Published in *Networks and Spatial Economics*, *2019*, vol. 19, no. 4, pp. 1215-1248, which should be cited to refer to this work.
DOI: 10.1007/s11067-018-9439-5

Euclidean distance versus travel time in business location: A probabilistic mixture of hurdle-Poisson models

Sabina Buczkowska [*] Nicolas Coulombel [†]
Matthieu de Lapparent [‡]

31 October 2014

Report TRANSP-OR 141031
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*]Université Paris-Est, DEST, French Institute of Science and Technology for Transport, Development and Networks (Ifsttar), 14-20 boulevard Newton, Cité Descartes, 77447 Marne la Vallée Cedex 2, France. sabina.buczkowska@ifsttar.fr (corresponding author)

[†]Université Paris-Est, LVMT, Ecole des Ponts ParisTech, 6-8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France. nicolas.coulombel@enpc.fr

[‡]Transp-OR, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. matthieu.delapparent@epfl.ch

Abstract

While the question of the specification of spatial weight matrix is now largely discussed in the spatial econometrics literature, the definition of distance has attracted less attention. The choice of the distance measure is often glossed over, with the ultimate use of the Euclidean distance. This paper investigates this issue in the case of establishments locating in the Paris region. Indeed, numerous works highlight the importance of transport infrastructure in the location model, which challenges the choice of the Euclidean distance in representing spatial effects. To compare the various distance measures, we develop a probabilistic mixture of hurdle-Poisson models for several activity sectors. Each model class uses a different definition of distance to capture spatial spillovers. The following distance measures are considered: Euclidean distance, two road distances (with and without congestion), public transit distance, and the corresponding travel times. Overall, the obtained results are in line with the literature regarding the main determinants of establishments' location. However, we find that for some activity sectors, such as construction, the peak road travel time for private vehicles is the most likely to correctly capture spatial spillovers, whereas for other sectors, such as real estate, the Euclidean distance slightly prevails. This tends to show that spatial spillovers are channeled by different means, depending on the activity sector. In addition, we find that the proposed mixture of hurdle-Poisson models that uses several latent classes performs significantly better than the "pure" hurdle-Poisson models based on a single distance measure, emphasizing the usefulness of our approach.

1 Introduction

The role played by a spatial weight matrix has long been a controversial aspect of spatial methods (Partridge *et al.*, 2012, LeSage and Pace, 2012, Vega and Elhorst, 2013). Numerous studies have attempted to determine which specification of the spatial weight matrix (W) best fits the data and to investigate the robustness of their results to different W specifications (e.g., Bell and Bockstell, 2000, Kostov, 2010). Investigations cover the definition of neighbors (rook or queen matrix, n nearest neighbors, etc.), the specification of the distance decay function, or the bandwith size[1]. Several studies reported that the weight matrix does play a role in spatial models and that two different choices of W may lead to significantly different estimates. Yet, LeSage and Pace (2012) found little evidence that estimates are sensitive to minor changes in specifications used for the spatial weight structure in these models if 1) estimates are correctly interpreted and rely on true partial derivatives, and 2) the model is well-specified. Changes in the spatial weight matrix specification may entail changes in measures of dispersion (e.g., t-Student statistics), but not significant differences in the coefficient estimates[2]. This result is as critical as sensitivity of estimates could be a good reason to consider spatial models as ill-conditioned. Overreaction to small changes in the weight matrix would therefore suggest a misspecification of the model.

Yet, a definition of distance has been the subject of less attention. When

[1]See Getis and Altstadt (2004) who summarize the typical well-known schemes that reaserchers follow to find a proper spatial dependence representation in the 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 n-th nearest neighbor distance, 5) ranked distances, 6) constrained weights for an observation equal to some constant, 7) all centroids within distance d, 8) n nearest neighbors, and so on. Some of the newer schemes are: 1) bandwidth distance decay (Fotheringham et al., 1996), 2) Gaussian distance decline (LeSage, 2003), and 3) "tri-cube" distance decline function (McMillen and McDonald, 2004).

[2]See Pace and LeSage (2008) who used this idea to develop a Hausman specification test for significant differences between OLS and SEM estimates. The authors state that under the null of correct specification, OLS and spatial error model estimates should be similar.

the spatial weight matrix is based on distance, the choice of the distance measure is often glossed over, with an ultimate preference for the Euclidean distance (e.g., Bhat et al., 2014). As noticed by McMillen and McDonald (2004) and emphasised by Billé and Arbia (2013), the use of a spatial contiguity matrix is often the starting point to specify the linkage between neighboring observations. Yet, it has the disadvantage of imposing a restrictive structure that can bias results when inappropriate, hence the importance of choosing the fitting distance measure. Corrado and Fingleton (2012) argued that the specification of W, including the choice of distance measure, should be supported by an economic theory. Thus, the Euclidean distance might not always be the most relevant one depending on the problem considered[3]. 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. Following this train of thought, several studies have considered alternative distance measures that are not purely based on topography (e.g., Conley and Ligon, 2002, Slade, 2005), including network distances and transport costs. However, there is little comparison with the geographical distance (Euclidean or great circle depending on the spatial scale), and when there is, it is based on the relative performances of two models, one based on the alternative distance and the other on the geographical distance.

This research proposes a new, flexible approach, where several distance measures may coexist and be combined instead of being systematically opposed. The methodology is based on a mixture of "mono-distance models", as described in the text below. This allows us to capture the diversity of agents' behavior, and provides a more direct and integrated way of comparing various distance measures with each other. We also aim to address the criticism of Vega and Elhorst (2013) that the choice of the spatial weight matrix is usually quite arbitrary, while it refers to the choice of the distance measure.

[3]Fingleton and Le Gallo (2008) stated for instance that "the spillover between areas are not simply a function of spatial propinquity, to the exclusion of other effects" and "it is more realistic to base it on relative economic distance."

The methodology is applied to the location choice of newly created establishments in the Paris region. Recent works have emphasized the importance of spatial effects in this context (Bhat et al., 2014, Buczkowska and Lapparent, 2014, Liviano-Solís and Arauzo-Carod, 2013, Liesenfeld et al., 2013, Lambert et al., 2010, Klier and McMillen, 2008). Yet, whenever the distance measure was used in the weight matrix to implement the spatial effects or spatial spillovers in location choice models, no discussion was provided on the choice of the distance measure itself and the Euclidean distance was utilized. As a large body of literature highlights the importance of the transport infrastructure in the location choice of establishments (reviewed by Arauzo-Carod et al., 2010, among others), this challenges the choice of the Euclidean distance to represent spatial effects. Distance measures based on the transport network might be more appropriate, as advocated by Combes and Lafourcade (2003, 2005), who stated that the Euclidean distance is only a proxy for the true physical distance. In reality, people or goods move along transport networks rarely going from point A to point B in a straight line. Congestion or speed limits may also cause drivers to make detours in order to reduce their travel time, which means that the fastest path may not be the shortest one[4].

