INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

DISCUSSION PAPER

2016/42

Automatic biclustering of regions and sectors

Haedo, C. and M. Mouchart

Automatic biclustering of regions and sectors

Christian HAEDO[a] and Michel MOUCHART[b]

[a] *Centro 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), Université catholique de Louvain, Belgium*

October 29, 2016

Abstract

This paper develops an automatic grouping procedure of regions and sectors based on a greedy biclustering algorithm, focused simultaneously on their relative regional specialization and relative industrial concentration. Based on an association measure for a contingency table (regions × sectors), this procedure enables to i) significantly reduce the size of the original table and obtain an optimal collapsed table with low level of information loss vis-à-vis the degree of global localization; and ii) identify the homogeneous regions according to the industrial structure in terms of sub- and over- specialization in large two-way contingency tables. The properties and results of the algorithm are discussed through the presentation of three applications, namely Argentina, Brazil and Chile. In particular, an object of discussion, in the case of Brazil, is the result that the number of cells of the original table is reduced by 99% while the lost global localization information is 23%.

Keywords: relative regional specialization, relative industrial concentration, biclustering, hierarchical clustering, correspondence analysis, large two-way contingency tables, permutation bootstrap.

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

Contents

1 Introducing the topic of this paper

Background. In the statistical analysis of specialization and concentration for understanding the regional pattern of the economic activity, the extension of the regions and of the sectors is known to have a prominent role. This paper is focused on simultaneously grouping regions and sectors according to their *relative* regional specialization and *relative* industrial concentration. Several challenges are at stake, three of which should be singled out. A first one consists in identifying regions with identical, or very similar, sectorial structure. These regions might be associated in view of developing a joint policy of between- and within-regional cooperation. At contrast, a second challenge consists in identifying regions with complementary sectorial structure. These regions might be vertically integrated in a chain of value built in view of between- and within-regional cooperation. It should be clear that any regrouping, of regions and/or of sectors, involves a loss of information. The third challenge consists in obtaining a regrouping involving a controled loss of information while the reduction of the overall dimension is substantial. This is indeed a crucial dilemma with large tables.

Following the Stochastic Independence Approach (SIA), as developed in Haedo and Mouchart (2012), these relative measures of specialization and of concentration are based on the comparison of two distributions, namely the profile (or, conditional distribution) of the region, or of the sector, and the corresponding marginal distribution. At the country level, the average of these relative measures provides a concept of *global localization*[1] of the economic activity.

The purpose of this paper is to extend in two directions the analysis of grouping for the evaluation characterization of global localization. Firstly, simultaneous groupings of regions and sectors, rather than separate ones are examined. Secondly, instead of considering arbitrarily pre-specified groupings, the algorithm provides an automatic construction of grouping aimed at giving optimal groupings according to a pre-specified criterion. The aim is to simultaneously regroup regions with a similar industrial structure in terms of relative over- and under-specialization and sectors with a similar regional pattern in terms of relative over- and under-concentration. Shortly said, we look for a summary of a large regions $\times$ sectors contingency table that keeps (almost) unchanged the measure of global localization of the economic activity. This paper is an extension of a chapter of Haedo (2009).

The algorithm developed in this paper is in the family of so-called biclustering (Mirkin, 1996), block clustering (Govaert and Nadif, 2008, 2010; Keribin *et al.* 2015), co-clustering (Govaert and Nadif, 2013), or two-mode clustering (Madeira and Oliveira, 2004; van Mechelen, Bock and de Boeck, 2004; Jagalur *et al.* 2007) which allows simultaneous clustering of the rows and columns of a matrix, an approach dating back to Hartigan (1972), Braverman *et al.* (1974), Govaert (1977), Bock

[1] "Global localization" has also been called "Polarization". Bickenbach and Bode call the global measures of localization measures of polarization in 2006 but of localization in 2008; in 2010 Bickenbach, Bode and Krieger-Boden also use localization. In the sequel "global localization" is only used.

(1979), Gilula (1986). Biclustering methods are aimed at designing in a same exercise a clustering of the rows and the columns of a large array of data. These methods are expected to be useful to summarize large data sets by dramatically smaller data sets with a similar structure.

Furthermore, the biclustering algorithm developed in this paper is also in the family of so-called deterministic procedures, that operate in the spirit of descriptive statistical methods or of data mining, see for instance, Duffy and Quiroz (1991), Lebart and Mirkin (1993), Govaert (1995), Tibshirani *et al.* (1999), Cheng and Church (2000), Tang *et al.* (2001), Ciampi, González Marcos and Castejón Lima (2005), Banerjee *et al.* (2007), Charrad (2009), Caldas and Kaski (2011), Benabdeslem and Allab (2013). Most of these works are based on the salient results about the links and the complementarity between clustering and factor analysis of contingency tables, reconciling two different accents: the symmetry of the roles of rows and columns in the process, and the property of distributional equivalence (Benzécri, 1973; Escofier, 1978; Jambu, 1978; Goodman, 1981, 1985; Hirotsu, 1983; Cazes, 1986; Gilula, 1986; Greenacre, 1988), which allows for a greater stability of the results when agglomerating elements with similar profiles.

For large tables, trying all possibilities of grouping is not feasible. Therefore, a greedy algorithm that only ensures a local optimum is preferred. This is obtained by means of a technique of Hierarchical Clustering (HC), according to a dendrogram approach, combined with a Correspondence Analysis (CA), thus the HCCA procedure. 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. When our algorithm looks for collapsing simultaneously regions and sectors, no restriction is considered about the regions or the sectors 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 sectors, no consideration of inter-sectorial relationship, nor of value chain, is operating because only a similar regional *relative* industrial concentration is at stake.

The specificity of the present algorithm is its genuine orientation towards the particular needs and motivation of the spatial analysis of regional specialization and of industrial concentration.

The order of exposition is as follows. After this first section giving an informal introduction to the topic of this paper, a second section provides a more explicit conceptual background. Third section describes the object of this paper, namely a fully automatic algorithm of simultaneous grouping of regions and sectors. Discussion of the properties and results of the algorithm is made through the presentation of three applications, namely Argentina, Brazil and Chile, in the fourth section. The last section gathers some concluding remarks.

2 The conceptual background of the algorithm

The finite framework for the analysis of specialization and concentration may be presented as follows. For a given country, a finite set of disjoint regions $i \in \mathcal{I} = \{1, ..., I\}$, and a finite set of sectors $j \in \mathcal{J} = \{1, ..., J\}$ are considered. For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, a number N_{ij} of primary units is observed; these could be for instance number of employees or number of establishments. These data refer to lattice or areal data as the N_{ij} are characteristics of the area defined by region i. Putting together these data, a two-way $I \times J$ contingency table $\mathbf{N} = [N_{ij}]$ is obtained, with the row, column and table totals denoted as follows:

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

The issues of specialization of regions in terms of sectors and of concentration of sectors within regions are to be analyzed from the contingency table $\mathbf{N}$ in terms of profiles, or conditional distributions, characterizing regions and sectors.

This paper is focused on simultaneously grouping regions and sectors according to their *relative* specialization and concentration, respectively. Following the Stochastic Independence Approach (SIA), as developed in Haedo and Mouchart (2012), the relative measures of specialization and of concentration are based on the comparison of two distributions, by means either of a distance or of a divergence. The term discrepancy is used to designate either one or the other one and $d(q \mid r)$ denotes the discrepancy of distribution q with respect to distribution r. When $d(\cdot \mid \cdot)$ is not symmetric, as may be the case with a divergence, the distribution r acts as a benchmark against which distribution q is to be calibrated evaluated.

