Mapping distributions in non-homogeneous space with distance-based methods

Éric Marcon, Florence Puech

Version: HAL-04345149v1

Mapping distributions in non-homogeneous space with distance-based methods

Éric Marcon[a,b] and Florence Puech[c,✉]

Abstract

Distance-based methods (DBMs) are frequently used to analyze spatial structures in economics. Results provided by DBMs are particularly effective for the precise detection of spatial concentration, dispersion or absence of significant patterns at any scale. The utility of plotting the results of DBMs in homogeneous space has already been shown. However, no consideration has been given to mapping results in non-homogeneous space. This paper aims to fill this gap. We provide a technique to map local values when using a relative DBM. We illustrate its advantages at first on a theoretical case and then on a real case drawing on contagious disease data on trees in a Parisian park. Data and R code are given for reproducible research. In both cases, we show that local plotting can enable a more accurate spatial characterization of the underlying patterns. To give an example, our empirical results on infested maple trees support evidence of the existence of a contagion disease because they appear to be located in areas where maples are relatively spatially concentrated.

JEL Classification: C10, C60, Q50

Keywords:
Distance-based methods, M-function, Spatial structure, Spatial distribution, Parisian trees, Contagion.

[a] *AgroParisTech, UMR Amap, Univ. Montpellier, Cirad, CNRS, INRAE, IRD, Montpellier, France.*
[b] *UMR EcoFoG, AgroParisTech, Cirad, CNRS, INRAE, Université des Antilles, Université de Guyane, Kourou, France.*
[c] *Université Paris-Saclay, INRAE, AgroParisTech, Paris-Saclay Applied Economics, F-91120, Palaiseau, France.*
✉ *Corresponding author: florence.puech@universite-paris-saclay.fr*
✎ *Authors are in alphabetical order.*

Introduction

Distance-based methods (DBM) are frequently used to analyze spatial structures in economics (Marcon and Puech, 2017). The results provided by these methods are highly effective for the precise detection of phenomena of spatial concentration, dispersion or the absence of significant patterns at any level. Rather than zoning space in separate units (regions etc.), space is considered within that framework as continuous, that is without any zoning. This is possible because, in such cases, estimates are only based on the distances that separate the entities under study (e.g., shops, productive establishments, accidents etc.). This understanding of space opens the door to a very detailed analysis of space without any bias. Today, both geolocalized and satellite data are increasingly available to researchers: as a consequence, in economics, DBM have been significantly developed over the last fifteen years (Arbia et al., 2021).[1] A series of DBM applications can be found in the economics literature to evaluate the location of manufacturing establishments (Sweeney and Feser, 1998), stores (Arbia et al., 2015), emergencies (Bonneu, 2007), patents (Arbia et al., 2008) etc. In a recent article, Piacentino et al. (2021, p.121) wrote that this kind of data has *"the potential to revolutionize spatial economic analysis"*. We strongly support that view. On the one hand, micro-geographic data can describe all spatial phenomena at any scale (even if it is very small). On the other hand, these precise results can henceforth be explained through an econometric analysis. The aim of this article is to contribute to the first step of the analysis; that is the descriptive one.

DBMs detect a departure from the null hypothesis (for example from a completely random location of entities) and, depending on the DBM used, the interpretation of the level of spatial concentration or dispersion may or may not be possible. Our attention in this paper is focused on a continuous mapping of DBM results. Under the assumption of a homogeneous space, that is the hypothesis of a constant density across the entire territory, Getis and Franklin (1987, p.476) proposed a spatial representation of the level of the average local value of DBM results based on maps with contour lines. However, as far as we are aware, under the non-homogeneous space hypothesis, a spatial representation of DBM results has not yet been proposed. This paper fills that gap for a given function, the M-function proposed by Marcon and Puech (2010). Our objective is to plot these observations so as to bring complementary local information to the results provided by DBMs. Firstly and most evidently, this helps to identify places where high or low values of the DBM occur. Secondly, continuous local plotting can help to obtain an improve spatial charac-

[1] Analyzing such kinds of data in continuous space undoubtedly presents a great advantage, because depending on the drivers considered, the relevant neighborhood is not the same. If we consider for example the explanation of industrial agglomeration, many factors should be taken into account and their intensity varies according to the distance. For instance, studies show that the benefits of externalities decrease rapidly with distance (very close neighborhoods in that case) whereas larger distances are considered for the explanation of input/output linkages (Rosenthal and Strange, 2020).

terization of the underlying patterns. Finally, this mapping can give support to an intuition that later analysis may confirm. In section 1, we explain the importance of exploring non-homogeneous space with DBMs. In section 2, we present how mapping results in non-homogeneous space. In section 3, we propose a concrete application on a dataset providing an inventory of contagious disease affecting trees in a Parisian park. Thanks to this mapping, we try to establish whether concentration favors contagion. Along with our examples (whether theoretical or empirical), a complete R code is given in the appendix to enable the reproducibility of the research.

1. Exploring non-homogeneous space with distance-based methods

1.1 The necessity to go beyond a spatial distribution of entities or density-maps

The first possibility that we have in mind for mapping micro-geographic data is to represent the distribution of the exact geographic positions of the entities under study. This is equivalent to a representation of mapped points in space. Many examples can be found in the economic literature: public schools and alcohol retailers in Picone et al. (2009), manufacturing plants in Aleksandrova et al. (2020) etc. This technique can detect certain location trends (such as clusters), but it is not informative if datasets contain an overly large number of observations. To solve this problem, one possibility is to map the density of the observations under study in two dimensions (Arbia et al., 2012, 2014; Coll-Martìnez et al., 2019; Moreno-Monroy and Cruz, 2016) or in three dimensions (Lang et al., 2020).

