On the Computation of Large Spatial Datasets With the M function

Eric Marcon
Florence Puech

2025 May 23

Introduction

Increasing access to large individual and spatial datasets and greater computing power have encouraged the development of statistical analysis tools for processing such data in the best possible way (Baddeley et al., 2016). Empirical studies at very fine geographical levels have thus been proposed in recent years for large datasets. Particular attention has been paid to detect the spatial structures (attraction, repulsion, independence) of individual spatialised data using analyses that are no longer based on zoned data but on geolocalised data. This type of approach has the advantage of preserving the exact positions of the entities analysed. It has been proven that any distance-based method (by considering space as continuous) circumvents statistical bias associated with to the Modifiable Areal Unit Problem – MAUP (Arbia, 1989; Openshaw & Taylor, 1979) due to discretising space into separate units. Various studies have shown how important it is to use this type of methodology in many fields, including geography (Deurloo & De Vos, 2008; Kukuliač & Horák, 2017; Sweeney & Feser, 1998), economics (Arbia, 1989; Marcon & Puech, 2003), ecology (Cressie, 1993; Lentz et al., 2011), biology (Dray et al., 2021), and so on.

In a recent article, Tidu et al. (2024) highlight the interest of a particular statistical measure, the *M* function proposed by Marcon & Puech (2010). This measure, which we will refer to as *M* in the remainder of the article, makes it possible to highlight spatial

structures within a spatialised distribution (attraction, repulsion, independence) from a study based on the distances separating the entities analysed. However, while this measure preserves all the richness of individual geolocated data, it requires a longer calculation time than other distance-based measures, since it is both a cumulative and relative measure (see Marcon & Puech, 2017 for a literature review on the advantages and limitations of a dozen existing distance-based measures). Tidu et al. (2024) propose to limit *M* calculation times by introducing a voluntary positioning error for the entities analysed. For example, in their study, industrial establishments in Sardinia (Italy) are no longer located at their exact postal address but at the centroid of their municipality. This repositioning reduces calculation times, as the number of possible distances between establishments is in fact limited to the distances separating the centroids of the municipalities. This approach is similar to that of Scholl & Brenner (2015) who proposed, for the K_d function (Duranton & Overman, 2005) which characterises spatial structures using another method, to approximate the distances between pairs of entities by grouping them into classes. The method of Scholl & Brenner (2015), implemented in the *dbmss* package (Marcon et al., 2015) for R (R Core Team, 2024) provides a considerable gain in computational performance with little loss of accuracy. On the other hand, the information loss due to the approximation of the location of objects should imply a loss of accuracy in the estimation of their interactions at the same scale, that must be assessed.

In our paper, we propose to test the effectiveness of Tidu et al. (2024)'s method and help the researchers to choose the appropriate method to characterise the spatial structure of quite large datasets. First, we show the advantages of using the *dbmss* package to estimate the *M* function on datasets with an order of magnitude of 100,000 points or less, and we show that the computation times become excessive beyond that, on a personal computer. We then study the effect of the geographical approximation of the locations of the entities analysed. This methodological work, based on a deliberately limited number of entity locations, enables us to quantify the extent of the deterioration in information that this approach creates. These performance tests provide a precise answer to the computational advantages and limitations of the *M* function as a function of the size of the datasets.

The layout of the article is as follows. The two first sections present the *M* function and the necessary data generated for the tests. Large point sets (of the order of several tens of thousands of points) that are either completely random or geographically concentrated are drawn. The third section details the use of the *dbmss* package to

calculate the *M* function and its confidence interval from a table giving 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 approximation which consists of grouping them together at the centre of the cells of a grid, following the approach of Tidu et al. (2024) which positions them at the centre of the administrative units in which they are located. In the last section, we conclude and discuss the advantages and the limits of an approximation of the locations on the results as well as on the computing time.

1 The *M* function

1.1 Main idea

Marcon & Puech (2010) introduced the *M* function that evaluates the dependence between geolocalized points without relying on a specific zoning of space. As any distance-based methods, the calculation of *M* is based on distances that separate entities under study (establishment, shops…). The idea of the *M* function is simple: it compares two proportions of neighbours of interest, a local one to a global one. The local one is defined as the proportion of neighbours of interest within a distance *r*. The global one is the same proportion but defines on the whole territory. This comparison of ratios allows the detection of:

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

This comparison of proportions of neighbours defines *M* as a *relative* distance-based measure in a strict sense (Marcon & Puech, 2017). The term *topographic* distance-based measures is preferred for those which used the surface area as a benchmark, as the well-known Ripley's *K* function (Ripley, 1976, 1977). Lastly, the *M* function is a *cumulative* distance-based method because the local environment is defined within a distance *r* and not at a distance *r*.

