DOES "DISTANCE" REALLY MATTER FOR DISTANCE-BASED MEASURES OF GEOGRAPHIC CONCENTRATION?

Florence Puech[a,d] **Nicolas Ovtracht** [a] **Pablo Jensen** [a,b,c] **Valérie Thiebaut** [a]

[a]Laboratoire d'Économie des Transports – LET (Université Lyon 2, CNRS, ENTPE), Lyon, France.

[b]Institut des Systèmes Complexes Rhône-Alpes (IXXI), Lyon, France.

[c]Laboratoire de Physique, École Normale Supérieure de Lyon and CNRS, Lyon, France.

[d]Corresponding author: Florence.Puech@univ-lyon2.fr

PRELIMINARY VERSION

16 MARCH 2010

ABSTRACT

Distance-based methods have been recently improved to evaluate the geographic concentration of economic activities. For these methods, the measurement of distances between all pairs of establishments is crucial. The Euclidean metric is systematically retained in the economic literature even though the location of activities is clearly constrained by the network (roads) or natural advantages (like rivers). In the article, we discuss the relevance of the Euclidean distance as a proxy of the actual network-based distance at an urban scale. It is shown on the Lyon area (France) that the significant relative geographic concentration or dispersion of stores is systematically underestimated when the Euclidean metric is used.

JEL Classification: C40, C81, L81, R14.
Keywords: Distance-based methods, M function, Urban network, Location of retail stores.

ACKNOWLEDGMENTS

We are grateful to the Chamber of Commerce and Industry of Lyon and Christophe Baume for providing us with data on stores. We thank the PLAINSUDD project (innovative numerical platforms of urban simulation for sustainable development) supported by the ANR "Villes Durables" which partly financed this study. We are grateful for helpful comments from Gilles Duranton, David Wong, Christine Thomas-Agnan, Florent Bonneu and participants of the Toulouse School of Economics-TSE Statistics seminar (March 10), the 8th Spatial Econometrics and Statistics workshop (Besançon, June 09) and the 56th North American Meetings of the Regional Science Association International (San Francisco, November 09).

1. INTRODUCTION

The question on ***the best way to evaluate the geographic distribution of economic activities*** has known a renewed interest in the last decade. The recent economic literature has underlined that the frequently used Gini or Ellison and Glaeser (1997) indexes are less powerful than alternative tools called "distance-based methods" (Combes *et al.* 2008). The latter rest on the point pattern theory (Diggle, 1983). They were initially imported from other scientific fields and then improved to be tractable for evaluating the distribution of activities (Duranton and Overman, 2005; Marcon and Puech, forthcoming). Distance-based methods offer a complete and precise analysis of the spatial distribution of entities (industrial establishments, stores…). The geographic concentration is not estimated from any administrative or predefined zoning but from bilateral distances between precise location of entities. The distance measurement between pairs of entities is thus crucial for providing an exact and detailed location pattern of activities. However, as far as we know, the definition of distance has rarely been discussed and the Euclidean metric is systematically chosen in economic studies (Arbia and Espa, 1996; Jensen 2006; Ó hUallacháin and Leslie, 2007; Arbia *et al.*, 2008; Fratesi, 2008; Arbia *et al.*, forthcoming). Nonetheless, it clearly appears that a good evaluation of the actual attraction or repulsion of stores constitutes an important stake to explain, in second time, the more accurately economic phenomena (like the real scope of agglomeration economies).[1]

Our paper completes recent developments made on distance-based methods for the appraisal of the geographic concentration of retail activities. ***Our aim is to check the robustness of the Euclidean distance as a proxy for the real distance on urban network.*** Modeling the potential attraction or repulsion of stores calls for the more accurate estimation of distances existing between them. At a city level, "as the crow flies" gives the first and naïve impression that it can not be the best candidate. The main reason is that interactions between individuals are constrained by various elements. The first issue is that stores are located on the street network. If one goes from one shop to another, he has to move on the urban network. Thus, joining any two network-constrained stores is more naturally captured by a network distance than a Euclidean one. The second issue is that every part of the network is definitely not accessible. This limitation

[1] Because technological externalities appear when people communicate, such externalities should theoretically be more important in the very close environment and their intensity should rapidly attenuate with distance. This result has empirically been verified by Jaffe *et al.* (1993) or Rosenthal and Strange (2001) with cluster-based methods and by Rosenthal and Strange (2003) using Euclidean distances. As far as we know, no study has used network-based distance to gauge the scope of agglomeration externalities.

refers for example to the existence of authorized access roads. A driver can not move freely: he has to avoid pedestrian streets or going up one-way streets. In what follows, an empirical analysis is provided to quantify how much geographic concentration results deviate according to the distance definition. Our aim is to test whether the use of the actual network significantly improves the evaluation of agglomeration or if "as the crow flies" is a good proxy of the network distance. Our work explores the agglomeration of non-eating retail stores in Lyon-Villeurbanne (France).[2] The real road network was simulated for any motorized vehicle trip. This transportation mode was chosen for two reasons. Firstly, travels by car face the most important number of constraints on a network (in comparison with pedestrian travels for example). Thus if differences exist, they should be greater by retaining the car-network movements. Secondly, it has also been proved that the share of the car mode for shopping purpose is far from being insignificant on the Lyon-Villeurbanne area.[3]

Our study consistently finds that the definition of distance matters in the evaluation of agglomeration of retail stores at the city-level. We go further by proposing consistent explanations to expound disparities in estimates. The remainder of the article is in four parts. It begins with the state of the art on distance-based methods. Then, we provide a presentation of the data and give an intuitive empirical example of the possible deviations depending on the distance definition. The next part contains the empirical approach and the explanation on how much distance definition influences the estimation of the agglomeration of stores. The last section concludes.

2. STATE OF THE ART

2.1. IN ECONOMICS

The use of distance-based methods for evaluating the geographic concentration of economic activities can be explained in two steps.

Firstly, several authors imported in economics measures largely employed in other scientific fields like ecology (see for example Barff, 1987; Arbia and Espa, 1996; Sweeney and Feser, 1998; Marcon and Puech, 2003). Those statistical methods are based on the point pattern analysis

[2] The Lyon-Villeurbanne area corresponds to the CBD of the Lyon metropolitan area (see Figure 1, page 8).
[3] According to the Household Travel Survey 2006 (*Enquête ménages déplacements*), a quarter of "house-shopping" travels over Lyon-Villeurbanne is made by car (SYSTRAL, 2007).

which the most famous tool is undoubtedly the Ripley's *K* function (1976, 1977). The geographic concentration is evaluated by counting the average number of neighbors of stores on a disk of a given radius *r*. The basic idea of this cumulative function is to test the observed spatial structure of a distribution against a random one (complete spatial randomness). If the actual distribution is more clustered at this radius than what it should be under the null hypothesis, the distribution is "spatially concentrated" (at this radius). On the opposite, if the distribution is more regular (at this radius *r*) than what we could expect under randomness, the distribution is defined "geographically dispersed". Finally, to characterize the complete structure of the stores distribution, this operation is then repeated for all possible radii.

Unfortunately, Ripley's *K* function does not respect all criteria for being a good measure of spatial concentration of activities. According to Duranton and Overman (2005), any measure should be able to (i) compare the geographic concentration results across industries, (ii) control for industrial concentration, (iii) control for the overall aggregation patterns of industries, (iv) test the significance of the results and, (v) keep the empirical results unbiased across geographic scales. The *K* function does not fulfill the properties (ii) and (iii). The first one because Ripley's *K* function is defined in homogenous space. This assumes the same density of neighbors all over the study area (hypothesis of a completely spatial random distribution of establishments: plants are distributed uniformly and independently). The hypothesis of homogeneity is not sustainable in the field of economics because human activities greatly aggregate in areas, towns etc.: complete spatial randomness is consequently not the best benchmark. Thus, property (ii) calls for a relative index to gauge the geographic concentration of industries. The property (iii) ensues from the (ii): Since the article of Ellison and Glaeser (1997), economic literature has steadily underlined all the importance for controlling industrial concentration (Combes and Overman, 2004). However, the *K* function does not integrate such quality.

