Unusual Spatial Patterns of Industrial Firm Locations Uncover their Social Interactions

Guy Kelman
Institute of Life Sciences
Hebrew University of Jerusalem
Jerusalem, Israel
Email: superk@cs.huji.ac.il

Marco Lamieri
Research dept.
Intesa SanPaolo
Milan, Italy
Email: marco.lamieri@intessanpaolo.com

Eran Manes
Jerusalem College of Technology
Jerusalem, Israel
Email: msemanes@gmail.com

David S. Bree
Department of Computer Science
University of Manchester
Manchester, UK
Email: davidSbree@gmail.com

***Keywords*-business networks; communication networks; social interaction; spatial statistics; stationary point processes; economy; industrial firms;**

***Abstract*—In this paper we report evidence from the Italian industrial sectors whereby firms that buy and sell are spatially distributed with a pattern that reflects the microeconomic powers at play. The main finding is that firms are neither clustered around population centers nor are they situated at random. Although geography has an important role in shaping the population map of Italy, the reasons for the positional pattern of buyers and sellers appear to be social. Geographic proximity between sellers and their buyers is supported by the excess in short-distance social ties.**

I. Introduction

The main aim of this paper is to test whether or not firms within an industrial sector are geographically located with association to their trade partnerships. Our major finding is that Italian firms in a given sector, cluster at a certain distance from their competitors, irrespective of their role as sellers or buyers, confirming to Hotelling's "linear city model" [1], where firms are positioned so as to maximize their market share. A second finding confirms that the number of seller-buyer links is indeed inversely proportional to the distance between the seller and the buyer.

One of the key features of networks is their distance metric. In its reduced form it is the minimal number of nodes it takes to pass when going from an origin to a destination node by following the links. In the case of real-world networks, the nodes are also situated on a 2-dimensional surface and their network location is usually unrelated to their geographical position. It is not surprising, therefore, that in network analysis the Euclidean metric drew little attention. Over time, it has become clear that actual bodies that are deposited on the surface of the Earth must sometimes be put in geographical context.

Especially in social networks, incuding trade-networks, frequent personal encounters are likely to be reflected in the social web [2]. So, for example, the distance of travel from one location to another could affect the decision of participants to find and maintain direct contacts. This is in support of the industrial knowledge spillovers that occurred in and around industrial parks., e.g. see [3] for a comprehensive account of the innovation boosts and growth of cities as explained by the covariate of knowledge spillover.

Knowledge spillover is a prominent factor that determines firm's spatial location choices and their tendency to cluster in order to fully capitalize on increasing returns. This idea is now a cornerstone in economic growth and R&D management, since the publication of the seminal paper by Krugman [4], which was followed by numerous other works showing that spatial concentration plays an important role in innovative activity, and providing robust empirical evidence in support of this theory. For an excellent literature review, the reader is referred to [5].

From an Economic perspective, freight may also be a major component affecting the mobility of goods, so environmental features such as hills and roads will therefore become relevant. Since this cost is generally shared with the customer, a sale would be priced lower in areas of dense cargo travel and flat lands. The reader could appreciate that competition between sellers exists with relation to the cost of the shipments, and to a lesser extent, with the distance of travel. Moreover, both the sellers and their buyers are expected to strategically position themselves on the map by factoring in the production and accessibility to the goods. Let us consider a production chain in which each producer is a seller of output material and a buyer of input inventories. Three financial forces act upon the producer: **(1)** the requirement to be in close proximity with its resources, **(2)** the desire to keep competitors at bay, and **(3)** the trivial

URI: http://hdl.handle.net/10125/41367
ISBN: 978-0-9981331-0-2

appeal to create or sit in dense populations of customers.

All the participants obey these three forces, conditional on their placement in the Leontief Input/Output model of trade between firms. For example: A Wheel and Tire factory will probably buy steel and rubber, in contrast to the small chance that it will buy aircraft seats. Following this intuitive example, the cost of moving steel may be high and the Wheel and Tire factory may find it useful to be situated next to steel works, and farther away from the potential client pool.

The reader may be tempted to inquire about the online ordering mechanisms that are so widespread nowadays. Do these come into play in the context of industrial trade-partnerships? Also, how has the digital age shaped the face of the industrial sectors, as it would appear that the distances between the buyers and their sellers bear no consequence to the buyers?. However, consider the industrial sectors that offer goods or services, rather than wholesale and retail. These industries do need to factor the transportation costs into the sale price, especially when the buyer is situated far away. The issue of transportation costs has a well documented distance barrier by which short distances are always favored [6]. This intra-national home bias was shown to shrink significantly but remains non-negligible when adding the possible social network effect, or information barrier as it is sometimes called. This barrier is defined by the ability of the firm's management to promote trade while exploiting their social ties that are inherently local [7] [8]. The bias due to network effects however, shows correspondence with the chosen unit area [6]. As a result the quantification of reduction in the intra-national home bias due to online ordering and network effects is limited. Our intuition is, therefore, that in comparison to the manufacturing industries, the consumer goods sectors (wholesale and retail) have responded in a more extreme fashion to the introduction of internet shopping.

