



Effects of missing data and locational errors on spatial concentration measures based on Ripley's K -function

Giuseppe Arbia^a, Giuseppe Espa^b, Diego Giuliani^c and Maria Michela Dickson^d

ABSTRACT

Effects of missing data and locational errors on spatial concentration measures based on Ripley's K -function. *Spatial Economic Analysis*. Measures based on Ripley's K -function are the preferred tools to test the concentration of individual agents in an economic space. In many empirical cases, however, the datasets contain different inaccuracies due to missing data or uncertainty about the location of the agents. Little is known thus far about the effects of these inaccuracies on the K -function. This paper sheds light on the problem through a theoretical analysis supported by Monte Carlo experiments. The results show that patterns of clustering or inhibition may be observed not as genuine phenomena but only as the effect of data imperfections.

KEYWORDS

spatial microeconometrics; spatial concentration; Ripley's K -function; geomasking; confidentiality; missing spatial data

摘要

遗失数据和区位误差对于根据雷普利 K 函数的空间集中测量之影响。 *Spatial Economic Analysis*。基于雷普利 K 函数的测量，是检验个别行动者在经济空间中的集中度时偏好的工具。但在诸多经验案例中，却因遗失的数据或行动者地点的不确定性，导致数据集包含了各种不精确性。这些不精确性对 K 函数的影响至今却所知甚少。本文透过由蒙地卡罗实验所支持的理论分析，对此一问题提出洞见。研究结果显示，集群或抑制的模式，或许不应被视为真实的现象，而仅是数据不完美的影响。


关键词

空间微观计量经济; 空间集中度; 雷普利 K 函数; 地理模糊处理; 匿名性; 遗失的空间数据


CONTACT

^a (Corresponding author)  giuseppe.arbia@unicatt.it


Catholic University of the Sacred Heart, Rome, Italy.

^b  giuseppe.espa@unitn.it


University of Trento, Trento, Italy.

^c  diego.giuliani@unitn.it

University of Trento, Trento, Italy.

^d  mariamichela.dickson@unitn.it

University of Trento, Trento, Italy.

 Supplemental data for this article can be accessed [10.1080/17421772.2017.1297479](https://doi.org/10.1080/17421772.2017.1297479)

RÉSUMÉ

Effets de données manquantes et d'erreurs de localisation sur les mesures de concentration spatiale basées sur la fonction K de Ripley. *Spatial Economic Analysis*. Les mesures fondées sur la fonction K de Ripley sont les outils de choix pour tester la concentration d'agents individuels dans un espace économique. Toutefois, dans un grand nombre de cas empiriques, les ensembles de données contiennent différentes inexactitudes attribuables à des données manquantes ou des incertitudes concernant la localisation des agents. Jusqu'à présent, les effets de ces inexactitudes sur la fonction K étaient bien peu connus. La présente communication fait la lumière sur ce problème, par le biais d'une analyse théorique qui s'appuie sur des expériences de Monte Carlo. Les résultats indiquent que l'on peut observer des modèles de regroupement ou d'inhibition non pas comme un authentique phénomène, mais simplement comme l'effet d'imperfections des données.

MOTS-CLÉS

micro-économétrie spatiale; concentration spatiale; fonction K de Ripley; géo-masquage; confidentialité; données spatiales manquantes

RESUMEN

Efectos de datos incompletos y errores de ubicación en las mediciones de concentración espacial basadas en la función K de Ripley. *Spatial Economic Analysis*. Las mediciones basadas en la función K de Ripley son las herramientas preferidas para comprobar la concentración de los agentes individuales en un espacio económico. Sin embargo, en muchos casos empíricos los grupos de datos contienen diferentes imprecisiones debido a una falta de datos o incertidumbre sobre la ubicación de los agentes. Hasta ahora se sabe poco sobre los efectos de estas imprecisiones en la función K . Este artículo arroja luz sobre el problema mediante un análisis teórico respaldado por experimentos de Monte Carlo. Los resultados muestran que los patrones de aglomeración o inhibición se podrían observar no como un fenómeno genuino sino como el efecto de las imperfecciones de los datos.

PALABRAS CLAVES

microeconomía espacial; concentración espacial; función K de Ripley; geo-enmascaramiento; confidentialidad; datos espaciales incompletos

JEL C2; C21; O1; O12; R1; R12

HISTORY Received May 2016; in revised form January 2017

INTRODUCTION

The analysis of the spatial concentration of economic agents is central to many microeconomic problems and may be invaluable in suggesting theoretical hypotheses concerning the role of economic agglomerations and the interactions among increasing returns, transportation costs and the location of production factors. For instance, in the analysis of firm demography, firm clustering can be explained by the Marshall–Arrow–Romer (MAR) externalities hypothesis (Arrow, 1962; Marshall, 1890; Romer, 1986), according to which the close proximity of similar firms increases the chances of human interaction, labour mobility and knowledge exchange, which, in turn, affect firm creation, development and survival. Most of the earlier empirical contributions on spatial concentration have considered data aggregated at a regional level, resulting in different ways of measuring the phenomenon (e.g., Arbia & Piras, 2009; Ellison & Glaeser, 1997; Maurel & Sédillot, 1999) and contradictory empirical results (Henderson, 2003; Mansfield, 1995; Rosenthal & Strange, 2003). In most of this literature, due to the aggregation of data at the regional level, conclusions are based on arbitrary definitions of the adopted jurisdictional spatial units, such that aggregating them in different ways produces different evidences. This is the essence of the so-called modifiable areal unit problem (MAUP) (Arbia, 1989). Furthermore,

theoretical models of the spatial behaviours of economic agents are generally grounded in the behaviours of individual economic agents (e.g., Hopenhayn, 1992; Krueger, 2003; Lazear, 2005) and can be tested empirically on regional aggregates only under the unrealistic assumption of homogeneous behaviours among regional agents. Indeed, the lack of theories to support regional econometric modelling is one of the deepest criticisms of spatial econometrics inference (Corrado & Fingleton, 2012; Pinkse & Slade, 2010), on the basis that, at most, regional data can help identifying technical relationships, with little or no possibility of drawing conclusions about cause–effect relationships.

The search for a statistical solution to the MAUP has led in recent years to an increasing interest in the use of micro-data and establishment-level information, thus eliminating the problem of defining a priori the level of geographical partition. This approach allows the use of individual-level variability in spatial concentration in order to test the various economic theories and to reconcile the contrasting empirical evidence found at the aggregate level. In this respect, the use of point pattern analysis methods (initially proposed by Ripley, 1977, and originally applied to epidemiological and ecological phenomena; Diggle, 2003; Getis & Boots, 1978; Getis & Franklin, 1987) was more recently extended to the analysis of economic phenomena by the contributions of Marcon and Puech (2010), Duranton and Overman (2005), and Arbia, Espa, and Quah (2008), amongst others. This gave rise to a new branch of econometrics sometimes referred to as *spatial microeconometrics* (Dubé & Legros, 2014). The use of micro-data in spatial microeconomics has also been fostered by the large and increasing availability of detailed georeferenced databases. For instance, the US Census Bureau's Longitudinal Business Database (LBD) contains annual observations for approximately 4 million firms and 70 million employees (Glaeser & Kerr, 2009). Similarly, the recently launched Billion Prices Project (Cavallo & Rigobon, 2016) collects daily geolocated retail prices from a large number of global retailers. However, while solving the MAUP, the use of granular micro-data raises entirely new problems of data quality that require proper evaluation. Indeed, when using individual geocoded data, we almost invariably encounter various forms of data imperfections (e.g., data based on a sample, missing data and data containing attribute and/or local errors) that can mask the real phenomena, even to the point of dramatically distorting inferential conclusions. This paper seeks to describe the effects of these imperfections in studying the spatial concentration of economic agents using state-of-the-art microgeographic measures, such as Duranton and Overman's (2005) and Marcon and Puech's (2010) indices. Since these measures are essentially extensions of Ripley's basic K -function, for simplicity and without loss of generality, we focus our discussion mainly on this tool. Consequently, the paper is structured as follows. The next section contains a short description of the K -function as a tool for testing the null hypothesis of random spatial distribution versus the alternative hypotheses of concentration or dispersion. The third section describes in more detail the data-quality issues that will be analysed, namely: missing data and locational errors. The fourth section discusses some of the theoretical expectations of estimating a K -function in the presence of data inaccuracies. The fifth section reports the results of a series of Monte Carlo simulations to support the theoretical expectations.

MEASURING SPATIAL CONCENTRATION USING MICROGEOGRAPHIC DATA: THE K -FUNCTION

Generally, in spatial statistics a point process is almost completely identified by its first- and second-order properties. The first-order properties can be described by the so-called intensity function $\lambda(x)$, defined as:

$$\lambda(x) = \lim_{|dx| \rightarrow 0} \{(E[N(dx)])/(|dx|)\},$$

where x denotes the coordinates of an arbitrary point; $N(\cdot)$ denotes the number of points in an infinitesimal area; and $|\cdot|$ denotes the surface area (Diggle, 2003). The second-order properties, in turn, can be described by the so-called second-order intensity function $\lambda_2(x, y)$, defined as:

$$\lambda_2(x, y) = \lim_{|dx|, |dy| \rightarrow 0} \{(E[N(dx)N(dy)])/(|dx||dy|)\},$$

where x and y denote the coordinates of two distinct and arbitrary points.

