Hal-05512154v2-Revised
August 29, 2026

Computation of Large Spatial Datasets with the M function

Eric Marcon[1] Florence Puech[2]

Abstract
Since agglomeration is a core question in regional science, spatial concentration measures are widely employed in that field to evaluate the spatial distribution of activities. Increasing access to large geo-referenced datasets, coupled with the development of computing power, has encouraged the search for suitable spatial statistical tools. Distance-based methods have been extensively developed to detect spatial concentration, dispersion or independence of entities at any distance and without any bias. Today, distance-based methods face a new challenge: they must be able to address very large micro-geographic datasets. Recently, Tidu et al. (2024) highlighted the qualities of Marcon and Puech's *M* function, a relative distance-based measure, and also expressed reservations about the computation time required. Herein, we explore two possible ways to reduce the computation burden of large geo-located datasets: approximating the position of points and thinning the point pattern. In both cases, the deterioration extent of the *M* results is estimated and discussed as the gains it provides in computation time, using the R software. We discuss implications of these findings in the field of regional science. We notably provide evidence that the individual location approximation generates information loss at substantially small distances, implying a trade-off between the smallest distance at which spatial interactions could be detected and computing performance. We also give support that random thinning is an efficient method to analyze large datasets with very good accuracy. The R code used in the article is given for the reproducibility of our results.

Keywords
Distance-based method, M-function, Performance test, R Package dbmss

[1]AgroParisTech, UMR AMAP, CIRAD, CNRS, INRAE, IRD, Univ Montpellier, Montpellier, France.

[2]Université Paris-Saclay, INRAE, AgroParisTech, Paris-Saclay Applied Economics, F-91120 Palaiseau, France.

***Corresponding author**: eric.marcon@agroparistech.fr,

Contents

Introduction 1

1 Spatial concentration measures and regional science 3

1.1 A close connection . . . 3

1.2 Recent developments . . . 3

2 The *M* function, for any geo-located dataset? 4

2.1 Main idea . . . 4

2.2 Definition . . . 4

2.3 Simulated data . . . 5
Drawing the points ▪ Gridding the space

2.4 Computing *M* using the *dbmss* package . . . 6
Point pattern ▪ Distance matrix

2.5 Computational performance . . . 6
Computing time ▪ Memory

2.6 Discussion: technical limit of computing *M* . . . 7

3 A guide to dealing with large datasets 7

3.1 First strategy: Approximating the position of points 7
Two main effects ▪ Test setting and results ▪ Discussion

3.2 Second strategy: Random thinning . . . 9
Test setting and results ▪ Discussion

4 Conclusion 10

Appendix 11

Acknowledgments 11

Introduction

Agglomeration of people and economic activities is a key topic in economics (Fujita et al., 1999; Fujita and Thisse, 2013; Glaeser, 2010). Because spatial concentration offers productivity gains due to cost reductions, economists and geographers in regional science have extensively studied clusters of economic activity, their consequences on territories and the importance of the determinants that explain them (Marshall, 1890; Duranton and Puga, 2004). Policymakers are also particularly interested in the agglomeration of activities, since a better understanding of the locations and extent of the agglomeration of activities allows them to develop more effective economic policies addressing regional growth disparities and the measures needed to promote economic convergence or greater territorial cohesion (see for example Baldwin et al., 2003).

Since industrial agglomeration is a core question in regional science, spatial concentration measures are widely employed to evaluate the agglomeration of activities (Combes and Overman, 2004). Increasing access to large spatial datasets and computing power has encouraged the development of statistical analysis tools for processing such data most efficiently (Baddeley et al., 2016). Empirical studies at highly detailed geographical levels have been proposed in recent years, see for example d'Aspremont et al. (2025) or

Baragwanath et al. (2026) based on satellite imagery. In particular, the detection of spatial structures (attraction, repulsion, and independence) of individual spatialized data using analyses that are no longer based on zoned data but on geo-located data has received an increasing attention. Such an approach has the advantage of preserving the exact positions of the entities analyzed contrarily to data aggregated at a given spatial level (for example regions, counties etc.). Since the 2000s, a series of spatial concentration measures based on distances between geo-localized entities studied has been proposed (see Marcon and Puech, 2017, for a complete review). By considering space as continuous, the advantage of these distance-based methods is to circumvent any statistical bias associated to a discretization of space into separate units, known as the MAUP-Modifiable Areal Unit Problem (Openshaw and Taylor, 1979; Arbia, 1989). Numerous studies have shown the importance of using such a methodology in social sciences (Arbia, 1989; Marcon and Puech, 2003; Sweeney and Arabadjis, 2022) or in natural and physical sciences (Cressie, 1993; Lentz et al., 2011; Dray et al., 2021). For example, Giuliani et al. (2014) published in this review an analysis based on continuous space on the technology manufacturing industry in the metropolitan areas of Milan and Turin in Italy. They showed all the importance of taking into account the exact geographic position of the plants to provide a complete description of the location patterns of high- and medium-high-technology manufacturing. Notably, they detected that establishments were spatially aggregated at distances greater than 500 meters in the metropolitan area of Milan.

One of the existing distance-based methods has been recently pointed out by Tidu et al. (2024): the M function, a particular statistical measure proposed by Marcon and Puech (2010). This distance-based method, hereafter referred to as M, makes it possible to highlight spatial structures within a spatialized distribution (attraction, repulsion, and independence) from a study based on the distances separating the analyzed entities. Various applications of M have been proposed for detecting industrial agglomeration. For example, Jensen and Michel (2011) and Floch et al. (2018) analyzed the spatial concentration of different facilities in French cities. At a regional level, Coll-Martínez et al. (2019) proposed an application of the M function on creative industries in a Spanish metropolitan area while more recently, Canello et al. (2025) employed M in particular to characterize the location patterns of footwear subcontractors in Italy. The M function has also been rapidly applied to other research topics such as the spatial distribution of ethnic groups (Deurloo and De Vos, 2008). However, while this measure preserves all the richness of individual geo-located data, it requires a longer calculation time than other distance-based measures, as it is a relative measure (see Marcon and Puech, 2017). This might be a limit of the use of the M function if datasets are very large. Tidu et al. (2024) proposed decreasing M calculation times by introducing a voluntary positioning error for the analyzed entities. In their study, industrial establishments in Sardinia (Italy) were not located at their exact addresses but at the centroid of their municipality. This repositioning reduces calculation times, as the number of possible distances between establishments is limited to the distances that separate the centroids of the municipalities. However, the information loss owing to the approximation of the location of objects should imply a loss of accuracy in the estimation of their interactions at the same scale, which needs to be assessed.

Herein, we propose testing the effectiveness of Tidu et al. (2024)'s method and help researchers in regional science choose the appropriate method to characterize the spatial structure of substantially large datasets. The article provides three main contributions to the literature. First, we show the advantages of estimating M on datasets with an order of magnitude of 100,000 points or less. On a personal computer, the computation time becomes excessive beyond that. Second, we study the effect of the geographical approximation of the locations of the analyzed entities in terms of the deterioration in information that this approach creates. Third, we show that an alternative method, i.e. random thinning of the point pattern (Sweeney and Konty, 2005; Arbia et al., 2017), is more efficient.