The possibility to detect exactly at which distance(s) the spatial concentration or dispersion appears coupled with the interpretation of the results open the way to describe very precisely the distribution of entities under study. Moreover, an easy-computation of *M* is possible thanks to the *dbmss* R package (Marcon et al., 2015). The *M* function was at first introduced in the field of economics and Marcon & Puech (2010) proved that *M* satisfies all of the requirements of Duranton & Overman (2005) for the evaluation of spatial distribution of industries. Since its introduction, various studies have described the spatial locations of industries by using *M* ; for example, Jensen & Michel (2011) study the location of shops at a urban level, Coll-Martínez et al. (2019) analyse the one of the creative industries at a metropolitan level etc. This methodology has also rapidly been applied in other sciences including biology (Fernandez-Gonzalez et al., 2005), geography (Deurloo & De Vos, 2008), ecology (Marcon et al., 2012) or seismology (Nissi et al., 2013). The *M* function is now included in general textbooks of spatial statistics as in Arbia et al. (2021), but it is less popular than K_d's function of Duranton & Overman (2005; see Chain et al., 2019).

1.2 Definition

The M function is based on the point process theory (Baddeley et al., 2016; Møller & Waagepetersen, 2004). This distance-based method has been developed to analyse interactions among entities in a context of heteregeneous space. It means that within this statistical framework, we consider that any entity analysed has not the same probability to locate everywhere on the territory (*first-order intensity property of the point pattern*). Then after controlling for space heterogeneity, we are able to identify interactions and thus detect spatial concentration or dispersion (*second-order intensity property of the point pattern*). Controlling for space heterogeneity is a consistent assumption for studying agglomeration of industries (see discussion of Duranton & Overman (2005) on that subject).

The definition of the *M* function it 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 the present article, we only consider the intra-type version of *M*: we study the spatial structure of neighbouring points of the same type (called *points of interest*) as the point at the centre of the disks of radius r. In a mathematical terms, let us denote:

- x_i^s, the geolocalised position of point i of the reference type s, at the centre of the disk (the point at which the neighbourhood is to be analyzed),
- x_j^s, the geolocalised position of a neighbour of interest j of the same type as point i,
- x_j, the geolocalised of a neighbour j of i, whatever its type,
- $w(.)$, the weight of a given neighbour. 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, whatever their type,
- $\mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right)$, the indicator function equal to 1 if x_j is in the neighbourhood of x_i^s, e.g., the distance between x_i^s and x_j is at most equal to r, 0 otherwise.
- $\mathbf{1}\left(\left\|x_i^s - x_j^s\right\| \leq r\right)$, the indicator function equal to 1 if the distance between x_i^s and x_j^s is at most equal to r, 0 otherwise.

The intra-type M function is defined as:

$$\hat{M}(r) = \sum_i \frac{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i^s - x_j^s\right\| \leq r\right) w\left(x_j^s\right)}{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i^s - x_j\right\| \leq r\right) w\left(x_j\right)} / \sum_i \frac{W_s - w\left(x_i^s\right)}{W - w\left(x_i^s\right)} \qquad (1.1)$$

A certain number of remarks has to be done. The first one is that the benchmark value of *M* is equal to 1, whatever the distance considered. It means that for any given radius *r*:

- if the estimated M result is above 1, 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 1, the local value of the ratio is lesser 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 thanks to Monte Carlo simulations following Marcon & Puech (2010). A risk level has to be 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, 2024) can be used to compute the M function. The Euclidean distance is generally preferred for the calculation of *M* but the *dbmss* package also proposed to used network-distances.

2 Data simulation

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

2.1 Drawing the points

A set of points is drawn by a Poisson process (whose expectation of the number of points is 5,000) in a square window of side 1. Each point is assigned a qualitative mark: 'Case' or 'Control'. 95% of points are 'Controls'. 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 points is completely random (*complete spatial randomness*: CSR), i.e. there is no simulation of attraction or dispersion of points which could generate spatial concentrations of points (aggregates) or, on the contrary, spatial regularities (dispersions). Sets of aggregated points can be drawn in a Matérn (1960) process.

