A unified framework for measuring industry spatial concentration based on marked spatial point processes

Christine Thomas-Agnan and Florent Bonneu

Toulouse School of Economics
GREMAQ

JMS2012

Objectives

Measure spatial concentration from micro-geographic data with

$$\text{locations} + \text{mass (mark)}$$

Related issues : measure of co-localisation, cluster detection.

Bibliography for micro-geographic data

- Duranton, G. and Overman, H.G. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies* **72** 1077-1106.
- Marcon, E. and Puech, F. (2010) Measures of the geographic concentration of industries : improving distance-based methods. *Journal of Economic Geography* **10(5)** 745-762.
- Combes P-J., Meyer T., and Thisse J-F. (2008) Measuring spatial concentration, in "Economic geography : the integration of regions and nations", Princeton university press.
- G. Espa, D. Giuliani and G. Arbia (2010). Weighting Ripley's K-function to account for the firm dimension in the analysis of spatial concentration, Department of Economics Working Papers 1012, Department of Economics, University of Trento, Italia.

Objectives as specified by DO

Duranton et Overman (2002) list 5 properties that a good measure of industrial spatial concentration should satisfy

1. DO1 The index must be comparable from one sector to the other (should not depend upon the number of firms in the sector)
2. DO2 The index must take into account the overall manufacturing geographical pattern (benchmark is not spatial homogeneity because geographic and demographic factors influence industrial location)
3. DO3 The index must take into account the structural differences of a particular sector / country ("degree of industrial concentration") i.e. take into account firm's sizes
4. DO4 The index must be independent of the geographical scale of observation (MAUP)
5. DO5 The index must be assorted with a level of statistical significance

Additional objectives as specified by BTA

1. BTA1 The index must be an empirical measure associated to a well identified theoretical characteristic. This last point is not satisfied by the current candidates in the literature. This point may allow to satisfy DO5 without using Monte Carlo methods.
2. BTA2 The index must take into account spatial inhomogeneity of a particular sector (for example fishing)
3. BTA3 The index must take into account a possible inhomogeneity of the distribution of firm's sizes in space.
4. BTA4 The index must have a known and constant benchmark in the absence of concentration.
5. BTA5 For testing concentration, a null hypotheses must be correctly specified.

Modelling a random point pattern

Tool : spatial point processes (PP) are models for a random spatial configuration of a random number of points N (for us :location of firms for different industrial sectors)

Spatial Inhomogeneity : some regions may have a mean number of points higher than others

Example : mountainous zones may be less populated

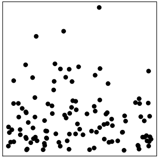
Spatial interaction : dependence between points locations

Example : competition for food may generate repulsion between animals positions, whereas for an infectious disease, contagion generates attraction between spatial occurences of the disease

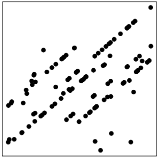
Marked PP : a random mark is associated to each position (for us : number of employees + sector)

Stationarity

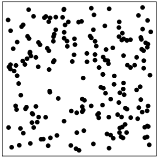
A PP is **stationary** if its law is invariant under translations of the configurations
A PP is **isotropic** if its law is invariant under the rotations of the configurations



(a) Non stationary

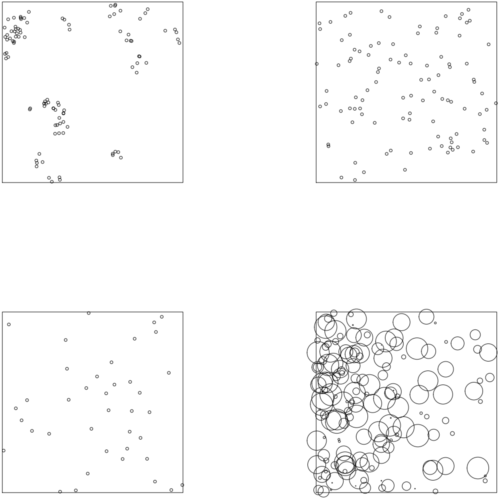


(b) Anisotropic



(c) Stationary and isotropic

Some examples of realizations



Order 1 characteristics of a PP

$N_X(B)$ is the number of points of PP X in B
Intensity measure

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

When Λ is absolutely continuous wrt the Lebesgue measure, one can write

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

where λ is called the **intensity function**

Order 2 characteristics of a PP : order 2 factorial moment measure

Order 2 factorial moment measure (mean number of points pairs with a point in A and the other in B)

$$\Lambda^{(2)}(A \times B) = \mathbb{E}(\sum_{u,v \in X : u \neq v} 1(u \in A, v \in B))$$