Formally, the K -function is an alternative description of the second-order properties of a spatial point process that can be adopted under the assumption of stationarity and isotropy (Diggle, 2003). A stationary and isotropic spatial point process is such that $\lambda_2(x, y) = \lambda_2(d)$, where $d = x - y$ is the Euclidean distance between the arbitrary locations x and y . A process of this kind is, therefore, characterized by the fact that the spatial interactions among points (which may identify the locations of certain kinds of economic agents) depend only on their distance and not on their specific positions. Furthermore, the second-order properties of such a process can be properly described by the K -function, which can be heuristically defined as follows:

$$K(d) = \lambda^{-1} E\{\text{number of further points falling at a distance } \leq d \text{ from an arbitrary point}\} \quad (1)$$

where λ denotes the first-order intensity (which is constant because of the stationarity assumption), which corresponds to the mean number of events per unitary area. Therefore, $\lambda K(d)$ indicates the expected number of farther points up to a distance d of a typical point (Ripley, 1977). In the empirical economic analyses, in which the data-generating point process is stationary and isotropic (i.e., when the territory is essentially homogeneous), the K -function properly quantifies the mean (global) level of spatial interactions among economic agents (e.g., firms or consumers) at each distance d .

The K -function is easier to estimate than $\lambda(d)$ and can be used to develop a formal test for the presence of spatial concentrations of points (e.g., firms) in a study area. When points are distributed at random in a study area, we say that they represent the hypothesis of complete spatial randomness (CSR). Formally, an observed CSR pattern is considered a realization of a homogeneous Poisson process (Diggle, 2003). Then, the $K(d)$ function is simply equal to the surface area of a circle of radius d . Consequently, the benchmarking value $K(d) = \pi d^2$ represents the null hypothesis of a random spatial distribution of points. Significant deviations from this reference value would favour the alternative hypothesis of spatial dependence. In particular, when $K(d) > \pi d^2$, we have a positive spatial dependence, or concentration, in which points tend to attract one another. Conversely, when $K(d) < \pi d^2$, we have evidence of a negative dependence and, hence, an *inhibition* or over-dispersion, in which points tend to inhibit one other (Figure 1). An unbiased estimator for $K(d)$ can be defined as follows:

$$\hat{K}(d) = \frac{|A|}{n(n-1)} \sum_{\langle i,j \rangle} \sum_{\langle l,m \rangle \neq \langle i,j \rangle} I(d_{\langle i,j \rangle, \langle l,m \rangle} < d) w_{\langle i,j \rangle, \langle l,m \rangle}^{-1} \quad (2)$$

where $|A|$ is the total surface of the study area A ; n represents the number of observed points; $d_{\langle i,j \rangle, \langle l,m \rangle}$ is the distance between the coordinate point $\langle i, j \rangle$ and the coordinate point $\langle l, m \rangle$; and $I(d_{\langle i,j \rangle, \langle l,m \rangle} < d)$ is an indicator function, such that $I = 1$ if $d_{\langle i,j \rangle, \langle l,m \rangle} < d$, and 0 otherwise. The weight $w_{\langle i,j \rangle, \langle l,m \rangle}$ is the proportion of the circumference of the circle centred on the coordinate point $\langle i, j \rangle$ and passing through the coordinate point $\langle l, m \rangle$, which lies within the study area (Boots & Getis, 1988). It must be introduced to mitigate any edge effects arising from the finite quality of the study area. A significance test to determine whether the events of interest

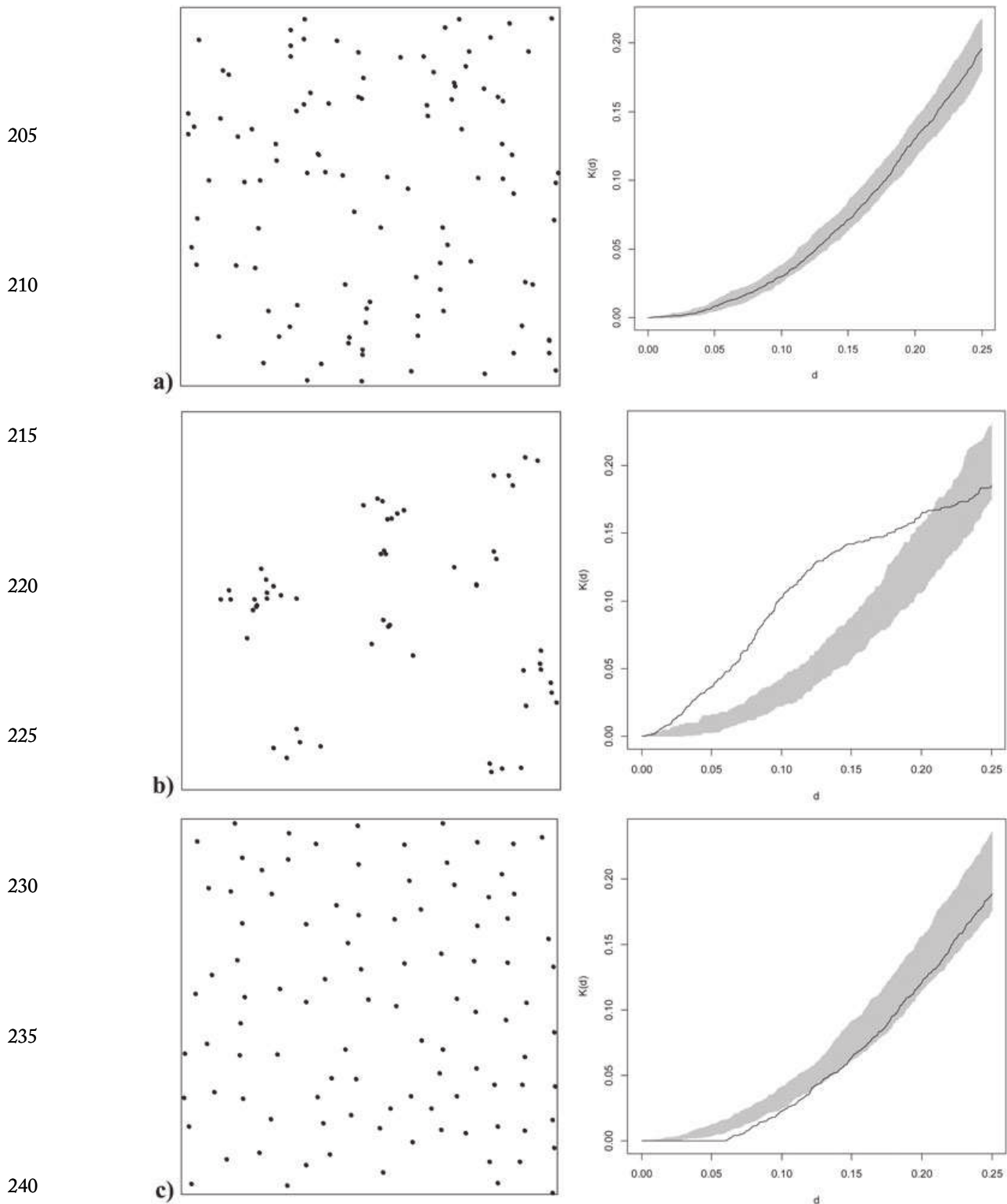


Figure 1. Map representation (first column) and corresponding K -function (second column) of three stylized spatial patterns of points in a unit square area: roes (a) complete spatial random pattern; (b) clustered pattern; and (c) inhibition pattern.

tend to concentrate in space may involve verifying whether, for some distance d , the functional $K(d)$ estimated on the observed point pattern is significantly different from πd^2 . Since the exact distribution of $K(d)$ is unknown, the test of the null hypothesis of the absence of spatial concentration is based on Monte Carlo-simulated confidence envelopes (Besag & Diggle, 1977).

MISSING SPATIAL DATA, UNCERTAIN GEOCODING AND ARTIFICIALLY INDUCED LOCATIONAL ERROR

When dealing with spatial micro-data, researchers often face datasets affected by inaccuracies in spatial information. On this topic, we often observe a certain degree of ambiguity in the literature, which requires some clarification. Indeed, in practical cases, problems of accuracy manifest in many different ways (Arbia, Espa, & Giuliani, 2016; Zimmerman, 2008): three of them are addressed here. First, the *missing spatial data problem* refers to a situation in which both the location and some measurements are unknown. We know of the presence of some individuals in a certain area, but we ignore exactly where they are and lack information about some or all their characteristics. Some individuals are simply not observed on the study area map. This situation is not uncommon in surveys in developing countries.

Second, *intentional locational errors* emerge when both the locations and the measurements of single individuals are known, but their position are geomasked a posteriori in order to preserve respondents' confidentiality.

Finally, *unintentional positional errors* (Arbia et al., 2016) refer to situations in which observations of individuals are available, but the individuals' locations are not known with certainty. For instance, we have a list of firms in a small area (e.g., a census tract) and observations on some of their statistical characteristics, but we do not know their exact addresses within the area. In such cases, it is common to assign individuals to the centroid of each area; however, this procedure generates uncertainty in the geocoding, producing positional errors.

The use of the K -function, as well as its more recent extensions, to characterize the spatial concentrations of economic agents is obviously affected by these data-quality issues. The aim of this paper is to make researchers aware of these problems and to understand how they can mask and distort real phenomena in practical cases. As far as we know, apart from an interesting study by Fehmi and Bartolome (2001) in the context of ecological analysis, no systematic study has yet been conducted on this topic.

