Analyzing the Localization of Retail Stores with Complex Systems Tools

Pablo Jensen

Institut des Systèmes Complexes Rhône-Alpes, IXXI-CNRS,
Laboratoire de Physique, Ecole Normale Supérieure de Lyon and LET-CNRS,
Université Lyon-2, 69007 Lyon, France

Abstract. Measuring the spatial distribution of locations of many entities (trees, atoms, economic activities, ...), and, more precisely, the deviations from purely random configurations, is a powerful method to unravel their underlying interactions. I study here the spatial organization of retail commercial activities. From pure location data, network analysis leads to a community structure that closely follows the commercial classification of the US Department of Labor. The interaction network allows to build a 'quality' index of optimal location niches for stores, which has been empirically tested.

1 Introduction

Walking in any big city reveals the extreme diversity of retail store location patterns. Fig. 1 shows a map of the city of Lyon (France) including all the drugstores, shoes stores and furniture stores. A qualitative commercial organisation 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 features and, more generally, the commercial logics of the spatial distribution of retail stores, seems a complex task. Many factors could play important roles, arising from the distincts characteristics of the stores or the location sites. Stores differ by product sold, surface, number of employees, total sales per month or inauguration date. Locations differ by price of space, local consumer characteristics, visibility (corner locations for example) or accessibility. One could reasonably think that to understand the logics of store commercial strategies, it is essential to take into account most of these complex features. This seems even more necessary for finding potentially interesting locations for new businesses.

However, in this paper, I show that location data suffices to reveal many important facts about the commercial organisation 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

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

N. Adams et al. (Eds.): IDA 2009, LNCS 5772, pp. 10–20, 2009.

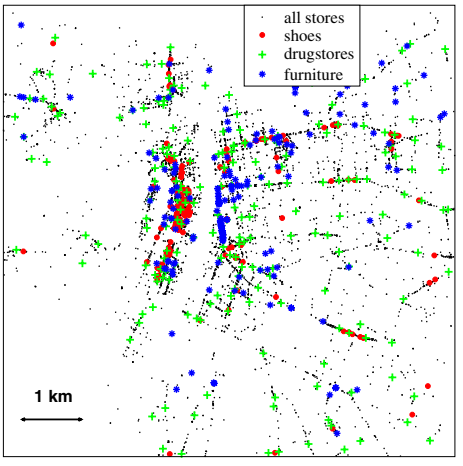

Fig. 1. Map of Lyon showing the location of all the retail stores, shoe stores, furniture dealers and drugstores

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 from solely *location* data, agree with the retailing "mantra": *the three points that matter most in a retailer's world are: location, location and . . . location.*

2 Quantifying Interactions between Activities

Measuring the spatial distribution of industries [1], atoms [2], trees [3] or retail stores [4,5] is a powerful method to understand the underlying mechanisms of their interactions. Several methods have been developed in the past to quantify the deviations of the empirical distributions from *purely random distributions*, supposed to correspond to the non-interacting case [6,7,8,9]. Recently, a method originally developed by G. Duranton and H. Overman [10], later modified by Marcon and Puech [11] has been proposed. Its main interest is that it takes as reference for the underlying space not a homogeneous one as for the former methods [6,7,8,9], but the overall spatial distribution of sites, thus automatically taking into account the many inhomogeneities of the actual geographical space. For instance, retail stores are inhomogeneously distributed because of rivers, mountains or specific town regulations (parks, pure residential zones, . . .). Therefore, it is interesting to take this inhomogeneous distribution as the reference when testing the random distribution of, for instance, bakeries, in town. Furthermore, by using precise location data (x and y coordinates), this method avoids all the well-known contiguity problems, summarized in the 'modifiable areal unit problem' [12,13,14,15]. However, the method has two main drawbacks:

1. the need of precise location data (i.e. x and y coordinates, and not only knowing that a site belongs to a given geographical area),

2. the need for Monte Carlo simulations in order to compute the statistical significance of the deviations from a random distribution.

