Methodological framework for the Analysis of Industrial Geographical Data.

Appendix for “La Geografía Industrial de América Latina y el Caribe”.[*][†]

(v150924)

Christian HAEDO[a] and Michel MOUCHART[b]

[a] *Centro Interuniversitario de Investigaciones sobre Desarrollo Económico, Territorio e Instituciones (CIDETI), Alma Mater Studiorum-Università di Bologna in the Argentine Republic*

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

[*] MIALC project commissioned by EU-LAC Foundation, Hamburg, Germany.

[†] This document has greatly benefited from comments of Dominique Peeters, Center for Operations Research and Econometrics (CORE), UCLouvain, Belgium.

Contents

1 Introduction

This report provides some hints that should help the reader understanding methodological issues underlying the treatment of areal data for the elaboration of regional economic policy.

Economic activity is spatially concentrated. Spatial concentration generates agglomeration economies, notably upstream-downstream linkages, which help firms become more productive. These positive effects involve a critical mass of workers and infrastructure, and dense networks of suppliers and collaborators. The central role of agglomeration economies on the spatial structure of the economy has inspired a large literature focused on trying to understand their causes or origins and the dynamic of these economies, as well as most adequate industrial policies for their consolidation and promotion wherever the industrial agglomerations are weak or nonexistent. It has been recognized that the industrial policy objectives can be better fulfilled if they are more sensitive to places and sectors in design and delivery (Donato, 2002; Nathan and Overman, 2013).

Agglomeration economies may appear in different geographical scales and may involve different disaggregation levels, and consequently, a certain space scale is not necessarily equivalent to another (Arbia, 1989, 2001; Krugman, 1991; Arbia and Espa, 1996; Anas *et al.*, 1998; Duranton and Overman, 2005; Arbia *et al.*, 2008, among others). In this report, we analyze data referring to a discrete space (lattice or areal data regarding administrative divisions), and the boundaries cannot be ignored, given that economic conditions can change abruptly due to changes in the tax system, in transportation costs, or to the impact of public policies at regional and sectorial levels.

The work underlying this report handles issues related to a discrete space, *i.e.* a space partitioned into a finite number of regions, along with a finite number of economic activities. The basic data have the form N_{ij} representing the a number of statistical units for a region i and an activity j. The labels i of the regions are arbitrary and incorporate no information neither on spatial contiguity nor on distance among regions. At a first stage, this analysis is "spaceless" and motivated by policy-making rather than by spatial diffusion issues. Indeed, the data N_{ij} provides no information about the localization of primary units *within* a region. At a second stage, problems of agglomerations, *i.e.* of clusters of contiguous regions, are introduced but these problems require *additional* data related to the distance, or contiguity, between the regions. This topic is the object of the last section of this report.

The results of most analyses are shown in Choropleth maps which enclose additional information in tables and graphics that are useful as a description and helps in understanding the different indexes and algorithm outputs. A Choropleth map or an area-value map is one of the most frequently used maps in geography (Robinson *et al.*, 1995) which reveals data patterns by showing the distribution of a chosen phenomenon within the selected area. In order to construct a Choropleth map, data is aggregated into classes that are represented in the map by shades of color. The greater the density of the color, the greater the density or value represented. While such generalization may undercover some details, it allows a quicker observation of patterns and variation, and provides a basis for posing analytical questions.

It is well known that when the areas or regions are not uniform, as in the case of this project which considers the different levels of administrative and political boundaries of each country, the Choropleth map fails to equate the visual importance of each region with its geographic area in comparison to a value indicator, giving sparsely areas great visual emphasis. This limitation can be solved by using the method of mesh/grid-square mapping (dividing the map into equal sized units/squares and then color each one according to the data being encoded), or by using dot grid maps (overlying a regular grid of circles in which each is sized accordingly to the value of the region the circle falls into), among others (for more details about these topics see Thompson et al. 1999; Bertin 2010; Liseikin 2010). In the following stages of the project it is foreseen to incorporate the aforementioned grid methods as an optional visualization of the results.

Next section gives a (slightly formal) description of the data to be analyzed along with some notational convention. Section 3 presents an overview of the main topics that are studied in this Atlas, namely the basic concepts around the regional specialization and the industrial concentration. This section is completed by introducing different indices useful for characterizing the manufacturing structure of the regions. The following section briefly explains the method to analyze the Regional Manufacturing Structure of the countries by a simultaneous grouping of regions and activities algorithm; the interested reader may find a more complete exposition in Haedo and Mouchart (2014). The last section briefly sketches a definition and a method for detecting Specialized Agglomerations, again the interested reader may find a more complete exposition in Haedo and Mouchart (2015); and ends with the explanation of the Manufacturing Competitiveness index.

2 Structure of the data and notation for the regions

This report treats data that refer to different countries of Latin America and the Caribbean. All these data share a similar structure. More specifically, 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. 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 of the regions is crucial for the availability of the data. It is to be noted that the labels, $i \in \mathcal{I}$ and $j \in \mathcal{J}$, are "not informative", meaning that the labeling system is arbitrary and does not embody any information.

For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, we observe the number N_{ij} of primary units; these could be typically a number of manufacturing employees or a number of manufacturing economic units. 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:

$$N_{i\cdot} = \sum\nolimits_{j=1}^{J} N_{ij}; \qquad N_{\cdot j} = \sum\nolimits_{i=1}^{I} N_{ij}; \qquad N_{\cdot\cdot} = \sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J} N_{ij} = \sum\nolimits_{j=1}^{J} N_{\cdot j} = \sum\nolimits_{i=1}^{I} N_{i\cdot} \tag{1}$$

This report is mostly concerned with the *relative industrial concentration*, *i.e.* the concentration of a given activity in a given region relatively to other regions and with the *relative regional specialization*, *i.e.* the shares of the different activities in a given region as compared that those shares at the country level. 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 [1]:

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

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

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

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

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

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

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

A high proportion of the data used in this work are presented in the form of these profiles, or conditional distributions, as a natural way of representing the regional structure of an industry or the industrial structure of a region.

For later use, 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:

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

We accordingly write:

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

The analysis may be conducted 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 $\ell : \mathcal{U} \rightarrow \Omega$ where $\ell(u)$ stands for the localization of u in Ω and an activity function $a \; : \; \mathcal{U} \rightarrow \mathcal{J}$ 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:

$$x_{ij}^{u} = 1\!\!1_{\{\ell(u) \in \Omega_i,\, a(u)=j\}} \tag{8}$$

[1] When the components of a vector are indexed by i (regions) or by j (activities), we use an arrow above the index that defines the components of the vector.

Clearly:

$$\sum_{u \in \mathcal{U}} x_{ij}^{u} = N_{ij} \tag{9}$$

Cartographic data

Country maps cover three levels of administrative and political boundaries: national administrative boundaries and first and second levels of sub-national administrative boundaries (for more details about the administrative and political boundaries of each country used in this project, see "Fuentes de datos/Data sources" in the bar menu of the project's website).

Manufacturing activities

The manufacturing activities for each country have been homologated in 21 divisions (Table 1) of the International Standard Industrial Classification of All Economic Activities (ISIC) Revision 4 of the United Nations (http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=17&Lg=1) as follows:

Table 1: *Homologated manufacturing activities*

Divisions ISIC Rev.4	**Description**
10 – 11	Manufacture of food products; Manufacture of beverages
12	Manufacture of tobacco products
13	Manufacture of textiles
14	Manufacture of wearing apparel
15	Manufacture of leather and related products
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
17	Manufacture of paper and paper products
18	Printing and reproduction of recorded media
19	Manufacture of coke and refined petroleum products
20 – 21	Manufacture of chemicals and chemical products; Manufacture of basic pharmaceutical products and pharmaceutical preparations
22	Manufacture of rubber and plastics products
23	Manufacture of other non-metallic mineral products
24	Manufacture of basic metals
25	Manufacture of fabricated metal products, except machinery and equipment
26	Manufacture of computer, electronic and optical products
27	Manufacture of electrical equipment
28 – 33	Manufacture of machinery and equipment n.e.c.; Repair and installation of machinery and equipment
29	Manufacture of motor vehicles, trailers and semi-trailers
30	Manufacture of other transport equipment
31 – 32	Manufacture of furniture; Other manufacturing
38	Waste collection, treatment and disposal activities; materials recovery

3 Regional manufacturing indicators

The evaluation of regional characteristics and performances can be facilitated by a reference to different indices. In the sequel, we shall use in particular the following ones.

The data classification method adopted to construct the Choropleth maps is the Fisher-Jenks optimization method, also called the Jenks Natural Breaks (Jenks, 1967). Jenks' method is the one-dimensional version of K-means clustering (Forgy, 1965). This data clustering method calculates groupings of data values based on the data distribution, seeking to reduce the variance within groups and maximize the variance between groups. The advantage of this classification is that it identifies real classes within the data. For more details about cluster analysis see Hartigan (1975), Kaufman and Rousseeuw (2005), Gan et al. (2007), Everitt et al. (2010), Aggarwal and Reddy (2013), among many others.

