Version: hal-05512154v1
February 16, 2026

Computation of Large Spatial Datasets with the M function

Eric Marcon[1] Florence Puech[2]

Abstract
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 in several scientific fields to detect spatial concentration, dispersion or independence of entities at any distance and without any bias. 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 propose a methodology that specifies the processing of large spatialized datasets with the *M* function using R software. The computational performance of *M* was conducted using two methods: (i) a precise evaluation of the computational time and memory requirements for geo-referenced data was conducted using the *dbmss* package in R via performance tests, and (ii) based on Tidu et al. (2024), we considered an approximation of the geographical positions of the entities. The deterioration extent of the *M* results was estimated and discussed as the gains it provides in computation time. We provided evidence that the individual location approximation generated information loss at substantially small distances, implying a trade-off between the smallest distance at which spatial interactions could be detected and computing performance. 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

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). Thus, empirical studies at highly detailed geographical levels have been proposed in recent years for large datasets. 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 increasing attention. Such an approach has the advantage of preserving the exact positions of the entities analyzed. Any distance-based method (by considering space as continuous) circumvents statistical bias associated with the Modifiable Areal Unit Problem (MAUP: Openshaw and Taylor, 1979; Arbia, 1989) owing to discretizing space into separate units. 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 exact sciences (Cressie, 1993; Lentz et al., 2011; Dray et al., 2021).

Tidu et al. (2024) highlighted 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. 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 (Marcon and Puech, 2017, reviewed the advantages and limitations of a dozen existing distance-based measures).

Tidu et al. (2024) proposed decreasing M calculation times by introducing a voluntary positioning error for the analyzed entities. For example, 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. This approach is similar to that of Scholl and Brenner (2015). They proposed the K_d function (Duranton and Overman, 2005), approximating the distances between pairs of entities by grouping them into classes. The method of Scholl and Brenner (2015), implemented in the *dbmss* package (Marcon et al., 2015) for R (R Core Team, 2025), provides a considerable gain in computational performance with minor loss of accuracy. 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 choose the appropriate method to characterize the spatial structure of substantially large datasets. First, we show the advantages of estimating M on datasets with an order of magnitude of 100,000 points or less. The computation times become excessive beyond that, using a personal computer. We studied the effect of the geographical approximation of the locations of the analyzed entities. This methodological study, based on a deliberately limited number of entity locations, quantified the extent of the deterioration in information that this approach created. These performance tests precisely provided the computational advantages and limitations of the M function as a function of the of the dataset size.

The layout of the article is mentioned here. The first two sections present the M function and the necessary data generated for the tests. Large point sets (in the order of several tens of thousands of points) that are either completely random or geographically concentrated are considered. The third section details the use of the *dbmss* package to calculate M and its confidence interval from a table that gives the position and characteristics of the points or a matrix of distances between them. The fourth section measures the performance of *dbmss* as a function of the size of the set of points, in terms of computing time and memory requirements. The fifth section tests the spatial approximation which involves grouping them at the center of the cells of a grid, following the approach of Tidu et al. (2024), which positions them at the center of the administrative units of their location. In the last section, we conclude and discuss the advantages and the limits of an approximation of the locations on the accuracy of the results and the computing time.

1. The *M* function

1.1 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. Similar to any 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 is preferred for those that use the surface area as a benchmark, as the well-known Ripley's K function (Ripley, 1976, 1977). The M function is also defined as a *cumulative* distance-based method because the local environment is appraised within 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. Marcon and Puech (2010) proved that this function satisfies all the requirements of Duranton and Overman (2005) for the evaluation of the spatial distribution of industries. Since its introduction, various studies have described the spatial locations of industries by using M; for example, Jensen and Michel (2011) studied the location of shops at an urban level, Coll-Martínez et al. (2019) analyzed creative industries at a metropolitan level. This methodology has also been rapidly applied in other domains including biology (Fernandez-Gonzalez et al.,

2005), geography (Deurloo and De Vos, 2008), ecology (Marcon et al., 2012), and seismology (Nissi et al., 2013). The M function is now included in general textbooks of spatial statistics such as Arbia et al. (2021) but it is less popular than the K_d function of Duranton and Overman (2005); see Chain et al. (2019).

1.2 Definition

The M function is based on the point process theory (Møller and Waagepetersen, 2004; Baddeley et al., 2016). This distance-based method 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 on the territory (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 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 of radius r. 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, 2025) 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. Data simulation

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.

2.1 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.