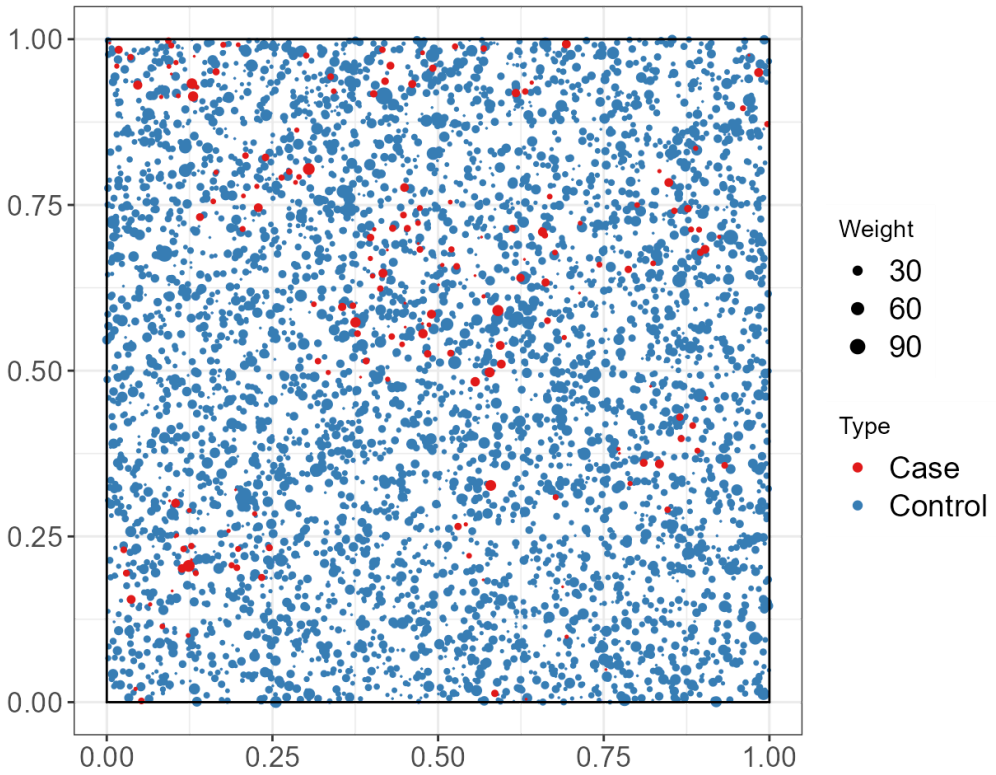

Figure 2.1: Random draw of a set of points where the Cases (in red) are aggregated and the Controls (in blue) are distributed completely randomly. The size of the points is proportional to their weight.

The Cases are shown in figure 2.1: the aggregates are clearly visible. The controls are distributed completely randomly.

2.2 Gridding the space

Let's consider the simulation of the Cases obtained by the Matérn process and cut the window into a grid. It simulates the usual approximation of the position of the points of an administrative unit to the position of its centre.

The approximated position of points is shown on the map in figure 2.2. Each cell now contains only one point of each type, whose weight is the sum of the weights of the individual points.

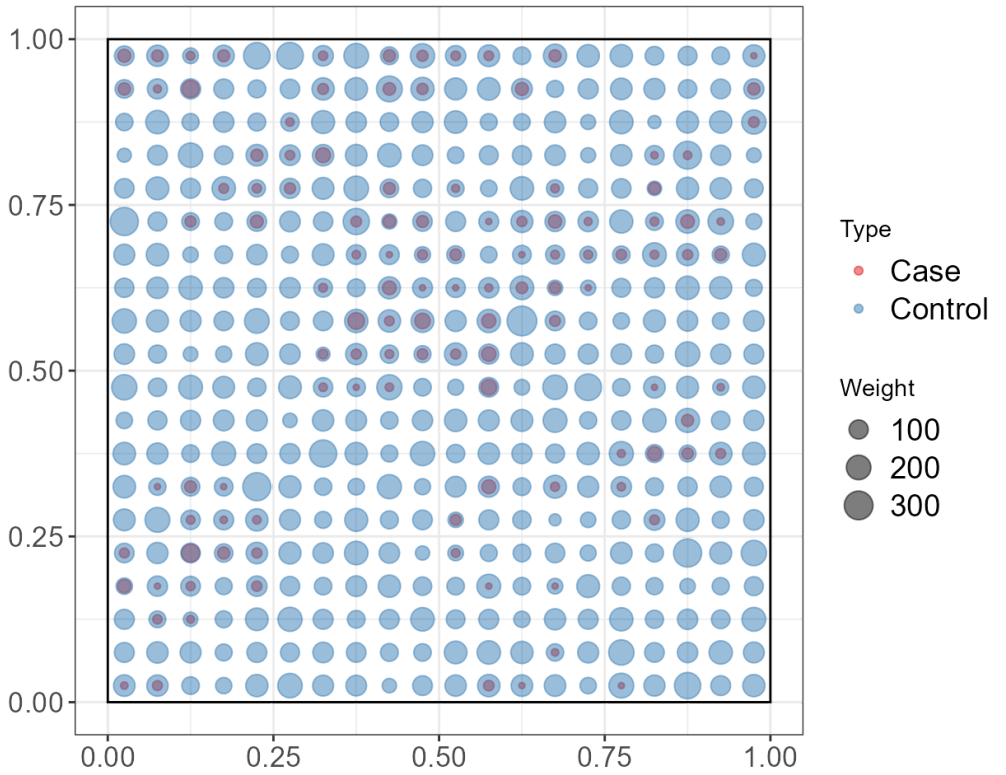

Figure 2.2: Repositioning of points in an arbitrary grid. The absence of Cases in a cell is easily detected (single-colour blue dot), as is the strong presence of Cases in a cell (two-colour dot, but predominantly red).