The limit is that mapped density can not provide any information as to the relations between entities (e.g., spatial attraction, repulsion or neither attraction nor repulsion). Distance-based methods solve this problem: their mapped results reveal location patterns and detect the intensity of relationships, at any scale.

1.2 Distance-based methods: basic concepts

Distance-based methods (DBMs) preserve the geolocalized information. This is possible because they rely on the exact position of entities (geographic coordinates) and on the individual characteristics under scrutiny (for example the gender of humans, sector of shops, circumference of trees etc). DBMs are particularly effective for testing whether there is any attraction or repulsion between entities under study (Floch et al., 2018; Sweeney and Arabadjis, 2022). Within that framework, space is considered as continuous and one's attention is only focused on the distance that separates pairs of entities. We test for whether there is any attraction or any repulsion between entities belonging to one group (what we refer to as localization) or two different groups (e.g., the phenomenon of co-localization, as per the vocabulary employed in the literature by Duranton and Overman, 2005). Technically, conclusions drawn from the results of distance-based methods rely on point process theory (Møller and Waagepetersen, 2004). Today, more than ten distance-based methods are used in the literature (Marcon and Puech, 2017). They are notably based on a set of hypotheses to correctly define the neighborhood of the points and the benchmark against which the observed distribution is compared. Thus, the question under study must always guide research in the direction of the appropriate function to use (Bickenbach and Bode, 2008). DBM results are represented as a plot of a function of distance and always compared to a simulated envelope representing the confidence interval of a null hypothesis to be tested. Depending on the DBM used, empirical values may or may not be interpreted; however, all DBMs can detect the intensity level of the spatial concentration or dispersion. The advantage of all DBMs is that they can detect a departure from the null hypothesis (e.g., a random distribution) at any scale and without statistical bias. This point is crucial because to grasp the interactions between entities, aggregating data up to a given level of space is not optimal. Many examples point out the loss of information by using areal data and illustrate the well-known Modifiable Areal Unit Problem (Openshaw and Taylor, 1979; Openshaw, 1984; Arbia, 1989). Finally, from a practical point of view, we can note that the evaluation of underlying structures of spatialized data (humans, shops, accidents etc.) can be carried out with R packages, for example with the **dbmss** package (Marcon et al., 2015).

1.3 Distance-based methods in homogeneous *vs.* non-homogeneous space

One of the main differences between DBMs is the definition of space they rely on (Marcon and Puech, 2010, 2017). Consider that the density of points is a function of space. If the hypothesis of a constant density everywhere across the area under study can be supposed, space is considered as homogeneous. In technical terms, we say that the point process is stationary. If not, space will be considered as non-homogeneous (or inhomogeneous). This consideration should be studied with care because it is one of the key indicators for choosing the most appropriate DBM to use. To give an example, the hypothesis of homogeneous space is limited for studying the location of firms (as noted by Duranton and Overman, 2005, footnote 24). Ripley's K function (Ripley, 1976, 1977) or its variations (such as the L function of Besag, 1977) is certainly the best known DBM. It is used under the hypothesis of a homogeneous space. Many applications can be found in various fields of research. This certainly explains why a technique for mapping the results of the K-function has already been proposed. Getis and Franklin (1987, p.476) proposed a spatial representation of the level of the average local value of DBM results based on maps with contour lines. More precisely, they analyze the spatial distribution of *ponderosa* pine trees in a square area (120m x 120m) of the Klamath National Forest in North Carolina in the United States. They plot the results of Besag's L function and map their results by using isolines. Due to the properties of Besag's L function, results can be compared across distances, making it particularly convenient for analyzing the spatial structure of the distribution of pines across the entire territory and identifying the spatial positions of clusters of pines. Hereinafter, we propose a technique to map points in that vein by using a DBM in non-homogeneous space for which the results may be also compared whatever the distance. A good candidate seems to be the *M* function: we explain why in the following section.

1.4 Presentation of the *M* function

1.4.1 Intuitive idea

The *M* function is a relative distance-based function proposed by Marcon and Puech (2010). The main idea of this function is to compare the local proportion of points of interest to the one observed over the entire territory. More precisely, if there are relatively more entities observed in the neighborhood of entities of interest than over the entire territory, the *M* function will detect a relative concentration of entities. On the contrary, if there are relatively less entities observed in the neighborhood of points of interest, a relative dispersion of entities will be detected by the *M* function. The null hypothesis is defined as the same distribution of points of interest as that of all points all over the territory. The simulation of a confidence envelope of the null hypothesis gives the significance of the results. Depending on the question under study, the researcher may define a distance for the relevant surroundings of an entity, for instance a neighborhood of 50 meters. If there is no privileged distance, the *M* function can be calculated for all distances (e.g., 5 meters, 10 meters, 15 meters etc.). The maximum distance is where we suspect no more potential interactions between entities will occur because they are too far from each other to interact. In general, Euclidean distance is the distance used in the analysis.

To give an example, suppose that we are interested in characterizing the location of one given species of trees, let us say maples, in a forest.

Step 1: For a series of surroundings defined as circular disks around each maple for example, measuring 5 meters, 10 meters, 15 meters etc., the relative proportion of maples compared to the other species of trees is counted.

Step 2: If the average local proportion of maples is greater than the one observed in the forest, we say that there is a relative spatial concentration of maples around maples (spatial attraction of maples). In contrast, if the local proportion of maples is lower than the one observed on the forest, we say that there is a relative dispersion of maples around maples (repulsion between maples).