More specifically, the relative specialization of region i is measured by a discrepancy $d(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$ that operates a comparison between its profile (or conditional distribution)[2] of the i-th row $p_{\vec{j}|i} = (p_{1|i}, \cdots, p_{j|i}, \cdots, p_{J|i})$ and the global row profile (or marginal distribution) taken as a benchmark of no specialization $p_{\cdot\vec{j}} = (p_{\cdot 1}, \cdots, p_{\cdot j}, \cdots, p_{\cdot J})$, where $p_{j|i} = \frac{N_{ij}}{N_{i\cdot}}$ and $p_{\cdot j} = \frac{N_{\cdot j}}{N_{\cdot\cdot}}$. Similarly, the relative concentration of sector j is measured by a discrepancy $d(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ that operates the comparison between the profile (or conditional distribution) of the j-th column $p_{\vec{i}|j} = (p_{1|j}, \cdots, p_{i|j}, \cdots, p_{I|j})$ and the global column profile (or marginal distribution) $p_{\vec{i}\cdot} = (p_{1\cdot}, \cdots, p_{i\cdot}, \cdots, p_{I\cdot})$, where $p_{i|j} = \frac{N_{ij}}{N_{\cdot j}}$ and $p_{i\cdot} = \frac{N_{i\cdot}}{N_{\cdot\cdot}}$. The discrepancy $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$ compares the actual bivariate distribution, on regions $\times$ sectors, with the product of their marginal distributions that represents the closest distribution revealing independence between regions and sectors, taken as a benchmark of a completely non-concentrated, or non-specialized, economy. This measure, taken as a measure of global localization of the country, characterizes the regional pattern of the economic activities, or equivalently the distribution of the sectors among the regions. As a matter of fact, this discrepancy represents

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

the (global) information provided by the contingency table $\mathbf{N}$ over the global localization of the economy. Table 1 presents a synthetic view of these different concepts. Considering the product

Table 1: *Some conventional definitions*

Discrepancy	**Measured concept**
$d(p_{\vec{j}\mid i} \mid p_{.\vec{j}})$	Relative specialization of region i
$d(p_{\vec{i}\mid j} \mid p_{\vec{i}.})$	Relative concentration of sector j
$d([p_{ij}] \mid [p_{i.}\, p_{.j}])$	Global localization

of the marginal distributions as a benchmark of a completely non-concentrated, or non-specialized, economy allows one to analyze the contribution of each cell (i, j) to the formation of the global localization of an economy. This contribution is a result of confronting the value of p_{ij} with that of $p_{i.}\, p_{.j}$. Table 2 presents three ways for that confrontation, at the cell level, along with a way of aggregating these measures into a global one that takes the form of a discrepancy $d([p_{ij}] \mid [p_{i.}\, p_{.j}])$. The second column of Table 2 also shows that each cell is equivalently characterized by the Location Quotient[3] LQ_{ij}. This quotient, well-known in the literature on relative regional specialization and on relative industrial concentration, may be written as follows:

$$LQ_{ij} = \frac{p_{ij}}{p_{i.}\, p_{.j}} \tag{2}$$

and is a local indicator, for the cell (i, j) of the contingency table, that reveals the following feature of sector j in region i:

$$\begin{array}{lllll} LQ_{ij} & = 1 & \text{or} & p_{ij} = p_{i.}\, p_{.j} & \text{non-specialization} \\ & > 1 & \text{or} & p_{ij} > p_{i.}\, p_{.j} & \text{over-specialization} \\ & < 1 & \text{or} & p_{ij} < p_{i.}\, p_{.j} & \text{under-specialization} \end{array} \tag{3}$$

Once a measure of global localization is considered as an adequate summary representation of the regional structure of the economic activity, an algorithm for optimal groupings of rows and columns should minimize the loss of that measure. The three approaches sketched in Table 2 suggest three families of that algorithm. The KL family uses the expression of "loss of information" because of its roots in information theory. The $\mathcal{X}^2$ family rather speaks of "inertia" and is the one used in this paper.

[3]The well stablished Location Quotient (Florence, 1939), is also known as the (estimated) *Hoover-Balassa coefficient* for the cell (i, j). More information may also be found *e.g.* in Haedo and Mouchart (2012).

Table 2: *Cell contribution to the global localization*

Three approaches	**At the cell level**	**Derived discrepancy**
χ^2-divergence or inertia (χ^2)	$= p_{ij} - p_{i\cdot}\, p_{\cdot j}$	$= \sum_i \sum_i \frac{(p_{ij} - p_{i\cdot}\, p_{\cdot j})^2}{p_{i\cdot}\, p_{\cdot j}}$
	$= p_{i\cdot}\, p_{\cdot j}\, (LQ_{ij} - 1)$	$= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j} (LQ_{ij} - 1)^2$
Kullback-Leibler divergence (KL)	$= \frac{p_{ij}}{p_{i\cdot}\, p_{\cdot j}}$	$= \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot}\, p_{\cdot j}}\right)$
	$= LQ_{ij}$	$= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, LQ_{ij} \log(LQ_{ij})$
Hellinger-distance (H)	$= \sqrt{p_{ij}} - \sqrt{p_{i\cdot}\, p_{\cdot j}}$	$= \frac{1}{2} \sum_1 \sum_j (\sqrt{p_{ij}} - \sqrt{p_{i\cdot}\, p_{\cdot j}})^2$
	$= \sqrt{p_{i\cdot}\, p_{\cdot j}}\ \left[\sqrt{LQ_{ij}} - 1\right]$	$= \frac{1}{2} \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, (\sqrt{LQ_{ij}} - 1)^2$

3 Biclustering of regions and sectors

3.1 Motivation of the algorithm

Our purpose is to summarize the original information contained in the complete contingency table $\mathbf{N} = [N_{ij}]$, in order to extract from the data the most relevant patterns of global localization.

The nature of the actual challenge should be kept in mind. In the case of Argentina, for instance, there are $I = 523$ regions. Using just the first 2 digits of the International Standard Industrial Classification of manufacturing sectors (ISIC-Rev.3.1) there are $J = 23$ sectors. In 1994, for example, the total number of employees is $N = 1083928$. Therefore the contingency table is a 523×23 matrix of 1083928 primary units spread in 12029 cells. Thus it should be expected that many cells have either a very small number of employees or no employee at all. The skeleton of the proposed algorithm may be viewed as follows. An optimal grouping of regions and sectors should compromise between two opposite desiderata: the collapsed table should be as small as possible but should also display a minimum loss of global localization of the country.

Collapsing tables means building tables of smaller dimension through aggregated regions (rows) and/or sectors (columns). The total number M of possible collapsed tables[4] for the $I \times J$ matrix $\mathbf{N}$ is:

$$M = \sum\nolimits_{(m_1 \ldots m_i \ldots m_l)} \binom{I}{m_1 \ldots m_i \ldots m_l} \times \sum\nolimits_{(n_1 \ldots n_i \ldots n_k)} \binom{J}{n_1 \ldots n_j \ldots n_k} \tag{4}$$

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, a greedy algorithm that only ensures a local optimum is preferred. In the present case,

[4]Equation (4) 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 (4), we have $n = I$ and $m = J$. More information may be found in Rota (1964), Gardner (1978), Branson (2000) or Sloane (2001).

it is based on Hierarchical Clustering (HC), according to a dendrogram approach, combined with a Correspondence Analysis (CA), thus an HCCA procedure.