The remainder of this article is organized as follows. The first section outlines the importance of agglomeration measurement in regional science. We also explain why an increasing number of researchers are working with very large micro-geographic datasets. The next section presents the M function and the necessary data generated for the tests. Consideration is given to large point sets (in the order of several tens of thousands of points) that are either completely random or geographically concentrated. We detail the use of the *dbmss* R package to calculate M from a table that gives the position and characteristics of the points or a matrix of distances between them. In this section, we also discuss the performance of *dbmss* with respect to the size of the set of points, in terms of computing time and memory requirements. The third section positions our approach by defining two strategies to reduce the computation burden of large geo-located datasets: approximating the position of points (grouping them at the center of the cells of a grid, following the approach of Tidu et al. (2024)) and thinning the point pattern. We discuss the advantages and the limits of both approaches. The last section concludes.

1. Spatial concentration measures and regional science

1.1 A close connection

In regional science, spatial concentration measures are widely employed to evaluate the agglomeration of activities (Combes and Overman, 2004). These statistical measures provide a simple way to describe the geographic distribution of industrial activities across regions. Industrial agglomeration generates productivity gains for firms and benefits for households (Fujita et al., 1999; Glaeser, 2010; Fujita and Thisse, 2013) but generates growth disparities between territories. Regional economic inequalities can be detected and quantified with spatial concentration measures. In that way, a lot of researchers have employed them for a better description and understanding of industrial clusters (see Feser and Sweeney, 2000; Giuliani et al., 2014; Kerr and Kominers, 2015, among others). Moreover, since spatial concentration measures are proven to be a relevant technique to describe agglomeration across territories, some studies rely on them to explain the mechanisms that generate the industrial clustering. The pioneer explanation of the agglomeration forces is due to Marshall (1890) who identified three main determinants of the industrial spatial concentration: a specialized labor pooling, the proximity of producers of intermediate goods (input sharing), and the opportunity to benefit from technological externalities (knowledge spillovers). These factors explain why firms concentrate their activities in a limited number of specific geographic areas: to reduce their costs and risks associated to their production. These three determinants are still widely accepted today, even though the modern view of these economic mechanisms proposed by Duranton and Puga (2004) is approached by a better matching, a better sharing, and a better learning for firms and households. A series of articles have empirically evaluated the importance of the determinants of agglomeration by regressing estimates of spatial concentration measures on the main factors that generate agglomeration. This technique constitutes a very simple way to quantify the importance of all of the determinants of industrial agglomeration: see for example the empirical studies of Rosenthal and Strange (2001), Ellison and Glaeser (1999), or the review of Rosenthal and Strange (2004)[1]. Estimating at best the importance of all factors at work is straightforward, notably for policymakers to provide effective economic policies in order to foster regional growth and to attenuate disparities between regions.

1.2 Recent developments

Based on the importance of the industrial geographic concentration, a great deal of attention has been paid in economics and in geography to develop statistical measures that are able to describe the spatial concentration precisely and without any statistical bias (Arbia et al., 2021). Today, many indices coexist in the literature but two distinct groups are clearly identified depending both on the nature of the data used and on their statistical treatment.

The first group of measures divides space into a given number of exclusive zones, for example data are aggregated at the county or region level. The well-known and widely used location quotient (Florence, 1972), the Gini (1912) index or the Ellison and Glaeser (1997) index belong to that group. It is out of the scope of this article to present them in details but a discussion of these spatial concentration measures is given in Sweeney and Feser (2004), Bickenbach and Eckhardt (2008) or Combes et al. (2008). Entropy measures also belong to that first group, even if they are less employed in spatial economics or in quantitative geography (see for example Brülhart and Traeger (2005) for an evaluation of spatial concentration of activities across regions with the Theil (1967) index). The common advantage of all of these measures is their ease of computation: they do not need intensive calculations, data are spatially aggregated so they are easy to find and process[2].

The shared limit of all of the previous indices is that they rest on a pre-defined zoning of space. As a consequence, all the information at a finer scale than the zoning delimitation is simply lost, preventing to study the short-distance interactions between firms, the face-to-face contacts between individuals etc. In addition, a series of studies put in light that all indices based on a discrete space are exposed to some statistical bias known as the Modifiable Areal Unit Problem-MAUP (Openshaw and Taylor, 1979; Arbia, 1989). For example, Arbia (2001) proved that the position of the frontiers of the zones or the geographical scale chosen for the analysis alter the findings on the spatial concentration levels. Finding the 'right' scale to analyze agglomeration factors remains an open issue in this framework (Rosenthal and Strange, 2020).

An answer to the MAUP issues was proposed by the development of a second group of measures that not divide space into a given number of zones. This second group encompasses all 'distance-based measures' (DBMs) based on point pattern analysis. In that case, the exact geographic position of firms is preserved and the spatial concentration levels of agglomeration is calculated from the distances between the geo-located firms studied. By avoiding aggregation, DBMs circumvent all statistical bias associated to a discretization of space into separate units. By considering space as continuous, DBMs evaluate the spatial concentration at all scales. One of the advantages of distance-based methods is to detect the exact distance at which spatial concentration (in case of at-

[1] This technique is also employed for understanding the determinants of coagglomeration (Ellison et al., 2010; Diodato et al., 2018).

[2] For example, Marcon (2019) employed European data (available at the following address: https://ec.europa.eu/eurostat/web/regions-and-cities) to measure industrial employment disparities across European regions.

traction) or dispersion (in case of repulsion) occur.

In the last decades, many developments have been proposed in geography and in economics to be able to grasp any definition of the spatial concentration (different benchmark distributions, weighting or not the plants by their number of employees etc.). Dozens of spatial concentration measures based on distances have been proposed (Marcon and Puech, 2017). Among the first applications of distance-based methods, the works of Barff (1987) or Sweeney and Feser (1998) are of particular interest. They investigated the relationship between the agglomeration of manufacturing activities and the characteristics of the plants (capital intensity, size, etc.). Sweeney and Feser (1998) notably gave support to an existing inverted U-shaped relation between the spatial manufacturing concentration and the size of plants in North Carolina (USA). Today, DBMs face a new challenge: they must be able to address the huge amount of micro-geographic data that are increasingly available for researchers (Piacentino et al., 2021). Dealing with satellite data is an opportunity to provide detailed location analysis of the economic activities, see for example recent studies of Bachtrögler-Unger et al. (2023), d'Aspremont et al. (2025) or Baragwanath et al. (2026). This kind of data opens new research questions both on the computational capacity to manage very large datasets and the possible reductions of computing time. In a recent paper Tidu et al. (2024) focused their attention on one of the existing distance-based methods, the M function (Marcon and Puech, 2010). In the next sections, we shall discuss the computing performances of the M function and we also try to appreciate its capacity to deal with very large datasets.

2. The *M* function, for any geo-located dataset?

2.1 Main idea

Marcon and Puech (2010) introduced the M function to evaluate the dependence between geo-located points without relying on a specific zoning of space. As a distance-based method, the calculation of M is based on distances that separate entities studied (establishments, shops…). The idea of M is simple: it compares two proportions of neighbors of interest, a local one to a global one. The local one is defined as the proportion of neighbors of interest within a distance r. The global one is the same proportion but defined on the entire territory. This comparison of ratios enables the detection of:

- spatial concentration (attraction) of entities if the proportion of local neighbors is greater than the one observed on the entire territory,
- spatial dispersion (repulsion) of entities if the relative proportion of local neighbors is lower than the one observed all over the territory,
- independence between entities if the local distribution of neighbors does not differ from the global one.