SOME THEORETICAL RESULTS

The effect of missing data points at random

In order to examine the effects of missing data on the estimation of the K -function, consider two generic points observed on the real plane, characterized respectively by the coordinates $\langle i, j \rangle$ and $\langle l, m \rangle$, so that the true pairwise inter-point Euclidean squared distances between the two generic points can be defined as $d_{\langle i, j \rangle, \langle l, m \rangle}^2 = (i - l)^2 + (m - j)^2$. Also, define the sample average distance as:

$$\bar{d}^2 = \frac{\sum_{\langle i, j \rangle} \sum_{\langle l, m \rangle} d_{\langle i, j \rangle, \langle l, m \rangle}^2}{n(n-1)/2} \quad (3)$$

Furthermore, consider first a case in which the observed points form a CSR pattern and the process of data elimination is completely random. Also assume, without loss of generalities, that we observe a random realization within a study area represented by a unitary quadrat defined with x - y coordinates ranging between -0.5 and 0.5 . In this case, it can be proven that (see Appendix A1 in the supplemental data online for a proof):

$$E(\bar{d}^2) = \frac{1}{3} \quad (4)$$

and, under the (often admittedly unrealistic) assumption of the independence of the distances:

$$\text{Var}(\bar{d}^2) \cong \frac{7}{90} \frac{2}{n(n-1)} \quad (5)$$

305 The presence of a positive correlation among the distances would further increase the variance and, consequently, the uncertainty. Finally, due to the central limit theorem, we also have:

$$\bar{d}^2 \approx N\left(\frac{1}{3}; \frac{7}{90} \frac{2}{n(n-1)}\right). \quad (6)$$

310 Equations (4) and (5) show that the expected sample distance is a constant determined by the space characteristics, while the variance decreases with the square of the sample size. The form of the variance expressed by equation (5) is a mere approximation of the actual variance, as it neglects the correlations among the $(1/2)n(n-1)$ distances, but negatively depends on n . The effects of missing data correspond to a case in which the sample size is reduced from n to, say, m units ($m < n$). In this sense, equation (5) shows that if we reduce n by eliminating m observations at random, $\text{Var}(\bar{d}^2)$ will increase, as could be expected due to the reduction of information and the consequent increase in uncertainty. This result is in line with Lang and Marcon's (2013) derivation of the variance of the K -function in the context of rectangular study regions, which showed that the variance decreases as the number of missing points increases.

The effect of missing clustered data points

325 In real cases, the hypothesis of data missing completely at random is sometimes implausible, since missing data are often spatially clustered. This is because in many social surveys entire portions of space (e.g., neighbourhoods) cannot be observed and, in general, the difficulties encountered in collecting accurate data change smoothly over space. When missing data are spatially clustered, the effects on the K -function are easier to describe. In fact, in this case, groups of points observed
 AQ2 at small distances are systematically erased, and the K -function can be described as a truncated CDF with a left truncation at $d > t$, where t is a threshold distance corresponding to the cluster dimension. In the presence of missing clustered data, the K -function will show the emergence of
 330 spurious inhibitory patterns at low distances below the truncation point t , which are compensated by spurious clustered patterns at high distances.

The effects of intentional locational error

335 An intentional locational error, as previously mentioned, emerges when, in the effort to preserve respondents' confidentiality, data that are precisely geocoded during a survey are then geomasked, assuming a certain random criterion. For example, the Demographic and Health Surveys (DHS) programme (USAID, 2013) collects household data and georeferences the locations of the observational units using GPS receivers with a positional accuracy of 15 m or less. However, to preserve
 340 confidentiality, the coordinates are then misdisplaced according to a random mechanism. A common procedure used in these circumstances is known as the random direction, random distance method (Collins, 2011; USAID, 2013) or uniform geomasking (Arbia, Espa, & Giuliani, 2015). This mechanism involves transforming an individual's coordinates by displacing them along a random angle (say θ) and a random distance (say δ), both obeying a uniform probability law $\delta^{iid} \approx U(0, \delta^*)$ and $\theta^{iid} \approx U(0, 360^\circ)$, with δ^* representing the maximum distance error and
 345 with δ and θ being mutually independent. If d^2 is the generic true square distance between two points, the error-contaminated squared distances (say d^{*2}) under this hypothesis can be proved to be (see Appendix A2 in the supplemental data online for the proof):

$$E(d^{*2}) = d^2 \frac{2}{3} \delta^{*2} \quad (8)$$

350

Equation (8) shows that an intentional locational error produces an expectation of an increase in the pairwise distances. By increasing the inter-point distances on which the calculation of the K -function is based, the introduction of an intentional locational error will thus produce spurious inhibitory patterns at all distances and the cancellation of true clustered patterns at small distances. The effect is expected to increase monotonically (i.e., with the maximum displacement distance in the random geomasking mechanism) with δ^* .

The effect of unintentional locational error

Finally, consider the case of unintentional locational errors, which emerge when data are available in a small area (e.g., a census tract), but the exact position of each individual within the study area is unknown. In such cases, a common procedure is to solve the problem by assigning each individual to the centroid of his or her area. For instance, in Italy, the National Statistical Institute (ISTAT) collects and disseminates data related to active firms (the ASIA archive; Cozzi & Filippini, 2012), including information on a set of economic variables and the geographical locations (in terms of latitude–longitude spatial coordinates) of the plants. Spatial coordinates are identified automatically based on street addresses; however, in some instances, they contain location errors. For example, during 2004–09, 15% of plants could not be precisely geocoded and had to be assigned to approximated geographic locations: 12% to the centroids of their municipalities (average surface 37 km²) and 3% to the centroids of the census tracks (the more accurate solution; average surface 1 km²). To study the effect of this procedure, we can approximate the district to which points are allocated (i.e., the census tract) to a square of side $l = \sqrt{S}$, where S is the surface area of the district. The observed distance following reallocation to the centroid (called d^{*2}), thus, introduces a maximum error of $l\sqrt{1/2}$. As a consequence, from equation (8), the expected contaminated distance is $d^{*2} \approx U(0, l\sqrt{1/2})$. Following the same reasoning presented above, this produces an inhibitory effect at small distances, approximately given by:

$$E(\bar{d}^2) = d^2 + \frac{2}{3}l\sqrt{\frac{1}{2}} = d^2 + \frac{l\sqrt{2}}{3}$$

with the effect increasing proportionally to l . Similarly, if we approximate the district using circles, the result is $d^{*2} \approx U(0, r)$. Here, r represents the radius of the circle, which can be approximately given by $r = S/2\pi$, where S is the surface area of the district. This produces an inhibitory effect governed by:

$$E(\bar{d}^2) = d^2 + \frac{2}{3}r$$

with the effect increasing proportionally with r .

However, the reallocation of points to the centroids of the districts also produces a clustering effect, since agents located at different points within the same district are all concentrated in the same centroid. This effect also increases with l (or r). In fact, when a geographical system comprises a small number of large districts, l (and r) will be larger, and several points will be concentrated in the same centroid. Thus, we expect the clustering effect to prevail on the inhibitory effect. The magnitude of the effect obviously also depends on the proportion of points affected by unintentional error.

SIMULATION EXPERIMENTS

Missing data

In this section the theoretical results presented above will be substantiated by a set of Monte Carlo experiments. First, in order to assess the consequences of clustered missing data on the

estimation of the K -function and the performance of the CSR test, 1000 complete spatially random point patterns of economic activities distributed on a unit square were generated as independent realisations of a homogenous Poisson process conditioned on 100 points. Setting the size of each pattern equal to a fixed number of economic activities makes the patterns directly comparable. In fact, simulating the homogenous Poisson process conditioning on a fixed total number of points actually yields the simulation of a binomial point process, which is a suitable approximation of the CSR hypothesis. From each of these artificial point patterns, which represent the unobserved point processes, new random datasets affected by missing spatial locations are obtained by erasing points from the clean point patterns according to a random mechanism mimicking missing clustered data. Specifically, locations are deleted randomly with a probability proportional to the values generated by a Gaussian random field with a variance of 0.5 and the exponential covariance function $\rho(d) = \exp(-d/0.25)$, where d represents distance (Figure 2).

For each missing dataset thus obtained, we estimate $K(d)$ and perform the CSR test at a 1% significance level. We then compute the bias and root mean square error (RMSE) of $\hat{K}(d)$ with respect to the true values and the type I error rate for the null hypothesis of CSR. In particular, if we use $\bar{K}_i = \sum_d K_i(d)$ to denote the sum over d of all values of $\hat{K}(d)$ for the i -th simulated pattern affected by missing locations, and $\bar{K}_{\text{TRUE}} = \sum_d \pi d^2$ to denote the sum over d of all values of $K(d)$ for the true 'clean' CSR pattern, we can define the bias and RMSE as follows:

$$\text{Bias} = \frac{\sum_i (\bar{K}_i - \bar{K}_{\text{TRUE}})}{1000}, \quad \text{RMSE} = \sqrt{\frac{\sum_i (\bar{K}_i - \bar{K}_{\text{TRUE}})^2}{1000}}.$$

The type I error rate is computed at a significance level of 1% as the proportion of simulated patterns with missing locations for which $\hat{K}(d)$ deviates the 1% CSR confidence bands at some values of d .