Point (1) is probably going to be less crucial as precisely spatialized data becomes more common. Moreover, it can be argued that, when only region-type data exists, it can be more convenient to locate all the sites at the region centroid and then apply the 'continuous' method, thus avoiding contiguity problems.

2.1 Definitions of the Spatial Indicators

The indicators that are studied here deal with the problem of quantifying deviations of empirical distribution of points from purely random and non-interacting distributions. One can be interested in the interaction of a set of points between themselves, or with some other set of points. From now on we shall work with two different types of points: A and B. We define two indicators, refered to as respectively the *intra* and *inter* coefficients [11], to characterize the (cumulative) spatial interaction between sites closer than a distance r. The *intra* coefficient is intended to measure the independence between points of type A, whereas the *inter* coefficient describes the type of interactions of fixed A points with random B points. One can also work with indicators characterizing the (differential) spatial distributions between distances r and $r + dr$ (with $dr \ll r$) [10]. Those last coefficients are potentially more sensitive to spatial variations of the distributions because they do not integrate features from 0 to r. We shall start by calculating the variance of the cumulative coefficient and then extend our results to other quantifiers of spatial distributions.

We shall use the following definitions and notations:

- one has N_t sites, of which N_A sites are of type A, and N_B sites are of type B,
- for any site S, one denotes by $N_t(S, r)$, $N_A(S, r)$ and $N_B(S, r)$ the number of respectively total, A and B sites that are at a distance lesser than r of site S, where site S is *not* counted, whichever its state.

The notation $N_A(D)$ (resp. $N_B(D)$) will denote the number of A (resp. B) sites in a subset D of T, T being the set of all the points.

In this discrete model, the locations of stores A and B are distributed over the total number of possible sites, with mutual exclusion at a same site. Therefore, the geographical characteristics of the studied area are carried by the actual locations of those possible N_t sites.

The coefficients that we introduce depend on the reference distance r, however we shall drop this dependency in the notations, unless when strictly necessary.

2.2 *Intra* Coefficient

Let us assume that we are interested in the distribution of N_A points in the set T, represented by the subset $\{A_i, i = 1 \dots N_A\} \subset T$. The reference law for this set, called *pure random distribution*, is that this subset is uniformely chosen at

random from the set of all subsets of cardinal N_A of T: this is equivalent to an urn model with N_A draws with no replacement in an urn of cardinal N_t.

Intuitively, under this (random) reference law, the local concentration represented by the ratio $N_A(A_i, r)/N_t(A_i, r)$ of stores of type A around a given store of type A should, in average, not depend on the presence of this last store, and should thus be (almost) equal to the global concentration N_A/N_t, this leads us to introduce the following *intra* coefficient:

$$M_{AA} = \frac{N_t - 1}{N_A(N_A - 1)} \sum_{i=1}^{N_A} \frac{N_A(A_i, r)}{N_t(A_i, r)} \tag{1}$$

In this definition, the fraction 0/0 is taken as equal to 1 in the right hand term. Under the *pure randomness hypothesis*, it is straightforward to check that the average of this coefficient is equal to 1: for all $r > 0$, we have $E[M_{AA}] = 1$.

We deduce a qualitative behaviour in the following sense: if the observed value of the *intra* coefficient is greater than 1, we may deduce that A stores tend to aggregate, whereas lower values indicate a dispersion tendency.

2.3 *Inter* Coefficient

In order to quantify the dependency between two different types of points, we set the following context: the set T has a fixed subset of N_A stores of type A, and the distribution of the subset $\{B_i, i = 1 \dots N_B\}$ of type B stores is assumed to be uniform on the set of subsets of cardinal N_B of $T \backslash \{A_1, \dots, A_{N_A}\}$. Just as in the *intra* case, the presence of a point of type A at those locations, under this reference random hypothesis, should not modify (in average) the density of type B stores: the local B spatial concentration $(N_B(A_i, r))\ /\ (N_t(A_i, r) - N_A(A_i, r))$ should be close (in average) to the concentration over the whole town, $(N_B)\ /\ (N_t - N_A)$. We define the *inter* coefficient as