Step 3: To test the significance of the results, at first we define each tree as a bundle of its characteristics (e.g., its species, its height, the circumference of its trunk etc.). Then, we simulate distributions by maintaining the exact positions of all trees but drawing the bundle of characteristics of trees randomly. For each iteration, the *M* function is calculated, the level of confidence defines the lower and upper bands of the confidence interval of the null hypothesis. A sufficient number of simulations is recommended to obtain the confidence interval. Finally, note that depending on the question, we can define a weight for each point of the distribution. In our example, it could be the circumference of the trunk or the basal area of the tree, i.e., the area of its trunk cut 1.3 meter above ground.

1.4.2 Definition of the *M* function and the local *M* function

The *M* function (Marcon and Puech, 2010) belongs to cumulative distance-based methods, that is the neighborhood of entities are analyzed up to a given distance rather than at a given distance. The *M* function (as is the case with all distance-based methods) is based on a strong mathematical foundation, that of point process theory (Møller and Waagepetersen, 2004; Baddeley et al., 2015). In less technical terms, let's say that the *M* function helps to characterize the relative spatial structures of entities under study by measuring the relative frequency of entities of interest up to each distance (denoted r), compared to the same ratio but defined in the whole area under study. In what follows, we will consider the data according to a given characteristic (species, sector etc.). We focus at first on the definition of the intra-type function: this means that the type of neighbors of interest of one given point is its own type only, denoted s. Let us denote:

- x_i^s, the position of point i of the reference type s, at the center of the disk i.e., the point at which the neighborhood is to be analyzed),
- x_j^s, the position of a neighbor of interest j of the same type as point i,
- x_j, the position of a neighbor j of i, whatever its type,
- $w(.)$, the weight of a given neighbor. For example, $w(x_j)$ defines the weight of a neighbor j of i.
- W_s, the total weight of the points x_j^s,
- W, the total weight of all points of the dataset, whatever their type,
- $\mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right)$, the indicator function equal to 1 if x_j is in the neighborhood of x_i^s, e.g., the distance between x_i^s and x_j is at most equal to r, 0 otherwise.
- $\mathbf{1}\left(\left\|x_i^s - x_j^s\right\| \leq r\right)$, the indicator function equal to 1 if the distance between x_i^s and x_j^s is at most equal to r, 0 otherwise.

The intra-type *M* function is estimated by the following equation:

$$\hat{M}(r) = \sum_i \frac{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i^s - x_j^s\right\| \leq r\right) w\left(x_j^s\right)}{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right) w\left(x_j\right)} \Big/ \sum_i \frac{W_s - w\left(x_i^s\right)}{W - w\left(x_i^s\right)} \tag{1}$$

As the *M* function compares two ratios (a local one to a global one), its reference value is 1 for any distance r considered. The significance of the results is obtained by Monte Carlo simulations after choosing the risk level (generally 1%, 5% or 10%) and the number of simulations. The **dbmss** package (Marcon et al., 2015) for the R software (R Development Core Team, 2022) is useful to compute the M function easily.

The intertype version of the function is immediately derived from the previous definitions, by replacing neighbors of interest of the same type s as the reference points by neighbors of another type t to obtain:

$$\hat{M}_{s,t}(r) = \sum_i \frac{\sum_{j\neq i} \mathbf{1}\left(\left\|x_i^s - x_j^t\right\| \leq r\right) w\left(x_j^t\right)}{\sum_{j\neq i} \mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right) w\left(x_j\right)} / \sum_i \frac{W_s}{W - w\left(x_i^s\right)} \tag{2}$$

Finally, note that the M function gives, at each distance r, the *average* value of the individual values calculated around every point i. As a consequence, the M function only returns, for all distances, the average spatial structure of points observed around all of the points of interest. Maps will preserve local values of the M function around every point i. We call them "individual values of M" hereinafter.

2. Mapping spatial structures in non-homogeneous space

2.1 Additional development needed in R software

The increasing availability of geo-referenced data opens the way for a more effective characterization of the spatial structure of entities: human beings, firms, plants, objects etc. (Baddeley et al., 2015). Consequently, new developments have been proposed to test new hypotheses of research. Moreover, a great number of developments were recently made to improve statistical methods and software applications for spatial analysis. The R software (R Development Core Team, 2022) now constitutes an essential tool in spatial analysis (Bivand et al., 2013). We are in line with this contemporary approach and propose two main developments. Firstly, we develop an extension of an existing R package: **dbmss** (Marcon et al., 2015; R Development Core Team, 2022) for mapping relative concentration. Secondly, our motivation is to both develop the methodology and facilitate the use of R code by researchers to reproduce this methodology if they are interested in its applications. The open data we used in the following empirical example helps the reader to reproduce the example, step by step, using its R code.

2.2 Mapping *M*'s individual values

The way to plot the M results is the following.

Step 1: Definition of the pertinent distance for the analysis. Firstly, the most relevant distance r should be chosen. It may be driven by the knowledge of the process under analysis (a meaningful distance of interaction between points) or a preliminary estimation of the function at all distances.

Step 2: Calculation of the individual M values. The M function must be computed at the chosen distance and all individual values maintained.

Step 3: Application of spatial smoothing. M individual results are geolocalized data: they are defined at the location of the points of interest only. To plot them across the whole area under study, we need to apply spatial smoothing.[2] The *Smooth.ppp()* function is available in the R package **spatstat** (Baddeley and Turner, 2005). It is based on kernel smoothing and requires the choice of an arbitrary bandwidth that is still as yet undefined (Lang et al., 2020). The combination of Gaussian kernel smoothing with the bandwidth proposed by Scott (1992) seems to be a good compromise for the purpose of our study.