When $\Lambda^{(2)}$ is absolutely continuous wrt the Lebesgue measure, one can write

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

Order 2 characteristics of a PP : pair correlation function

It is defined by

$$g(x, y) = \frac{\lambda^{(2)}(x, y)}{\lambda(x)\lambda(y)}$$

with the convention $\frac{a}{0} = 0$ if $a \geq 0$.
A PP is said to be "second order reweighted stationarity" when g is translation invariant

Order 2 characteristics of a PP : Ripley's K function

If X is "second order reweighted stationary" and isotropic, the **Ripley's K function** is defined by

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

In this case, $\lambda K(r)$ is the mean number of points within radius r of the origin given that the origin belongs to the configuration.
Under CSR (PPP : Poisson homogeneous process) :
$K(r) = \pi r^2$ and $g(r) \equiv 1$

Estimators of PP characteristics (isotropic)

Under homogeneity assumption

$$\hat{\lambda}(x) = \frac{N}{\mid \mathcal{X} \mid}$$

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

where $w_{i,j}$ is a boundary correction factor

Estimators of PP characteristics (isotropic)

Under inhomogeneity assumption

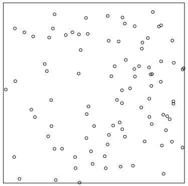
$$\hat{\lambda}(x) = \sum_{\xi \in \mathcal{X}} \kappa((x - \xi)/h)/h$$

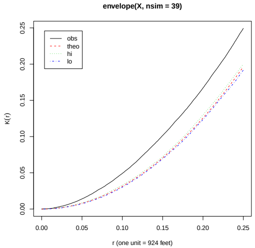
$$\hat{K}_{inhom} = \frac{1}{\mid \mathcal{X} \mid} \sum_{i \neq j} w_{i,j,r} \frac{1(\| x_i - x_j \| \leq r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)}$$

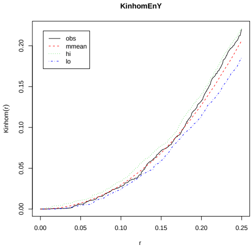
where $w_{i,j,r}$ is a boundary correction factor

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

Use of K function to test CSR







Characteristics of a marked PP

Let (X, M) be a marked PP, homogeneous for positions, and let $f(m_1, m_2)$ be a weighting function, we define a weighted version of $\alpha^{(2)}$ by

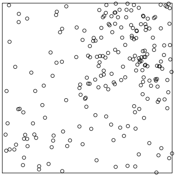
$$\alpha_f^{(2)}(A \times B) = \mathbb{E}\left[\sum_{u,v \in X : u \neq v} f(m_1, m_2) \boldsymbol{1}_A(u) \boldsymbol{1}_B(v)\right].$$

When $\alpha^{(2)}$ is absolutely continuous wrt the Lebesgue measure, one can write

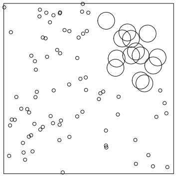
$$\alpha_f^{(2)}(A \times B) = \int_A \int_B \rho_f^{(2)}(u, v) du dv$$

then $\rho_f^{(2)}$ is called second order product density of X for weighting scheme f.

Order 1 concentration

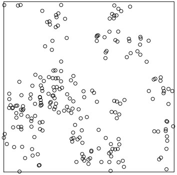


Inhomogeneity of positions



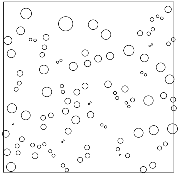
Inhomogeneity of marks conditionally on positions

Order 2 concentration

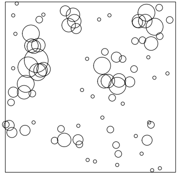


Aggregation of positions

Order 2 concentration



Constructed marks : distance between each point and its nearest neighbor



Constructed marks : number of neighbors at *dist* ≤ 0.1

The Duranton-Overman index (2005)

Based on inter-distances $\| x_i - x_j \|$

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

Can be compared to the PR density estimator associated to the replicated PP of positions.

