Spatial Economic Analysis
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rsea20

Measuring and Testing Spatial Mass Concentration with Micro-geographic Data

Florent Bonneu & Christine Thomas-Agnan
Published online: 11 Aug 2015.

To cite this article: Florent Bonneu & Christine Thomas-Agnan (2015) Measuring and Testing Spatial Mass Concentration with Micro-geographic Data, Spatial Economic Analysis, 10:3, 289-316

To link to this article: http://dx.doi.org/10.1080/17421772.2015.1062124

Spatial Economic Analysis, 2015
Vol. 10, No. 3, 289–316, http://dx.doi.org/10.1080/17421772.2015.1062124

Measuring and Testing Spatial Mass Concentration with Micro-geographic Data

FLORENT BONNEU & CHRISTINE THOMAS-AGNAN

(*Received January 2014; accepted July 2014*)

ABSTRACT *We address the question of measuring and testing industrial spatial concentration based on micro-geographic data with distance-based methods. We discuss the basic requirements for such measures and we propose four additional requirements. We also discuss the null assumptions classically used for testing aggregation of a particular sector and propose an alternative point of view. Our general index measure involves a cumulative and a non-cumulative version. This allows us to propose an alternative version of the Duranton–Overman index with a proper baseline as well as a cumulative version of this same index. We present simulations to evaluate the respective powers of this new approach and the classical ones.*

Mesure et tests de concentration spatiale d'une masse à l'aide de données micro-géographiques

RÉSUMÉ *nous nous penchons sur la question de la mesure et des tests de concentration spatiale industrielle basés sur des données micro-géographiques à l'aide de méthodes basées sur la distance. Nous discutons des exigences de base pour ces mesures, et présentons quatre exigences supplémentaires. Nous nous penchons également sur les hypothèses nulles utilisées traditionnellement pour tester l'agrégation d'un certain secteur, et proposons un point de vue alternatif. Notre mesure générale de l'indice comporte une version cumulative et non cumulative, ce qui nous permet de proposer une version alternative de l'indice de Duranton–Overman, avec une valeur de référence, ainsi qu'une version cumulée de ce même indice. Nous présentons des simulations permettant d'évaluer les puissances respectives de cette nouvelle approche et des approches classiques.*

Christine Thomas-Agnan, GREMAQ, Toulouse School of Economics, Toulouse, France. Email: Christine.Thomas@tse-fr.eu (to whom correspondence should be sent). Florent Bonneu, LMA Laboratory, University of Avignon, Avignon, France. Email: Florent.Bonneu@univ-avignon.fr.
This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).
No potential conflict of interest was reported by the authors.

Medición y prueba de la concentración de más espacial con datos microgeográficos

RESUMEN *tratamos la cuestión de medir y probar la concentración espacial industrial con base en datos microgeográficos con métodos basados en la distancia. Discutimos los requisitos básicos para realizar dichas medidas y proponemos cuatro requisitos adicionales. También hablamos sobre las suposiciones nulas que se suelen usar para para probar la agrupación de un sector particular y proponemos un punto de vista alternativo. Nuestra medida del índice general supone una versión acumulativa y una no acumulativa. Esto nos permite proponer una versión alternativa del índice de Duranton–Overman con una línea de base apropiada, al igual que una versión acumulativa de este mismo índice. Presentamos simulaciones para evaluar los poderes respectivos de este nuevo enfoque y los clásicos.*

使用微观地理数据测量和测试空间质量集聚

摘要: 我们设法采用基于距离的方法, 根据微观地理数据解决测量和测试工业空间集聚的问题。 我们讨论了此类测量的基本要求, 并提出了四项额外要求。 我们也探讨了通常用于测试特定行业聚合的零假设, 并提出了另一种观点。 我们的一般指数测量分为累积版本和非累积版本。这让我们提出一项设有合适基线的 Duranton–Overman 指数替代版本, 以及同一指数的累积版本。我们提出了评估这一新方法和经典方法各自优势的模拟.

KEYWORDS: *Spatial concentration; marked point processes; agglomeration; spatial clusters*

JEL CLASSIFICATION (2000): C12; C13; C15; C21; R12

1. Introduction

1.1. Literature

The question of measuring spatial mass concentration is encountered in many fields. This topic has received a lot of attention in the economics literature with the concentration of industrial settlement on which we focus in this paper. Similarly, in forestry it is interesting for example to study the spatial concentration of biomass from the knowledge of trees (or plants) locations and sizes. In epidemiology, when studying the spatial concentration of cancer cases, it is important to account for the number of cases in each hospital as a mark.

In economic geography, Krugman's theory states that 'instead of spreading out evenly around the world, production will tend to concentrate in a few countries, regions, or cities, which will become densely populated but also have higher levels of income'. There is empirical evidence that jobs and industries are clustered in a small number of regions. There are several mechanisms that induce this agglomeration. First of all, plants locate near to each other because of agglomeration spillovers or local amenities. Returns to scale induce industries to concentrate their production in a small number of business units and there is interdependence between firm's location choices (snowball effect mechanism).

There are numerous motivations for studying the geographic concentration of economic sectors. Such a measure allows to understand the determinants of localization, compare different sectors with respect to agglomeration/dispersion and predict the evolutions of localization. A similar question is that of co-localization and interactions between sectors for which measures can be generally derived from the former. Another related issue is that of cluster detection but we do not consider this problem in the present paper.

Until 2000, all studies about geographic concentration of economic activity use areal data for measuring spatial concentration. The precise location of firms is not available and the data only consists in aggregated counts over administrative zones. There is a large literature on this topic with many measures including the Herfindahl index, the locational Gini index (which is the Gini index of the localization ratio), the Ellison–Glaeser index, the Maurel-Sédillot's index and many others. However, these measures depend upon the aggregation level [modifiable areal unit problem (MAUP)] and, most importantly, they do not take geography into account in the sense that a permutation of the sites does not affect the measure. A good description of the drawbacks of these approaches is found in Arbia (2001).

A new vein of this literature arises in the years 2000 considering the treatment of micro-geographic data. This type of data usually consists in the precise location of firms together with a size measure such as the number of employees. Duranton and Overman (2005) introduce a measure based on the distribution of inter-distances between firms. It will be referred to subsequently as the Duranton–Overman (DO) index. Marcon and Puech (2003, 2010; Marcon et al., 2012) introduce another measure based on Ripley's *K*-function that we will refer to as the Marcon–Puech (MP) index. Combes et al. (2006) survey this literature. With the tools of point process theory, Giuliani et al. (2014) use a model-based approach to assess concentration with an index based on a weighted version of Ripley's *K*-function that we will refer to as the Espa-Giuliani-Arbia (EGA) index.

1.2. Basic Requirements

First of all, we should make clear that the problem is not only that of measuring firm's locations spatial concentration. The classical Ripley's *K*-function can be used for this purpose. We address the problem of taking into account firm's sizes in the measure. Indeed a mass characteristic is attached to each firm (like the number of employees or the capital) and the question of interest is that of spatial mass concentration and not spatial location concentration.

Duranton and Overman (2005) list five properties that a good measure of industrial geographic concentration should satisfy:

- Requirement [DO1]: The index must be comparable from one sector to the other. This implies that the measure should not depend upon the number of firms of a given sector nor upon the scale of the firm's sizes.
- Requirement [DO2]: The index must take into account the overall manufacturing geographical pattern. Indeed, the absence of concentration should not correspond to spatial homogeneity of locations because obviously geographic and demographic factors influence industrial location.

- Requirement [DO3]: The index must control for industrial concentration. Indeed, the problem of measuring the concentration of the firm's sizes should be distinguished from that of their spatial location concentration.
- Requirement [DO4]: The index must be independent of the geographical scale of observation. This is related to the so-called MAUP: the fact that aggregations over different geographical subdivisions of space may lead to diverging conclusions about the concentration pattern. This pleads for a method based on micro-geographic data versus the classical indices based on areal data.
- Requirement [DO5]: The index must be assorted with a level of statistical significance.

In this paper, we argue that the following four additional requirements should be added to the previous list:

- Requirement [BT1]: The index must be an empirical measure associated to a well-identified theoretical characteristic. The satisfaction of this requirement allows for correct statistical inference about the significance of the results (see Combes & Overman, 2004).
- Requirement [BT2]: The index must take into account spatial inhomogeneity of a particular sector. The factors influencing the inhomogeneity of locations can vary from sector to sector (think about fishing for example).
- Requirement [BT3]: The index must have a known and constant benchmark in the absence of concentration (under the null hypotheses). This requirement is stated by Combes and Overman (2004).
- Requirement [BT4]: For testing concentration, a null hypotheses must be correctly specified.

Concerning [BT1], if the purpose was just to construct a test statistic, it would not be necessary to link the statistic to a well-identified theoretical characteristic. But a concentration measure is also used as an indicator of concentration which should also make sense for the underlying theoretical process, condition necessary for requirement [BT4] to be satisfied. We will see that the DO index as well as the MP index and the EGA index are all inspired from the marked point process theory but only Giuliani et al. (2014) explicitly link their measure to a well-identified statistical parameter. Satisfying this requirement could allow to satisfy [DO5] without resorting to Monte Carlo methods.