This comparison of proportions of neighbors defines M strictly as a *relative* distance-based measure (Marcon and Puech, 2017). The term *topographic* distance-based measures designates those that use the surface area as a benchmark, as the well-known Ripley's K function or L function (Ripley, 1976, 1977; Besag, 1977). The K_d function (Duranton and Overman, 2005) is an *absolute* DBM since it has no benchmark at all. The M function is also defined as a *cumulative* distance-based method because the local environment is appraised up to a distance r rather than at a distance r. The possibility to detect exactly at which distance(s) the spatial concentration or dispersion appears coupled with the interpretation of results opens the way to precisely describe the distribution of entities under study. An easy computation of M is possible owing to the *dbmss* R package (Marcon et al., 2015).

M was first introduced in the field of economics and geography by Marcon and Puech (2010). They proved that this function satisfies all the requirements of Duranton and Overman (2005) for the evaluation of the spatial distribution of industries. Since then, various studies have described the spatial locations of industries by using it: for example, Jensen and Michel (2011) studied the location of shops at an urban level and Coll-Martínez et al. (2019) analyzed creative industries at a metropolitan level. This methodology has also been rapidly applied in other domains (Fernandez-Gonzalez et al., 2005; Deurloo and De Vos, 2008; Marcon et al., 2012; Nissi et al., 2013). The M function is now included in general textbooks of spatial statistics such as Arbia et al. (2021).

2.2 Definition

The M function is based on the point process theory, in line with Ripley (1977). The M function was developed to analyze interactions among entities in a heterogeneous space. In other words, within this statistical framework, we consider that any entity analyzed does not have the same probability to locate everywhere (the *first-order property of the point pattern* is its intensity). Subsequently, upon controlling for space heterogeneity, we can identify interactions and detect spatial concentration or dispersion (the *second-order property of the point pattern*). Space heterogeneity is a consistent assumption for studying the agglomeration of industries (refer to the discussion of Duranton and Overman, 2005, on that subject).

The definition of the univariate (or 'intra-type') version of M is as follows. It compares the relative proportion of entities of interest up to each distance r to the same ratio but defined over the entire territory under study. In this article, we only consider the intra-type version of M: we study the spatial structure of neighboring points of the same type (called *points of interest*) as the points at the centers of the disks (called *reference points*) of radius r, as opposed to the bivari-

ate (intertype) version that addresses points of interest of a different type from the reference points. In mathematical terms, we denote the terms mentioned below:

- x_i^s, the location of point i of the reference type s, at the center of the disk (the point whose neighborhood is to be analyzed).
- x_j^s, the location of a neighbor j of the same type as point i.
- x_j, the location of a neighbor j of i, regardless of its type.
- $w(.)$, the weight of a given neighbor. In that sense, $w(x_j)$ 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, regardless of their type.
- $\mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right)$, the indicator function is equal to one 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, zero otherwise.

The intra-type M function is defined as (1).

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

Several remarks must be made. The first remark is that the benchmark value of M is equal to one, regardless of the distance considered. It means that for any radius r:

- If the estimated M result is above one, the local value of the ratio is greater than the global one: a spatial concentration of entities of type s within that radius is thus detected.
- If the estimated M result is under one, the local value of the ratio is lower than the global one: a spatial dispersion of entities of type s within that radius is thus detected.

The second remark concerns the significance of the results. A confidence interval can be generated using Monte Carlo simulations following Marcon and Puech (2010). A risk level is chosen (for example 5%) as well as the number of simulations. The greater the number of simulations, the longer is the duration of the calculation of M. Third, the package *dbmss* (Marcon et al., 2015) on the R software (R Core Team, 2026) can be used to compute the M function. In most studies, the Euclidean distance is preferred to calculate M, but the *dbmss* package can also be used for network distances. We discuss that point hereinafter for large datasets.

2.3 Simulated data

The datasets we consider in this article were obtained by simulation. The R code is given in the appendix, which allows perfect reproducibility of the examples treated.

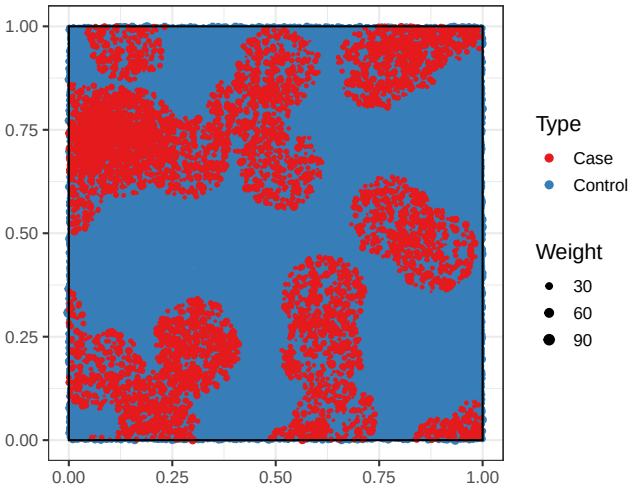

Figure 1. Random drawing of a set of points where the Cases (red) are aggregated and the Controls (blue) are distributed completely randomly. The size of the points is proportional to their weight.

2.3.1 Drawing the points

A set of points is drawn using the Poisson process (whose expectation of the number of points is 100,000) in a square window of side one. Each point is assigned a qualitative mark: 'Case' or 'Control'. 95% of points are Controls. Additionally, 5% are Cases, whose spatial structure is studied. The weight of the points is drawn from a gamma distribution with free shape and scale parameters.

In this example, the drawing of Controls points is completely random (*complete spatial randomness*: CSR), i.e., there is no simulation of attraction or dispersion. On the contrary, the spatial distribution of Cases is aggregated, which means that Cases are spatially concentrated (like clusters). Practically, sets of aggregated points are drawn in a Matérn (1960) process.

Cases shown in Figure 1 indicate visible aggregates. Controls are distributed completely randomly. A careful analysis of Figure 1 shows a limited number of tiny white spaces on the square window, indicating locations with no points (whatever the type). The scarcity of empty spaces is because of the high number of simulated points.