3.2 Singular Value Decomposition and Correspondence Analysis

The matrix $\mathbf{R} = [p_{ij} - p_{i\cdot}\, p_{\cdot j}]$ is a matrix of residuals between an observed quantity p_{ij} and an expected one, under an hypothesis of independence, $p_{i\cdot}\, p_{\cdot j}$. Let $\mathbf{P} = \frac{\mathbf{N}}{N_{\cdot\cdot}}$ be the probability matrix corresponding to $\mathbf{N}$, $\mathbf{r} = (p_{i\cdot})$ the vector of row marginals, $\mathbf{c} = (p_{\cdot j})$ the vector of column marginals and $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices formed with the row marginals and column marginals, respectively. The matrix of the residuals may be conveniently written as: $\mathbf{R} = \mathbf{P} - \mathbf{r}\mathbf{c}' = [p_{ij} - p_{i\cdot}p_{\cdot j}]$. Later on, it will be rather worked on the matrix of the standardized residuals defined as:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\mathbf{R}\mathbf{D}_c^{-1/2} \qquad s_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}. \tag{5}$$

The standardized residual s_{ij} is connected to the location quotient LQ_{ij} by the following relationship:

$$s_{ij} = \sqrt{p_{i\cdot}p_{\cdot j}}\,[LQ_{ij} - 1] \tag{6}$$

Therefore the sign of the standardized residual gives exactly the same information on the cell (i, j) as the position of the location quotient with respect to the pivot value 1. The standardized residual may be also be viewed as an homothetic transformation of the location quotient, by a factor $\sqrt{p_{i\cdot}p_{\cdot j}}$. This scale factor may be motivated by the problems raised by small areas, and small sectors, in line with the works of Moineddin, Beyene and Boyle (2003), O'Donoghue and Gleave (2004) and Guimarães, Figueiredo and Woodward (2003, 2009).

When elaborating a simultaneous grouping of regions and sectors, a Singular Value Decomposition (SVD) of a matrix operates simultaneously on the rows and the columns. More specifically, a SVD of $\mathbf{S}$ may be written as:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}' \tag{7}$$

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 $K = min(I-1, J-1)$, the dimension of $\mathbf{U}$ is $I \times K$, of $\mathbf{V}$ is $J \times K$ and where $\mathbf{D}_\lambda$ is accordingly $K \times K$; moreover $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = I_{(K)}$. The space $\mathbb{R}^K$ is called the *factor space*. By so doing, SVD decomposes simultaneously the linear space generated by a matrix and the χ^2-statistic, or inertia. This is precisely the way Correspondence Analysis (CA) operates. More explicitly, at each step of the decomposition, a new set of coordinates for both regions and sectors is obtained, in which the χ^2-divergence becomes the classical Euclidean distance, or Pythagorean distance. Therefore, the χ^2-divergence between two objects is equivalent to the Pythagorean distance between the two objects as represented in the factor space of the CA. The SVD decomposes the total inertia so that the first axis is associated to the greatest proportion

of the total inertia, the second axis to the second largest proportion and so on. In other words, CA defines a sequence of subspaces containing an increasing proportion of the total inertia; for more information see *e.g.* Benzécri (1973, 1992), Lebart, Morineau and Warwick (1984), Greenacre (1984, 1993, 2007).

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

$$d_{\chi^2}([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}]) \;=\; \phi^2 \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} \;-\; p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \;=\; tr\mathbf{S}'\mathbf{S} \;=\; \sum_{k=1}^{K} \lambda_k^2 \tag{8}$$

The principal coordinates for the rows (regions) are defined as:

$$\mathbf{F} \;=\; \mathbf{D}_r^{-1/2}\,\mathbf{U}\,\mathbf{D}_\lambda \;=\; [f_{ik}] \quad I \times K \qquad f_{ik} \;=\; p_{i\cdot}^{-1/2}\,\lambda_k\,u_{ik} \tag{9}$$

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

$$\mathbf{F}'\mathbf{D}_r\mathbf{F} \;=\; \mathbf{D}_\lambda^2 \qquad i.e. \quad \sum_i p_{i\cdot} f_{ik}^2 \;=\; \lambda_k^2 \tag{10}$$

thus equation (10) 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 $I_i \;=\; p_{i\cdot} f_{ik}^2$. Similarly, the principal coordinates for the columns (sectors) are defined as:

$$\mathbf{G} \;=\; \mathbf{D}_c^{-1/2}\,\mathbf{V}\,\mathbf{D}_\lambda \;=\; [g_{jk}] \quad J \times K \qquad g_{jk} \;=\; p_{\cdot j}^{-1/2}\,\lambda_k\,v_{jk} \tag{11}$$

where g_{jk} represents the score of sector 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:

$$\mathbf{G}'\mathbf{D}_c\mathbf{G} \;=\; \mathbf{D}_\lambda^2 \qquad i.e. \quad \sum_j p_{\cdot j} g_{jk}^2 \;=\; \lambda_k^2 \tag{12}$$

thus equation (12) decomposes the k-th eigenvalue of $\mathbf{S}'\mathbf{S}$ according to the contribution of each sector j, where $I^j \;=\; p_{\cdot j} g_{jk}^2$ measures the contribution of the sector j.

The SVD of $\mathbf{S}$ will be used in the following spirit. Let $\mathbf{D}_{\lambda(m)}$ be the principal submatrix of $\mathbf{D}_\lambda$ corresponding to the first m eigenvalues λ_k and $\mathbf{U}_{(m)}$ 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, see *e.g.* Eckart and Young, 1936). For each $m \;=\; 1, \ldots, K$, the algorithm will consider a sequence of hierarchical clusterings corresponding to a sequence of improved approximations of $\mathbf{S}$. Therefore, the SVD of $\mathbf{S}$ provides a decomposition of the total global localization measure ϕ^2 in terms of the contributions of each factor k and of the contribution of the regions i, respectively the sectors j:

$$\phi^2 \;=\; \sum_i I_i \;=\; \sum_i \sum_k p_{i\cdot}\, f_{ik}^2 \;=\; \sum_j I^j \;=\; \sum_j \sum_k p_{\cdot j}\, g_{jk}^2 \tag{13}$$

For more information, see *e.g.* Mardia, Kent and Bibby (1979), Jobson (1992) and Greenacre (2007).

Let us write the rows and columns profiles as follows:

$$\mathbf{D}_r^{-1}\,\mathbf{P} \;=\; \left[\frac{p_{ij}}{p_{i\cdot}}\right] \;=\; [p_{j|i}] \qquad\qquad \mathbf{P}\,\mathbf{D}_c^{-1} \;=\; \left[\frac{p_{ij}}{p_{\cdot j}}\right] \;=\; [p_{i|j}] \tag{14}$$

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 concentration of sector j:

$$d_{\chi^2}(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) \;=\; \sum_j \frac{(p_{j|i} \,-\, p_{\cdot j})^2}{p_{\cdot j}} \;=\; [p_{\vec{j}|i} - p_{\cdot\vec{j}}]' D_c^{-1} [p_{\vec{j}|i} - p_{\cdot\vec{j}}] \;=\; \sum_k f_{ik}^2 \tag{15}$$

$$d_{\chi^2}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \;=\; \sum_i \frac{(p_{i|j} \,-\, p_{i\cdot})^2}{p_{i\cdot}} \;=\; [p_{\vec{i}|j} - p_{\vec{i}\cdot}]' D_r^{-1} [p_{\vec{i}|j} - p_{\vec{i}\cdot}] \;=\; \sum_k g_{jk}^2 \tag{16}$$

Therefore, the decomposition of the Total inertia as a measure of global localization, in (13), may also be written in terms of average relative specialization or concentration:

$$\phi^2 \;=\; \sum_i p_{i\cdot}\, d_{\chi^2}(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) \;=\; \sum_j p_{\cdot j}\, d_{\chi^2}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \tag{17}$$

More details, under a Stochastic Independence Approach (SIA), are given in Haedo and Mouchart (2012).