The Marcon-Puech index (2010)

MP note that i_{DO} does not account for order 1 inhomogeneity. They propose to use the union of all the available sectors to perform this correction.

$$I_{MP}(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j \mathbf{1}(\|x_{i,s} - x_{j,s}\| \leq r)}{\sum_{j=1,j\neq i}^{N} m_j \mathbf{1}(\|x_{i,s} - x_j\| \leq r)} / \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j}{\sum_{j=1,j\neq i}^{N} m_j} \quad \forall r > 0,$$

$I_{MP}(r)$ can be written $J_{MP}(r)/J_{MP}(\infty)$ where

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

Hypotheses H0 for DO and MP

- Simulations are done conditionally upon the positions
- Marks (sector + number of employees) are randomly reassigned to the observed positions.

This simulation framework is not compatible with BTA3.

The weaknesses of MP and DO

1. there are no theoretical characteristics clearly associated to these indices (cf BTA1)
2. the possible dependence between marks and positions is not incorporated in the index formula (cf BTA3)
3. DO does not take into account inhomogeneity of location intensity (cf BTA2)
4. no clear benchmark for DO (cf BTA4)
5. no edge correction (implies bias for large r)
6. underlying assumption that all sectors are issued from the same type of process ("overall manufacturing")(cf simulations under H_0)

The theoretical characteristics : order 1

In the non stationary case, for any weight function k, we introduce the weighted intensity measure α_k

$$\alpha_k(D) = \mathbb{E} \sum_{u \in X} k(m) \mathbf{1}_D(u).$$

For $k(m) = m$, $\alpha_k(D)$ is the expected number of employees in D whereas $\Lambda(D)$ was the expected number of firms in D.
If $\alpha_k(D) = \int_D \lambda_k(u) du$ then λ_k is the weighted intensity function for weighting function k.

The theoretical characteristics : order 2

For two weighting functions k and q, and for a multiplicative scheme $f(m_1, m_2) = k(m_1)q(m_2)$ we introduce the weighted measure $\beta_f^{(2)}$, corresponding to the unweighted $\alpha^{(2)}$,

$$\beta_f^{(2)}(A \times B) = \mathbb{E}\left[\sum_{u,v \in X: u \neq v} \frac{f(m_1, m_2)}{\lambda_k(u)\lambda_q(v)} \mathbf{1}_A(u)\mathbf{1}_B(v)\right]$$

with $\lambda_k(x) > 0$ and $\lambda_q(x) > 0$ ps for all $x \in A$. If $\beta_f^{(2)}(A \times B) = \int_A \int_B g_f(u, v) du dv$ then g_f is the weighted pair correlation function for weighting function f.

The Bonneu-Thomas-Agnan index : non cumulative version

Non cumulative version for all $r > 0$

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

with

$$\hat{\lambda}_k(x) = \hat{\lambda}(x)\mathbb{E}[k(\hat{M})|X]$$

NB : the index can be calculated under the assumption of homogeneity of the intensity of positions as well as under the assumption of inhomogeneity $\mapsto$ two estimators BThom and BTinhom.

Null hypotheses and estimations

Null hypotheses :
$H0$: Poisson point process for positions with marks depending only on their own position.

Estimations :
For a given sector, we estimate :
1) The intensity of positions λ is estimated locally by a non parametric kernel method or by an non parametric iterative and adaptative method based on Voronoï cells.
2) The expectation of the mark conditionally on the position is estimated by a non-parametric kernel method or by an non parametric iterative and adaptative method based on Voronoï cells.

Simulations under the null hypotheses

1) We generate a realization of a Poisson PP with the same intensity as in the estimation step.
2) For each point of the realization, we estimate the conditional cumulative distribution function of the mark conditionally on the position by a non-parametric kernel method. We then simulate a mark realization from this cdf.