2.3.2 Gridding the space

The simulation of the Cases obtained by the Matérn process is considered and the window is split into a 20 x 20 square grid. This partition simulates the approximation of the position of the points of an administrative unit to the position of its center. Moreover, this grid size is consistent with that of Tidu et al. (2024), which facilitates a comparison of our results. The choice of the optimal level of the grid remains an open question, as Arbia et al. (2021) noticed (p.109): '*Unfortunately, the choice of the partitioning scheme is usually arbitrary and an optimal criterion to guide this choice is not available.*'

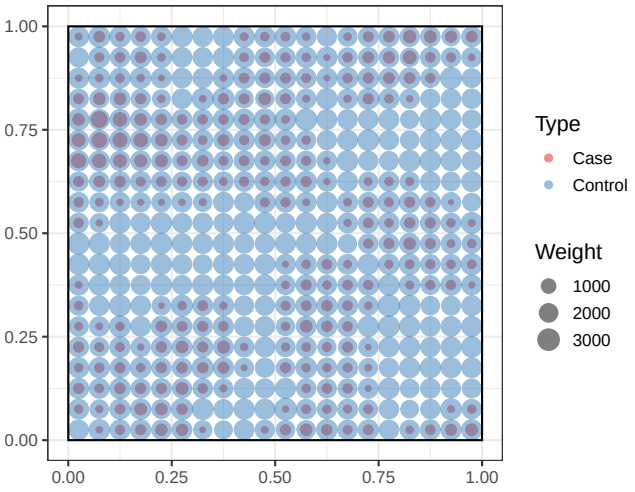

Figure 2. Repositioning of points in an arbitrary grid. The absence of Cases in a cell is easily detected (single-color blue dot), as is the strong presence of Cases in a cell (two-color dot, but predominantly red).

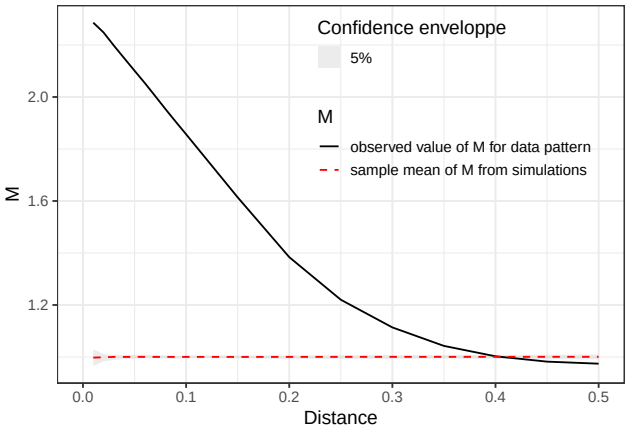

Figure 3. Value of M as a function of the distance from the reference point. The 95% confidence envelope, obtained from 100 simulations, appears in gray and is centered on the value of one.

The approximated position of points is depicted on the map presented in Figure 2. Each cell comprises only one point of each type, whose weight is the sum of the weights of the individual points.

M values can be calculated from the original point set or its approximation.

2.4 Computing *M* using the *dbmss* package

In the *dbmss* package, the data are a set of points or a distance matrix. The set of points in Figure 1 is used.

2.4.1 Point pattern

The `Mhat()` function in the *dbmss* package is used to estimate the M function. The theoretical reference value for M is one, as this function relates the proportion of Cases up to a distance r to that observed across the entire window. The aggregation of Cases will be highlighted by values of $M > 1$ (the relative presence of Cases is greater locally than across the entire window) and the dispersion of Cases by values < 1. The Euclidean distance is used to estimate the distance between points.

Figure 3 shows that M detects an agglomeration of Cases, which is in line with the simulation of this type of point (Controls having a completely random location on the window). The advantage of a function based on distances is clearly visible: for any distance, the level of spatial concentration is precisely estimated. It enables the detection of distances at which the attraction phenomena occur and are the most important (for functions whose values can be compared at different radii, such as M). In addition to estimating the M function, the `Menvelope()` function can be used to calculate its global confidence interval (Duranton and Overman, 2005) under the null hypothesis of random point location. The confidence interval of the null hypothesis is centered on one, and is narrow because of the large dataset. Lastly, the necessary simulations can be parallelized to save time.

2.4.2 Distance matrix

Matrices can be used to process non-Euclidean distances (transport time, road distance, etc.) which cannot be represented by a set of points. The `Mhat()` and `MEnvelope()` functions are the same, and provide the same results irrespective of the data form (point set or distance matrix). Details are given in the appendix.

The size of distance matrices is the square of their number of points, i.e. 10^{10} cells for 100,000 points. R does not handle them because it relies on integer values to index them: square matrices cannot be larger than 46340 rows and columns, i.e., the square root of the largest supported integer value. This does not allow dealing with large datasets. Thus, this approach is not explored further in this paper.

2.5 Computational performance

The use of M to characterize the spatial structure of large sets of points might be limited by its computing time or the memory required. In this section, we investigate these two potential limitations.

2.5.1 Computing time

The distances between all pairs of points must be calculated to estimate M. Therefore, it is expected that the calculation time will increase as the square of the number of points. The time required for the exact calculation is evaluated for a range of numbers of points (Figure 4a).

The calculation time is related to the size of the set of points by a power law. It increases less rapidly than the square of the number of points. It can be estimated precisely ($R^2 = 0.98$) using the mentioned relation: $t = t_0(n/n_o)^p$, where t denotes the average time for n points (e.g., 2.88 seconds for 100,000 points), knowing the time t_0 for n_0 points, and p is the power relation (here: 1.5).

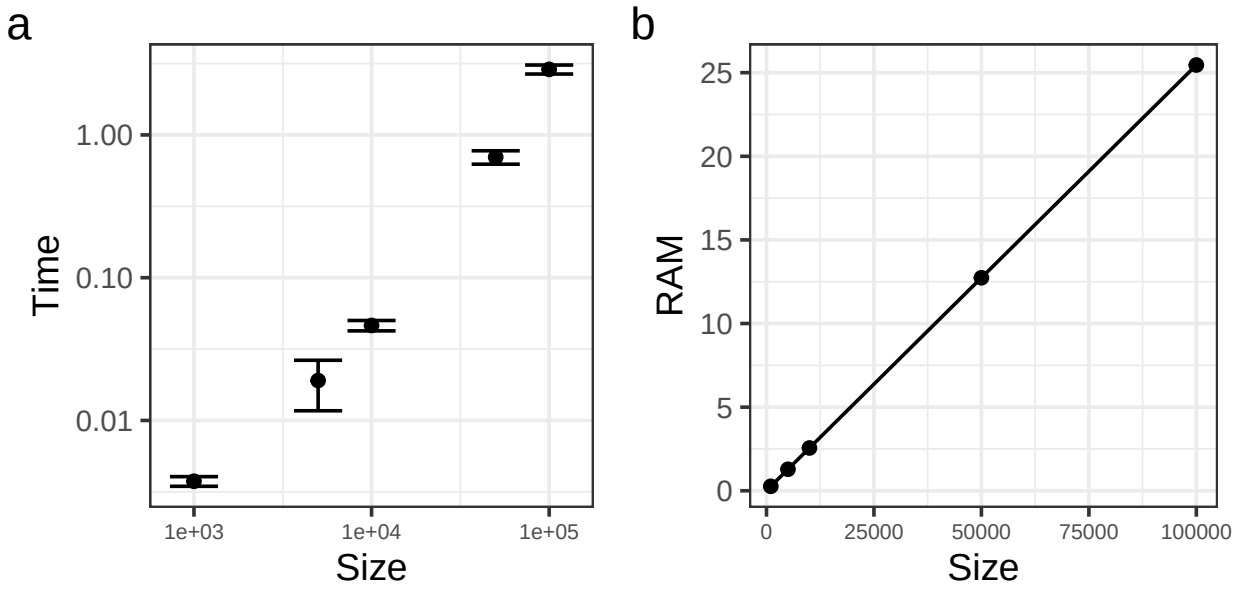

Figure 4. (a) Calculation time (seconds) and (b) memory required (MB) to estimate M as a function of the size of the set of points. The measures were repeated 10 times. The bars represent the +/- 1 standard deviation interval.