Now, with respect to competition there have been many applications following Hotelling's "linear city model" [1]. The assertion backed by observation is that given a set of uniformly dispersed customers, competing sellers optimize their geographical location in such a way that maximizes their market share and places them in the geometric center, closest to any potential customer. In the next simplest instance there are two sellers selling identical goods situated along a single (linear) street of length 1: the market will split in half and both sellers will position themselves in the middle of the street. Consequently, they will be close to each other. Maintaining market power, on the other hand, will drive these sellers away from each other, i.e. at $1/4$ and $3/4$ of the street's length where, again, each of them controls exactly half of the market. Any relocation of one seller will encourage the other to move closer to the middle, gaining market power. This trade-off between a firm's market power and the price competition reaches an equilibrium with the optimal positioning of the two sellers. Generalizations to non-uniform spread of customers were also studied [9], and yet the sellers generate a similar market splitting response to competition, under more restrictive setups.

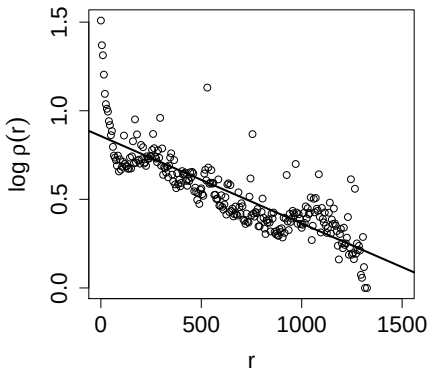


Figure 1. The density $\rho(r)$, or ensemble average number of customers in any ring of distance r around the 129,584 sellers in our business network. The number of customers is 345,403 totalling in $1,117,029$ links to buyers. A reference fit to $\log \rho(r)$ is plotted in solid. Outliers can be explained by intuitive geographic and demographic features such as commute distance (40km), region boundary (200km), Milan-Rome and Rome-Palermo (550km), and long-range monopolist firms (800-1200km). The fitted slope is -0.000506, declining one order of magnitude over a distance of three orders magnitude.

Going back to the Italian business network, the sellers and the buyers constitute the network's directional links, and from the financial statements data set we can locate the operational addresses of these firms[1].

In the analysis that follows we define the search area in several ways: one would be a rectangular frame overlaid on the map of Italy, the other would be the area (and shape) of Italy in the whole. Inside these search areas we deploy methods of second moment statistics of spatial-statistics processes in order to find whether or not there is lack of homogeneity, i.e. clustering of points/nodes. The software that we use is from the R package `spatstat` [10].

II. Methods

As part of obtaining prior knowledge on the geographic layout of the network we turn our attention to the density of customers with distance. This, as indicated in figure 1, could be approximated by a constant. On this evidence we base our primary theoretical assumption that industrial customers are uniformly visible to the seller with little regard to distance.

A. Point Processes and Spatial Analysis

A *point process* is a random process that places points in space and a *spatial point process* is a point process that lays points out in the Euclidean space. The points usually specify objects of study such as shrubs in the savanna, drops

[1]The address of the bank branch where the firm's headquarters is

in a cloud or lighting strikes on a surface. Patterns that emerge from this positioning process can be analyzed using statistical models. In this context, a point process model is used to check, for example, spatial homogeneity, dispersion, centrality and spatial correlation.

Similar to the dynamics in a network, the bottom-up approach is to model the rules for construction of spatial patterns. So, the most appealing methodology in spatial analysis of point processes is to work with the moments of the distance distribution. In the following sections we will describe different spatial point processes. For each process we will define the quantities used for spatial analysis, and highlight how the process is to be compared to the null hypothesis, also termed the *Poisson Point Process*.

The Poisson point process in 2 dimensions: Given that n points should be dispersed inside a region W, we divide it into sub-regions $W = \{\cup A_i\}$ such that $p_i = w_{A_i}/w_W$ is the proportion of A_i's area to the total area. The probability that the number of points k that fall inside the sub-region A_i has a binomial distribution.

$$\Pr(n_{A_i} = k) = \binom{n}{k} p_i^k (1-p_i)^{n-k} \tag{1}$$

Without loss of generality we will abbreviate $n_{A_i} = n$. Now, the mean of the binomial distribution is $\lambda = np$ and so $p = \lambda/n$. Putting this into (1) and assuming $n \to \infty$ we arrive at the Poisson distribution

$$Pr(n_{A_i} = k) = \frac{e^{-\lambda}}{k!}\lambda^k \tag{2}$$

for all sub-regions A_i. The quantity λ is termed *the intensity* of the point process.

