I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/32

Specialized Agglomerations with Areal Data: Model and Detection

HAEDO, C. and M. MOUCHART

Specialized Agglomerations with Areal Data: Model and Detection *

Christian HAEDO[a] and Michel MOUCHART[b]

[a] *Centro Interuniversitario de Investigaciones sobre Desarrollo Económico, Territorio e Instituciones (CIDETI), University of Bologna in Argentine Republic*

[b] *Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA) and Center for Operations Research and Econometrics (CORE), UCLouvain, Belgium*

August 8, 2013

Abstract

This paper develops new statistical and computational methods for the automatic detection of spatial clusters displaying an over- or under- relative specialization spatial pattern. A probability model provides a space partition into clusters representing homogenous portions of space as far as the probability of locating a primary unit is concerned. A cluster made of contiguous regions is called an agglomeration. A greedy algorithm detects specialized agglomerations through a model selection criteria. A random permutation test evaluates whether the contiguity property is significant. Finally this algorithm is run on Argentinean data. Evaluating the proposed methodology concludes the paper.

Keywords: relative specialization, specialized agglomeration, areal (lattice) data, spatial clustering, spatial cluster detection, permutation bootstrap.

Corresponding Author: Michel MOUCHART, ISBA, voie du Roman Pays 20 - L1.04.01, B-1348 Louvain-la-Neuve(B), e-mail: michel.mouchart@uclouvain.be

*Michel Mouchart gratefully acknowledges financial support from IAP research network grant $nrP6/03$ of the Belgian government (Belgian Science Policy). Both authors gratefully acknowledge the financial support of CIDETI, that promoted intercontinental cooperation. The help of Cristian S. Rocha, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (Arg.) has been instrumental in shaping the algorithm developed in this paper and is deeply acknowledged. A special thank is due to Vicente Donato for the impetus he gave to the development of the topic of this paper. The authors are also grateful to Dominique Peeters (UCL) for several clarifying discussions during the elaboration of this paper.

Contents

1 Introduction

This paper aims at making a contribution to the detection of agglomerations understood as a cluster of *contiguous* regions. In the New Economic Geography (NEG), the concept of agglomeration is used as complementary evidence of the existence and significance of the role played by externalities, *i.e.* firm-external economies but internal to their regions of localization. Simultaneously, dynamic intra-industrial economies -within the same activity- or inter-industrial -among different activities- make likely the presence of these external effects; see *e.g.* Fujita, Krugman and Venables (2001), Martin and Ottaviano (2001), Fujita and Thisse (2002), Rossi-Hansberg (2005), Combes, Mayer and Thisse (2008), Duranton, Martin, Mayer and Mayneris (2010), among others.

This paper focuses the attention on the relative industrial concentration of each activity separately, measured through the discrepancy between the activity specific distribution of the regions (*i.e.* the distribution of regions conditional on the activity), and the overall distribution of the regions (*i.e.* the marginal distribution of the regions); detail on this concept in *e.g.* Haedo and Mouchart (2012). We are not looking for the detection of one particular agglomeration, we rather look for a cluster scheme, *i.e.* a finite partition of regions such that each cluster is an agglomeration (thus a cluster of contiguous regions, possibly a single region) and such that the agglomerations are the most contrasted with respect to their industrial concentration. Here, the concept of specialized agglomeration refers to the homogeneity, within the agglomeration, of the regions relatively to industrial concentration and to the heterogeneity between the agglomerations; thus for a particular activity, some agglomeration are over-specialized, other ones are under-specialized.

The identification of specialized agglomerations has then some of the following motivations: i) the hypotheses regarding the nature of increasing returns; ii) the policies for economic growth and development, that often involves agglomerations and productivity; and iii) an understanding of regional inequalities, that matters for the same reasons as do inequalities across people and across nations; for more on this issue see *e.g.* Quah (2002), Quah and Simpson (2003).

This paper makes a dual contribution to the field of agglomeration detection. Firstly we develop a statistical model for the formation of agglomerations. We rely on a structural modeling approach; on this approach see *e.g.* Mouchart, Russo and Wunsch (2009). In particular, we endeavour to structure the model around distributions characterized by parameters meaningful from a spatial economics point of view. Based on a three-step structure of individual multiple choices, the formation of agglomerations is accordingly modeled through a multinomial distribution, where each agglomeration is characterized by a probability of individual location that is different among agglomerations and is depending on the activity. Secondly, we develop a greedy algorithm for the construction of a cluster scheme that is coherent with the statistical model just developed and that relies on a natural concept of specialization. This algorithm is initialized by postulating a null hypothesis of no specialized agglomeration, *i.e.* an initial cluster scheme of a unique cluster made of all regions. The algorithm

proceeds by adding one-by-one contiguous regions to find the most significant clusters. To do so, we compute the statistical significance (p-value) of potential clusters through a model selection criterion based on the highest values of BIC on the null hypothesis H_0 of no specialized agglomeration. A third step evaluates whether the contiguity property is significant: this is achieved by a bootstrapping method based on random permutation. Finally, the algorithm is run on a set of Argentinean data. The paper is concluded by evaluating the achievements of the proposed methodology. This paper started from and considerably extends the third chapter of Haedo (2009).

Spatial cluster detection is an important issue in many fields, ranging for instance from astrophysics to forest ecology, including epidemiology and many others. There is accordingly quite a vast and rapidly expanding literature. In particular, Kulldorff (1999), Lawson, Biggeri, Böhning, Lesaffre, Viel and Bertollini (1999), Wakefield, Kelsall and Morris (2000), Lawson (2001) and many of his later contributions have had a substantial influence on this field. Let us try to to suggest how this paper may be situated in this literature. Although this paper is basically motivated by economic geography (or, spatial economics), a wider view may be useful in order to better grasp the role and relevance of specific hypotheses, in particular by comparing what has been achieved in other contributions.

Important families of spatial cluster detection might be pointed out, in particular density-based clustering method and scan statistic. About the first family, two of the most well known are DBSCAN propose by Ester, Kriegel, Sander, Xu (1996), and CLIQUE proposed by Agrawal, Gehrke, Gunopulus, and Raghavan (1998). Each of these works by finding small dense regions and aggregating these high-density regions together in a bottom-up strategy. Han, Kamber, and Tunget (2001) and Neil (2006) provide a survey of these and other clustering methods. In this paper, our algorithm is closely related to the greedy heuristic search to locate dense regions proposed by Friedman and Fisher (1999) in the density-based clustering method context. The "bump hunting" iteratively removes or adds some portions of the data in such a way that density is maximized. They work in a discretized continuous space, with the basic data in geo-localized form, whereas we work in a strictly discrete space, with areal or lattice data.

In the second family, the approach is closely related to the spatial scan statistic proposed by Openshaw, Craft and Birch (1988), Besag and Newell (1991), Kuldorff and Nagarwalla (1995), Kuldorff (1997), Kulldorff, Tango and Park (2003), Gangnon and Clayton (2003), Neill, Moore and Cooper (2006), that have been used to detect disease cluster in epidemiology, and in Huang, Kulldorff and Gregorio (2007), and Cook, Gold and Li (2007) for survival data. These procedures operate on a continuous space and use moving circular or rectangular windows, scanning the complete space. Spatial location of regions will be done by their centroids, which mark the centre of the total area. These centroids are usually not taken as the geometrical centre, but are weighted by the actual population location within the region. In contrast, this paper works on a fixed set of polygons. The

spatial scan performs tests for a large set of overlapping spatial regions, and this overlap creates a complex dependency structure for the multiple tests to compute the statistical significance of each selected region at each step. In this paper, model selection procedure, based on the BIC criterion, is used in the algorithm for building the clusters. In most papers, significance testing is concerned with a null hypothesis of spatial uniformity (in order to detect agglomerations) whereas in this paper we are concerned with the significance of the contiguity condition, hence the reliance on permutation bootstrapping. Neill, Moore and Cooper (2006) compute the statistical significance in a Bayesian setting for space-time cluster through the posterior probability of each cluster, and find regions where an alternative hypothesis is likely or has low posterior probability. They also use moving rectangular windows with a specific stopping rule based on score bounds.

Another family of procedures is based on local decompositions of global autocorrelation measures, such as Moran's I coefficient decomposed into a Local Indicator of Spatial Association (LISA) proposed for Anselin (1995). In the case of a one-dimensional continuous space, this approach is contrasted with a non-parametric estimation of the intensity of a non-homogenous Poisson process applied in a problem of road accidents in Flahaut, Mouchart, San Martin and Thomas (2003). With auxiliary information, LISA is able to locate spatial associations, such as hot spots and cool spots (Sokal, Oden and Thomson 1998). The model-based of Zhang and Lin (2007) test may complement LISA, able also to account for heterogeneous population sizes. A residual Moran's I test for spatial autocorrelation can also be performed to detect spatial clustering for unexplained variations (Cliff and Ord 1981).

2 Selection of a best specialized agglomerations scheme

In this section, we expose a structural model of agglomeration formation and an algorithm used to identify a best cluster scheme of specialized agglomerations. We start by describing the data base underlying the algorithm along with some notational device, firstly at the level of the regions, next at the level of the clusters of regions.

2.1 Structure of the data and notation for the regions

For a given country, let us consider a finite set of administrative regions $i \in \mathcal{I} = \{1, ..., I\}$, and a finite set of activities $j \in \mathcal{J} = \{1, ..., J\}$. The administrative nature of the regions refers to two aspects. Firstly, the number of regions is finite by necessity. Secondly, the boundaries of the regions are designed exogenously, *i.e.* independently of the problem under analysis; in particular the areas of the different regions are typically quite heterogeneous. Furthermore, this administrative nature is crucial for the availability of the data.