2.5.2 Memory

The memory used is evaluated for the same data sizes (Figure 4b). **The memory required increases linearly with the number of points and is never critical for point set sizes that can be processed within reasonable times**. This highlights Tidu et al. (2024)'s conclusion regarding the power and computation time required when using M on large datasets.

2.6 Discussion: technical limit of computing *M*

The first conclusion is that the computation burden of estimating M on large datasets might be a problem. The calculation time for M is < 3 seconds for a set of 100,000 points on a modern computer[3] and requires 25 MB of RAM. Therefore, calculating a confidence interval from 1,000 simulations requires less than 50 minutes. For a set of five million points, the power law predicts around 10 minutes of computing time. Accordingly, 1,000 simulations would take around 7 days.

Owing to parallelization, a calculation server would drastically increase performance, but at the cost of the complexity of implementation that limits its use. **If we limit ourselves to the computing power of a personal computer, exact calculation is fully justified for data of the order of 100,000 points**: less than an hour is sufficient to calculate confidence intervals. Since parallelizing the simulations is offered with no effort by the *dbmss* package, this time can be reduced by a considerable factor based on the available hardware, e.g., by a factor of 2 to 6 using modern multicore central processing units. Beyond that, approximating the location reduces the size of the set of points to the number of locations selected. It may be up to 100,000 locations to keep the computing time acceptable, regardless of the size of the original dataset.

[3]The results presented here were obtained on a GitHub-hosted Mac OS runner with a virtual 3-core Apple M1 (Virtual), similar to a fast laptop computer.

3. A guide to dealing with large datasets

Computation time becomes critical when very large datasets are considered, as shown in the previous section. The M function is often used to test a point pattern against a null hypothesis, requiring simulations of a null point pattern and the same computation for each of them, hence multiplying the computing time. Here, we explore two possible ways to reduce the computational burden: approximating the position of points, and thinning the point pattern. Our purpose is to identify the potential biases caused by both strategies, and consequently identify the best practice.

3.1 First strategy: Approximating the position of points

3.1.1 Two main effects

Unambiguously, approximating the position of the points generates a first loss of information: in each grid cell, the distance between all the points is set to zero. Grouping points at the centroid of cells erases any spatial structure under the grid size. Consequently, this first statistical bias must be investigated in details.

The second loss of information is due to the distance between two points in different cells, which is approximated using the distance between the centroids of the two cells. The resulting error in the estimation of M may appear at a scale of the order of the mag-

nitude of the size of the cells, which should decrease with distance, when the relative size of the cells becomes negligible. This effect due to approximation must be studied with care.

3.1.2 Test setting and results

The effect of the location approximation is tested on a set of aggregated points, similar to the real Tidu et al. (2024) data, and on a completely random point pattern. Both comprise 100,000 points such that groups include 250 points (100,000 points divided by 400 groups) on average. 5% are Cases.

20 sets of completely random and of aggregated points were simulated. The exact calculation and the calculation on the grid points were performed on each set of points to evaluate the effect of the approximation.

The mean values of the estimates of M are presented in Figure 5. The size of the grid cells is equal to 0.05. All neighbors at distances less than this threshold are placed at zero distance. The estimate of the corresponding M function (Grouped M) is constant in that case, up to this threshold. When the actual point pattern is not structured, artifactual patterns are generated at a small scale. In contrast, the local aggregation of the Matérn pattern is detected but underestimated.

Tidu et al. (2024) tested the effect of grouping the points by the correlation between exact and approximated M values. We refrained from following them because the main source of variation of the grouped-point M values is that of the grouping itself: when the number of points per group is small, groups considerably vary between simulations, and the correlation is weak. It increases with the size of the data (an illustration is given in the appendix). A substantially high correlation does not exclude systematic errors. Therefore it is not considered an appropriate statistic here.

The M function is basically used as a test against the independence of point locations. To assess the impact of grouping the points on the test result, Figure 6 shows its average p-value, computed among 20 simulations of each point pattern. At each distance, the M value is compared with that obtained with simulated point patterns based on the null hypothesis, which is rejected if the actual value is outside of the 95% central quantiles of the simulations.

In the case of CSR, the p-value to reject independence around 50%, regardless of the distance, and whether the points are grouped or not. In the case of aggregated patterns, the null hypothesis is rejected correctly when the points are grouped. At small scale, below the size of the cells, M values are actually that of the cell size: the test is correct because the point process we used is aggregated at all scales. The point pattern might have been repulsive at the small scale, but this information is destroyed by grouping the points: M and their p-values are not reliable below the size of the cells.

Finally, the study of Arbia et al. (2017) must be mentioned. They proposed a first evaluation of the consequences of an *'unintentional positional error'* owing to the uncertainty of the location of a part of the studied entities, e.g., when their address is not accurate. In that scenario, these uncertain geo-localized entities are placed at the centroid of the zone considered, exactly as in Tidu et al. (2024). Applying different distance-based methods on a real case (Italian manufacturing firms in the province of Trento), Arbia et al. showed that the error measurement was less important than expected. For M, their explanation rests on the definition of a relative measure: a compensation effect of positional errors is suspected between the local and the global ratios.

3.1.3 Discussion

We can easily understand that applying a grid is an attractive technique to save calculation time, but the risk is missing the geographic scale at which the spatial phenomenon occurs. **Since all information below the grid size is lost, the approximation might or might not be acceptable depending on the research question and the spatial scale at which interactions occur. Above the grid size, the M function is less affected by the approximation and the gap becomes rapidly negligible**. Let us come back to a core question in regional science given in section 1, the industrial clustering, to illustrate the challenge of finding the optimal grid size. Knowing the real extent of agglomeration economies is not an easy task but it has been proven that they decay rapidly as the distance increases (Rosenthal and Strange, 2003, 2020). A careful use of location approximation is thus at first recommended if short-distance interactions are suspected, such as information externalities or contagion phenomena. Moreover, empirical studies show that the scale at which plants agglomerate varies substantially. Many factors influence the findings: the industry studied, the distance-based measure employed, the area considered, etc. To give some examples, Giuliani et al. (2014) found that plants in high- and medium-high-technology manufacturing sectors were spatially aggregated at distances greater than 500 meters in the metropolitan area of Milan while no significant pattern we observed in the same industry in the metropolitan area of Turin. In a pioneer study, Sweeney and Feser (2004) detected that, in Los Angeles, the highest level of spatial concentration of establishments in electronics appeared within a distance of 1 kilometer while for textiles plants the agglomeration was greatest within a distance of 4 kilometers. In their seminal article, Duranton and Overman (2005) showed that location patterns were very different from an industry to another by taking four *ad hoc* industries in the United Kingdom (basic pharmaceuticals, pharmaceutical preparations, other agricultural and forestry machinery, and machinery for textile, apparel and leather production). These results are perfectly in line with Feser and Sweeney