Distance distribution: The next definition relevant to our discussion is derived from the distance metric $d_{a,b} \geq 0$. This metric is the Euclidean distance between points a and b. Generalizing this we could estimate the distance between a and a set of points X as the minimal distance between a and any point inside X. $d_{a,X}$ is called the *contact distance*. An important property of this quantity is that point a must have neighbors in any distance *greater* than its contact distance. If we consider the circular sub regions $A \subset W$ of radius r centered at a point $a \in W$ then

$$d_{a,X} \leq r \quad \Longleftrightarrow \quad n_{u(a,r)} > 0 \tag{3}$$

noting that the expected number of points in a circular region $u(a,r)$ is

$$\mathbf{E}[n_{u(a,r)}] = \rho_0 \pi r^2$$

where $\rho_0 = n_W/w_W$ is the density of the system. Immediately following this notation we could write the cumulative probability distribution of the contact distance as

$$\begin{aligned}\Pr(d_{a,X} \leq r) &= \Pr(n_{u(a,r)} > 0) = \\ &= 1 - e^{-\mathbf{E}[n_{u(a,r)}]} = \\ &= 1 - e^{-\rho_0 \pi r^2}\end{aligned} \tag{4}$$

This distribution function is denoted by $F(r)$ and called the *empty space function*. $F(r)$ depends on the radius of observation, and could easily be estimated from the data by counting the points that fall within a radius r around the origin a.

Nearest Neighbor Distance Distribution and the J function: The nearest neighbour distance distribution function is close in its definition to the empty space function. It is the cumulative probability distribution with respect to a circular region in the geographic space, as opposed to the contact distribution that observes the space from the perspective of the central point, around which neighbors are located. The NNDD is the number of balls of radius smaller or equal to r that remain empty of points. Formally, consider a reference point a in a point process $x \in X$. The NNDD is the distance distribution of points in $X \backslash \{x\}$ of which the contact distance is r:

$$\begin{aligned}\text{NNDD}(r) &= \Pr(d_{a,X\backslash\{x\}} \leq r) \\ &= \Pr(n_{u(a,r)\backslash\{x\}} > 0)\end{aligned} \tag{5}$$

In a stationary Poisson point process, NNDD(r) is identical to the empty space function $F(r)$.

The contrast between NNDD and the empty space function is emphasized by what is commonly known as the J function. This is the proportion of the complementing cumulative distribution functions of F and NNDD.

$$J(r) = \frac{1 - \text{NNDD}(r)}{1 - F(r)}$$

The J function often admits a clear analytic solution even in situations when F and NNDD cannot be estimated. Its definition is restricted to cases where $r \geq 0$ such that $F(r) < 1$.

Another immediate property of the J function is that it evaluates to unity for Poisson processes. However, we must also note that due to the closeness of $F(r)$ and NNDD(r) fluctuations may be amplified.

Var - Mean ratio: The variance to mean ratio, denoted I, is one of the simplest measures of how dispersed a point process is: The random placement of points that could be described by a Poisson distribution has its variance equal the mean

$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda$$

thus, a quick estimation of var/mean allows us to reject the possibility of complete independence of the point locations. Inspecting the limiting cases is helpful in classification of the distribution: roughly, when $I = 0$ this is a degenerate distribution where a single value is chosen with probability 1. In the range $0 < I < 1$ is a process with a binomial distribution, and $I > 1$ is generated by a negative-binomial (discrete) distribution where points cluster together due to interactions that increase with I.

The method includes a procedure of quadrature of the 2-dimensional space by which the search area is partitioned into equal sized square (quadrats) of length L on the side. As we decrease L each quadrat confines less points. The procedure will terminate at $L = 1km$, then an estimation is made in how many different values of L will I be equal to unity. A Poisson point process will be such that $I = 1$ for a wide variety of quadrat sizes.

It is well worth noting that reversing the reasoning of $I = 1$ is not always possible, and one cannot assume a Poisson point process by backtracking from an evidence of $I = 1$ [11], [12].

The Pair Correlation (Radial) Distribution Function and Ripley's K: The reason to give further thought on point processes is that we would like to pin down a possible clustering of points in space, and the derivations in previous sections may not be enough to quantify the nature of this point process.

The radial distribution function and Ripley's K function are both second order statistics. The full derivation of these functions is given in [13], [14], and [15]. What is worthy of thorough explanation are the intuitions for these measures and their outcomes.

The pair correlation function in the Euclidean space $g(r)$ is an estimator for the strength of interaction between points. This information is used to find if there is regularity in the point process, or otherwise a clustering mechanism controls the spatial positioning. $g(r)$ is applicable to dynamic point processes as well as stationary ones. It is positive and depends on the radius around an origin point a with no relevance to the position of this origin. It measures the ratio of the point count over what would be expected by chance in a ring of width dr at radius r. For a given radius r this is identical to the Pearson correlation coefficient. The estimation of $g(r)$ from data can be written as

$$g(r) = \frac{N_a(r)}{N_{Pois}(r)} = \frac{N_a(r)}{2\pi r dr \cdot \rho_0} \tag{6}$$

where $N_a(r)$ is the count of points x found in the ring $[r, r + dr]$, and $N_{Pois}(r)$ is the expected number of points in an identical ring. This expectation requires knowledge on the bulk density ρ_0 of the points in W.

If the points that we observe result from a Poisson process, the number of points in the ring dr at r should be monotonously increasing with r, independent of the point of origin a. Thus, given a set of points inside a search area we may choose each, in turn, to be the origin and estimate the pair correlation while iterating through the other points.

