PHYSICAL REVIEW E **74**, 035101(R) (2006)

Network-based predictions of retail store commercial categories and optimal locations

Pablo Jensen
Laboratoire de Physique, CNRS UMR 5672, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France and Laboratoire d'Economie des Transports, CNRS UMR 5593, ISH-Université Lyon-2, 14, Avenue Berthelot, 69007 Lyon, France
(Received 3 March 2006; published 8 September 2006)

I study the spatial organization of retail commercial activities. These are organized in a network comprising "antilinks," i.e., links of negative weight. From pure location data, network analysis leads to a community structure that closely follows the commercial classification of the U.S. Department of Labor. The interaction network allows one to build a "quality" index of optimal location niches for stores, which has been empirically tested.

DOI: 10.1103/PhysRevE.74.035101 PACS number(s): 89.65.Lm, 05.65.+b, 89.75.Hc

Walking in any big city reveals the extreme diversity of retail store location patterns. Figure 1 shows a map of the city of Lyon (France) including all the drugstores, shoe stores, and furniture stores. A qualitative commercial organization is visible in this map: shoe stores aggregate at the town shopping center, while furniture stores are partially dispersed on secondary poles and drugstores are strongly dispersed across the whole town. Understanding this kind of feature and, more generally, the commercial logic of the spatial distribution of retail stores, seems a complex task. Many factors could play important roles, arising from the distinct characteristics of the stores or the location sites. Stores differ by product sold, surface, number of employees, total sales per month, and inauguration date. Locations differ by price of space, local consumer characteristics, visibility (corner locations, for example), and accessibility. Only by taking into account most of these complex features of the retail world can we hope to understand the logic of store commercial strategies, let alone finding potentially interesting locations for new businesses.

Here I show that location data suffices to reveal many important facts about the commercial organization of retail trade [1]. First, I quantify the interactions among activities and group them using network analysis tools. I find a few homogeneous commercial categories for the 55 trades in Lyon, which closely match the usual commercial categories: personal services, home furniture, food stores, and apparel stores. Second, I introduce a quality indicator for the location of a given activity and empirically test its relevance. These results, obtained solely from location data, agree with the retailing "mantra": the three points that matter most in a retailer's world are location, location, and location.

Finding meaningful commercial categories. To analyze in detail the interactions of stores of different trades, I start from the spatial pair correlations. These functions are used to reveal store–store interactions, as atom–atom interactions are deduced from atomic distribution functions in materials science [2]. Tools from that discipline cannot be used directly, though, because the underlying space is not homogeneous: the density of consumers is not uniform and some town areas cannot host stores, as is clearly seen in the blank spaces of the map (due to the presence of rivers, parks, or residential spaces defined by town regulations).

A clever idea [3] is to take as reference a random distribution of stores located on the array of *all existing* sites (black dots in Fig. 1). This is the best way to take into account the geographical peculiarities of each town automatically. I then use the "M index" [4] to quantify the spatial interactions between categories of stores. The definition of M_{AB} at a given distance r is straightforward: draw a disk of radius r around each store of category A, count the total number of stores (n_{tot}), the number of B stores (n_B), and compare the ratio n_B/n_{tot} to the average ratio N_B/N_{tot}, where N refers to the total number of stores in town. If this ratio, averaged over all A stores [5], is larger than 1, this means that A "attracts" B, otherwise that there is repulsion between these two activities [6]. I have chosen $r=100m$ as this represents a typical distance a customer accepts to walk to visit different stores [7].

I can now define a network structure of retail stores. The nodes are the 55 retail activities (Table I). The weighted [8] links are given by $a_{AB}\equiv\ln(M_{AB})$, which reveal the spatial attraction or repulsion between activities A and B [9]. This retail network represents first a social network with quantified "antilinks," i.e., repulsive links between nodes [10]. The antilinks add to the usual (positive) links and to the absence of any significant link, forming an essential part of the network. If only positive links are used, the analysis leads to

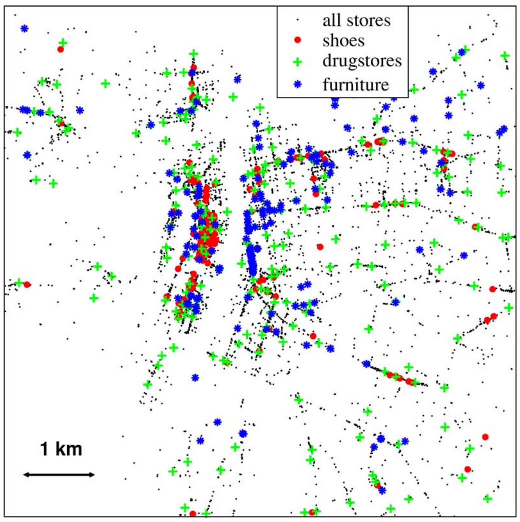

FIG. 1. (Color online) Map of Lyon showing the location of all the retail stores, shoe stores, furniture dealers, and drugstores.