Step 4: Mapping the results A development in the **dbmss** (version 2.8) R package is proposed in order to proceed with this final step. It is thus possible to visualize the smoothing results of the local M function plotted on a map. Contour lines delimit areas with the same level of spatial concentration. The number of rows and columns of the grid defines the resolution of the map and should be chosen with care.

2.3 Theoretical example

We consider a study area defined as a 1-by-1 window where two types of points are located. The first type of points defines the *controls*. Their generation respects the hypothesis of a non-homogeneous distribution of points. The second type of points defines the *cases*, clustered in space. We focus our interest on the cases: the controls added to the cases constitute the benchmark distribution.

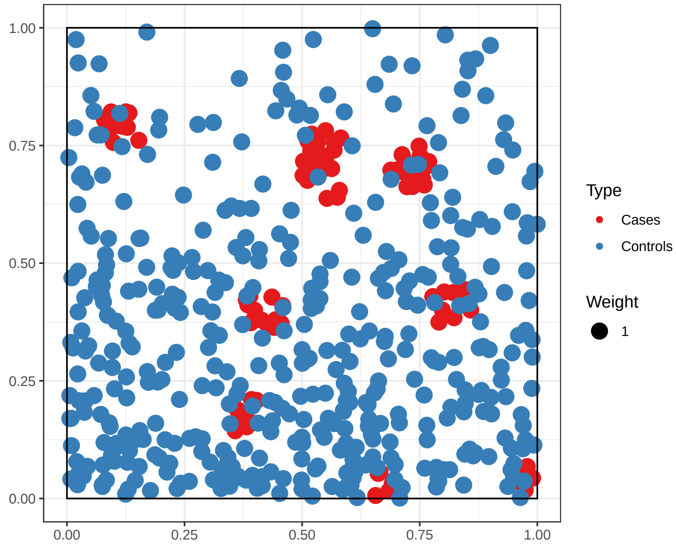
A step-by-step presentation of the R code in this example is available in the appendix.

In figure 1a, red points identify the cases while the blue ones represent the controls. All point weights are fixed to 1, whatever their type. The distribution of controls is simulated by a non-homogeneous Poisson process defined by an arbitrary density function (see appendix for details). 435 controls are indicated in figure 1a. The distribution of cases is simulated from a Matérn (1960) process that produces, on average, 10 randomly located clusters of radius 0.05 and containing 10 cases. In figure 1a, 97 cases are plotted: aggregates of cases can be easily detected.

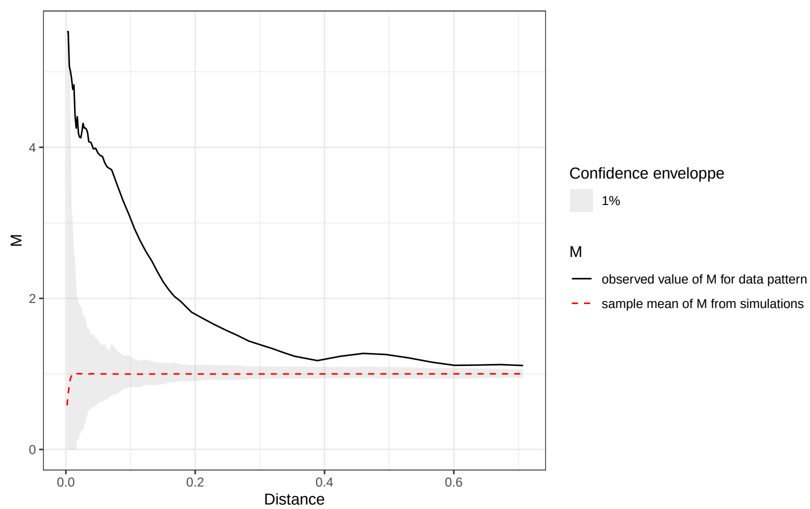
In figure 1b, the M function for the case points is plotted by using the **dbmss** R package (Marcon et al., 2015; R Development Core Team, 2022). The global confidence envelope is obtained by 1,000 simulations at 1% risk level. For all distances, the M cases plot is above the confidence interval: M detects a relative spatial concentration of cases for all radii considered. The maximum value of M is obtained for a distance around 0.05, which corresponds to the simulated size of the clusters. One can see that the M plot decreases as the distance r increases: the lack of cases outside the clusters and the presence of controls leads to a decrease of relative spatial concentration of cases. Irregularities in the decreasing of the

[2] We thank the anonymous referee that suggests the spatial smoothing technique as opposed to kriging, used in a previous version.

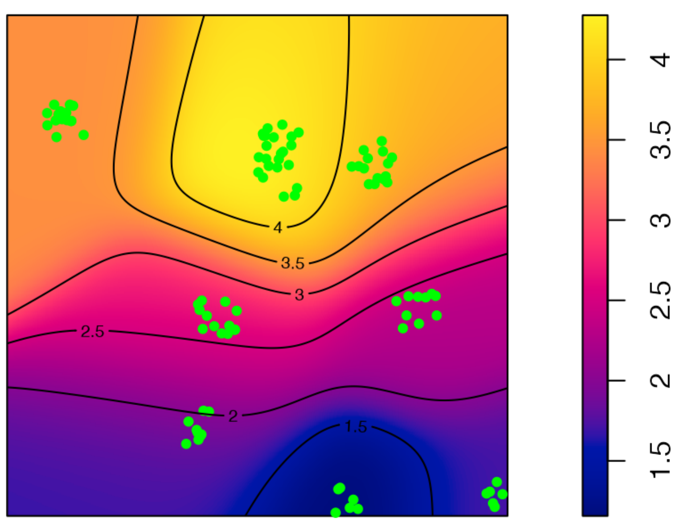
Figure 1. Theoretical example



(a) Theoretical distribution of points with multiple patterns.