3.1 Specialization and concentration: basic concepts

3.1.1 General approach

The practitioner may like to be reminded that the analysis of specialization and of concentration raises, at the conceptual level, two different issues to be carefully distinguished.

One issue consists in characterizing the dispersion, or concentration, of distributions on unordered categorical variables such as regions or activities. In the theory of information it is established that the distribution of maximal dispersion, or minimal concentration, is the uniform distribution; for instance, on the regions it would be: $P(i) = \frac{1}{I} \;\; i = 1, 2, \cdots I$. A measure of dispersion may accordingly be obtained though a measure of the discrepancy (*i.e.* a distance or a divergence, see below) between the uniform distribution, taken as a benchmark of maximal dispersion, and the distribution of interest. A possible alternative approach is to introduce an order, "natural" or artificial, among the categories and to rely on methods based on Lorentz curve and Gini coefficients. In most cases, however, the ordering is defined only relatively to the distribution to be analyzed and is therefore not intrinsic to the set of possible categories. In this book, we generally give preference to the first approach.

Another issue is to focus the attention on one among three possible levels of analysis. A first level, to be called an *absolute* concept (of specialization or of concentration), analyzes the dispersion of a distribution in itself without reference to the contingency with another variable. We then obtain absolute concepts of specialization or of concentration, either for a whole country or for a specific region i or a specific activity j. A second level, to be called a *relative* concept (of specialization or of concentration), confronts the conditional distribution of the activities within a region with the marginal distribution of the activities for the country, taken as a benchmark of no relative specialization, or the conditional distribution of the regions for a given activity with the marginal distribution of the regions for the country, taken as a benchmark of no relative concentration. A third level, to be called a *global* concept confronts the actual bivariate distribution, on regions × activities, with the product of their marginal distributions that represents the closest distribution revealing independence between regions and activities, taken as a benchmark of a completely non-concentrated, or non-specialized, country. This measure may be called a measure of “global localization” or, following Perroux (1950), of “polarization”; this book uses both terms interchangeably.

In Table 2 we have summarized, under some conventional headings, these different levels of analysis, denoting by $d(p \mid q)$ an arbitrary divergence or distance between distributions p and q.

Table 2: *Some conventional definitions*

Technique	**Measured concept**
$d(p_{\cdot\vec{j}} \mid [1/J])$	Absolute industrial homogeneity
$d(p_{\vec{i}\cdot} \mid [1/I])$	Absolute regional homogeneity
$d(p_{\vec{j}\mid i} \mid [1/J])$	Absolute specialization of region i
$d(p_{\vec{i}\mid j} \mid [1/I])$	Absolute concentration of activity j
$d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$	Relative specialization of region i
$d(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$	Relative concentration of activity j
$d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$	Global localization, or polarization, of the country

Absolute homogeneity refers to the spread of the (marginal) distribution of the regions $p_{\vec{i}\cdot}$ or of the activities $p_{\cdot\vec{j}}$. *Absolute regional specialization*, is a feature of the distribution of activities across a region $p_{\vec{j}\mid i}$, and a region is said to be absolutely specialized if a few activities concentrate a large share of the region. This may be the case, for instance, when an activity is considerably larger than others at a country level. *Relative regional specialization* of a region shows up when an area has a greater proportion of a particular activity than the proportion of that activity in the whole territory. In other words, relative regional specialization compares an area share of a particular activity with the activity share at the country level, and is accordingly measured through a discrepancy $d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$, thus relatively to the marginal distribution $p_{\cdot\vec{j}}$. The similar comment is also valid for absolute and relative industrial concentration.

In order to introduce the concept of *global localization* or *polarization*, imagine the following (artificial) experiment. Draw randomly one primary unit from the $N_{\cdot\cdot}$ ones and classify the drawn primary unit into the region and the activity. The probability of drawing a primary unit from the cell (i, j) is evidently p_{ij}. Within this framework, the absence of global localization may be viewed as a stochastic independence between the row and the column criteria: for instance, in every region, there would be a same probability that a randomly drawn individual is active in any specific activity. Thus, global localization may be viewed as an association between the region and the activity variables. This suggests to measure the degree of global localization through a statistic that might be used for testing independence in a contingency table: this is precisely operated by the discrepancy $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$.

Remark

These concepts are based only on information contained in the contingency table. In some cases, it may be of interest to use as a benchmark distribution a distribution relative to an exogenous variable. For instance, for distributions on the regions, such as $p_{\vec{i}}$ or $p_{\vec{i}|j}$, a benchmark distribution might be the distribution of the areas or of the populations of the regions. For instance, Mori and Smith (2011) have proposed that the benchmark of the areas might be used as a basis for an hypothesis of a purely non concentrated activity. Similarly, the distribution of the populations of the regions may be used, in epidemiology, as a benchmark for the non-concentration of a disease of interest. ■

3.1.2 An additional note on international comparisons

When comparing the regional structure of several countries, the divergences, shown up in Table 2, are evaluated country-wise. It should be emphasized that these comparisons may crucially depend on the choice of a particular discrepancy, or dissimilarity, among distributions. This work makes reference to three discrepancies of particular interest, namely the Hellinger distance and the Kulback-Leibler and χ^2 divergences. In the finite case of two distributions $q = (q_1, \cdots, q_n)$ and $r = (r_1, \cdots, r_n)$, these discrepancies are defined as follows:

$$d^2_H(q_{\vec{i}} \mid r_{\vec{i}}) = \frac{1}{2}\sum_{i=1}^{I}(\sqrt{q_i} - \sqrt{r_i})^2 \qquad \text{Hellinger-distance} \tag{10}$$

$$d_{\chi^2}(q_{\vec{i}} \mid r_{\vec{i}}) = \sum_{i=1}^{I} r_i \left(\frac{q_i}{r_i} - 1\right)^2 \qquad \chi^2 - \text{divergence, or inertia} \tag{11}$$

$$d_{KL}(q_{\vec{i}} \mid r_{\vec{i}}) = \sum_{i=1}^{I} q_i \log\left(\frac{q_i}{r_i}\right) \qquad \text{Kullback-Leibler divergence} \tag{12}$$

Experience shows that the ranking among regions or activities may be robust with respect to changes of the discrepancy in some cases but may also crucially depend on it in other cases. However, the ranking of the measures of polarization is generally robust. For some interesting examples, see Haedo and Mouchart (2012).

3.1.3 Local approach

A natural start for the analysis of relative concepts is a local one, where one examines whether a cell (i, j) reveals over- or under-specialization, or equivalently whether a cell (i, j) reveals over- or under-concentration. For this purpose, the well-established *Location Quotient* (Florence, 1939), also known as the (estimated) *Hoover-Balassa coefficient* for the cell (i, j), may be written in several equivalent forms as:

$$LQ_{ij} = \frac{N_{ij}/N_{i\cdot}}{N_{\cdot j}/N_{\cdot\cdot}} = \frac{N_{ij}/N_{\cdot j}}{N_{i\cdot}/N_{\cdot\cdot}} = \frac{N_{ij}N_{\cdot\cdot}}{N_{i\cdot}N_{\cdot j}} = \frac{p_{ij}}{p_{i\cdot}\ p_{\cdot j}} = \frac{p_{j|i}}{p_{\cdot j}} = \frac{p_{i|j}}{p_{i\cdot}} \tag{13}$$

The last three equalities in (13) express the same concept through proportions, *i.e.* independently of $N_{\cdot\cdot}$ that represents the number of observations.

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

$$\begin{array}{lllll} LQ_{ij} & = 1 & \text{or} & p_{ij} = p_{i\cdot}\, p_{\cdot j} & \text{non-specialization} \\ & > 1 & \text{or} & p_{ij} > p_{i\cdot}\, p_{\cdot j} & \text{over-specialization} \\ & < 1 & \text{or} & p_{ij} < p_{i\cdot}\, p_{\cdot j} & \text{under-specialization} \end{array} \tag{14}$$

where "non-specialization" corresponds to a local contribution to the row-column independence. It should be clear from (14) that a discrepancy between the distributions $[p_{ij}]$ and $[p_{i\cdot}\, p_{\cdot j}]$ is equivalent to a discrepancy between the matrix $[LQ_{ij}]$ and a corresponding matrix of one's. Note that LQ_{ij} is valued in $[0, +\infty)$ and that $p_{i\cdot}\, p_{\cdot j} > 0$. The last two equalities in (13) emphasize that the specialization is an issue concerning the global structure at a country level: thus the absence of specialization of a cell (i, j) means that, *relative to the distribution in the country*, activity j is not over-(nor under-) represented in region i *and* that region i is not over-(nor under-) represented for activity j. Thus, "location" points to the fact that LQ_{ij} is localized in the cell (i, j).

Remark.