On the basis of former work by Buczkowska and Lapparent (2014), we extend their model by estimating a mixture of hurdle-Poisson models whereby two latent classes are used. Each class uses a different definition of distance to capture spatial spillovers. To our knowledge, this is the first formulation and application of spatial count data models of the choice of location, wherein various other than Euclidean distance measures are investigated to build the spatial distance weight matrices for different activity sectors. Besides the Euclidean distance, the proposed distance measures are two road distances (with or without congestion), the public transit distance, and the corresponding travel times. As noted above, the mixture

[4]As stressed by Nguyen et al. (2012), many distance-based weighting functions have been proposed be used in the weight matrix. It is always assumed that the inter-centroid distance from site i to site j is the same as the distance from site j to site i (see also Miaou and Sui, 2004), which may not be the case in reality.

allows several distance measures to coexist within a same model [5]. We contribute to the existing literature on location modeling, opening up a discussion and a new direction for empirical explorations using appropriate econometric tools and putting more consideration on the definition of distance. Overall, we find results that are in line with the literature regarding the main determinants of firm location. However, we find that for some activity sectors, such as construction, the peak road travel time for private vehicles is the most likely to correctly capture spatial spillovers, whereas for other sectors, such as real estate, the Euclidean distance slightly prevails. This tends to show that spatial spillovers are channeled by different means depending on the activity sector. Moreover, we find that the proposed mixture of hurdle-Poisson models that uses several latent classes performs significantly better than "pure" hurdle-Poisson models based on a single distance measure, emphasizing the usefulness of our approach.

The present paper is organized as follows. In Section 2, we review the literature relevant to our topic. Next, we describe the data in Section 3, and develop our parametric statistical model in Section 4. In Section 5, we present and discuss the results of our research. In the final section, we conclude and point out to a possible extension of the proposed approach.

2 Literature review

The analysis of firm location choices has attracted considerable attention in the past decades. In a recent survey, Arauzo-Carod et al. (2010) reviewed over fifty papers on location choice modeling with a focus on location decisions of new industrial establishments or firms. They described the establishment/firm location determinants, the econometric methods used in

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

these investigations, and their main findings[6].

However, only recently has the importance of spatial effects in this context has been emphasized (Bhat et al., 2014, Buczkowska and Lapparent, 2014, Liviano-Solís and Arauzo-Carod, 2013, Liesenfeld et al., 2013, Lambert et al., 2010, Klier and McMillen, 2008). As shown by Nguyen et al. (2012), an establishment does not act in isolation during its decision-making processes and is likely to be influenced by other establishments located nearby. When choosing an appropriate place in which to set up on 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 that are close to one another. Jayet (2001) proved the existence of interactions among units located in space and demonstrated that their intensity decreases with distance. Despite the existence of these spatial effects, they are most often completely ignored in the analysis of the unit location. There is 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-

[6]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 (+), transport infrastructure (+), road infrastructure (+), distance to highway (-), rail infrastructure (-), airports facilities (+), travel time to airport (-), energy costs (+,-), and environmental regulation (-).

Solís and Arauzo-Carod, 2013, Liesenfeld et al., 2013, Lambert et al., 2010, Klier and McMillen, 2008).

Klier and McMillen (2008) proposed a model with a spatially weighted dependent variable to analyze location decisions of auto supplier plants in the US (discrete choice framework). They accounted 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) developed the Spatial Autoregressive Poisson model and assessed 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. (2013) proposed an 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. Efficient importance sampling is a high-dimension MC integration technique based on simple LS approximations used to maximize the numerical accuracy of MC likelihood estimation. The accuracy of EIS likelihood evaluation is computationally feasible even for large sample sizes, such as 5000 and more.

Bhat et al. (2014) formulated a spatial multivariate model to predict the count of new businesses at a county level in the state of Texas considering the business location decisions by the industrial sector. It allows for a better recognizion of the industry specific determinants. The authors accommodated overdispersion and excess zero problems. They accounted for the unobserved factors that simultaneously affect the county-level count of new businesses in different sectors and spatial dependence effects across counties.

However, whenever a distance measure was used in the weight matrix to implement the spatial effects or spatial spillovers in the location choice

model, no discussion was provided on the choice of the distance measure itself. Bhat et al. (2014) tested different specifications of the weight matrix in spatial models, including inverse distance, inverse of the square of the distance, and inverse of the cube of the distance between counties. Yet, they did not concentrate on the distance definition. Lambert et al. (2010) proposed, among others, a row-standardized inverse distance matrix based on the Euclidian distance between the nearest neighbors. Liviano-Solís and Arauzo-Carod (2013) considered a distance matrix such that $w_{ij} = 1/d_{ij}$, where d_{ij} is the Euclidean distance between the municipalities i and j. In the paper by Buczkowska and Lapparent (2014), spatial spillovers were modeled as: $\mathbf{x_{i,s}} = \ln\left(\sum_{j=1}^{I} \mathbf{e}^{-\mu d_{i,j}} \mathbf{z_{j,s}}\right)$, where $\mathbf{z_{j,s}}$ is an attribute of the municipality that applies to activity sector s or the number of pre-existing establishments from this sector, $d_{i,j}$ - the Euclidean distance between the centroids of municipalities i and j.

For all these reasons, we find it necessary to open up a discussion on the distance definition used in the location choice models.

3 Data

In this section, we describe the possible distance measures that can be used in the models, matrices computation, and statistical sources of other data used in the models.

3.1 Distance measures

In mathematics, computer science and the graph theory, a distance matrix is a two-dimensional array containing distances, taken pairwise, between a set of N points. This matrix has a size of NxN. The Euclidean distance is the "ordinary" distance between two points that one would measure with a ruler (Dattorro, 2005). Euclidean distance matrices have five properties: 1) nonnegativity, 2) self-distance, 3) symmetry, 4) triangle inequality, and 5) relative-angle inequality.

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 in economics; it is easily available to boot. Combes and Lafourcade (2003, 2005) claimed that any Euclidean distance can only be regarded as a proxy for the actual physical distance, though. 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 a 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) move along a transport network; they rarely go from point A to point B in a straight line. For instance, car users may only drive on the existing road network, hence the well-known example of the Manhattan distance. Congestion or speed limits may also cause drivers to make detours in order to reduce their travel time, meaning that the shortest path may not be the fastest one. As noticed by Combes and Lafourcade (2005), researchers could get inspired by work of geographers or transport planners who have developed more accurate measures such as distances and travel times matrices derived from Geographic Information Systems. Yet, those mainly focus on specific transport planning purposes.

Based on these considerations, several authors advocated the use of "real" distance measures based on a transport network over geographical distance measures, Euclidean and great circle alike (Combes and Lafourcade, 2005, Graham, 2007, Duran-Fernandez, 2008, Weisbrod, 2008, Faber, 2014). This point is especially cogent when it comes to the location choice of economic establishments, for which the role of a transport infrastructure is now well-known (Arauzo-Carod et al., 2010). As detailed in Section 4, the modeling framework by Buczkowska and Lapparent (2014) has therefore been modified in order to consider alternative "transport distances" in addition to the Euclidean distance. Those are namely: two road distances (with or without congestion), the public transit distance, and the corresponding travel times. Given the size of the Paris region (12,000 km2), the great circle distance is close to the Euclidean distance within our study area and is therefore not included in our analysis.

In practice, several studies have compared whether and to what ex-

tent crow-fly measures (Euclidean and/or great circle) differ from "real" distance measures based on transport networks. For instance, Chalasani et al. (2005) looked 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. In the same line, Rietveld et al. (1999) studied the relationship between travel time and travel distance for car commuters in the Netherlands. They examined the following distance measures: 1) distance as the crow flies between the centroids of the zone of origin and the zone of destination of a trip; 2) shortest travel time by car between these same points, computed with a route planner on the basis of travel time minimization, as well as the corresponding trip length; and 3) the actual travel time reported by the respondent. In France, Combes and Lafourcade (2005) compared the great circle distance, the real distance and the real time based on the real transport network, 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)[7]. This point will be carefully considered and will ultimately lead us to restrict the set of distance measures used in our analysis.

3.1.1 Computation of matrices

As indicated above, this research compares the standard Euclidean distance matrix with transport distance matrices. All matrices are of size 1 690 000 (1300 by 1300), since we measure the distances between all the 1300 municipalities of the Paris region. Euclidean distances were computed in Quantum GIS based on the latitude and longitude coordinates of the centroids of the municipalities.