[BT2] is a refinement of [DO2] in the sense that it extends the inhomogeneity recognition to inhomogeneity across sectors. The meaning of 'taking into account' in this context is that the inhomogeneity can be directly incorporated into the index so that the inhomogeneous index measures the second order concentration effects discounted from the first order ones as we will see in Section 6.3. We will show that the DO index as well as the MP index do not correctly satisfy [BT2]. Arbia et al. (2012) incorporate inhomogeneity in the framework of firm's location concentration (without mass characteristic) but Giuliani et al. (2014; which includes mass characteristic) do not correct for inhomogeneity.

With respect to [BT3], the MP index has a constant benchmark but not the DO one. In Giuliani et al. (2014), the benchmark value depends upon some

parameters and hence is not constant. The existence of a benchmark is important to allow comparison between sectors and/or between regions within a sector for example.

Indeed, as stated in [BT4], a null assumption should be stated in terms of a theoretical parameter. The distribution of the test statistic under the null should be known or at least it should be possible to simulate from this distribution. As we have seen with [BT1], the absence of clear specification of a theoretical parameter in the former literature is related to the absence of clear definition of the theoretical meaning of spatial concentration. Giuliani et al. (2014) use a specific point process model to reach this goal. Indeed, we will explain which aspects we believe are not entirely satisfactory in the simulation framework used in Duranton–Overman and Marcon–Puech for testing spatial concentration and will argue that there is no clear statement of the null assumption in their work. We propose an alternative approach for this purpose.

It is first necessary to present the mathematical tools of the spatial point process theory in Section 2. In Section 3, we discuss the different faces of spatial concentration and distinguish several types. Section 4 is devoted to the definitions of three classical indices based on inter-point distances and a presentation of their imperfections. In Section 5, we introduce our family of indices. We show how this family is related to the DO, MP and EGA indices, and how this relation sheds light on the mentioned imperfections. We show how this new point of view allows to introduce a modified version of the DO index which has a clear benchmark. This relationship also allows to make a minor correction in the EGA index which appears as a homogeneous version of the cumulative BT index for a particular weighting scheme.

We discuss the testing framework in Section 6. Finally in Section 7, we present some simulated examples to illustrate our arguments.

2. The Relevance of Spatial Point Pattern Models

2.1. Marked Spatial Point Processes

As we already observed, the tools from spatial point pattern models (PP hereafter) have inspired most of the industrial spatial concentration literature using micro-geographic data. Sweeney and Feser (1998) use Diggle and Chetwynd's *D* function (1991). Ripley's *K*-function is mentioned in Arbia (2001) and Marcon and Puech (2003). The need to take into account firm's sizes leads to consider more complex models which are called marked spatial point patterns. The spatial distribution of firms together with their sizes can be modelled using spatial point patterns associated with possibly several marks: the size and the sector for example. In this section, we briefly review the main theoretical concepts. Spatial point processes (PP) are models for a random spatial configuration of a random number N of points (for us: location of firms). One talks about a marked PP when a random mark is associated to each position (for us the mass characteristic, for example number of employees and sector of each firm). Mathematically, let $\mathcal{X}$ be a subset of R^2, a configuration of n points of $\mathcal{X}$ is a non-ordered set of n points $x = \{x_1, \cdots, x_n\}$. A PP model is a model for a random configuration with a random countable number N of points (possibly zero or infinity), repetitions being

allowed. Two mathematical approaches exist for this theory: they are based on locally finite random sets of points of $\mathcal{X}$ or alternatively on random measures on $\mathcal{X}$ and we refer the reader to Moller and Waagepetersen (2004) or to Illian et al. (2008) for precise definitions and properties.

Two important aspects of the description of these processes are spatial inhomogeneity and spatial interaction. Spatial inhomogeneity relates to the fact that some regions may have a mean number of points higher than others, for example when studying the spatial distribution of population, mountainous zones may be less populated. Spatial interaction relates to the dependence between points in pairs of locations. For example, the competition for food may generate repulsion between animals positions, whereas when looking at infectious disease cases, contagion generates attraction between spatial occurrences of a disease.

Spatial interaction is illustrated in Figure 1 with simulated realizations of such processes. In the centre, the process is a homogeneous Poisson process which embodies homogeneity and absence of interaction between points. On the right of Figure 1 is an aggregated process with interaction between the locations of an attraction type. On the left of Figure 1 is a regular process with interaction between the locations of a repulsion type. The circles on this figure will be commented upon later.

Two invariance properties will play a role later: stationarity and isotropy. A PP is stationary or homogeneous if its law is invariant under translations of the configurations. A PP is isotropic if its law is invariant under the rotations of the configurations. Figure 2 illustrates the notion of non-stationarity on the left panel and the notion of anisotropy on the right panel.

Because marks and locations are both random, their joint distribution has to be modelled with a marked point process model. Let M be a space for marks and, for each configuration X, let m_X be a random variable with values in M. Then one says that (X,m_X) is a marked PP with mark space M. In practice, we consider the case M finite (it is the case when the mark is the firm's sector) or M subset of R^p (it is the case when the mark is the firm's number of employees or the firm's capital).

Figure 3 presents a realization of an inhomogeneous marked Poisson PP with independent marks. It is usual to draw circles which show the mark through their radius.

2.2. Characteristics of a Marked Spatial Point Process

As for other types of stochastic elements, one can define characteristics of order one and two for point processes. Let us first consider the case of an unmarked PP. The order one characteristic of a PP is given by its intensity and captures large-scale

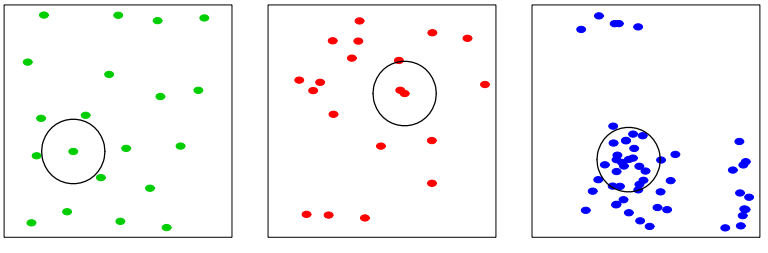

Figure 1. From left to right: regular PP, Poisson homogeneous PP and aggregated PP.

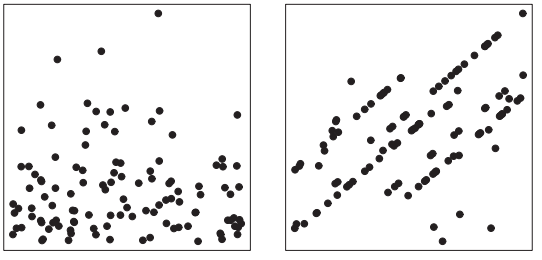

Figure 2. Left: non stationarity, Right: anisotropy.

variations of the process. For a sub-region B of $\mathcal{X}$, let $N_X(B)$ be the number of points of the PP X in B. The intensity measure for a sub-region B is defined by the expected number of points of X in B:

$$\Lambda(B) = \mathbb{E}(N_X(B)).$$

When this measure is absolutely continuous with respect to the Lebesgue measure, the intensity function λ is defined by:

$$\Lambda(B) = \int_B \lambda(u)du,$$

and can be interpreted as follows: $\lambda(u)du$ is the probability of occurrence of a point in the infinitesimal ball of *centre* u and radius du.

The intensity is constant in the homogeneous Poisson model and equal to the total number of points divided by the area of the region.

The order two structure of a PP which characterizes the small-scale variations can be specified by several tools. The order two factorial moment measure counts the mean number of point pairs with a point in A and the other in B:

$$\Lambda^{(2)}(A \times B) = \mathbb{E}\left(\sum_{u,v \in X: u \neq v} I(u \in A, v \in B)\right)$$

When this measure $\Lambda^{(2)}$ is absolutely continuous with respect to the Lebesgue measure, one can write:

$$\Lambda^{(2)}(A \times B) = \int_A \int_B \rho(u,v)dudv$$

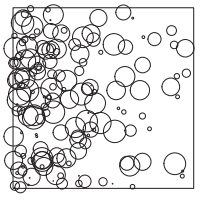

Figure 3. Marked Inhomogeneous Poisson PP.

where $\rho(u,v)$dudv can be interpreted as the probability of joint occurrence of a point in the infinitesimal ball of *centre u* with radius d*u* and of a point in the infinitesimal ball of *centre v* and of radius *du* The function ρ is named as the second order product density function.

Another way of characterizing the second order structure is through the pair correlation function which is related to ρ by:

$$g(x,y) = \frac{\rho(x,y)}{\lambda(x)\lambda(y)} \tag{1}$$

with the convention $\frac{a}{0} = 0$ if $a \geq 0$. A PP is said to be 'second order reweighted stationarity' when the function g is translation invariant $g(x,y) = g(\| x - y \|)$ with a slight abuse of notation.

At last, a third way of characterizing the second order structure is through the Ripley's *K*-function. In the stationary case, $\lambda K(r)$ is the mean number of points within radius r of the origin given that the origin belongs to the configuration (λ being the constant intensity). On Figure 1, a circle of radius r centred on a configuration point illustrates the fact that the *K*-function counts the mean number of points within a given radius of a point of the configuration. In the more general 'second order reweighted stationary' case, the Ripley's *K*-function can be defined through its relationship with the pair correlation function by:

$$K^{inhom}(r) = \pi \int_0^r ug(u)du,$$

Even though we mentioned that spatial homogeneity of locations is not a good reference for the concentration problem (see [DO2]), it is important to define this assumption in order to understand the testing procedures introduced later. The assumption of complete spatial randomness (CSR) is embodied by the Poisson homogeneous process or for which we have $K(r) = \pi r^2$ and $g(r) \equiv 1$.