2.2 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,

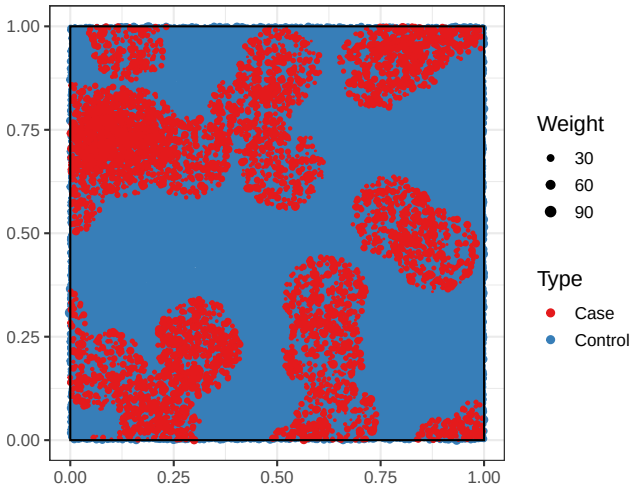

Figure 1. Random drawing of a set of points where the Cases (red) are aggregated and the Controls (blue) are distributed completely randomly. The size of the points is proportional to their weight.

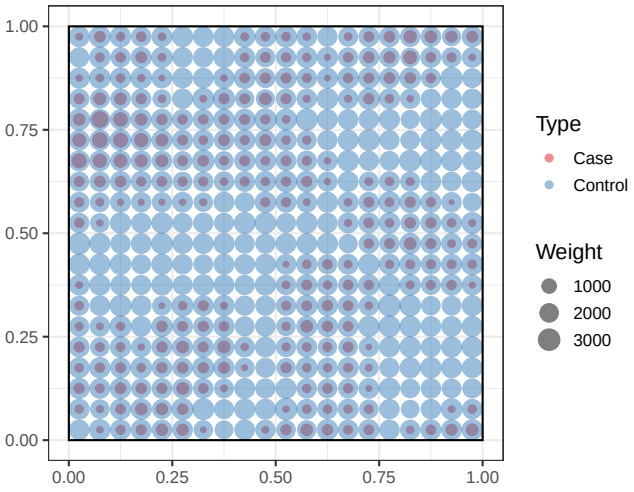

Figure 2. Repositioning of points in an arbitrary grid. The absence of Cases in a cell is easily detected (single-color blue dot), as is the strong presence of Cases in a cell (two-color dot, but predominantly red).

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.*'

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.

3. Computing *M* using the *dbmss* package

3.1 Necessary data

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

3.2 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.

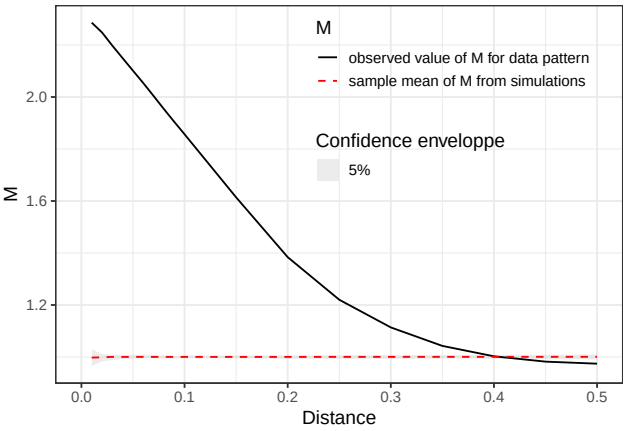

Figure 3. Value of M as a function of the distance from the reference point. The 95% confidence envelope, obtained from 100 simulations, appears in gray and is centered on the value of one.

3.3 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 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.

4. 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.

4.1 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., 3.05 seconds for 100,000 points), knowing the time t_0 for n_0 points, and p is the power relation (here: 1.5).

4.2 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.

5. Effects of approximating the position of points

5.1 Two main effects

Unambiguously, approximating the position of the points generates a loss of information: in each grid cell, the distance between all the points is set to zero, and the distance between two points in different cells is approximated using the distance between the centroids of the two cells. The first consequence is that **no spatial structure in each cell can be detected because the information is completely lost.** The merits of any distance-based methods are detecting the exact geographic scale(s) where any patterns of aggregation, dispersion or independence of points exist. Grouping points at the centroid of cells erases any spatial structure under the size grid. The same limit under the size grid is faced for distributions with multiple patterns of points (that is a coexistence of attraction and dispersion of points depending on the distance considered). Tidu et al. (2024) gathered an average of 320 points in every 377 municipalities of Sardinia. However, the local structure of these points was overlooked. Moreover, we suspect **bias in M estimation based on a scale of the order of the magnitude of the size of the cells**, which should decrease with distance, when the relative size of the cells becomes negligible.