3.3 Automatic optimal collapsed table: the algorithm based on HCCA

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

An overview

In line with (15) and (16), the distance between regions or sectors is provided by means of a "square of weighted Euclidean distances" among profiles. The dissimilarity between the profiles of two regions i and i' or two sectors j and j' is accordingly measured as follows:

$$\sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} \,-\, p_{j|i'}\right)^2 \;=\; [p_{\vec{j}|i} \,-\, p_{\vec{j}|i'}]' D_c^{-1} [p_{\vec{j}|i} \,-\, p_{\vec{j}|i'}] \;=\; d^2_{D_c}(p_{\vec{j}|i} \mid p_{\vec{j}|i'}) \tag{18}$$

$$\sum_i \frac{1}{p_{i\cdot}} \left(p_{i|j} \,-\, p_{i|j'}\right)^2 \;=\; [p_{\vec{i}|j} \,-\, p_{\vec{i}|j'}]' D_r^{-1} [p_{\vec{i}|j} \,-\, p_{\vec{i}|j'}] \;=\; d^2_{D_r}(p_{\vec{i}|j} \mid p_{\vec{i}|j'}) \tag{19}$$

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

$$\frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} \,+\, p_{i'\cdot}} \sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} \,-\, p_{j|i'}\right)^2 \;=\; \frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} \,+\, p_{i'\cdot}} [p_{\vec{j}|i} \,-\, p_{\vec{j}|i'}]' D_c^{-1} [p_{\vec{j}|i} \,-\, p_{\vec{j}|i'}] \tag{20}$$

The global localization of an economy decreases as a consequence of clustering and this loss of information is reduced by clustering the most similar regions or sectors. Thus the algorithm chooses

pairs of regions i and i' and pairs of sectors j and j' minimizing the measures of dissimilarity (18) and (19). The two rows are then merged by summing their frequencies and the profile and mass are recalculated. The same measure of difference as (20) 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 denoted 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 partition. 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:

$$\mathcal{I}_{(0)} = \{\{1\},\{2\},\ldots,\{I\}\} \qquad \mathcal{J}^{(0)} = \{\{1\},\{2\},\ldots,\{J\}\} \tag{21}$$

$$\mathcal{I}_{(I)} = \{\{1,2,\ldots,I\}\} \qquad \mathcal{J}^{(J)} = \{\{1,2,\ldots,J\}\} \tag{22}$$

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

First step: building collapsed tables from HCCA

• *Work on the rows (regions).* For $m = 1,2,\ldots,K$:

Consider the first m columns of $\mathbf{F}$, let $\mathbf{F}_{(m)} = (f_{\vec{i}1},\ldots,f_{\vec{i}l},\ldots,f_{\vec{i}m})$ $I \times m$, where $f_{\vec{i}l}$ represents the l-th column of $\mathbf{F}$, and obtain a hierarchical clustering of the rows of $\mathbf{F}_{(m)}$, corresponding to the rows of $\mathbf{S}$, as follows. Let $\mathcal{I}_{(n,m)}$ $n = 0,\ldots,I$, with $\forall m : \mathcal{I}_{(0,m)} = \mathcal{I}_{(0)}$ and $\mathcal{I}_{(I,m)} = \mathcal{I}_{(I)}$, 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 $\mathcal{I}_{(n-1,m)}$. Thus $\forall m$, there are only $I-1$ relevant levels of the hierarchical clustering, graphically represented by a dendrogram.

• *Work on the columns (sectors).* For $m = 1,2,\ldots,K$:

Repeat the same with the columns of $\mathbf{G}$, namely $\mathbf{G}_{(m)} = (g_{\vec{j}1},\ldots,g_{\vec{j}l},\ldots,g_{\vec{j}m})$ $J \times m$ where $g_{\vec{i}l}$ represents the l-th column of $\mathbf{G}$, and obtain a dendrogram through a hierarchical clustering of the rows of $\mathbf{G}_{(m)}$, corresponding to the columns of $\mathbf{S}$, as follows. Let $\mathcal{J}^{(h,m)}$ $h = 0,\ldots,J$, with $\forall m : \mathcal{J}^{(0,m)} = \mathcal{J}^{(0)}$ and $\mathcal{J}^{(J,m)} = \mathcal{J}^{(J)}$, be the nested sequence of partitions of sectors, starting with $\mathcal{J}^{(0)}$ and with each following cluster obtained as an optimized clustering scheme based on $\mathcal{J}^{(h-1,m)}$. Thus $\forall m$, 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 $(I-1)(J-1)$ collapsed tables, repeat the operation for each m, obtain accordingly $(I-1)(J-1)K$ collapsed tables $T^{(m)}_{\mathcal{I}_{n,m} \times \mathcal{J}^{h,m}}$ and calculate the corresponding inertia $\phi^2\left(T^{(m)}_{\mathcal{I}_{n,m} \times \mathcal{J}^{h,m}}\right)$.

Second step: identifying an optimal collapsed table through bootstrapping

Having built the array A of $(I-1)(J-1)K$ collapsed tables, the final question is: which of the collapsed tables is better in the sense of an optimal compromise between a smallest table that preserves

the highest global localization (*i.e.* association) possible? Permutation bootstrapping provides a tool for a suitable compromise.

• *Bootstrapping*:

Let us consider whether a particular table $T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}$ is optimal in the sense alluded above. At least, one 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 characterized by the three characteristics (n, h, m). Here, $\mathcal{I}_{n,m}$ 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 $\pi_r(\mathcal{I}_{n,m})$ for the image of the partition $\mathcal{I}_{n,m}$ transformed by π_r. Similarly, let π^c be a permutation defined on $\mathcal{J}$ and its image $\pi^c(\mathcal{J}^{h,m})$.

• *Optimal collapsed table*:

Given (π_r, π^c), one may define a transformed table $T^{(m)}_{\pi_r(\mathcal{I}_{n,m})\times\pi^c(\mathcal{J}^{h,m})}$, following the same partition scheme as the optimized table $T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}$ with shuffled labels, and compute a corresponding inertia $\phi^2\left(T^{(m)}_{\pi_r(\mathcal{I}_{n,m})\times\pi^c(\mathcal{J}^{h,m})}\right)$. Note that the transformed table $T^{(m)}_{\pi_r(\mathcal{I}_{n,m})\times\pi^c(\mathcal{J}^{h,m})}$ has a same dimension as $T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}$; thus their inertia are comparable. The difference $\phi^2\left(T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}\right) - \phi^2\left(T^{(m)}_{\pi_r(\mathcal{I}_{n,m})\times\pi^c(\mathcal{J}^{h,m})}\right)$ 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. Thus, the difference

$$\psi(n, h, m) \;=\; \phi^2\left(T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}\right) \;-\; \mathbb{E}_B\,\phi^2\left[T^{(m)}_{\pi_r(\mathcal{I}_{n,m})\times\pi^c(\mathcal{J}^{h,m})}\right] \tag{23}$$

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 optimal collapsed table, $\mathcal{I}_{n^*,m^*} \times \mathcal{J}^{h^*,m^*}$ by solving the maximization problem:

$$(n^*, h^*, m^*) \;=\; arg\,max_{n,h,m}\;\psi(n, h, m), \tag{24}$$

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

• *Stopping rules*:

Two stopping rules of the algorithm may be considered, corresponding to two different approaches to the issue of optimal collapsing. The applications that are now coming, correspond to (24) and terminate the computation once identified the collapsing providing the largest difference between the global localization $\phi^2\left(T^{(m)}_{\mathcal{I}_{n,m}\times\mathcal{J}^{h,m}}\right)$ of a given table and the bootstrap average of randomly shuffled labels of a same level of the dendrogram. Another stopping rule would introduce an overall objective function built through a weighting of the size of a table and its corresponding global localization measure. In that case, the algorithm computes this objective function at each level of the dendrogram and retains that one with the highest score.

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. ■

4 Applications: Argentina, Brazil and Chile

4.1 An overview of the applications

Preliminary works with simulated data had provided encouraging results about the performance of this algorithm, motivating a further step with real data. Here, the purpose of the applications is to analyze the overall degree of global localization of Argentina, Brazil and Chile using for each of them the optimal collapsed tables algorithm and to evaluate how far the functioning of the algorithm provides additional information on the regional structure of the economic activity.