(b) *M*-cases plot.



(c) Map of *M*-case local values within a distance of 0.1.
Green dots indicate cases. Cold to hot colors represent the increasing level of local concentration.

M plot signal the presence of (aggregated) cases around the case clusters.

The *M* function gives an average estimation of the relative spatial concentration or relative spatial dispersion of cases over the whole territory. As a consequence, if we only have figure 1b to refer to, and no other information is available, no observations can be made as to: (*i*) the location of clusters on the area and (*ii*) the local variations of the *M* values. However, for example, if two clusters of cases are closely located, locally there is a greater level of spatial concentration that may prove useful to plot for the purpose of our study. Mapping individual *M* values provides the answer.

In figure 1c, local *M* values are plotted with contour lines, all cases are represented by red points. A distance (r) of 0.1 is chosen for mapping. We can now easily detect different levels of spatial concentration or dispersion all over the area. *M* values are readable on the contour lines. Cold colors represent low levels of *M* local values while warm colors indicate high levels. The relative spatial concentration of cases is the greatest in areas in the north of the map (assumed to be at the top of the figure) where the controls are less present. Cases located in the south of the map are relatively less concentrated. Note that plotting the density of cases in that example would lead to a very different conclusion, as the controls are not homogeneously distributed. The density map is given in section 1.4 of the appendix.

3. Empirical application on Parisian trees

3.1 Motivation

To illustrate the potential of our method, we test an ecological hypothesis on a Parisian park, *Parc Omnisport Suzanne Lenglen* (POSL) located in the 15th arrondissement. Trees are subject to contagious fungal diseases. In ecology, an abundant body of literature addresses the effect of the level of biodiversity on the severity of such diseases. For example, an extensive study by Nguyen et al. (2016) conducted on 16 tree species in European forests confirms the hypothesis that biodiversity (measured as the number of tree species) decreased the incidence of disease in conifers. Rutten et al. (2021) showed that increased local biodiversity decreased the foliar fungal pathogen infestation rate in subtropical forests, while Saadani et al. (2021) confirmed that severity decreased. The main mechanism was the dilution of host species in the local neighborhood, i.e., a smaller relative concentration of potential host trees at short distance.

Recently, some maple trees in the POSL were contaminated with sooty bark disease caused by the fungus *Cryptostroma corticale*[3]. Only maples were infested. 23 maples among the 1,472 trees (including 529 maples) in the park were logged to eliminate infested trees and limit the contagion. By opportunity, 25 decaying trees (including 3 non-infested maples) were logged at the same time. In what follows, we investigate whether the local spatial concentration of maples spreads the contagion. A map of the relative concentration of maple trees and the location of logged trees is informative on two points. We can test for (*i*) whether infested trees were located in areas with high concentration of maple trees and (*ii*) whether or not decaying trees followed the same pattern.

[3] Personal communication from the crew of the park.

3.2 Data and study area

Our data is extracted from "Paris open data", available at: https://opendata.paris.fr.[4]

The positions and the characteristics of trees located in the city of Paris are given. The online database is very large: more than 200,000 trees are inventoried. Our empirical work is conducted on a park located in the south of Paris. The "Parc Omnisport Suzanne Lenglen" (POSL) in the 15th arrondissement is particularly suitable for our analysis. Firstly, a pathogen agent affects only one species (maple trees) among numerous located in the park. Secondly, the presence of sports facilities (rugby, football, basketball, tennis etc.) and paths gives little support to the hypothesis of a homogeneous space for the distribution of trees in this park. In our empirical study, we use the geolocation of trees i.e., the exact geographical position of trees), their genus, species and circumference. The latter characteristic enables the calculation of the basal area of trees (taken as their weight). We complete this database with another source, also available via Paris Open Data, providing data relating to trees felled and the reasons they were cut down.

The distribution of the trees in the POSL in February 2021 is given in figure 2. Coordinates on both axis are in meters (Lambert 93 projection).[5] The non-homogeneous distribution of trees is evident. Different species are present but the most frequent one is maples (*Acer spp.*). Among the 1,472 trees in the POSL, 48 were logged. Only two reasons are given for cutting down POSL trees. The first is an irreversible decline. The second is the presence of a pathogen agent: infested maples must be logged to avoid contagion by the *Cryptostroma corticale* fungus (Koukol et al., 2015).