5.2 Test setting

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.

20 sets of completely random and of aggregated points (100,000 points with 5% of Cases) 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.

5.3 Results

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, some artifactual aggregation

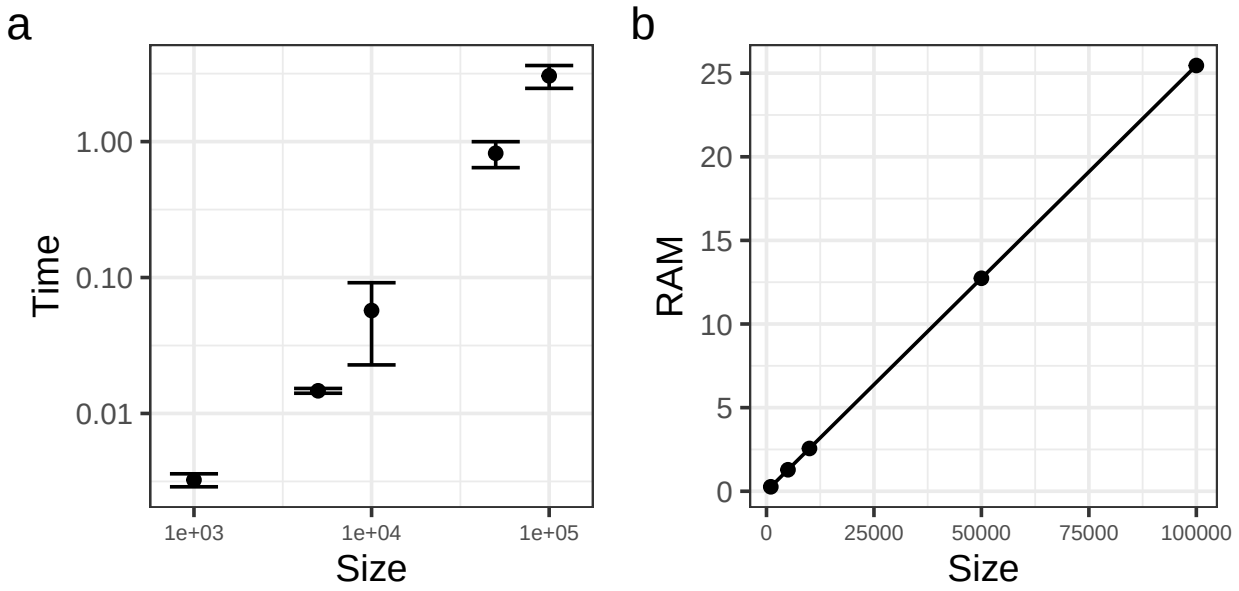

Figure 4. (a) Calculation time (seconds) and (b) memory required (MB) to estimate M as a function of the size of the set of points. The measures were repeated 10 times. The bars represent the ± 1 standard deviation interval.