For all of these reasons, after having imported distance-based methods, in a second step several authors improved existing tools. They introduce not topographic but relative concentration[4] measures to get round the homogeneous space problem. Both issues are then solved and a large use in economics of these relative tools can now be considered. For example, Marcon and Puech (forthcoming) introduced the *M* function, as a generalization of the *K* function in non-homogenous space. The idea is as follows. Let us denote by N_t the total number of stores on the

4 "Topographic concentration" is defined as in Brülhart and Traeger (2005): the spatial distribution of activities is evaluated according to physical space. The "relative concentration" is taken in the sense of Haaland *et al.* (1999) that is considering the distribution of another variable as a benchmark.

territory and by N_s the one of a given sector *S*. Around each plant *i* of the sector *S*, a disk of radius *r* is drawn. The *M* function compares the average local proportion of *S*-stores in the aggregate activity within a distance *r,* $N_S(A_i,r)/N_t(A_i,r)$, against the same proportion but defined on the entire territory N_S/N_t. We can easily prove (Jensen and Michel, forthcoming) that the *M* function can be defined as: $M_S(r) = \frac{N_t - 1}{N_S(N_S - 1)} \sum_i^{N_S} \frac{N_S(A_i, r)}{N_t(A_i, r)}$

At this radius *r*, the *S-sector* is said geographically concentrated (respectively dispersed), if the average proportion of the *S-stores* compared to the whole industry is greater (respectively lesser) than the one on the entire territory. The construction of a confidence interval of the null hypothesis allows testing the significance of the results: we check whether around *S-stores*, the relative location of *S-stores* follows the same pattern as the one observed on the entire territory.

Duranton and Overman (2005) have also recently proposed a distance-based method in non-homogeneous space, called the *K*-density function. Their measure is not a cumulative function but a probability density function. Counting neighbors *at* a radius *r* (and not on a disk) is here privileged. This function, smoothed and normalized, have a strong potential to be largely employed in economics (Duranton and Overman, 2008; Ellison *et al.*, forthcoming). To sum up, the *M* and *K*-density functions constitute two complementary measures to evaluate accurately the geographic concentration of economic activities (see Marcon and Puech (forthcoming) for theoretical examples).

2.2. In geography

Geographers (or people specialized in geographic issues) are logically at the leading edge of progress for capturing space. The first distance-based method for evaluating network constrained distributions was proposed in that field by Okabe and Yamada (2001). They called it the "network *K*-function". The three main characteristics of this measure are as follows. Firstly, it constitutes an extension of the original Ripley's *K* function. Thus the Okabe and Yamada's measure is defined as a cumulative function. Secondly, the distance between any two entities is not evaluated thanks to the Euclidean distance but the shortest path on the network. Lastly, the construction of the null hypothesis is obtained by simulating entities independently distributed *over the network* according to a uniform probability.

Yamada and Thill (2005) show recently that the network has a significant impact on the evaluation of the spatial structure of network-constrained distributions. To prove this, they simulate random distributions of traffic accidents in three large regions in the State of New York. The network *K*-function confirms the random location of accidents on the network. However, Yamada and Thill empirically show that possibly Ripley's *K*-function computed with Euclidean distance "*mistakenly detects a clustered pattern even though the observed pattern is actually random with respect to underlying network*" (p.155). This result invites to proceed with attention for evaluating the spatial structure of any network-constrained distributions.

2.3. RECONCILE BOTH APPROACHES

As far as we know, the street network has never been included in the field of economics to evaluate the geographic concentration of activities with distance-based methods. One explanation is that economists generally analyze the distribution of activities with distance-based methods at larger scale than the city-level. They then generally justify the use of Euclidean distance thanks to two arguments (Duranton and Overman, 2005). The first one is that the incidences of the curvature of the earth are not enough to create a bias at the scale of interest. The second one rests on the findings of Combes and Lafourcade (2005). In that paper, authors prove that there is a very high correlation (0.97) between Euclidean distances and generalized transport costs computed from real truck transport data on a simulated network containing 5,000 arcs on the French metropolitan territory. We can expect that the curvature of the earth's problem is all the more insignificant at the city level. However, increasing the precision of the network has an unclear impact on the pertinence of the Euclidean distance proxy. On one hand, the more the number of arcs, higher the number of constraints for moving from one location to another. In that case, the real network distance for joining any two locations should be longer than Euclidean one. On the other hand, if the number of arcs is very high, the shortest path between two locations could be closer to a straight line. In the latter case, Euclidean distance should be a good proxy. In what follows simulated trips will be done on the real motorized car network at the city-level including more than 12,000 road sections. The results shall bring a first answer to that dilemma.

To conclude, we have shown that economists have recently improved the intrinsic properties of distance-based methods to gauge agglomeration more precisely. At the same time, geographers progress on the apprehension of the real distance. In the next section, we reconcile both approaches by retaining economists and geographers' advances to test the robustness of the results of the geographic concentration of economic activities in non-homogeneous space. As we previously underlined, estimates obtained on an unconstrained network (with Euclidean distances) will be confronted to a constrained one (by considering motorized vehicles distances).

3. DATA

Several databases are used to estimate the real car-accessibility of retail stores in Lyon-Villeurbanne. In the following sub-sections, we explain the construction of the network in the Lyon metropolitan area and we present data on retail stores.

3.1. URBAN NETWORK

Simulations are made by MOSART (© LET), a powerful decision making tool for private and public urban mobility authorities.[5] The main objective of this project is measuring and viewing services levels offered by different transport networks in the whole Lyon metropolitan area. The road data is highly detailed: more than 12,000 road sections are included on Lyon-Villeurbanne. Every type of communication routes is taken into account (from alleyways to highways) and traffic flows are simulated considering network constraints (one/two-way streets, pedestrian streets…). Retail stores are localized at their exact position. They are then fixed to the road network by a connector. 10,167 connectors are activated on the studied area. In our analysis, we only simulate movements for motorized vehicles. Any travel by car can be simulated from any starting point in the Lyon metropolitan area. The shortest path between two retail stores is chosen and every movement is estimated by its length (in meters).

3.2. RETAIL STORES

The database comes from the Chamber of Commerce and Industry of Lyon. It registers the exact geographic position of all stores in Lyon-Villeurbanne for the year 2005. From this database, we retain individual information on the industrial classification and the precise location of stores (given by the Lambert coordinates of the French projection system). The focus is made on the spatial distribution of the non-eating retail trade industry. The database is composed of 3,389 non-eating retail shops belonging to 19 sectors (corresponding to the French industrial NAF classification 52.3 and 52.4). The description of the sectors and several summary statistics on the intra-sectoral distribution of stores are given in Table 1. Figure 1 illustrates the Lyon-

[5] MOSART (MOdelling and Simulation of Accessibility to netwoRks and Territories) is a numerical platform of modelling based on a traffic allocation model (VISUM from PTV), a highly detailed network (NAVTEQ) and GIS analysis and databases.

Villeurbanne area where the non-eating retail stores are located. There is a high density of shops in the peninsula especially near the City Hall. The relief is particularly interesting in terms of network constraints because the Rhône and Saône rivers cross the area and two hills overhang the city of Lyon (namely the "Croix-Rousse" in the north of the City Hall and "Fourvière" in the west of the peninsula).