The Bonneu-Thomas-Agnan index : cumulative version

Cumulative version

$$I_{BT}(r) = \hat{K}_f(r) = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \frac{k(m_i)q(m_j)\mathbb{1}(\|x_i - x_j\| \leq r)}{|A \cap (A - x_i + x_j)|\hat{\lambda}_k(x_i)\hat{\lambda}_q(x_j)} \quad \text{pour tout } r >$$

Consequences for the Duranton-Overman index

We establish a link between the Duranton-Overman index and a theoretical characteristic g_f (the weighted pair correlation function)

$$i_{DO}(r) = \frac{2\pi r}{|A|}\hat{g}_f(r)$$

hence we derive a natural normalization of this index
with a clear benchmark : under H_0 we have $g_f \equiv 1$
We can also propose a cumulative version of this index

$$I_{DO}(r) = \frac{\sum\sum_{j\neq i} m_i m_j \mathbb{1}(\|x_i - x_j\| \leq r)}{\sum\sum_{j\neq i} m_i m_j} = \frac{\hat{K}_f(r)}{|A|}$$

$$\hat{K}_f(r) = |A|\frac{\sum\sum_{j\neq i} m_i m_j \mathbb{1}(\|x_i - x_j\| \leq r)}{\sum\sum_{j\neq i} m_i m_j}$$

Consequences for the Marcon-Puech index

Comparing

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

and

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

for $k(m) = m$ and $q(m) = 1$, we understand that the correction for inhomogeneity of the location intensity of sector s is missing in the MP index.

Framework of the simulated scenarios

We simulate two sectors, non necessarily of the same type.

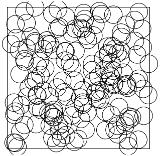
We compare

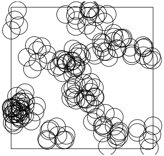
- the normalized DO index (non cumulative version)
- the cumulative MP index
- the indices BThom and BTinhom (non cumulative versions)

They all have a benchmark of 1 under H_0.

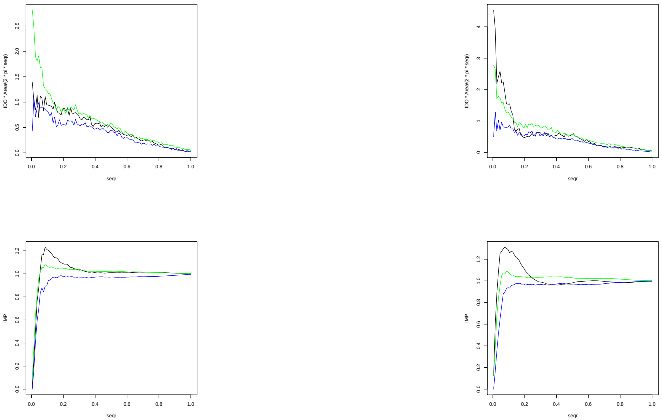
Scénario 1

Two sectors :
1) Homogeneous Poisson with constant marks.
2) Aggregated process with constant marks.



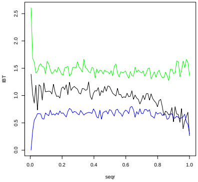


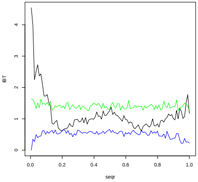
The indices DO and MP for scenario 1

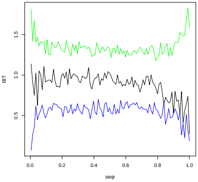


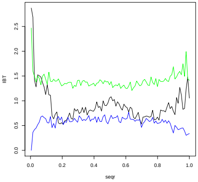
The indices DO et MP detect concentration of sector 2

The indices BThom and BTinhom for scenario 1









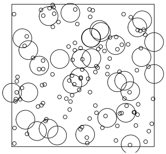
The indices BThom et BTinhom correctly detect that the origin of concentration of sector 2 comes from second order.

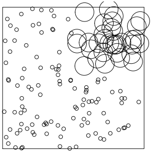
Scénario 2

Two sectors :

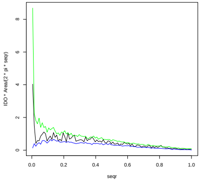
1) Homogeneous Poisson with random marks independent from positions.