3.3 Are infested maples located in areas where the relative spatial concentration of maples is high?

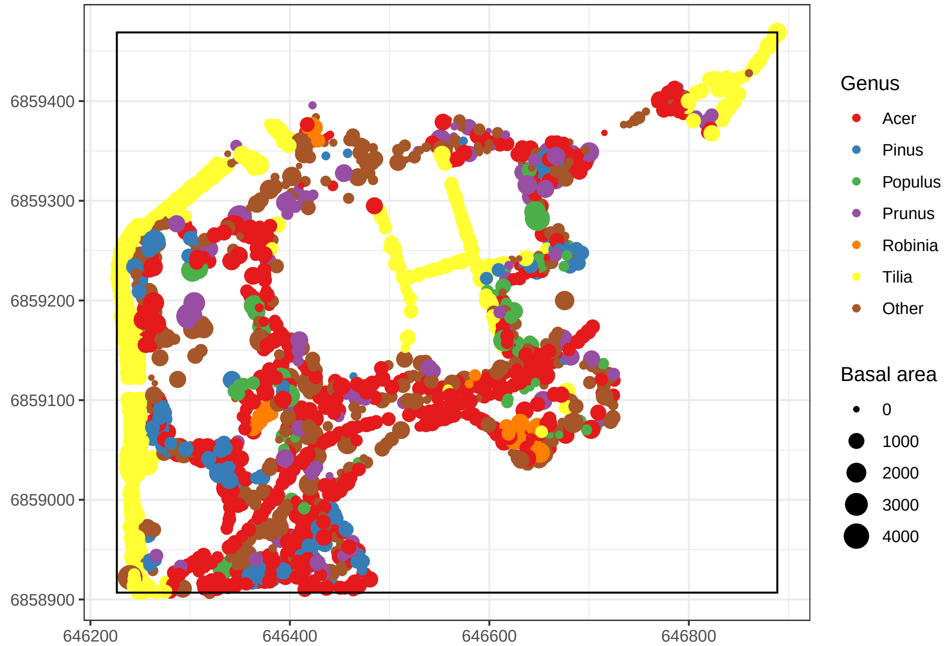
Our empirical work aims to answer the following question: are infested maples located more frequently around maples?

To answer this question, firstly we need to evaluate the relative spatial concentration of maples in the park: empirical results will confirm whether or not there is a spatial concentration of maples and at what scale of observation it occurs. This helps us to choose the pertinent distance for mapping the local spatial concentration estimates a second time. All details are given in the appendix as well as the R code.

[4] The two databases are: Arbres, Direction des Espaces Verts et de l'Environnement - Ville de Paris, 7 February 2022, under license ODbL and Arbres à abattre pour raison sanitaires et essence de remplacement, Direction des Espaces Verts et de l'Environnement - Ville de Paris, 3 February 2022, under license ODbL.

[5] For improved readability in figure 2, the visualization of trees is magnified. The area of the points is proportional to the the basal area of the trees.

Figure 2. Spatial distribution of trees in the POSL in 2021



Note: Distribution based on Paris Open Data. Lambert 93 projection is used. The basal area gives the weight of tree.

The M plot of the relative spatial concentration of maples around maples is shown in figure 3 (prior to any felling carried out). The maximum distance considered is 30 meters, 1,000 simulations were generated for the confidence interval and the associated risk level is 5%. M detects a relative spatial concentration of maples around maples for all distances up to 30 meters. We then follow the mapping technique developed in section 2.2. We chose a distance of 15 meters to map the M results as per the literature on contagion among trees that focuses on immediate neighbors (Hantsch et al., 2014), with this distance appearing to be a good candidate according to the M local estimates[6]. The answer of our first research question seems to be that almost all infested maples are located in the areas where maples are relatively the most spatially concentrated. This result is in line with a contagion disease.

In section 2.3 of the appendix, we provide a complementary analysis to test more directly the dilution effect. Our approach is to focus on the relative proportion of maples in the neighborhood of infested trees (Hantsch et al., 2014). The intertype M function is used to map it. The conclusion is the same as that of our first question above.

This example underlines the usefulness of distance-based methods within that framework. Firstly, because the contagion may appear at very small scales: this is confirmed by the M plot in figure 3. Secondly, and without doubt, aggregate data mask interactions. The shapes of the contour lines of spatial concentration call for a precise analysis of the area and the use of appropriate statistical methods that may correctly reveal the underlying spatial structures. Finally, one can say that a limit of our analysis is that we focus only on trees located in the park and we do not take into account trees positioned for instance at the border of the park (e.g., in the surrounding streets etc.). This is with no doubt an edge-effect, but we consider it of secondary importance in our case, namely because the distance of interest ($r = 15m$) is small.

[6] For sake of transparency, in the appendix we also provide the M local results for different distances. Results systematically corroborate our conclusion.

Conclusion and ways of research

This paper provides evidence on the relevance of plotting spatial concentration results in continuous space. Its attention is focused on relative spatial concentration and on a new methodological development with available data and R code for reproducible research.

Future research could be developed in both directions. Firstly, the empirical analysis proposed in this article comes from the field of ecology. In economics, estimates resulting from local mapping would be of great interest: as we outlined, the M function has been deployed in studies into the location of activities. Local mapping in continuous space will undoubtedly be useful for the precise identification of industrial clustering (e.g., shapes of agglomerations and intensity of interactions). This could encourage deeper analysis in these particular areas in the spirit of the article by Kerr and Kominers (2015). The second possible research avenue is to compare the effects of local spatial concentration to that of local spatial diversity. Whatever the field of research, an interesting development could be achieved by theoretically linking