Figure 1 : Study area: Lyon-Villeurbanne

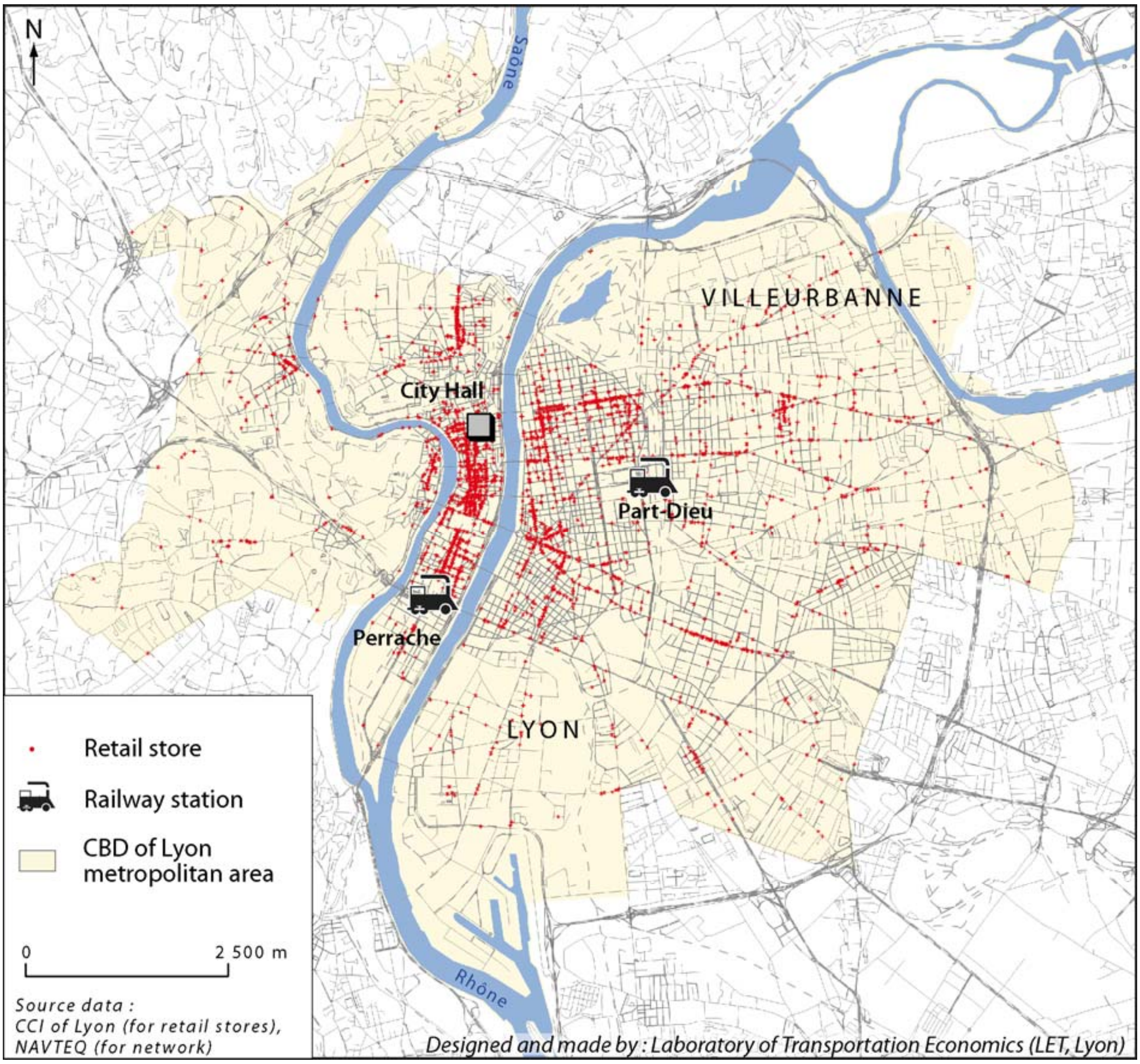


Table 1 : Summary statistics on the 19 non-eating sectors in the Lyon-Villeurbanne area

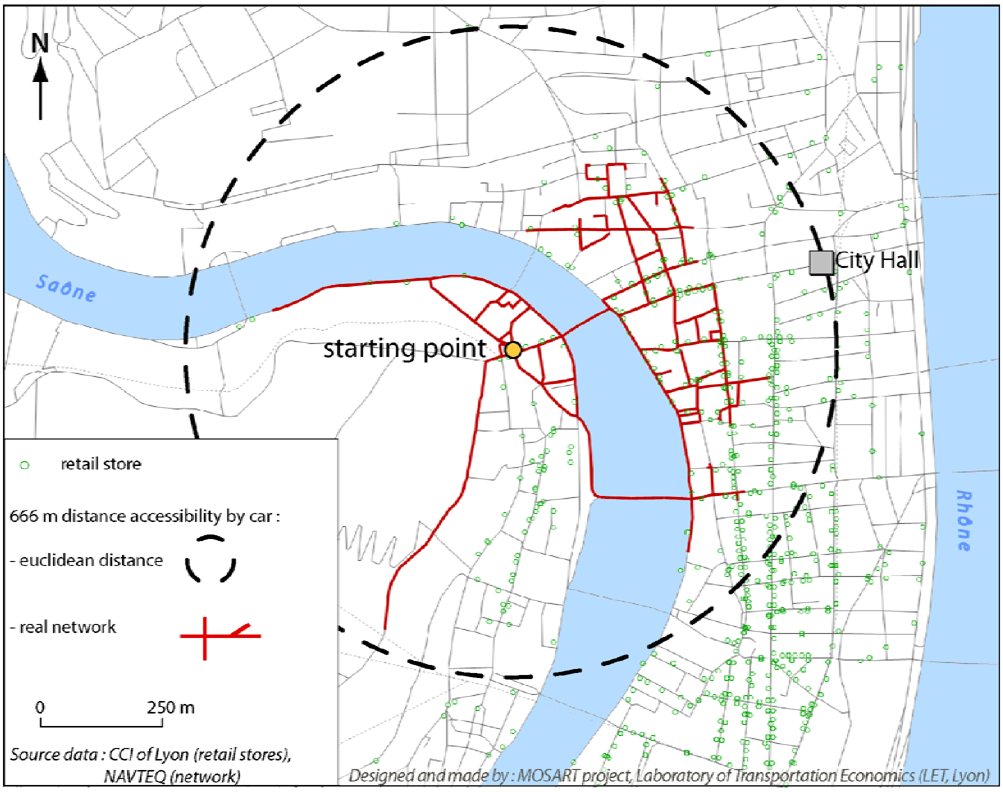
NAF Code	Sector	Nb. stores	Intra-sectoral distance (meters)			
			Median		Mean	
			Eucl.	Car	Eucl.	Car
52.3A	Dispensing chemists	220	2,974	3,856	3,161	4,106
52.3C	Retail sale of medical and orthopaedic goods	71	2,986	3,911	3,106	3,990
52.3E	Retail sale of cosmetic and toilet articles	111	2,293	3,037	2,472	3,233
52.4A	Retail sale of textiles	100	1,685	2,417	2,053	2,756
52.4C	Retail sale of clothing	927	1,567	2,156	1,758	2,370
52.4E	Retail sale of footwear	169	1,598	2,205	1,836	2,440
52.4F	Retail sale of fine leather goods and of travel articles	38	1,687	2,250	1,701	2,304
52.4H	Retail sale of furniture	146	1,896	2,515	2,142	2,753
52.4J	Retail sale of household equipment and articles	199	1,971	2,741	2,324	3,123
52.4L	Retail sale of electrical household appliances and radio and television goods	173	2,513	3,347	2,776	3,596
52.4N	Retail sale of hardware	62	2,762	3,540	2,915	3,804
52.4P	Retail sale of do-it-yourself (DIY)	5	688	878	804	1,008
52.4R	Retail sale of books, newspapers and stationery	229	2,429	3,237	2,659	3,484
52.4T	Retail sale of optics and photography	143	2405	3,152	2,572	3,369
52.4U	Retail sale of floor-covering and wall-covering	22	1,758	2,388	2,206	2,804
52.4V	Retail sale of clocks and jewellery	122	1,747	2,424	2,012	2,670
52.4W	Retail sale of sport and leisure articles	108	2,180	2,939	2,377	3,160
52.4X	Retail sale of flowers	174	2,850	3,724	3,025	3,945
52.4Z	Sundry retail sale in specialized stores	370	2,086	2,805	2,355	3,085

4. HOW MUCH DISTANCE-NETWORK REALLY MATTERS?

4.1. AN EMPIRICAL EXAMPLE FOR A MOTORIZED VEHICULE TRIP

Let us start with a simple example to shed the light on the relevance of the distance definition. Figure 2 zooms on the City Hall area in Lyon. The urban network can easily be recognized: dotted lines show pedestrian streets or footbridges and plain black lines indicate roads or bridges authorized for cars and urban transportation. As we pointed out (section 3.1), car-travel simulations take into account the direction of streets even if it is not shown in Figure 2. Non-eating retail shops are represented by green circles (whatever the sector) and they are located at their exact position on the network.

Figure 2: Example of a 666m car-accessibility from a given shop



Case 1- We firstly simulate all travels of 666 meters that can be reached by car from a starting point located in the west of the peninsula (at the bottom of Fourvière hill). These routes are indicated in red in Figure 2. One can easily observe that all feasible trips on the car-network do not draw a disc centered on the starting point. This result somewhat intuitive is not without consequences. On one hand, the network constraints consistently limit the number of paths accessible by car. For example, the south area of the point of departure is unapproachable by car

("Vieux Lyon" pedestrian area). As a result, even if shops are closely located to the starting point, they could be inaccessible by taking into account network constraints. On the other hand, natural advantages seem potentially important to the stores accessibility. From the starting point for instance the only possibility to have an access to the peninsula is to go over a bridge authorized for motorized cars.

Case 2- Let us now consider all accessible stores in a radius equal to 666 meters from the starting point. This case is illustrated in Figure 2 by all shops located inside the dotted black disk. The comparison of stores accessibility clearly emphasized important differences considering the car-travel distance (case 1) or the Euclidean distance (case 2). Retaining the Euclidean distance leads to an obvious overestimation of the actual number of stores that can be reached from the starting point. This leads to a first observation: ***the number of shops (absolute concentration) will be higher in Case 2 because the Euclidean distance is unconstrained. In other words, the Euclidean metric goes "too far" than the real accessibility permits.*** However, as we underlined in section 2.1, one criterion for evaluating the geographic industrial concentration states that a *relative* concentration measure is more pertinent. It is very difficult to assess if that overestimation still remains to the naked eye on a simple map of activities because firstly there are numerous shops belonging to various sectors located in the city center and secondly network constraints are the same whatever the sector. The analysis given in the following section provides the answer.