$$M_{AB} = \frac{N_t - N_A}{N_A N_B} \sum_{i=1}^{N_A} \frac{N_B(A_i, r)}{N_t(A_i, r) - N_A(A_i, r)} \tag{2}$$

where $N_A(A_i, r)$, $N_B(A_i, r)$ and $N_t(A_i, r)$ are respectively the A, B and total number of points in the r-neighbourhood of point A_i (not counting A_i), i.e. points at a distance smaller than r. It is straightforward to check that for all $r > 0$, we have $E[M_{AB}] = 1$.

We can also deduce a qualitative behaviour in the following sense: if the observed value of the *inter* coefficient is greater than 1, we may deduce that A stores have a tendency to attract B stores, whereas lower values mean a rejection tendency.

3 Analyzing Retail Stores Interactions

I now analyze in detail the interactions of stores of different trades, using the coefficients defined above.

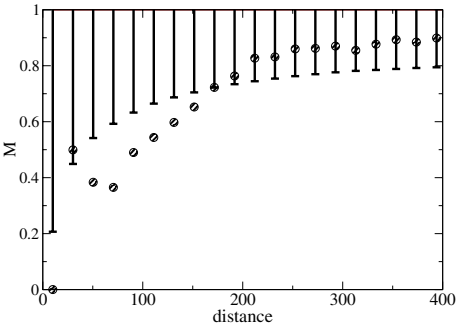

Fig. 2. Evolution of the *intra* coefficient for bakeries in the city of Lyon with respect to r, and (half) confidence interval with $\alpha = 0.05$. Data from CCI Lyon.

The figure 2 shows the practical importance of variance calculations for economic interpretations of the data. Although M_{AA} remains well below the reference value (i.e. 1), bakeries are significantly dispersed only until $150m$. For longer distances, their spatial locations approach a random pattern.

In the two following tables, I present other examples of interaction coefficients at $r = 100m$, together with the confidence intervals, for Paris, thanks to data kindly provided by Julien Fraîchard from INSEE. Table 1 shows the most aggregated activities.

Table 1. The most aggregated activities

activity	a	confidence interval at 95 %
textiles	5.27366	[0.979 , 1.021]
second-hand goods	3.47029	[0.9951 , 1.0049]
Jewellery	2.81346	[0.987 , 1.013]
shoes	2.60159	[0.9895 , 1.0105]
furniture, household articles	2.49702	[0.9846 , 1.0154]

Overall, the same activities are concentrated in Lyon and Paris. A simple economical rationale behind the concentrations or dispersions of retail activities is the following. Locating many stores at similar locations has two contradictory effects. First, it increases the attractiveness of the neighborhood by multiplying the offers. Second, it divides the generated demand among the stores. For some activities, the increase in demand is so high that it compensates the competition for customers. This is the case for when stores offer differentiated goods. Inversely, for stores offering more comparable products (such as bakeries), concentration does not increase the demand, and therefore would lead to a strong decrease in profit.

To illustrate the *inter* coefficient, I show in Table 2 the couples of activities that attract the most each other.

Table 2. The highest attractions between activities

activity 1	activity 2	a	confidence interval at 95 %
clothes	shoes	2.23216	[0.9978 , 1.0022]
Jewellery	Leather articles	2.12094	[0.984 , 1.016]
second-hand goods	household articles	2.10815	[0.9917 , 1.0083]
meat	fruits, vegetables	1.85213	[0.9906 , 1.0094]

4 Finding Retail Stores Communities

From the interaction coefficients measured above, one can define a network structure of retail stores. The nodes are the 55 retail activities (Table 3). The weighted[2] links are given by $a_{AB} \equiv \log(M_{AB})$, which reveal the spatial attraction or repulsion between activities A and B[3]. This retail network represents the first a social network with quantified "anti-links", i.e. repulsive links between nodes[4]. The anti-links 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 different results, which are less satisfactory (see below).