For a marked PP, one needs to extend these definitions. These extensions are introduced and studied in the homogeneous case by Schlather (2001) and Illian et al. (2008).

Let (X, M) be a marked PP, homogeneous for positions. Let $k(m), q(m)$ be univariate weight functions and $f(m_1, m_2)$ be a bivariate weighting function which will be specified functions of the marks.

An order one characteristic called the mark-sum intensity measure Λ_k is given by:

$$\Lambda_k(B) = \mathbb{E} \sum_{u \in X} k(m_u) I_B(u).$$

For example, for $k(m) = m$, $\Lambda_k(B)$ is the expected number of employees in B whereas $\Lambda(B)$ is the expected number of firms in B. If this measure has a density with respect to Lebesgue measure $\Lambda_k(B) = \int_B \lambda_k(u)du$, then λ_k is the weighted intensity function for weighting function k.

When the weighting scheme is multiplicative $f(m_1, m_2) = k(m_1)q(m_2)$, one can define similarly a weighted version of the second order factorial moment measure

$\Lambda^{(2)}$ given by:

$$\Lambda_f^{(2)}(A \times B) = \mathbb{E}\left[\sum_{u,v \in X: u \neq v} k(m_u) q(m_v) I_A(u) I_B(v)\right].$$

When $\Lambda_f^{(2)}$ is absolutely continuous with respect to Lebesgue measure, one can write:

$$\Lambda_f^{(2)}(A \times B) = \int_A \int_B \rho_f(u, v) du dv$$

and ρ_f is called second order product density of X for weighting scheme f. A weighted version of Equation (1) yields a weighted version of the pair correlation function:

$$g_f(x, y) = \frac{\rho_f(x, y)}{\lambda_k(x) \lambda_q(y)} \tag{2}$$

and a weighted version of the Ripley's K-function under the assumption that g_f is translation invariant. This assumption is an adjustment of the classical second order reweighted stationarity assumption to the existence of an f-weighting scheme.

$$K_f(r) = \pi \int_0^r u g_f(u) du. \tag{3}$$

2.3. Estimating the Theoretical Characteristics

The estimation of these theoretical characteristics has been extensively studied under several assumptions and we refer the reader to Moller and Waagepetersen (2004) and Illian et al. (2008) for details. Let us just recall here the basic estimators that will be used in the sequel. Under the assumption of homogeneity, one can estimate the constant intensity λ from one realization of the process by:

$$\hat{\lambda} = \frac{N}{|W|}, \tag{4}$$

where N is the total number of points in W and $|W|$ is the area of the observation window W. Similarly, one can estimate in this case the Ripley's K-function by:

$$\hat{K}(r) = \frac{|W|}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} w_{i,j} I(\| x_i - x_j \| \leq r)$$

where $w_{i,j}$ is a boundary correction factor to take into account discs partially included in the region given by:

$$w_{i,j} = \frac{1}{|W \cap (W - x_i + x_j)|} = \frac{1}{|(W + x_i) \cap (W + x_j)|}$$

The corresponding estimate of the g function is given by:

$$\hat{g}(r) = \frac{|W|}{2\pi r N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{i,j} \frac{1}{h} \kappa\left(\frac{r - \|x_i - x_j\|}{h}\right). \tag{5}$$

In the inhomogeneous case, one can estimate the intensity by the following non-parametric estimator:

$$\hat{\lambda}^{inhom}(x) = \sum_{i=1}^{N} \frac{1}{h} \kappa\left(\frac{\|x - x_i\|}{h}\right) \tag{6}$$

where κ is a given kernel density function and h a given bandwidth. It is known (see for example Baddeley et al. (2000) and Diggle et al. (2007) that it is difficult to estimate both first and second order structure non-parametrically from a single realization since clusters due to attractive interactions are indistinguishable from clusters due to peaks in the first order structure. In the case of simulated data (as in Section 7), the way out for this problem is to use two separate bandwidths with a larger one for the first order structure (see more details about the bandwidth selection in Section 7). In real data applications, it is of course wiser to model the intensity using covariates thus imposing underlying assumptions on the smoothness of the intensity, as in Espa et al. (2013). Of course this type of approach is subject to mis-specification errors. In the second order reweighted stationary case, the following is an estimator of the inhomogeneous Ripley's K-function:

$$\hat{K}^{inhom}(r) = \frac{1}{|W|} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{i,j} \frac{I(\| x_i - x_j \| \leq r)}{\hat{\lambda}^{inhom}(x_i)\hat{\lambda}^{inhom}(x_j)}$$

and the pair correlation function can be estimated by:

$$\hat{g}^{inhom}(r) = \frac{1}{2\pi r} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{i,j} \frac{\frac{1}{h}\kappa\left(\frac{r - \|x_i - x_j\|}{h}\right)}{\hat{\lambda}^{inhom}(x_i)\hat{\lambda}^{inhom}(x_j)}$$

In the marked PP case, assuming that marks are independent from positions, we have that $\lambda_k(x) = \lambda(x)E(k(m_X))$, and one can thus estimate the weighted intensity function for example by:

$$\hat{\lambda}_k(x) = \hat{\lambda}(x)\overline{k(m_X)}, \tag{7}$$

where $\hat{\lambda}$ can be understood as Equation (4) in the homogeneous positions case and as Equation (6) in the inhomogeneous positions case and where $\overline{k(m_X)}$ is the empirical mean of the transformed marks.

Similarly, in the second order reweighted stationary and isotropic case, one can estimate the weighted version of the pair correlation function by:

$$\widehat{g_f}(r) = \frac{1}{2\pi r} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{h^{-1}\kappa\left(\frac{r - \|x_i - x_j\|}{h}\right) k(m_i)q(m_j)}{|W \cap (W - x_i + x_j)|\hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)}, \tag{8}$$

Table 1. Error rates in 500 repetitions of scenario 1

	DO	MP	BThom	BTinhom
Local envelopes test				
PPhom (%)	32	100	11	3
PPinhom (%)	0	0	0	1
Deviation test				
PPhom (%)	1	100	4	3
PPinhom (%)	3	0	29	1

where $\hat{\lambda}$ can take the two different forms, Equations (4) or (6), leading to two versions of this estimators $\widehat{g_f}^{inhom}$ and $\widehat{g_f}^{hom}$ and the weighted version of the Ripley's K function by:

$$\hat{K}_f = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{k(m_i)q(m_j)I\left(\left\|x_i - x_j\right\| \leq r\right)}{|W \cap (W - x_i + x_j)|\hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)} \tag{9}$$

leading similarly to the two estimators $\hat{K}_f^{hom}$ and $\hat{K}_f^{inhom}$ depending upon which form between Equations (4) and (6) has been selected to estimate the intensity.

3. The Different Faces of Spatial Concentration

In this section, we discuss the definition of spatial concentration for locations and for mass and distinguish between several types. People have an intuitive idea of what a concentrated pattern is but it may be more difficult to define as it first seems. We can define spatial concentration of firm's locations as the fact that firms are more aggregated in space than in a random pattern. The reverse situation of inhibition when firms are more scattered than in a random case does not lead to spatial concentration. For mass concentration, it can be described as the fact that the employees are more aggregated in space than in a random pattern defined by a homogeneous distribution is space.

Figure 4 shows three examples of spatially concentrated marked processes. In the first two cases (left and centre), we chose constant marks on purpose to start with a simple situation. In the left panel, the point process is an inhomogeneous Poisson PP and the concentration aspect of the configuration is due to the inhomogeneity of positions (order one) and not to interaction (order two): Arbia et al. (2012) call it apparent contagion. In the centre panel, the concentration aspect is due to interaction of aggregation type and thus we say that it is a

Table 2. Error rates in 500 repetitions of scenario 2

	DO	MP	BThom	BTinhom
Local envelopes test				
PPhom (%)	53	100	8	3
Aggregated (%)	0	0	0	38
Deviation test				
PPhom (%)	13	100	4	3
PPinhom (%)	16	0	1	38

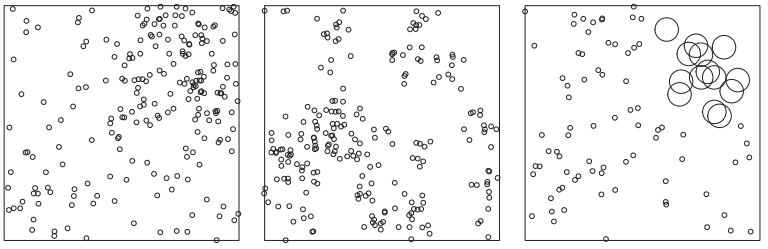

Figure 4. Several types of concentration.

concentration of order 2: Arbia et al. (2012) call it true contagion. The left and centre panels bring face to face two types of concentration: that due to inhomogeneity of locations (order one) and that due to interactions between locations (order two). It is important to understand that the difference between the two types is that, if one were able to observe a second realization of the same process, the clusters would appear in the approximate same locations in the order one concentration case, corresponding to peaks of intensity, whereas they could arise in very different locations in the second type.

If we now oppose the centre and right panels, we see that concentration of the mass in the centre panel is due to constant marks with aggregated positions, whereas in the right panel mass distribution is inhomogeneous in space but located at homogeneously distributed positions resulting in a third kind of concentration due to marks. Giuliani et al. (2014) oppose these two cases using the terms of clustering of firms for the centre case and clustering of economic activities for the right one.