4.2. EMPIRICAL ROBUTESS OF DISTANCE-BASED MEASURES

We compute the *M* function (as defined in the section 2.1.) to evaluate the relative concentration of the 19 non-eating activities in Lyon-Villeurbanne area. In a first step, Euclidean bilateral distances between each pair of shops are retained. Then, we consider the shortest path by car between every pair of establishments respecting car-network constraints. The *M* function has been computed every 100 meters up to the median distance between any two shops on the studied area (2,766 meters) as recommended by Duranton and Overman (2005).[6] The significance of the results is given by the construction of a confidence interval of the null hypothesis. As we previously underlined, the global tendency of the non-eating retail stores location constitutes the benchmark distribution against which the spatial structure of the sub-

[6] Results of the *M* functions are shown from a 200m radius and not 100m because the average length of the connectors that join up stores to the urban network is 47m. This may induce an imprecision at the first radius (100m). Estimates are thus given from the second step of computation (200m).

sectors are compared to. To investigate the significance of the results of the *M* function for this sector, the Monte-Carlo method is used. We carry out as follows. Consider a particular non eating retail activity. A series of random and independent distributions of stores (included in our database) is generated on the actual locations of stores. Then we choose a threshold (α) and we define a (1-α) confidence interval of the *M* function associated for each distance. In the paper, the confidence interval is computed by 1,000 simulations at a 95% confidence level. We only investigate location pattern of sectors that present a significant departure from randomness (concentration or dispersion). The comparison of both Euclidean and car-*M* plots per sector provides a series of interesting results.

OBSERVATION 1: EUCLIDEAN DISTANCE UNDERESTIMATES THE ACTUAL RELATIVE INDUSTRIAL CONCENTRATION OR DISPERSION

Eight sectors show a significant geographic concentration for both definitions of distances: the retail sales of cosmetic and toilet articles (52.3E), clothing (52.4C), footwear (52.4E), furniture (52.4H), electrical household appliances and radio and television goods (52.4L), hardware (52.4N), flowers (52.4X) and specialized stores (52.4Z). One sector depicts a mix pattern of localization: dispensing chemists (52.3A) are significantly dispersed at short distances and concentrated at larger distances. *M* plots for several sectors are given in the following figures.[7]

Figure 3 : M plots for furniture (52.4H)

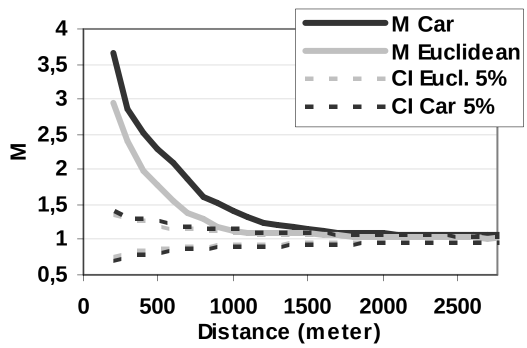
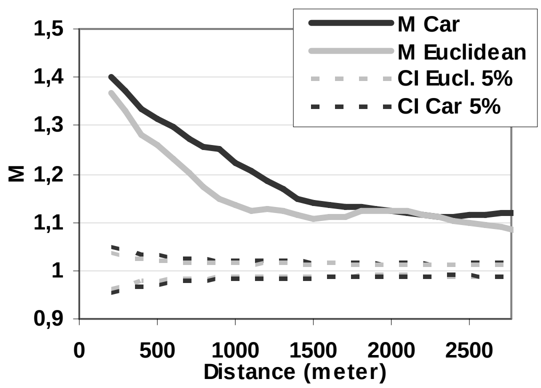


Figure 4 : M plots for clothing (52.4C)



7 In order to save space, *M* plots for the other sectors are given in the appendix.

Figure 5 : M plots for cosmetic and toilet articles (52.3E)

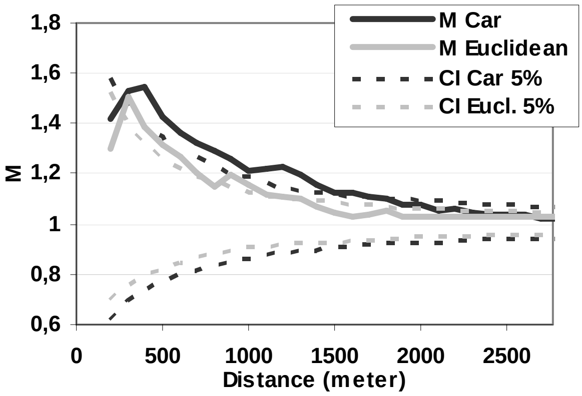
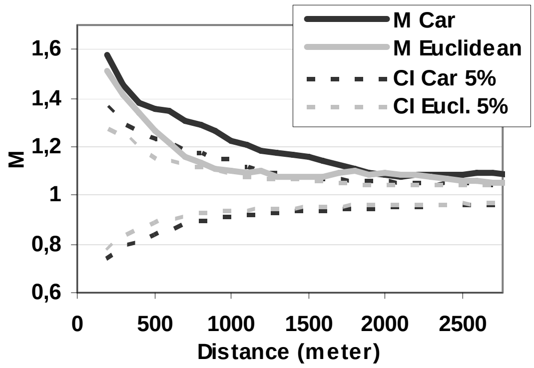
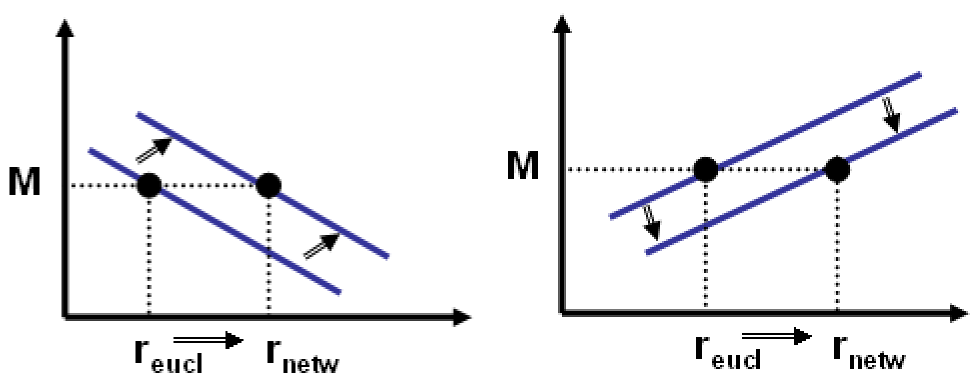


Figure 6 : M plots for footwear (52.4E)



What can we learn from these examples? In Figure 2, we underlined that in absolute terms, the Euclidean distance overestimates the real number of accessible stores on the urban network. This result is in line with Yamada and Thill (2004) where they highlight that Ripley's *K* function increases the over detection of clustering patterns. However, **in relative terms for all sectors that present a significant pattern of location, the Euclidean-*M* plot *systematically* indicates lower level of agglomeration.** Graphically Euclidean *M*-plots are below car-*M* plots whatever the scale of observation. **In other words, *M* plots reveal that the Euclidean metric *systematically underestimates* the real level of the relative concentration or dispersion phenomenon.** The explanation is straightforward. As we previously underlined, the Euclidean metric could be understood as an "unconstrained distance" (straight line) unlike the network path that respects at the same time the network structure and the feasible movements over it. For a given travel, the Euclidean distance is thus superior or equal to the one over the network. As a result, switching from the Euclidean distance to the network one implies that the Euclidean distance has to be multiplied by a factor superior or equal to one. In Figure 7, this is illustrated for two simple theoretical *M* plots: a decreasing one and an increasing one.

Figure 7: Two theoretical *M* plots



Switching from a given Euclidean distance (r_{eucl}) to its corresponding network distance (r_{netw}) implies a translation of the plots on the right. In case of a positive slope of the *M*-plot, the value of *M* will be lower if the network distance is used. On the opposite, the value of *M* will be greater if the network distance is employed in case of a negative trend of the *M*-plot. Excepting for the Dispensing chemists sector (52.3A) shown in Figure 21, significant *M*-plots present only one main trend. A negative one is observed for agglomeration whereas a positive slope is shown for dispersion.[8] In both cases and, as expected, this leads to an underestimation of the actual concentration or dispersion if the Euclidean distance is used. Moreover, the fact that the Euclidean *M*-results underestimate the actual level of agglomeration or dispersion of stores reveals that the real stores-accessibility is better captured by distance-based methods that take into account the network. This result is important for economic issues for instance if one want to gauge the determinants of agglomeration of stores. Our results support that urban space has not to be considered as homogeneous (as the Euclidean distance does): the urban network tends to "distort" the territory and in consequence may have an influence on the geographic concentration appraisal.