Ripley's K function is another familiar second order statistic. Intuitively, the K function is equivalent to the variance of the sample. With reference to the radial distribution it could be written as

$$\frac{dK(r)}{dr} = 2\pi r g(r) \tag{7}$$

in other words, $K(r)$ is proportional to the integrated form of the pair correlation. In particular it is the expected number of points within distance r of a given point chosen to be the origin. For example, the K function of a Poisson process is $K(r) = \lambda^{-2}\pi r^2$. From this expression it is straightforward to arrive back at (7).

The major advantage of K over $g(r)$ is that it is invariant to random translation and thinning of the points in the point process. A 'thinning' process is similar to an observation missing at random. It is obtained by selecting at random a subset of points Y from a point process X.

Communication networks and their geographic impact

In networks there is no spatial structure. Even so, there may be much to learn by analyzing the geographical structure of the nodes (or agents) in cases where such information exists [16]. The network gives additional information on whether two points a and b truly interact with one another. Further, in communication networks, where information is passed between the nodes, travel distances may be skewed by the added link information. The realization of distances between points is thus amplified across existing links while rendering other paths unutilized.

The limiting cases to consider are the fully connected graph and on the other extreme, a sparse random graph or Erdös-Rényi (ER). In fully connected networks, the distance between any two nodes is 1, link information may be superfluous, and so the distances are dominated by the point process. A sparse ER network, can be constructed by beginning with a fully connected graph of N nodes and removing links at random. This structure causes the average path length to increase above unity, up until the graph forms a linear chain with an expected path length $N/2$. In such networks the path lengths are dominated by the network structure rather than the geographic distance [17].

B. The data

Every node in our network is a firm that has an address in Italy. The data for this trade-network were made available from a large Italian bank, and comprise two data sets:

- Time series of individual firm balance sheets (and Profit & Loss statements) in the 8 years between 2002 and 2009. These data contain information that allows one to know the financial status of a firm. It will hereafter be abbreviated BS.
- Bank-mediated credit transactions of trading partners in the year 2007. Each record contains 3 fields of interest: The identity of the seller, the buyer and the total face value of the all trade-credit transactions.

In total, BS holds balance sheets of 1.3 million firms over the 8 year-window, and on average, 700,000 firms are represented in any given year. The overlap along the timeline is approximately 300,000. In 2007 there were 703,858 firms with net-sales greater than zero (potential suppliers) and

601,535 firms had purchases greater than zero (potential buyers).

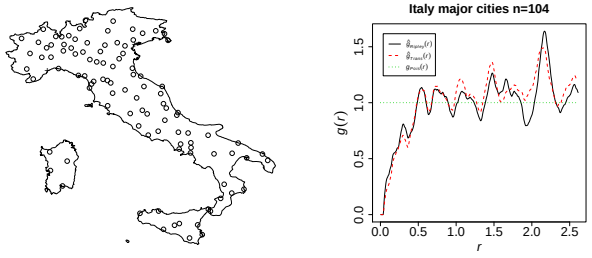
In the TC data set there are 1,578,812 firms, connected by 7,290,072 links. When intersected with the firms in BS we obtained a total of 345,403 firms connected by 2,874,830 links. This makes a ratio of approximately 1:8 nodes over links. 273,726 of the firms in the TC data are suppliers (have incoming links), and of the joined data set TC+BS, 140,580 are suppliers. If we remove the suppliers that are linked to buyers without BS data, we are left with 129,584 suppler firms, all of whom have at least one buyer with BS information. We call this set M. 122,728 of the suppliers in M (94%) have outgoing links and therefore are buyers and sellers. The remaining 215,819 firms are buyers only.

An individual firm (node) assumes attributes from the BS such as firm size, credit-rating, financial costs or industrial classification. The address of each firm can be represented as a marker on a map of Italy.

III. Results

We begin our exploration by testing the tendency of selected subsets of firms in the trade-network to cluster. The first subset consists of **the 104 major cities in Italy**.

In the past, cities were built along waterways, and thereafter served as major crossroads. They are considered efficient geographic centers for resources and trade opportunities. Naturally, we would expect that buyers and sellers will exhibit the tendency to stay close to city centers, and thus the position of cities should serve as a mediating variable.



(a) (b)

Figure 2. the radial distribution of 104 major cities in Italy. The periodic structure is visible at 50km (= 0.5 latitudinal degree), 70km (= $50 \cdot \sqrt{2}$), and their multiples thereafter.

The two panels in figure 2 represent the visualization of the 104 major Italian cities, and the radial distribution function $g(r)$ estimated inside the coastline contour. This is a replication of a famous result of the spatial pattern of cities [18]. Further statistics of this point process appear in figure 3. The subscript *Ripley* stands for the ordinary isotropic-corrected estimate which corrects for edge effects that introduce a possible non-uniformity of the point pattern due to the shape of the sampling window, *Trans* is the translation-corrected estimate that is less suited in this measurement on the map of Italy since we were using a polygonal search area [14]. Last, *Pois* is the theoretical $g(r)$ (the radial distribution function of the Poisson point process).