Of course some situations may involve a concentration due to an interplay between positions and marks. We see that there are several types of concentration and that the marks may or may not induce this concentration. Mass concentration is indeed a complicated interaction between the locations and the marks sizes. We will concentrate in this paper on the case when marks are independent from positions (the case of the right panel of Figure 4 is then excluded) leaving the other case for a further paper. We claim that not only it is important to measure concentration but also that it may be relevant to determine which type of concentration is present in the data at hand. In an applied perspective, as we have seen previously, we need decision rules to separate empirically between first order intensity/concentration and second order structure/concentration. It is often a matter of scale in the sense that order one concentration is a larger scale phenomenon, whereas order two is small scale one. When we can use a parametric model of intensity, we may also consider that what comes under order one is what can be explained by contextual variables whereas order two contains all other unobserved factors.

4. Indices Based on Inter-points Distances

In this section, we recall the definitions of the classical indices based on inter-point distances: the DO index (Duranton & Overman, 2005), the MP index (Marcon & Puech, 2003, 2010) and EGA index (Giuliani et al., 2014). We use a unified notation in order to ease the comparisons and we discuss their imperfections. Let $x_{i,s}$ denote the location of firm $i (i = 1, \cdots, n)$ of sector $s (s = 1, \cdots, S)$ and let $m_{i,s}$ be the corresponding mark (to illustrate we will use a mark equal to the number of employees).

4.1. The DO, MP and EGA indices

The DO index is defined for each sector separately hence we drop momentarily the sector index. It is a non-cumulative index defined for any $r > 0$ by:

$$i_{DO}(r) = \frac{\sum_{i=1}^{N_s} \sum_{j=1,j\neq i}^{N_s} h^{-1}\kappa\left(\frac{r-\|x_i-x_j\|}{h}\right) m_i m_j}{\sum_{i=1}^{N_s} \sum_{j=1,j\neq i}^{N_s} m_i m_j},$$

When the mark is a count, which is the case for the number of employees, it can be compared to the Parzen–Rosenblatt density estimator associated to a replicated point process of positions (number of replications equal to the mark) considering points positions as i.i.d.

Starting from the fact that i_{DO} does not account for order one inhomogeneity of locations, Marcon and Puech (2010) propose to perform this correction by using the union of all the available sectors as a reference. Note that no correction is then possible if there is only one sector available. The MP index is a cumulative index defined for any $r > 0$ by:

$$I_{MP}(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j I\left(\|x_{i,s} - x_{j,s}\| \leq r\right)}{\sum_{j=1,j\neq i}^{N} m_j I\left(\|x_{i,s} - x_j\| \leq r\right)} \Big/ \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j}{\sum_{j=1,j\neq i}^{N} m_j},$$

$I_{\mathrm{MP}}(r) > 1$ indicates that there are proportionally more employees close to plants of sector s within a radius r than in the whole area. Note that $I_{\mathrm{MP}}(r)$ can be written $J_{\mathrm{MP}}(r)/J_{\mathrm{MP}}(\infty)$ where,

$$J_{MP}(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j I\left(\|x_{i,s} - x_{j,s}\| \leq r\right)}{\sum_{j=1,j\neq i}^{N} m_j I\left(\|x_{i,s} - x_j\| \leq r\right)}.$$

$J_{\mathrm{MP}}(r)$is the average proportion of employees of sector s among all sectors within a given radius r.

Giuliani et al. (2014) propose to use a weighted Ripley's K-function defined as follows for any $r > 0$:

$$I_{EGA}(r) = \sum_{i=1}^{N_s} \sum_{j=1,j\neq i}^{N_s} \frac{m_i m_j I\left(\|x_{i,s} - x_{j,s}\| \leq r\right)}{|W \cap (W - x_{i,s} + x_{j,s})| N \hat{\lambda} \hat{\mu}^2}, \tag{10}$$

where W is an observation window, $\hat{\mu}$ is an estimator of the mean value of the mark and $\hat{\lambda}$ is an estimator of the mean value of the intensity of locations. They associate to this empirical measure a corresponding theoretical EGA and they derive a closed form formula for it in the framework of a particular log-Gaussian Cox model which they use for testing concentration.

4.2. The Imperfections of the Classical Indices

Let us first explain the weaknesses of these classical indices, postponing the discussion about the imperfections of the corresponding testing strategies to Section 6:

(1) Except for EGA, these indices are introduced as purely empirical quantities and there are no theoretical characteristics clearly associated to them hence they do not satisfy requirement [BT1].
(2) With respect to the [DO2] requirement, the DO index takes location inhomogeneity into account in the simulation framework (with the fact that locations remain unchanged), but it certainly does not incorporate inhomogeneity in the formula of the index itself. The MP index tries to take it into account in the measure itself but we will show in Section 5.4 that this correction is not entirely satisfactory.
(3) DO and EGA do not take into account inhomogeneity of location intensity of a particular sector hence do not satisfy requirement [BT2]. MP avoid this problem by considering relative indices.
(4) There is no clear benchmark for DO (cf. [BT3]); the benchmark for EGA in the log-Gaussian Cox model depends upon some parameters.
(5) There is no edge correction for DO (which implies bias for large r).

5. Introducing the Family of BT Indices

In an attempt to correct some of these imperfections, we present an approach using some theoretical characteristics of spatial marked point processes which will allow us to cast the previous approaches in a same mould and to point at their respective weaknesses. In this paper, we will consider that marks can be assumed to be independent from positions.

We propose to construct the indices as estimators of the following two characteristics to measure spatial mass concentration: a non-cumulative measure corresponding to the weighted pair correlation function (Equation 2) and a cumulative measure corresponding to the weighted Ripley's K-function (Equation 3).

For a given choice of multiplicative weighting scheme, we introduce the non-cumulative BT index by:

$$i_{BT}(r) = \hat{g}_f(r) = \frac{1}{2\pi r} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{h^{-1}\kappa\left(\frac{r-\|x_i-x_j\|}{h}\right) k(m_i)q(m_j)}{|W \cap (W - x_i + x_j)| \hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)} \tag{11}$$

with the weighted intensity function λ_k being estimated by Equation (7).

Our index is an estimator of the theoretical g_f characteristic. It is defined at any distance $r > 0$. It is important to note that this index can be calculated under the assumption of homogeneity of the intensity of positions as well as under the assumption of inhomogeneity using one of the two estimators of the intensity, Equations (4) or (6), and this leads to two versions of our index called $i_{\mathrm{BT}}^{\mathrm{hom}}$ and $i_{\mathrm{BT}}^{\mathrm{inhom}}$ thereafter. In the homogeneous case, the square of the intensity appears in the denominator and can be estimated by a slightly different version of Equation (4) which is unbiased for λ^2 (see Illian et al., 2008) namely:

$$\widehat{\lambda^2} = \frac{N(N-1)}{|W|^2}. \tag{12}$$

The intensity is estimated for each sector separately so that requirement [BT2] is satisfied.

5.1. The BT Index: Cumulative Version

For a given multiplicative weighting scheme, a corresponding cumulative version of the BT index is given by the following estimator of the weighted K-function, defined at any distance $r > 0$:

$$I_{BT}(r) = \hat{K}_f(r) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{k(m_i)q(m_j)I\left(\left\|x_i - x_j\right\| \leq r\right)}{|W \cap (W - x_i + x_j)|\hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)}. \tag{13}$$

In the case that $\| x_i - x_j \|$ is small compared to the diameter of W, the border correction term approaches $|W|$ so that we can consider that a version without border correction is obtained by substituting $|W|$ for $|W \cap (W - x_i + x_j)|$.

As for the non-cumulative one, this index can be calculated under the assumption of homogeneity of the intensity of positions as well as under the assumption of inhomogeneity using one of the two estimators of the intensity, Equations (4) or (6), and this leads to two versions of this cumulated index called $I_{\mathrm{BT}}^{\mathrm{hom}}$ and $I_{\mathrm{BT}}^{\mathrm{inhom}}$ thereafter.

5.2. Consequences for the DO Index

In this section, we establish a link between the DO index and the classical estimate of the weighted pair correlation function g_f for the following choice of weighting scheme $k(m) = m$ and $q(m) = m$. Indeed for this choice, we have the following result (see Appendix for a proof) when considering the homogeneous BT index without border correction:

$$i_{DO}(r) = \frac{2\pi r}{|W|}\hat{g}_f(r) = \frac{2\pi r}{|W|} i_{\mathrm{BT}}(r).$$

This formula induces a natural normalization of the DO index $\frac{|W|}{2\pi r} i_{\mathrm{DO}}(r) = i_{\mathrm{BT}}(r)$ with a clear benchmark since we will see in the next section that under our proposed H_0 assumption we have $g_f \equiv 1$.

This link allows us to propose a cumulative version of the DO index:

$$I_{DO}(r) = \frac{\sum\sum_{j\neq i} m_i m_j I\left(\left\|x_i - x_j\right\| \leq r\right)}{\sum\sum_{j\neq i} m_i m_j} = \frac{\hat{K}_f(r)}{|W|}$$

5.3. Consequences for the MP Index

Comparing,

$$J_{\mathrm{MP}}(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1, j\neq i}^{N_s} m_j I\left(\left\|x_{i,s} - x_{j,s}\right\| \leq r\right)}{\sum_{j=1, j\neq i}^{N} m_j I\left(\left\|x_{i,s} - x_j\right\| \leq r\right)}$$

and

$$I_{\mathrm{BT}}(r) = \hat{K}_f(r) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{k(m_i)q(m_j)I\left(\left\|x_i - x_j\right\| \leq r\right)}{|W \cap (W - x_i + x_j)|\hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)}$$

for $k(m) = m$ and $q(m) = 1$, we first see that in the stationary case, the two indices are related by $I_{\mathrm{BT}}(r) = \frac{|W|}{N} J_{\mathrm{MP}}(r)$. Moreover, ignoring the bias correction term (which was also proposed in some versions of the MP index), and focusing on the denominator, we understand that the correction for inhomogeneity of the location intensity of sector s is missing in the MP index (see details in Appendix).