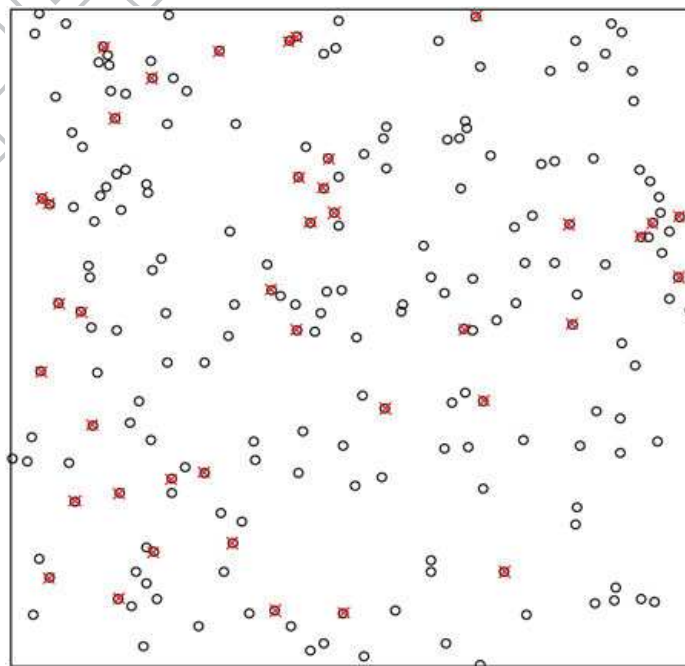
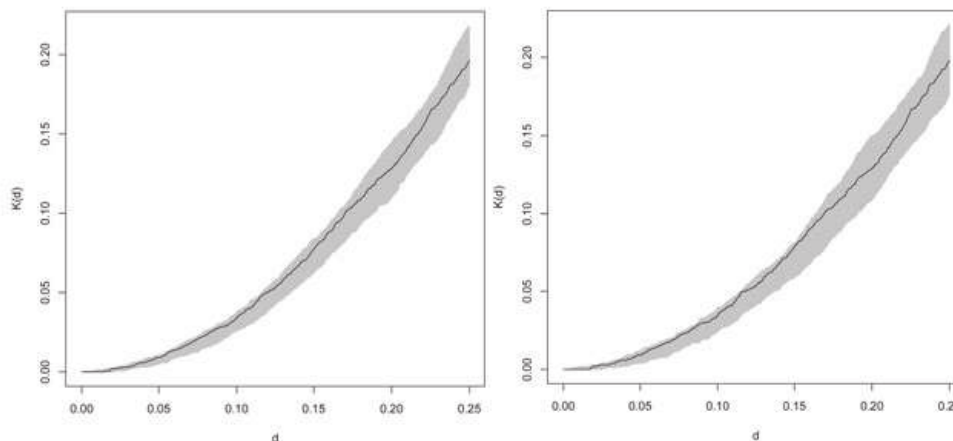


Figure 2. Pattern of clustered missing points in a unit square area. Crossed points are unobservable. In this simulation, when a point is randomly deleted, other points in the neighbourhood are also cancelled with a probability proportional to $\rho(d) = \exp(-d/0.25)$.

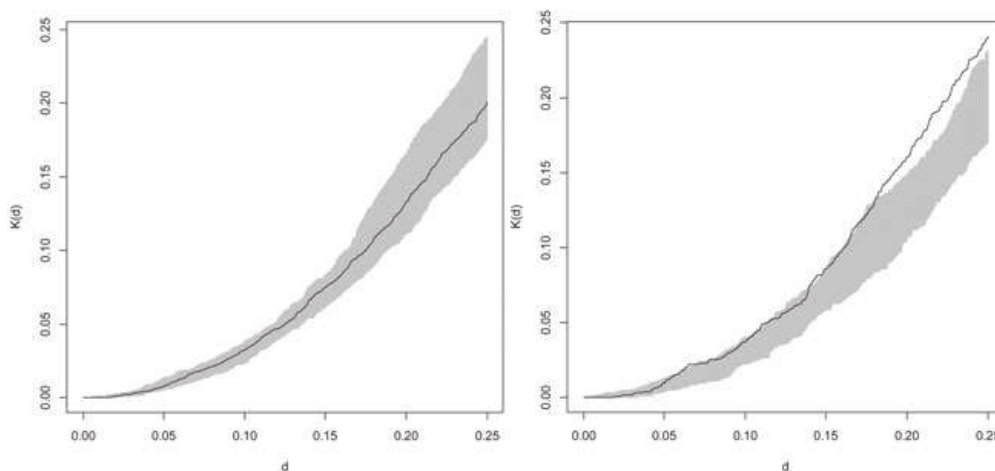
PMP = 10%



a) True pattern

b) Pattern with missing locations

PMP = 40%



a) True pattern

b) Pattern with missing locations

locations

Figure 3. Effects of missing clustered data on the K -function. Empirical K -function and simulated 1% confidence bands: (a) true pattern; and (b) pattern with missing locations.

To illustrate visually the effects of missing clustered data, Figure 3 compares the graphs of the CSR test based on the K -function for both a simulated pattern and the same pattern affected by clustered missing data with a given proportion of missing points (PMP). Confirming the above theoretical findings, the graphs suggest that the presence of clustered missing data does not produce any significant effects in the estimation of the K -function when $PMP = 0.1$; however, when $PMP = 0.4$, it tends to overestimate the K -function at high distances and, as a consequence, tends to suggest the presence of spurious patterns of spatial concentration. The expected emergence of spurious inhibitory patterns at low distances (see above) is less evident within our simulation parameters and does not produce significant results.

Table 1 summarizes the main simulation results for different levels of PMP and shows that the presence of missing data leads to an overestimation of the K -function. Both the bias and the RMSE of the estimated K -function increase with the percentage of missing points, while the type I error rate is substantially unaffected by the presence of missing data. This result reflects the particular random mechanism chosen for the missing data, which tends to create 'holes' of points in the patterns. While the number of holes affects the precision of the estimates, it does not alter the global spatial pattern and, hence, does not influence the type I error rate.

Table 1. Empirical bias, root mean square error (RMSE) of the K -function estimator and type I error rate for the complete spatial randomness (CSR) test with clustered missing data.

PMP	Missing spatial information		
	Bias	RMSE	Type I error rate
0.1	0.148	1.648	0.345
0.2	0.240	1.875	0.317
0.3	0.441	2.325	0.345
0.4	0.919	2.808	0.303

Intentional locational error

To corroborate the above findings concerning the effects of intentional locational errors on the estimation of the K -function and the associated inference, we again generate 1000 complete spatial random point patterns of 100 points each on a unit square. In this case, however, for each artificial 'true' point pattern we displace the observations using a random mechanism. In particular, we use a procedure that replicates the method described by, for example, USAID (2013). In each simulated point pattern, for each point's location we select a random angle and a random distance. The random angle is generated according to a uniform distribution $U(0, 360^0)$, while the random distance is extracted from a uniform distribution $U(0, \delta^*)$, where δ^* is a simulation parameter that can vary over a fixed range. In our simulations, we let δ^* vary between 0.05 and 0.25, considering that, in the unit square, the maximum inter-point distance equals the diagonal, or $\sqrt{2}$. Thus, the maximum displacement errors we consider range between approximately 3% and 17% of the maximum distance. As in case of missing data, for each *geomasked* point pattern, we estimate $K(d)$ and perform a Monte Carlo-based CSR test at a 1% significance level. We then compute the RMSE of $\hat{K}(d)$ obtained over the simulation runs. From the above, we expect the RMSE to increase with the maximum displacement distance. Similarly, in order to analyse the effects of geomasking on the CSR test, for each simulation run we calculate the significance of the test and again reject the null if $\alpha \leq 0.01$. We can then compute the number of times we wrongly reject the true null hypothesis $K(d) = \pi d^2$ and monitor how the type I error rate, thus evaluated, changes by increasing the parameter δ^* . Figure 4 illustrates the effects of locational errors due to intentional geomasking for two different values of the maximum displacement distance; Table 2 summarizes all the results of the Monte Carlo experiments.

As expected from the theoretical findings, Figure 6 clearly shows the presence of spurious inhibitory patterns at different distances when a location error is introduced and the maximum displacement distance is $\delta^* = 0.25$. Similarly, Table 2 provides evidence that the empirical bias is negative for all values of the geomasking maximum displacement distance, δ^* , thus confirming that errors due to intentional geomasking tend to underestimate the degree of spatial concentration and to introduce spurious patterns of inhibition. As expected, increasing the maximum displacement distance increases both bias and RMSE in absolute terms. Furthermore, the probability of type I errors almost doubles when moving from $\delta^* = 0.05$ to 0.25. As a consequence, preserving confidentiality by means of a random displacement mechanism has the potential to conceal the presence of spatial clusters and to provide inaccurate evidence toward inhibitory patterns.