For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, we observe the number N_{ij} of primary units; these could be typicallly

a number of employees or a number of firms. Thus we obtain a two-way $I \times J$ contingency table $\mathbf{N} = [N_{ij}]$ that also produces row, column and table totals denoted as follows:

$$N_{i\cdot} = \sum_{j=1}^{J} N_{ij}; \qquad N_{\cdot j} = \sum_{i=1}^{I} N_{ij}; \qquad N_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij} = \sum_{j=1}^{J} N_{\cdot j} = \sum_{i=1}^{I} N_{i\cdot}. \quad (1)$$

This paper is concerned with the relative industrial concentration, *i.e.* the concentration of one activitiy in some regions. Following Haedo and Mouchart (2012), the basic tools for these analyses are derived from the profiles provided by the contingency table $\mathbf{N} = [N_{ij}]$; more explicitly:

- region i may be characterized by the profile (or conditional distribution) of the i-th row:

$$p_{\vec{j}|i} = (p_{1|i}, \cdots, p_{j|i}, \cdots, p_{J|i}) \qquad p_{j|i} = \frac{N_{ij}}{N_{i\cdot}} \quad (2)$$

  to be compared with the global row profile (or marginal distribution):

$$p_{\cdot\vec{j}} = (p_{\cdot 1}, \cdots, p_{\cdot j}, \cdots, p_{\cdot J}) \qquad p_{\cdot j} = \frac{N_{\cdot j}}{N_{\cdot\cdot}} \quad (3)$$

- similarly, activity j may be characterized by the profile (or conditional distribution) of the j-th column:

$$p_{\vec{i}|j} = (p_{1|j}, \cdots, p_{i|j}, \cdots, p_{I|j}) \qquad p_{i|j} = \frac{N_{ij}}{N_{\cdot j}} \quad (4)$$

  to be compared with the global column profile (or marginal distribution):

$$p_{\vec{i}\cdot} = (p_{1\cdot}, \cdots, p_{i\cdot}, \cdots, p_{I\cdot}) \qquad p_{i\cdot} = \frac{N_{i\cdot}}{N_{\cdot\cdot}} \quad (5)$$

It may be convenient to refer to the areas rather than to the (arbitrary) labels of the regions. Thus we also denote the country, considered as an area, as Ω and the disjoint regions as Ω_i. Evidently, the regions Ω_i provide a partition of the country Ω:

$$\Omega_i \neq \emptyset \qquad \Omega_i \cap \Omega_{i'} = \emptyset \ (i \neq i') \qquad \bigcup_{i=1}^{I} \Omega_i = \Omega \quad (6)$$

We accordingly define:

$$p_{i\cdot} = p(\Omega_i) \quad (7)$$

2.2 Notation for clusters of regions

Let us operate a partition of the I regions into M "grouped regions", to be called "g-regions" for the ease of exposition. This regrouping may be written in terms of the labels:

$$\mathcal{I} = \{1, 2, \cdots, I\} = \bigcup_{m=1}^{M} \mathcal{I}_m \qquad \mathcal{I}_m \cap \mathcal{I}_{m'} = \emptyset \ (m \neq m') \qquad \#(\mathcal{I}_m) = I_m \qquad \sum_m I_m = I \quad (8)$$

Let us define accordingly

$$N_{m\cdot} = \sum_{i \in \mathcal{I}_m} N_{i\cdot} \qquad N_{m,j} = \sum_{i \in \mathcal{I}_m} N_{ij} \quad (9)$$

Using g to denote relative frequencies on the space of the g-regions, we successively define:

$$g_{m\cdot} \;=\; \sum_{i\in\mathcal{I}_m} p_{i\cdot} = \frac{N_{m\cdot}}{N_{\cdot\cdot}} \qquad g_{m|j} \;=\; \sum_{i\in\mathcal{I}_m} p_{i|j} = \frac{N_{m,j}}{N_{\cdot j}} \tag{10}$$

$$g_{\vec{m}\cdot} \;=\; (g_{1\cdot},\cdots,g_{m\cdot},\cdots,g_{M\cdot}) \qquad g_{\vec{m}|j} \;=\; (g_{1|j},\cdots,g_{m|j},\cdots,g_{M|j}) \tag{11}$$

$$p_{i\cdot|m} \;=\; \frac{p_{i\cdot}}{g_{m\cdot}}\;\mathbb{1}_{\{i\in\mathcal{I}_m\}} = \frac{N_{i\cdot}}{N_{m\cdot}}\;\mathbb{1}_{\{i\in\mathcal{I}_m\}} \qquad p_{i|j,m} \;=\; \frac{p_{i|j}}{g_{m|j}}\;\mathbb{1}_{\{i\in\mathcal{I}_m\}} = \frac{N_{ij}}{N_{m,j}}\;\mathbb{1}_{\{i\in\mathcal{I}_m\}} \tag{12}$$

This regrouping may also be viewed in terms of a cluster scheme of areas, $C = \{\mathcal{I}_{1|C},\cdots,\mathcal{I}_{m|C},\cdots,\mathcal{I}_{M|C}\}$, consisting of disjoint regional clusters:

$$\Omega_{\mathcal{I}_{m|C}} \;=\; \bigcup_{i\in\mathcal{I}_{m|C}} \Omega_i \qquad \text{with} \;\; \bigcup_{m=1}^{M} \Omega_{\mathcal{I}_{m|C}} \;=\; \Omega \tag{13}$$

Thus, when we want to make explicit the role of a particular cluster scheme C, we also write, instead of $g_{m\cdot}$:

$$g_{m\cdot|C} \;=\; g(\Omega_{\mathcal{I}_{m|C}}) = \sum_{i\in\mathcal{I}_{m|C}} p_{i\cdot}$$

Two extremes cases should be noticed:

a) $M = I$: no grouping; *i.e.* each region remains isolated, more explicitly:

$$C_{min} = \{\{1\},\{2\},\cdots,\{I\}\} \qquad \mathcal{I}_{m|C_{min}} = \{i\} \qquad m \longleftrightarrow i \tag{14}$$

$$g_{\vec{m}\cdot|C_{min}} = p_{\vec{i}\cdot} \qquad g_{\vec{m}|j;C_{min}} = p_{\vec{i}|j} \qquad p_{i\cdot|m;C_{min}} = \mathbb{1}_{\{\{i\}=\mathcal{I}_{m|C_{min}}\}} \tag{15}$$

$$p_{i|j,m;C_{min}} = \mathbb{1}_{\{\{i\}=\mathcal{I}_{m|C_{min}}\}}\,\mathbb{1}_{\{N_{ij}>0\}} \tag{16}$$

b) $M = 1$: maximal grouping, *i.e.* each region enters a unique g-region, more explicitly:

$$C_{max} = \{1,2,\cdots,I\} \qquad \mathcal{I}_{m|C_{max}} = \mathcal{I} \qquad m \equiv 1 \tag{17}$$

$$g_{\vec{m}\cdot|C_{max}} = g_{\vec{m}|j;C_{max}} = \{1\} \qquad p_{i\cdot|m;C_{max}} = p_{i\cdot} \qquad p_{i|j,m;C_{max}} = p_{i|j} \tag{18}$$

2.3 The multinomial model

We now develop a model in terms of the $N_{\cdot\cdot}$ primary units labeled by u, *i.e.* $u \in \mathcal{U}$ with $\#(\mathcal{U}) = N_{\cdot\cdot}$ along with a localization function $\ell : \mathcal{U} \to \Omega$ where $\ell(u)$ stands for the localization of u in Ω and an activity function $a : \mathcal{U} \to \mathcal{J}$ where $a(u)$ stands for the activity of the primary unit u. For each pair (i,j), the primary unit u is associated with a binary variable:

$$x_{ij}^u = \mathbb{1}_{\{\ell(u)\in\Omega_i,\, a(u)=j\}}. \tag{19}$$

Clearly:

$$\sum_{u \in \mathcal{U}} x_{ij}^u = N_{ij} \tag{20}$$

The algorithm developed in Section 2.5 provides a flexibility to adjust the regions to be taken into account when modeling a particular activity j; thus, for a given j it may be specified that only the regions with $N_{ij} > 0$ or only the regions with N_{ij} larger than some pre-specified limit (*e.g.* 5) will be the object of modeling. For a particular activity j we eventually have I_j regions to be taken into account and we define an $A_j \times I_j$ matrix $X_j = [x_{ij}^u]$ where $i = 1, \cdots, I_j$ in columns (up to an eventual reordering of the regions, putting first the relevant ones) and u, in rows, runs over the set $\mathcal{A}_j$ of the primary units entering the relevant regions for the analysis of the sector of activity j with $A_j = \#(\mathcal{A}_j)$. As X_j is an incidence matrix with elements equal to x_{ij}^u for $u \in \mathcal{A}_j$, the sum of each row is equal to 1.

We now introduce a model aimed at representing how the data have been generated and eventually may be interpreted; the notation now distinguishes unknown parameters, in Greek letters, and functions of data (estimators or statistics) in Latin letters although we also use Greek letters with hat for estimators. We start with an arbitrary cluster scheme $C = \{\mathcal{I}_{1|C}, \cdots, \mathcal{I}_{m|C}, \cdots, \mathcal{I}_{M|C}\}$.

The stochastic model accordingly involves three categorical random elements, namely: regions (i), activity (j) and cluster (m). When drawing randomly a statistical unit u from the universe $\mathcal{U}$, we therefore need to specify a trivariate distribution $\pi_{i,j,m}$. The process is decomposed as follows:

1. individual u selects an activity j according to a distribution $\pi_{\cdot j \cdot}$;
2. conditionally on j, individual u selects a cluster m according to a distribution $\gamma_{m|j;C}$;
3. conditionally on (j, m), individual u selects a region Ω_i within cluster $\Omega_{\mathcal{I}_{m|C}}$ according to a distribution $\rho_{i|j,m;C}$.