The values of the *M* function can now be calculated from the original point set or its approximation after recentring.

3 Computing *M* with the *dbmss* package

3.1 Necessary data

In the *dbmss* package, the function is applied to a set of points or a distance matrix. The set of points in figure 2.1 is used. The distance matrix between all the pairs of its points is calculated to form the data on which the performance tests will be carried out.

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 1, as this function relates the proportion of Cases up to a distance r to that observed over the entire window. The aggregation of Cases will be highlighted by values of *M* greater than 1 (the relative presence of Cases is greater locally than over the whole window) and the dispersion of Cases by values less than 1. We observe (figure 3.1) that *M* detects an agglomeration of Cases, which is in line with the simulation of this type of point (the controls having a completely random location on the window). The advantage of a function based on distances is clearly visible: it allows us to detect exactly at which distance(s) 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 & Overman, 2005) under the null hypothesis of random point location. It allows parallezing the necessary simulations. The result is shown in figure 3.1.

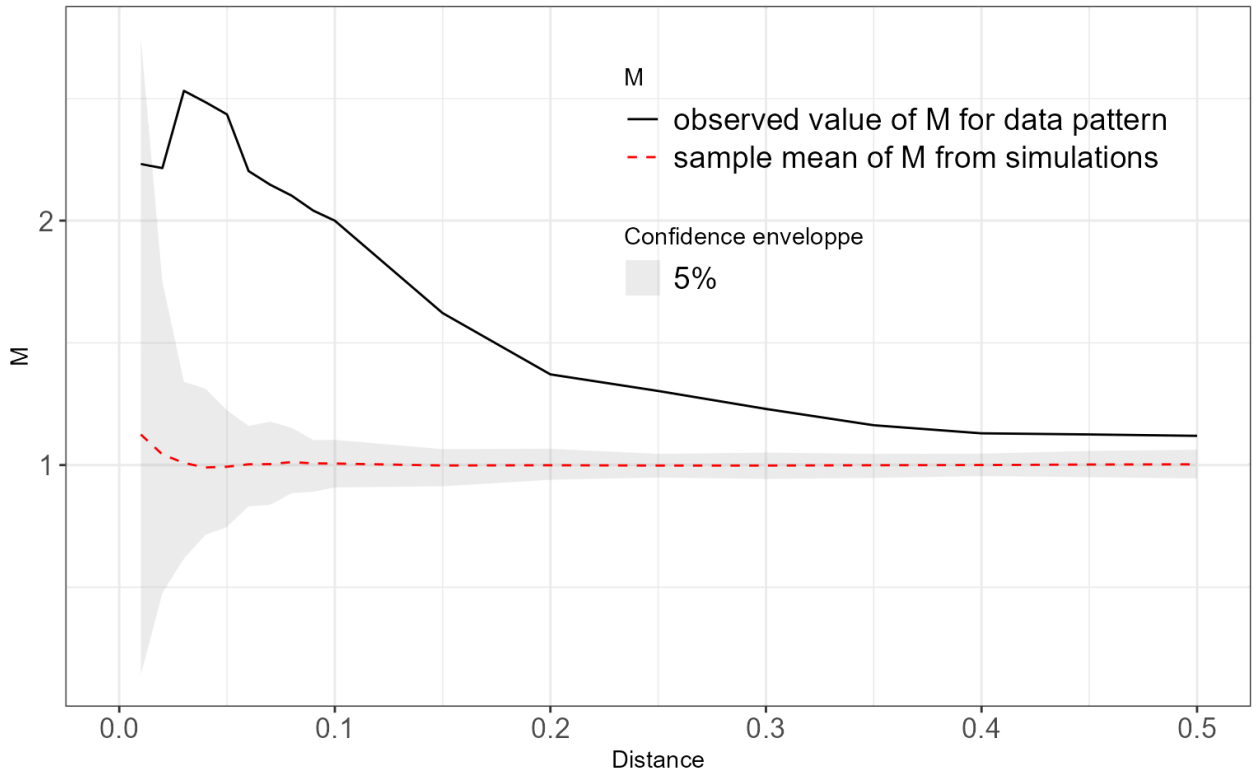

Figure 3.1: Value of M as a function of distance from the reference point. The confidence interval, simulated at 95%, appears in grey and is centred on the value 1.

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 whatever the form of the data used here (point set or distance matrix).

4 Computational performance

The use of the *M* function to characterise the spatial structure of large sets of points may be limited by the computing time or memory required.