TABLE I. Retail store groups obtained from Pott's algorithm. Our groups closely match the categories of the U.S. Department of Labor Standard Industrial Classification (SIC) system. Group 1 corresponds to personal services, 2 to food stores, 3 to home furniture, 4 to apparel and accessory stores, and 5 to used merchandise stores. The columns correspond to group number, activity name, satisfaction, correlation with population density (U stands for uncorrelated, P for population correlated) and finally number of stores of that activity in Lyon. To save space, only activities with more than 50 stores are shown.

group	activity	s	pop. corr.	N_{stores}
1	bookstores and newspapers	1.00	U	250
1	repair of electronic household goods	0.71	P	54
1	make up, beauty treatment	0.68	P	255
1	hairdressers	0.67	P	844
1	power laundries	0.66	P	210
1	drug stores	0.55	P	235
1	bakery (from frozen bread)	0.54	P	93
2	other repair of personal goods	1.00	U	111
2	photographic studios	1.00	P	94
2	delicatessen	0.91	U	246
2	grocery (surface <120 m^2)	0.77	P	294
2	cakes	0.77	P	99
2	miscellaneous food stores	0.75	P	80
2	bread, cakes	0.70	U	56
2	tobacco products	0.70	P	162
2	hardware, paints (surface <400 m^2)	0.69	U	63
2	meat	0.64	P	244
2	flowers	0.58	P	200
2	retail bakeries (home made)	0.47	P	248
2	alcoholic and other beverages	0.17	U	67
3	computer	1.00	P	251
3	medical and orthopedic goods	1.00	U	63
3	sale and repair of motor vehicles	1.00	P	285
3	sport, fishing, camping goods	1.00	U	119
3	sale of motor vehicle accessories	0.67	U	54
3	furniture, household articles	0.62	U	172
3	household appliances	0.48	U	171
4	cosmetic and toilet articles	1.00	U	98
4	jewelry	1.00	U	230
4	shoes	1.00	U	178
4	textiles	1.00	U	103
4	watches, clocks, and jewelry	1.00	U	92
4	clothing	0.91	U	914
4	tableware	0.83	U	183
4	opticians	0.78	U	137
4	other retail sale in specialized stores	0.77	U	367
4	other personal services	0.41	U	92
4	repair of boots, shoes	−0.18	U	77
5	second-hand goods	0.97	U	410
5	framing, upholstery	0.81	U	135

different results, which are less satisfactory (see below).

To divide the store network into communities, I adapt Pott's algorithm [11]. This algorithm interprets the nodes as magnetic spins and groups them in several homogeneous magnetic domains to minimize the system energy. Antilinks can then be interpreted as antiferromagnetic interactions between the spins. Therefore, this algorithm naturally groups the activities that attract each other, and places trades that repel into different groups. A natural definition [11,12] of the satisfaction ($-1 \leq s_i \leq 1$) of site i to belong to group σ_i is

$$s_i \equiv \frac{\sum_{j \neq i} a_{ij} \pi_{\sigma_i \sigma_j}}{\sum_{j \neq i} |a_{ij}|}, \tag{1}$$

where $\pi_{\sigma_i \sigma_j} \equiv 1$ if $\sigma_i = \sigma_j$ and $\pi_{\sigma_i \sigma_j} \equiv -1$ if $\sigma_i \neq \sigma_j$.

To obtain the group structure, I run a standard simulated annealing algorithm [13] to maximize the overall site satisfaction:

$$K \equiv \sum_{i,j=1,55; i \neq j} a_{ij} \pi_{\sigma_i \sigma_j}. \tag{2}$$

Pott's algorithm divides the retail store network into five homogeneous groups (Table I, note that the number of groups is not fixed in advance but a variable of the maximization). This group division reaches a global satisfaction of 80% of the maximum K value and captures more than 90% of positive interactions inside groups. Except for one category ("repair of shoes"), our groups are communities in the strong sense of Ref. [12]. This means that the grouping achieves a positive satisfaction for every element of the group. This is remarkable since hundreds of "frustrated" triplets exist [14]. Taking into account only the positive links and using the modularity algorithm [15] leads to two large communities, whose commercial interpretation is less clear.