In short, we consider as structural the following decomposition:

$$\pi_{i,j,m|C} = \pi_{\cdot j} \;\; \gamma_{m|j;C} \;\; \rho_{i|j,m;C} \qquad\qquad i \in \mathcal{I}_{m|C} \tag{21}$$

Notice that, from (13) we have

$$\gamma_{m|j;C} = \sum_{i \in \mathcal{I}_{m|C}} \pi_{i|j} \qquad\qquad \gamma_{m|C} = \sum_{i \in \mathcal{I}_{m|C}} \pi_{i\cdot\cdot} \tag{22}$$

Remark. In the limit case of a one-term partition, *i.e.* $M = 1$ and $C = \{\mathcal{I}\}$, we have: $\gamma_{m|j;C} = \gamma_{m|C} = 1$, $\rho_{i|j,m;C} = \pi_{i|j}$ and (21) boils down to $\pi_{i,j} = \pi_{\cdot j} \, \pi_{i|j}$. ■

In next subsection we design a procedure for identifying specialized agglomeration, by means of a cluster scheme C different for each activity j. Therefore, we do not discuss the specification of $\pi_{\cdot j}$

and conduct the whole analysis conditionally on j. The concept of specialized agglomeration is to be built progressively. As a first step, we use, as an hypothesis maintained throughout the paper:

$$H_m : \qquad \rho_{i|j,m;C} = \rho_{i|m;C} \quad \text{with} \sum_{i \in \mathcal{I}_{m|C}} \rho_{i|m;C} = 1 \tag{23}$$

This hypothesis of conditional independence, namely $i \perp\!\!\!\perp j \mid m; C$, means that once an individual has selected an activity j, he selects a cluster m likely to be suitable for his activity j and when, conditionally on his choice (j, m), he selects a region, within the cluster m, he considers that *within* $\Omega_{\mathcal{I}_{m|C}}$ the regions exert an attraction independent of his sector of activity. Thus cluster m is considered as specialized as long as it is selected conditionally on j *and* that the area within $\Omega_{\mathcal{I}_{m|C}}$ is homogeneous as far as attractivity relative to activity j is concerned. In next subsection, we pursue making the concept of specialized agglomeration more explicit. Moreover, because:

$$\rho_{i|m;C} = \frac{\pi_{i\cdot\cdot}}{\gamma_{m|C}}, \tag{24}$$

the maintained hypothesis also assumes that the attractivity of region i, within cluster m, only depends of its general (or marginal) size $\pi_{i\cdot\cdot}$, relatively to the cluster size, $\gamma_{m|C}$. This hypothesis is built within a multinomial setting and is different from Mori and Smith (2011) condition of "Complete spatial randomness" assuming that (our notation) $\rho_{i|m;C}$ is equal to the ratio between the area of Ω_i and the area of $\Omega_{\mathcal{I}_m}$. Thus in (24) all the activities are taken into account through $\pi_{i\cdot\cdot} = \sum_j \pi_{ij\cdot}$; in other words the role of $\pi_{i\cdot\cdot}$ in $\rho_{i|m;C}$ is to provide a proxy for the set of characteristics of region i being favorable to the development of the activity in general. From now on, we explicitly write that the number of clusters depends on the cluster scheme under consideration; thus we shall write $M(C)$ instead of M.

Under our maintained hypothesis we have:

$$\pi_{i|j,m;C} = \gamma_{m|j;C}\ \rho_{i|m;C} \qquad \forall i \in \mathcal{I}_{m|C} \tag{25}$$

and, for a given j, the probability of the data matrix X_j may be written as

$$\begin{aligned} p(X_j \mid C) &= \prod_{u \in \mathcal{U}} \prod_{1 \le m \le M(C)} \prod_{i \in \mathcal{I}_{m|C}} \pi_{i|j,m;C}^{x_{ij}^u} && (26)\\ &= \prod_{1 \le m \le M(C)} \prod_{i \in \mathcal{I}_{m|C}} [\gamma_{m|j;C}\ \rho_{i|m;C}]^{N_{ij}} && (27)\\ &= \left[\prod_{1 \le m \le M(C)} \gamma_{m|j;C}^{N_{m,j|C}} \right] \cdot [\, b(X_j) \,] && (28) \end{aligned}$$

where

$$\begin{aligned} N_{m,j|C} &= \sum_{i \in \mathcal{I}_{m|C}} N_{ij} && (29)\\ b(X_j) &= \prod_{1 \le m \le M(C)} \prod_{i \in \mathcal{I}_{m|C}} \rho_{i|m;C}^{N_{ij}} && (30) \end{aligned}$$

As the parameter of the factor $b(X_j)$ does not depend on j, we may factorize the likelihood function as

$$L(X_j \mid C) = L_1(\gamma_{\vec{m}|j;C} \mid X_j)\ L_2([\rho_{i|m;C}] \mid X_j) \tag{31}$$

2.4 The concept of specialized cluster

The previous section proposes a modeling for the formation of an arbitrary cluster scheme of regions. In this section, we want to identify specialized clusters relatively to a specified activity j and propose a measure of the intensity of the specialization of a country for the specified activity j. We first make operational the concept of cluster specialization.

The starting point of our analysis is the well established (estimated) *Hoover-Balassa Local Quotient* for the cell (i, j) that may be written indifferently in several forms, as follows:

$$LQ_{ij} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i.}\ \hat{\pi}_{.j}} = \frac{\hat{\pi}_{j|i}}{\hat{\pi}_{.j}} = \frac{\hat{\pi}_{i|j}}{\hat{\pi}_{i.}} = \frac{N_{ij}\ N_{..}}{N_{i.}\ N_{.j}} \tag{32}$$

This local quotient reveals the following feature of activity j in region i:

$$\begin{aligned} LQ_{ij} \quad &= 1 \quad \text{non-specialization} \\ &> 1 \quad \text{over-specialization} \\ &< 1 \quad \text{under-specialization} \end{aligned} \tag{33}$$

where the "non-specialization" corresponds *locally* - *i.e.* at the level of the cell (i, j)- to the row-column independence. Indeed, the last two equalities in (32) stress the point that specialization is an issue concerning the global structure at the country level: the absence of specialization of a cell (i, j) means that, *relatively to the marginal distributions*, activity j is not over-(nor under-) represented in region i (*i.e.* there is locally no relative regional specialization) *and* that region i is not over-(nor under-) represented for the activity j (*i.e.* there is locally no relative industrial concentration). On the duality between regional specialization and industrial concentration see Haedo and Mouchart (2012, Table 1). Thus, "local" points to the fact that LQ is localized in a cell (i, j).

At the level of regions, the concept of industrial concentration for a given activity j compares the conditional distributions of the regions $\pi_{\vec{i}|j}$ for activity j with a benchmark distribution of the regions, say $b_{\vec{i}} = (b_1, \cdots, b_i, \cdots, b_I)$, and is measured by means of some discrepancy $d(\pi_{\vec{i}|j} \mid b_{\vec{i}})$.

Remark. As indicated, with more detail in Haedo and Mouchart (2012), the absolute industrial concentration takes a uniform distribution, *i.e.* $b_i = \frac{1}{I}\ \forall i$, as a benchmark and therefore considers that a not concentrated activity would be spread uniformally over the regions. The relative industrial concentration takes the marginal distribution $\pi_{\vec{i}.}$ as a benchmark, *i.e.* $b_i = \pi_{i.}$. In that case, the marginal distribution is taken as an indicator of the global attractivity of a region and a non-concentrated activity would the spread over the regions according to the same marginal distribution $\pi_{\vec{i}.}$. Other benchmark distributions might be and have been considered, in particular: $b_i = \frac{A_i}{\sum_i A_i}$

where A_i is the area of region i (see *e.g.* Mori and Smith 2011) or $b_i = \frac{P_i}{\sum_i P_i}$ where P_i is the population of region i (see *e.g.* Lawson, Biggeri, Böhning, Lesaffre, Viel and Bertollini 1999, Lawson 2001, among others). Note that in epidemiology, the population P_i may represent the population at risk for a given disease, in which case $b_i = \frac{P_i}{\sum_i P_i}$ may be a reasonable benchmark for defining an epidemic concentration. ■

Now we want to identify specialized clusters relatively to a specified activity j. Here a cluster $\mathcal{I}_m$ is over-specialized (resp. under-specialized) with respect to activity j when the $\pi_{i|j}$'s for $i \in \mathcal{I}_m$ are significantly greater (resp. smaller) than the country-wide average $\pi_{i\cdot}$, taken as a benchmark (remember that $\pi_{i\cdot}$ is an average of the $\pi_{i|j}$'s, *i.e.* $\pi_{i\cdot} = \sum_j \pi_{i|j}\ \pi_{\cdot j}$) and in view of (32) this is equivalent to the local quotients being significantly larger, or smaller, than 1. Under the maintained hypothesis (23), it may be seen from (26)-(28) that the identification of specialized clusters and the construction of a cluster scheme C of specialized clusters is to be based on the properties of $\gamma_{m|j;C}$.

A (fully) non-specialized cluster $\mathcal{I}_m$ is a cluster where $LQ_{ij} = 1$ for $\forall i \in \mathcal{I}_m$ (equivalently, $\pi_{i|j} = \pi_{i\cdot}$ or $\pi_{j|i} = \pi_{\cdot j}$). This hypothesis, extended to each cluster m, *i.e.* $LQ_{ij} = 1\ \forall i \in \mathcal{I}$, implies, because of (22):

$$H_0 : \qquad \gamma^{(0)}_{m|j;C} = \gamma^{(0)}_{m|C} \qquad m = 1, \cdots, M \tag{34}$$

This hypothesis means that for the activity j and for the cluster scheme C, there is no industrial concentration on the whole country. The maximum likelihood estimation under H_0 is therefore:

$$\widehat{\gamma}^{(0)}_{m|j;C} = \widehat{\gamma}^{(0)}_{m|C} = \frac{N_{m\cdot|C}}{N_{\cdot\cdot}} \tag{35}$$

The absence of industrial concentration for activity j, underlying (35), is relative to a particular cluster scheme C. The estimated log likelihood under H_0 is

$$\ln \widehat{L}^{(0)}(X_j \mid C) = \sum_{1 \le m \le M(C)} N_{m,j|C}\ \ln\left(\frac{N_{m\cdot|C}}{N_{\cdot\cdot}}\right) + \ln a(X_j) \tag{36}$$

where $a(X_j)$ corresponds to the term $L_2([\rho_{i|m;C}] \mid X_j)$ in (31) and gathers the terms unaffected by the null hypothesis. As an alternative hypothesis H_1, the parameter $\gamma_{m|j;C}$ is left unconstrained and maybe estimated as

$$\widehat{\gamma}^{(1)}_{m|j;C} = \frac{N_{m,j|C}}{N_{\cdot j}} \tag{37}$$

Thus when the alternative hypothesis is assumed for each cluster $\mathcal{I}_m$ of a cluster scheme C, the estimated log likelihood under H_1 is

$$\ln \widehat{L}^{(1)}(X_j \mid C) = \sum_{1 \le m \le M(C)} N_{m,j|C}\ \ln\left(\frac{N_{m,j|C}}{N_{\cdot j}}\right) + \ln a(X_j) \tag{38}$$