Figure 6(d) suggests, however, that the effects of these kinds of locational errors on the estimate of $K(d)$ may mitigate when $d > \delta^*$, since it shows that $\hat{K}(d)$ returns within the confidence bands as d approaches the value of δ^* . To explore this issue better, the results of the simulations are also analysed separately below and over the geomasking distance. In particular, Table 3 reports the type I error rates of the CSR test, computed distinctly for the two cases, and confirms that the

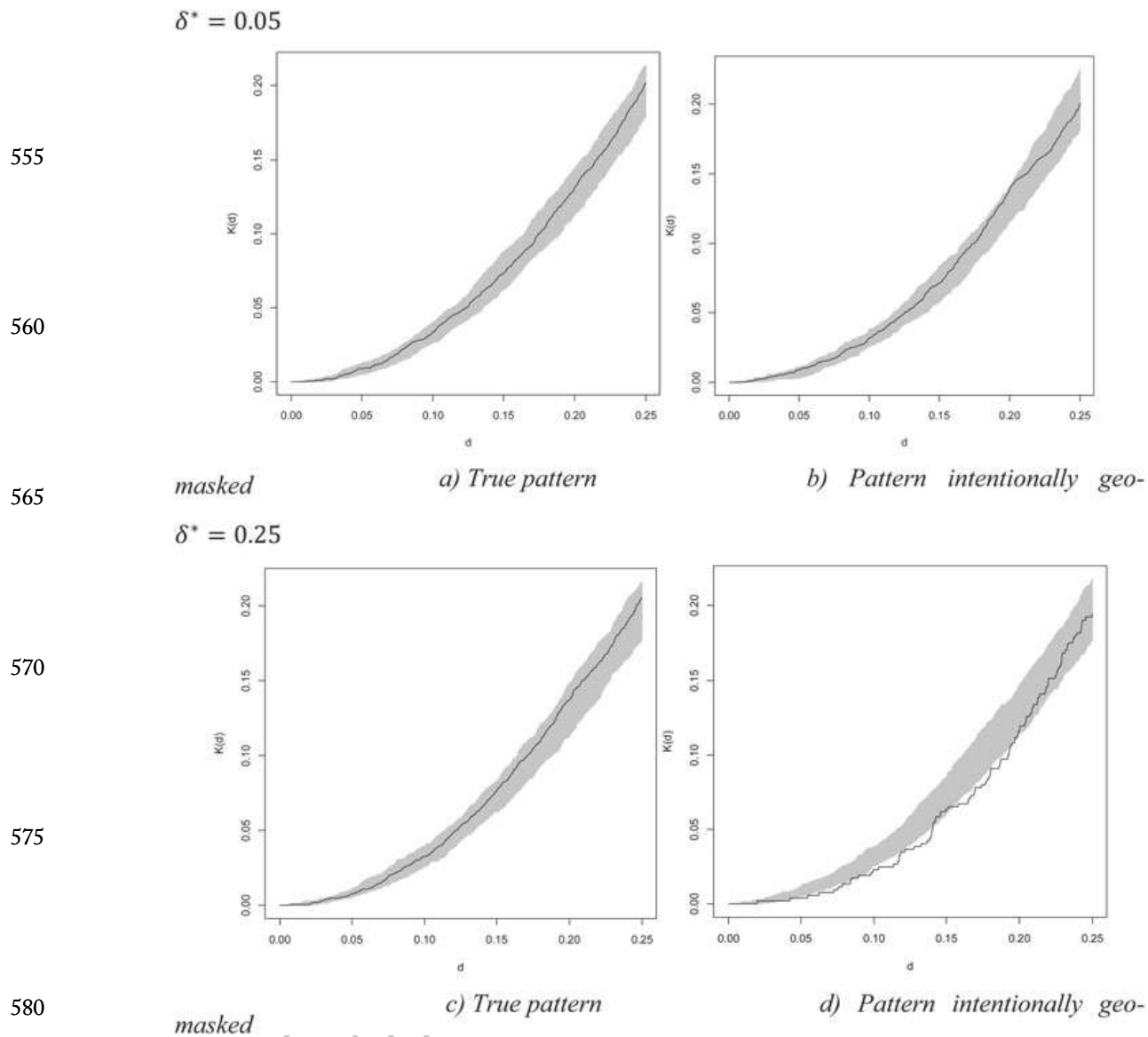


Figure 4. Effects of intentional locational errors on the complete spatial randomness (CSR) test. Empirical K -function and simulated 1% confidence bands: (a) true pattern; and (b) pattern intentionally geomasked.

consequences for inference are less dramatic. This evidence is interesting because it suggests that even when data are affected by locational errors at small distances, the analysis at wider distances may still be valid.

Table 2. Empirical bias, RMSE of the K -function estimator and type I error rate for the CSR test under locational errors generated by random geomasking.

δ^*	Intentional locational error		
	Bias	RMSE	Type I error rate
0.05	-0.447	1.487	0.379
0.10	-1.368	2.066	0.435
0.15	-2.814	3.275	0.560
0.20	-4.518	4.838	0.609
0.25	-6.110	6.380	0.668

Table 3. Type I error rate for the CSR test under locational errors generated by random geomasking computed distinctly for $d \leq \delta^*$ and $d > \delta^*$

δ^*	Type I error rate $d \leq \delta^*$	Type I error rate $d > \delta^*$
0.05	0.259	0.232
0.10	0.346	0.229
0.15	0.438	0.208
0.20	0.556	0.194

Unintentional locational error

Finally, to assess the consequences of uncertain geocoding on the estimation of the K -function and the performance of the CSR test, we follow a procedure similar to that used above. First, we generate 1000 CSR point patterns, each with 100 points. Then, from each of these artificial patterns we obtain a new dataset by selecting a proportion of cases (PMP) affected by unintentional locational error and assign each of the selected points to the centroid of its original subarea. The subareas are defined by partitioning the unit area into a 4×4 regular square lattice grid. Again, locations affected by the unintentional locational error are randomly selected with a probability proportional to the values generated by a Gaussian random field with a variance of 0.5 and the exponential covariance function $\rho(d) = \exp(-d/0.25)$, where d represents distance (Figure 4). However, instead being simply deleted, the points are reassigned to the centroid. We then again estimate the $K(d)$ function and perform the Monte Carlo-based CSR test at a 1% significance level. Figure 5 compares the CSR test in a simulated 'true' pattern with its counterpart in a pattern affected by unintentional locational errors. The graphs suggest that in our Monte Carlo experiment the tendency to create spurious clustering patterns prevails over the opposite tendency to produce inhibition patterns due to the use of coarse partitioning in which all affected points concentrate around only 16 centroids. The effect is obviously clearer when PMP = 0.4, in which case we move 40 points and distribute them across only 16 centroids.

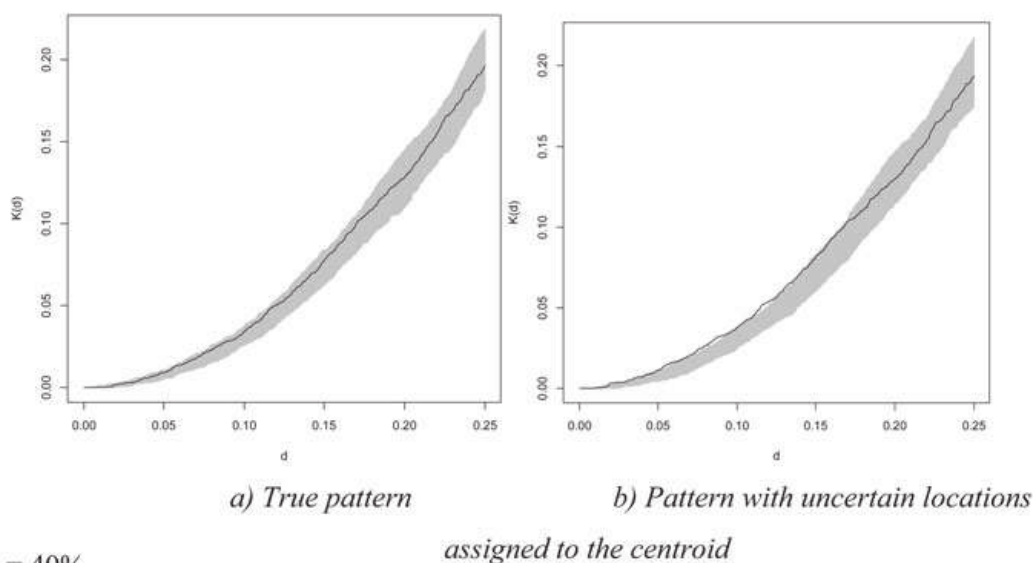
Table 4 summarizes the main Monte Carlo results for different proportions of relocated points (PMP). Apart from the usual (and theoretically hypothesized) result of an increase in PMP increasing the absolute bias and the RMSE, Table 4 also shows that in our simulation experiment the consequences of unintentional locational errors are more severe than those observed with missing data and intentional locational errors. Specifically, the type I error rate for the CSR test increases nearly to 1 when 20% of points are reallocated to the centroids of the square lattice grid. This happens when the reference partition is very coarse, as in our simulation experiment.

However, Figure 5(d) suggests that the undesirable effects of unintentional locational errors on $\hat{K}(d)$ may decrease as d increases over the order of magnitude of the errors. In this simulation's settings, in which error-affected locations are moved to the centroids of their respective squared subareas, the maximum possible locational error is equal to half the diagonal of the squared subarea. Since the subareas are here defined by partitioning the larger unit area into a 4×4 regular square lattice grid – and, hence, consist of squares sized 0.25×0.25 – the maximum possible locational error is $(1/4\sqrt{2})/2 \cong 0.177$. Figure 5(d) shows, indeed, that $\hat{K}(d)$ returns within the confidence bands after approximately this value. Corroborating this insight, Table 5 reports the type I error rates of the CSR tests for $d \leq 0.177$ and $d > 0.177$. The results summarized have important practical consequences since they indicate that analyses at distances consistently greater than the locational error may produce valid inferences.