Two arguments ascertain the commercial relevance of this classification. First, the grouping closely follows the usual categories defined in commercial classifications, such as the U.S. Department of Labor Standard Industrial Classification System [16] (see Table I). It is remarkable that, starting exclusively from location data, one can recover most of such a significant commercial structure. Such a significant classification has also been found for Brussels and Marseilles stores (to be presented elsewhere), suggesting the universality of the classification for European towns. There are only a few exceptions, mostly nonfood proximity stores which belong to the "food store" group. Second, the different groups are homogeneous in relation to correlation with population density. The majority of stores from groups 1 and 2 (18 out of 26) locate according to population density, while most of the remaining stores (22 out of 29) ignore this characteristic [17]. Exceptions can be explained by the small number of stores or the strong heterogeneities [18] of those activities.

From interactions to location niches. Thanks to the quantification of retail store interactions, I can construct a mathematical index to automatically detect promising locations for retail stores. The basic idea is that a location that resembles the average location of the actual bakeries might well be a good location for a new bakery. To characterize the average environment of activity i, I use the average number of neighbor stores (inside a circle of radius 100 m) of all the activities j, thus obtaining the list of average $\overline{n_{ij}}$. I then use the network matrix a_{ij} to quantify deviations from this average. For example, if an environment lacks a bakery (or other shops that are usually repelled by bakeries), this should increase the suitability of that location. The quality $Q_i(x,y)$ of an environment around (x,y) for an activity i as

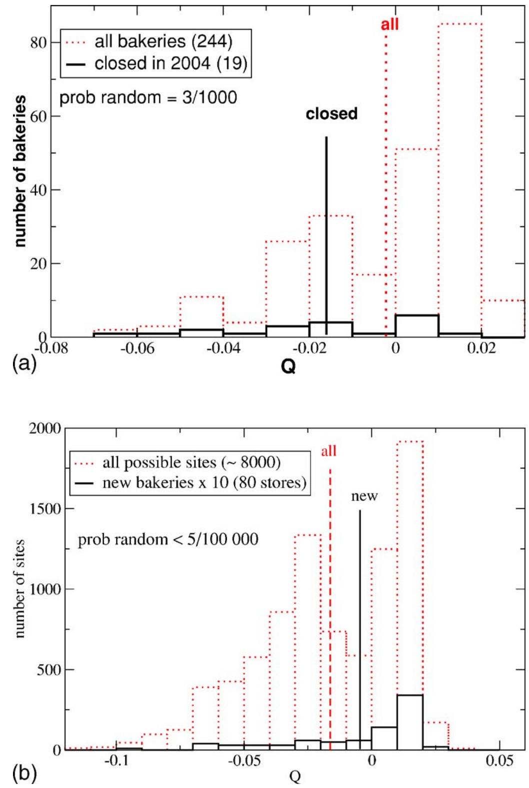

FIG. 2. (Color online) The landscape defined by the quality index is closely correlated to the location decisions of bakeries. (a) The 19 bakeries that closed between 2003 and 2005 had an average quality of -2.2×10^{-3} to be compared to the average of all bakeries (4.6×10^{-3}), the difference being significative with probability 0.997. Taking into account the small number of closed bakeries and the importance of many other factors in the closing decision (family problems, bad management, etc.), the sensitivity of the quality index is remarkable. (b) Concerning the 80 new bakeries in the 2005 database (20 truly new, the rest being an improvement of the database), their average quality is -6.8×10^{-4}, to be compared to the average quality of all possible sites in Lyon (-1.6×10^{-2}), a difference significant with a probability higher than 0.9999.

$$Q_i(x,y) \equiv \sum_{j=1,55} a_{ij}[n_{ij}(x,y) - \overline{n_{ij}}] \tag{3}$$

where $n_{ij}(x,y)$ represents the number of neighbor stores around x,y. To calculate the location quality for an existing store, one removes it from town and calculates Q at its location.