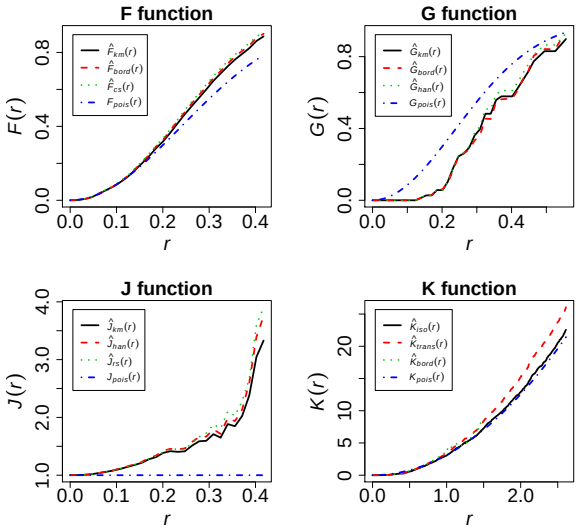


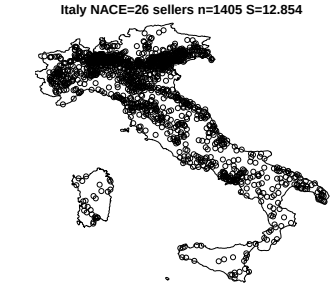
Figure 3. relevant statistics of the spatial pattern estimated from the 104 major cities in Italy. Distances are in latitudinal degrees (1 deg = approximately 100km). The Nearest Neighbor degree distribution is G, the empty space function is F, K is Ripley's K-function and $J = (1 - G)/(1 - F)$

The nearest neighbor degree distribution $G(r)$ in figure 3 and the radial distribution $g(r)$ in figure 2(b) give a clear indication of a hard-core disc structure at a radius of 30-50km since the values at r smaller than 0.5 sit well below the Poisson reference lines $G_{pois}(r)$ (the blue dashed line in figure 3 and the green dashed line in figure 2(b)). The empty space (F) and the K-functions follow the Poisson reference line and so point to a possibility that the process is random, especially in small radii. However, the J-function that amplifies the deviation of NNDD from the empty space function, shows that the randomness occurs only at small radii ($r < 0.05^o$ or approximately 5 km), much smaller than the hard core disc reflected by the NNDD. In this set of plots, the subscripts *pois*, *iso* and *trans* are the same as in figure 2(b), *bord* (equivalently *rs*) is a sample reduction such that points nearing the border of the search area will not be considered (will be useful for the large samples in later analyses), *km* is the Kaplan-Meier correction that estimates the distribution using a spectral method on the reduced samples near the edge [19], and *cs* (equivalently *han*) is the Chiu-Stoyan correction that uses the border-method, i.e. counting samples about each origin point up to the respective maximal radius that does not touch on the border of the search area.

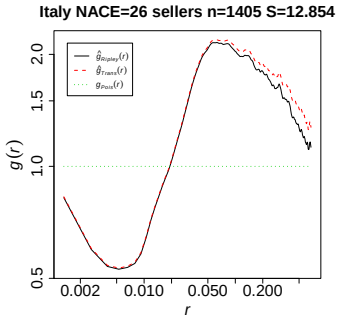
The conclusion is that the major cities in Italy are positioned in a regular construct, with a characteristic distance between them that is strictly bounded from below. This also supports the historical view of Italy; the 104 cities were territorial capitals until 20th century, with distinct

boundaries that were guarded and managed by similar sized populations. By nature of their population sizes, the areas of these territories were more or less identical. This may be the reason for the hard-core disc structures.

We could attribute the lack of sensitivity of the K-function to the fact that the analysis was performed inside the boundary of the Italian peninsula, a closed, peculiar shaped curve, making it harder to contain edge effects. We found it useful to place these statistics against the ones obtained inside a smaller, rectangular, search area to show the difference. In further analyses we will exercise this method.



(a)



(b)

Figure 4. an illustration of the sampling overflow that may occur in radial distribution function calculations. g(r) peaks in the range 5-10km and drops sharply elsewhere. This indicates a high probability of locating any two firms of this industrial code at 10km apart.

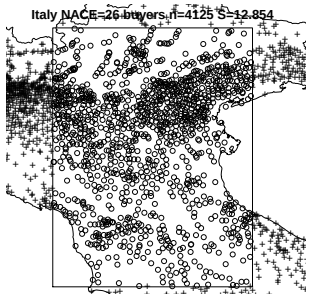
Every firm has an industrial classification that marks the core trade of the firm (and what products they may sell) with good approximation. This classification serves as coordinates in the Leontief Input/Output model of trade estimations between sectors in the industry. Furthermore, this classification is important to financial institutions and regulators that may identify sectors to target their policies. The sellers in any single industrial sector can be considered as competitors. There are 51 identified sectors in the Italian industry.

Figures 4, 6, and 7 display several radial statistics of seller firms in industrial code 26: 'Manufacture of other non-metallic mineral products'. The main trade of this industry is the fabrication of products made of glass, ceramics, and clay. The proportion of sellers or buyers with this industrial code is 1.5% of the total of firms in our network, so this sector may be regarded as representative of Italian industrial sectors.