4.1 Computing time

Calculating the distances between all pairs of points is necessary to estimate *M*. The calculation time is therefore expected to increase as the square of the number of points. The calculation time required for the exact calculation is evaluated for a range of numbers of points (figure 4.1a).

The calculation time is related to the size of the set of points by a power law. It increases less quickly than the square of the number of points. It can be estimated very precisely ($R^2 = 0.98$) by the relation $t = t_0(n/n_o)^p$ where t is the estimated time for n points (e.g.: 5.42 seconds for 100,000 points) knowing the time t_0 for n_0 points and p is the power relation (here: 1.7).

Using a distance matrix may seem an efficient way of saving computation time, but in reality calculating distances is extremely fast and the whole process from a matrix is ultimately more time-consuming. The median execution time is equal to 14 milliseconds for estimating the *M* function from a set of 5,000 points or 23 milliseconds for the corresponding distance matrix.

4.2 Memory

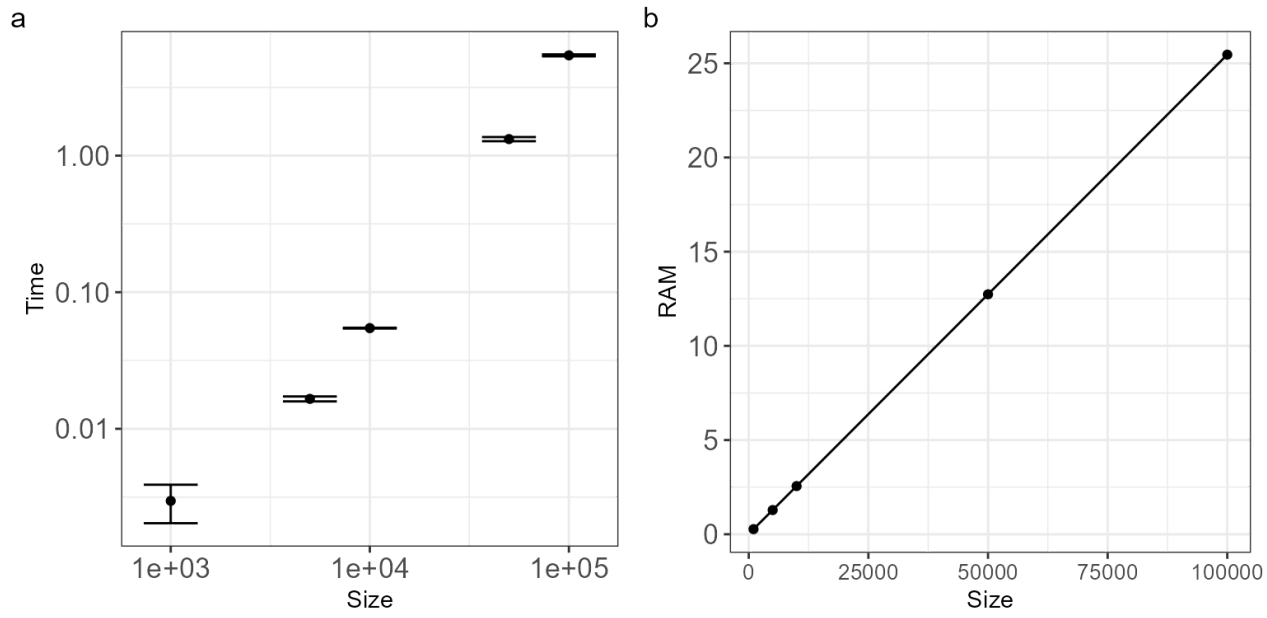

Figure 4.1: Calculation time (a) in seconds and memory required (b) in MB to estimate the M function as a function of the size of the set of points. The bars represent the ± 1 standard deviation interval.

The memory used is evaluated for the same data sizes (figure 4.1b). **The memory required increases linearly with the number of points and is never critical for point set sizes that can be processed in reasonable times.** This highlights Tidu et al. (2024)'s conclusion about the power and computation time required when using *M* on large datasets. The memory used by `Dtable` objects to calculate M from a distance matrix is much greater: it is that of a numerical matrix, of the order of 8 bytes times the number of points squared, i.e. 800 MB for 10,000 points only. As the calculation time is not reduced by this approach, its use should be reserved for non-Euclidean distances.

5 Effects of approximating the position of points

Clearly, approximating the position of the points results in 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 by the distance between the centroids of the two cells. We therefore expect a severe error in the estimation of *M* on a small scale (of the order of magnitude of the size of the cells) and an error that decreases with distance, when the relative size of the cells decreases. The effect of the location approximation is first tested on a set of aggregated points, similar to the real Tidu et al. (2024) data. Secondly, the case of an unstructured set of points is considered.

5.1 Case of an aggregated distribution (Matérn)

100 sets of aggregated points (5,000 points with 5% of Cases) are simulated. To evaluate the effect of the position approximation, the exact calculation and the calculation on the grid points are performed on each set of points.