5.4. Consequences for the EGA Index

For the weighting scheme given by $f(m_1, m_2) = m_1 m_2$, if we compare:

$$I_{BT}(r) = \hat{K}_f(r) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{k(m_i) q(m_j) I\left(\left\|x_i - x_j\right\| \leq r\right)}{|W \cap (W - x_i + x_j)| \hat{\lambda}_k(x_i) \hat{\lambda}_q(x_j)}$$

and

$$I_{EGA}(r) = \sum_{i=1}^{N_s} \sum_{j=1, j\neq i}^{N_s} \frac{m_i m_j I\left(\left\|x_i - x_j\right\| \leq r\right)}{|W \cap (W - x_i + x_j)| N \hat{\lambda} \hat{\mu}^2}, \tag{14}$$

we find that:

- the EGA index is an homogeneous (location intensity) version of the cumulative BT index;
- there is a minor mistake in its denominator $|W| I_{\mathrm{EGA}} = I_{\mathrm{BT}}$, which has no impact in their paper since they have $|W| = 1$.

6. Testing Strategy

We now turn attention to the definition of a null hypotheses for testing mass concentration and to the testing strategy. Our two main concerns about the testing strategy in the classical approach are that there is no clear null hypotheses identified and that there is unstated assumption that all sectors originate from the same process.

6.1. The Null Hypotheses

The question we want to test is that of absence of mass concentration and we need to specify a clear null hypotheses corresponding to this idealistic situation.

For the classical DO and MP approaches, the proposed test of absence of concentration is based on the following Monte Carlo framework. M permutations of the observed firm's locations are randomly chosen for all sectors altogether. The marks (size for DO and couples size and sector for MP) are then reallocated to the permuted locations. Both in the Duranton–Overman and the Marcon–Puech framework, the simulations are done conditionally upon the positions: marks (sector and number of employees) are randomly reassigned to the observed positions. This same procedure is used in Illian et al. (2008) for testing the assumptions of 'independent marking' (also called 'random labelling', case of uncorrelated marks) and that of geostatistical marking (case of correlated marks) but with distinct test statistics. We believe this approach is only valid for the case when all sectors originate from a single model. Indeed, one finds this as an unstated

assumption in the classical approach that all sectors are issued from the same type of process, the 'overall manufacturing' process. However, if each sector has its own intensity or dependence structure, the fact of mixing these processes in the simulations generates confounding effects. For the EGA approach, H_0 corresponds to the nullity of one parameter in the log-Gaussian Cox model but we claim that even under this restriction on the parameters, the log-Gaussian Cox Process may exhibit concentration.

Ideally, we would like to use the hypotheses $g_f = 1$ as H_0. For constant marks, it boils down to $g = 1$ which is not equivalent to the CSR assumption but it is implied by this assumption (we come back to this problem in Section 7). Indeed deviations of g_f from 1 may arise as the result of large marks in some regions or as the result of aggregated locations (or as a combination of both). With this choice of null, the fact that our index (which is also our test statistic) is an estimator of g_f implies that we fulfil the constant benchmark requirement [BT4]. The difficulty however is that unlike in the case $g = 1$, one does not know how to simulate under the assumption $g_f = 1$ if we do not further restrict the process. The strategy we propose is to simulate under a more restricted null hypotheses of a Poisson PP model for positions with independent marks (following the same distribution throughout space) for which we know how to simulate realizations. We allow this Poisson process to be homogeneous or not, leading to two versions of the null H_0^{hom} and H_0^{inhom} and therefore to two versions of the simulations scenario. For the simulations under the null, we generate realizations of a Poisson PP with the intensity given by Equation (4) if we are testing H_0^{hom} and given by Equation (6) if we are testing H_0^{inhom}. In a real application, one could use a model based on covariates for estimating the intensity instead of Equation (6). Before introducing our testing strategy, we recall in the next section how the Ripley's *K*-function can be used to test for CSR. In the introduction, we argued that CSR was not a good benchmark for studying spatial concentration of industrial location but this is just a preliminary step in order to better understand the tools introduced later.

6.2. Using the K*-function to Test for Complete Spatial Randomness*

Figure 5 shows a realization of an inhomogeneous Poisson process on the left panel. The central panel shows the ordinary *K*-function and the right panel the inhomogeneous *K*-function: both are displayed together with an empirical envelope obtained by Monte Carlo simulations of a Poisson process with intensity estimated from the data (using Equation (4) in the central panel and Equation (6) in the right one). The central *K*-curve is out of the envelope, whereas the right K-curve is inside the envelope: the estimation of the *K*-function in the central panel does not take into account inhomogeneity, whereas this is done using the inhomogeneous *K* estimator on the right panel. The fact that the curve is outside the envelope in the central panel is not due to the presence of interaction but rather due to inhomogeneity. A parallel can be done with a time series situation when the unaccounted presence of a trend may reveal a wrong serial correlation. The fact that the curve is back in the envelope on the right panel is coherent with the fact that the inhomogeneous Poisson model does not exhibit order two

Figure 5. Use of K to test CSR.

interaction. This procedure in two steps allows to distinguish between inhomogeneity of intensity from order two interaction between the locations.

6.3. Using the Weighted K-function to Test for Concentration

Following the classical approach for testing CSR of locations described in the previous section, we propose a two-step procedure in order to separate concentration of order one from concentration of order two:

- Test H_0^{hom}:

(1) if accept: conclude that there is no concentration;
(2) if reject: go to next step.

- Test H_0^{inhom}:

(1) if accept: conclude that there is significant concentration of order 1 (apparent contagion);
(2) if reject: conclude that there is significant concentration of order 2 (true contagion).

We simulate each point process corresponding to each sector separately with a homogeneous Poisson model in case of H_0^{hom} and an inhomogeneous Poisson model in case of H_0^{inhom} after estimating its intensity. This allows for sectors with different intensity-driven processes and therefore to satisfy requirement [BT2]. In the simulations of next section, the intensity of positions λ is estimated locally by a non-parametric kernel method or by a non-parametric iterative and adaptive method based on Voronoï cells. The choice of a global bandwidth is difficult in practice, notably with very wide variations in the intensity function λ as mentioned in Diggle et al. (2007). Indeed, the method proposed in Berman and Diggle (1989) of minimizing an estimation of the mean square error (MSE) of $\hat{\lambda}$ produces in our case a value of h close to zero. Therefore, we tune manually the value of h based on the optimal value obtained by the MSE criterion taking into account the fact that the first order concentration appears at a larger geographical scale than the second order one. We also tested an adaptive estimation of the intensity by averaging several estimations of the intensity which are constants in each Voronoï cell of their random Dirichlet tessellation. This non-kernel method also yields good results. For the estimation of our concentration index itself, we compute the optimal h defined by Stoyan's rule of thumb (1995, for the homogeneous case) and adapt it so that it is smaller than the value of h used for the estimation of the intensity. The expectation of the mark is estimated by the empirical mean of marks. We do not mix the sectors in a permutation framework as in the DO or MP approach.

Due to the fact that the test statistic is a function of the distance r, we face a multiple testing problem and we have two options. A first option is to use the local envelopes to build a global test for which we do not control the global nominal level. For a given local nominal level α, we select at each distance r the α and $1-\alpha/2$ quantile among the M realizations of the index at r: this defines the lower and upper local envelopes. We reject the null when the observed curve gets out of the upper envelope at least once. Note that we use a single sector at a time.

A second option is to do a deviation test. We compute for each simulated process and for the observed one the maximum over the distances of the absolute

value of the difference between the index of the given process and the mean index over all the simulations. We then compute an empirical significance level for the observed deviation in the distribution of the simulated deviations and take a decision with a given nominal level.

7. Simulations

We devise some simulations to compare our testing strategy with the classical DO and MP indices approaches. We do not include the EGA index in the comparison because it can be viewed as a homogeneous version of our cumulative index (once corrected from the mentioned minor mistake). We simulate two sectors, non-necessarily of the same type with respect to spatial homogeneity and interaction: in scenario 1, we have a homogeneous Poisson process versus an inhomogeneous Poisson process; and in scenario 2, a homogeneous Poisson versus an aggregated Matern process. Scenario 3 illustrates an exceptional case. Because the possible interplay between marks and positions may obscure the comparisons, we only focus here on the case of marks independent from positions and following a discrete uniform distribution. We compare the following indices:

- the DO index (original version, non-cumulative)
- the cumulative MP index
- the indices BThom and BTinhom (non-cumulative versions)

MP, BThom and BTinhom all have a benchmark of 1 under H_0. The envelopes are based on $M = 1000$ replications. The confidence level is a local 5% level. We first present graphs of the indices on one realization of the processes before proceeding to the analysis of the comparative performance of the corresponding tests on replications of these simulated processes.