The log-likelihood-ratio statistic to test H_0 against H_1, under H_m is

$$
\begin{aligned}
T(X_j \mid C) &= -2 \left[\ln \widehat{L}^{(0)}(X_j \mid C) - \ln \widehat{L}^{(1)}(X_j \mid C) \right] && (39)\\
&= -2 \left[\sum_{1 \leq m \leq M(C)} N_{m,j|C} \ \ln \left(\frac{N_{m\cdot|C}}{N_{\cdot\cdot}} \right) - \sum_{1 \leq m \leq M(C)} N_{m,j|C} \ \ln \left(\frac{N_{m,j|C}}{N_{\cdot j}} \right) \right] && (40)\\
&= 2 \left[\sum_{1 \leq m \leq M(C)} N_{m,j|C} \ \ln \left(\frac{N_{m,j|C}/N_{\cdot j}}{N_{m\cdot|C}/N_{\cdot\cdot}} \right) \right] && (41)\\
&= 2 \left[N_{\cdot j} \sum_{1 \leq m \leq M(C)} p_{m\cdot|j} \ \ln \left(\frac{p_{m\cdot|j}}{g_{m\cdot}} \right) \right] && (42)\\
&= 2 \ N_{\cdot j} \ d(p_{\vec{m}|j} \mid g_{\vec{m}\cdot}) && (43)
\end{aligned}
$$

where $d(\cdot \mid \cdot)$ is the (non-symmetric) Kullback-Leibler divergence between two distributions. Therefore, the test statistic (43) may be viewed as a measure of relative industrial concentration of activity j among the clusters (more on this concept in Haedo and Mouchart 2012).

The argument of the logarithms in (41) is the local quotient obtained after clustering the regions; more explicitly it is the cross product ratio in the 2×2-table:

$$
\begin{bmatrix}
N_{m,j|C} & N_{m\cdot|C} \\
N_{\cdot j} & N_{\cdot\cdot}
\end{bmatrix}
$$

From (32), it is seen that the argument of the logarithms in (42) is also equal to $p_{\cdot j|m}/p_{\cdot j}$ and stresses the fact that the local quotient for cluster m and activity j is independent of the area of cluster m and of the weight $g_{m\cdot}$ of the cluster m; this is at variance from Mori and Smith (2011).

In (34), H_0 represents $M(C) - 1$ restrictions for a fixed j under the condition that the sum in m is equal to 1. Therefore under H_0, the test statistics $T(X_j \mid C)$ is asymptotically distributed as a chi-square distribution with $M(C) - 1$ degrees of freedom. The asymptotic p-value for this likelihood-ratio test is given by

$$
p - value = 1 - F_{M(C)-1}(T(X_j \mid C)) \tag{44}
$$

where $F_{M(C)-1}$ denotes the cumulative distribution function for the chi-square distribution with $M(C) - 1$ degrees of freedom. The null hypothesis is rejected when the value of the likelihood-ratio test is sufficiently large, or when the corresponding p-value is sufficiently small.

2.5 Detecting a specialized agglomerations scheme

Up to now, we have developed a "space free" analysis as long as the labels of the region is arbitrary and convey no information on the localization of the regions. Now we introduce an idea of distance-based pattern by means of the concept of agglomerations that are clusters made of neighboring regions. The simplest case is obtained when neighboring regions is interpreted as contiguous regions.

In that case, only regrouping contiguous regions is of interest; therefore each $\Omega_{\mathcal{I}_{m|C}}$ should be a connected set of regions. The contiguity matrix (or weights matrix) W formally expresses the proximity links existing between all pair of regions.

The elements of the $I \times I$-matrix W are obtained as the values of the following function:

$$w : \mathcal{I} \times \mathcal{I} \longrightarrow \{0, 1\} \qquad \text{where} \quad w(i_1, i_2) = \mathbb{1}_{\{i_1 \text{ and } i_2 \text{ are contiguous}\}} \tag{45}$$

Note that W is symmetric ($W = W'$) with 1's on the main diagonal and the sum of the rows (or, of the columns) minus 1 is equal to the number of contiguous regions of each region in the set $\mathcal{I}$. Therefore, the set of regions contiguous to a cluster $\mathcal{I}_{m|C}$ may be written as:

$$v(\mathcal{I}_{m|C}) = \{i_1 \in \mathcal{I} \setminus \mathcal{I}_{m|C} \mid \exists\, i_2 \in \mathcal{I}_{m|C} : w(i_1, i_2) = 1\} \tag{46}$$

Remark on the W matrix. The concept of contiguity underlying (45) deserves to be made more precise. Contiguity may mean at least one point common in the boundaries of the two contiguous regions, in which case W is a first order queen weights matrix, or contiguity may mean a partly common frontier with more than one point, in which case W is a first order rook weights matrix; for more information on weights matrices see O'Sullivan and Unwin (2010). In this paper we choose the rook form. ■

Our final objective is to construct a cluster scheme C made of specialized agglomerations, that maximizes the heterogeneity among agglomerations up to a penalization on the number of parameters. In this context, a set of agglomerations is "best" (or, close to best) as long as the agglomerations are the most different possible for their specialization with respect to activity j. Here, the heterogeneity among clusters is measured by the divergence $d(p_{\vec{m}|j} \mid g_{\vec{m}\cdot})$, as given in equation (43).

It should be noticed that a particular cluster scheme generates a particular model; indeed, from (21), a cluster scheme C may be viewed as a model of agglomeration formation. Therefore, selecting a cluster scheme out of a set of possible schemes may be treated as a problem of model selection. As reminded in Appendix A, a natural model selection procedure may be based on the Bayesian information criterion (BIC), that naturally involves the divergence (43). Thus we consider the optimization problem:

$$C^* = \arg\max_C BIC(X_j \mid C) \tag{47}$$

where, in the *max* operation, C is running over all possible partitions of I into agglomerations (*i.e.* clusters of contiguous regions).

A sketch of the algorithm.

The number of possible cluster schemes can be enormous for an even modest number of basic regions. Thus, it is necessary to consider limited search procedures that yield reasonable approximations to best cluster schemes. Our approach is essentially an elaboration of the basic ideas

of the scan method proposed by Besag and Newell (1991) in which, given the set of basic regions, we start with individual regions and progressively add contiguous regions to find the most significant cluster, evaluating all possible cluster schemes that can be formed from these regions. This algorithm may also be viewed in the family of hierarchical divisive clustering algorithms (for an interesting synthesis of this topic, one may consult `http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html`, see also `http://nlp.stanford.edu/IR-book/html/htmledition/divisive-clustering-1.html`). Experience suggested that the divisive structure of this algorithm is likely to be suitable for taking due account of the constraint of contiguity.

For each activity j, we develop a greedy forward algorithm that uses the BIC as a selection criterion. We start with a baseline configuration $\mathcal{I}_{[j]}$ made of all regions with $N_{ij} > 0$; we might also specify N_{ij} larger than some specified minimum. For the sake of exposition only, we shall assume that $\mathcal{I}_{[j]} = \mathcal{I} \ \forall j$.

Remark. In some cases, the selection of individuals is made on those localized in regions where $LQ_{ij} > 1+\varepsilon$ or $LQ_{ij} < 1-\varepsilon$, for some $\varepsilon \geq 0$ (for fixed j), with the objective of analyzing separately over- or under- specialized agglomerations (see *e.g.* Neill, Moore and Cooper 2006). ■

The algorithm generates, as a first (myopic) version, a sequence of configuration $C^*_{[k]}$ as follows:

Step 0. As an initial step, $C^*_{[0]}$ is the one-term partition:

$$C^*_{[0]} = \{\mathcal{I}\} = C_{max} \tag{48}$$

In this case, $BIC(X_j \mid C^*_{[0]}) = T(X_j \mid C^*_{[0]}) = 0$, because $M(C) = 1$.

Step 1. The first step chooses the region which forms a separated one region cluster and that maximizes the value of $BIC(X_j \mid C)$. There are I possible regions to choose from. For each cluster scheme under consideration $M(C) = 2$ as the outcome of this first step is a two clusters configuration, namely one cluster formed by the chosen region, say $\{i^*_{[1]}\}$, and the other cluster consisting of all the remaining regions, $\mathcal{I} \setminus \{i^*_{[1]}\}$. This second cluster is a "residual" cluster in the sense that it is not composed of homogenous regions only, as far as specialization is concerned, but rather is to be used as a reservoir from which a new region will be extracted in the following step. For this cluster the condition of contiguity is not required. At the first step, the optimal configuration has accordingly the following form:

$$C^*_{[1]} = \{\{i^*_{[1]}\}, \mathcal{I} \setminus \{i^*_{[1]}\}\} \tag{49}$$

This step is completed by evaluating $BIC(X_j \mid C^*_{[1]})$. If, however, $\pi_{i|j} = \pi_{i\cdot} \ \forall i$ for a given j, then $C^*_{[1]} = C^*_{[0]}$ and the algorithm stops.