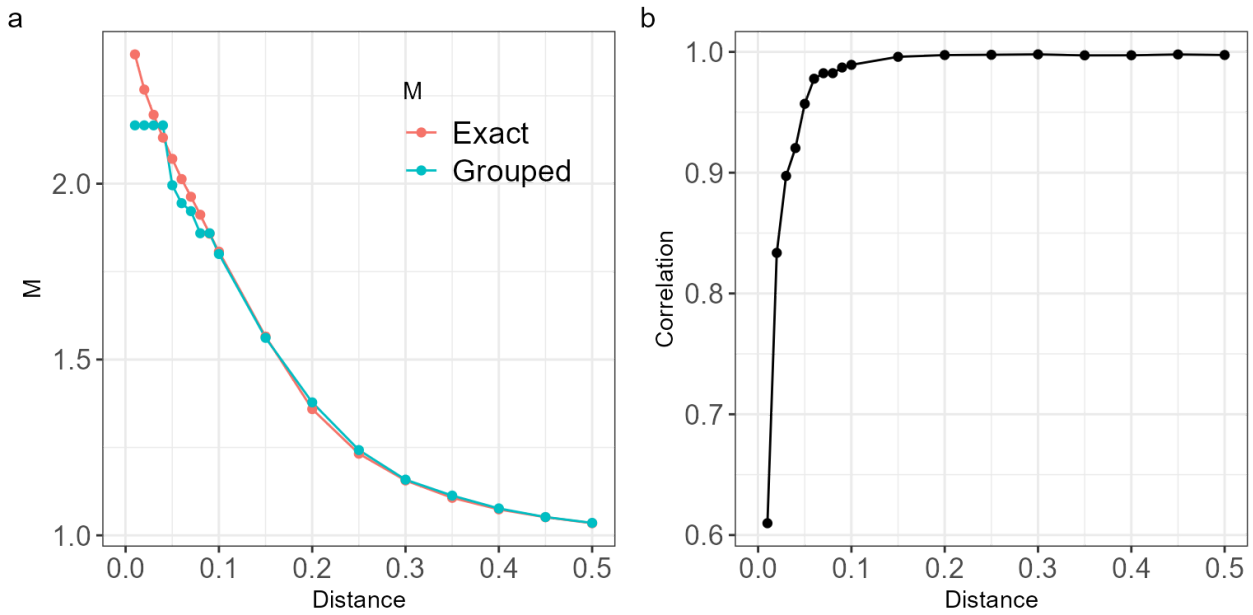

Figure 5.1: Average estimate of the M function from the exact position of the points compared with the values obtained by grouping the points (a) and correlation between them (b). Cases form aggregates of radius 0.1.

The mean values of the estimates of *M* are presented in figure 5.1a. The size of the grid cells is equal to 0.05. All neighbours at distances less than this threshold are placed at zero distance: the estimate of the function is constant up to this threshold and small-

scale aggregation is underestimated. The correlation between the *M* values estimated by each method is calculated at each distance (figure 5.1b).

The correlation is very close to 1, and the estimated values very similar, as soon as the distance taken into account exceeds the grid cell: the approximation is not a problem if the interactions between the points are studied beyond this distance. The information on interactions at short distances, i.e. within each grid cell, is lost, or, more precisely, approximated by its value at the grid scale.

5.2 Case of a completely random distribution (CSR)

The same simulations are run with a completely random set of points. The exact calculation and the calculation on the grid points are carried out on each set of points.