Simulation experiments with real data

The simulation experiments conducted in the previous sections are based on simplified stylized spatial distributions of events. While they are useful for supporting the theoretical arguments

PMP = 10%



PMP = 40%

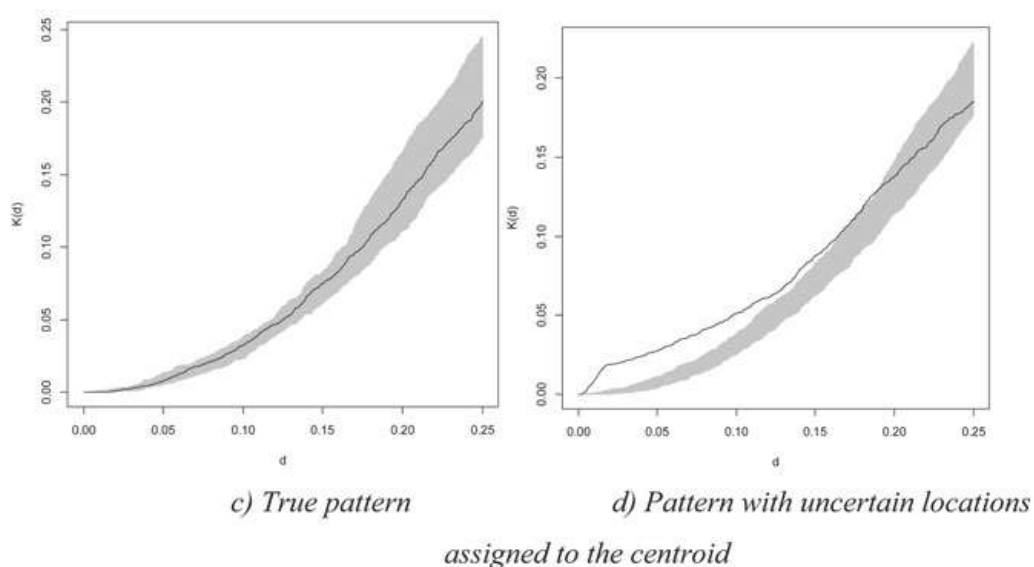


Figure 5. Effects of unintentional locational errors on the CSR test. Empirical K -function and simulated 1% confidence bands: (a) true pattern; and (b) pattern with uncertain locations assigned to the centroid.

Table 4. Empirical bias, RMSE of the K -function estimator and type I error rate for the CSR test under unintentional location errors.

PMP	Uncertain spatial information		
	Bias	RMSE	Type I error rate
0.1	0.145	1.446	0.572
0.2	0.194	1.423	0.994
0.3	0.317	1.513	1
0.4	0.200	1.489	1
1.0	-1.086	1.964	1

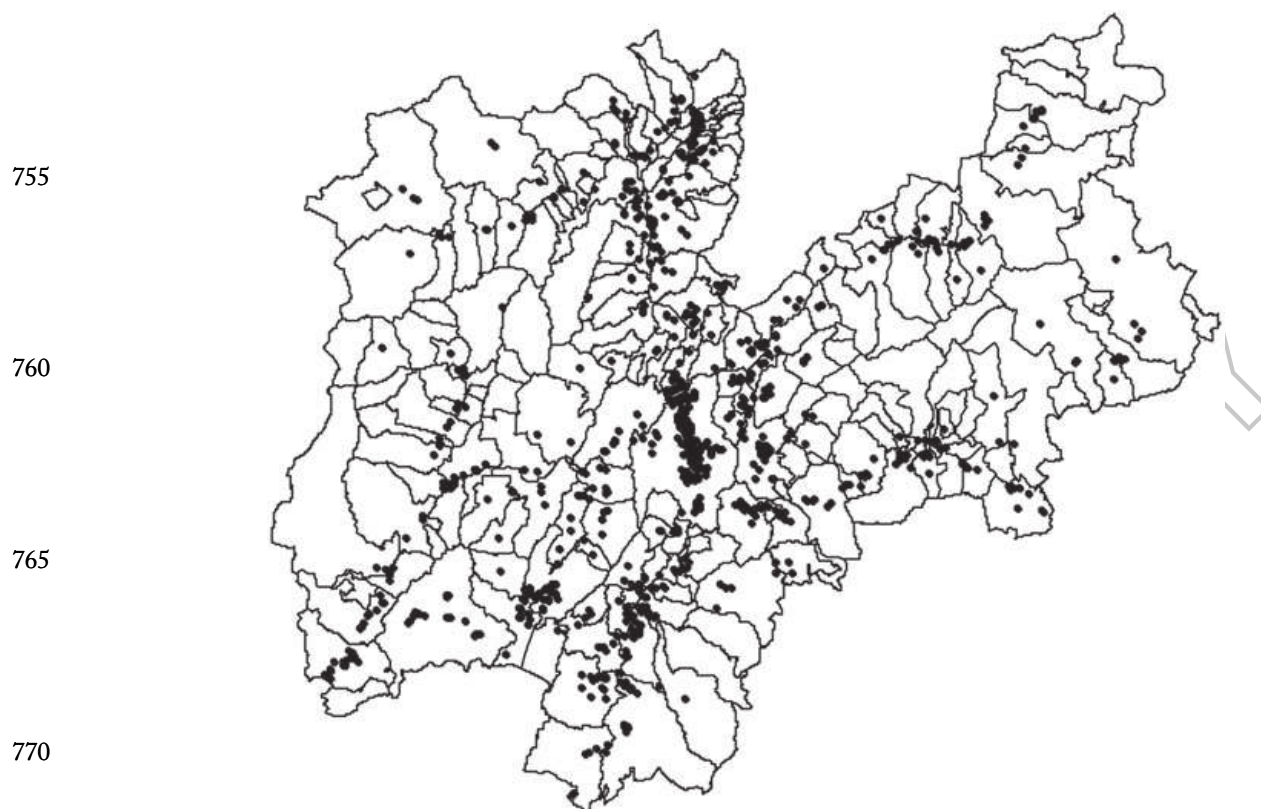
Table 5. Type I error rate for the CSR test under unintentional location errors computed distinctly for $d \leq 0.177$ and $d > 0.177$.

PMP	Type I error rate $d \leq 0.177$	Type I error rate $d > 0.177$
705 0.1	0.542	0.093
0.2	0.994	0.162
0.3	1	0.392
0.4	1	0.773
710 1.0	1	1

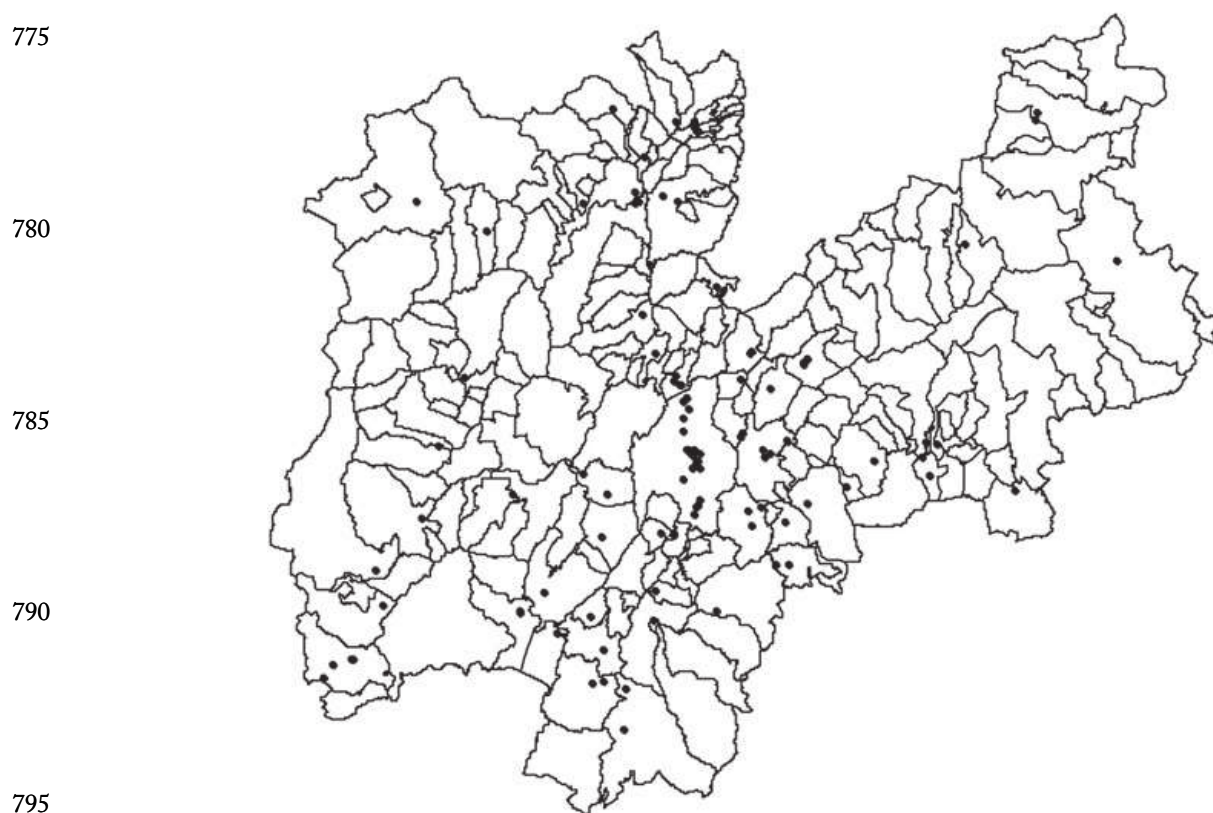
715 developed above, they do not properly represent real location patterns of economic agents. Indeed, the CSR hypothesis clearly represents the null hypothesis of the absence of spatial interactions amongst economic agents only under the assumption of spatial homogeneity, which is rarely attainable in practice. For example, in the context of the locations of production plants of a specific sector of activities, the locational choices of economic agents are typically constrained by environmental, administrative and economic limits that lead to violations of the CSR hypothesis even in absence of actual spatial concentrations or dispersions. In such cases, a more appropriate (and commonly used) null hypothesis is the absence of *relative* concentration or dispersion (Marcon & Puech, 2012). According to this benchmark, spatial concentration (or dispersion) of a specific industry is detected when its plants are more concentrated (or dispersed) than those of the whole economy. The most popular distance-based measures of relative spatial concentration in economic geography and spatial economics are Duranton and Overman's (2005) K_d and Marcon and Puech's (2010) M functions, which are essentially sophisticated extensions of the K -function to inhomogeneous spatial point patterns.