Finally, for the non-retail stores agglomeration on Lyon-Villeurbanne, *M* estimates do not generally greatly differ from one or the other definition of distance. This encourages us to calculate approximately differences in the results according to the distances used.

<u>OBSERVATION 2:</u> DIFFERENCES IN ESTIMATES ARE ROUGHLY EXPLAINED BY A FACTOR EQUAL TO $\sqrt{2}$.

Is it possible to find a unique correction factor to switch from the Euclidean to the network distance? The answer is far from being easy.[9] Several features may create a gap in the estimates. First of all, the configuration of the network has a decisive impact on the results. To make things clearer, consider two stores. If they are positioned at two opposite angles of a square-block or a rectangular-block, the factor correction will not be the same. Additionally, in the real life, things may be worse because the actual network could be partly composed of a mix of square, rectangular, circum-radial and other grid-structures! The second reason that could explain a gap between the Euclidean and network distances is the existence of constrained movements on the

[8] Firstly because *M* is a cumulative function so positive/negative local variations may exist but they never reverse the main decreasing or increasing trend of the curve. Secondly, the more the distance (r) raises the nearer *M* curve from the benchmark value (1).

[9] There is an important literature on the best way to estimate distances on the network with mathematical functions (see for example Perreur and Thisse, 1974; Love and Morris, 1979; 1988; Love and Walker, 1994; Brimberg *et al.,* 1995, 2007).

network (one way street for example). Let us consider two stores on the network and the shortest path to link them. Due to movements constraints for joining both stores, imagine that a detour is compulsory. For example, the trip begins on the opposite direction of the target (say the other store) in order to take an authorized street. This introduces a gap in the inter-stores distance measurement using the straight line (Euclidean distance) or the network-based travel. As a consequence, the correction factor is difficult to estimate.

After proceeding by trial and error, our estimations give some support that car-travel estimations are well-explained by increasing the Euclidean distance by a correction of $\sqrt{2}$. This "first-order factor" is somewhat a good correction on the network-configuration of the Lyon-Villeurbanne. It rests on the regularity of the urban network given by the Manhattan metric (all called rectilinear metric). The intuitive explanation is as follows. Consider two stores located on a network and the shortest path to join them. If the grid of the network is based on square-blocks, get though rectilinear distance to the diagonal distance implies a inflating correction factor of $\sqrt{2}$. For the nine retail activities that present a significant departure from randomness, both Euclidean and car-*M* plots and the one obtained by inflating the Euclidean distance by $\sqrt{2}$ are represented below.

Figure 8 : M plots for dispensing chemists (52.3A)

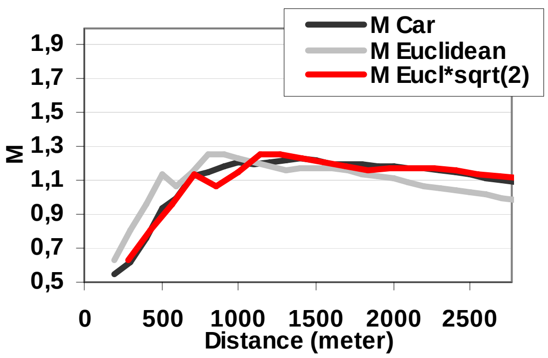


Figure 9 : M plots for cosmetic and toilet articles (52.3E)

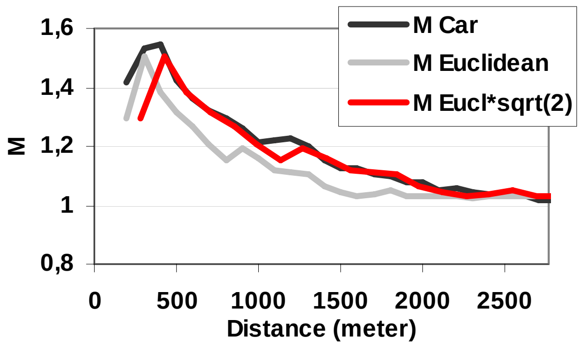


Figure 10 : M plots for clothing (52.4C)

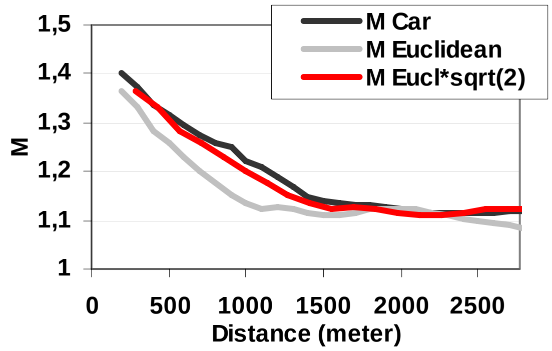


Figure 11 : M plots for footwear (52.4E)

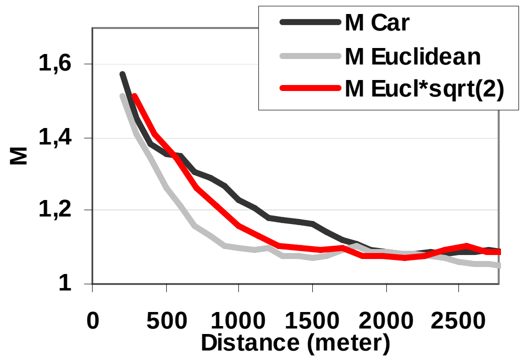


Figure 12 : M plots for furniture (52.4H)

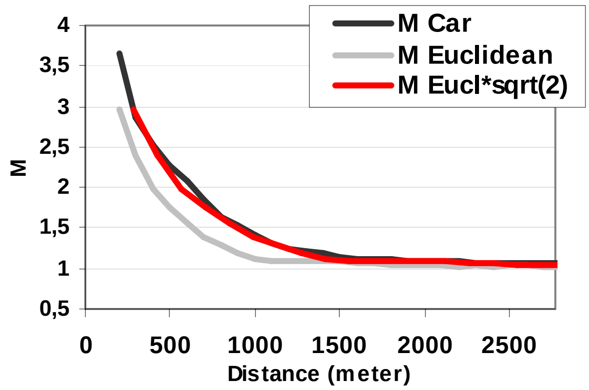


Figure 13 : M plots for electrical household appliances and radio and television goods (52.4L)

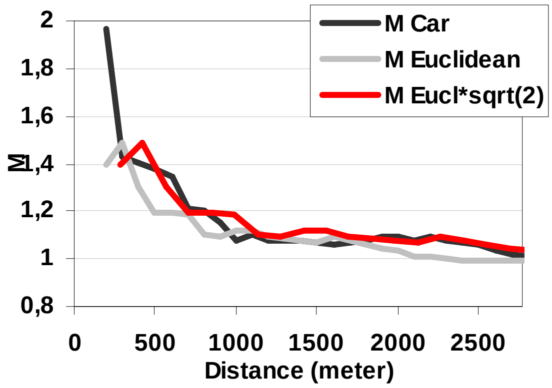


Figure 14 : M plots for hardware (52.4N)

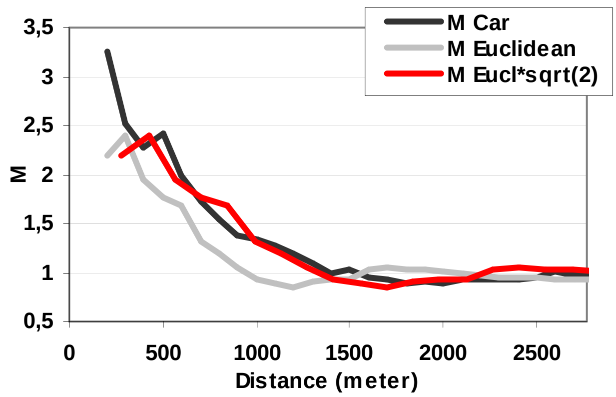


Figure 15 : M plots for flowers (52.4X)