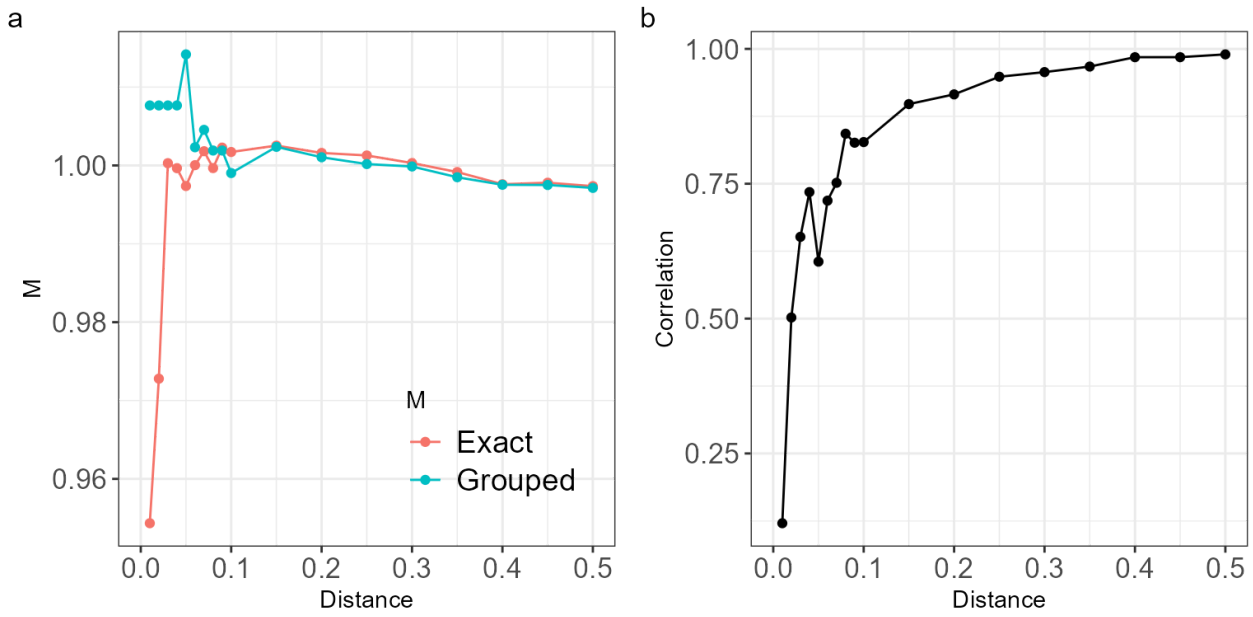

Figure 5.2: Average estimate of the M function from the exact position of the points compared with the values obtained by grouping the points (a) and correlation between them (b). Both Cases and Controls are drawn in a Poisson process.

The average values are shown in figure 5.2a. The mean value of *M* is equal to 1 at all distances by construction: Cases and Controls are distributed completely randomly. The approximations are relatively small in value (a few percent) but artefactual aggregation is

generated at small scale. As the real value of *M* varies little around 1, the correlations are much weaker (figure 5.2b) in the absence of spatial structure than in the aggregated case.

6 Discussion and conclusion

Our results are in line with Tidu et al. (2024)'s article, which mentions strong correlations between *M* values computed from exact and approximated Italian company location data. Since the spatial structure of their data is probably an intermediate case between the two cases dealt with in our article (aggregated and random theoretical distributions), the results provided by our two contributions are complementary.

The computation burden of estimating the *M* function on large datasets may be an issue. The calculation time for *M* is around 5 seconds for a set of 100,000 points on a laptop (Intel i7-1360P 2.20 GHz processor), and requires 25 MB of RAM. Calculating a confidence interval from 1,000 simulations therefore takes less than two hours. For a set of five million points, the power law predicts around an hour of computing time. 1,000 simulations would then take more than one month. Thanks to parallelization, a calculation server would drastically increase performance, but at the cost of a 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 10^5 points**: a few hours are enough to calculate confidence intervals. Since parallelizing the simulations is offered with no effort by the *dbmss* package, this time can be reduced by a factor depending on the available hardware, say 2 to 6 with modern multicore CPU's.

Beyond that, approximating the location reduces the size of the set of points to the number of locations selected. Again, it may be up to 10^5 locations to keep computing time acceptable, whatever the size of the original dataset. The choice of the size of the grid (or the administrative scale of the aggregation of points in Tidu et al., 2024) must be done according to the scale of the interactions under study: they can not be characterized correctly at distances below it.

So, **approximation on location can be considered to save computation time, given the strong correlation observed between the values of *M* on exact and approximated data, but the scale of the grid must be fine enough to be informational at small distances**. Choosing the approximation scale is a tradeoff 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
(https://ericmarcon.github.io/MLargeDataSets/Appendix.pdf)

Acknowledgements

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*. Kluwer.
Arbia, G., Espa, G., & Giuliani, D. (2021). *Spatial microeconometrics*. Routledge, Taylor & Francis Group.
Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial point patterns: Methodology and applications with R*. CRC Press.

Chain, C. P., Santos, A. C. D., Castro, L. G. D., & Prado, J. W. D. (2019). Bibliometric analysis of the quantitative methods applied to the measurement of industrial clusters. *Journal of Economic Surveys*, *33*(1), 60–84. https://doi.org/10.1111/joes.12267 (https://doi.org/10.1111/joes.12267)

Coll-Martínez, E., Moreno-Monroy, A.-I., & Arauzo-Carod, J.-M. (2019). Agglomeration of creative industries: An intra-metropolitan analysis for Barcelona. *Papers in Regional Science*, *98*(1), 409–432. https://doi.org/10.1111/pirs.12330 (https://doi.org/10.1111/pirs.12330)

Cressie, N. A. (1993). *Statistics for spatial data*. John Wiley & Sons.