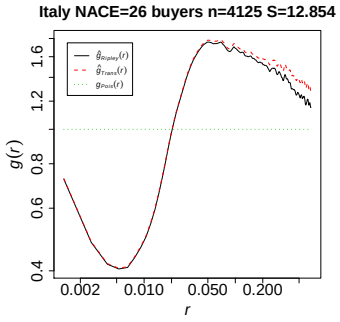
Evidence shows that the firms keep each other at a distance. It is important to note that the findings reported for this subset were reproduced in all the other industrial codes, with small variations in the minimal distance between neighboring firms.

We can offer the following interpretation of these results: Firms that are placed in a competitive market may follow Hotelling's model trying to control the price of the good while maximizing their market power.

This evidence also provides the following important insight: firms do not necessarily position themselves inside city centers (where the minimal distance is in the order of 30-50km). Rather, as visually displayed in the zoomed map (fig. 6) we can clearly detect the positioning alongside major roads and other geographical features (Northern Italy is mountainous, so cities are positioned alongside rivers on the way to the mountain-pass areas)



(a)



(b)

Figure 5. spatial analysis in a window of size 2.5×3.5 degrees restricts the number of analyzed points. Inside this window there are 1258 of the 4125 buyer firms. Panel 5(b) gives the radial distribution function $g(r)$ in log scale; note it is both above and below the estimated Poisson process. The smaller than would be expected by chance pair-correlation, with minimum at 7km, and the larger than what would be expected by chance, at $r >$ 20km, supports the finding in figure 4 that repulsion exists between firms in the same industry.

A third subset of interest is the groups of **buyer firms in a single industry**. If they buy in competitive markets, do they cluster around resources too? Figure 5 is a display of the

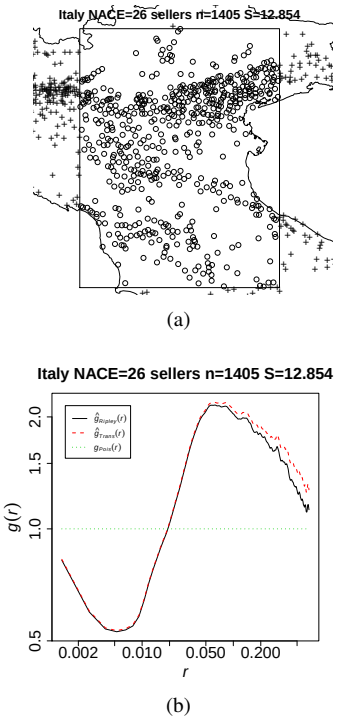


Figure 6. With axes similar to 5, we display the analysis of seller positions in the non-metallic mineral products industry. 558 of the 1405 seller firms are in the rectangular window. A pair correlation pattern in 6(b) of below and above the estimated Poisson process indicates that a repulsive force keeps firms at a distance, as it does in the previous figure.

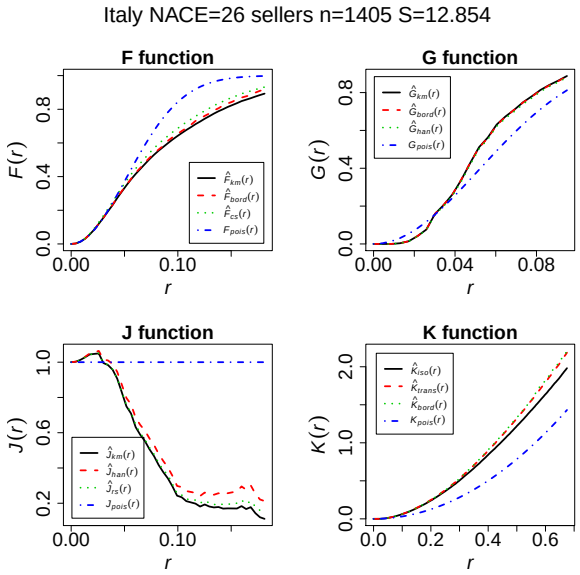


Figure 7. statistics of the spatial pattern created by the seller firms in industrial code 26. Axis labels follow the conventions in figure 3 and the search area is like in panel 6(a). G infers that it is an inhomogeneous point process with a hard core disc of 3km, inhomogeneity is indicative of the placement along roadsides. K suggests a departure from Poisson with a tendency to cluster. The J function indicates the dominance of the NNDD (G function) in all distances.