To divide the store network into communities, I adapt the "Potts" algorithm[5] [19]. This algorithm interprets the nodes as magnetic spins and groups them in several homogeneous magnetic domains to minimize the system energy. Anti-links can then be interpreted as anti-ferromagnetic 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 [19,20] of the satisfaction ($-1 \leq s_i \leq 1$) of site i to belong to group σ_i is:

[2] Important differences introduced by including weighted links are stressed for example in [16].

[3] For a pair interaction to be significant, I demand that both a_{AB} and a_{BA} be different from zero, to avoid artificial correlations [17]. For Lyon's city, I end up with 300 significant interactions (roughly 10% of all possible interactions), of which half are repulsive.

[4] 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 sq. meter is too expensive downtown for car stores) or different location strategies. For introductory texts on retail organization ans its spatial analysis, see [18] 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.

[5] Note that the presence of anti-links 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 then of a γ coefficient here.

$$s_i \equiv \frac{\sum_{j \neq i} a_{ij} \pi_{\sigma_i \sigma_j}}{\sum_{j \neq i} |a_{ij}|} \tag{3}$$

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 [21] to maximize the overall site satisfaction:

$$K \equiv \sum_{i,j=1,55; i \neq j} a_{ij} \pi_{\sigma_i \sigma_j} \tag{4}$$

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 maximisation). 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. [20]. This means that the grouping achieves a positive satisfaction for every element of the group. This is remarkable since hundreds of "frustrated" triplets exist[6]. Taking into account only the positive links and using the modularity algorithm [22] 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, as the U.S. Department of Labor Standard Industrial Classification System[7] (see Table 1). 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, Paris and Marseilles stores, suggesting the universality of the classification for European towns. There are only a few exceptions, mostly non-food 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[8]. Exceptions can be explained by the small number of stores or the strong heterogeneities[9] of those activities.

[6] 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.

[7] See for example the U.S. Department of Labor Internet page: http://www.osha.gov/pls/imis/sic_manual.html.

[8] To calculate the correlation of store and population density for a given activity, I count both densities for each of the 50 Lyon's 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%.

[9] 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 4 different categories: it is a French commercial structure!

Table 3. 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 $< 120m^2$)	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 $< 400m^2$)	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 orthopaedic 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	Jewellery	1.00	U	230
4	shoes	1.00	U	178
4	watches, clocks and jewellery	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

5 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. Let's

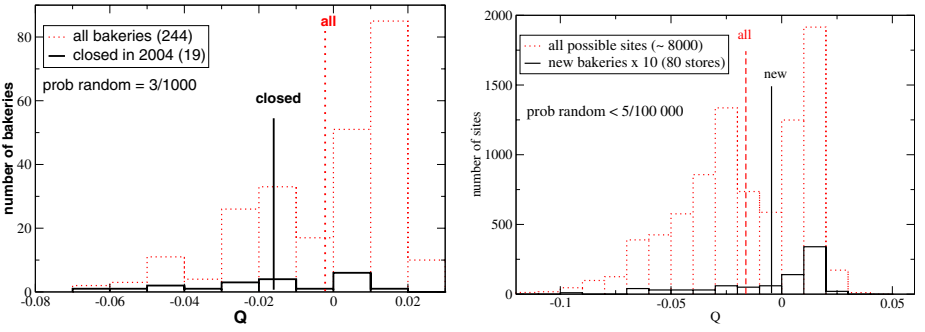

Fig. 3. The landscape defined by the quality index is closely correlated to the location decisions of bakeries. (a) The 19 bakeries that closed between 2003 and 2005 had an average quality of -2.2 x 10^{-3} to be compared to the average of all bakeries (4.6 x 10^{-3}), the difference being signifcative 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...), 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 x 10^{-4}, to be compared to the average quality of all possible sites in Lyon (-1.6 x 10^{-2}), a difference significant with probability higher than 0.9999.

take the example of bakeries. The basic idea is that a location that gathers many activities that are "friends" of bakeries (i.e. activities that attract bakeries) and few "ennemies", might well be a good location for a new bakery. The quality $Q_A(x, y)$ of an environment around (x,y) for an activity A as:

$$Q_A(x, y) \equiv \sum_{B=1,55} N_B(x, y) \tag{5}$$

where $N_B(x, y)$ 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 test empirically 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. 3) 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.