Transport matrices include the network distance and travel time matrices for the road network and the public transit network. These matrices

[7]Combes and Lafourcade (2005) developed a methodology to compute transport costs based on the transport network, encompassing the characteristics of the infrastructure, vehicle and energy used, labor, insurance, tax and general charges borne by transport carriers. The level of correlation falls when considering time series, emphasizing changes in travel conditions over time.

are computed by means of two assignment models, one for each transport mode. Assignment models, which are also sometimes called network models, simulate the route choice behavior of individuals on a transport network. In road models, congestion plays a major role. As more individuals 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 the 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 (in time and money), which depend on the characteristics of the services that are used (frequency, speed, etc.)[8]. Assignment models are primarily applied to determine the usage of road infrastructure or of transit routes for a given time period (typically during the morning or evening peak periods), but they can also serve to derive the shortest path between any O-D pair, and the corresponding 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 - which is the one most frequently used, included here - also depends on the network characteristics, and firstly on 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 Ile-de-France (DRIEA Ile-de-France, 2008), which were adapted to run with the TransCAD software. Due to data availability issues, the two original models were calibrated for different years, 2008 for the road model and 2009 for the transit model[9]. The

[8]For some transit lines, congestion (Lapparent and Koning, 2014) and/or service unreliability (Benezech and Coulombel, 2013) may also be an important component of the generalized cost of travel. It is seldomly considered in standard transit assignment models, however, as introducing either of these items drastically increases the model complexity.

[9]Transport matrices being relatively stable over time at a regional scale, especially in the Ile-de-France where the transport networks are already well developed, this one year difference should have a very limited impact on our results.

road network, which comprises 65,692 links, includes all the main roads in the Paris region. It is strongly radial, yet with three concentric bypasses (Figure 1). The public transit model includes 62,102 links and similarly all the main transit lines of the Paris region. The public transit network is even more radial than the road network, as the vast majority of the heavy transit lines passes through Paris.

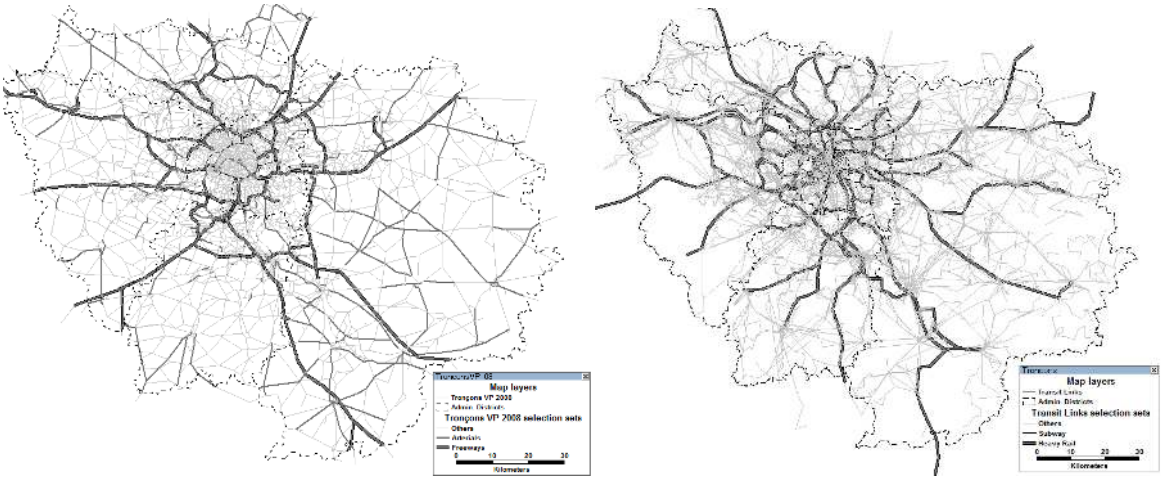

Figure 1: Road network (left figure) and public transit network (right figure) in the Paris region

For each transport mode, we derive two matrices: the shortest travel time matrix, i.e., the minimum travel time between each O-D pair, and the associated network distance matrix. Compared to the Euclidean distance, shortest travel time matrices only have three properties: 1) nonnegativity, 2) self-distance, and 3) triangle inequality[10]. They 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 one-way roads, assymetric congestion patterns, or different service frequencies based on the line direction in case of transit, and so on. Because the fastest path is not necessarily the shortest path (cf. example in Figure 2), the network distance that we compute is greater than the

[10]In other words, the shortest travel time is not a proper distance based on the mathematical definition of distance, but only a "quasi-metrics".

shortest path distance. For the same reason, the triangle inequality may be violated, and our network distance matrices only satisfy 1) nonnegativity and 2) self-distance.

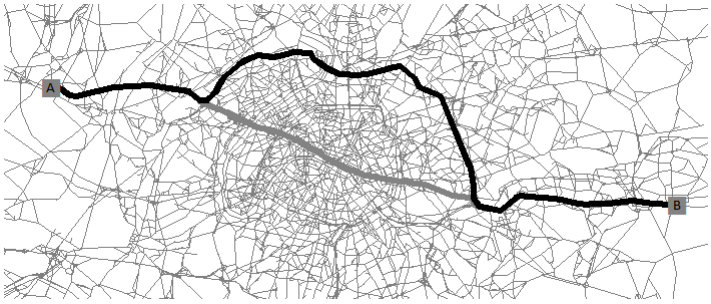

Figure 2: Example of shortest versus fastest path from point A to B

The transport matrices are computed 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 Ile-de-France, 2008). In the case of private vehicles, 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 shipping are concentrated during this period.

3.1.2 Comparison of various distance measures

As pointed out above in this section, several studies find a strong correlation between geographical and transport distance measures. Our data lead to similar conclusions (Table 3.1.2). The Euclidean distance (ED) is very 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). These values are in line with the literature. Combes and Lafourcade (2005) found a correlation coefficient

of 0.990 between great circle distance and road network distance, and Rietveld et al. (1999) of 0.966 between Euclidean distance and road network distance. However, we find that the level of correlation falls markedly with distance. For municipalities which are distant by less than 10 km from one another according to the ED, the correlation between ED and DistVhMph and between ED and DistVhOph are equal to 0.868 and 0.869, respectively (Table 2). Considering the range 40-50 km, the same values drop to 0.541 and 0.586. The mean detour factor, which measures the ratio between the network distance and the Euclidean distance, is equal to 1.287 when there is no congestion. It is slightly higher for the morning peak hours (1.294), reflecting the fact that individuals make additional detours to avoid congestion. Things are quite different when considering 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 less dense, especially in the peri-urban and in the rural area, than the road network, but also more radial (hence a higher detour factor).

The comparison of ED with travel times leads to the same observations. For the road network, we find the levels of correlation between the ED and travel times equal to 0.952 for the morning peak period (TtVhMph) and 0.974 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 as low as 0.452. Last, the correlation levels are again significantly lower when computed by increasing distance interval, and sharply decrease with distance.

All in all, we find that road distance measures (length and travel time) are strongly correlated with the ED at first glance, but less correlated when disaggregating the O-D pairs by distance interval. While the corresponding values are not reported here for the sake of concision, we also find that road network distances (with and without congestion) are strongly correlated with each other, even for a given distance interval, and that they are also relatively strongly correlated with their associated travel time measure. On the contrary, congested and free-flow travel times are less correlated, especially when considering distance intervals. Road network distances were therefore discarded in the subsequent analysis. Similarly, public transport

measures, distance and travel time alike, were tested but yielded poor results, and were thus also discarded[11].