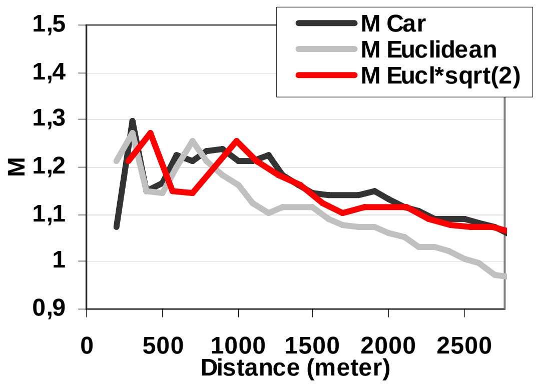
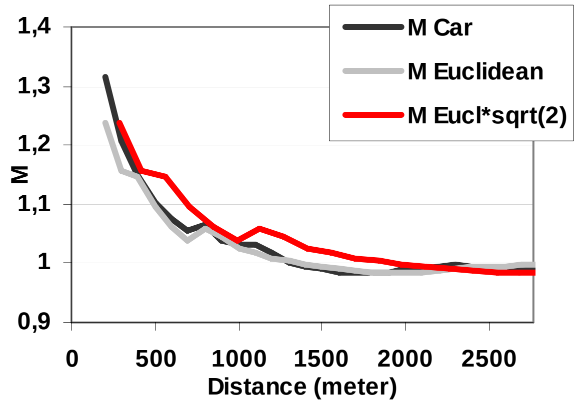


Figure 16 : M plots for specialized stores (52.4Z)



As we can observe, for a great number of activities, the *M*-car plot and the one obtained with the inflating Euclidean distance meet more or less. Cases of dispensing chemists (52.3A), clothing (52.4C), cosmetic and toilet articles (52.3E), furniture (52.4H) or hardware (52.4N) are particularly striking. It is noteworthy that since there is no difference in the estimates of *M*-car of *M*-Euclidean for specialized stores (52.4Z), this operation was unsuitable. In that case, the proposed correction (if it is done) introduced an overestimation of the results.

An insight to the deviations of two other specific cases is not without interest. The first one is given by the *M* plots for footwear (52.4E, see Figure 11). The $\sqrt{2}$ correction factor undoubtedly improves the Euclidean geographic concentration results. However on the interval 600m-1700m, the correction is not perfect. This sector has indeed a very specific location pattern because their retail stores are highly localized on the peninsula (see Figure 17). One can expect that rivers constraints may have a severe impact on estimates. To test this, our strategy is not to change the actual location of stores but the car urban network. On this modified network, footbridges can be taken by cars, bridges can be crossed in both ways and one more bridge was added in the South of the urban center to increase the accessibility to the peninsula in that area. The new *M*-car is called "*M* all bridges" in Figure 18. The bridges effect on the agglomeration evaluation is clearly visible and, explains to a certain extent the observed gap between M plots on the 600m-1700m interval. This sector really constitutes an exception because this modification of the network does not generally have a significant incidence on the estimates.

Figure 17 : Location of footwear retail stores in the Lyon Villeurbanne area.

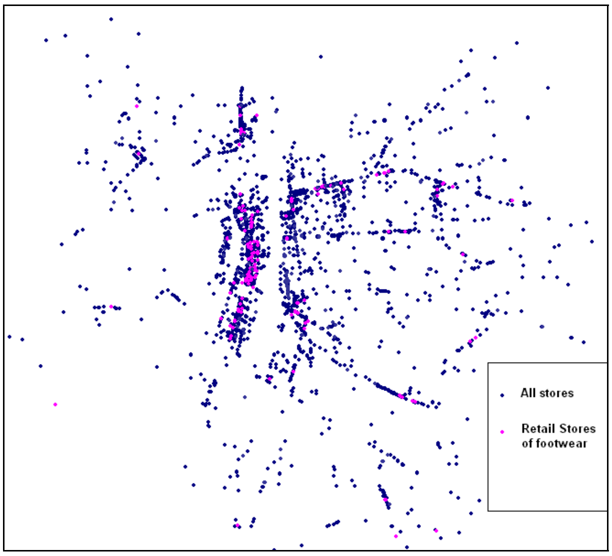
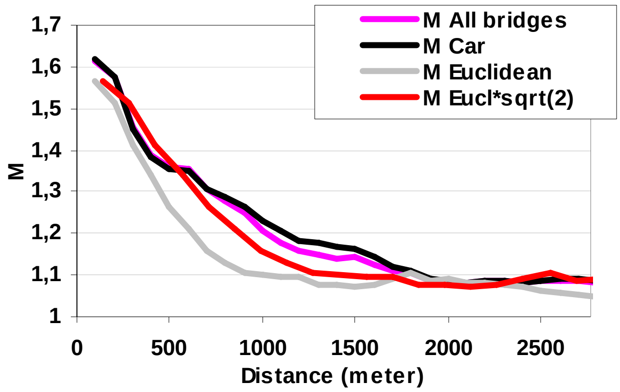
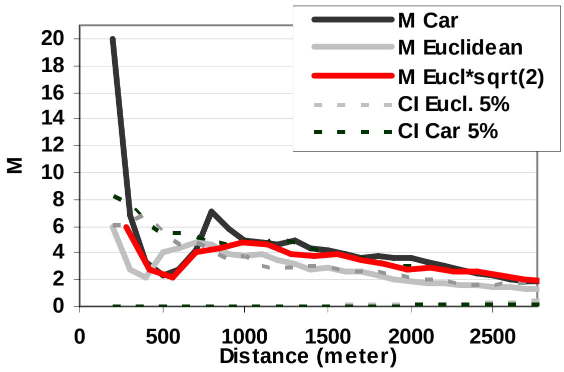


Figure 18 : M plots for footwear (52.4E)



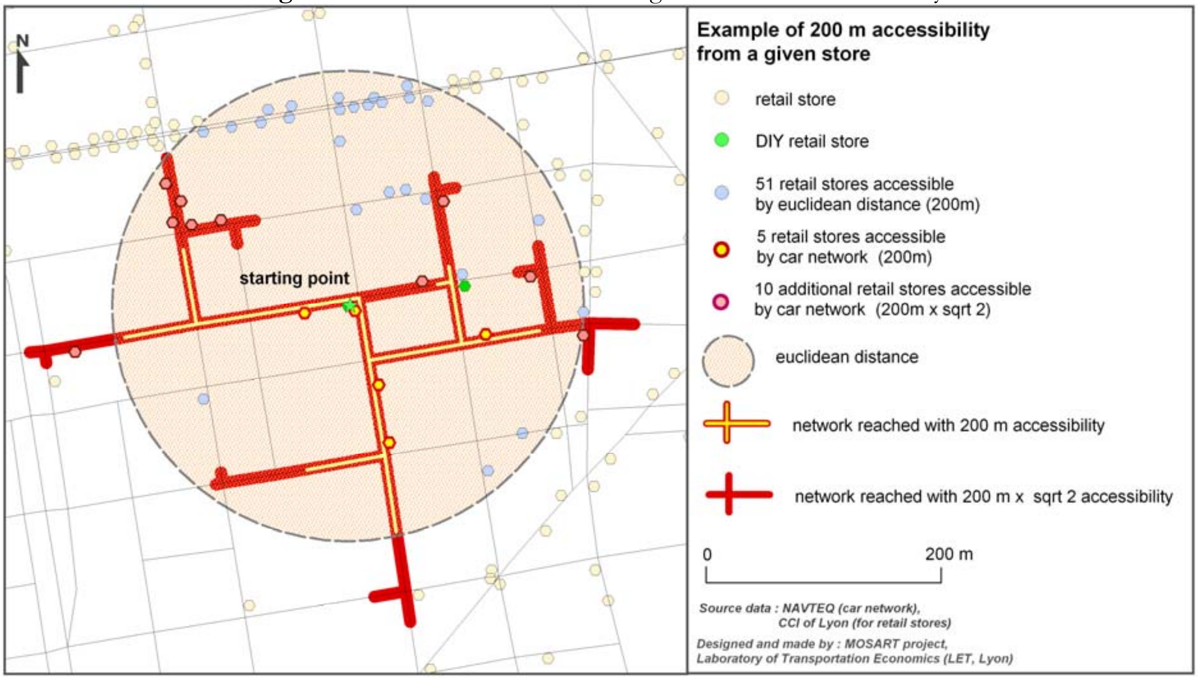
The second case explored in details is the DIY retail stores (52.4P). This sector has the particularity to present non-significant results of concentration for almost all ranges of distances excepting for the car *M*-plot at very low distances (Figure 19). At a radius of 200m, a huge deviation in the estimates can be observed: the Euclidean *M* plot reaches 6 but is non-significant and the car-*M* plot is highly significant and reaches an approximate value of 20.