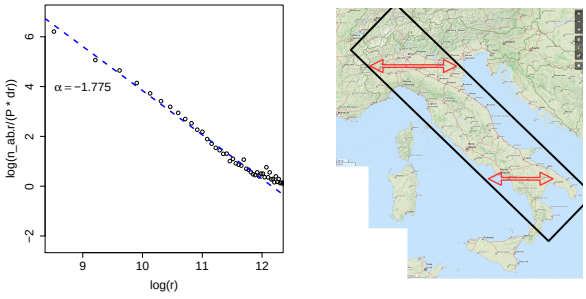


Figure 8. the radial distribution of sellers and their buyers on double logarithmic axes, clearly indicating a clustering effect. In this panel r is measured in kilometers. The right panel illustrates the large aspect of the search area and that it sits slanted respective to the roads. The width of Italy is 250km but the roads follow along the latitudes. For practical purposes the width is 300km. The map in this panel was taken from `www.openstreetmap.org`

radial distribution function of industrial code 26 buyers; the buyers of ceramic, glass and clay products such as wall and electrical insulators, optical fibers, and bricks as intermediate inputs. We can detect a similar positioning pattern in this group of firms that suggests again a hard core process. The interpretation of this evidence could be twofold:

- buyers are sellers, and if they buy a product in one industry, they will compete over their own customers in their own industries, so in any given area, buyers of a single product will act by responding to competitive forces inside their own markets.
- buyers and sellers share the positioning pattern due to another, possibly latent, mediating variable. One suggestion to this alternative is that online ordering systems did penetrate these industries, and so less costs are incurred by approaching a distant seller. This calls for a differential study between the eras of pre- and post-internet revolution which the current data does not cover.

The two options are related in the sense that a possible mediating variable could be the mere fact that these buyers are competing for market power in their own industries. We can just make a note that it is rare to find a buyer inside the same industry as his/her seller. But since buyers always purchase inventories from many sellers, it is hard to follow a single path that flows from one seller to one buyer.

An established fact in our data is that 94% of the sellers are also buyers. It is therefore easier to accept the latter option while assuming, with the necessary caution, that other buyers (that are not sellers) do in fact produce and sell, but our data does not show them as such. There could be many reasons for missing network information [20], [21], [22].

Let us now consider the relationship between the sellers and their marked buyers. Seller or buyer pools, independently, exhibit a characteristic minimal distance. Any buyer pool Y may purchase a certain product from sellers of

industrial class X. However, finding the exact seller (or in reverse - recruiting the exact buyer) is a marketing task and generally takes a lot of effort in the seller's account. The next subgroup to analyze is therefore the **sellers and their direct buyers**. In this analysis we place each seller in the origin, and measure the radial distances of the direct buyers to this seller. Then, the distances are classified irrespective of any individual seller (we ignore the labelling or position of any specific seller).

The estimation of the radial distribution function uses the counts of buyers in any ring of width dr created around all sellers. The results are shown on double-logarithmic axes in figure 8. From this figure we see that the pair correlation is scale-free up to the width of the search area (250-300km), and the interpretation is that clustering occurs around sellers irrespective of their reach (the distance to their farthest customer).

The existence of many extremely small distances between sellers and their buyers indicates that geographical clustering drives the distance distribution towards the origin rather than network clustering. The fact that the distances are small means that sellers tend to write contracts with trade partners that operate close-by.

In previous studies that tried to overlay social phenomena on a geographic map, it was discovered that clustering occurs owing to the ease of social encounters in the localities [2].

The rejection of a Poisson process between sellers and buyers, or between sellers or buyers individually was established by means of G, K, F-functions and the pair correlation function. The regular grid formed by the cities supports the fact that communication patterns in the industry determine the pull forces of resources, much more than legacy locations of centers of operation.

The var-to-mean ratio methods are less interesting than this last result. Quadrat analysis of var/mean ratio was also performed inside the subgroups that were defined above, and the result of the quadrat analysis of buyers connected to a seller is given in figure 9. The departure from var/mean $= 1$ occurs in very small radii, and then reaches high proportions later on, with a peak at $L = 300$km, a low at $L = 550$km, another peak at $L = 750$km, a low at $L = 1050$km and the again a peak at $L = 1200$km. The peaks (350, 750 and 1200) are an expression of sharp increases of variance with no significant change in the mean value of counts inside the quadrats. These may be due to the geography of the search area, or as mentioned before, due to restrictive elements of the numerical method: there is but a single quadrat of $L = 1200$km, and there are no more than 4 quadrats of 750km on the edge, of which two are heavily occupied by paths between sellers and their buyers (extending from Rome to each side of the country). The peak at 350km is again a result of the existence of major roads on the east-west axis, where some quadrats have frequent occurrence of paths between sellers and buyers, otherwise quadrats are empty of such paths.

Other partitions of the firm data set did not provide extra information on top of the two main results: **(1)** that the var/mean ratio is equal to unity in extremely small radii and nowhere else, rejecting the possibility of a Poisson point process that occurs uniformly in all areas of the map, and **(2)** that the search area is restrictive, and natural or legal boundaries are evident.

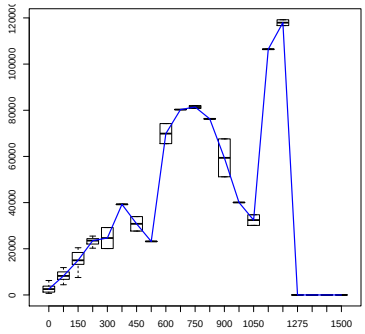


Figure 9. var/mean ratio of quadrats with edge size L. $0 \leq L \leq 1500$ is incremented 75km at a time. The magnitude of the peaks is less important than the quadrat size at which they occur, since the number of quadrats that fill the map drops with L.