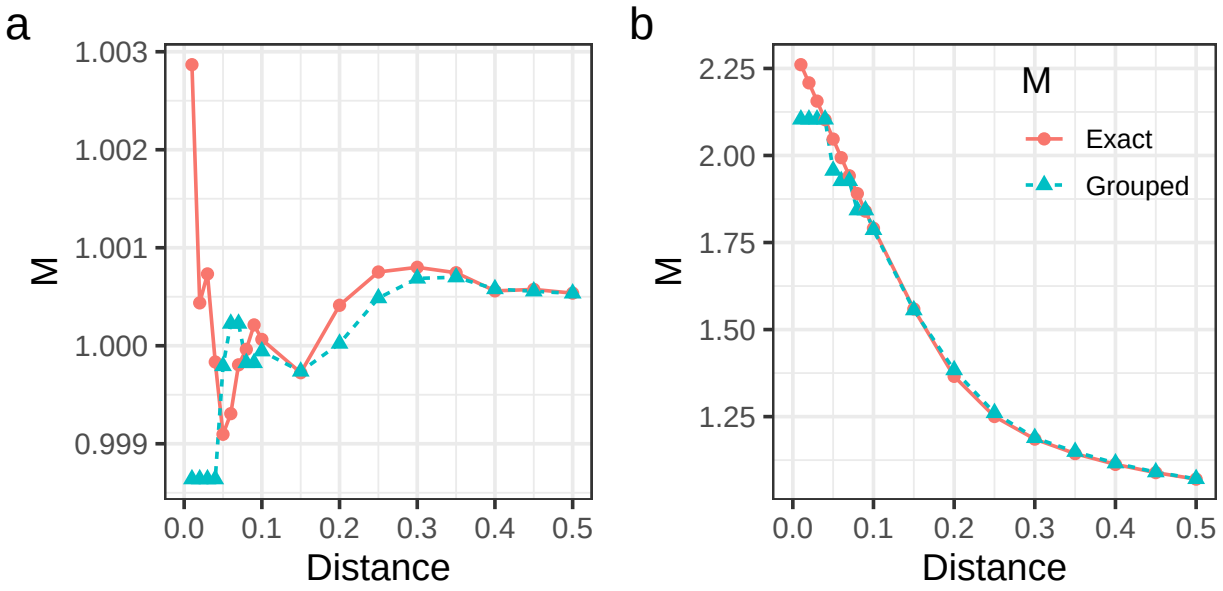

Figure 5. Average estimate of M from the exact position of the points compared with the values obtained by grouping the points for CSR (a) and Matérn (b) point patterns.

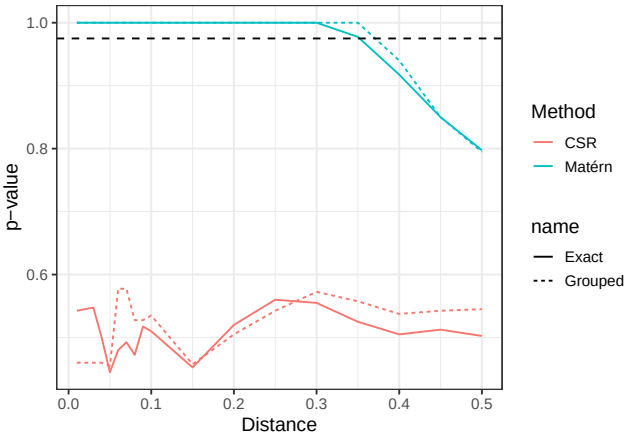

Figure 6. p-values to reject the null hypothesis (H0) of independence of the point locations tested by the M function. The colors of the curves represent CSR or aggregated point patterns. Solid lines provide the exact p-values of M, and dotted lines provide the values estimated on grouped point patterns. The horizontal, dotted line corresponds to the significance threshold, i.e., 97.5%, to reject H0 in a 5% risk-level two-tailed test.

(2002)'s findings for the United States who provided a very detailed analysis for a large number of industries in fourteen American metropolitan areas (see tables on pages 245-246 in their article).

In sum, there is no general rule to decide an appropriate grid size. Grouping the points for computing performance may result in missing the geographical scale at which agglomeration economies occur: this a classical MAUP's issue. The interest of DBMs is actually to avoid this issue by preserving the information contained in the location of points.

3.2 Second strategy: Random thinning

3.2.1 Test setting and results

Random thinning consists of reducing the number of points by applying the same probability of being kept to each point. The resulting dataset is thus a representative sample of the actual one.

Compared to approximating the position of points, random thinning does not degrade the information of each point but reduces the amount of data. The acceptability of the method can be tested rigorously by simulating randomly thinned point patterns and comparing the distribution of their M function to that of the original pattern. A million-point pattern is drawn with 5% of aggregated Cases and thinned with down to 10% of its size. The M values of the original and 100 thinned patterns are compared. Figure 7a shows the mean value and the confidence envelope of the M function applied to thinned patterns. As expected, random thinning causes no bias. The relative Root Mean Squared Error (RMSE), that is the expectation of the error caused by thinning, is shown on Figure 7b. Since the bias is zero, it is simply the coefficient of variation of the simulated M values at each distance. It is very small, around 2%, at short distances and

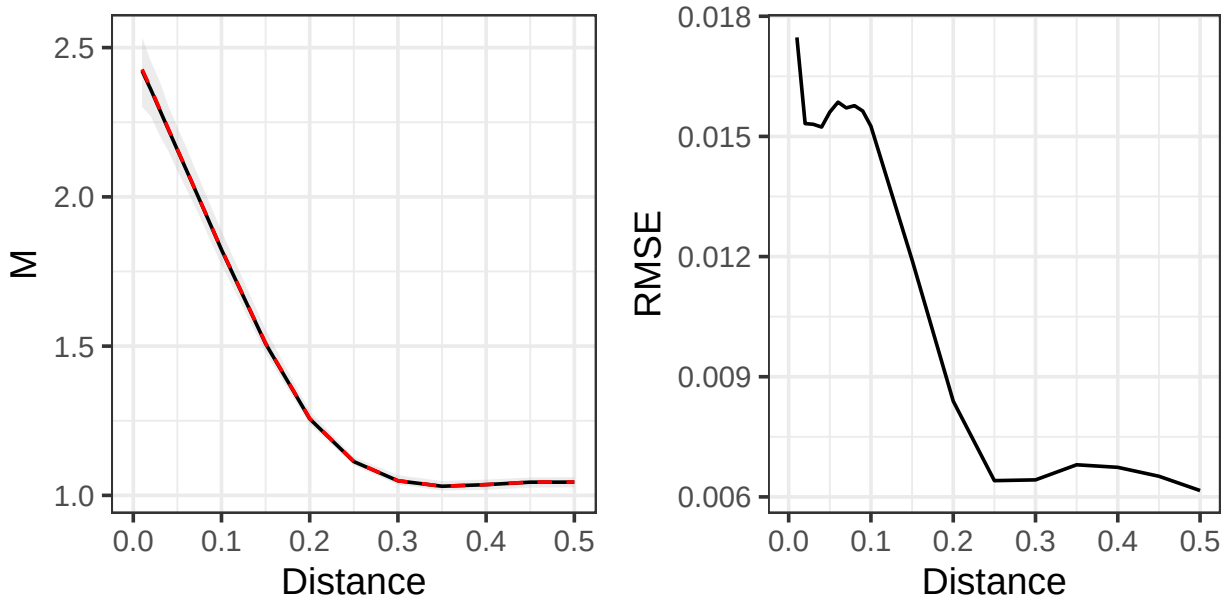

Figure 7. (a) Confidence envelope of the estimation of the M function after thinning a 1-million-point pattern with aggregated Cases to 10% of its size. The mean M values among thinned patterns (dotted, red line) can't be distinguished from the exact values (black line). The grey envelope corresponds to the 95% confidence level. (b) relative Root Mean Squared Error of the estimation due to thinning, i. e. the expected relative error of estimation, always below 2%.

negligible at large distances.