Recent works, among others by Moineddin *et al.*(2003), O'Donoghue and Gleave (2004), Guimarães *et al.*(2003 and 2009) or Haedo (2009), have drawn the attention on the so-called "small area problem" for the location quotient. Indeed, when a region i is very small, as compared with other regions of a country, the value of its location quotient may not be put on an equal foot with the location quotients relative to other regions. In Section 5.3, we shall weight the *log* of the location quotient with the corresponding N_{ij}, see equation (78), with the effect of representing more adequately the impact of a small area in the characterization of the relative industrial concentration of a given activity. ■

Manufacturing regional specialization

For each country, we have been evaluated the regional specialization of each region i, $Mre_{[i\cdot]}$, for both primary units: manufacturing economic units, $[Meu_{ij}]$, and manufacturing employment, $[Mem_{ij}]$.

$$Mre_{[i\cdot]} = \sum\nolimits_{j=1}^{J} p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot}\; p_{\cdot j}}\right) \tag{15}$$

And for international comparisons of the global regional specialization or global localization levels:

$$Mre = \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot}\; p_{\cdot j}}\right) \tag{16}$$

The Choropleth maps for each country show the values of $Mre_{[i\cdot]}[Meu_{ij}]$ and $Mre_{[i\cdot]}[Mem_{ij}]$, both at last available period, grouping the over-specialized regions i into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

Regional concentration of manufacturing activities

For each country, we have been evaluated the regional concentration of each activity j separately, Mrc_{ij}, for both primary units: manufacturing economic units, $[Meu_{ij}]$, and manufacturing employment, $[Mem_{ij}]$, as follows

$$Mrc_{ij} = p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot}\; p_{\cdot j}}\right) \tag{17}$$

And for international comparison of the regional concentration of every single activity j,

$$Mrc_{[\cdot j]} = \sum\nolimits_{i=1}^{I} p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot}\; p_{\cdot j}}\right) \tag{18}$$

The Choropleth maps of each country show the values of $Mrc_{ij}[Meu_{ij}]$ and $Mrc_{ij}[Mem_{ij}]$ for each activity j separately, both at last available period, grouping the over-specialized regions i into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

3.2 Further indicators

Population

For each country, the number of people, either for a country, Pop, or for a region i, Pop_i, are evaluated as follows:

$$Pop = \sum\nolimits_{i=1}^{I} Pop_i \tag{19}$$

The Choropleth map of each country shows the values of Pop_i at last available period, aggregated into three classes: high, medium and low, using Jenks Natural Breaks classification method.

Relative Variation of the Population

For each country, the relative variation of the number of people, either for Pop or for Pop_i, has been evaluated by computing its values at two different instants, namely t_1 and t_2 with $t_1 < t_2$, as follows:

$$RVPop_{i|t_1,t_2} = \left(\frac{Pop_{i|t_2}}{Pop_{i|t_1}} - 1\right) \times 100 \tag{20}$$

$$RVPop_{t_1,t_2} = \left(\sum\nolimits_{i=1}^{I} \frac{Pop_{i|t_2}}{Pop_{i|t_1}} - 1\right) \times 100. \tag{21}$$

The Choropleth map of each country shows the values of $RVPop_{i|t_1,t_2}$, aggregated into three classes: increased, no changes and decreased.

Manufacturing economic units

For each country and for each pair $(i,j) \in \mathcal{I} \times \mathcal{J}$, one may evaluate the number Meu_{ij} of manufacturing economic units. Thus we obtain a two-way $I \times J$ contingency table $\mathbf{N} = [Meu_{ij}]$ that also produces row, column and table totals denoted as follows:

$$Meu_{i\cdot} = \sum\nolimits_{j=1}^{J} Meu_{ij} \tag{22}$$

$$Meu_{\cdot j} = \sum\nolimits_{i=1}^{I} Meu_{ij} \tag{23}$$

$$Meu = \sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J} Meu_{ij} = \sum\nolimits_{i=1}^{I} Meu_{i\cdot} = \sum\nolimits_{j=1}^{J} Meu_{\cdot j} \tag{24}$$

The Choropleth maps of each country show the values of $Meu_{i\cdot}$ and of $Meu_{\cdot j}$ separately for each activity j, both at last available period, aggregated into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

Relative Variation of the Manufacturing economic units

For each country, the relative variation of the manufacturing economic units has been evaluated by computing $\mathbf{N} = [Meu_{ij}]$ at two different instants, t_1 and t_2 with $t_1 < t_2$, as follows:

$$RVMeu_{ij|t_1,t_2} = \left(\frac{Meu_{ij|t_2}}{Meu_{ij|t_1}} - 1\right) \times 100 \tag{25}$$

$$RVMeu_{[i\cdot]|t_1,t_2} = \left(\sum\nolimits_{j=1}^{J} \frac{Meu_{ij|t_2}}{Meu_{ij|t_1}} - 1\right) \times 100 \tag{26}$$

$$RVMeu_{[\cdot j]|t_1,t_2} = \left(\sum\nolimits_{i=1}^{I} \frac{Meu_{ij|t_2}}{Meu_{ij|t_1}} - 1\right) \times 100 \tag{27}$$

$$RVMeu_{t_1,t_2} = \left(\sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J} \frac{Meu_{ij|t_2}}{Meu_{ij|t_1}} - 1\right) \times 100 \tag{28}$$

The Choropleth maps of each country show the values of $RVMeu_{[i\cdot]|t_1,t_2}$ and of $RVMeu_{[\cdot j]|t_1,t_2}$ separately for each activity j, aggregated into three classes: increased, no changes and decreased.

Manufacturing availability of local enterprise resources

For each region i of a country, one may evaluate the manufacturing availability of local enterprise resources, $Maler_{[i\cdot]}$, as the ratio between the manufacturing economic units, $Meu_{i\cdot}$, and the population, Pop_i:

$$Maler_{[i\cdot]} = \frac{Meu_{i\cdot}}{Pop_i} \tag{29}$$

The Choropleth map of each country shows the values of $Maler_{[i\cdot]}$ at last available period, aggregated into three classes: high, medium and low, using Jenks Natural Breaks classification method.

Manufacturing employment

For each country and for each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, we observe the number Mem_{ij} of manufacturing jobs held. Thus we obtain a two-way $I \times J$ contingency table $\mathbf{N} = [Mem_{ij}]$ that also produces row, column and table totals denoted as follows:

$$Mem_{i\cdot} = \sum\nolimits_{j=1}^{J} Mem_{ij} \tag{30}$$

$$Mem_{\cdot j} = \sum\nolimits_{i=1}^{I} Mem_{ij} \tag{31}$$

$$Mem = \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} Mem_{ij} = \sum\nolimits_{i=1}^{I} Mem_{i\cdot} = \sum\nolimits_{j=1}^{J} Mem_{\cdot j} \tag{32}$$

The Choropleth maps of each country show the values of $Mem_{i\cdot}$ and of $Mem_{\cdot j}$ separately for each activity j, both at last available period, aggregated into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

Relative Variation of the Manufacturing employment

For each country, the relative variation of the manufacturing jobs held has been calculated by computing $\mathbf{N} = [Mem_{ij}]$ at two different instants, t_1 and t_2 with $t_1 < t_2$, as follows:

$$RVMem_{ij|t_1,t_2} = \left(\frac{Mem_{ij|t_2}}{Mem_{ij|t_1}} - 1 \right) \times 100 \tag{33}$$

$$RVMem_{[i\cdot]|t_1,t_2} = \left(\sum\nolimits_{j=1}^{J} \frac{Mem_{ij|t_2}}{Mem_{ij|t_1}} - 1 \right) \times 100 \tag{34}$$

$$RVMem_{[\cdot j]|t_1,t_2} = \left(\sum\nolimits_{i=1}^{I} \frac{Mem_{ij|t_2}}{Mem_{ij|t_1}} - 1 \right) \times 100 \tag{35}$$

$$RVMem_{t_1,t_2} = \left(\sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} \frac{Mem_{ij|t_2}}{Mem_{ij|t_1}} - 1 \right) \times 100 \tag{36}$$

The Choropleth maps of each country show the values of $RVMem_{[i\cdot]|t_1,t_2}$ and of $RVMem_{[\cdot j]|t_1,t_2}$ separately for each activity j, aggregated into three classes: increased, no changes and decreased.