Figure 3. *M* function plot for POSL's maples

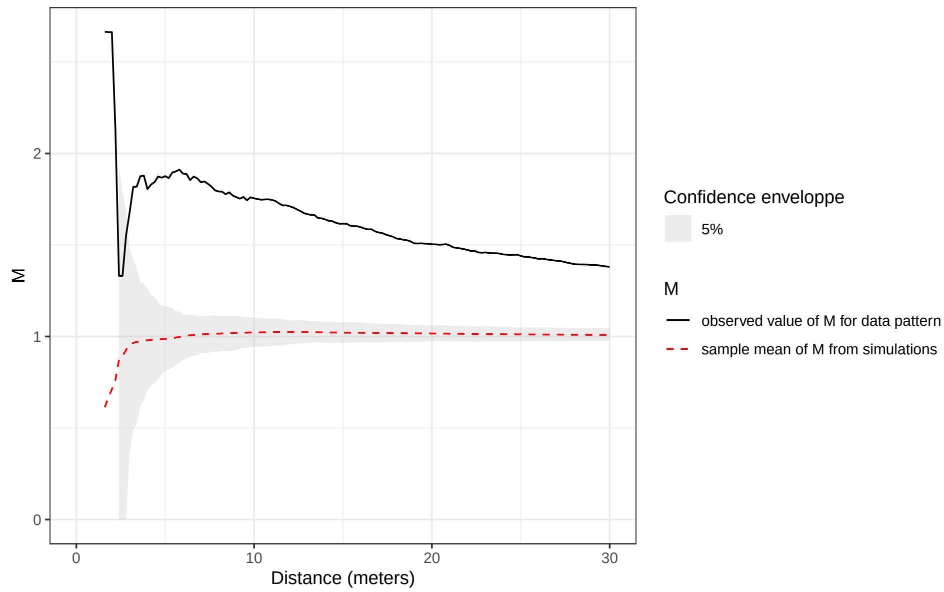
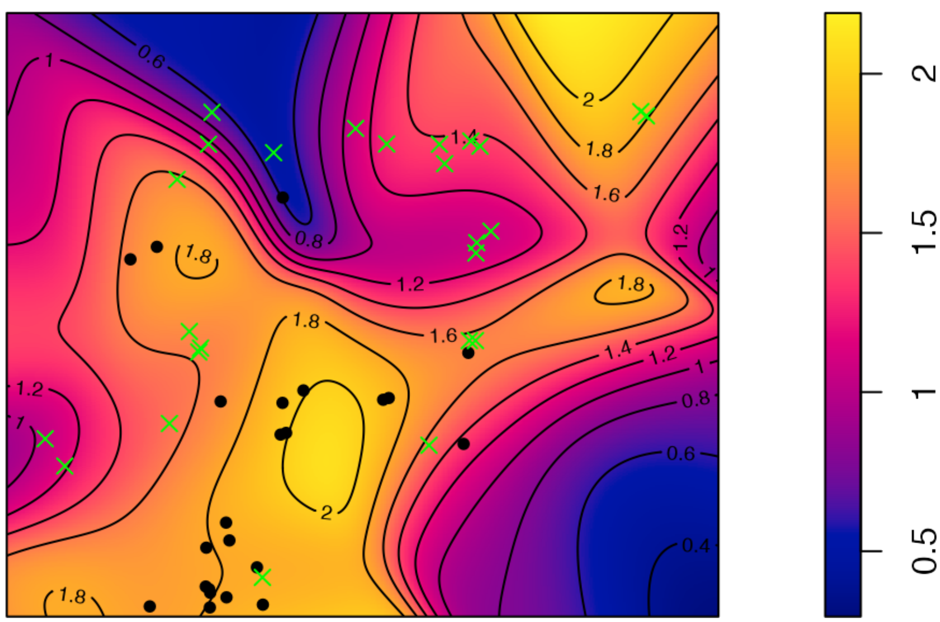


Figure 4. Map of the *M* local values of the POSL's maple trees with the location of logged trees



Lecture: Black points indicate the position of infested maples while the irreversibly decaying trees are represented with green crosses. Cold to hot colors represent the increasing level of the relative local concentration within a distance of 15 meters.

two concepts: the *propagation effect* relative to the level of spatial concentration and the *dilution effect* relative to local diversity. Mechanisms at work should then be disentangled or perhaps unified as per the work carried out by Marcon (2019).

Appendix

R code is available at the following address:
https://ericmarcon.github.io/JSPE-D-23-00002

Acknowledgments

We are sincerely grateful to two anonymous referees for their helpful comments. This research was carried out when Florence Puech was visiting Paris-Saclay Applied Economics - PSAE (INRAE-AgroParisTech). We would also like to thank seminar participants at the 20th International Workshop in Spatial Econometrics and Statistics held in Lille (France) and the seminar at the Department of Geography of the University of California in Santa Barbara (UCSB). Florence Puech gratefully acknowledges financial support from INRAE and Université Paris-Saclay (MERR program). Eric Marcon is supported by labex CEBA, ANR-10-LABX-0025.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aleksandrova, E., Behrens, K., and Kuznetsova, M. (2020). Manufacturing (co)agglomeration in a transition country: Evidence from Russia. *Journal of Regional Science*, 60(1):88–128.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Kluwer, Dordrecht.

Arbia, G., Cella, P., Espa, G., and Giuliani, D. (2015). A micro spatial analysis of firm demography: the case of food stores in the area of Trento (Italy). *Empirical Economics*, 48(3):923–937.

Arbia, G., Dickson, M. M., Gabriele, R., Giuliani, D., and Santi, F. (2021). On the spatial determinants of firm growth: A microlevel analysis of the Italian SMEs. In Colombo, S., editor, *Spatial Economics Volume II: Applications*, pages 89–120. Springer International Publishing, Cham.

Arbia, G., Espa, G., Giuliani, D., and Dickson, M. M. (2014). Spatio-temporal clustering in the pharmaceutical and medical device manufacturing industry: A geographical micro-level analysis. *Regional Science and Urban Economics*, 49:298–304.

Arbia, G., Espa, G., Giuliani, D., and Mazzitelli, A. (2012). Clusters of firms in an inhomogeneous space: The high-tech industries in Milan. *Economic Modelling*, 29(1):3–11.

Arbia, G., Espa, G., and Quah, D. (2008). A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics*, 34(1):81–103.

Baddeley, A. J., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. Chapman and Hall/CRC. 810 pages.