The spatial units are the lower level political-administrative jurisdictions called departments (523), municipalities (5138) and communes (342) of Argentina, Brazil and Chile, respectively. After eliminating those without employees, there are remaining 462, 5138 and 249 spatial units for Argentina, Brazil and Chile, respectively. The data, related to the number of employees in the manufacturing industry, were obtained from of the National Institutes of Statistics and Censuses of Argentina (INDEC-1994: 1083928 employees), Brazil (IBGE-1998: 6018445 employees), and Chile[5] (INE-2005: 446613 employees), respectively. The sectors classification in Table 3 refers to the first 2 digits (divisions) of the International Standard Industrial Classification (ISIC-Rev.3.1) of manufacturing industry (22 divisions, after combining sectors 36 and 37 into a unique sector).

[5]The available data for Chile refer to the firms with 5 or more employees.

Table 3: *Sectors of the manufacturing industry (Divisions ISIC-Rev.3.1)*

Sector	Description
15	Manufacture of food products and beverages
16	Manufacture of tobacco products
17	Manufacture of textiles
18	Manufacture of wearing apparel; dressing and dyeing of fur
19	Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear
20	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
21	Manufacture of paper and paper products
22	Publishing, printing and reproduction of recorded media
23	Manufacture of coke, refined petroleum products and nuclear fuel
24	Manufacture of chemicals and chemical products
25	Manufacture of rubber and plastics products
26	Manufacture of other non-metallic mineral products
27	Manufacture of basic metals
28	Manufacture of fabricated metal products, except machinery and equipment
29	Manufacture of machinery and equipment n.e.c.
30	Manufacture of office, accounting and computing machinery
31	Manufacture of electrical machinery and apparatus n.e.c.
32	Manufacture of radio, television and communication equipment and apparatus
33	Manufacture of medical, precision and optical instruments, watches and clocks
34	Manufacture of motor vehicles, trailers and semi-trailers
35	Manufacture of other transport equipment
36	Manufacture of furniture; manufacturing n.e.c. (includes division 37: recycling)

Table 4 shows a summary of the results obtained from the three measures of global localization- as proposed through the SIA in Haedo and Mouchart (2012)- for the original and for the optimal collapsed tables, the number of cells, and the resulting loss of information about the level of global localization.

Table 4: *Summary of the results*

Measure	Original table			Optimal collapsed table			Lost level of global localization (%)		
	Argentina	**Brazil**	**Chile**	**Argentina**	**Brazil**	**Chile**	**Argentina**	**Brazil**	**Chile**
$d_{\mathcal{X}^2}$	2.1580	3.1345	3.4363	1.6532	2.4162	2.8584	23.39	22.91	16.82
d_{KL}	0.5049	0.7420	0.8870	0.3176	0.4928	0.6759	37.10	33.58	23.80
d_H	0.1300	0.1894	0.2600	0.0713	0.1073	0.1773	45.17	43.39	31.80
# of cells	10164	113036	5478	595	884	450			
	(462x22)	(5138x22)	(249x22)	(35x17)	(52x17)	(30x15)			
Reduction # of cells				94.15%	99.22%	91.79%			

While the absolute values of these measures are not comparable because they use different scales[6], the measures of global localization in the original tables show that Chile has a higher level of global localization, followed by Brazil and Argentina, respectively. This order is not an effect of the number of cells for which one has: Chile $<$ Argentina $<$ Brazil. The three measures of global localization on the optimal collapsed tables also show the same order between these countries, although in this case the different measures show a decline in its absolute values as a result of the loss of information arising from simultaneous grouping regions and sectors. The loss in the level of global localization is quite similar for Argentina and Brazil, while Chile clearly shows a minor loss of information. The loss of information seems very reasonable vis-à-vis a substantial reduction in the size of the tables (more than 90%). Note however that the percentages of the loss of information clearly depend on the underlying divergences and systematically give a same ordering: $d_{\mathcal{X}^2} < d_{KL} < d_H$. That $d_{\mathcal{X}^2}$ be associated with the smallest loss of information is coherent with the fact that this version of the algorithm is based on optimizing the loss in terms of $d_{\mathcal{X}^2}$.

The following Figure 1 corresponds to the steps of the algorithm where the optimal collapsed tables of Argentina, Brazil and Chile (in columns) are found using B=1000 permutation bootstraps and gives, through 4 figures (in rows), more insight on the meaning of the final results.

This figure has been constructed as follows:

- *First line.* The row (regions) dendrograms of the HCCA method, based on the number of the first eigenvalues selected in the step of the algorithm where the optimal collapsed tables were found. The red lines in the row dendrogram show the cutting branches selected automatically.

- *Second line.* Same as first line but now for the columns (sectors). Here, the cutting branches are more visible than for the rows (regions) dendrogram because of the much smaller number of sectors.

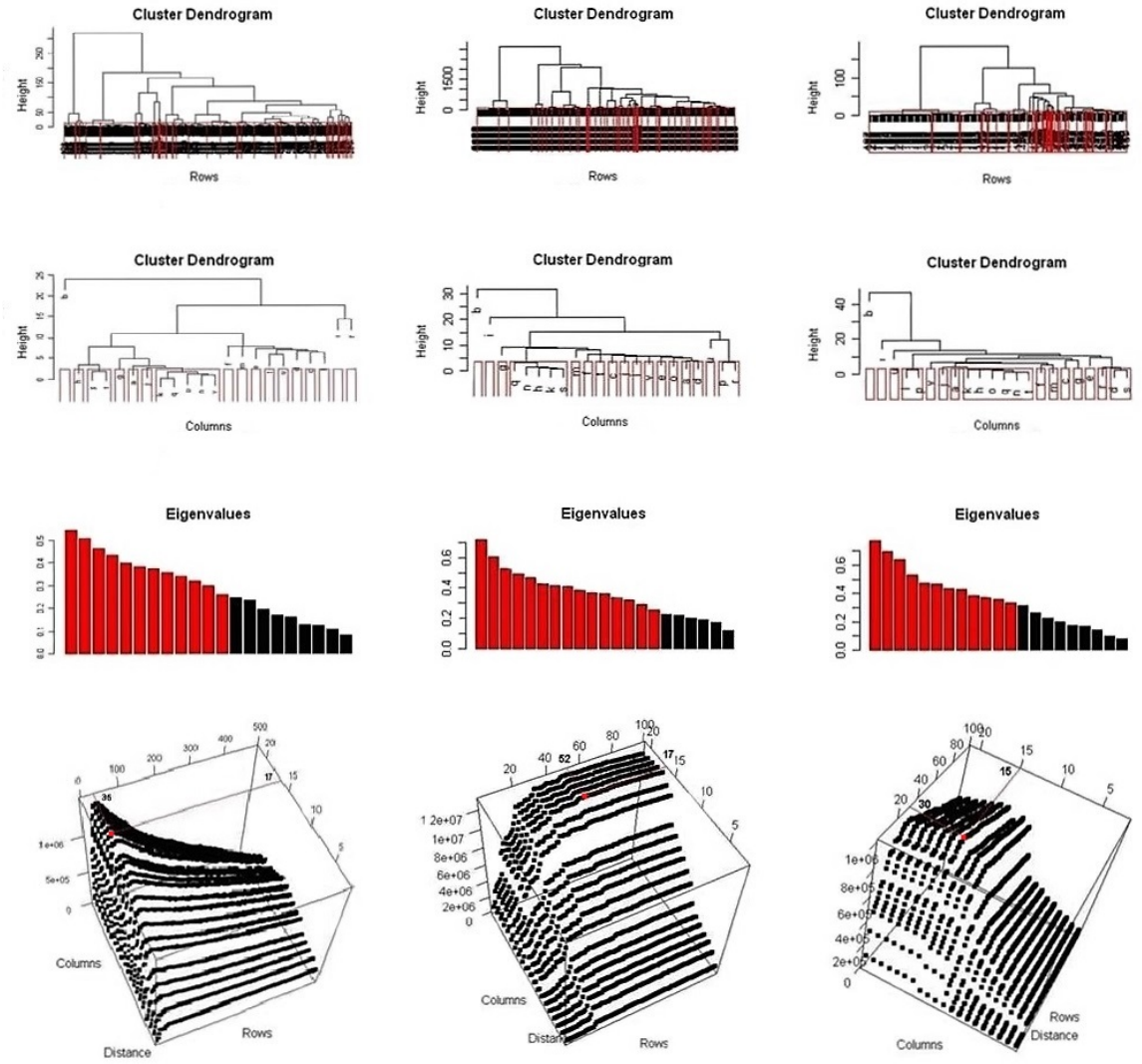
- *Third line.* Bar graph of the eigenvalues. The red bars indicate the first eigenvalues selected in the step of the algorithm where the optimal collapsed tables were found: the first 12 out of 21 for Argentina and Chile and the first 15 out of 21 for Brazil.