3.2.2 Discussion

The random thinning technique has been already explored by Sweeney and Konty (2005) on theoretical examples, and also on the location of North Carolina manufacturing establishments. They utilized others DBMs: the K function (Ripley, 1976, 1977) and the D function (Diggle and Chetwynd, 1991). Both are topographic DBMs but the latest that allows controlling for heterogeneity of space. Sweeney and Konty (2005) showed that the K and D functions are invariant under random thinning.

So is the M function: it is obtained by dividing a number of neighbors of interest by a total number of neighbors. Both terms are reduced in the same proportion by random thinning, keeping the ratio unchanged. As Sweeney and Konty (2005) already noted for the K function, the variance of the estimation of M increases when the number of points decreases. Thinning a 100,000-point dataset to 1000 points results in a very fast and unbiased estimation of the M function but its RMSE is close to 100% at short distance (Figure in supplementary material). Finally, our results provided on large datasets confirm previous findings of Arbia et al. (2017). They investigated the consequences of missing clustered data on the M function. In their study, data were not degraded homogeneously but some data were locally missing. Their empirical application on manufacturing firms located in Trento (Italy) provided evidence that M findings are 'quite robust' in that case.

To sum up, random thinning is an efficient method to reduce huge, untractable datasets to large, computable ones, but is not an appropriate way to drastically reduce computing time by thinning large datasets to small ones.

4. Conclusion

In this article, we provide some answers to a new recent challenge faced by all DBMs: the huge amount of micro-geographic data that are increasingly available for researchers. We focus on one particular DBM, the M function of Marcon and Puech (2010), a DBM recently pointed out by Tidu et al. (2024). We think that our results will be useful for researchers in regional science who work with geo-localized data, notably for studying the industrial clustering.

Our study highlighted several important points. We support that studying 100,000 locations or less keeps the M computing time acceptable, regardless of the size of the original dataset. In all other cases, the location approximation or the random thinning can be considered. Regarding the errors generated in the estimates of M when the approximation of location is used, our findings support those of Tidu et al. (2024), which mentions strong correlations between M values computed from exact and approximated Italian company location data. The spatial approximation problem originates from the loss of information regarding possible interactions at distances smaller than the size of the grid. Random thinning is thus the good option to deal with very large datasets, avoiding any statistical bias but suffering from excessive variance if the resulting point pattern is small. It is thus recommended to thin very large point patterns down to the largest tractable size possible, typically

100,000 points considering the performance of current personal computers. Even if it was out of the scope of this study to test the performance of other DBMs, we provide also detailed comparisons between widely used DBMs in the appendix: the M function, the D function of Diggle and Chetwynd (1991) and the K_d function of Duranton and Overman (2005).

Appendix

R code is available at the following address: https://ericmarcon.github.io/MLargeDataSets/Appendix.pdf

Acknowledgments

We thank the editor and three anonymous referees for helped us improve the article. We also thank participants of the SEW 2025 and especially Vincenzo Nardelli for helpful advice. Eric Marcon benefited from an "Investissement d'Avenir" grant managed by the Agence Nationale de la Recherche (LABEX CEBA, ref. ANR-10-LBX-25) and Florence Puech gratefully acknowledges financial support from INRAE.

References

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

Arbia, G. (2001). Modelling the geography of economic activities on a continuous space. *Papers in Regional Science 80*(4), 411–424.

Arbia, G., G. Espa, and D. Giuliani (2021). *Spatial Microeconometrics*. Routledge Advanced Texts in Economics and Finance. London and New York: Routledge, Taylor & Francis Group.

Arbia, G., G. Espa, D. Giuliani, and M. M. Dickson (2017). Effects of Missing Data and Locational Errors on Spatial Concentration Measures Based on Ripley's K-Function. *Spatial Economic Analysis 12*(2-3), 326–346.

Bachtrögler-Unger, J., M. Dolls, C. Krolage, P. Schüle, H. Taubenböck, and M. Weigand (2023). EU cohesion policy on the ground: Analyzing small-scale effects using satellite data. *Regional Science and Urban Economics 103*, 103954.

Baddeley, A., E. Rubak, and R. Turner (2016). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton London New York: CRC Press.

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano, and F. Robert-Nicoud (2003). *Economic geography and public policy*. Princeton University Press.

Baragwanath, K., G. Hanson, A. K. Khandelwal, C. Liu, and H. Park (2026). Using satellite imagery to measure the impacts of new highways: An application to India. *Journal of International Economics 159*, 104201.

Barff, R. A. (1987). Industrial Clustering and the Organization of Production: A Point Pattern Analysis of Manufacturing in Cincinnati, Ohio. *Annals of the Association of American Geographers 77*(1), 89–103.

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

Bickenbach, F. and B. Eckhardt (2008). Disproportionality Measures of Concentration, Specialization, and Localization. *International Regional Science Review 31*(4), 359–388.

Brülhart, M. and R. Traeger (2005). An account of geographic concentration patterns in Europe. *Regional Science and Urban Economics 35*(6), 597–624.

Canello, J., F. Vidoli, E. Fusco, and N. Giudice (2025). Identifying and Mapping Industrial Districts Through a Spatially Constrained Cluster-Wise Regression Approach. *Journal of Regional Science 65*(2), 403–428.

Coll-Martínez, E., A.-I. Moreno-Monroy, and J.-M. Arauzo-Carod (2019). Agglomeration of Creative Industries: An Intra-metropolitan Analysis for Barcelona. *Papers in Regional Science 98*(1), 409–432.

Combes, P.-P., T. Mayer, and J.-F. Thisse (2008). *Economic Geography, The Integration of Regions and Nations*. Princeton: Princeton University Press.

Combes, P.-P. and H. G. Overman (2004). The spatial distribution of economic activities in the European Union. In J. V. Henderson and J.-F. Thisse (Eds.), *Handbook of Urban and Regional Economics*, Volume 4, Chapter 64, pp. 2845–2909. Amsterdam: Elsevier. North Holland.

Cressie, N. A. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

d'Aspremont, A., S. Ben Arous, J.-C. Bricongne, B. Lietti, and B. Meunier (2025). Satellites turn "concrete": Tracking cement with satellite data and neural networks. *Journal of Econometrics 249*, 105923.

Deurloo, M. C. and S. De Vos (2008). Measuring Segregation at the Micro Level: An Application of the M Measure to Multi-Ethnic Residential Neighbourhoods in Amsterdam. *Tijdschrift voor economische en sociale geografie 99*(3), 329–347.