[11]Two facts might account for the poor performance of the public transit measures. 1) Establishments/firms might focus on road travel conditions because it is the predominant transport mode for freight or for intrametropolitan business trips. 2) The strong spatial irregularities of public transport measures, in particular in the most distant parts of the metropolitan area, may make them unsuitable to model spatial spillovers.

Table 1: Correlations between Euclidean distance, crow flies (free-flows), travel times during morning peak hours

Basic statistics	ED[1]	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 measure	ED	DistVhMph	DistVhOph	DistTcMph	TtVhMph	TtVhOph	TtTcMph
ED	1						
DistVhMph	0.987	1					
DistVhOph	0.988	0.992	1				
DistTcMph	0.635	0.620	0.631	1			
TtVhMph	0.952	0.962	0.956	0.558	1		
TtVhOph	0.974	0.975	0.978	0.656	0.961	1	
TtTcMph	0.452	0.435	0.446	0.949	0.382	0.483	1

[1]Euclidean distance (ED); Road distance with congestion (DistVhMph) and without congestion - free-flows (DistVhOph), average distance travelled by public transport during the morning peak period (DistTcMph), and the corresponding travel times (TtVhMph, TtVhOph, TtTcMph, respectively).

Table 2: Correlations between Euclidean distance and transport distances by increasing range (in km)

Distance measure	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
DistVhOph	0.868	0.785	0.696	0.633	0.586	0.555	0.517
DistTcMph	0.355	0.323	0.311	0.283	0.244	0.207	0.193
TtVhMph	0.701	0.488	0.381	0.320	0.290	0.278	0.279
TtVhOph	0.763	0.634	0.541	0.469	0.428	0.401	0.377
TtTcMph	0.354	0.215	0.179	0.146	0.117	0.101	0.088

3.2 Statistical sources of other data

Many 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. Data on the stock of establishments are given for the 1st of January 2007. In our sample, 763 131 pre-existing establishments were registered on the market until the 1st of January 2007. The number of newly created establishments in 2007 equals to 87 974. Data are pooled across activity sectors. In the current paper, we select and analyze three sectors: construction (Constr), special, scientific, technical activities (SpecSci), and real estate (RealEst). 13.8% of all the newly created establishments in the year 2007 belong to the construction sector (12 115), further 15 282 new units to the special, scientific and technical activities sector (17.%), and 4 683 (5.3%) to the real estate sector.

Detailed description of other data used in the models that describe, among others, the structure of population and employment, the proximity to retail, services, universities and schools, public transport and highways, and the levels of prices and taxes, with their sources can be found in the paper by Buczkowska and Lapparent (2014). We limit their presentation to the summary table (Table 3).

Table 3: Description of potential explanatory variables

Variable and its expected sign	Description
Establishments from respective sector (+)[1]	Number of pre-existing establishments from the analyzed sector within a particular municipality divided by the surface of municipality (km2)[2]
Large establishments from all sectors (-)	Number of large pre-existing establishments with fifty employees or more divided by the surface of municipality (km2)
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
Trips home-work, intelectual workers (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 made by white-collar workers
Offices (+)	Fraction of a municipality's surface dedicated to offices
Shops (+)	Fraction of a municipality's surface dedicated to shops
Vacant land (+)	Fraction of a municipality's vacant land available for new investments
Residential area (+)	Fraction of a municipality's land dedicated to the residential area
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
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)

[1](+) 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.

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

4 Econometric model: discrete mixture of hurdle-Poisson models

Below our statistical formulation is described in detail.

4.1 Motivation

The Paris region is highly heterogeneous[12], 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 aggregate at the municipality level data for the Paris region, 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 real estate activities 738, and in special, scientific, technical activities 569 out of 1300 possible municipalities. These findings are similar to the remark made by Liviano-Solís and Arauzo-Carod (2012) 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 and the overdispersion (the excess of conditional variance over the conditional mean), the

[12]The Paris region is one of the most important metropolises in the world. It is Europe's most populated region. While the physical area represents only 2.2% of the surface of France, over 19% of the country's population reside in this area (11.7 million). The GDP of the region amounts to 29% of total French GDP (IAU IdF, 2014). The Paris region's economy is dynamic, innovative, and competitive with a large share of senior professionals, the high density of company headquarters and over 5,6 million jobs distributed across the region. The Paris region's economy is also diversified. Ile-de-France is divided into 1300 municipalities that cover the city of Paris and its suburbs. Very large differences in population and employment densities are to be found between Paris and its outer periphery (see: http://www.iau-idf.fr/lile-de-france/un-portrait-par-les-chiffres/population.html).

hurdle model can be suggested. In this paper, we respond to the complaint voiced by Liviano-Solís and Arauzo-Carod (2013) and Bhat et al. (2014) who noticed that scholars have not fully explored the hurdle model technique yet when analysing location phenomena. Consequently, the empirical evidence (for comparisons purposes) is still scarce. We will try to fill this gap in the business location modeling literature taking into account scarcely two existing papers 1) by Liviano-Solís and Arauzo-Carod (2013) and 2) by Buczkowska and Lapparent (2014).

Liviano-Solís and Arauzo-Carod (2013) found that the hurdle approach fits their industrial sector location data better than the zero-inflated approach. The authors compared several models: Poisson, negative binomial, zero-inflated versions of these models, hurdle Poisson (HP) and hurdle negative binomial (HNB). They showed that the hurdles (HP and HNB) are the models whose expected number of zero counts match the observed zero counts, and that the distribution of the HNB model is the one that best fits the data under study. They concluded that the use of a HNB clearly improves the explanatory power of the econometric estimations, and they suggested that the analysis of firm location behaviour should consider the following factors: 1) the existence of a threshold that allows a site to be chosen by at least one firm and 2) the number of times that this site is chosen by the total population of plants during the analysed period.

Buczkowska and Lapparent (2014) tested various count data models: Poisson, zero-inflated Poisson, zero-inflated (tau) Poisson, negative binomial, zero-inflated negative binomial, and hurdle Poisson models. Having estimated 84 nested and non-nested count data models for various activity sectors, the authors found that the hurdle models are preferable for taking into account the presence of excess zeros and for dealing with overdispersion. 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, as already stated in the paper, one does not know what type of spatial measure is the most appropriate one to characterize spatial spillovers. All these motivations presented in this section and the results described by Liviano-Solís and Arauzo-Carod (2013) and by Buczkowska

and Lapparent (2014) justify our decision to develop a discrete mixture of hurdle-Poisson models wrapping the spatial measures in a common statistical framework of analysis. In our application, we consider that the mixture is the same for every locations. In that, we accept that mixing is done independently of local peculiarities. We obvioulsy agree that it might differ from one location to another and that we consider a somewhat restrictive point of view. Further generalization is left aside for future research work.