Figure 19 : M plots for DIY stores (52.4P)



This gap can be easily explained by a detailed analysis of the data; the number of DIY stores (5) allows such a work. We focus on one particular DIY store (green star in Figure 20) located on the study area. Within a 200m-accessibility by car or "as the crow flies", the number of accessible DIY stores is the same (green stores). However, the accessibility to the non-DIY stores greatly differs (comparison of blue and yellow stores). In relative terms, the level of concentration will consequently be markedly higher for the car travel in comparison to the Euclidean one. The *M* function, defined as the sectoral averaged value of individual relative concentration, perfectly captures that phenomenon. As we can show in Figure 19, the Euclidean-*M* plot is below the car-M plot at this distance. In Figure 20, we additionally represent all feasible car-travels within a 283m-distance $(\approx 200\text{m} \times \sqrt{2})$. All things being equal, the value of the car-*M* plot should decrease since (i) the number of non-DIY stores increases and (ii) the number of DIY-shops is the same. This is confirmed by our findings (Figure 19).

Figure 20 : Zoom on the location on given DIY store accessibility



To sum up, the analysis provided on our on data shows that a $\sqrt{2}$ correction factor generally improves the accuracy in estimating the actual travel car distances. However, several important warnings were underlined:

- Firstly, the unique proposed correction factor (for all sectors and distances) has to be considered in better cases as a "rough correction factor". In the Lyon-Villeurbanne area, the urban streets configuration is not a completely square-based network (existence of ring roads for example).
- Secondly, as the deviation from Euclidean metric to car-travel metric can not be systematically distinguished, an empirical investigation of the agglomeration degree seems compulsory to be able to distinguish sectors for which differences in estimates remain.
- Thirdly, a particular attention has to be drawn on any potential highly specific location pattern of activities. We illustrated that by two examples: the DIY shops and the retail stores of footwear.

5. CONCLUSION

In this paper, we evaluate the importance of the notion of the distance used in the analysis of the geographic concentration of activities. We provide evidence that the definition matters at an urban scale if one wants to evaluate the agglomeration phenomenon thanks to distance-based methods. Nonetheless, a raw correction factor was proposed to switch from the Euclidian metric to the actual travel-car metric in the Lyon-Villeurbanne area. Our results certainly pave the way for further investigations. On one hand, our findings could be confronted to other studies that replicate our work in other cities (possibly with different urban networks). On the other hand, there are a possible number of extensions of our approach, for example, by taking into account the inter-stores travel time-distance.

REFERENCES

Arbia G. and G. Espa, 1996. Statistica economica territoriale. Cedam, Padua.

Arbia G., Espa G. and D. Quah, 2008. A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. Empirical Economics 34: 81-103.

Arbia G., Espa G., Giuliani D. and A. Mazzitelli. Detecting the existence of space–time clustering of firms, Regional Science and Urban Economics, forthcoming.

Barff, R.A., 1987. Industrial Clustering and the Organization of Production: A Point Pattern Analysis of Manufacturing in Cincinnati, Ohio. Annals of the Association of American Geographers 77: 89-103.

Brimberg J., Love R.F. and J.H. Walker (1995). The effect of axis rotation on distance estimation. European Journal of Operation Research 80: 357-364.

Brimberg J., Love R.F. and J.H. Walker (2007). Estimations of travel distances with the weighted l_p norm: Some empirical results. Journal of Transport Geography 15: 67-72.

Brülhart M. and R. Traeger (2005). An Account of Geographic Concentration Patterns in Europe." Regional Science and Urban Economics 35: 597-624.

Combes P.-P. and M. Lafourcade, 2005. Transport Costs: Measures, Determinants and Regional Policy Implications for France, Journal of Economic Geography 5, 319-349.

Combes P.-P, Mayer T. and J.-F. Thisse, 2008. Economic Geography: The Integration of Regions and Nations, Princeton University Press.

Combes P.-P. and H.G. Overman, 2004. The spatial distribution of economic activities in the European Union. in: J.V. Henderson, and J.-F. Thisse, (Eds.), Handbook of Urban and Regional Economics, Elsevier. North Holland, Amsterdam.

Diggle P.J., 1983. Statistical analysis of spatial point patterns. Academic Press, London.

Duranton G. and H.G. Overman, 2005. Testing for Localisation Using Micro-Geographic Data. Review of Economic Studies 72: 1077-1106.

Duranton G. and H.G. Overman, 2008. Exploring the Detailed Location Patterns of UK Manufacturing Industries using Microgeographic Data. Journal of Regional Science 48: 213-243.

Ellison G. and E.L. Glaeser, 1997. Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. Journal of Political Economy 105: 889-927.

Ellison G., Glaeser E.L. and W.R. Kerr. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. The American Economic Review, *forthcoming*.

Fratesi U., 2008. Issues in the measurement of localization. Environment and Planning A 40: 733-758.

Haaland J.I., Kind H.J., Midelfart-Knarvik K.H. and J. Torstensson (1999). What determines the economic geography of Europe? Centre for Economic Policy Research. Discussion paper 2072.

Jaffe M., Tratjenberg and R. Henderson, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations, Quarterly Journal of Economics 108: 577-598.

Jensen P., 2006. Network-based predictions of retail store commercial categories and optimal locations, Physical review. E 74(2) R035101.

Jensen P. and J. Michel. Measuring spatial dispersion: exact results on the variance of random spatial distributions. Annals of Regional Science, *forthcoming.*

Love R.F. and J. Morris, 1979. Mathematical models of road travel distances. Management Science 25: 130-139.

Love R.F. and J. Morris, 1988. On estimating road distances by mathematical functions. European Journal of Operational Research 36: 251-253.

Love R.F. and J.H. Walker, 1994. An empirical comparison of block and round norms for modelling actual distances. Location Science 2: 21-43.

Marcon E. and F. Puech, 2003. Evaluating the Geographic Concentration of Industries Using Distance-Based Methods. Journal of Economic Geography 3: 409-428.

Marcon, E. and F. Puech. Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods. Journal of Economic Geography, *forthcoming.*

Ó hUallacháin B. and T.F. Leslie, 2007, Producer Services in the Urban Core and Suburbs of Phoenix, Arizona. Urban Studies 44: 1581-1601.

Okabe A. and I. Yamada, 2001. The K-Function Method on a Network abd Its Computational Implementation, Geographical Analysis 33: 271-289.

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

Ripley B.D., 1976. The Second-Order Analysis of Stationary Point Processes. Journal of Applied Probability 13: 255-266.

Ripley B.D., 1977. Modelling Spatial Patterns. Journal of the Royal Statistical Society B 39: 172-212.

Rosenthal S.S. and W.C. Strange, 2001. The Determinants of Agglomeration. Journal of Urban Economics 50, 191-229.

Rosenthal S.S. and W.C. Strange, 2003. Geography, Industrial Organisation, and Agglomeration. Review of Economics and Statistics 85: 377-393.

Sweeney S.H. and E.J. Feser, 1998. Plant Size and Clustering of Manufacturing Activity. Geographical Analysis 30: 45-64.

SYTRAL, 2007. Enquête ménages déplacements – Principaux résultats pour l'agglomération lyonnaise, Fiche technique Conférences des maires « Lyon-Villeurbanne », 16p. Available: http://www.sytral.fr/fileadmin/template/images/user_upload/PUBLICATIONS/INSTITUTION/EMD/EMD_CM_LyonVilleurbanne.pdf

Yamada I. and J.-C. Thill, 2004. Comparison of planar and network K-functions in traffic accident analysis, Journal of Transport Geography 12: 149-158**.**

APPENDIX

Note: M plots of the Retail sale of cosmetic and toilet articles (52.3E), clothing (52.4C), footwear (52.4E), furniture (52.4H) and the DIY stores are given in the text. They are not reproduced in the Appendix.

Figure 21 : M plots Dispensing chemists (52.3A)

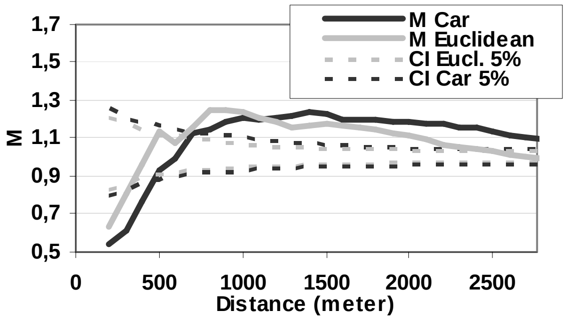


Figure 22 : M plots Retail sale of medical and orthopaedic goods (52.3C)

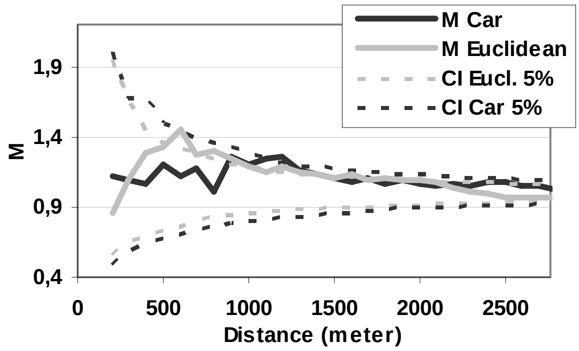


Figure 23 : M plots for Retail sale of textiles (52.4A)