Step 2. The second step looks for a second region, say $i^*_{[2]}$. The algorithm starts by examining each region i_2 in $\mathcal{I} \setminus \{i^*_{[1]}\}$ and evaluates the BIC-criterion for the corresponding cluster

scheme $C_{[1],i_2}$ formed from $C^*_{[1]}$ and i_2 as follows. If $i_2 \in v(\{i^*_{[1]}\})$ - see (46)- , two cluster schemes are possible: either configuration a, namely $C_{[1],i_2,a} = \{\{i^*_{[1]}, i_2\},\ \mathcal{I} \setminus \{i^*_{[1]}, i_2\}\}$ (with $M(C) = 2$) or configuration b, namely $C_{[1],i_2,b} = \{\{i^*_{[1]}\}, \{i_2\},\ \mathcal{I} \setminus \{i^*_{[1]}, i_2\}\}$ (with $M(C) = 3$). The algorithm retains $C_{[1],i_2}$ as a configuration with the highest BIC. If $i_2 \notin v(\{i^*_{[1]}\})$, only $C_{[1],i_2} = \{\{i^*_{[1]}\}, \{i_2\},\ \mathcal{I} \setminus \{i^*_{[1]}, i_2\}\}$ is possible and $M(C) = 3$. The optimum $i^*_{[2]}$ and the corresponding optimum cluster scheme are now defined as:

$$i^*_{[2]} = \arg \max_{i_2 \in \mathcal{I} \setminus \{i^*_{[1]}\}} BIC(X_j \mid C_{[1],i_2}) \tag{50}$$

$$C^*_{[2]} = C_{[1],i^*_{[2]}} \tag{51}$$

Note that we may write $C^*_{[2]}$ as follows:

$$C^*_{[2]} = \{\mathcal{I}^*_{1|[2]}, \cdots, \mathcal{I}^*_{M(C^*_{[2]})|[2]}\} \tag{52}$$

where $M(C^*_{[2]})$, equal either to 2 or to 3, is the number of elements of the best cluster scheme at the end of step 2, and $\mathcal{I}^*_{M(C^*_{[2]})|[2]} = \mathcal{I} \setminus \{i^*_{[1]}, i^*_{[2]}\}$ is a residual cluster at the end of step 2. If the maximizer does not improve the BIC-criterion of the previous step, the algorithm stops.

Next steps. Each following step of the algorithm proceeds according to the same structure as that of step 2. More specifically, at the end of step k, the cluster scheme has the form

$$C^*_{[k]} = \{\mathcal{I}^*_{1|[k]}, \mathcal{I}^*_{2|[k]}, \cdots, \mathcal{I}^*_{M(C^*_{[k]})|[k]}\} \tag{53}$$

where $M(C^*_{[k]})$ is the number of agglomerations after k steps and $\mathcal{I}^*_{M(C^*_{[k]})|[k]}$ is a residual cluster, thus $M(C^*_{[k]}) \leq k$. Note that some agglomerations might be contiguous without being merged into a unique one because of too different values of the local quotients: this is decided through the BIC-criterion that balances the effect on the divergence against the penalization for the number of parameters, see equations (43) and (65).

Stopping rule. The algorithm stops when no choice of region from the residual cluster of the preceding step provides an increase in the criterion. If we write k^* for the last step before stopping the algorithm, we simplify the notation by writing C^* instead of $C^*_{[k^*]}$.

More on the stopping rule.

The proposed algorithm is a heuristic one: it economizes on a complete enumeration of all possible partitions but does not ensure that the absolute optimal selection has been achieved. It is also myopic because the algorithm stops as soon as the value of the criterion deteriorates for the first time. The performance of such an algorithm is to be evaluated from two criteria. Firstly, the proximity of the obtained solution with respect to the true optimal solution should be appreciated in terms of the value of the objective function (*i.e.* is $BIC(C^*)$ close to the BIC of the true optimum?) and in terms of the configuration of the final partition (*i.e.* is the partition C^* close to the true

optimum partition?). Secondly, once the number of regions becomes sizable, the computing time becomes an important issue, particularly in view of the simulations to be performed through the permutation bootstrap, to be exposed in the next subsection. These two criteria crucially depend on adopting a more or less sophisticated stopping rule. Neil (2006) proposed quite an interesting discussion on the elaboration of a stopping rule, in a however quite different analytical context. As a preliminary phase in the elaboration of the stopping rule we first let the algorithm run on simulated data till the exhaustion of the residual cluster (*i.e.* till $\mathcal{I}^*_{M(C^*_{[k]})|[k]}$ has a simple region) with a moderate number of regions (roughly, between 20 and 100). We observed that (i) the value of the optimum BIC as a function of the number of steps is roughly unimodal but with local peaks of minor importance (ii) once the optimum BIC substantially decreases, the number of agglomerations $M(C^*_{[k]})$ increases dramatically giving, in BIC, more weight to the penalization of the number of parameters (producing even negative values of the final BIC's) and producing agglomerations with too small numbers of regions for being significant. Thus stopping the algorithm once the value of BIC decreases for the first time is probably too myopic a rule.

In order to protect for a situation where a first local maximum be followed by higher values of BIC and to avoid a complete exhaustion of the residual cluster, we introduce a more sophisticated stopping rule as follows. We define "h", as a parameter of backward horizon *i.e.* , the number of past criterion values that enters the stopping rule; later we discuss the specification of h. Write, as before, k for the label of the step of the algorithm $k = 1, 2, \ldots$, and y_k for the value of the criterion at the end of step k. Once $k \geq h$, consider a linear trend of the last h realizations, *i.e.* evaluate a least squares linear trend of y on the $h+1$ past steps. The slope of this trend is proportional to the covariance $\frac{1}{h+1}\sum_{i=k-h}^{k}(y_i - \bar{y})(i - \bar{i})$. The algorithm stops once the slope becomes null or negative and the optimum values is obtained by examining the $h+1$ last values of y_k, namely

$$k^* = \arg\max_h \{y_k, y_{k-1}, \ldots, y_{k-h}\} \tag{54}$$

$$y^* = y_{k^*} = \max\{y_k, y_{k-1}, \ldots, y_{k-h}\} \tag{55}$$

The specification of the horizon h may require some trials. Our experience suggests that $h = 60$ is a reasonable choice, ensuring a fair protection against a local maximum followed by higher values of the criterion or against a situation where the objective function would have a very flat behavior for a sizeable number of steps, followed by a steeper ascent behavior.

A related issues is the penalization imposed by $M(C)$, the number of agglomerations in the cluster scheme. AIC weights $(M(C)-1)$ by a factor of -2 while BIC uses a factor of $-\ln(N_{\cdot\cdot})$ (see Appendix A for details). The motivation of BIC is to improve the asymptotic convergence toward a "true" model. This motivation is however irrelevant in the present case. Indeed, what would be the meaning of a "true" agglomeration or of a "false positive"? As a matter of fact we are looking for a "reasonable" cluster scheme with a number of agglomerations that is neither too high nor too

low and with agglomerations homogeneous as far as relative concentration is concerned. Thus we have in mind no structural model explaining the frontiers of agglomeration. If this modeling effort were the purpose of the research then the present algorithm might be viewed as an exploratory step for orienting the search for a reasonable model. Experiences show that basing the stopping rule on the p-value of $T(X_j \mid C)$ produces cluster schemes with too many agglomerations, many of them too small: there is a definite need for penalizing $M(C)$. The choice between AIC and BIC may depend both on I, the number of regions, and on $N_{..}$, the number of primary units. We recommended some explorations of this issue before embarking to the bootstrapping phase.

2.6 Testing for spurious specialized agglomerations schemes

Consider a cluster scheme that has been observed, or determined, for a given activity j for which there is some degree of industrial concentration, as measured by an index of relative industrial concentration of the form $d(p_{\vec{i}|j} \mid p_{i.})$. The question is to try to understand why the industrial concentration of activity j tends to cluster into some agglomerations. As a first step it is natural to ask whether the contiguity among regions is a significant factor of clustering into over- or under-specialized agglomerations. Formally, we want to test the null hypothesis that the vector of the local quotients, for a fixed activity j, is invariant for the group of permutations of its coordinates.

Permutation bootstrap is a standard methodology for facing such a question. Indeed, redistributions of $LQ's$ among all regions without replacement has been used when assessing spatial dependence between neighbouring regions, see *e.g.* Manly (1991), Zoellner and Schmidtmann (1999), Good (2000) and Lawson (2001). Each redistribution is simulated independently of the contiguities among regions, thus independently of the matrix W, and if for each simulation we run the algorithm and compute the optimal BIC, then we may appreciate where the optimal $BIC(X_j \mid C^*)$ is localized relatively with the distribution of the simulated BIC. In particular, one may decide that when $BIC(X_j \mid C^*)$ is far in the tail of the distribution of the simulated BIC, it is a signal that contiguity is a significant factor of clustering. The significance of $BIC(X_j \mid C^*)$ is accordingly evaluated thought the bootstrap distribution. More specifically, let us write $BIC^b(X_j \mid C_b^*)$ for the BIC of the optimal cluster scheme obtained as a result of the b-th simulation and $\widehat{F}_{BIC}^B$ for the empirical distribution function of $BIC^b(X_j \mid C_b^*)$ obtained after B simulations. The empirical p-value of $BIC(X_j \mid C^*)$ is therefore:

$$\begin{aligned} p-\text{value}\,[BIC(X_j \mid C^*)] &= 1 - \widehat{F}_{BIC}^B(BIC(X_j \mid C^*)) \\ &= \frac{1}{B}\sum_{1\leq b\leq B} \mathbb{1}_{\{(BIC^b(X_j|C_b^*)>BIC(X_j|C^*)\}} \end{aligned} \tag{56}$$

Thus the bootstrap p-value is, in general, simply the proportion of the bootstrap test statistics $BIC^b(X_j \mid C_b^*)$ that are more extreme than the observed test statistic $BIC(X_j \mid C^*)$: rejecting

the null hypothesis whenever p-value $[BIC(X_j \mid C^*)] < \alpha$ is equivalent to rejecting it whenever $BIC(X_j \mid C^*)$ exceeds the 1-α quantile of $\widehat{F}^B_{BIC}$.