Manufacturing quality of local enterprise resources

For each region i of a country, one may evaluate the manufacturing quality of local enterprise resources, $Mqler_{[i\cdot]}$, as the ratio between the manufacturing employment, $Mem_{i\cdot}$, and the manufacturing economic units, $Meu_{i\cdot}$:

$$Mqler_{[i\cdot]} = \frac{Mem_{i\cdot}}{Meu_{i\cdot}} \tag{37}$$

The Choropleth map of each country shows the values of $Mqler_{[i\cdot]}$ at last available period, aggregated into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

Manufacturing level

For each region i of a country, one may evaluate a manufacturing level, $Ml_{[i\cdot]}$, as the ratio of the rate of manufacturing employment in region i and the corresponding rate for the country:

$$Ml_{[i\cdot]} = \frac{\frac{Mem_{i\cdot}}{Pop_i}}{\frac{Mem}{Pop}} \tag{38}$$

The Choropleth map of each country shows the values of $Ml_{[i\cdot]}$ at last available period, aggregated into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

Manufacturing performance

For each region i of a country, the manufacturing level, $Ml_{[i\cdot]}$, may be evaluated by computing its $Ml_{[i\cdot]}$ at two different instants, t_1 and t_2 with $t_1 < t_2$. Hence, the $Mp_{[i\cdot]}$ distinguish between different scenarios or classes of $Ml_{[i\cdot]|t_1,t_2}$ as follows:

1 = *raising industrialized regions*, when $Ml_{[i\cdot]|t_2} > Ml_{[i\cdot]|t_1} > 1$;

2 = *declining industrialized regions*, when $Ml_{[i\cdot]|t_1} > Ml_{[i\cdot]|t_2} > 1$;

3 = *new industrialized regions*, when $Ml_{[i\cdot]|t_2} > 1 > Ml_{[i\cdot]|t_1}$;

4 = *desindustrializing regions*, when $Ml_{[i\cdot]|t_1} > 1 > Ml_{[i\cdot]|t_2}$;

5 = *developing industrialized regions*, when $1 > Ml_{[i\cdot]|t_2} > 0.60$, and $Ml_{[i\cdot]|t_2} > Ml_{[i\cdot]|t_1}$;

6 = *unindustrialized regions*, when $0.60 > \max\{Ml_{[i\cdot]|t_1}, Ml_{[i\cdot]|t_2}\}$.

where the threshold value 0.60 is somewhat arbitrary but aimed at indicating a movement of clear significance.

The Choropleth maps of each country show the classes of $Mp_{[i\cdot]}$.

4 Regional manufacturing structure: simultaneous grouping of regions and activities

4.1 Background

The purpose of this section is to extend in two directions the usual analysis of a given contingency table. Firstly, we want to examine simultaneous groupings of regions and activities, rather than separate ones. Secondly, instead of considering arbitrarily pre-specified groupings we look for an automatic construction of grouping aimed at providing optimal groupings according to a pre-specified criterion. From an economic geography point of view, the discrepancy $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$, taken as a measure of polarization, or global localization, of the country, summarizes the spatial pattern of the economic activities, or equivalently the distribution of the activities among the regions. We aim to simultaneously regroup regions with a similar manufacturing structure in terms of relative sub- and over-specialization and activities with a similar spatial pattern in terms of relative sub- or over-concentration. Shortly said, we want to build an algorithm for an automatic summary of a possibly large regions × activities contingency table that keeps (almost) unchanged the polarization of the economy.

When the algorithm looks for collapsing regions or activities, no restriction is considered about the regions or the activities to be clusterized. Thus, for the regions, no criteria of contiguity, or of some distance-based pattern, is operating because the algorithm is not looking for agglomerations, in the sense of clustering "neighboring" regions. The clusters to be elicited are of a structural nature, *i.e.* clusters of regions with a similar *relative* regional specialization pattern, or similar sectorial structure, irrespectively of their geographical localization. Similarly, when collapsing activities, no consideration of inter-sectorial relationship, nor of value chain, is operating because only a similar *relative* manufacturing concentration is at stake.

Our purpose is to summarize the original information, *i.e.* the complete contingency table $\mathbf{N} = [N_{ij}]$, to extract the most relevant patterns of specialization in the data. The actual challenge should be kept in mind. In the case of Argentina, for instance, there are $I = 511$ regions. Using the homologated manufacturing activities of Table 1 there are $J = 21$ activities. In 2004, for example, the total number of manufacturing employees was $N = 955,965$. Thus the contingency table is a 511 × 21 matrix of 955,965 primary units spread in 10,731 cells. It should be expected that many cells have either a very small number of manufacturing employees or no employee at all. The skeleton of the proposed algorithm may be viewed as follows. An "optimal" grouping of regions and activities should compromise between two opposite desiderata: the collapsed table should be as small as possible but should also display a minimum loss of polarization of the country.

Collapsing tables means building tables of smaller dimension through aggregated regions (rows) and/or activities (columns). The total number M of possible collapsed tables[2] for the $I \times J$ matrix $\mathbf{N}$ is

$$M = \sum\nolimits_{(m_1\ldots m_i\ldots m_l)} \binom{I}{m_1\ldots m_i\ldots m_l} \times \sum\nolimits_{(n_1\ldots n_i\ldots n_k)} \binom{J}{n_1\ldots n_j\ldots n_k} \tag{39}$$

where $l \leqslant I-1$, $k \leqslant J-1$, $m_1 + \ldots + m_i + \ldots + m_l < I$ and $n_1 + \ldots + n_j + \ldots + n_k < J$.

For I and J large, as in the present case, M is huge and trying all possibilities is not feasible. Therefore, we look for a greedy algorithm that only ensures a local optimum. This is obtained by means of a technique of hierarchical clustering, according to a dendrogram approach, combined with a correspondence analysis. Finally, at each step of the tree, permutation bootstrapping is used as a test that the envisaged regrouping performs better than if it had been generated randomly.

[2] Equation (39) may also be written as a product of two Bell numbers $B_n = \sum_{(m_1\ldots m_i\ldots m_l)}\binom{n}{m_1\ldots m_i\ldots m_l} = \sum_{0\leq k\leq n}\binom{n}{k}$ where $l \leqslant n-1$, $m_1 + \ldots + m_i + \ldots + m_l < n$. The Bell number is the sum of Stirling numbers of the second kind $S(n,k)$ that are equal to the number of partitions with k elements of a set with n members. Thus, the Bell number represents the total number of partitions of a set of n elements. In equation (39), we have $n = I$ and $m = J$. More details may be found in Rota (1964), Gardner (1978), Branson (2000) or Sloane (2001).

4.2 Singular value decomposition

Correspondence Analysis is based on a classical result in matrix theory, namely the Singular Value Decomposition (SVD). Let $\mathbf{P} = \frac{\mathbf{N}}{N_{\cdot\cdot}}$ be the probability matrix corresponding to $\mathbf{N}$. Let $\mathbf{r} = (p_{i\cdot})$ the vector of row marginals and $\mathbf{c} = (p_{\cdot j})$ the vector of column marginals. Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices formed with the row marginals and column marginals, respectively. Let us also consider the matrix of the residuals: $\mathbf{R} = \mathbf{P} - \mathbf{r}\mathbf{c}' = [p_{ij} - p_{i\cdot}p_{\cdot j}]$ and the matrix of the standardized residuals:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\mathbf{R}\mathbf{D}_c^{-1/2} \qquad s_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}. \tag{40}$$

A SVD of $\mathbf{S}$ may be written as:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}' \tag{41}$$

where $\lambda = (\lambda_1, \ldots, \lambda_K)$ is the vector of the strictly positive singular values, or eigenvalues, of $\mathbf{S}$ organized in descending order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ with $K = min(I-1, J-1)$, and where $\mathbf{D}_\lambda$ is accordingly $K \times K$, moreover $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = I_{(K)}$. In this decomposition, the dimension of $\mathbf{U}$ is $I \times K$, of $\mathbf{V}$ is $J \times K$.

The SVD of $\mathbf{S}$ will be used in the following spirit. Let us consider $\mathbf{D}_{\lambda(m)}$ be the principal submatrix of $\mathbf{D}_\lambda$ corresponding to the first m eigenvalues λ_k, let $\mathbf{U}_{(m)}$ and $\mathbf{V}_{(m)}$ be the submatrices made of the first m columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. The least-squares rank m approximation of $\mathbf{S}$ is obtained as: $\mathbf{S}_{(m)} = \mathbf{U}_{(m)}\,\mathbf{D}_{\lambda(m)}\,\mathbf{V}'_{(m)}$ (Eckart-Young theorem). Thus the sequence $\mathbf{S}_{(m)} \quad m = 1, \ldots, K$ is a sequence of improved approximations of $\mathbf{S}$.

The χ^2-divergence between $[p_{ij}]$ and $[p_{i\cdot}p_{\cdot j}]$, or Total Inertia, may be written as:

$$d_{\chi^2}([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}]) = \phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} = tr\mathbf{S}'\mathbf{S} = \sum_{k=1}^{K} \lambda_k^2 \tag{42}$$

where λ_k^2 also represents the k-th eigenvalues of $\mathbf{S}'\mathbf{S}$. This inertia, being a measure of the polarization of the country, may be viewed as a global measure of the information provided by the contingency table $\mathbf{P}$.

The principal coordinates for the rows (regions) are:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\,\mathbf{U}\,\mathbf{D}_\lambda = [f_{ik}] \quad I \times K \qquad f_{ik} = p_{i\cdot}^{-1/2}\,\lambda_k\,u_{ik} \tag{43}$$

where f_{ik} represents the score of region i in the k-th dimension of the factor space $I\!R^K$. Later on, we shall systematically use the decomposition of $\mathbf{F}$ into its I-dimensional columns denoted[3] as $\mathbf{F} = [f_{\vec{i}1}, \cdots, f_{\vec{i}K}]$. It may be checked that:

$$\mathbf{F}'\mathbf{D}_r\mathbf{F} = \mathbf{D}_\lambda^2 \qquad i.e. \quad \sum_{1 \leq i \leq I} p_{i\cdot}\, f_{ik}^2 = \lambda_k^2 \tag{44}$$

Thus, equation (44) decomposes the k-th eigenvalue of $\mathbf{S}'\mathbf{S}$, also the k-th component of the inertia, according to the contribution of each region i, namely $I_i = p_{i\cdot}f_{ik}^2$.