7.1. Scenario 1

Scenario 1 has two sectors:

- Sector 1 is homogeneous Poisson with intensity 100 and uniform marks on $\{0, \cdots, 50\}$.
- Sector 2 is inhomogeneous Poisson with uniform marks on $\{0, \cdots, 50\}$ with intensity function given by $\lambda(x, y) = \frac{500}{1-\exp(-5)} \exp(5x)$.

Sector 2 has the same expected number of points as sector 1 and Figure 6 shows one realization of these two processes with sector 1 on the left panel.

Figure 7 presents the graphs of the DO index for the two sectors and Figure 8 presents the MP index for the two sectors with sector 1 on the left panel. We can see that DO and MP detect concentration of sector 2. MP concludes that sector 1 is also concentrated which is not true.

Figure 9 presents the graphs of the BThom index for the two sectors and Figure 10 presents the BTinhom index for the two sectors. For sector 1, BThom and BTinhom curves are both inside the envelope which is compatible with the homogeneous Poisson nature of the locations together with the uniform independent marks. For sector 2, we can see that BThom and BTinhom correctly detect that the origin of its concentration comes from first order since the BThom

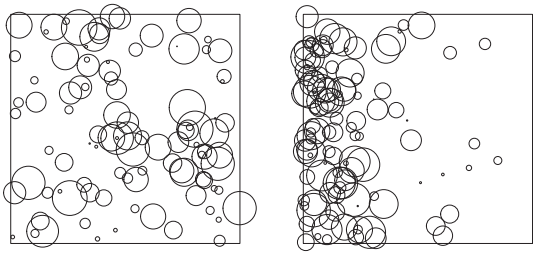

Figure 6. Scenario 1: the two sectors.

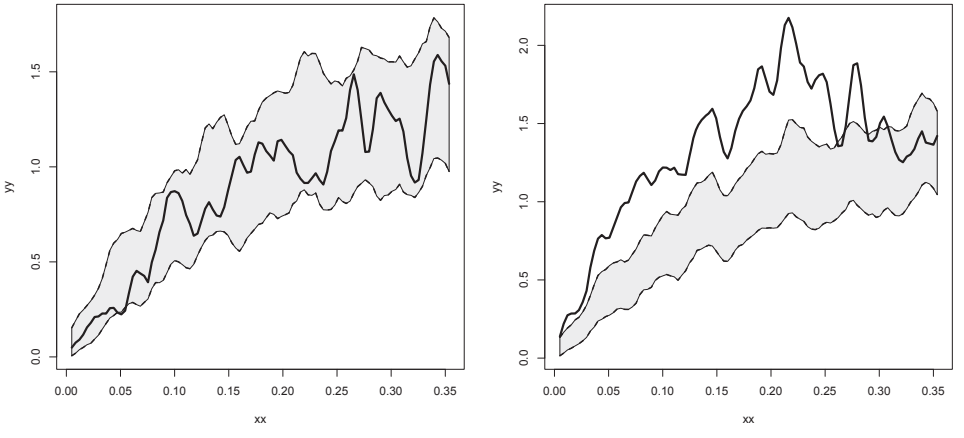

Figure 7. DO index for scenario 1.

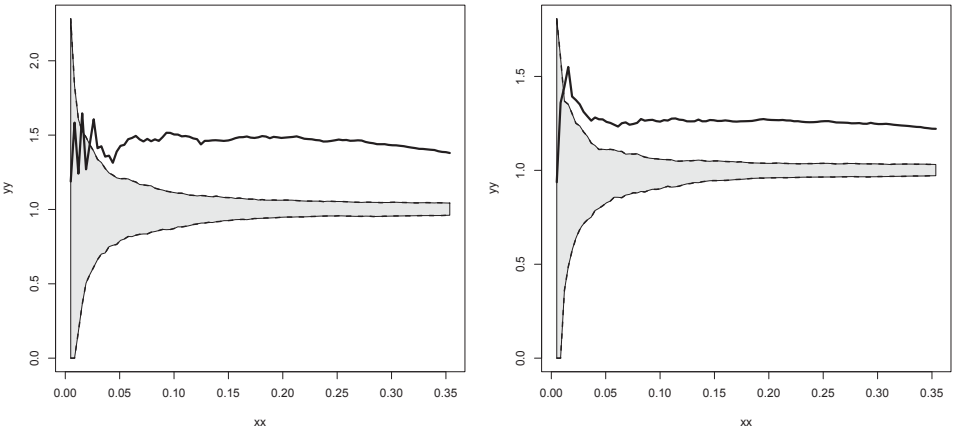

Figure 8. MP index for scenario 1.

curve lies outside the envelope and the BTinhom is inside (compare with Figure 5).

We then analyse the 500 repetitions of the simulated scenario 1 and Table 1 contains the percentage of error for the tests based on local envelopes and for the

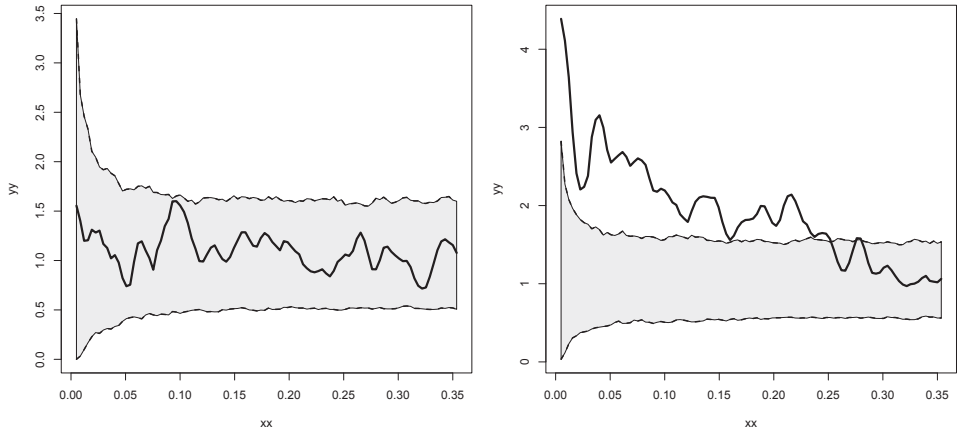

Figure 9. BThom index for scenario 1.

deviation tests. It is easier to make an error for sector 1 than for sector 2. The error rate is clearly higher for DO and MP than for BThom and BTinhom. Recall also that BThom and BTinhom should be used jointly one after the other and that the conclusion is more informative since it reveals not only concentration but also its nature.

7.2. Scenario 2

Scenario 2 has two sectors:

- Sector 1 is homogeneous Poisson with intensity 100 and uniform marks on $\{0, \cdots, 50\}$.
- Sector 2 is a Matern process (parent process: homogeneous Poisson with intensity 10, children process: homogeneous Poisson in a disc of radius 0.1) and uniform marks on $\{0, \cdots, 50\}$.

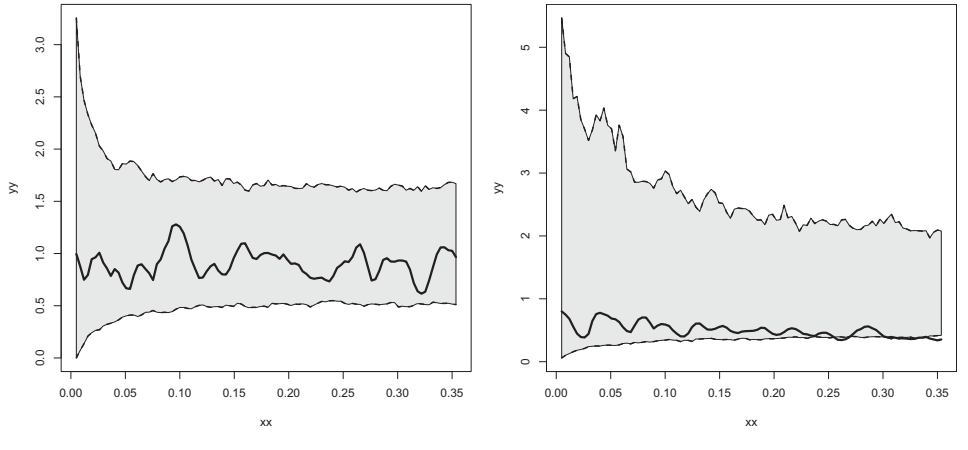

Figure 10. BTinhom index for scenario 1.

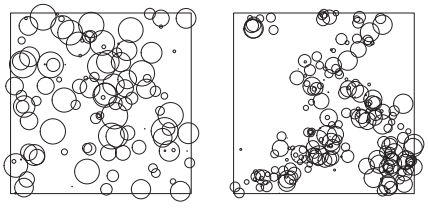

Figure 11. Scenario 2: the two sectors.

Figure 11 shows one realization of these two processes with sector 1 on the left panel. Figure 12 presents the graphs of the DO index for the two sectors and Figure 13 presents the MP index for the two sectors. We can see that DO and MP detect concentration of sector 2. MP concludes that sector 1 is also concentrated which is not true.

Figure 14 presents the graphs of the BThom index for the two sectors and Figure 15 presents the BTinhom index for the two sectors. We can see that BThom and BTinhom correctly detect that the origin of concentration of sector 2 comes from second order.

We then run 500 simulations of scenario 2 and Table 2 contains the percentage of error of the tests based on local envelopes and of the deviation tests. The conclusions are similar than for scenario 1.

7.3. Scenario 3

Scenario 3 has two sectors:

- Sector 1 is homogeneous Poisson with intensity 100 and uniform marks on $\{0, \cdots, 50\}$.
- Sector 2 is Non-Poisson process described in Baddeley et al. (2000) and such that $g = 1$ and uniform marks on $\{0, \cdots, 50\}$.