3 Application to a pair of activities in Argentina

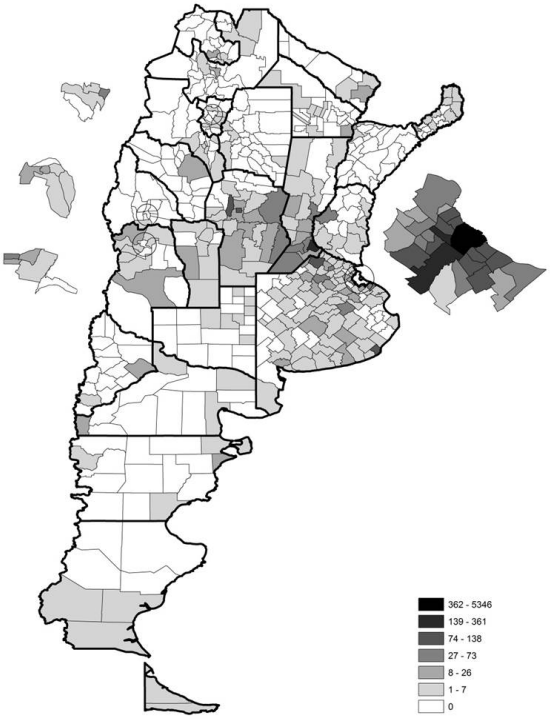
Let us now experiment this algorithm on Argentine data. The data base is made of an $I \times J$-contingency table $\mathbf{N} = [N_{ij}]$ where $I = 511$ is the number of regions (partidos/departamentos), $J = 23$ the number of activities and N_{ij} the number of firms of less than 200 employees; these data come from the Firms Directory of *Fundación Observatorio PyME* (small and medium firms) of Argentina for the year 2010, at the level of the first 2-digit (divisions) of the International Standard Industrial Classification of manufacturing activities (ISIC- Rev.3.1, items 15 to 37). We want to detect and compare the partitions into specialized agglomerations for two activities, namely Manufacture of wearing apparel, dressing and dyeing of fur ($j = 18$) and Manufacture of machinery and equipment n.e.c. ($j = 29$).

Figure 1 gathers some basic information on the industrial structure of the small and medium firms in Argentine. Above in the figure, two maps give the spatial distribution of the primary units for activities 18, distributed across 250 regions, and 29, distributed across 253 regions; below a map gives the same information for all the manufacturing activities. For these three maps, the intensity of darkness in each region is proportional to the number of primary units according to a discretized scale provided at the South-East of each map. Finally a table provides the same information in a numerical format with the regions (partidos/departamentos) aggregated into the 24 provinces. This table permits an evaluation of the relative weight of the two activities of interest with respect to the global economy. Thus in terms of small and medium firms activity 18 represents 8.3 % and activity 29 represents 4.5 %.

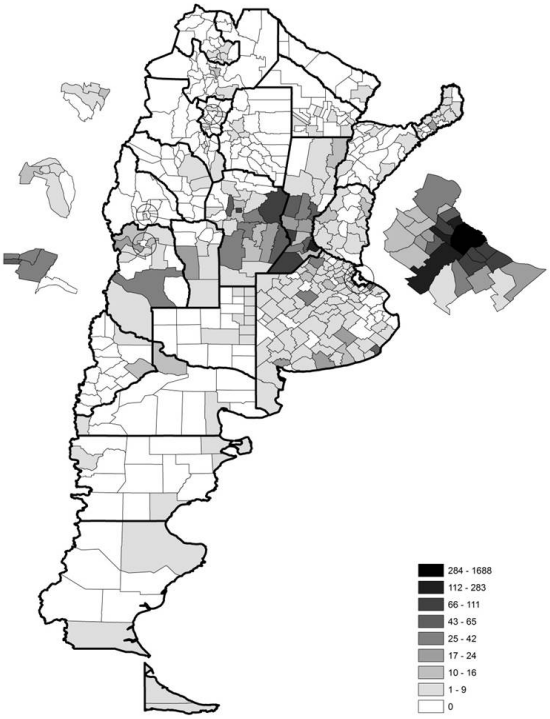
The results from the analysis of the two activities are provided in Figure 2, under a specification of the parameters of the algorithm as given in Appendix B. On the left-hand side is given a map of Argentine with the over-specialized agglomerations in dark and on the right-hand side some figures characterizing the numerical aspects of the algorithm. East-side of the maps of the country there is a map of the region around the city of Buenos Aires (the so-called GBA: Great Buenos Aires) at a lower scale than the maps of the country. Similarly, on the west side, lower scale maps are given for the regions around three cities, namely (from north to south) Tucumán, San Juan and Mendoza.

The best specialized agglomeration scheme shows that in GBA only 1 agglomeration with 5 regions is over-specialized in activity 29 but 3 agglomerations (two with 1 region and one with 8 regions) are over-specialized in activity 18. For the country in general the numbers of over-specialized agglomerations in activity 18 and in activity 29 are roughly the same, namely 6 and 7, but their spatial distribution is quite different. Indeed, activity 18 is mostly concentrated in the center of

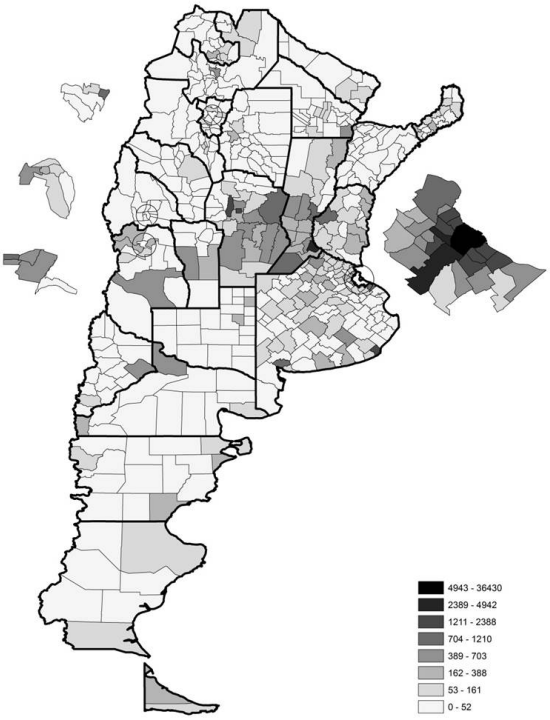
Activity j=18



Activity j=29



Total

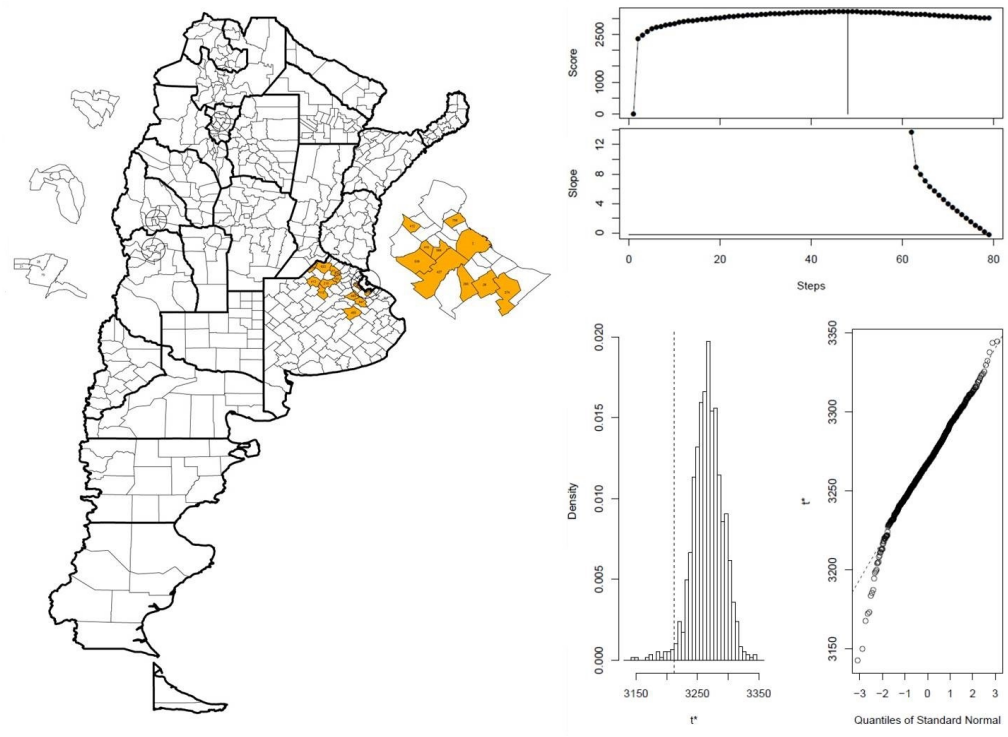


Total for provinces

Province	Primary units			Total
	Act18	Act29	Others	
Buenos Aires	2836	1722	35164	39722
Capital Federal	5346	1688	29396	36430
Catamarca	19	4	285	308
Chaco	30	13	959	1002
Chubut	22	14	781	817
Córdoba	426	517	8680	9623
Corrientes	16	11	635	662
Entre Ríos	47	54	2344	2445
Formosa	12	2	244	258
Jujuy	13	12	533	558
La Pampa	14	23	667	704
La Rioja	8	2	219	229
Mendoza	133	227	4081	4441
Misiones	25	46	1483	1554
Neuquén	13	13	686	712
Río Negro	19	20	915	954
Salta	26	18	876	920
San Juan	40	26	1127	1193
San Luis	30	24	762	816
Santa Cruz	6	4	340	350
Santa Fe	595	802	10263	11660
Santiago del Estero	9	4	558	571
Tierra del Fuego	8	6	258	272
Tucumán	46	25	1185	1256
Total	**9739**	**5277**	**102441**	**117457**