Deurloo, M. C., & De Vos, S. (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. https://doi.org/10.1111/j.1467-9663.2008.00465.x (https://doi.org/10.1111/j.1467-9663.2008.00465.x)

Dray, N., Mancini, L., Binshtok, U., Cheysson, F., Supatto, W., Mahou, P., Bedu, S., Ortica, S., Than-Trong, E., Krecsmarik, M., Herbert, S., Masson, J.-B., Tinevez, J.-Y., Lang, G., Beaurepaire, E., Sprinzak, D., & Bally-Cuif, L. (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. https://doi.org/10.1016/j.stem.2021.03.014 (https://doi.org/10.1016/j.stem.2021.03.014)

Duranton, G., & Overman, H. G. (2005). Testing for localisation using micro-geographic data. *Review of Economic Studies*, *72*(4), 1077–1106. https://doi.org/10.1111/0034-6527.00362 (https://doi.org/10.1111/0034-6527.00362)

Fernandez-Gonzalez, R., Barcellos-Hoff, M. H., & Ortiz-de-Solorzano, C. (2005). A tool for the quantitative spatial analysis of complex cellular systems. *IEEE Transactions on Image Processing*, *14*(9), 1300–1313. https://doi.org/10.1109/tip.2005.852466 (https://doi.org/10.1109/tip.2005.852466)

Jensen, P., & Michel, J. (2011). Measuring spatial dispersion: Exact results on the variance of random spatial distributions. *The Annals of Regional Science*, *47*(1), 81–110. https://doi.org/10.1007/s00168-009-0342-3 (https://doi.org/10.1007/s00168-009-0342-3)

Kukuliač, P., & Horák, J. (2017). W function: A new distance-based measure of spatial distribution of economic activities. *Geographical Analysis*, *49*(2), 199–214. https://doi.org/10.1111/gean.12120 (https://doi.org/10.1111/gean.12120)

Lentz, J. A., Blackburn, J. K., & Curtis, A. J. (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. https://doi.org/10.1371/journal.pone.0021830 (https://doi.org/10.1371/journal.pone.0021830)

Marcon, E., & Puech, F. (2003). Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, *3*(4), 409–428. https://doi.org/10.1093/jeg/lbg016 (https://doi.org/10.1093/jeg/lbg016)

Marcon, E., & Puech, F. (2010). Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, *10*(5), 745–762. https://doi.org/10.1093/jeg/lbp056 (https://doi.org/10.1093/jeg/lbp056)

Marcon, E., & Puech, F. (2017). A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, *62*, 56–67. https://doi.org/10.1016/j.regsciurbeco.2016.10.004 (https://doi.org/10.1016/j.regsciurbeco.2016.10.004)

Marcon, E., Puech, F., & Traissac, S. (2012). Characterizing the relative spatial structure of point patterns. *International Journal of Ecology*, *2012*(Article ID 619281), 11. https://doi.org/10.1155/2012/619281 (https://doi.org/10.1155/2012/619281)

Marcon, E., Traissac, S., Puech, F., & Lang, G. (2015). Tools to characterize point patterns: dbmss for R. *Journal of Statistical Software*, *67*(3), 1–15. https://doi.org/10.18637/jss.v067.c03 (https://doi.org/10.18637/jss.v067.c03)

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

Møller, J., & Waagepetersen, R. P. (2004). Statistical inference and simulation for spatial point processes. In *Monographs on statistics and applies probabilities* (Vol. 100). Chapman and Hall.

Nissi, E., Sarra, A., Palermi, S., & Luca, G. (2013). The application of m-function analysis to the geographical distribution of earthquake sequence. In A. Giusti, G. Ritter, & M. Vichi (Eds.), *Classification and data mining* (pp. 271–278). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28894-4_32 (https://doi.org/10.1007/978-3-642-28894-4_32)

Openshaw, S., & Taylor, P. J. (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). Pion.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ripley, B. D. (1976). The foundations of stochastic geometry. *Annals of Probability*, *4*(6), 995–998. https://www.jstor.org/stable/2242958 (https://www.jstor.org/stable/2242958)

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *39*(2), 172–212. https://www.jstor.org/stable/2984796 (https://www.jstor.org/stable/2984796)

Scholl, T., & Brenner, T. (2015). Optimizing distance-based methods for large data sets. *Journal of Geographical Systems*, *17*(4), 333–351. https://doi.org/10.1007/s10109-015-0219-1 (https://doi.org/10.1007/s10109-015-0219-1)

Sweeney, S. H., & Feser, E. J. (1998). Plant size and clustering of manufacturing activity. *Geographical Analysis*, *30*(1), 45–64. https://doi.org/10.1111/j.1538-4632.1998.tb00388.x (https://doi.org/10.1111/j.1538-4632.1998.tb00388.x)

Tidu, A., Guy, F., & Usai, S. (2024). Measuring Spatial Dispersion: An Experimental Test on the *M*-Index. *Geographical Analysis*, *56*, 384–403. https://doi.org/10.1111/gean.12381 (https://doi.org/10.1111/gean.12381)