[3] When the components of a vector are indexed by i (regions) or by j (activities), we use an arrow above the index that defines the components of the vector.

Similarly, the principal coordinates for the columns (activities) are:

$$\mathbf{G} \;=\; \mathbf{D}_c^{-1/2}\,\mathbf{V}\,\mathbf{D}_\lambda \;=\; [g_{jk}] \quad J \times K \qquad g_{jk} \;=\; p_{\cdot j}^{-1/2}\,\lambda_k\, v_{jk} \tag{45}$$

where g_{jk} represents the score of activity j in the k-th dimension of the factor space $\mathbb{R}^K$. Similarly, the decomposition of $\mathbf{G}$ into its J-dimensional columns is denoted as $\mathbf{G} \;=\; [g_{\vec{j}1}, \cdots, g_{\vec{j}K}]$. Here also:

$$\mathbf{G}'\mathbf{D}_c\mathbf{G} \;=\; \mathbf{D}_\lambda^2 \qquad i.e. \;\; \sum_j p_{\cdot j} g_{jk}^2 \;=\; \lambda_k^2 \tag{46}$$

Thus, equation (46) decomposes the k-th eigenvalue of $\mathbf{S}'\mathbf{S}$ according to the contribution of each activity j, where $I^j \;=\; p_{\cdot j} g_{jk}^2$ measures the contribution of the activity j.

Summarizing, the SVD of $\mathbf{S}$ provides a decomposition of the total polarization ϕ^2 in terms of the contributions of each factor k and of the contribution of the regions i, respectively the activities j:

$$\phi^2 \;=\; \sum_i I_i \;=\; \sum_i \sum_k p_{i\cdot}\, f_{ik}^2 \;=\; \sum_j I^j \;=\; \sum_j \sum_k p_{\cdot j}\, g_{jk}^2 \tag{47}$$

For more details, see Mardia *et al.*(1979), Jobson (1992) and Greenacre (2007).

Let us write the rows and columns profiles as follows:

$$\mathbf{D}_r^{-1}\,\mathbf{P} \;=\; \left[\frac{p_{ij}}{p_{i\cdot}}\right] \;=\; [p_{j|i}] \qquad\qquad \mathbf{P}\,\mathbf{D}_c^{-1} \;=\; \left[\frac{p_{ij}}{p_{\cdot j}}\right] \;=\; [p_{i|j}] \tag{48}$$

Comparing, by means of a divergence, a profile with the corresponding marginal distribution provides a measure of relative specialization of region i and of relative industrial concentration of activity j:

$$d_{\chi^2}(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) \;=\; \sum_j \frac{(p_{j|i} \,-\, p_{\cdot j})^2}{p_{\cdot j}} \;=\; [p_{\vec{j}|i} - p_{\cdot\vec{j}}]' D_c^{-1} [p_{\vec{j}|i} - p_{\cdot\vec{j}}] \;=\; \sum_k f_{ik}^2 \tag{49}$$

$$d_{\chi^2}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \;=\; \sum_i \frac{(p_{i|j} \,-\, p_{i\cdot})^2}{p_{i\cdot}} \;=\; [p_{\vec{i}|j} - p_{\vec{i}\cdot}]' D_r^{-1} [p_{\vec{i}|j} - p_{\vec{i}\cdot}] \;=\; \sum_k g_{jk}^2 \tag{50}$$

Therefore, the decomposition of the Total inertia as a measure of polarization, in (47), may also be written in terms of average relative concentration or specialization:

$$\phi^2 \;=\; \sum_i p_{i\cdot}\, d_{\chi^2}(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) \;=\; \sum_j p_{\cdot j}\, d_{\chi^2}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \tag{51}$$

Equations (42) and (51) may also be interpreted in terms of divergences between row or columns profiles, or conditional distributions. More details, under a stochastic independence approach, are given in Haedo and Mouchart (2012).

4.3 Best collapsed table: the algorithm

Background

The concept of distance between regions or activities is provided by means of a "square of weighted Euclidean distances" (Greenacre 2011) among profiles.

Thus, the similarity between the profiles of two regions i and i' or two activities j and j' is measured as follows:

$$\sum_j \frac{1}{p_{\cdot j}} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}} \right)^2 = \sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} - p_{j|i'} \right)^2 = [p_{\vec{j}|i} - p_{\vec{j}|i'}]' D_c^{-1} [p_{\vec{j}|i} - p_{\vec{j}|i'}] \tag{52}$$

$$\sum_i \frac{1}{p_{i\cdot}} \left(\frac{p_{ij}}{p_{\cdot j}} - \frac{p_{ij'}}{p_{\cdot j'}} \right)^2 = \sum_i \frac{1}{p_{i\cdot}} \left(p_{i|j} - p_{i|j'} \right)^2 = [p_{\vec{i}|j} - p_{\vec{i}|j'}]' D_r^{-1} [p_{\vec{i}|j} - p_{\vec{i}|j'}] \tag{53}$$

The polarization of an economy decreases as a consequence of clustering and this loss of information is reduced by clustering the most similar regions or activities. Thus the algorithm chooses pairs of regions i and i' and pairs of activities j and j' minimizing the measures of dissimilarity (52) and (53).

Following Ward (1963)'s approach, the pair of regions (i, i') that gives the least decrease in inertia is identified by the pair of rows (i, i') which minimize the following measure:

$$\frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} + p_{i'\cdot}} \sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} - p_{j|i'} \right)^2 = \frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} + p_{i'\cdot}} [p_{\vec{j}|i} - p_{\vec{j}|i'}]' D_c^{-1} [p_{\vec{j}|i} - p_{\vec{j}|i'}] \tag{54}$$

The selected two rows are then merged by summing their frequencies and the profile and mass are recalculated. The same measure of difference as (54) is calculated at each stage of the clustering. We also operate similarly for merging two columns.

A collapsed table is characterized by two partitions: a partition $\mathcal{I}_*$ of the rows and a partition $\mathcal{J}_*$ of the columns. Thus a collapsed table is noted as $T_{\mathcal{I}_* \times \mathcal{J}_*}$ and is obtained by merging the rows and the columns of the original table according to the relevant partitions.

Hierarchical clustering, of the rows or of the columns, generates a nested sequence of $(I + 1)$ partitions of the rows and $(J + 1)$ partitions of the columns, with the first and the last ones being:

$$\mathcal{I}_{(0)} = \{\{1\}, \{2\}, \ldots, \{I\}\} \qquad \mathcal{J}^{(0)} = \{\{1\}, \{2\}, \ldots, \{J\}\} \tag{55}$$

$$\mathcal{I}_{(I)} = \{\{1, 2, \ldots, I\}\} \qquad \mathcal{J}^{(J)} = \{\{1, 2, \ldots, J\}\} \tag{56}$$

The other not extreme $(I - 1)$ and $(J - 1)$ partitions corresponds to the levels of a dendrogram.

In this section, we give the essentials of this algorithm. We shall use the example, in next section, to provide further details on the working of the algorithm.

First step: building collapsed tables.

Work on the rows. For $k = 1, 2, \ldots, K$:

Consider the first k columns of $\mathbf{F}$, let $\mathbf{F}_{(k)} = (f_{\vec{i}1}, \ldots, f_{\vec{i}l}, \ldots, f_{\vec{i}k})$ $I \times k$, where $f_{\vec{i}l}$ represents the l-th column of $\mathbf{F}$, and obtain a dendrogram through a hierarchical clustering of the rows of $\mathbf{F}_{(k)}$, corresponding to the rows of $\mathbf{S}$, as follows. Let $\mathcal{I}_{(n,k)}$ $n = 0, \ldots, I$, with $\forall k : \mathcal{I}_{(0,k)} = \mathcal{I}_{(0)}$ and $\mathcal{I}_{(I,k)} = \mathcal{I}_{(I)}$, be the nested sequence of partitions of regions, starting with $\mathcal{I}_{(0)}$ and with each following cluster obtained as an optimized clustering scheme based on $\mathcal{I}_{(n-1,k)}$. Thus $\forall k$, there are only $I - 1$ relevant levels of the hierarchical clustering.