- *Fourth line.* The 3-dimensional plots of the differences $\psi(n, h, m)$ for each level of the rows and columns dendrograms, for $(I-1)(J-1)K$ collapsed tables, based on the number of the first eigenvalues selected in the step of the algorithm where the optimal collapsed tables were found. A red point indicates the maximum of $\psi(n^*, h^*, m^*)$.

In the sequel, the working of the algorithm is analyzed separately for each country in order to show how these results have been obtained. Indeed, the way this algorithm is operating is an important ingredient for the interpretation of the final results. Different aspects of the algorithm are

[6]See Haedo and Mouchart (2012) for more information on this issue.

Figure 1: *Algorithmic results of Argentina (first column), Brazil (second column) and Chile (third column)*



selected for each country. The final results for the three countries are presented in increasing order of dimension: Chile, Argentina and Brazil, with an interest for the performance on the algorithm facing, at the start, an increasing number of regions along with a constant number of sectors.

4.2 Optimal collapsed table for Chile

From Table 4 and the third column of Figure 1, 249 regions are grouped into 30 g-regions, while the 22 sectors are grouped into 15 g-sectors, obtained by grouping 10 sectors as follows: 18 and 33 into $GS1$, 22, 25, 28, 29, 31 and 34 into $GS2$, and 26 and 30 into $GS3$, and leaving the other 12 sectors

ungrouped. These groupings reduce by 92% the number of cells of the original table (from 5478 to 450 cells) with a lost level of global localization, measured by $d_{\mathcal{X}^2}$, of 17%.

Tables 5 and 6 show the standardized residuals and the number of primary units, namely employees, by sector for six arbitrarily selected g-regions of Chile. As seen from (6), up to a scale factor given by the the square root of the product of the marginal distributions, these standardized residuals provide a same information as the location quotient, regarding the over- or under-specialization of each cell (i, j). The algorithm precisely aims at grouping regions according to their relative specialization, rather than to the volume of their employment only. Indeed, (17) shows how the total inertia is decomposed into a weighted average of relative specializations or concentrations. For each of these six g-regions, the sector of main over-specialization is provided by the maximum of the corresponding row of the standardized residuals in Table 5 and marked with red numbers. The corresponding numbers for the employment are also marked in red in Table 6.

Table 5: *Standardized residuals of the six selected g-regions for Chile*

g-region	**g-sector**														
	15	16	17	$GS1$	19	20	21	$GS2$	23	24	$GS3$	27	32	35	36
$GR3$	-30	-2	-9	-11	-7	-15	-10	-22	-4	173	-9	-14	-1	-6	-9
$GR4$	-64	-5	-25	-28	-20	-40	-27	-5	-12	12	3	278	-3	-16	-25
$GR6$	-49	-3	-15	-18	-12	-26	-16	-38	-7	-18	-15	291	-2	-10	-16
$GR8$	-55	-7	39	54	125	-37	-6	37	-15	10	-14	-15	19	-21	1
$GR13$	-1	-6	-5	-25	-10	-44	-11	-1	-13	123	13	-24	-3	-17	-14
$GR15$	-32	-3	1	-15	196	-19	-6	23	-6	-15	-7	-19	-1	15	3

$GS1$: sectors 18 and 33.

$GS2$: sectors 22, 25, 28, 29, 31 and 34.

$GS3$: sectors 26 and 30.

The g-regions 8 and 15, formed by 7 and 2 regions respectively, are over-specialized in sector 19. The 3864 and the 2080 employees of these two g-regions 8 and 15 respectively gather together, at the national level, 67% of the total employment of that sector. The g-region 8 is also over-specialized, to a lesser extent, in $GS1$, 17, $GS2$, 32 and 24, respectively. Instead, g-region 15 is also over-specialized in $GS2$ and 35. These differences explain why g-regions 8 and 15 are not aggregated into a unique g-region.

Table 6: *Number of employees of the six selected g-regions for Chile*

g-region	**g-sector**														
	15	16	17	$GS1$	19	20	21	$GS2$	23	24	$GS3$	27	32	35	36
$GR3$	0	0	0	0	0	0	0	0	0	2574	0	0	0	0	0
$GR4$	1666	0	0	36	0	103	5	3582	0	1953	726	11894	0	16	22
$GR6$	96	0	0	0	0	0	0	6	0	119	9	7220	0	0	0
$GR8$	5667	0	2212	3483	3864	899	994	9323	0	2937	634	1618	80	5	1146
$GR13$	8050	0	552	223	236	43	522	4417	0	6754	1122	647	0	13	374
$GR15$	429	0	161	0	2080	38	100	1700	0	90	76	0	0	192	213

$GS1$: sectors 18 and 33.

$GS2$: sectors 22, 25, 28, 29, 31 and 34.

$GS3$: sectors 26 and 30.

The g-regions 3 and 13, formed by 4 and 5 regions respectively, are over-specialized in sector 24. The 2574 and the 6754 employees of the g-regions 3 and 13 respectively gather together, at the national level, 28% of the total employment that the sector. Except for the g-region 13 which is also over-specialized in sector $GS3$, both of these g-regions are under-specialized in the remaining sectors. Finally, g-regions 4 and 6, formed by 9 and 10 regions respectively, are over-specialized in sector 27. The 11894 and 7220 employees of g-regions 4 and 6 respectively gather together, at the national level, 60% of the employment of that sector. Except for the g-region 4 which is also over-specialized in sector 24, both of these g-regions are under-specialized in the remaining sectors. The high concentration of employment in the main sectors of over-specialization of those selected g-regions is in line with the observation that employment has a high tendency to co-localize in specialized regions, suggesting the presence of economies of agglomeration (see *e.g.* Haedo and Mouchart 2015a, for a recent result in that direction). Figures 2 and 3 show the location of the selected g-regions for two parts of Chile: one map for the central and another one for the north and south. These maps illustrate the fact that the algorithm does not include any restriction of contiguity or of distance, within the g-regions. But neighboring g-regions, for instance g-regions 4 and 6 in Figure 3, and neighboring regions within a g-region show an effect of specialized agglomeration, comforting some of the findings of Haedo and Mouchart (2015a).