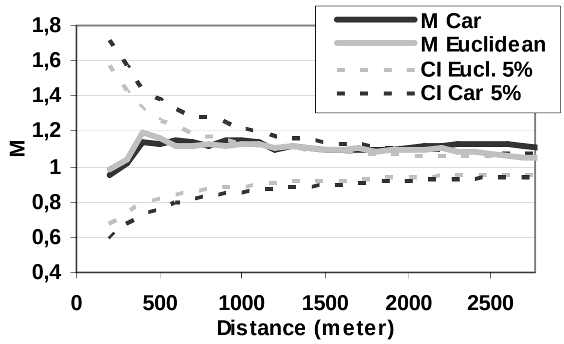


Figure 24 : M plots for Retail sale of fine leather goods and of travel articles (52.4F)

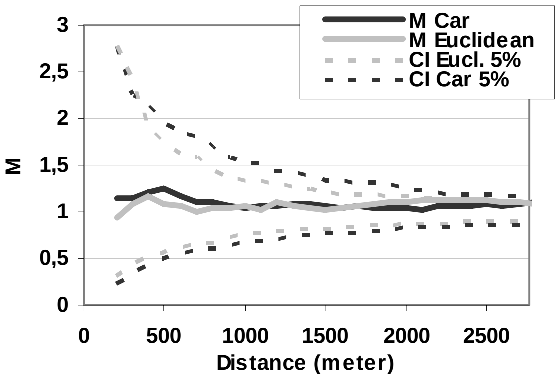


Figure 25 : M plots for Retail sale of household equipment and articles (52.4J)

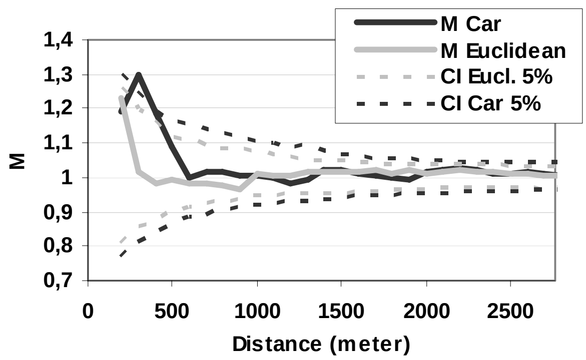


Figure 26 : M plots for Retail sale of electrical household appliances and radio and television goods (52.4L)

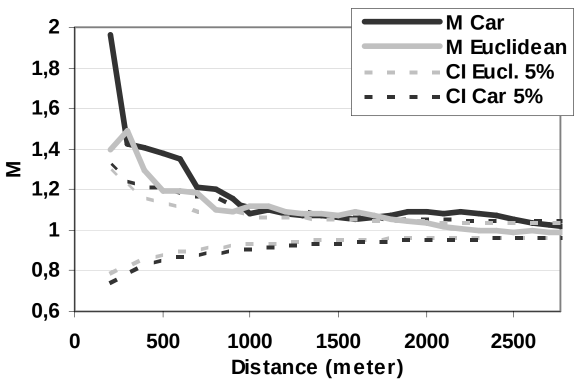


Figure 27 : M plots for Retail sale of hardware (52.4N)

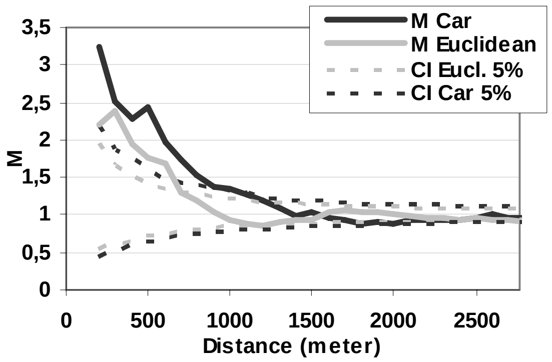


Figure 28 : M plots for Retail sale of books, newspapers and stationery (52.4R)

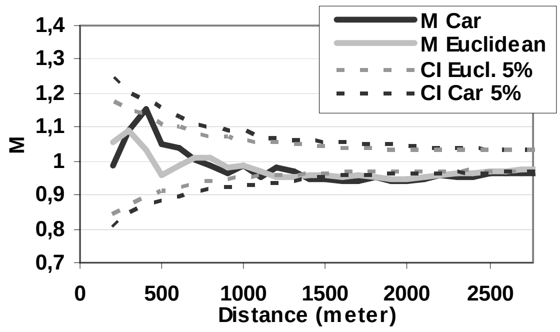


Figure 29 : M plots for Retail sale of books, newspapers and stationery (52.4R)

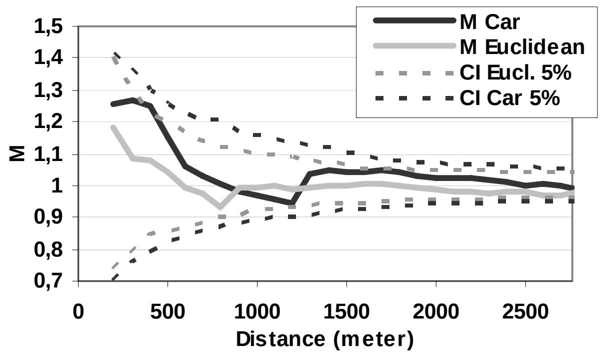


Figure 30 : M plots for Retail sale of floor-covering and wall-covering (52.4U)

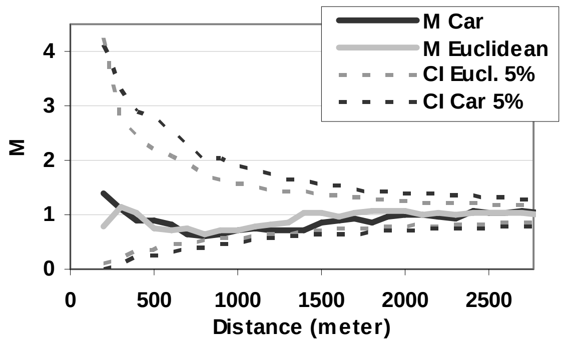


Figure 31 : M plots for Retail sale of clocks and jewellery (52.4V)

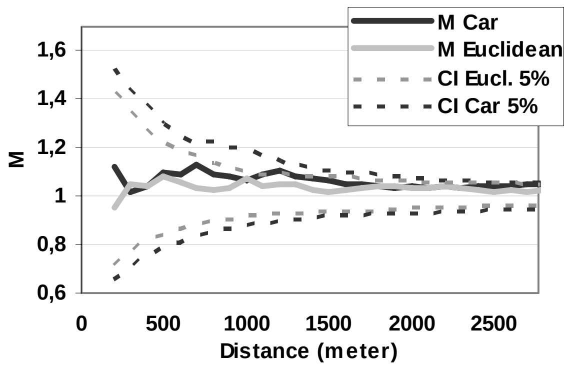


Figure 32 : M plots for Retail sale of sport and leisure articles (52.4W)

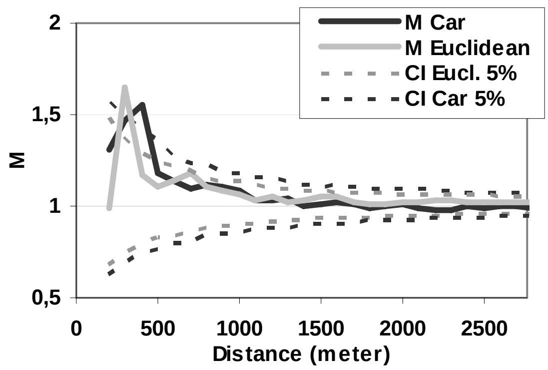


Figure 33 : M plots for Retail sale of flowers (52.4X)

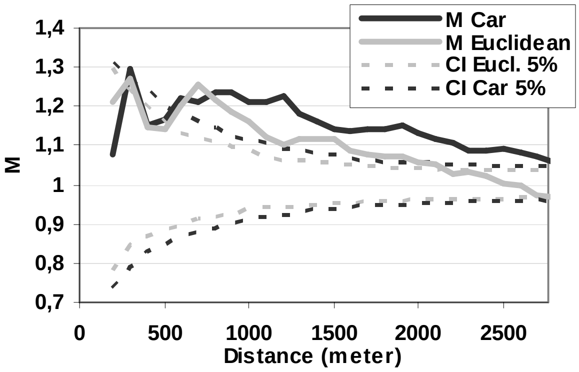


Figure 34 : M plots for Sundry retail sale in specialized stores(52.4Z)