4.2 Model specification

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

$$\begin{aligned}\ell\left(d_l, y_l | \mathbf{x}_{l,m}; \theta_{1,m}, \theta_{2,m}\right) = \\ \left(1 - p\left(y_l > 0 | \mathbf{x}_{l,m}; \theta_{1,m}\right)\right)^{1-d_l} \times \\ \left(p\left(y_l > 0 | \mathbf{x}_{l,m}; \theta_{1,m}\right) h\left(y_l | y_l > 0; \mathbf{x}_{l,m}; \theta_{2,m}\right)\right)^{d_l},\end{aligned} \tag{1}$$

where

$$d_l = \begin{cases} 0 \text{ if } y_l = 0 \\ 1 \text{ otherwise} \end{cases}. \tag{2}$$

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

Probability that location l has one or more new establishments that locate at it is based on a latent profit variable: establishments locates 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:

$$\Pi\left(\mathbf{x}_{l,m}; \theta_{1,m}\right) = \mathbf{x}'_{l,m}\theta_{1,m} + \varepsilon_{l,m}. \tag{3}$$

We assume that the error terms $\varepsilon_{l,m}$ are *iid* Logistic with a location parameter equal to 0 and a scale parameter equal to 1. It is well known that, for identification purpose, we have to assume that the scale parameter of the distribution of the error terms is fixed to some given value, here

1. It implies that the values of parameters $\theta_{1,m}$ are not sensible. Their signs and significance matter. The probability to observe one or more new establishments locating at l is then defined as:

$$p\left(y_l > 0 | \mathbf{x}_{l,m}; \theta_{1,m}\right) = \frac{1}{1 + \exp\left(-\mathbf{x}'_{l,m}\theta_{1,m}\right)}. \tag{4}$$

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:

$$h\left(y_l | y_l > 0; \mathbf{x}_{l,m}; \theta_{2,m}\right) = \frac{\lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right)^{y_l} \exp\left(-\lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right)\right)}{y_l!\left(1 - \exp\left(-\lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right)\right)\right)}, \tag{5}$$

where the rate of occurence is parametrically defined as:

$$\lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right) = \exp\left(\mathbf{x}'_{l,m}\theta_{2,m}\right). \tag{6}$$

4.3 Full information maximum (log-)likelihood function

Considering the M types of spatial measures together, we define as $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_M)$, $\sum_{m=1}^{M} \pi_m = 1$, the probability to belong to a type of spatial measure. The full information maximum likelihood estimator (FIMLE) is based on maximizing the following marginal log-likelihood function with respect to unknown parameters $\boldsymbol{\pi}$ and θ conditionally to observed data $\mathbf{x}_{.,l} = (\mathbf{x}_{1,l}, \cdots, \mathbf{x}_{M,l})$:

$$\ell\left(\theta, \boldsymbol{\pi} | \mathbf{y}_{.}, \mathbf{x}_{.,.}\right) = \sum_{l=1}^{L} \ln\left(\sum_{m=1}^{M} \pi_m \ell\left(d_l, y_l | \mathbf{x}_{m,l}; \theta_{1,m}, \theta_{2,m}\right)\right). \tag{7}$$

4.4 Partial effects

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

$$\mathbb{E}\left(y_l | \mathbf{x}_{l,.}; \theta\right) = \sum_{m=1}^{M} \pi_m \frac{p\left(y_l > 0 | \mathbf{x}_{l,m}; \theta_{1,m}\right)}{1 - \exp\left(-\lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right)\right)} \lambda\left(\mathbf{x}_{l,m}; \theta_{2,m}\right) \tag{8}$$

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 spatial measure m. The effect on the expected number of new establishments that locate at l with respect to a variation of it is defined as:

$$\begin{array}{l} \frac{\partial \mathbb{E}(y_l|\mathbf{x}_{l,.})}{\partial z_l} = \\ \sum_{m=1}^{M} \pi_m \frac{\partial p(y_l>0|\mathbf{x}_{l,m};\theta_{1,m})}{\partial z_{l,m}} \mathbb{E}\left(y_l|y_l>0,\mathbf{x}_{l,m};\theta_{2,m}\right) + \\ \sum_{m=1}^{M} \pi_m p\left(y_l>0|\mathbf{x}_{l,m};\theta_{1,m}\right) \frac{\partial \mathbb{E}(y_l|y_l>0,\mathbf{x}_{l,m};\theta_{2,m})}{\partial z_{l,m}} \end{array} \tag{9}$$

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

4.5 Posterior class assignment probabilities

Another interesting point is that, once the model is estimated, one also may compute posterior class assignment probabilities, i.e probability of spatial measure m contingently on location l:

$$\psi_{j|l} = \frac{\pi_j \ell\left(d_l, y_l|\mathbf{x}_{j,l};\theta_{1,j},\theta_{2,j}\right)}{\sum_{m=1}^{M} \pi_m \ell\left(d_l, y_l|\mathbf{x}_{m,l};\theta_{1,m},\theta_{2,m}\right)} \tag{10}$$

By doing so, we update our "knowledge" about which spatial measure is appropriate for a location l using observed (aggregate) choices of establishments. Such a result gives us a clue to the probability distribution of types of spatial measure m given location l. It does not state which is to be used but which is the most likely to be considered.

4.6 Spatial spillovers

We discuss below the structure of the matrix of observed explanatory variables. This matrix contains variables directly concerning either location l or sector s and the number of pre-existing establishments in the relative sector s. We account also for the characteristics of the surrounding areas

and the stock of establishments located nearby when modeling the spatial spillovers as follows:

$$x_{l,s} = \ln\left(\sum_{j=1}^{L} e^{-\mu d_{l,j}} z_{j,s}\right), \tag{11}$$

where $z_{j,s}$ is an attribute of the municipality that applies to sector s or is the number of already existing establishments from this sector. μ is fixed to 1 and $d_{l,j}$ is the distance between the centroids of municipalities l and j.

Extending the paper by Buczkowska and Lapparent (2014), where $d_{l,j}$ was the Euclidean distance, we consider alternative distance measures when building the spatial distance weight matrices, namely the travel time by car during the morning peak hour and the off-peak period, and evaluate their performance for different activity sectors.

5 Results

The results are organized in two subsections: the first one presents and discusses the models and the parameter estimates, while the second one focuses on the class assignments probabilities.

5.1 Estimates

We estimate the hurdle-Poisson mixture model with two latent classes for three selected sectors: 1) construction, 2) special, scientific, and technical activities, and 3) real estate activities. Furthermore, we consider two alternative cases for the mixture: in the first mixture, the two classes are based on the Euclidean distance (ED) and the peak road travel time (TtVhMph), while in the second mixture, the first class is based on the off-peak road travel time (TtVhOph) and the second class is again based on TtVhMph. For reminder, the class indicates which distance measure is used to compute spatial spillovers. However, we do not systematically present all cases for the sake of concision, to focus on the most illustrative ones.

The full set of parameter estimates is presented in Table 4 for the construction sector and the mixture model for the case: ED with TtVhMph.

Estimates of the second part of the model (truncated-at-zero Poisson distribution) are presented for all three sectors for the mixture models in Annex (Table 7)[13].

Focusing first on the hurdle part, one can observe that the peak road travel time seems to provide better results than the Euclidean distance. More parameter estimates are significant and have the expected sign for class 2 (peak road travel time) than for class 1 (Euclidean distance). For instance, one may expect that the amount of vacant land increases the probability to cross the hurdle, i.e., that at least one establishment in the construction sector locates in the municipality[14]. The sign of the associated parameter should consequently be positive, which is the case for class 2 but not for class 1. One estimate, distance to highway, does not present the expected sign for class 2, but is actually not significant. This being said, we find results that are in line with Buczkowska and Lapparent (2014) and the literature in general. In particular, the presence of establishments from the same sector in the vicinity increases the probability that at least one establishment settles in the municipality. Conversely, large establishments or high real estate prices act as deterrents to the implantation of new establishments.

We now turn to the results of the truncated-at-zero Poisson parts of the mixture models, this time for all three sectors. We observe for all sectors marked localization patterns: the greater the presence of establishments from a given sector, the greater the number of newly created units of this sector locating within the same area. Conversely, the presence of large establishments tends to repel new establishments. High real estate prices (of shops or offices depending on the sector considered)[15] 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 choice deci-

[13]The hurdle parameters are not presented for the sake of concision

[14]There are at least two reasons to think so. First, more vacant land means that it will be easier for one establishment to settle there. Second, more vacant land means potentially more constructions in the future, which should attract construction firms.