Figure 2: *Map 1 for Chile: location of the six selected g-regions (central)*

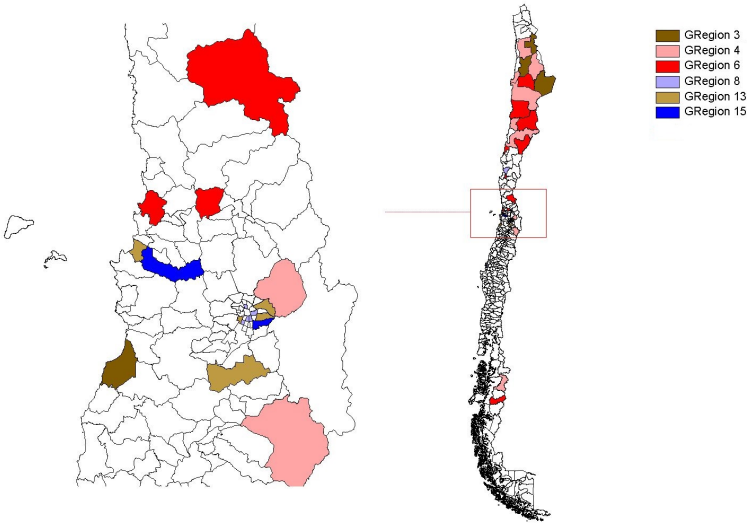
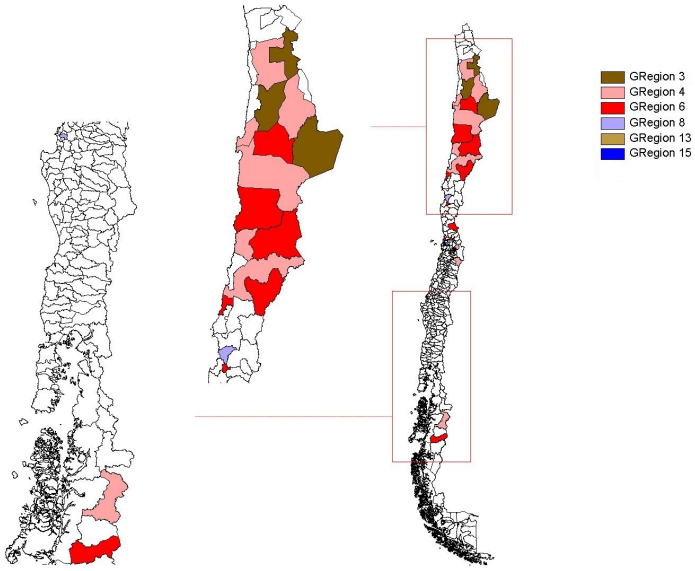


Figure 3: *Map 2 for Chile: location of the six selected g-regions (north and south)*



4.3 Optimal collapsed table for Argentina

From Table 4 and the first column of Figure 1, 467 regions are grouped into 35 g-regions, while the 22 sectors are grouped into 17 g-sectors obtained by grouping 7 sectors as follows: 25, 28, 29, 31 and 36 into $GS1$ and 30 and 33 into $GS2$, and leaving the other 15 sectors ungrouped. These groupings

reduce by 94% the number of cells of the original table (from 10164 to 595 cells) with a lost level of global localization, measured by $d_{\mathcal{X}^2}$, of 23%.

Table 7: *Number of employees of the four selected g-regions for Argentina*

g-region	**g-sector**																
	15	16	17	18	19	20	21	22	23	24	$GS1$	26	27	$GS2$	32	34	35
$GR2$	3819	6	1496	6779	424	258	479	1750	4	2147	3300	384	68	281	114	587	21
$GR15$	851	0	193	3255	289	149	1	151	0	4	827	61	21	8	26	81	0
$GR16$	1409	0	732	140	59	120	462	200	784	639	2125	454	13022	5	281	1328	60
$GR35$	485	0	224	24	1	164	0	76	0	85	827	61	0	0	3379	215	0

$GS1$: sectors 25, 28, 29, 31 and 36.

$GS2$: sectors 30 and 33.

From the 35 g-regions of Argentina, Table 7 show the number of employees for four arbitrarily selected g-regions. It may be observed that g-regions 2 and 15, formed by 7 and 10 regions respectively, display a high employment mainly in sectors 15, 18 and $GS1$ (red values on Table 7). Similarly for the g-regions 16 and 35.

Table 8 show the standardized residuals by sector for the same four g-regions. The two g-regions 2 and 15 are not aggregated because of differences of relative specialization between these two g-regions, namely that the g-region 2 is over-specialized mainly in sectors 18, 22 and 24 (positive standardized residuals), while g-region 15 is over-specialized mainly in sectors 18 and 19. For the remaining sectors, these g-regions are essentially under-specialized relatively to different sectors (negative, or small positive, standardized residuals).

Table 8: *Standardized residuals of the four selected g-regions for Argentina*

g-region	**g-sector**																
	15	16	17	18	19	20	21	22	23	24	$GS1$	26	27	$GS2$	32	34	35
$GR2$	-27	-9	9	191	-15	-15	-2	24	-11	20	-28	-20	-25	5	-7	-23	-12
$GR15$	-18	-5	-7	189	4	-2	-12	-7	-6	-19	-16	-13	-13	-7	-4	-16	-7
$GR16$	-58	-10	-13	-26	-27	-20	-3	-25	55	-20	-44	-18	452	-14	4	-44	-9
$GR35$	-26	-5	-5	-14	-15	0	-12	-11	-6	-14	-14	-12	-14	-7	449	-8	-7

$GS1$: sectors 25, 28, 29, 31 and 36.

$GS2$: sectors 30 and 33.

The main sectors, in terms of employment, correspond to the main over-specialized sectors of the g-regions. Indeed, from Table 8, the sectors 18, 27 and 32 display the highest levels of relative concentration (*i.e.* highest level of positive standardized residuals) in g-regions of highest employment, as may be guessed from Table 7. More explicitly, g-regions 2 and 15 are over-specialized in sector 18 and gather, at the national level, 22% of the total employment of that sector. Similarly, g-region

16, over-specialized in sector 27, gathers 35% of the total employment of that sector and g-region 35, over-specialized in sector 32, gathers 32% of the total employment of that sector. Similarly to Chile, the high concentration of employment in the main sectors of over-specialization shows a high tendency to co-localize employment in specialized regions. Figures 4 and 5 show the location of the selected g-regions for two parts of Argentina: one map for the central-east and another one for the central-west. Similarly to Chile, neighboring g-regions, for instance g-regions 2, 16 and 35 in Figure 4, and neighboring regions within a g-region show an effect of specialized agglomeration.

Figure 4: *Map 1 for Argentina: location of the four selected g-regions (central-east)*

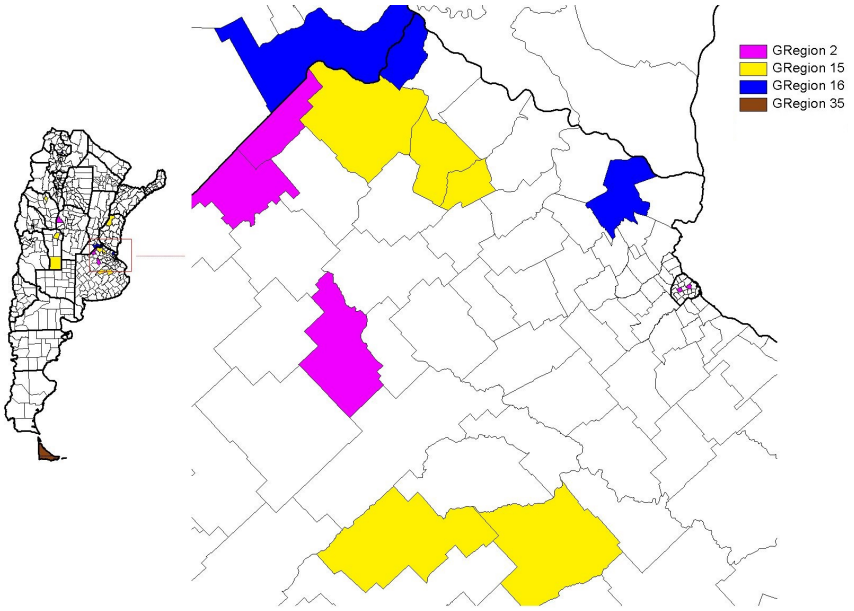
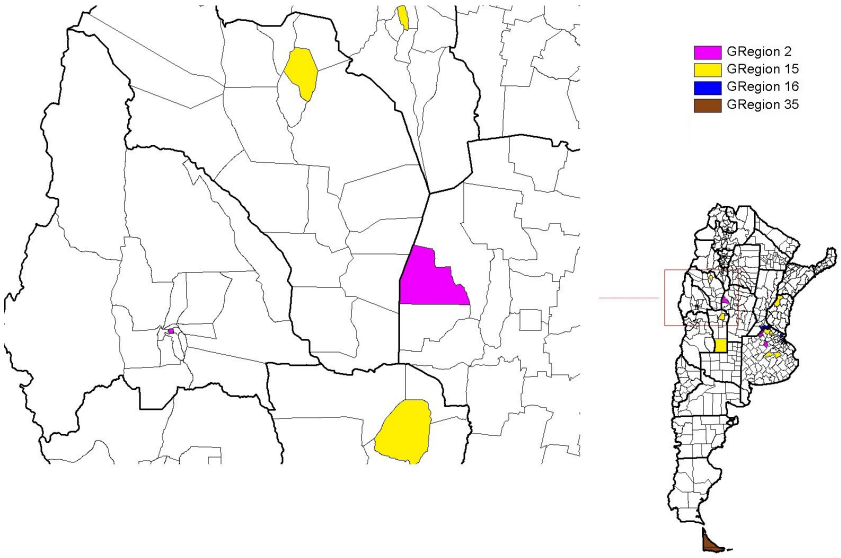