Work on the columns. For $k = 1, 2, \ldots, K$:

Repeat the same with the columns of $\mathbf{G}$, namely $\mathbf{G}_{(k)} = (g_{\vec{j}1}, \ldots, g_{\vec{j}l}, \ldots, g_{\vec{j}k})$ $J \times k$ where $g_{\vec{il}}$ represents the l-th column of $\mathbf{G}$, and obtain a dendrogram through a hierarchical clustering of the rows of $\mathbf{G}_{(k)}$, corresponding to the columns of $\mathbf{S}$, as follows. Let $\mathcal{J}^{(m,k)}$ $m = 0, \ldots, J$, with $\forall k : \mathcal{J}^{(0,k)} = \mathcal{J}^{(0)}$ and $\mathcal{J}^{(J,k)} = \mathcal{J}^{(J)}$, be the nested sequence of partitions of activities, starting with $\mathcal{J}^{(0)}$ and with each following cluster obtained as an optimized clustering scheme based on $\mathcal{J}^{(m-1,k)}$. Thus $\forall k$, there are only $J - 1$ relevant levels of the hierarchical clustering.

Building collapsed tables.

For each level of the rows and columns dendrograms, build the $(I-1) \times (J-1)$ collapsed tables $T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}}$ and calculate the corresponding inertia $\phi^2 \left(T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}} \right)$.

Second step: identifying an optimal collapsed table.

Having built the array A of $\#(A) = (I-1)(J-1)K$ collapsed tables, the final question is: which of the collapsed tables is better in the sense of a best compromise between a smallest table that preserves the highest polarization (*i.e.* association) possible? Permutation bootstrapping provides a tool for a suitable compromise.

Bootstrapping.

Let us consider whether a particular table $T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}}$ is "best" in the sense alluded above. At least, we should check that this table is not dominated by a table obtained through a random shuffling of the labels (of rows and/or of columns) based on a same level of the dendrogram. The optimized tables from the dendrograms are completely identified by the three characteristics (n, m, k). Here, $\mathcal{I}_{n,k}$ is a partition $\mathcal{I}$ with $I - n$ elements, let $\{\mathcal{I}_1, \ldots, \mathcal{I}_{I-n}\}$. Let π_r be a permutation defined on $\mathcal{I}$, *i.e.* $\pi_r : \mathcal{I} \to \mathcal{I}$, bijective and let us write $\pi_r(\mathcal{I}_{n,k})$ for the image of the partition $\mathcal{I}_{n,k}$ transformed by π_r. Similarly, let π^c be a permutation defined on $\mathcal{J}$ and its image $\pi^c(\mathcal{J}_{m,k})$. Given (π_r, π_c), one may define a transformed table $T^{(k)}_{\pi^r(\mathcal{I}_{n,k}) \times \pi^c(\mathcal{J}_{m,k})}$, following the same partition scheme as the optimized table $T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}}$ with shuffled labels, and compute a corresponding inertia $\phi^2 \left(T^{(k)}_{\pi^r(\mathcal{I}_{n,k}) \times \pi^c(\mathcal{J}_{m,k})} \right)$. Note that the transformed table $T^{(k)}_{\pi^r(\mathcal{I}_{n,k}) \times \pi^c(\mathcal{J}_{m,k})}$ has a same dimension as $T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}}$; thus their inertia are comparable. The difference $\phi^2 \left(T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}} \right) - \phi^2 \left(T^{(k)}_{\pi^r(\mathcal{I}_{n,k}) \times \pi^c(\mathcal{J}_{m,k})} \right)$ is an effect of the label shufflings of the rows and of the columns. The permutation bootstrap is obtained by generating randomly the permutations (π_r, π_c) and evaluates the average, denoted as $\mathbb{E}_B$, of the corresponding inertia. The difference

$$\psi(n, m, k) = \phi^2 \left(T^{(k)}_{\mathcal{I}_{n,k} \times \mathcal{J}_{m,k}} \right) - \mathbb{E}_B \, \phi^2 \left[T^{(k)}_{\pi^r(\mathcal{I}_{n,k}) \times \pi^c(\mathcal{J}_{m,k}} \right] \tag{57}$$

represents how much the optimized table has gained, in inertia, relatively to a table with a same cluster scheme but with randomly shuffled individuals and variables. The algorithm terminates by defining the best collapsed table, $\mathcal{I}_{n^*,k^*} \times \mathcal{J}_{m^*,k^*}$ through the solution of the maximization problem:

$$(n^*, m^*, k^*) \;=\; arg\,max_{n,m,k}\;\psi(n, m, k), \tag{58}$$

balancing by so-doing the trade-off between the association degree and the table dimension.

Remark. In general, a clustering of a table involves a loss of information, measured by a decrease of the inertia. In the extreme cases, $T_{\mathcal{I}_{(I)}\times\mathcal{J}_{(J)}}$ is a 1×1 table representing a maximum level of clustering and maximum loss of information, whereas $T_{\mathcal{I}_{(0)}\times\mathcal{J}_{(0)}}$ is a $I \times J$ table representing the original table with no loss of information. In both cases, bootstrapping is irrelevant. ■

Finally, the Choropleth map of each country shows the grouped regions (g-regions) obtained from the optimal collapsed table at the last available period of the data.

5 Identification of specialized agglomerations

This section develops new statistical and computational methods for the automatic detection of agglomerations displaying an over- or under- relative specialization spatial pattern. A probability model is used to provide a basis for a space partition into clusters representing homogeneous portions of space as far as the probability of locating a economic 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. As a preliminary step we first present the notation for clusters of regions.

5.1 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:

$$\mathcal{I} \;=\; \{1, 2, \cdots, I\} \;=\; \bigcup_{m=1}^{M} \mathcal{I}_m \qquad \mathcal{I}_m \cap \mathcal{I}_{m'} \;=\; \emptyset \;\; (m \neq m') \qquad \#(\mathcal{I}_m) \;=\; I_m \qquad \sum_m I_m \;=\; I \tag{59}$$

Let us define accordingly

$$N_{m\cdot} = \sum_{i\in\mathcal{I}_m} N_{i\cdot} \qquad N_{m,j} = \sum_{i\in\mathcal{I}_m} N_{ij} \tag{60}$$

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

$$g_{m\cdot} \;=\; \sum_{i\in\mathcal{I}_m} p_{i\cdot} = \frac{N_{m\cdot}}{N_{\cdot\cdot}} \qquad g_{m|j} \;=\; \sum_{i\in\mathcal{I}_m} p_{i|j} = \frac{N_{m,j}}{N_{\cdot j}} \tag{61}$$

$$g_{\vec{m}\cdot} = (g_{1\cdot}, \cdots, g_{m\cdot}, \cdots, g_{M\cdot}) \qquad g_{\vec{m}|j} = (g_{1|j}, \cdots, g_{m|j}, \cdots, g_{M|j}) \tag{62}$$

$$p_{i\cdot|m} = \frac{p_{i\cdot}}{g_{m\cdot}}\, 1\!\!1_{\{i\in\mathcal{I}_m\}} = \frac{N_{i\cdot}}{N_{m\cdot}}\, 1\!\!1_{\{i\in\mathcal{I}_m\}} \qquad p_{i|j,m} = \frac{p_{i|j}}{g_{m|j}}\, 1\!\!1_{\{i\in\mathcal{I}_m\}} = \frac{N_{ij}}{N_{m,j}}\, 1\!\!1_{\{i\in\mathcal{I}_m\}} \tag{63}$$

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:

$$\Omega_{\mathcal{I}_{m|C}} = \bigcup_{i\in\mathcal{I}_{m|C}} \Omega_i \qquad \text{with} \quad \bigcup_{m=1}^{M} \Omega_{\mathcal{I}_{m|C}} = \Omega \tag{64}$$

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

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

5.2 A structural model

We now introduce a model aimed at representing how the data have been generated and eventually may be interpreted; the notation eventually 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 involves three categorical random elements, namely: regions (i), activity (j) and cluster (m). When drawing randomly a economic 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:

$$\pi_{i,j,m|C} = \pi_{\cdot j\cdot}\ \ \gamma_{m|j;C}\ \ \rho_{i|j,m;C} \qquad\qquad i \in \mathcal{I}_{m|C} \tag{65}$$

Notice that, from (64) we have

$$\pi_{i\cdot|j} = \frac{\pi_{i,j,\cdot}}{\pi_{\cdot,j,\cdot}} \qquad\qquad \gamma_{m|j;C} = \sum_{i\in\mathcal{I}_{m|C}} \pi_{i|j} \qquad\qquad \gamma_{m|C} = \sum_{i\in\mathcal{I}_{m|C}} \pi_{i\cdot\cdot} \tag{66}$$

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 this work:

$$H_m : \qquad \rho_{i|j,m;C} = \rho_{i|m;C} \quad \text{with} \sum_{i \in \mathcal{I}_{m|C}} \rho_{i|m;C} = 1 \tag{67}$$

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. Moreover, because:

$$\rho_{i|m;C} = \frac{\pi_{i\cdot\cdot}}{\gamma_{m|C}}, \tag{68}$$

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, $\gamma_{m|C}$. Thus in (68) 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 an activity in general. Under our maintained hypothesis we have:

$$\pi_{i,m|j;C} = \gamma_{m|j;C}\ \rho_{i|m;C} \qquad \forall i \in \mathcal{I}_{m|C} \tag{69}$$

The algorithm, to be sketched in Section 5.4, 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 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 $X_j = [x^u_{ij}]$ where $i = 1, \cdots, I_j$ in columns and u, in rows, runs over the set $\mathcal{A}_j$ of the economic units entering the relevant regions for the analysis of the activity j with $A_j = \#(\mathcal{A}_j)$. As X_j is an incidence matrix with elements equal to x^u_{ij} for $u \in \mathcal{A}_j$, the sum of each row is equal to 1.