[15]Price levels for shops or offices can be treated as a proxy for the average price that an establishment needs to pay to set up on the market.

sions 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. Proximity to public transportation is also an important criterion in the real estate sector. On the other hand, proximity to the highway network did not turn out to be significant. One interpretation is that real estate establishments act more locally and settle preferentially in dense areas with good access to public transit, with customers maybe more prone to come by foot or by public transit than by the highway.

Now 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 establishments, such as schools, universities, hospitals and clinics. They also prefer municipalities where people both live and work. Special, scientific and technical activities look for areas with good access to the intellectual workforce, and close to other academic establishments and to offices. Last, the presence of high-income households increases the probability that new establishments from the real estate sector settle in the area.

In order to check the robustness of our model, we compare the parameter estimates of the positive count parts of the two mixture models described at the beginning of this subsection (see Figure 3). We observe that the estimates of most of the variables tend to behave in a similar way. In particular, the parameter estimates for the class TtVhMph is little sensitive to the choice of the other class (ED or TtVhOph), which tends to validate the robustness of our model.

In addition to the mixture of hurdle-Poisson models based on two latent classes presented above, we run two "pure" hurdle-Poisson models, a first one based on ED and a second one based on TtVhMph[16]. We then calculate and compare the Bayesian Information Criterion (BIC) of the mixture model with the BIC levels of the "pure" HP models. The results are reported in Tables 4 and 5. We stress that all these models use the same

[16]See Buczkowska and Lapparent (2014) for more details on the "pure" hurdle-Poisson model.

Table 4: Hurdle-Poisson mixture model for the construction sector: ED (Class 1) and TtVhMph (Class 2)

Matrix:	ED (Class 1)		TtVhMph (Class 2)	
Hurdle part	Estimate	T-Statistics	Estimate	T-Statistics
Constant	-4,267	-0,48[1]	29,572 ***	4,28
Estab. from respective sector[2]	1,341	1,49	0,344 **	2,18
Large estab. from all sectors	0,341	1,05	-0,096 *	-2,03
Trips home-work	-1,223	-0,92	1,530 ***	4,71
Shops and offices	-0,211	-0,94	0,166 ***	3,59
Vacant land	-2,895 *	-1,74	0,424 ***	4,05
Universities and schools	0,609	1,06	0,275 ***	3,14
Hospitals and clinics	-0,677	-1,38	0,103 **	2,46
Distance to highway	-2,770 *	-1,76	0,046	0,37
Public transport	0,111	0,48	0,135 ***	3,43
Residence tax	6,973 **	2,06	0,135	0,40
Price of shops (log)	-7,331 *	-1,72	-8,892 ***	-2,89
Poisson part: Positive counts	Estimate	T-Statistics	Estimate	T-Statistics
Constant	23,194 ***	40,70	19,842 ***	8,21
Estab. from respective sector	1,296 ***	24,28	1,140 ***	16,34
Large estab. from all sectors	-0,114 ***	-4,00	-0,092 ***	-4,58
Trips home-work	2,022 ***	30,07	1,385 ***	24,88
Shops and offices	0,115 ***	4,59	0,245 ***	12,21
Vacant land	0,077 ***	3,00	0,002	0,12
Universities and schools	0,559 ***	12,96	0,476 ***	12,83
Hospitals and clinics	0,055 ***	3,13	0,056 ***	5,39
Distance to highway	-0,136 ***	-4,96	-0,083 ***	-4,94
Public transport	0,098 ***	6,16	0,080 ***	7,22
Residence tax	-0,381 ***	-4,07	-0,213 **	-2,68
Price of shops (log)	-4,342 ***	-26,63	-4,107 ***	-3,96
Pi (probability of class 1)	0,322 ****	13,70		
#Parameters	2 x 12 x2			
#Observations	1300,000			
Objective function	-2790,61			
BIC	5692,37			

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

[2]See Table 3 for the description of variables.

Table 5: Simple hurdle-Poisson models for the construction sector. Two models are run independently for ED and then for TtVhMph

Matrix:	ED		TtVhMph	
Hurdle part	Estimate	T-Statistics	Estimate	T-Statistics
Constant	13,792 ***	7,19	15,779 ***	3,30
Estab. from respective sector	0,287 **	2,07	0,174 ***	2,63
Large estab. from all sectors	-0,031	-0,80	0,009	0,35
Trips home-work	0,984 ***	4,70	0,778 ***	4,82
Shops and offices	0,066	1,60	0,080 ***	3,10
Vacant land	0,023	0,27	0,046	0,74
Universities and schools	0,402 ***	4,88	0,062	1,54
Hospitals and clinics	-0,029	-0,98	0,008	0,37
Distance to highway	-0,210 **	-2,31	-0,167 **	-2,13
Public transport	0,092 ***	3,01	0,072 ***	3,35
Residence tax	1,135 ***	4,53	1,124 ***	5,52
Price of shops (log)	-4,266 ***	-6,98	-6,222 ***	-2,84
Poisson part: Positive counts	Estimate	T-Statistics	Estimate	T-Statistics
Constant	22,375 ***	68,41	16,471 ***	33,97
Estab. from respective sector	1,553 ***	45,47	1,147 ***	48,29
Large estab. from all sectors	-0,256 ***	-13,99	-0,087 ***	-8,84
Trips home-work	1,689 ***	46,49	1,470 ***	53,28
Shops and offices	0,355 ***	21,49	0,166 ***	18,41
Vacant land	-0,040 ***	-2,61	-0,001	-0,15
Universities and schools	0,450 ***	18,28	0,432 ***	27,69
Hospitals and clinics	0,024 **	2,44	0,039 ***	7,08
Distance to highway	-0,159 ***	-10,06	-0,089 ***	-9,73
Public transport	0,222 ***	19,55	0,101 ***	20,67
Residence tax	-0,185 ***	-3,46	-0,229 ***	-6,11
Price of shops (log)	-4,633 ***	-53,34	-2,494 ***	-11,79
#Parameters	2 x 12		2 x 12	
#Observations	1300		1300	
Log-Likelihood	-3849,25		-3842,30	
AIC	7746,50		7732,60	
AICC	7747,40		7733,50	
BIC	7870,60		7856,60	

set of variables and the same number of observations (1300). However, the number of parameters doubles when using the mixture of the hurlde-Poisson models (48 parameters) in comparison to the "pure" HP models (24 parameters). We find that the mixture model proposed in this paper performs significantly better than the "pure" hurdle-Poisson models based on a single distance measure. For the construction sector, the BIC is equal to 5692.37 for the mixture model, with a reduction of more than 2000 compared to the "pure" HP models (with BIC of 7870.60 for ED and 7856.60 for TtVhMph)[17].

Overall, we find that the mixture of hurdle-Poisson models is relevant as it performs significantly better than pure HP models. The consideration of alternative distance measures to the Euclidean distance even provided better results for the hurdle part. Last, regarding the significance and sign of our parameter estimates, most of our results are conform to our review of the literature, which is again a sign of the robustness of our results. The problematic variables turn out to be availability of the vacant land and the proximity to residential areas, for which, depending on the distance measure used, the signs turned out to be not always positive as expected based on our survey of the literature (if still significant). By tracking changes on the objective function, we see however, that these variables have little effect on the objective function, thus on the choices of the establishments. Still, this point should be further investigated in the future.