Baddeley, A. J. and Turner, R. (2005). Spatstat: an r package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42.

Besag, J. E. (1977). Comments on Ripley's paper. *Journal of the Royal Statistical Society*, B 39(2):193–195.

Bickenbach, F. and Bode, E. (2008). Disproportionality measures of concentration, specialization, and localization. *International Regional Science Review*, 31(4):359–388.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. UseR! Series, Springer, 2nd ed., NY.

Bonneu, F. (2007). Exploring and modeling fire department emergencies with a spatio-temporal marked point process. *Case Studies in Business, Industry and Government Statistics*, 1(2):139–152.

Coll-Martìnez, E., Moreno-Monroy, A.-I., and Arauzo-Carod, J.-M. (2019). Agglomeration of creative industries: An intra-metropolitan analysis for Barcelona. *Papers in Regional Science*, 98(1):409–431.

Duranton, G. and Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4):1077–1106.

Floch, J.-M., Marcon, E., and Puech, F. (2018). Spatial distribution of points. In Loonis, V. and de Bellefon, M.-P., editors, *Handbook of Spatial Analysis, Theory and Application with R*, number 131 in INSEE Méthodes, pages 71–111. Insee-Eurostat.

Getis, A. and Franklin, J. (1987). Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68(3):473–477.

Hantsch, L., Bien, S., Radatz, S., Braun, U., Auge, H., and Bruelheide, H. (2014). Tree diversity and the role of non-host neighbour tree species in reducing fungal pathogen infestation. *Journal of Ecology*, 102(6):1673–1687.

Kerr, W. R. and Kominers, S. D. (2015). Agglomerative forces and cluster shapes. *The Review of Economics and Statistics*, 97(4):877–899.

Koukol, O., Kelnarová, I., and Černý, K. (2015). Recent observations of sooty bark disease of sycamore maple in Prague (Czech Republic) and the phylogenetic placement of *Cryptostroma corticale*. *Forest Pathology*, 45(1):21–27.

Lang, G., Marcon, E., and Puech, F. (2020). Distance-based measures of spatial concentration: Introducing a relative density function. *The Annals of Regional Science*, 64:243–265.

Marcon, E. (2019). Mesure de la biodiversité et de la structuration spatiale de l'activité économique par l'entropie. *Revue économique*, 70(3):305–326.

Marcon, E. and Puech, F. (2010). Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, 10(5):745–762.

Marcon, E. and Puech, F. (2017). A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, 62:56–67.

Marcon, E., Traissac, S., Puech, F., and Lang, G. (2015). Tools to characterize point patterns: dbmss for R. *Journal of Statistical Software*, 67(3):1–15.

Matérn, B. (1960). Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut*, 49(5):1–144.

Moreno-Monroy, A. I. and Cruz, G. A. G. (2016). Intra-metropolitan agglomeration of formal and informal manufacturing activity: Evidence from Cali, Colombia. *Tijdschrift voor economische en sociale geografie*, 107(4):389–406.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*, volume 100 of *Monographs on Statistics and Applies Probabilities*. Chapman and Hall.

Nguyen, D., Castagneyrol, B., Bruelheide, H., Bussotti, F., Guyot, V., Jactel, H., Jaroszewicz, B., Valladares, F., Stenlid, J., and Boberg, J. (2016). Fungal disease incidence along tree diversity gradients depends on latitude in European forests. *Ecology and Evolution*, 6(8):2426–2438.

Openshaw, S. (1984). The modifiable areal unit problem. CATMOG - Concepts And Techniques in Modern Geography 38, Geo Abstracts University of East Anglia.

Openshaw, S. and Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In Wrigley, N., editor, *Statistical Applications in the Spatial Sciences*, pages 127–144. Pion, London.

Piacentino, D., Arbia, G., and Espa, G. (2021). Advances in spatial economic data analysis: methods and applications. *Spatial Economic Analysis*, 16(2):121–125.

Picone, G., Ridley, D., and Zandbergen, P. (2009). Distance decreases with differentiation: Strategic agglomeration by retailers. *International Journal of Industrial Organization*, 27(3):463–473.

R Development Core Team (2022). R: A Language and Environment for Statistical Computing.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society*, B 39(2):172–212.

Rosenthal, S. S. and Strange, W. C. (2020). How close is close? The spatial reach of agglomeration economies. *Journal of Economic Perspectives*, 34(3):27–49.

Rutten, G., Hönig, L., Schwaß, R., Braun, U., Saadani, M., Schuldt, A., Michalski, S. G., and Bruelheide, H. (2021). More diverse tree communities promote foliar fungal pathogen diversity, but decrease infestation rates per tree species, in a subtropical biodiversity experiment. *Journal of Ecology*, 109(5):2068–2080.

Saadani, M., Hönig, L., Bien, S., Koehler, M., Rutten, G., Wubet, T., Braun, U., and Bruelheide, H. (2021). Local Tree Diversity Suppresses Foliar Fungal Infestation and Decreases Morphological but Not Molecular Richness in a Young Subtropical Forest. *Journal of Fungi*, 7(3):173.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Chapman & Hall/CRC Interdisciplinary Statistics. John Wiley & Sons, Inc.

Sweeney, S. and Arabadjis, S. (2022). Spatial point patterns. In Rey, S. J. and Franklin, R. S., editors, *Handbook of Spatial Analysis in the Social Sciences*, chapter 15, pages 262–276. Edward Elgar Publishing.

Sweeney, S. H. and Feser, E. J. (1998). Plant size and clustering of manufacturing activity. *Geographical Analysis*, 30(1):45–64.