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. It is shown, in Haedo and Mouchart (2015), that the probability of the data matrix X_j may be factorized into:

$$p(X_j \mid C) = \left[\prod_{1 \le m \le M(C)} \gamma_{m|j;C}^{N_{m,j|C}} \right] \cdot [\, b(X_j)\,] \tag{70}$$

where

$$N_{m,j|C} = \sum_{i \in \mathcal{I}_{m|C}} N_{ij} \qquad\qquad b(X_j) = \prod_{1 \le m \le M(C)} \prod_{i \in \mathcal{I}_{m|C}} \rho_{i|m;C}^{N_{ij}} \tag{71}$$

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

$$L(X_j \mid C) = L_1(\gamma_{\vec{m}|j;C} \mid X_j)\ L_2([\rho_{i|m;C}] \mid X_j) \tag{72}$$

5.3 The concept of specialized cluster in a structural model

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 $\pi_{i\cdot}$ (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 (13) this is equivalent to the location quotients being significantly larger, or smaller, than 1. Under the maintained hypothesis (67), it may be seen from (70) 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 (66):

$$H_0: \qquad \gamma^{(0)}_{m|j;C} = \gamma^{(0)}_{m|C} \qquad m = 1, \cdots, M \tag{73}$$

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:

$$\widehat{\gamma}^{(0)}_{m|j;C} = \widehat{\gamma}^{(0)}_{m|C} = \frac{N_{m\cdot|C}}{N_{\cdot\cdot}} \tag{74}$$

The absence of industrial concentration for activity j, underlying (74), is relative to a particular cluster scheme C. The estimated log likelihood under H_0 is

$$\ln \widehat{L}^{(0)}(X_j \mid C) = \sum_{1 \le m \le M(C)} N_{m,j|C}\ \ln\left(\frac{N_{m\cdot|C}}{N_{\cdot\cdot}}\right) + \ln a(X_j) \tag{75}$$

where $a(X_j)$ corresponds to the term $L_2([\rho_{i|m;C}] \mid X_j)$ in (72) 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

$$\widehat{\gamma}^{(1)}_{m|j;C} = \frac{N_{m,j|C}}{N_{\cdot j}} \tag{76}$$

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

$$\ln \widehat{L}^{(1)}(X_j \mid C) = \sum_{1 \le m \le M(C)} N_{m,j|C}\ \ln\left(\frac{N_{m,j|C}}{N_{\cdot j}}\right) + \ln a(X_j) \tag{77}$$

It is shown in Haedo and Mouchart (2015) that the log-likelihood-ratio statistic to test H_0 against H_1, under H_m may be written as:

$$T(X_j \mid C) = 2\left[\sum_{1 \le m \le M(C)} N_{m,j|C}\ \ln\left(\frac{N_{m,j|C}/N_{\cdot j}}{N_{m\cdot|C}/N_{\cdot\cdot}}\right)\right] = 2\ N_{\cdot j}\, d(p_{\vec{m}|j} \mid g_{\vec{m}\cdot}) \tag{78}$$

where $d(\cdot \mid \cdot)$ is the (non-symmetric) Kullback-Leibler divergence between two distributions. Therefore, the test statistic (78) 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). Note that the argument of the logarithms in (78) is the location quotient obtained after clustering the regions. Notwithstanding the existence of agglomeration economies, economic units' location decision is modeled as independent of these agglomeration economies as the model is essentially a static one that does not present the dynamics of agglomeration formation. The weight $N_{m,j|C}$ in (78) of the logarithm of LQ may be viewed as a solution to a small areas problem in line with the works of Moineddin *et al.*(2003), O'Donoghue and Gleave (2004) and Guimarães *et al.*(2003 and 2009).

In (73), 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

$$p - value = 1 - F_{M(C)-1}(T(X_j \mid C)) \tag{79}$$

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.

5.4 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:

$$w : \mathcal{I} \times \mathcal{I} \longrightarrow \{0,1\} \qquad \text{where} \quad w(i_1, i_2) = \mathbb{1}_{\{i_1 \text{ and } i_2 \text{ are contiguous}\}} \tag{80}$$

Note that W is symmetric ($W = W'$) with 1s 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:

$$v(\mathcal{I}_{m|C}) = \{i_1 \in \mathcal{I} \setminus \mathcal{I}_{m|C} \mid \exists\, i_2 \in \mathcal{I}_{m|C} : w(i_1, i_2) = 1\} \tag{81}$$

Remark on the W matrix. The concept of contiguity underlying (80) 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). ■

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 (78).

It should be noticed that a particular cluster scheme generates a particular model; indeed, from (65), 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. A natural model selection procedure may be based on the Bayesian information criterion (BIC), that naturally involves the divergence (78). Thus we consider the optimization problem:

$$C^* = \arg\max_{C} BIC(X_j \mid C) \tag{82}$$

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

$$BIC(X_j \mid C) = T(X_j \mid C) - (M(C) - 1)\ \ln\,(N_{\cdot\cdot}) \tag{83}$$

with $T(X_j \mid C)$ as defined in (78).

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, generally made of all regions with $N_{ij} > 0$; in some cases, we might also select the regions with N_{ij} larger than some specified minimum. The algorithm generates, in 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:

$$C^*_{[0]} = C_{max} \tag{84}$$

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 and a second cluster consisting of all the remaining regions. The 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. This step is completed by evaluating $BIC(X_j \mid C^*_{[1]})$. If $BIC(X_j \mid C^*_{[1]}) = BIC(X_j \mid C^*_{[0]})$, the algorithm stops.

Step 2. The second step looks for a second region, to be chosen from the $I - 1$ remaining regions of the previous step, by maximizing the configuration criteria. Therefore, the outcome of $C^*_{[2]}$ will depend on the cluster configuration of the previous step $C^*_{[1]}$. At most three clusters configurations ($M(C) = 3$) should have been formed: i) one cluster with the region chosen at the first step; ii) another cluster with the region chosen at the second step; and iii) a third cluster with the remaining regions. If however the two regions chosen at the first two steps are contiguous then the number of clusters is two ($M(C) = 2$): one with the two chosen regions and the other one with the remaining regions. 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. Note that some agglomerations might be contiguous without being merged into a unique one because of too different values of the location 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 (78) and (83).

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 $C^*_{[k^*]}$. As such, this stopping provides a "myopic" algorithm that stops once the objective decreases for a first time. The actual stopping rule of the algorithm is completed so as to protect against the possibility of local optima; more detail are given in Haedo and Mouchart (2015).

Finally, the Choropleth map of each country shows the over-specialized agglomerations resulting of the optimal cluster scheme of each activity j separately at last available period.

5.5 Testing the role of contiguity for specialized agglomerations

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 $d(p_{\vec{\imath}|j} \mid p_{i\cdot})$. 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. Indeed, as a difference from geo-localized data, lattice data provides no information on the intra-regional localizations. But a major role of contiguity, among regions, in the formation of specialized agglomerations provides evidence that, for a specific activity, localization economies, inside an agglomeration, have more impact. Formally, we want to test the null hypothesis that the vector of the location quotients, for a fixed activity j, is invariant for the group of permutations of its coordinates.

The 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 (2006). 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 $\widehat{F}^B_{BIC}$ 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:

$$\begin{aligned} p\text{-value}[BIC(X_j \mid C^*)] &= 1 - \widehat{F}^B_{BIC}(BIC(X_j \mid C^*)) \\ &= \frac{1}{B} \sum_{1 \le b \le B} 1\!\!1_{\{(BIC^b(X_j|C^*_b) > BIC(X_j|C^*)\}} \end{aligned} \tag{85}$$

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 $\widehat{F}^B_{BIC}$.

5.6 The main parameters of the algorithm

The algorithm treats each activity j independently and requires the specification of several parameters. The main parameters are:

- contiguity matrix: simple first order queen or rook matrix; see Remark on the W matrix in Section 5.4) (selected: 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 (selected: $N_{ij} > 4$);

- 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 $N_{\cdot\cdot}$ (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 (selected: 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 (selected: sign = TRUE; more explicitly: if $i^*_{[1]}$ and i_2 are aggregated into a same cluster, then the log of the location 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 (selected: B = 1000);
- criterion: BIC or AIC (selected: 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 (selected: change of sign);

  - secondly the width of the window within which is evaluated the change of sign or the slope of the criterion (selected: h = 60).