5.2 Class assignment probabilities

We can now tackle our main research question, i.e., which distance measure is the most appropriate to capture spatial spillovers in our establishment location choice model. Again, we focus the analysis on the same three

[17]For the real estate sector, the BIC of the mixture model equals to 3628.71. The Bayesian Information Criteria for the "pure" hurdle-Poisson models based on one class, ED or TtVhMph, are at the level of 4497.30 and 5079.90, respectively. For the special, scientific, technical activities sector, the BIC of the HP mixture model is 5270.61, which is almost 4000 less than the BIC of the "pure" HP models based on ED or TtVhMph, for which BIC equals to 8839.90 and 9090.10, respectively.

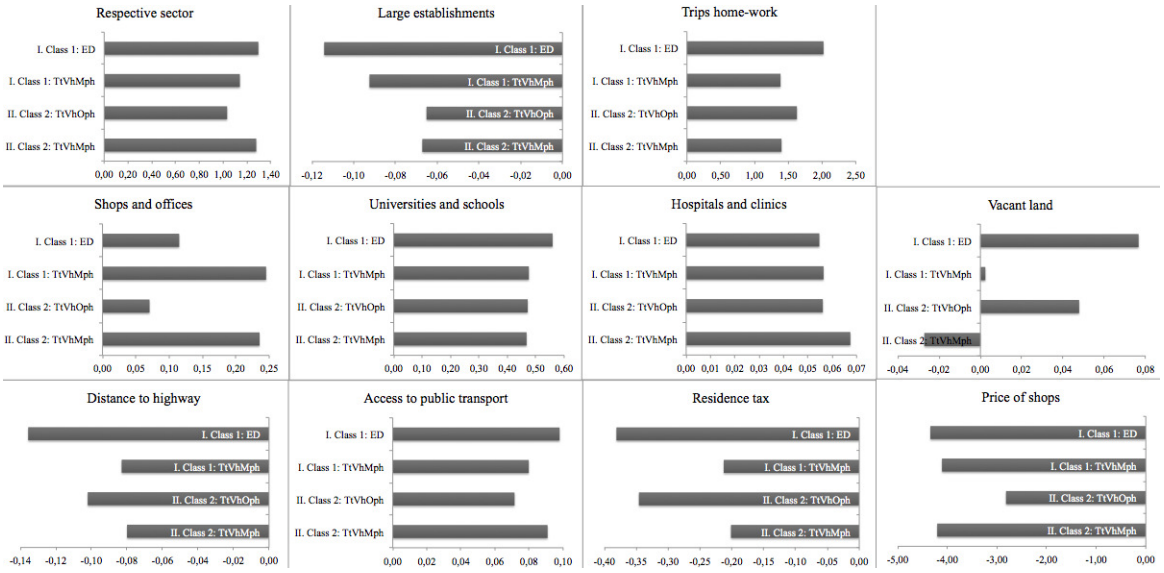

Figure 3: Selected example of the estimates of the truncated-at-zero Poisson parts of the mixture models for the construction sector. Comparison of two cases for models that use two classes. The first case (I.) includes ED (class 1) and TtVhMph (class 2) matrices. In the second case (II.), TtVhOph (class 1) and TtVhMph (class 2) coexist.

economic sectors, construction, real estate, and special, scientific and technical activities, and we consider two alternative mixtures regarding the distance measures: Euclidean distance (ED) with peak road travel time (TtVhMph), and off-peak road travel time (TtVhOph) with peak road travel time (TtVhMph). The estimated class assignment probabilities are reported in Table 6. For the construction sector, Pi is equal to 0.322 in case 1 (ED with TtVhMph) and to 0.321 in case 2 (TtVhOph with TtVhMph). Therefore, the peak road travel time is in both cases the most likely to adequately capture spatial spillovers in our HP model. We find the same result for special, scientific and technical activities, with Pi equal to 0.283 and 0.222 in case 1 and 2, respectively. On the other hand, for the real estate sector, the value of Pi is 0.522 when the two classes are based on ED and TtVhMph, and 0.639 when they are based on TtVhOph and TtVhMph.

For the first two sectors, the predominance of peak road travel time 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 Euclidean distance for the real estate sector tends to emphasize that spatial effects are channelled not only through the road mode but also and mainly through other modes, such as walk, public transit, or even communication modes.[18] Regarding the fact that we find relatively similar results for the Euclidean distance and the off-peak road travel time, this probably stems from the higher level of correlation between the two than between the Euclidean distance and the peak road travel time (see Table 2).

As indicated in Section 4, we may then compute the posterior probabilities for each of the 1300 municipalities of the Paris region. For the sake of concision, we do so only for the construction and real estate sectors, and for the case 1 (ED and TtVhMph). The results are presented in Figure 4. Overall, it is clear that the peak road travel time prevails in more municipalities for the construction sector, while the situation is more mixed for the real estate sector. This being said, no clear spatial patterns appear at this stage. The proximity of highway tends to be associated with the predominance of the peak road travel time, there are some 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 road travel time. One possible interpretation would be that when density is high enough, the market size allows establishment to operate at a more local scale, while in the least dense parts of the metropolitan areas, establishments must increase their market area and thus rely more heavily on the car use. These points call for more investigation, which will be the object of our further work.

[18] Again, one possible interpretation is that real estate establishments might operate at a more local scale, settling preferentially in dense areas with good access to public transit, with customers maybe more prone to come by foot or by public transit than by the highway.

Figure 4: 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) and real estate sector (lower figure)

Table 6: Estimated probability of belonging to class 1 (Pi): Comparison across cases and sectors

	Case I	Class 1: ED Class 2: TtVhMph		Case II	Class 1: TtVhOph Class 2: TtVhMph	
Sector[1]	Pi[2]	T-Statistics[3]	Convergence	Pi	T-Statistics	Convergence
Constr	0.322***[4]	13.70	Satisfied	0.321		Satisfied[5]
SpecSci	0.283***	9.12	Satisfied	0.222***	10.90	Satisfied
RealEst	0.522***	11.12	Satisfied	0.639***	16.12	Satisfied

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

[2]The level of estimated probability Pi inferior to 0.5 indicates that the distance measure of the second class has a larger probability to be the appropriate measure to account for spatial spillovers as compared to the distance measure of the first class.

[3]Convergence stands for Convergence criterion.

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

[5]Convergence criterion has been satisfied, yet, the standard error was not reported for this particular case.

6 Conclusions

We contribute to the existing literature on location choice models, opening up a discussion and a whole new direction of empirical exploration using appropriate econometric tools and the more carefully considered distance definition for location analysis. To compare the various distance measures, we developed a probabilistic mixture of hurdle-Poisson models that use two latent classes for several activity sectors. We applied it to the location decisions of establishments that wish to set up on the market. Each class used a different definition of distance to capture spatial spillovers. The following distance measures were considered at first: the Euclidean distance, two road distances (with and without congestion), the public transit distance, and the corresponding travel times. After restricting the set of tested measures due to the correlation issues, we estimated several mixture models for the Paris region.

Based on the performed analyses we drew four main conclusions. 1) Overall, the obtained results are in line with the literature regarding the main determinants of establishment location. 2) Based on the Bayesian

Information Criteria (BIC), we found that the proposed mixture of hurdle-Poisson models that uses two latent classes performs significantly better than the "pure" hurdle-Poisson models based on a single distance measure, emphasizing the usefulness of our approach. By using the mixture hurdle-Poisson model we considerably decreased the level of BIC up to 42%. 3) From the overall level of estimated probabilities Pi, we observed that for some activity sectors, such as construction, the peak road travel time is the most likely to correctly capture spatial spillovers, whereas for other sectors, such as real estate, the Euclidean distance slightly prevails. This tends to show that spatial spillovers are channeled by different means depending on the activity sector. Our analyses showed that for some transport-oriented sectors, such as construction, for which a good transport infrastructure is tremendously important, it seems more reasonable to consider travel times instead of an Euclidean distance measure in the establishments location models. As stressed by Bhat *et al.* (2014), a good roadway network is extremely important for businesses in some sectors for unhindered delivery of raw materials from other regions to the business locations and finished products from business locations to the markets. For other sectors, which do not rely so heavily on the transport infrastructure and which search the proximity to the potential client or user, such as real estate, the Euclidean distance tends to perform well to account for the linkage between neighboring areas. 4) In addition, by allowing different distance measures to coexist within a hurdle-Poisson mixture model, the hurdle part of the model that uses the appropriate distance matrix significantly improves.