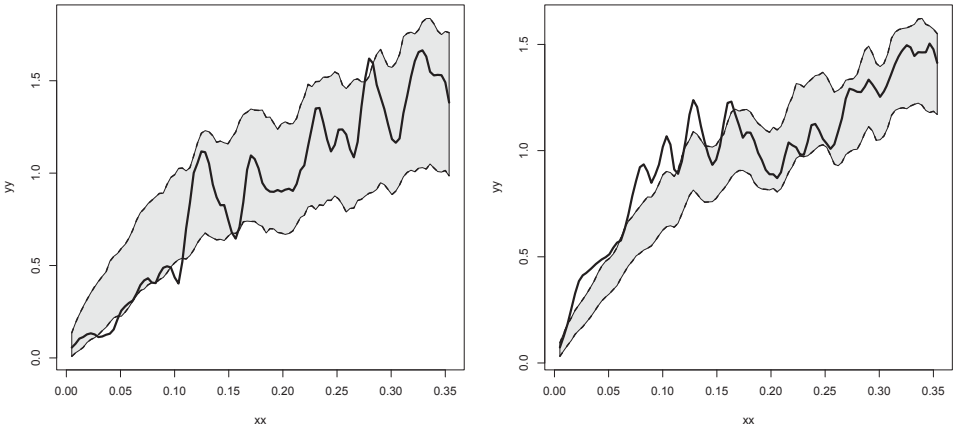

Figure 12. DO index for scenario 2.

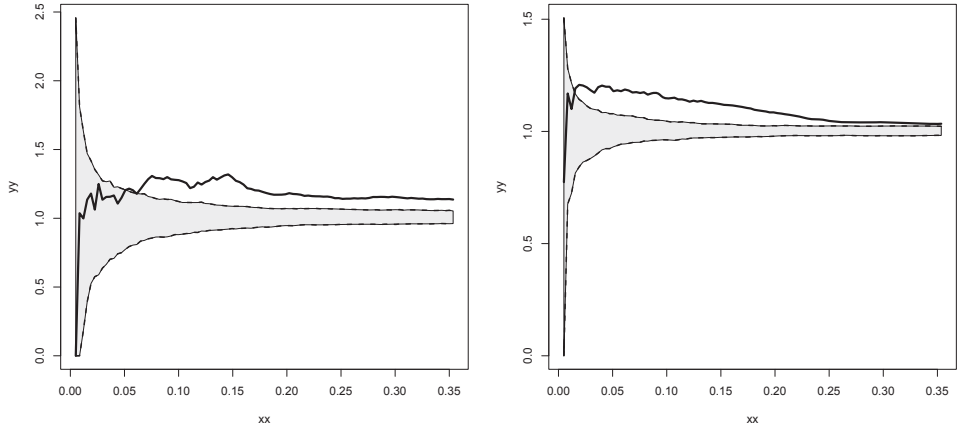

Figure 13. MP index for scenario 2.

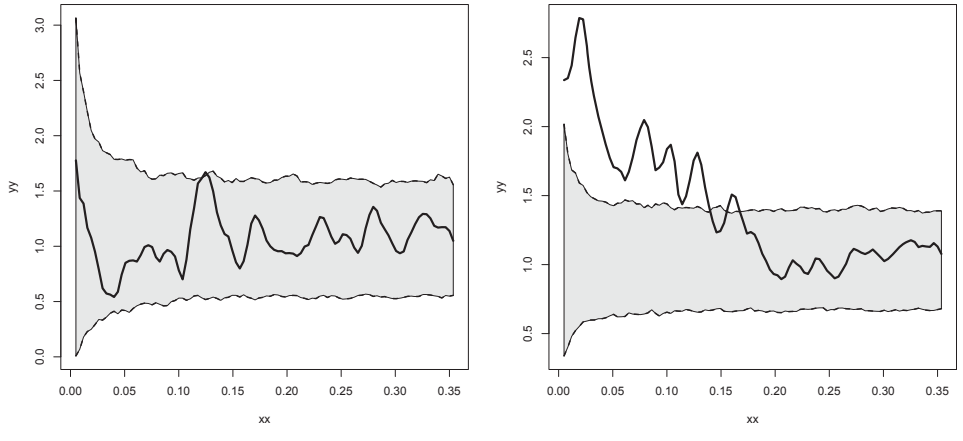

Figure 14. BThom index for scenario 2.

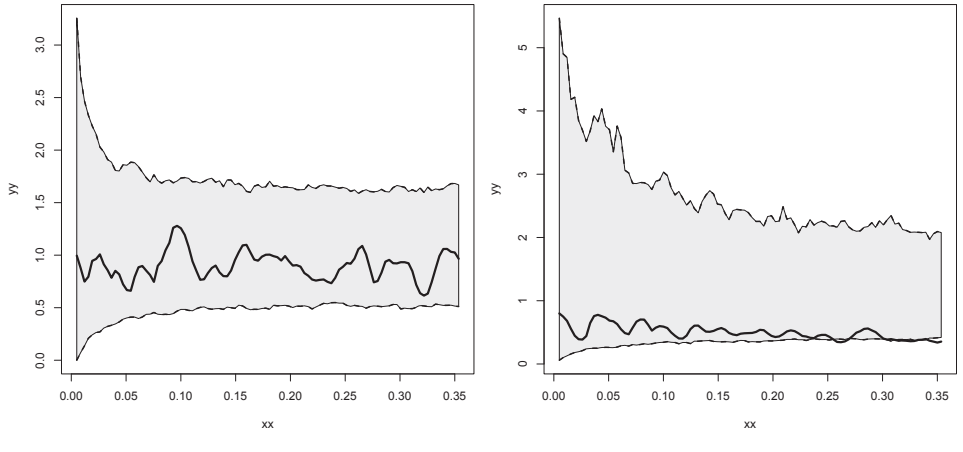

Figure 15. BTinhom index for scenario 2.

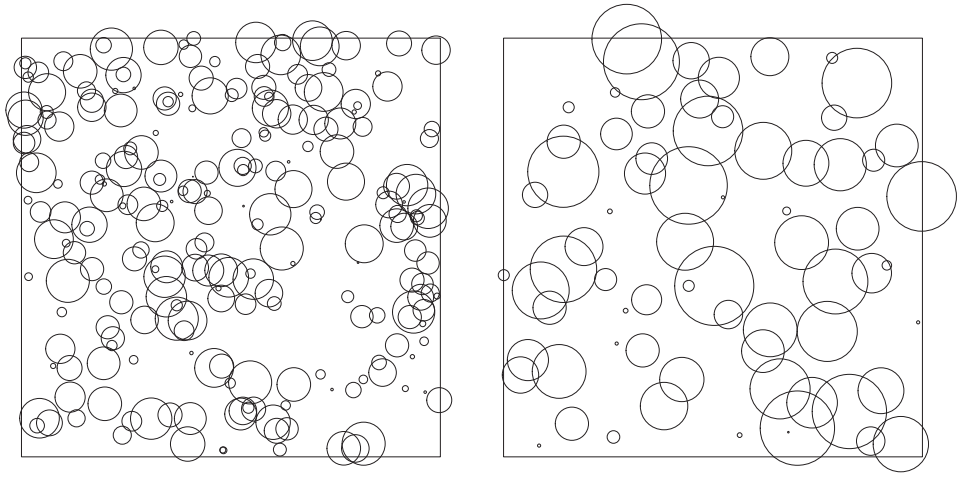

Figure 16. Scenario 3: the two sectors.

Note that sector 2 satisfies $g_f = 1$ but the process is not Poisson hence presents interaction between the locations. However, there is no concentration effect as can be seen on the realization shown on Figure 16.

Figure 17 presents the graphs of the DO index for the two sectors and Figure 18 presents the MP index for the two sectors. We can see that the indices DO and MP do not detect any concentration for sectors 1 and 2.

Figure 19 presents the graphs of the BThom index for the two sectors. We can see that BThom does not detect any concentration for sector 2.

8. Conclusion

We have introduced a family of spatial concentration indices which subsumes the classical DO, MP and Espa et al. indices. The BT indices have a cumulative and non-cumulative version. The BT indices satisfies the nine objectives DO1 to DO5 and BT1 to BT4. The relationships between the classical indices and our family allow to introduce a normalized version of the DO index with a proper benchmark

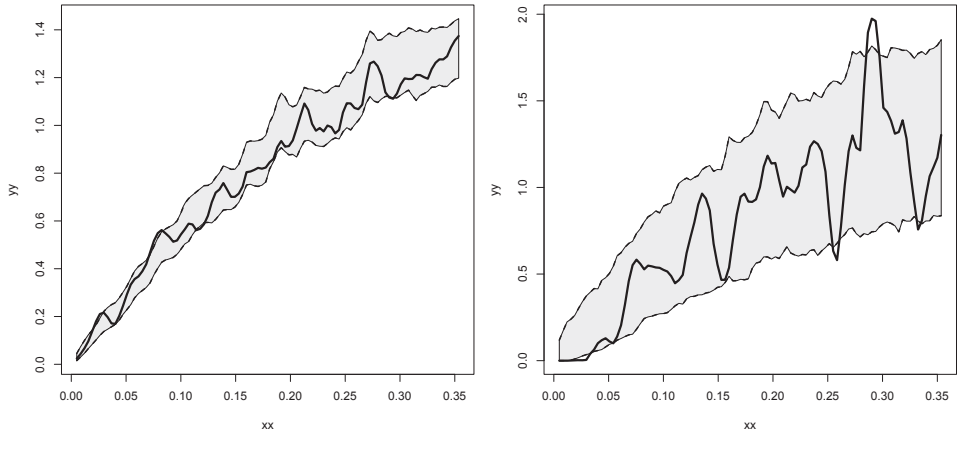

Figure 17. DO index for scenario 3.