725 In this section, in order to assess the effects of missing data and locational errors on more realistic empirical situations, we consider the observed spatial distribution of the single-plant manufacturing firms located in the province of Trento (Italy) in 2009 (Figure 6). In particular, we assume that we are interested in studying the spatial pattern of the economic sector of food products (Figure 6(b)). Since the assumption of spatial homogeneity does not hold in this context, the CSR hypothesis is not an appropriate null hypothesis for the absence of spatial interactions amongst economic agents, and the K -function is not a suitable measure for detecting spatial concentration. Therefore, in order to detect the genuine spatial concentration generated by the interactions among food production firms, while controlling for the heterogeneity of their geographic territory, we must use the K_d or M functions, which use the spatial distribution of all manufacturing (Figure 6(a) as the null distribution). The graphs reported in Figure 7 show the behaviour of the estimated $K_d(d)$ and $M(d)$, along with the 99% confidence level simulation-based envelopes referring to the null hypothesis of absence of spatial dependence. The computations are done using the `dbmss` package (Marcon, Traissac, Puech, & Lang, 2015) of the R statistical software package.¹ We omit a detailed explanation and interpretation of the two methods while referring for this to other papers.² For the scope of the present paper it is sufficient to say that, like the CSR test based on the K -function, upward (downward) deviations of the estimated $K_d(d)$ and $M(d)$ functions indicate significant levels of relative spatial concentration. None of the graphs in Figure 7 shows significant deviations from the null hypothesis, implying that the spatial distribution of food production firms is not significantly different (i.e., more concentrated or more dispersed) from the spatial distribution of all manufacturing firms.

745 In order to evaluate the consequences of missing data on the estimation and inference of K_d and M , we generate 1000 spatial point patterns of firms affected by missing spatial locations. Furthermore, to control properly for the degree of missing data, each of these simulated patterns is first obtained by randomly erasing a given proportion (PMP) of food-production firms from



a) All manufacturing plants (1007 observations)



b) Plants from the food product sector (106 observations)

Figure 6. Location of 1007 single-plant manufacturing firms in the municipalities of the province of Trento (Italy) in 2009: (a) all manufacturing plants (1007 observations); and (b) plants from the food product sector (106 observations).

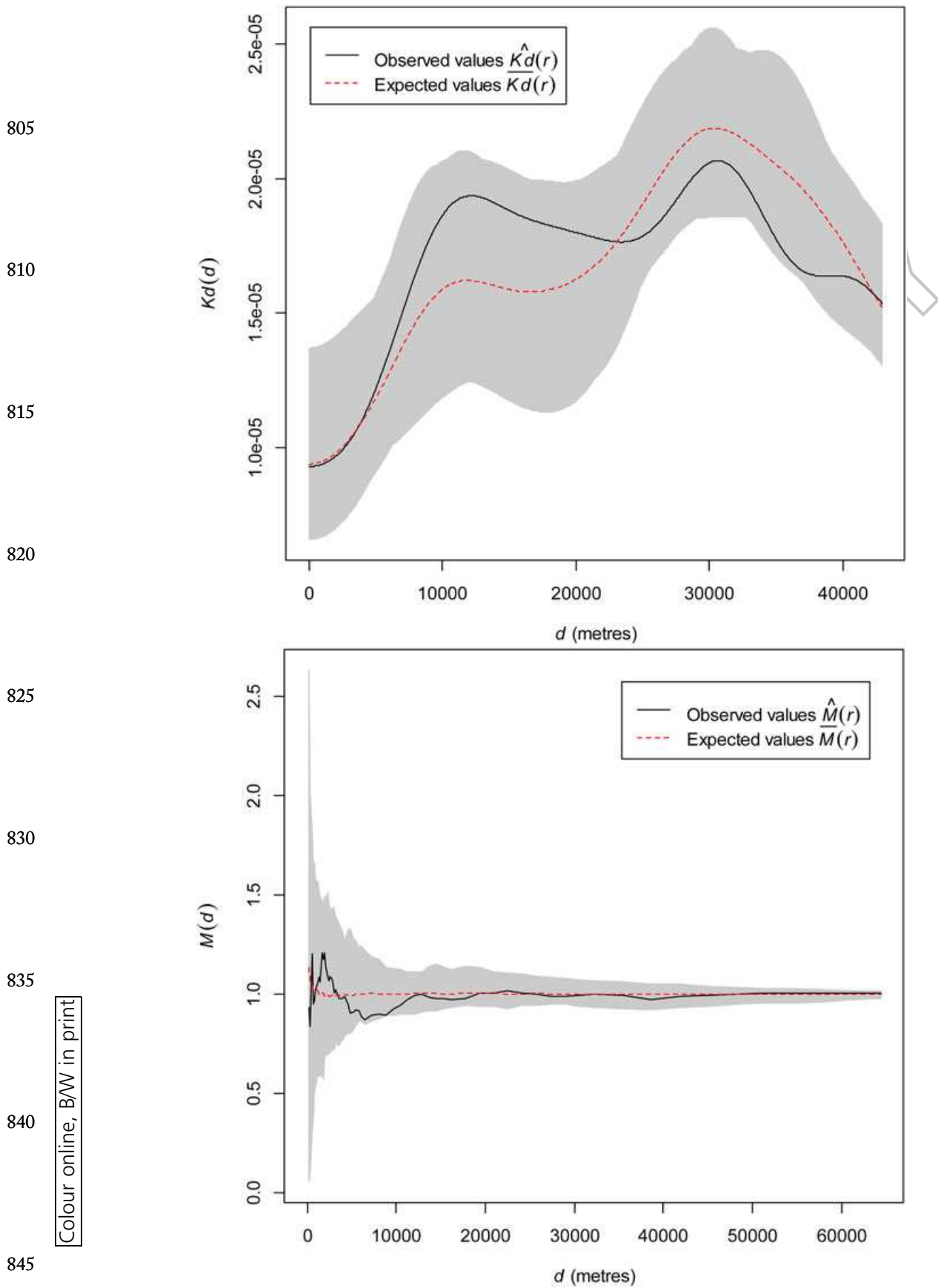


Figure 7. Behaviour of $\hat{K}_d(d)$ and $\hat{M}(d)$ and the corresponding 99.9% confidence bands for the food product sector in the province of Trento (Italy) in 2009.

the observed dataset (Figure 6(b)). Then, we randomly remove the same proportion (PMP) of firms from the other manufacturing sectors located in the same municipalities of the erased food production firms. In this way the firms of interest and their benchmark (i.e., all manufacturing firms) are affected by the same level of missing data within the same municipalities. Moreover, for both spatial patterns, we use the same random missing mechanism employed above based on a Gaussian random field.³

On the other hand, to evaluate the consequences of unintentional locational error, we create 1000 spatial point patterns of firms affected by uncertain geocoding by simply reassigning the previously erased firms to the centroids of the municipalities to which they belong.

For every spatial pattern of food production firms with missing and uncertain locations, we estimate K_d and M and perform the corresponding test of absence of spatial interactions at a 1% significance level. We then compute the type I error rate for the null hypothesis as the proportion of spatial patterns for which the spatial localization tests wrongly reject the hypothesis of absence of spatial dependence. In this case, the bias and RMSE are not appropriate measures of estimation uncertainty, since the values of K_d cannot be meaningfully interpreted and can only be used to perform the graphical test. Table 6 summarizes the results for different levels of PMP and shows that the effects of missing data are still important, though less severe, even in the context of a real distribution of firm data that does not follow the CSR hypothesis. In particular, the M function is quite robust to missing data and provides a largely reliable inference, as can be observing the type I error rates that are not very high in this case. If both the locations of interests and the benchmark locations are affected by the same mechanism for generating missing data (a fairly realistic situation), the effects tend to compensate for the data deficiencies and the consequences are limited.

With regard to the case of unintentional locational error, the simulation results suggest the same conclusion drawn above. The effects of uncertain geocoding are more serious than those of missing data. As a consequence, exploiting coarsened areal spatial information and assigning the missing locations to the centroids of their corresponding subareas is not more efficient than simply neglecting the points, because this strategy tends to create spurious spatial clustering.

Finally, to evaluate the consequences of intentional locational errors, we replicate the simulation experiment for the analysis of CSR patterns. Therefore, we again generate 1000 geomasked versions of the observed pattern of single-plant manufacturing firms (where all the 1007 locations are randomly displaced) for each given value of the geomasking maximum displacement distance δ^* .

Table 7 presents the results for the type I error rate of the spatial localization test for the food product sector based on the use of K_d and M according to several geomasking maximum displacement distances (namely, $\delta^* = 1, 5, 10, 15$ and 20 km). Even in this case, the effects of locational

Table 6. Type I error rate for the localization test based on the use of K_d and M under missing clustered data unintentional location errors.