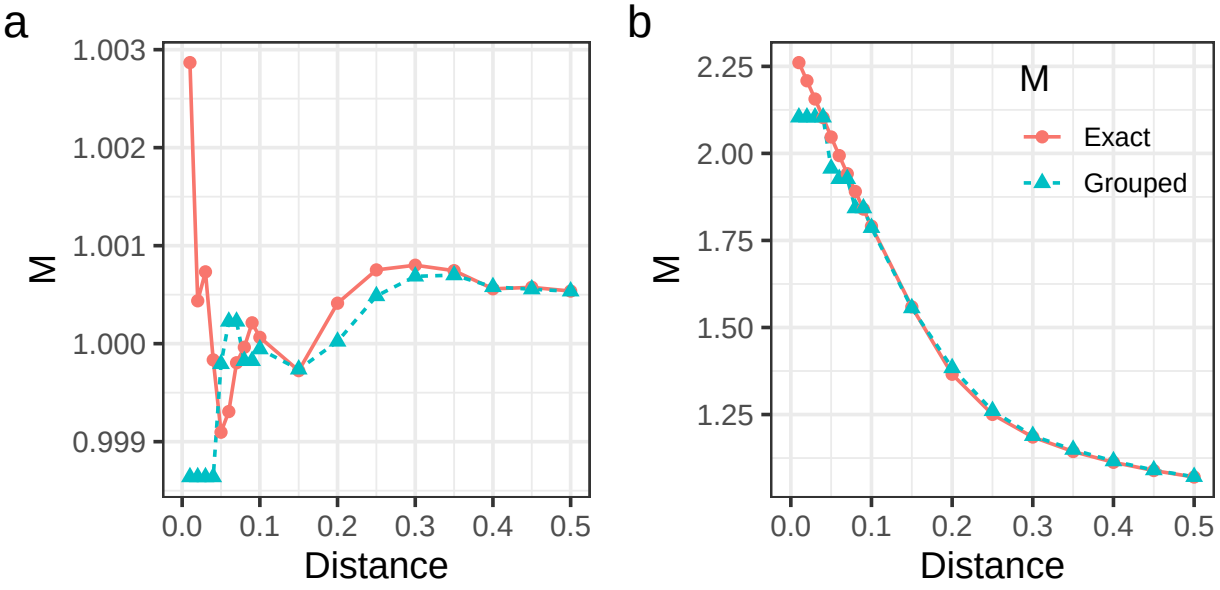

Figure 5. Average estimate of M from the exact position of the points compared with the values obtained by grouping the points for CSR (a) and Matérn (b) point patterns.

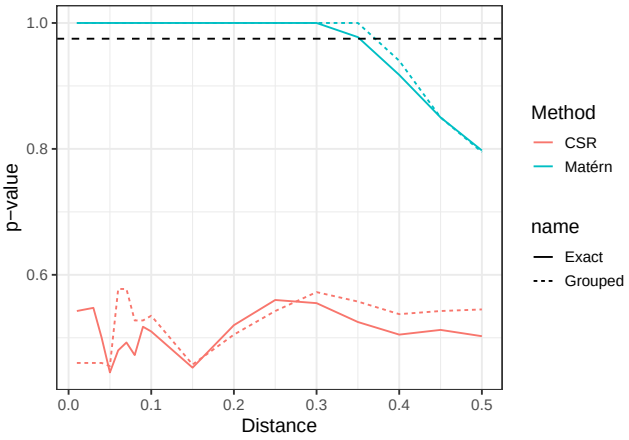

Figure 6. p-values to reject the null hypothesis (H0) of independence of the point locations tested by the M function. The colors of the curves represent CSR or aggregated point patterns. Solid lines provide the exact p-values of M, and dotted lines provide the values estimated on grouped point patterns.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

is generated at a small scale. In contrast, the local aggregation of the Matérn pattern is underestimated. Figure 5 shows that at substantially short distances the grouped M plot is below the exact M plot obtained using the actual position of points.

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.

We chose the mentioned test. 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 point patterns simulated 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.

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.

6. Discussion and conclusions

The computation burden of estimating M on large datasets might be a problem. The calculation time for M is < 4 seconds for a set of 100,000 points on a modern computer[1] and requires 25 MB of RAM. Therefore, calculating a confidence interval from 1,000 simulations requires less than 67 minutes. For a set of five million points, the power law predicts around 15 minutes of computing time. Accordingly, 1,000 simulations would take around 10 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. The size of the grid (or the administrative scale of the aggregation of points in Tidu et al., 2024) should be chosen based on the scale of the interactions under study: they cannot be characterized correctly at distances below it.

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. This situation needs to be analyzed attentively.

Moreover, Arbia et al. (2017) investigated applying distance-based methods to spatial datasets that include positional errors. They proposed a first evaluation of the consequences of an *'unintentional positional error'* owing to the uncertainty of the location

[1]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.

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). Arbia et al. showed on a real case (Italian manufacturing firms) that the error measurement was less severe than expected. Their explanation rests on the definition M, a relative measure: a compensation effect of positional errors is suspected between the local and the global ratios.

An approximation of the spatial locations might be considered to save computation time, given the strong agreement observed between M values on exact and approximated data above the grid size, but the information related to interactions at smaller scales is lost. As we previously noticed, it is challenging to choose the optimal size of the grid, requiring a certain degree of caution in the use of the approximated locations. For example, careful use of location approximation is recommended if short-distance interactions are suspected, such as information externalities or contagion phenomena. Choosing the approximation scale is a trade-off between accuracy, i.e., a small distance threshold above which results are accurate, and speed with a coarse grid.

Appendix

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

Acknowledgments

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. The authors thank participants of the SEW 2025 and especially Nardelli Vincenzo for helpful advice.

References

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

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.

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.

Chain, C. P., A. C. D. Santos, L. G. D. Castro, and J. W. D. Prado (2019). Bibliometric analysis of the quantitative methods applied to the measurement of industrial clusters. *Journal of Economic Surveys 33*(1), 60–84.

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.

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

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.

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.

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.

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.

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. 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.

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

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

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.

R Core Team (2025). *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.

Scholl, T. and T. Brenner (2015). Optimizing distance-based methods for large data sets. *Journal of Geographical Systems 17*(4), 333–351.

Sweeney, S. 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.

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