape activities
81.1	Combined facilities support activities
81.2	Cleaning activities
81.3	Landscape service activities
82	Office administrative, office support and other business support activities
82.1	Office administrative and support activities
82.2	Activities of call centres
82.3	Organization of conventions and trade shows
82.9	Business support service activities n.e.c.
Section O	PUBLIC ADMINISTRATION AND DEFENCE; COMPULSORY SOCIAL SECURITY
84	Public administration and defence; compulsory social security
84.1	Administration of the State and the economic and social policy of the community
84.2	Provision of services to the community as a whole
84.3	Compulsory social security activities
Section P	EDUCATION
85	Education
85.1	Pre-primary education
85.2	Primary education
85.3	Secondary education
85.4	Higher education
85.5	Other education
85.6	Educational support activities
Section Q	HUMAN HEALTH AND SOCIAL WORK ACTIVITIES
86	Human health activities
86.1	Hospital activities
86.2	Medical and dental practice activities
86.9	Other human health activities
87	Residential care activities

Code	Activity sector, NAF nomenclature - Rev. 2
87.1	Residential nursing care activities
87.2	Residential care activities for mental retardation, mental health and substance abuse
87.3	Residential care activities for the elderly and disabled
87.9	Other residential care activities
88	Social work activities without accommodation
88.1	Social work activities without accommodation for the elderly and disabled
88.9	Other social work activities without accommodation
Section R	ARTS, ENTERTAINMENT AND RECREATION
90	Creative, arts and entertainment activities
90.0	Creative, arts and entertainment activities
91	Libraries, archives, museums and other cultural activities
91.0	Libraries, archives, museums and other cultural activities
92	Gambling and betting activities
92.0	Gambling and betting activities
93	Sports activities and amusement and recreation activities
93.1	Sports activities
93.2	Amusement and recreation activities
Section S	OTHER SERVICE ACTIVITIES
94	Activities of membership organizations
94.1	Activities of business, employers and professional membership organizations
94.2	Activities of trade unions
94.9	Activities of other membership organizations
95	Repair of computers and personal and household goods
95.1	Repair of computers and communication equipment
95.2	Repair of personal and household goods
96	Other personal service activities
96.0	Other personal service activities
Section T	ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS; UNDIFFERENTIATED GOODS- AND SERVICES-PRODUCING ACTIVITIES OF HOUSEHOLDS FOR OWN USE

Code	Activity sector, NAF nomenclature - Rev. 2
97	Activities of households as employers of domestic personnel
97.0	Activities of households as employers of domestic personnel
98	Undifferentiated goods- and services-producing activities of private households for own use
98.1	Undifferentiated goods-producing activities of private households for own use
98.2	Undifferentiated service-producing activities of private households for own use
Section U	ACTIVITIES OF EXTRATERRITORIAL ORGANIZATIONS AND BODIES
99	Activities of extraterritorial organizations and bodies
99.0	Activities of extraterritorial organizations and bodies

Table 6.1: Explanation of French classification of activities (NAF nomenclature - Rev. 2)

Modeling location problems requires an understanding of the real-world operations that are to be reflected in the model. Models need not to reflect every aspect of the real-world operations. In fact, parsimonious models are generally better than complex inscrutable models. The ability to know what must be incorporated into a model and what can safely be treated as exogenous is both an ART and a SCIENCE. [...] The purpose of modeling is to identify the tradeoffs between the objectives while capturing as much of the richness of the real-world problem as is necessary to ensure the credibility of the modeler and model itself.

- M.S. Daskin, "Network and Discrete Location: Models, Algorithms, and Applications"

A LaTeX template that can be used to format a dissertation with this look and feel has been released under the permissive AGPL license, and can be found online at github.com/asm-products/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.