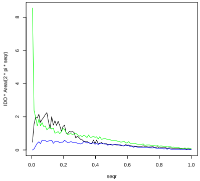
2) Homogeneous Poisson with marks depending upon the positions.

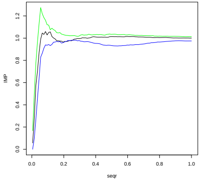


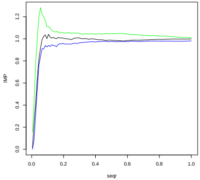


The indices DO et MP for scenario 2



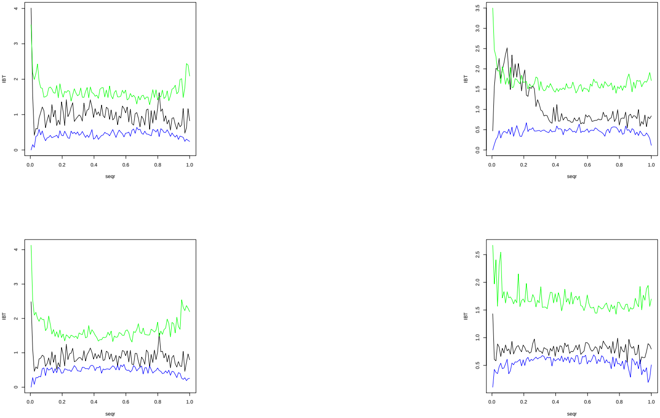






The index DO detects concentration for sector 2 and MP does not detect anything.

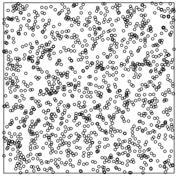
The indices BThom and BTinhom for scenario 2

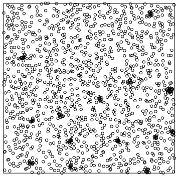


The index BThom detects concentration and BTinhom does not hence the origin of this concentration of sector 2 comes from the first order.

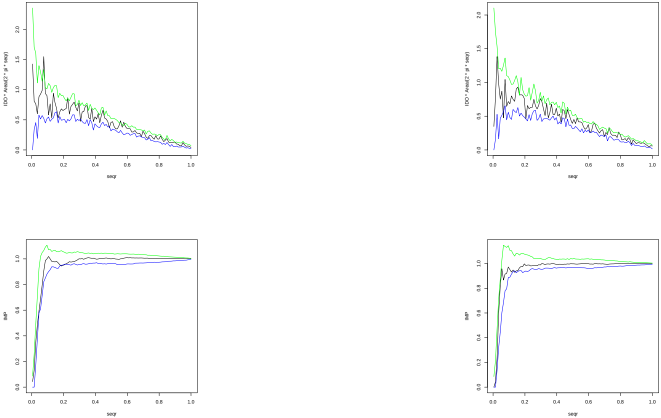
Scénario 3

$g_f = 1$ but the process is not Poisson.
Two sectors :
1) Homogeneous Poisson and constant marks.
2) Non-Poisson process described in BMW2000 and such that $g = 1$.



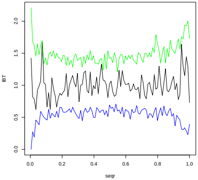


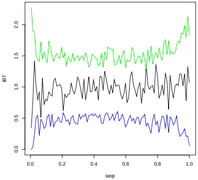
The indices DO and MP for scenario 3

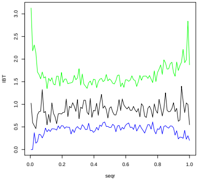


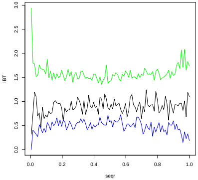
The indices DO and MP do not detect any concentration for sector 2

The indices BThom and BTinhom for scenario 3









The indices BThom and BTinhom do not detect any concentration for sector 2.

Conclusion

- the BT index satisfies the ten objectives DO1 to DO5 and BT1 to BT5
- depend upon r : advantage or disadvantage ?
- choice of weighting scheme f ?
- interpretation
- what to do for scenario 3 ?
- same tools can be used to study co-localization