Figure 1: *Argentinean manufacture activities*

Activity j=18



Activity j=29

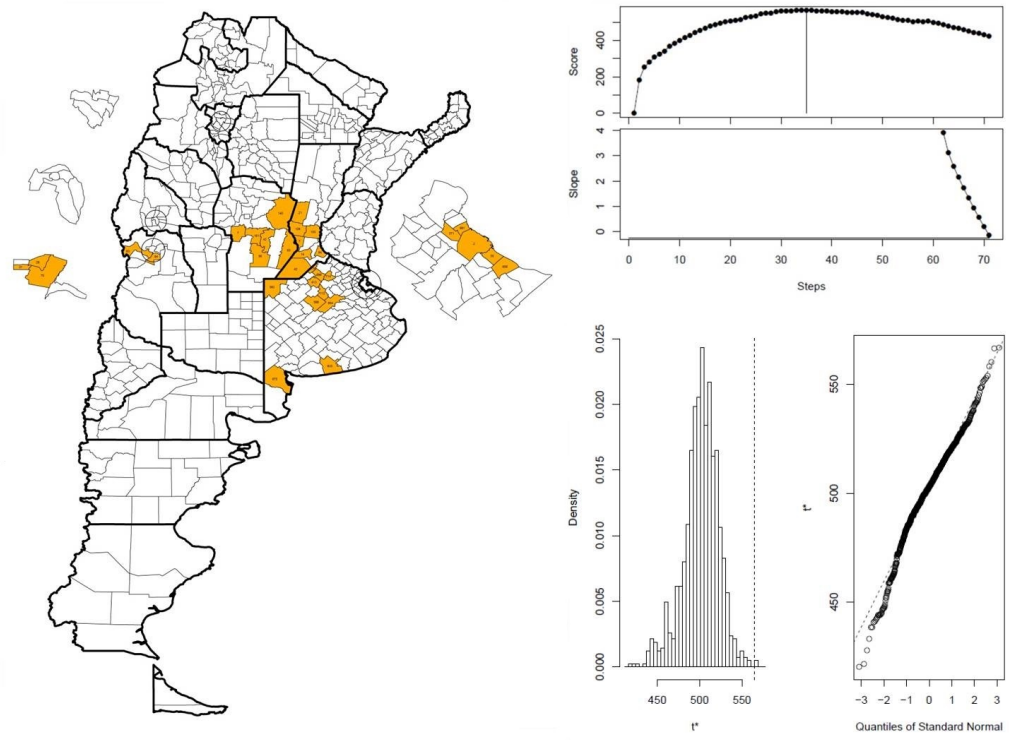


Figure 2: *Specialized agglomerations for two argentinean manufacture activities*

the country in 24 regions of the GBA and in the rest of the province of Buenos Aires whereas activity 29 is more spread over the country in 33 regions, located in the GBA, in the rest of the province of Buenos Aires and in the provinces of Santa Fe, Córdoba and Mendoza. Moreover, the over-specialized agglomerations of activity 18 gathers 67 % of firms and 65 % of employees of the whole activity 18 at a national level whereas activity 29 gathers 61 % of firms and 58 % of employees of the whole activity 29. These results show that in both activities the firms have a high tendency to co-localize in specialized agglomerations.

It is of interest to compare Figures 1 and 2. Indeed Figure 1 gives a picture of the agglomerations in terms of the *absolute* concentration of the activities whereas Figure 2 provides a similar picture in terms of the *relative* concentration of the activities.

This greedy algorithm is aimed at detecting specialized agglomerations the actual meaning of which depends, among other issues, on some characteristics of the numerical behavior of the algorithm. On the right-hand side of each map of Figure 2 are shown four figures illustrating, from North to South, the following aspects:

1. The evolution of the objective function, namely the BIC, at each step of the algorithm with a mark for its maximum value (vertical line). In both cases the curve is roughly unimodal although for activity 29 there are some local maxima. The maximum value of the objective function was found at step 48 (value 3212) and at step 35 (value 565) for activities 18 and 29 respectively.

2. The control for local *vs.* global maxima is operated by means of a window of $h = 60$ steps: from the 61-th step on, a least-squares slope is evaluated at each step for the last 60 steps and the algorithm is stopped once the slope change the sign (and therefore crosses the zero value). Once the algorithm is stopped the maximum is located within the last 60 values of the objective function. The second figure provides the trajectory of the slope, starting from the 61-th step. The change of sign of the slope was obtained in steps 79 and 71 for activities 18 and 29 respectively.

3. As mentioned in Section 2.6, an important issue for interpreting the detected agglomerations is to evaluate how far the contiguity among regions is a significant factor of clustering. This issue is faced by means of a bootstrap test of invariance for the group of permutations of the coordinates. The third figure provides an histogram of the results of $B = 1000$ bootstrap replications with a mark (vertical line) corresponding to the maximum obtained at the optimizing phase. These simulations show that the contiguity is a significant factor of the clustering for the activity 29 and not significant for activity 18. In view of the observations made from the maps in Figure 2, this suggests that the central territories of Argentine have specifically attractive factors for activity 18 but not for activity 29; this might explain that for activity

18 contiguity is not a significant factor of clustering, location in the central part being more important, whereas for activity 29 contiguity its more important, probably due to factors of local attractivity.

4. Finally, the fourth figure presents the same data as the third one on a cumulative form under the scale of a standard normal distribution. These last two figures reveal that the distributions of the bootstrap replications is clearly unimodal, bell-shaped and roughly normal without noticeable pathologies in the tails.

4 Concluding remarks

The role of contiguity in the formation of specialized agglomeration is multifaceted. The application to Argentinean data suggests that it may be different for different activities. The bootstrap test of invariance for coordinates permutation therefore provides an information complementary to the information provided by the optimization result and permits to detect different types of specialized agglomerations. This is also an issue on the very concept of agglomeration. From a narrow spatial approach, one could consider that the contiguity is so an essential ingredient that a region may be incorporated in an agglomeration even if its relative specialization is different from that of the agglomeration and this incorporation is left to be decided by the optimizing process. But if the contiguity is of a different nature, one might require that each region within a given agglomeration should display a same sign for the *log* of its Local Quotient. This is indeed a flexibility of the algorithm one parameter of which, namely "sign", specifies whether the regions entering an agglomeration should, or should not, have a same sign of $log\, LQ$. Another flexibility of the algorithm allows the choice between two types of the distance underlying the condition of contiguity, namely a distance "rook" or "queen" (see Appendix B).

This algorithm detects *relative* rather than absolute specialization; on the distinction between the two types of specialization, see *e.g.* Haedo and Mouchart (2012). As long as we interpret relative specialization for a given activity as a hint for the presence of factors favorable to the development of that activity, the information provided by the algorithm is of interest for an investor willing to invest in that activity and also suggests to investigate in order to uncover what and why are these favorable factors. Once these favorable factors have been identified one may consider to further develop those factors within the already specialized agglomerations and/or one may evaluate whether it might be advisable to develop those factors in non-specialized regions.

This paper is focused on a particular problem of spatial economics but the literature cited in the Introduction makes clear that the proposed structural model and the algorithmic detection have a potentially wide field of applications in spatial statistics.

References

Agrawal, R., Gehrke, J., Gunopulus, D., and Raghavan, P. (1998), Automatic subspace clustering of high dimensional data for data mining applications. In Proc. of the ACM/SIGMOD Int. Conference on Management of Data: 94-105.

Anselin, L. (1995), Local indicator of spatial association - LISA. *Geographical Analysis* **27**: 93-115.

Akaike, H. (1974), A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716-723.

Besag, J. and Newell, J. (1991), The detection of clusters in rare diseases. *Journal of the Royal Statistical Society* **154**: 143-155.

Burnham, K., and Anderson, D. (2002), *Model Selection and Multi-Model Inference*. New York: Springer-Verlag Inc.

Burnham, K., and Anderson, D. (2004), Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods and Research* **33**: 261-304.

Claeskens, G., and Hjort, N. (2008), *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Cliff, A., and Ord, J. (1981), *Spatial Processes: Models And Applications*. London: Pion.

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

Cook, A., Gold, D., and Li, Y. (2007), Spatial cluster detection for censored outcome data. *Biometrics* **63**: 540-549.

Diggle, P. (2000), Overview of statistical methods for disease mapping and its relationship to cluster detection. In P. Elliott, J.C. Wakefield, N.G. Best, and D.G. Briggs (eds.), *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

Duranton, G., Martin, P., Mayer, T., and Mayneris, F. (2010), *The Economics Of Clusters: Lessons From The French Experience*. New York: Oxford University Press.

Ester, M., Kriegel, H-P., Sander, J., and Xu, X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), AAAI Press.

Flahaut, B., Mouchart, M., San Martin, E., and Thomas, I. (2003), The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. *Accident Analysis and Prevention* **35**: 991-1004.

Friedman, J., and Fisher, N. (1999), Bump hunting in high dimensional data. *Statistics and Computing* **9**: 1-20.

Fujita, M., Krugman, P., and Venables, A. (2001), *The Spatial Economy. Cities, Regions, and International Trade*. Cambridge: MIT Press.

Fujita, M. and Thisse, J-F. (2002), *Economics of Agglomeration. Cities, Industrial Location, and Regional Growth*. Cambridge: Cambridge University Press.

Gangnon, R., and Clayton, M. (2003), A hierarchical model for spatially clustered disease rates. *Statistics in Medicine* **22**: 3213-3228.

Gómez-Rubio, V., Ferrándiz-Ferragud, J., and López-Quílez, A. (2005), Detecting clusters of disease with R. *Journal of Geographical Systems* **7**: 189-206.

Good, P. (2000), *Permutation Tests- A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Han, J., Kamber, M., and Tung, A. (2001), Spatial clustering methods in data mining: a survey. In H. Miller and J. Han (eds.), *Geographic Data Mining and Knowledge Discovery*. London and New York: Taylor and Francis.

Haedo, C. (2009), Measure of Global Specialization and Spatial Clustering for the Identification of "Specialized" Agglomeration. Ph.D. thesis, Bologna: Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna (IT). `http://amsdottorato.cib.unibo.it/1735/1/Christian_Haedo_tesi.pdf`

Haedo, C., and Mouchart, M. (2012), A stochastic independence approach for different measures of concentration and specialization, CORE Discussion Paper 2012/25, UCL Louvain-la-Neuve (B). `http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2012_25web.pdf`

Huang, L., Kulldorff, M., and Gregorio, D. (2007), A spatial scan statistics for survival data. *Biometrics* **63**: 109-118.

Kulldorff, M. (1997), A spatial scan statistic. *Communications in Statistics-Theory and Methods* **26**: 1481-1496. 799-810.