In the current exercise we tested the mixture model using only two classes. The number of latent classes could be increased, provided that one finds additional distance measures that are both relevant from an economic point of view and not excessively correlated with the ones already used. The proposed specification can also be applied in other fields, such as the residential location or land-use models, this whenever the Euclidean distance does not seem the most appropriate distance measure to account for the relationship between neighboring observations.

Anselin (2010) described the evolution of the field of spatial econometrics, arguing that it moved from the margins of applied regional science

to the mainstream of econometric methodology. Now that the field has reached maturity, Anselin (2010) asked what will come next? What are the exciting new directions and challenges that have only been partially addressed? He saw at least three: 1) The complex dynamics that result in the existence of spatial interaction are still poorly reflected in model specifications; 2) The second challenge is to deal with the analysis of massive data sets (e.g., geographical and individual scale); 3) The final challenge parallels the previous ones and refers to the computational techniques needed to handle the complex interactions in large data sets. In keeping with Anselin, we will follow this direction in our future work. Drawing from individual business data, we will try to incorporate some spatial interactions in the location choice models.

References

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

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.

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

Billé A.G. and G. Arbia, 2013, Spatial discrete choice and spatial lim-

ited dependent variable models: A review with an emphasis on the use in regional economics. *ArXiv e-prints*.

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.

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

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.

Dattorro, J., 2005. Convex Optimization & Euclidean Distance Geometry $M\epsilon\beta oo$ *PublishingUSA*, v2010.01.05. https://ccrma.stanford.edu/ dattorro/EDM.pdf.

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

Duran-Fernandez, R., 2008. Gravity, distance, and traffic flows in Mexico. *Working Paper 1023*, Transport Studies Unit, Oxford University.

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

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.

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.

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

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

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

Klier, T., 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.

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

Lapparent, M. (de) and M. Koning, 2014. Travel discomfort-time tradeoffs in Paris subway: an empirical analysis using interval regression models, *Working Paper*.

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, an: S. J. Rey. Heidelberg: Springer.

LeSage, J.P. and R.K. Pace, 2012. The biggest myth in spatial econometrics, *Working Paper*.

LeSage, J. P., 2014. What regional scientists need to know about spatial econometrics. *Working Paper*.

Liesenfeld, R., J.-F. Richard, and J. Vogler, 2013. Analysis of Discrete Dependent Variable Models with Spatial Correlation. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2196041.

Liviano-Solís, D. and J.-M. Arauzo-Carod (2012). Industrial Location and Spatial Dependence: An Empirical Application. *Regional Studies*, DOI:10.1080/00343404.2012.675054.

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

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. Flora, and S. J. Rey. Heidelberg: Springer.

Miaou S. and D. Sui, 2004. Implications of changing demographic and socioeconomic structures on highway safety: a texas initiative. *Final report*, 19-21.

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

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

Pace, R. K. and J. P. LeSage, 2008. A Spatial Hausman Test, *Economics Letters*, 101(3), 282-284.

Partridge, M.D., M. Boarnet, S. Brakman, and G. Ottaviano 2012. Introduction: Whither Spatial Econometrics? *Journal of Regional Science*, 52(2), 167-171.

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

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.

Slade, M. E., 2005. The Role of Economic Space in Decision Making,

ADRES Lecture. *Annales D'Economie et de Statistique*, 77, 1-20.

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.

Appendix

Table 7: Mixture of Hurdle-Poisson models for construction, special, scientific technical activities, and real estate activities. Case: ED (1st class) with TtVhMph (2nd class). Truncated-at-zero Poisson parts reported.

Constr	Estimate	T-statistics	SpecSci	Estimate	T-statistics	RealEst	Estimate	T-statistics
Class 1 Matrix:	ED		Class 1 Matrix:	ED		Class 1 Matrix:	ED	
Positive counts								
Constant	23,194 ***[1]	40,70	Constant	14,043 ***	23,17	Constant	16,327 ***	20,49
Respec. sector[2]	1,296 ***	24,28	Respec. sector	0,490 ***	8,17	Respec. sector	1,729 ***	11,77
Large estab.	-0,114 ***	-4,00	Large estab.	-0,185 ***	-5,92	Large estab.	-0,453 ***	-7,41
Trips HW	2,022 ***	30,07						
			Trips HW, intel.	0,932 ***	22,98			
			Offices	0,039 ***	3,10			
Shops, offices	0,115 ***	4,59				Shops, offices	0,598 ***	13,07
Vacant land	0,077 ***	3,00						
						Resid. area	-0,173 **	-2,40
Univ., schools	0,559 ***	12,96	Univ., schools	0,323 ***	7,28			
Hospitals, clinics	0,055 ***	3,13						
Dist. highway	-0,136 ***	-4,96	Dist. highway	-0,318 ***	-11,14			
Public transp.	0,098 ***	6,16	Public transp.	0,222 ***	10,31	Public transp.	0,425 ***	17,24
Residence tax	-0,381 ***	-4,07				Residence tax	-0,573 ***	-3,53
						log Income	0,688 ***	5,12
			log Office Price	-2,429 ***	-15,74			
log Shop price	-4,342 ***	-26,63				log Shop price	-3,085 ***	-11,82
Class 2 Matrix:	TtVhMph		Class 2 Matrix:	TtVhMph		Class 2 Matrix:	TtVhMph	
Positive counts								
Constant	19,842 ***	8,21	Constant	12,419 ***	9,19	Constant	10,481 ***	8,05
Respec. sector	1,140 ***	16,34	Respec. sector	0,118 *	1,74	Respec. sector	0,710 ***	3,02
Large estab.	-0,092 ***	-4,58	Large estab.	0,0004	0,02	Large estab.	-0,094 **	-2,09
Trips HW	1,385 ***	24,88						
			Trips HW, intel.	0,638 ***	18,77			
			Offices	0,080 ***	9,42			
Shops, offices	0,245 ***	12,21				Shops, offices	0,418 ***	6,65
Vacant land	0,002	0,12						
						Resid. area	0,239 **	2,67
Univ., schools	0,476 ***	12,83	Univ., schools	0,388 ***	14,71			
Hospitals, clinics	0,056 ***	5,39						
Dist. highway	-0,083 ***	-4,94	Dist. highway	-0,00004	0,00			
Public transp.	0,080 ***	7,22	Public transp.	0,078 ***	6,92	Public transp.	0,366 ***	5,55
Residence tax	-0,213 **	-2,68				Residence tax	-0,850 ***	-3,47
						log Income	1,099 ***	5,61
			log Office Price	-2,423 ***	-4,16			
log Shop price	-4,107 ***	-3,96				log Shop price	-1,513 ***	-12,58
Pi	0,322 ****	13,70	Pi	0,283 ***	9,12	Pi	0,522 ***	11,12
Convergence criterion	Satisfied		Convergence criterion	Satisfied		Convergence criterion	Satisfied	
#Parameters	2 x 12 x2		#Parameters	2 x 9 x2		#Parameters	2 x 9 x2	
Objective function	-2790,608		Objective function	-2593,621		Objective function	-1772,672	

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

[2]See Table 3 for the description of variables.