Diggle, P. J. and A. G. Chetwynd (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics 47*(3), 1155–1163.

Diodato, D., F. Neffke, and N. O'Clery (2018). Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics 106*, 1–26.

Dray, N., L. Mancini, U. Binshtok, F. Cheysson, W. Supatto, P. Mahou, S. Bedu, S. Ortica, E. Than-Trong, M. Krecsmarik, S. Herbert, J.-B. Masson, J.-Y. Tinevez, G. Lang, E. Beaurepaire, D. Sprinzak, and L. Bally-Cuif (2021). Dynamic Spatiotemporal Coordination of Neural Stem Cell Fate Decisions Occurs through Local Feedback in the Adult Vertebrate Brain. *Cell Stem Cell 28*(8), 1457–1472.e12.

Duranton, G. and H. G. Overman (2005). Testing for Localisation Using Micro-Geographic Data. *Review of Economic Studies 72*(4), 1077–1106.

Duranton, G. and D. Puga (2004). Micro-Foundations of Urban Agglomeration Economies. In J. V. Henderson and J.-F. Thisse (Eds.), *Cities and Geography*, Volume 4 of *Handbook of Regional and Urban Economics*, Chapter Chapter 48, pp. 2063–2117. Elsevier.

Ellison, G. and E. L. Glaeser (1997). Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy 105*(5), 889–927.

Ellison, G. and E. L. Glaeser (1999). The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration? *The American Economic Review 89*(2), 311–316.

Ellison, G., E. L. Glaeser, and W. R. Kerr (2010). What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *The American Economic Review 100*(3), 1195–1213.

Fernandez-Gonzalez, R., M. Barcellos-Hoff, and C. Ortiz-de-Solorzano (2005). A Tool for the Quantitative Spatial Analysis of Complex Cellular Systems. *IEEE Transactions on Image Processing 14*(9), 1300–1313.

Feser, E. J. and S. H. Sweeney (2000). A test for the coincident economic and spatial clustering of business enterprises. *Journal of Geographical Systems 2*(4), 349–376.

Feser, E. J. and S. H. Sweeney (2002). Theory, methods and a cross-metropolitan comparison of business clustering. In M. Philip (Ed.), *Industrial Location Economics*, pp. 222–257. Edward Elgar Publishing.

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

Florence, P. S. (1972). *The Logic of British and American Industry: A Realistic Analysis of Economic Structure and Government* (3rd ed.). London: Routledge & Kegan Paul.

Fujita, M., P. Krugman, and A. J. Venables (1999). *The Spatial Economy: Cities, Regions, and International Trade*. The MIT Press.

Fujita, M. and J.-F. Thisse (2013). *Economics of Agglomeration: Cities, Industrial Location, and Globalization* (2 ed.). Cambridge University Press.

Gini, C. (1912). *Variabilitá e mutabilitá*, Volume 3. Universitá di Cagliari.

Giuliani, D., G. Arbia, and G. Espa (2014). Weighting Ripley's K-Function to Account for the Firm Dimension in the Analysis of Spatial Concentration. *International Regional Science Review 37*(3), 251–272.

Glaeser, E. L. (2010). *Agglomeration Economics*. National Bureau of Economic Research, University of Chicago Press.

Jensen, P. and J. Michel (2011). Measuring Spatial Dispersion: Exact Results on the Variance of Random Spatial Distributions. *The Annals of Regional Science 47*(1), 81–110.

Kerr, W. R. and S. D. Kominers (2015). Agglomerative Forces and Cluster Shapes. *The Review of Economics and Statistics 97*(4), 877–899.

Lentz, J. A., J. K. Blackburn, and A. J. Curtis (2011). Evaluating Patterns of a White-Band Disease (WBD) Outbreak in Acropora palmata Using Spatial Analysis: A Comparison of Transect and Colony Clustering. *PLoS ONE 6*(7), e21830.

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 F. Puech (2003). Evaluating the Geographic Concentration of Industries Using Distance-Based Methods. *Journal of Economic Geography 3*(4), 409–428.

Marcon, E. and F. Puech (2010). Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods. *Journal of Economic Geography 10*(5), 745–762.

Marcon, E. and F. Puech (2017). A Typology of Distance-Based Measures of Spatial Concentration. *Regional Science and Urban Economics 62*, 56–67.

Marcon, E., F. Puech, and S. Traissac (2012). Characterizing the Relative Spatial Structure of Point Patterns. *International Journal of Ecology 2012*, 619281.

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

Marshall, A. (1890). *Principle of Economics*. London: Macmillan.

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

Nissi, E., A. Sarra, S. Palermi, and G. Luca (2013). The Application of M-Function Analysis to the Geographical Distribution of Earthquake Sequence. In A. Giusti, G. Ritter, and M. Vichi (Eds.), *Classification and Data Mining*, Chapter 32, pp. 271–278. Springer Berlin Heidelberg.

Openshaw, S. and P. J. Taylor (1979). A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem. In N. Wrigley (Ed.), *Statistical Applications in the Spatial Sciences*, pp. 127–144. London: Pion.

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

R Core Team (2026). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripley, B. D. (1976). The Foundations of Stochastic Geometry. *Annals of Probability 4*(6), 995–998.

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 39*(2), 172–212.

Rosenthal, S. S. and W. C. Strange (2001). The Determinants of Agglomeration. *Journal of Urban Economics 50*(2), 191–229.

Rosenthal, S. S. and W. C. Strange (2003). Geography, Industrial Organisation, and Agglomeration. *The Review of Economics and Statistics 85*(2), 377–393.

Rosenthal, S. S. and W. C. Strange (2004). Evidence on the Nature and Sources of Agglomeration Economies. In J. V. Henderson and J.-F. Thisse (Eds.), *Cities and Geography*, Volume 4 of *Handbook of Regional and Urban Economics*, Chapter 49, pp. 2119–2171. Elsevier.

Rosenthal, S. S. and W. C. Strange (2020). How Close Is Close? The Spatial Reach of Agglomeration Economies. *Journal of Economic Perspectives 34*(3), 27–49.

Sweeney, S. H. and S. Arabadjis (2022). Spatial Point Patterns. In S. J. Rey and R. Franklin (Eds.), *Handbook of Spatial Analysis in the Social Sciences*, pp. 262–276. Edward Elgar Publishing.

Sweeney, S. H. and E. J. Feser (1998). Plant Size and Clustering of Manufacturing Activity. *Geographical Analysis 30*(1), 45–64.

Sweeney, S. H. and E. J. Feser (2004). Business Location and Spatial Externalities: Tying Concepts to Measures. In M. F. Goodchild and D. G. Janelle (Eds.), *Spatially Integrated Social Science*. Oxford University Press.

Sweeney, S. H. and K. J. Konty (2005). Robust point-pattern inference from spatially censored data. *Environment and Planning. A 37*(1), 141–159.

Theil, H. (1967). *Economics and Information Theory*. Rand McNally & Company.

Tidu, A., F. Guy, and S. Usai (2024). Measuring Spatial Dispersion: An Experimental Test on the *M*-Index. *Geographical Analysis 56*(2), 384–403.