6 Conclusions, Perspectives

Practical applications of Q are under development together with Lyon's Chamber of Commerce and Industry. A software called *LoCo* reads the location data of the town(s) under investigation and gives in a few seconds the top quality regions. this can help retailers to find good locations and/or city mayor's in improving commercial opportunities on specific town sectors. In a word, LoCo pumps the cleverness of social actors, inscribed in the "optimal" town configuration, and uses it to help finding good locations. Whether the actual store configuration is optimal or not is an open question. Clearly, no one expects all retailers to be able to choose the "best" location. However, one could argue that those that have selected bad locations perish, leading to a not too bad overall configuration. This analysis suggests a crude analogy with the Darwinian selection process, with variation and selection, which would be interesting to discuss further.

References

1. Hoover, E.M.: Location Theory and the Shoe and Leather Industries. Harvard University Press, Cambridge (1937)
2. Egami, T., Billinge, S.: Underneath the Bragg Peaks: Structural Analysis of Complex Materials. Pergamon Materials Series (2003)
3. Ward, J.S., Parker, G.R., Ferrandino, F.J.: Long-term spatial dynamics in an old-growth deciduous forest. Forest ecology and management 83, 189–202 (1996)
4. Hoover, E.M., Giarratani, F.: An Introduction to Regional Economics (1984)
5. Jensen, P.: Network-based predictions of retail store commercial categories and optimal locations. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 74 (2006)
6. Ripley, B.D.: The second-order analysis of stationary point processes. Journal of Applied Probability 13, 255–266 (1976)
7. Besag, J.E.: Comments on ripley's paper. Journal of the Royal Statistical Society B 39, 193–195 (1977)
8. Ellison, G., Glaeser, E.L.: Geographic concentration in us manufacturing industries: A dartboard approach. Journal of Political Economy 105, 889–927 (1997)
9. Maurel, F., Sedillot, B.: A measure of the geographic concentration of french manufacturing industries. Regional Science and Urban Economics 29, 575–604 (1999)
10. Duranton, G., Overman, H.G.: Testing for localisation using micro-geographic data. The Review of Economic Studies 72, 1077 (2005)
11. Marcon, E., Puech, F.: Measures of the geographic concentration of industries: improving distance-based methods (2007)
12. Yule, G.U., Kendall, M.G.: An Introduction to the Theory of Statistics. Griffin, London (1950)
13. Unwin, D.J.: Gis, spatial analysis and spatial statistics. Progress in Human Geography 20, 540–551 (1996)
14. Openshaw, S.: The Modifiable Areal Unit Problem. Geo Books, Norwich (1984)
15. Briant, A., Combes, P.P., Lafourcade, M.: Do the size and shape of spatial units jeopardize economic geography estimations (2007)
16. Barthelemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A.: Rate equation approach for correlations in growing network models. Physica A 346 (2005)

17. Marcon, E., Puech, F.: Measures of the geographic concentration of industries: Improving distance-based methods (2007)
18. Berry, B.J., Parr, J.B., Epstein, B.J., Ghosh, A., Smith, R.H.: Market Centers and Retail Location: Theory and Application. Prentice-Hall, Englewood Cliffs (1988)
19. Reichardt, J., Bornholdt, S.: Detecting fuzzy communities in complex networks with a potts model. Phys. Rev. Lett. 93 (2004)
20. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Publ. Natl. Acad. Sci. USA 101, 2658–2663 (2004)
21. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671 (1983)
22. Newman, M.E.J., Girvan, M.: Community structure in social and biological networks. Proceedings of the National Academy Science USA 69, 7821–7826 (2004)