5.7 Manufacturing competitiveness

For each region i of a country, a composite manufacturing competitiveness index, $Mc_{[i\cdot]}$, may be evaluated as follows:

$$Mc_{[i\cdot]} = NMaler_{[i\cdot]} + NMqler_{[i\cdot]} + NMl_{[i\cdot]} + NMp_{[i\cdot]} + NMre_{[i\cdot]}[Meu_{ij}] + NMre_{[i\cdot]}[Mem_{ij}] + NMae_{[i\cdot]} \tag{86}$$

where $NMre_{[i\cdot]}$ is the regional specialization of regions i (15) based on manufacturing economic units $[Meu_{ij}]$ and manufacturing employment $[Mem_{ij}]$ and $Mae_{[i\cdot]}$ is the number of over-specialized agglomerations $\mathcal{I}_{m|C}$ of all activities j of which region i is a part (for more details see Section 5).

In the composite $Mc_{[i\cdot]}$ index, the indicator $Mp_{[i\cdot]}$ may be reclassified as follows:

$$\begin{aligned} 2 =&\quad \text{when } Mp_{[i\cdot]} = 1 \text{ or } Mp_{[i\cdot]} = 3; \\ 1 =&\quad \text{when } Mp_{[i\cdot]} = 2 \text{ or } Mp_{[i\cdot]} = 5; \\ 0 =&\quad \text{when } Mp_{[i\cdot]} = 4 \text{ or } Mp_{[i\cdot]} = 6. \end{aligned}$$

Finally, every indicator of the composite $Mc_{[i\cdot]}$, generically named $X_{[i\cdot]}$, may be normalized as follows:

$$N(X_{[i\cdot]}) = \frac{X_{[i\cdot]} - \min\{X_{[i\cdot]}\}}{\max\{X_{[i\cdot]}\} - \min\{X_{[i\cdot]}\}} \times 100 \tag{87}$$

Thus, each indicator is bounded between 0 and 100 points, while the $Mc_{[i\cdot]}$ is bounded between 0 and 700 points.

The Choropleth map of each country shows the values of $Mc_{[i\cdot]}$ at last available period, aggregated into three classes: high, medium and low, using the Jenks Natural Breaks classification method.

For international comparisons, the $Mc_{[i\cdot]}$ has been computed for the global country simply dividing it by the number of regions i with al least one economic unit.

6 A retrospective view

This Atlas gathers a set of several contributions. As a first step, data are collected for a number of countries of this continent. These data refer basically to the employment and the firms in the manufacturing sectors. This process is continuously ongoing as long as data for new countries are still in the process of collection. Then, the collected data are scrutinized for evaluating the design of the collecting process and thereafter summarized by means of a set of (rather standard) descriptive measures and of illustrative maps. This Atlas also makes original contributions in the field of detecting *relatively* specialized groups of regions and/or activities, for which two innovative algorithms are used.

A first algorithm, to be called the Agglomerative Algorithm (AgA), constructs a partition of the regions into agglomerations, *i.e.* clusters of contiguous regions, characterized by a property of relative specialization with respect to a particular activity. A second algorithm, to be called a Bi-clustering Algorithm (BcA), proposes a simultaneous regrouping of regions and activities with an objective of a better synthetic view of the regional manufacturing structure of an economy.

These contributions have lead to a number of insightful developments. Of particular interest is comparing the spatial structure of a given activity in different countries or of different activities in a given country.

AgA has been shown to be particularly well suited for comparing the spatial structure of a given activity in different countries or of different activities of a given country:

- The final bootstrap step of AgA has been used to analyze more in depth the role of contiguity in the formation of agglomeration, in particular for evaluating the presence of intra- or inter-regional localization economies.
- The agglomerations identified, for a given activity, by AgA have been compared in terms (i) of the dimension of firms, (ii) of the distribution of activities within the agglomeration, (iii) of the intensity of the specialization of the agglomeration, in terms of the average of the location quotients LQ_{ij} of the agglomeration or in terms of the global LQ_{ij} of the agglomeration or of the log of the location quotient weighted by the corresponding N_{ij}, (iv) of the other activities where the agglomeration is also specialized .

Given that the set of activities is the same for every country, BcA allows one to compare the best collapsed tables for two countries following one of two complementary strategies. A first one considers every g-region with all activities and compute the discrepancy between the activity distributions for every pair of g-regions. A second strategy is based on an ordering of the activities for each g-region and evaluates a rank correlation for the activities of every pair of g-regions. The ordering of the activities may be either the order of the weight of the activities or the intensity of the specialization of each activity within each g-region.

References

Aggarwal, C., and Reddy, C. (2013) (eds.), *Data Clustering: Algorithms and Applications.* Boca Raton, Florida: Chapman & Hall/CRC.

Anas, A., Arnott, R., and Small, K. (1998), Urban spatial structure. *Journal of Economic Literature* **36**: 1426-1464.

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

Arbia, G. (2001), The role of spatial effects in the empirical analysis of regional concentration. *Journal of Geographical Systems* **3**: 271-281.

Arbia, G., and Espa, G. (1996), *Statistica Economica Territoriale.* Padova: CEDAM.

Arbia, G., Espa, G., and Quah, D. (2008), A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics* **34**: 81-103.

Bertin, J. (2010), *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands, CA: Esri Press.

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

Branson, D. (2000), Stirling numbers and Bell numbers: their role in combinatorics and probability. *Mathematical Scientist* **25**: 1-31.

Donato, V. (2002), Políticas públicas y localización industrial en Argentina. CIDETI Working Paper 2002/01, Alma Mater Studiorum-Università di Bologna in the Argentine Republic.

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

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2010), *Cluster Analysis.* Chichester: John Wiley & Sons.

Florence, P. (1939), *Report of the Location of Industry.* Political and Economic Planning, London, UK.

Forgy, E.W. (1965), Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21**: 768-769.

Gan, G., Ma, C., and Wu, J. (2007), *Data Clustering: Theory, Algorithms, and Applications.* Philadelphia: ASA-SIAM Series on Statistics and Applied Probability.

Gardner, M. (1978), The Bells: versatile numbers that can count partitions of a set, primes and even rhymes. *Scientific American* **238**: 24-30.

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

Greenacre, M. J. (2007), *Correspondence Analysis in Practice.* Boca Raton, Florida: Chapman & Hall/CRC.

Greenacre, M. J. (2011), A simple permutation test for clusteredness. Working Paper 555, Barcelona GSE Working Paper Series.

Guimarães, P., Figueiredo, O., and Woodward, D. (2003), A tractable approach to the firm location decision problem. *Review of Economics and Statistics* **84**: 201-204.

Guimarães, P., Figueiredo, O., and Woodward, D. (2009), Dartboard tests for the location quotient. *Regional Science and Urban Economics* **39**: 360-364.

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 (I).

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

Haedo, C., and Mouchart, M. (2014), Automatic grouping of regions and activities *Alias: Best Collapsed Tables*, in progress.

Haedo, C., and Mouchart, M. (2015), Specialized Agglomerations with Lattice Data: Model and Detection. *Spatial Statistics* **11**: 113-131.

Hartigan, J.A. (1975), *Clustering Algorithms.* New York: John Wiley & Sons.

Jenks, G.F. (1967), The data model concept in statistical mapping. *International Yearbook of Cartography* **7**: 186-190.

Jobson, J. (1992), *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* New York: Springer-Verlag.

Kaufman, L., and Rousseeuw, P. J. (2005), *Finding Groups in Data. An Introduction to Cluster Analysis.* Hoboken, New Jersey: John Wiley & Sons.

Krugman, P. (1991), Increasing returns and economic geography. *Journal of Political Economy* **99**: 483-499.

Lawson, A. (2006), *Statistical Methods in Spatial Epidemiology.* New York: Wiley.

Liseikin, V.D. (2010), *Grid Generation Methods.* Dordrecht: Springer.

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

Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis.* London: Academic Press.

Moineddin, R., Beyene, J., and Boyle, E. (2003), On the location quotient confidence interval. *Geographical Analysis* **35**: 249-256.

Mori, T., and Smith,T. (2011), An industrial agglomeration approach to central place and city size regularities. *J. Reg. Sci.* **51**: 694-731.

Nathan, M., and Overman, H. (2013), Agglomeration, clusters, and industrial policy. *Oxford Review of Economic Policy* **29**: 383-404.

O'Donoghue, D., and Gleave, B. (2004), A note on methods for measuring industrial agglomeration. *Regional Studies* **38**: 419-427.

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

Perroux, F. (1950), Economic space: theory and applications. *Quarterly Journal of Economics* **64**: 89-104.

Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Jon Kimerling, A., and Guptil, S.C. (1995), *Elements of Cartography.* New York: John Wiley & Sons.

Rota, G-C. (1964), The number of partitions of a set. *American Mathematical Monthly* **71**: 498-504.

Sloane, N.J.A. (2001), Bell numbers. In Hazewinkel, M. (Ed.), *Encyclopedia of Mathematics.* New York: Springer.

Thompson, J.F., Soni, B.K., and Weatherill, N.P.(1999) (eds.), *Handbook of Grid Generation.* Boca Raton, Florida: CRC Press.

Ward, J. H., Jr. (1963), Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**: 236-244.

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.