As often in social contexts, it is difficult to empirically test the relevance of our quality index. In principle, one should open several bakeries at different locations and test whether those located at the "best" places (as defined by Q) are on average more successful. Since it may be difficult to fund this kind of experiment, I use location data from two years, 2003 and 2005. It turns out (Fig. 2) that bakeries closed between these two years are located on significantly lower quality sites. Inversely, new bakeries (not present in the 2003 database) do locate preferently on better places than a random choice would dictate. This stresses the importance of location for bakeries, and the relevance of Q to quantify the interest of each possible site. Possibly, the correlation would be less satisfactory for retail activities whose locations are not so critical for commercial success. Practical applications of Q are under development together with Lyon's Chamber of Commerce and Industry: advice to newcomers on good locations, advice to city mayors on improving commercial opportunities on specific town sectors, etc. From a physicist's point of view, it would be interesting to define store-store interaction "potentials" by analogy with those used for atomic species. Finally, new tools are needed to describe networks containing antilinks, starting with a basic one: "how to define a node degree?"

---

[1] Christophe Baume and Frederic Miribel (commerce chamber, Lyon) have kindly provided extensive location data for the 8500 stores of the city of Lyon.

[2] See, for example, T. Egami and S. Billinge, *Underneath the Bragg Peaks: Structural Analysis of Complex Materials* (Pergamon Materials Series, Boston, 2003).

[3] G. Duranton and H. G. Overman, *Review of Economic Studies* **72** (4), 1077 (2005).

[4] E. Marcon and F. Puech, http://team.univ-paris1.fr/teamperso/puech/textes/Marcon-Puech_ImprovingDistance-BasedMethods.pdf (accessed Sept. 7th 2005).

[5] One could argue that the average is dominated by the denser regions, thus eliminating the influence of peripheral areas. This effect exists, even if it is partially corrected through the ponderation by the total number of stores. I have tried several other statistical representations of the relative concentration, such as the mode or the median, but none performed as well as the average. The median, for example, fails because most A stores have no B stores around them, leading to mostly null interaction coefficients.

[6] To ascertain the statistical significance of the repulsion or attraction, I have simulated 800 random distributions of n_B stores on all possible sites, calculating for each distribution the n_B/n_{tot} ratio around the same A locations. This gives the statistical fluctuations and allows us to calculate how many times the random ratio deviates from 1 as much as the real one. I assume that if there are less than 3% random runs that deviate more than the real one, the result is significant (97% confidence interval).

[7] Alternatively, one can fully count stores closer than 50 m and linearly decrease the counting coefficient until 150 m. This leads to similar results.

[8] Important differences introduced by including weighted links are stressed, for example, in M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Physica A **346**, 34 (2005).

[9] For a pair interaction to be significant, I demand that both a_{AB} and a_{BA} be different from zero, to avoid artificial correlations [4]. For Lyon, I end up with 300 significant interactions (roughly 10% of all possible interactions), of which half are repulsive.

[10] While store–store attraction is easy to justify (the "market share" strategy, where stores gather in commercial poles, to attract costumers), direct repulsion is generally limited to stores of the same trade which locate far from each other to capture neighbor costumers (the "market power" strategy). The repulsion quantified here is induced (indirectly) by the price of space (the square meter is too expensive downtown for car stores) or different location strategies. For introductory texts on retail organization ans its spatial analysis, see B. J. L. Berry *et al.*, *Market Centers and Retail Location: Theory and Application* (Prentice-Hall, Englewood Cliffs, NJ, 1988) and the Web book on regional science by E. M. Hoover and F. Giarratani, available at http://www.rri.wvu.edu/WebBook/Giarratani/contents.htm

[11] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004). Note that the presence of antilinks automatically ensures that the ground state is not the homogeneous one, when all spins point into the same direction (i.e., all nodes belong to the same cluster). Then, there is no need of a γ coefficient here.

[12] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).

[13] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).

[14] A frustrated (A,B,C) triplet is one for which A attracts B, B attracts C, but A repels C, which is the case for the triplet shown in Fig. 1.

[15] M. E. J. Newman and M Girvan, Phys. Rev. E **69**, 026113 (2004).

[16] See, for example, the U.S. Department of Labor Internet page http://www.osha.gov/pls/imis/sic_manual.html (accessed Sep. 28th, 2005).

[17] To calculate the correlation of store and population density for a given activity, I count both densities for each of the 50 Lyon sectors. I then test with standard econometric tools the hypothesis that store and population densities are uncorrelated (zero slope of the least squares fit), with a confidence interval of 80%.

[18] Several retail categories defined by the Commerce Chamber are unfortunately heterogeneous: for example, "bookstores and newspapers" refers to big stores selling books and CDs as well as to the proximity newspaper stand. Instead, bakeries are precisely classified in four different categories: it is a French commercial structure!