Kulldorff, M. (1999), Spatial scan statistics: models, calculations, and applications. In J. Glaz and N. Balakrishnan (eds.), *Scan Statistics and Applications*. Boston: Birkhäuser.

Kulldorff, M., and Nagarwalla, N. (1995), Spatial disease cluster: detection and inference. *Statistics in Medicine* **14**: 799-810.

Kulldorff, M., Tango, T., and Park, P. (2003), Power comparisons for disease clustering tests *Computational Statistics and Data Analysis* **42**: 665-684.

Lawson, A. (2001), *Statistical Methods in Spatial Epidemiology.* New York: John Wiley & Sons.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J-F., and Bertollini, R.(eds.) (1999), *Disease Mapping and Risk Assessment for Public Health.* Chichester: John Wiley & Sons.

Manly, B. (1991), *Randomization and Monte Carlo methods in biology.* London: Chapman and Hall.

Martin, P., and Ottaviano, G. (2001), Grow and agglomeration. *International Economic Review* **42**: 947-968.

Mori, T., and Smith, T. (2011), An industrial agglomeration approach to central place and city size regularities. *Journal of Regional Science* **51**: 694-731.

Mouchart, M., Russo, F., and Wunsch, G. (2009), Structural modelling, exogeneity, and causality. In H. Engelhardt, H-P. Kohler, and A. Fürnkranz-Prskawetz (eds.), *Causal Analysis in Population Studies: Concepts, Methods, Applications.* Dordrecht: Springer.

Neill, D. (2006), Detection of Spatial and Spatio-Temporal Clusters. Ph.D. thesis, Pittsburgh: School of Computer Science, Carnegie Mellon University (US).

Neill, D., Moore, A., and Cooper, G. (2006), A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems* **18**: 1003-1010.

Openshaw, S., Craft, A.W., Charlton, M., and Birch, J.M. (1988), Investigation of leukaemia cluster by use of a geographical analysis machine. *Lancet* **1**: 272-273.

O'Sullivan, D., and Unwin, D.J. (2010), *Geographic Information Analysis*, 2nd Edition. New Jersey: John Wiley & Sons.

Quah, D. (2002), Spatial agglomeration dynamics. *American Economic Review* **92**: 247-252.

Quah, D., and Simpson, H. (2003), *Spatial Cluster Empirics.* London: Mimeo.

Rossi-Hansberg, E. (2005), A spatial theory of trade. *American Economic Review* **95**: 1464-1491.

Schwartz, G. (1978), Estimating the dimension of a model. *Annals of Statistics* **6**: 461-464

Sokal, P., Oden, N., and Thomson, B. (1998), Local spatial autocorrelation in a biological model. *Geographical Analysis* **30**: 411-432.

Wakefield, J., Kelsall, J., and Morris, S. (2000), Clustering, cluster detection and spatial variation in risk. In P. Elliot, J. Wakefield, N. Best, and D. Briggs (eds.), *Spatial Epidemiology. Methods and Applications*. Oxford: Oxford University Press.

Yang, Y. (2005), Can the strengths of AIC and BIC be shared? *Biometrika* **92**: 937-950.

Zhang, T., and Lin, G. (2007), A decomposition of Moran's I for clustering detection. *Computational Statistics and Data Analysis* **51**: 6123-6137.

Zoellner, I., and Schmidtmann, I. (1999), Empirical studies of cluster detection: different cluster tests in application to German cancer maps. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, and J-F. Viel (eds.), *Disease Mapping and Risk Assessment for Public Health*. Chichester: John Wiley.

Appendix A: Model selection criteria

A good model selection technique essentially involves a balance between goodness of fit and complexity, typically measured by counting the number of parameters in the model. Given the set of basic regions, it would of course be desirable to compare all possible cluster schemes that can be formed from these regions, and then to identify a best cluster scheme. But this amount to evaluate as many schemes as there are partitions of the finite set $\mathcal{I} = \{1, 2, \cdots, I\}$. Even with contemporary computing facilities, this becomes an impossible task once I, the number of basic regions, become sizable.

We develop an algorithmic procedure, based on a greedy approach and, at each step, on a model selection criterion, balancing the goodness of fit and the number of parameters. There are a lot of model selection procedures (for more details see Diggle 2000, Burnham and Anderson 2002 and 2004, Kulldorff, Tango and Park 2003, Yang 2005, Gómez-Rubio, Ferrándiz-Ferragud and López-Quílez 2005, Claeskens and Hjort 2008, among others) and three of them deserve some attention. Firstly, the critical alpha criterion has the advantage of a unique asymptotic distribution, namely uniform on the unit interval, irrespectively of the number of degrees of freedom; however a drawback of the criterion is to require a comparison among very small numbers. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) have also been widely discussed. In what follows we briefly summarize the essential elements of those criteria.

The Akaike information criterion

Developed in Akaike (1974), AIC is founded in information theory where the loss of the information is evaluated by means of the Kullback-Leibler divergence. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. Thus the criterion to be minimized is given by

$$AIC^{(1)}(X_j \mid C) = 2\ (M(C) - 1) - 2\ \ln\ (\widehat{L}^{(1)}(X_j \mid C)) \tag{57}$$

If we now denote the AIC value for the random benchmark model by

$$AIC^{(0)}(X_j \mid C) = -2\ \ln\ (\widehat{L}^{(0)}(X_j \mid C)) \tag{58}$$

(because $M = 1$ under H_0). Then reformulating this measure in terms of AIC-differences from this benchmark, we obtain:

$$\begin{aligned} AIC(X_j \mid C) &= AIC^{(0)}(X_j \mid C) - AIC^{(1)}(X_j \mid C) && (59)\\ &= -2\ \ln\ [\widehat{L}^{(0)}(X_j \mid C) - \widehat{L}^{(1)}(X_j \mid C)] - 2\ (M(C) - 1) && (60)\\ &= T(X_j \mid C) - 2\ (M(C) - 1) && (61) \end{aligned}$$

and hence the best cluster schemes C is now that with larger value of $AIC(X_j \mid C)$.

The Bayesian information criterion

The Bayesian information criterion (BIC) was introduced by Schwartz (1978). This BIC has proved to yield a consistent estimator of the true model whenever it is included among the candidate models. It is based, in part, on the likelihood function, and it is closely related to AIC. The best model is that with the smallest

$$BIC^{(1)}(X_j \mid C) = (M(C) - 1)\ \ln\,(N_{..}) - 2\ \ln\,(\widehat{L}^{(1)}(X_j \mid C)) \tag{62}$$

Reformulating this measure in terms of BIC-differences from the benchmark equation (58), we obtain:

$$\begin{aligned} BIC(X_j \mid C) &= BIC^{(0)}(X_j \mid C) - BIC^{(1)}(X_j \mid C) && (63)\\ &= -2\ \ln\,[\widehat{L}^{(0)}(X_j \mid C) - \widehat{L}^{(1)}(X_j \mid C)] - (M(C) - 1)\ \ln\,(N_{..}) && (64)\\ &= T(X_j \mid C) - (M(C) - 1)\ \ln\,(N_{..}) && (65) \end{aligned}$$

and the best cluster schemes C is that with larger value of $BIC(X_j \mid C)$.

These model selection techniques differ in the way they penalize the larger numbers of clusters and offers alternative model-selection criteria to resolve potential over-fitting problems. From equations (61) and (65) it is seen that, contrary of BIC, AIC penalizes the number of parameters independently of $N_{..}$. For this property, BIC might be prefered although many authors argue that AIC has theoretical and practical/performance advantages over BIC (for more details see Burnham and Anderson 2004, and Yang 2005). More importantly, maximizing AIC or BIC is equivalent to maximize the divergence $d(p_{\vec{m}|j} \mid g_{\vec{m}\cdot})$, up to different penalizations for the number of parameters.

Appendix B: The main parameters of the algorithm

The algorithm treats each activity j independently and requires the specification of several parameters. The values used in the application on Argentinean data are identified by: *Arg.* The main parameters are:

- contiguity matrix: simple first order queen or rook matrix; see Remark on the W matrix in Section 2.5) (*Arg*: rook);

- selection of the regions i:
  - either according to N_{ij}: either all regions, or only those with $N_{ij} > 0$, or only those with $N_{ij} >$ some fixed quote; ignoring the regions where $N_{ij} = 0$ actually accelerates the processing of the algorithm (*Arg*: $N_{ij} > 4$);
  - or according to the sign of $log\, LQ$: all, only this with $log\, LQ > 0$ (over-specialized) or only those with $log\, LQ < 0$ (under-specialized), possibly with 0 replaced by $+ or - \varepsilon$.
  Note: both criterion may be combined or not;

- specification of the total number of primary units to be taken into account in the evaluation of the criterion: either $N_{..}$ (coded as δ TRUE), or the sum of the N_{ij} corresponding to the regions actually selected in the previous step (coded as δ FALSE). The case "FALSE" actually ignores the unselected regions although present in the country (*Arg*: δ = TRUE);

- parameter "sign": "sign" = TRUE when the agglomerations contain only regions with a same sign of the $log\, LQ$, *i.e.* only over-specialized or only under-specialized regions, otherwise "sign" is FALSE (*Arg*: sign = TRUE; more explicitly: if $i^*_{[1]}$ and i_2 are aggregated into a same cluster, then the log of the local quotients $LQ_{i^*_{[1]},j}$ and $LQ_{i_2,j}$ have the same sign; *i.e.* both regions $i^*_{[1]}$ and i_2 are either over-specialized or under-specialized).
  Note: the parameter "sign" is activated at each step of the algorithm but if, in the second parameter, the regions have been selected on the sign of $log\, LQ$, the parameter "sign" is always TRUE;

- number of bootstrap replications (*Arg*: B = 1000);

- criterion: BIC or AIC (*Arg*: BIC);

- two parameters for the stopping rule:
  - firstly a selection between stopping according to a change of sign in the trajectory of the criterion or according to the local slope of the trajectory of the criterion (*Arg*: change of sign);
  - secondly the width of the window within which is evaluated the change of sign or the slope of the criterion (*Arg*: h = 60).