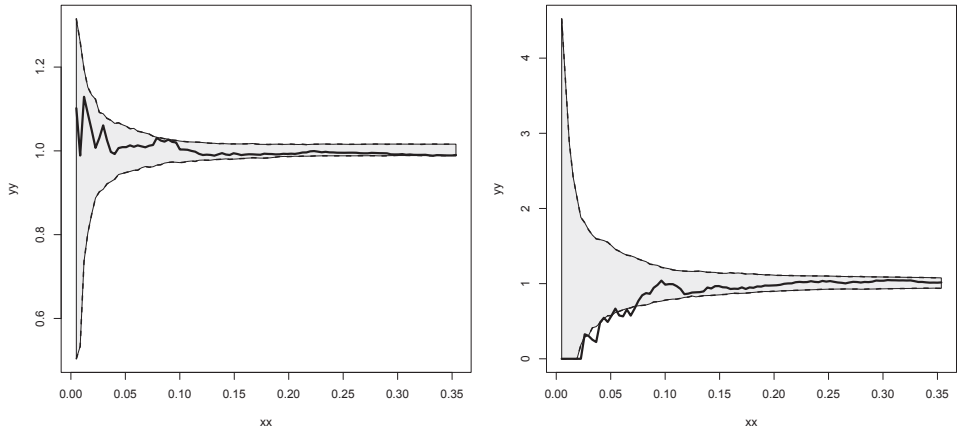

Figure 18. MP index for scenario 3.

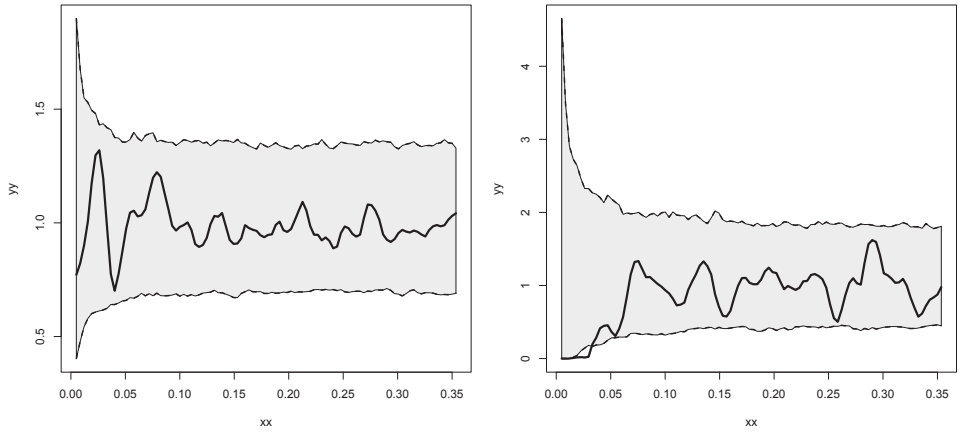

Figure 19. BThom index for scenario 3.

in the absence of concentration and to correct some weaknesses of the other two indices. The simulations show that these new indices yield good results concerning tests error rates. The perspectives for further discussion are the choice of weighting scheme f related to the concrete interpretation of the indices and the proper monitoring of the global level of the tests based on local envelopes. Another important perspective which is the object of current further research is to treat the case of dependence between marks and positions. This implies a different method of estimation for the weighted intensity function as well as a different simulation procedure under the null hypotheses.

References

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

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

Baddeley, A. J., Moller, J. & Waagepetersen, R. (2000) Non-and semi-parametric estimation of interaction in inhomogeneous point patterns, *Statistica Neerlandica*, 54(3), 329–350.

Berman, P. & Diggle, P. J. (1989) Estimating weighted integrals of the second-order intensity of a spatial point process, *Journal of the Royal Statistical Society. Series B.*, 51, 81–92.

Combes, P.-P., Mayer, T. & Thisse, J.-F. (2006) *Chapitre 10. Mesurer la concentration spatiale* [Chapter 10. Measuring spatial concentration]. Document Universite Paris 1.

Combes, P.-P. & Overman, H. (2004) The spatial distribution of economic activities in the European Union. In J. Henderson, J. F. Thisse (eds.) *Handbook of Urban and Regional Economics*, pp. 2845–2909, Elsevier, Amsterdam.

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

Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G. & Gooding, S. (2007). Second-order analysis of inhomogeneous spatial point processes using case-control data, *Biometrics*, 63(2), 550–557.

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

Espa, G., Arbia, G. & Giuliani, D. (2013). Conditional versus unconditional industrial agglomeration: disentangling spatial dependence and spatial heterogeneity in the analysis of ICT firms' distribution in Milan, *Journal of Geographical Systems*, 15(1), 31–50.

Giuliani, D., Arbia, G. & Espa, G. (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.

Illian, J., Illian, P, Stoyan, H. & Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*, Wiley, Statistics in practice, New York.

Marcon, E. & Puech, F. (2003) Evaluating the geographic concentration of industries using distance-based methods, *Journal of Economic Geography*, 3(4), 409–428.

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

Marcon, E., Puech, F. & Traissac, S. (2012). Characterizing the relative spatial structure of point patterns, *International Journal of Ecology*, 2012, 11 pages.

Moller, J. & Waagepetersen, R.P. (2004) *Statistical Inference and Simulation for Spatial Point Processes*, vol. 100. Chapman & Hall CRC, London.

Schlather, M. (2001) On the second-order characteristics of marked point processes, *Bernoulli*, 7(1), 99–117.

Stoyan, D. & Stoyan, H. (1995) *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*, John Wiley and Sons, New-York.

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

Appendix

Proof of Link between BT and DO

$$i_{DO}(r) = \frac{\sum_i \sum_{j>i} h^{-1} w\left(\frac{r-\left\|x_i-x_j\right\|}{h}\right) m_i m_j}{\sum_i \sum_{j>i} m_i m_j} = \frac{\sum_i \sum_{j\neq i} h^{-1} w\left(\frac{r-\left\|x_i-x_j\right\|}{h}\right) m_i m_j}{\sum_i \sum_{j\neq i} m_i m_j}$$

We define classical estimators $\hat{\rho}_f^{(2)}$, $\widehat{\lambda^2}$ and $\hat{\mu}^2$ for respectively $\rho_f^{(2)}$, λ^2 and $\mu^2 := E[M]^2$, in the stationary case, with the following formulae:

$$\hat{\rho}_f^{(2)}(r) = \frac{1}{2\pi r|W|} \sum_i \sum_{j\neq i} h^{-1} w\left(\frac{r-\left\|x_i-x_j\right\|}{h}\right) m_i m_j$$

Illian et al. (2008, eq.5.3.54, p. 354)

$$\widehat{\lambda^2} = \frac{N(N-1)}{|W|^2}$$

Illian et al. (2008, eq.4.3.34, p. 231)

$$\hat{\mu}^2 = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} m_i m_j$$

adapted from Illian et al. (2008, eq.5.3.48, p. 353)

Consequently, we have:

$$\begin{aligned} i_{DO}(\mathrm{r}) &= \frac{2\pi r |W| \hat{p}_f^{(2)}(r)}{|W^{(2)}| \frac{N(N-1)}{|W|^2} \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} m_i m_j} \\ &= \frac{2\pi r}{|W|} \frac{\hat{p}_f^{(2)}(r)}{\widehat{\lambda^2} \hat{\mu}^2} = \frac{2\pi r}{|W|} \hat{g}_f(r) = \frac{2\pi r}{|W|} i_{\mathrm{BT}} \end{aligned}$$

Proof of Link between BT and MP

Several conditions are needed to establish a link between I_{BT} and I_{MP} in the same way as I_{BT} and I_{DO}. We present this link in a stationary framework, with $f(m_1, m_2) = m_2$:

$$\begin{aligned} I_{BT} &= \frac{1}{|W|} \sum_{i=1}^{N_s} \sum_{j=1, j \neq i}^{N_s} \frac{m_j I\left(\left\|x_{i,s} - x_{j,s}\right\| \leq r\right)}{\hat{\hat{\lambda}}_q^2} \\ &= \frac{1}{|W|} \sum_{i=1}^{N_s} \sum_{j=1, j \neq i}^{N_s} \frac{m_j I\left(\left\|x_{i,s} - x_{j,s}\right\| \leq r\right)}{\frac{N(N-1)}{|W|^2} \hat{\mu}} \end{aligned}$$

where $\hat{\mu}$ is estimated by $\frac{1}{N-1} \sum_{j=1, j \neq i}^{1} m_j I\left\|x_{i,s} - x_j\right\| \leq r$.

We have:

$$I_{\mathrm{BT}} = \frac{|W|}{N} \sum_{i=1}^{N_s} \frac{\sum_{j=1, j \neq i}^{N_s} m_j I\left(\left\|x_{i,s} - x_{j,s}\right\| \leq r\right)}{\sum_{j=1, j \neq i}^{N} m_j I\left(\left\|x_{i,s} - x_j\right\| \leq r\right)} = \frac{|W|}{N} J_{\mathrm{MP}}(r)$$