PMP	Missing spatial location		Uncertain geocoding	
	Type I error rate of the test based on ...		Type I error rate of the test based on ...	
	K_d	M	K_d	M
0.1	0.516	0.214	0.601	0.495
0.2	0.550	0.166	0.597	0.554
0.3	0.504	0.177	0.551	0.629
0.4	0.525	0.203	0.578	0.636

Table 7. Type I error rate for the localization test based on the use of K_d and M under intentional location errors (geomasking).

δ^*	Type I error rate of the test based on K_d	Type I error rate of the test based on M	
905	1 km	0.471	0.417
	5 km	0.192	0.335
	10 km	0.173	0.263
	15 km	0.158	0.255
910	20 km	0.137	0.243

errors on the validity of inferential conclusions based on the K_d and M functions are less serious (although still relevant) than those obtained when the localization test is based on the simple Ripley's K -function. This is certainly due to the fact that the former functions constitute relative measures of spatial concentration where the locational errors affecting the spatial distributions of the sector of interest and other economic activities tend to compensate. This compensation effect seems to become stronger when the maximum displacement distance increases.

SUMMARY AND CONCLUSIONS

Spatial microeconomic approaches (inconceivable until only a few decades ago) are now increasingly feasible due to the availability of very large georeferenced databases on individual agents that are no longer limited to the archives, administrative records or panels traditionally employed in official surveys, but also increasingly include alternative data sources, such as satellite and aerial photographs, drones, crowd sourcing, cell phones and many others. The availability of these detailed geographical databases makes it possible now (and increasingly so in future) to model individuals' economic behaviour in space, thus improving our ability to forecast economic trends. The use of micro-data, however, hides pitfalls connected to the presence of non-sampling errors of various natures. This paper aims to shed light on the dangers of drawing inferences based on the spatial concentration features of maps of economic agents using Ripley's K -function and related measures (such as Duranton and Overman's K_d and Marcon and Puech's M functions) when datasets are affected by the presence of missing data and/or intentional and unintentional locational errors. The main results for Ripley's K -function are as follows:

- The presence of randomly missing data introduces spurious clustered patterns at all distances.
- The presence of clustered missing data introduces both spurious inhibitory patterns at low distances and spurious clusters of points at high distances.
- When individuals' locations are geomasked for confidentiality, spurious inhibitory patterns emerge at all distances, as do cancellations of true clustered patterns at small distances.
- If individuals' locations are uncertain and individual points are allocated to the centroids of their districts, we observe two contrasting effects whose weights depend on the dimension of the district and the proportion of points (PMP) affected by uncertainty. In very large districts with large PMPs, the dominating effect is the creation of spurious clusters around the centroids. Conversely, in very small districts with more precise allocations, there is a prevalence of inhibitory patterns created by points' reallocations.

Moreover, on the basis of a simulation study of a real georeferenced dataset of economic activities, it is also possible to conclude that the previous conclusions can be largely extended to the K_d

and M functions. However, since the K_d and M functions are both relative measures of concentration based on comparisons of a specific pattern of interest and a benchmark distribution, the effects of missing data, intentional and unintentional locational errors are slightly mitigated by the compensation effects of the errors affecting the two patterns under consideration.

The results obtained should increase researchers' awareness of the possible dangers of incorrect inferences, thus enhancing their consciousness when selecting the database on which they base their empirical analysis. This study should also make data producers more conscious of the severe consequences of their choices when reallocating uncertain data points or when geomasking them with the intent to preserve confidentiality.

NOTES

¹ For the dbmss package, freely download at <https://CRAN.R-project.org/package=dbmss/>.

² For example, Marcon and Puech (2012) have an interesting critical overview of the main distance-based approaches.

³ With a variance of 0.5 and the exponential covariance function $\rho(d) = \exp(-d/0.25)$.

AQ3 DISCLOSURE STATEMENT

AQ4 No potential conflict of interest was reported by the authors.

REFERENCES

- Arbia, G. (1989). Statistical effects of spatial data transformations: A proposed general framework. In M. F. Goodchild & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 162–169). London: Taylor & Francis.
- Arbia, G., Espa, G., & Giuliani, D. (2015). Measurement errors arising when using distances in microeconomic modelling and the individuals' position is geo-masked for confidentiality. *Econometrics*, 3, 709–718. doi:10.3390/econometrics3040709
- Arbia, G., Espa, G., & Giuliani, D. (2016). Dirty spatial econometrics. *Annals of Regional of Science*, 56, 177–189. doi:10.1007/s00168-015-0726-5
- Arbia, G., Espa, G., & Quah, D. (2008). A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics*, 34, 81–103. doi:10.1007/s00181-007-0154-1
- Arbia, G., & Piras, G. (2009). A new class of spatial concentration measures. *Computational Statistics and Data Analysis*, 53, 4471–4481. doi:10.1016/j.csda.2009.07.003
- Arrow, K. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 29, 155–173. doi:10.2307/2295952
- Besag, J., & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 26, 327–333. doi:10.2307/2346974
- Boots, B. N., & Getis, A. (1988). *Point pattern analysis*. London: Sage Scientific Geography Series Vol. 8
- Cavallo, A., & Rigobon, R. (2016). *The Billion Prices project: Using online data for measurement and research*, NBER Working Paper No. 22111, March 2016.
- Collins, B. (2011). *Boundary respecting point displacement*. Arlington: Python Script, Blue Raster LLC.
- Corrado, L., & Fingleton, B. (2012). Where is the economics in spatial econometrics. *Journal of Regional Science*, 52, 210–239. doi:10.1111/j.1467-9787.2011.00726.x
- Cozzi, M., & Filipponi, D. (2012). *The new geospatial business register of local units: Potentiality and application areas*, 3rd Meeting of the Wiesbaden Group on Business Registers – International Roundtable on Business Survey Frames, Washington D.C., 17–20 September, 2012.
- Diggle, P. J. (2003). *Statistical analysis of spatial point patterns* (2nd ed.). London: Edward Arnold.
- Dubé, J., & Legros, D. (2014). *Spatial econometrics using microdata*. New York: Wiley.

- Duranton, G., & Overman, H. G. (2005). Testing for localisation using micro-geographic data. *Review of Economic Studies*, 72, 1077–1106. doi:10.1111/0034-6527.00362
- Ellison, G., & Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105, 889–927. doi:10.1086/262098
- 1005 Fehmi, J. S., & Bartolome, J. W. (2001). A grid-based method for sampling and analysing spatially ambiguous plants. *Journal of Vegetation Science*, 12, 467–472. doi:10.2307/3236998
- Getis, A., & Boots, B. (1978). *Models of spatial processes*. Cambridge: Cambridge University Press.
- Getis, A., & Franklin, J. (1987). Second-order neighbourhood analysis of mapped point patterns. *Ecology*, 68, 473–477. doi:10.2307/1938452
- 1010 Glaeser, L., & Kerr, W. R. (2009). Local industrial conditions and entrepreneurship: How much of the spatial distribution can we explain? *Journal of Economics and Management Strategy*, 18, 623–663. doi:10.1111/j.1530-9134.2009.00225.x
- Henderson, J. V. (2003). Marshall's scale economies. *Journal of Urban Economics*, 53, 1–28. doi:10.1016/S0094-1190(02)00505-3
- 1015 Hopenhayn, H. A. (1992). Entry, exit and firm dynamics in long run equilibrium. *Econometrica*, 60, 1127–1150. doi:10.2307/2951541
- Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal*, 113, F34–F63. doi:10.1111/1468-0297.00098
- Lang, G., & Marcon, E. (2013). Testing randomness of spatial point patterns with the Ripley statistic. *ESAIM: Probability and Statistics*, 17, 767–788. doi:10.1051/ps/2012027
- 1020 Lazear, E. P. (2005). Entrepreneurship. *Journal of Labor Economics*, 23, 649–680. doi:10.1086/491605
- Mansfield, E. (1995). Academic research underlying industrial innovations: Sources, characteristics, and financing. *Review of Economics and Statistics*, 77(1), 55–65. doi:10.2307/2109992
- Marcon, E., & Puech, F. (2010). Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, 10, 745–762. doi:10.1093/jeg/lbp056
- 1025 Marcon, E., & Puech, F. (2012). A typology of distance-based measures of spatial concentration, technical report
- Marcon, E., Traissac, S., Puech, F., & Lang, G. (2015). Tools to characterize point patterns: dbmss for R. *Journal of Statistical Software*, 67, 1–15.
- Marshall, A. (1890). *Principles of economics*. London: Macmillan.
- 1030 Maurel, F., & Sédillot, B. (1999). A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics*, 29, 575–604. doi:10.1016/S0166-0462(99)00020-4
- Pinkse, J., & Slade, M. E. (2010). The future of spatial econometrics. *Journal of Regional Science*, 50, 103–117. doi:10.1111/j.1467-9787.2009.00645.x
- Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, B*, 39, 172–212.
- 1035 Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, 94, 1002–1037. doi:10.1086/261420
- Rosenthal, S. S., & Strange, W. C. (2003). Geography, industrial organization and agglomeration. *Review of Economics and Statistics*, 85, 377–393. doi:10.1162/003465303765299882
- 1040 USAID. (2013). *Geographical displacement procedure and georeferenced data release policy for the demographic and health surveys*, DHS Spatial Analysis Report, 7 September 2013.
- Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*, 64, 262–270. doi:10.1111/j.1541-0420.2007.00870.x

1045

1050