IV. Conclusion

The use of second moment spatial analyses and subsetting of the data to different species were used previously in the context of trees in the wood or galaxy catalogs. The investigation in this paper is unique in the sense that our data combines the three features: inside the search area of Italy we obtain **(a)** the geographical positions of $345,403$ firms, **(b)** their industry (species), and **(c)** information on the ones that are connected.

We show the feasibility of intersecting geographical data of the nodes with network connectivity to retrieve the relationship between the occurrence of links and the spatial positions of the nodes in Euclidean space. This is especially true in social networks, and today evidence accumulates in the literature to confirm that the affinity of participants in a social network is reciprocal to the distance between them.

A possible outlook is to consider the applicability of this method to financial institutions: the points on the map mark buyer firms that were brought to the attention of the bank by the borrowers. If the general pattern is a scale-free clustering of buyers around a seller, then a 'hole' (or peak/low) in the radial distribution function at radius r may mean that invoices from buyers at radius r were deliberately put aside [23].

References

[1] H. Hotelling, "Stability in competition," *The Economic Journal*, vol. 39, no. 153, pp. 41–57, March 1929.

[2] D. A. Coleman and J. C. Haskey, "Marital distance and its geographical orientation in england and wales, 1979," *Transactions of the Institute of British Geographers*, vol. 11, no. 3, pp. 337–355, 01 1986. [Online]. Available: http://www.jstor.org/stable/621794

[3] M. H. Best, "Greater bostons industrial ecosystem: A manufactory of sectors," *Technovation*, vol. 39, pp. 4–13, 2015.

[4] P. Krugman, "Increasing returns and economic geography," *The Journal of Political Economy*, vol. 99, no. 3, pp. 483–499, 1991.

[5] D. B. Audretsch and M. P. Feldman, "Knowledge spillovers and the geography of innovation," *Handbook of regional and urban economics*, vol. 4, pp. 2713–2739, 2004.

[6] R. Hillberry and D. Hummels, "Trade responses to geographic frictions: A decomposition using micro-data," *European Economic Review*, vol. 52, no. 3, pp. 527–550, 2008.

[7] D. L. Millimet and T. Osang, "Do state borders matter for us intranational trade? the role of history and internal migration," *Canadian Journal of Economics/Revue canadienne d'économique*, vol. 40, no. 1, pp. 93–126, 2007.

[8] P.-P. Combes, M. Lafourcade, and T. Mayer, "The trade-creating effects of business and social networks: evidence from france," *Journal of international Economics*, vol. 66, no. 1, pp. 1–29, 2005.

[9] B. Gupta, D. Pal, and J. Sarkar, "Spatial cournot competition and agglomeration in a model of location choice," *Regional Science and Urban Economics*, vol. 27, no. 3, pp. 261–282, 1997.

[10] A. Baddeley and R. Turner, "spatstat: An r package for analyzing spatial point patterns," *Journal of Statistical Software*, vol. 12, no. 6, pp. 1–42, 2005. [Online]. Available: http://www.jstatsoft.org/v12/i06/

[11] C. J. Krebs, *Population fluctuations in rodents*. University of Chicago Press, 2013, ch. 6.

[12] M. R. Dale, "Spatial pattern analysis in plant ecology," *Ecology*, vol. 88, pp. 366–370, 1999.

[13] P. J. Diggle, P. J. Ribeiro, and O. F. Christensen, "An introduction to model-based geostatistics," in *Spatial statistics and computational methods*, J. Møller, Ed. Springer Verlag, 2003, ch. 2.

[14] B. D. Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 172–212, 1977.

[15] A. Baddeley, *Spatial Point Processes and their Applications*, University of Western Australia, School of Mathematics & Statistics, Nedlands WA 6009, Australia, 2010.

[16] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1–3, pp. 1–101, 2 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037015731000308X

[17] S. Sreenivasan, T. Kalisky, L. A. Braunstein, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Effect of disorder strength on optimal paths in complex networks," *Physical Review E*, vol. 70, no. 4, p. 046133, 2004.

[18] L. GLASS and W. R. TOBLER, "General: Uniform distribution of objects in a homogeneous field: Cities on a plain," *Nature*, vol. 233, no. 5314, pp. 67–68, 09 1971. [Online]. Available: http://dx.doi.org/10.1038/233067a0

[19] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[20] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[21] C. J. Rhodes and P. Jones, "Inferring missing links in partially observed social networks," *J Oper Res Soc*, vol. 60, no. 10, pp. 1373–1383, 10 2008. [Online]. Available: http://dx.doi.org/10.1057/jors.2008.110

[22] J. A. Smith and J. Moody, "Structural effects of network sampling coverage i: Nodes missing at random," *Social networks*, vol. 35, no. 4, pp. 652–668, 2013.

[23] G. Kelman, D. Bree, E. Manes, M. Lamieri, N. Golo, and S. Solomon, "Dissortative from the outside, assortative from the inside: Social structure and behavior in the industrial trade network," in *Proceedings of the 48th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Computer Society Press, 2015, Jan 2015, p. 10.