Figure 5: *Map 2 for Argentina: location of the four selected g-regions (central-west)*



4.4 Optimal collapsed table for Brazil

From Table 4 and the second column of Figure 1, 5138 regions are grouped into 52 g-regions, while the 22 sectors are grouped into 17 g-sectors obtained by grouping 7 sectors as follows: 22, 25, 28, 31 and 33 into $GS1$ and 30 and 32 into $GS2$, and leaving the other 15 sectors ungrouped. These groupings reduce by 99% the number of cells of the original table (from 113036 to 884 cells) with a lost level of global localization, measured by $d_{\mathcal{X}^2}$, of 23%. These results indicate a powerful capacity of adaptation of the algorithm for the case of a large number of regions.

For Brazil, standardized residuals are first commented at a graphical level before discussing numerical tables in order to get further information on how the algorithm operates in relation to the concept of global localization. Figure 6 depicts the standardized residuals for the 113036 cells of the original table. Most of the standardized residuals are around zero. Therefore, it is to be expected that most of the cells are not significantly over-specialized or under-specialized, suggesting that the degree of specialization can be explained with much less information.

Figure 6: *Standardized residuals of the original table of Brazil*

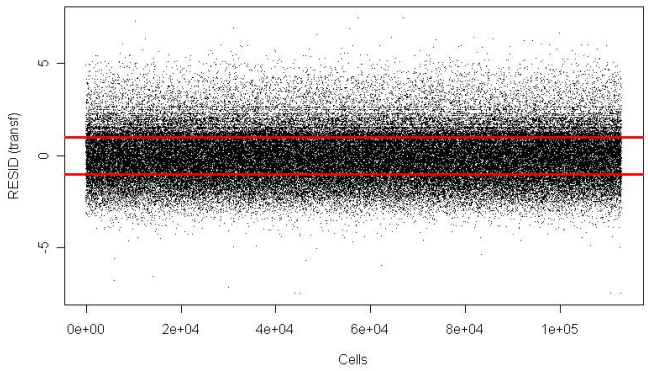
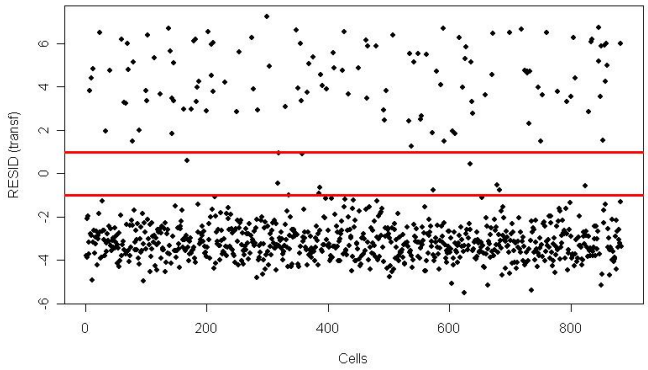


Figure 7 shows the standardized residuals for the 884 cells of the optimal collapsed table obtained by grouping rows and columns so as to make the most extreme possible standardized residual values, *i.e.* to make the over-specialized and the under-specialized phenomenon more apparent. As a matter of fact, grouping into g-regions and g-sectors succeeds in identifying homogenous regions of similar relative specialization structure.

Figure 7: *Standardized residuals of the optimal collapsed table of Brazil*



Tables 9 and 10 show the number of employees and the standardized residuals by g-sectors for four arbitrarily selected g-regions of Brazil. As for the previous countries, the main over-specialization of each g-region is provided by the maximum of each row of Table 10, marked with red numbers. The corresponding numbers for the employment are also marked in red in Table 9. The g-region 27, formed by 41 regions, is mainly over-specialized in the g-sector $GS2$. The 22070 employees of this g-sector gather, at the national level, 23% of the total employment of that sector. The g-region 27 is also over-specialized, to a lesser extent, in sector 35 and g-sector $GS1$. The g-regions 46 and 49, formed by 118 and 73 regions respectively, are over-specialized in sector 19. The 164653

Table 9: *Number of employees of the four selected g-regions for Brazil*

g-region	g-sector																
	15	16	17	18	19	20	21	$GS1$	23	24	26	27	29	$GS2$	34	35	36
$GR27$	9515	17	3179	2309	120	1448	1379	29664	42	2655	3298	925	3160	22070	2681	6411	2921
$GR46$	4347	6	1773	2368	164653	1279	2877	10934	2	2436	2233	945	2314	30	228	215	4663
$GR49$	27479	49	4560	8415	91625	4179	3648	17523	86	2626	6848	962	6752	101	860	273	6744
$GR51$	24738	25	8702	12400	1753	5276	6093	75143	1410	19449	15397	9363	29945	4345	122629	1407	16976

$GS1$: sectors 22, 25, 28, 31 and 33.

$GS2$: sectors 30 and 32.

employees of g-region 46 and the 91625 of g-region 49 gather together, at the national level, 70% of the total employment of that sector. Both g-regions are under-specialized, with different levels, in all remaining sectors. Finally, g-region 51, formed by 48 regions, is over-specialized in sector 34. The 122629 employees of this sector gather, at the national level, 43% of the total employment of that sector. The g-region 51 is also over-specialized, to a lesser extent, in sector 29 and g-sector $GS1$. Similarly to Chile and Argentina, the high concentration of employment in the main sectors of over-specialization of those selected g-regions shows a high tendency to co-localize employment in specialized regions.

Table 10: *Standardized residuals of the four selected g-regions for Brazil*

g-region	g-sector																
	15	16	17	18	19	20	21	$GS1$	23	24	26	27	29	$GS2$	34	35	36
$GR27$	-58	-15	-23	-66	-73	-41	-20	91	-21	-35	-28	-33	-35	537	-26	197	-35
$GR46$	-171	-23	-85	-118	1373	-81	-31	-141	-34	-83	-87	-64	-92	-56	-96	-37	-68
$GR49$	-36	-20	-51	-64	761	-44	-14	-94	-30	-75	-36	-59	-45	-52	-84	-33	-40
$GR51$	-161	-30	-72	-111	-135	-84	-31	27	-15	-3	-36	-8	50	-18	809	-31	-30

$GS1$: sectors 22, 25, 28, 31 and 33.

$GS2$: sectors 30 and 32.

Figures 8 and 9 show the location of the selected g-regions for two parts of Brazil: one map for the south and another one for the north-east. Similarly to Chile and Argentina, neighboring g-regions of Brazil, for instance g-regions 46 and 49 both over-specialized in sector 19, show an effect of specialized agglomeration (southern part of Figure 8).

Figure 8: *Map 1 for Brazil: location of the four selected g-regions (south)*

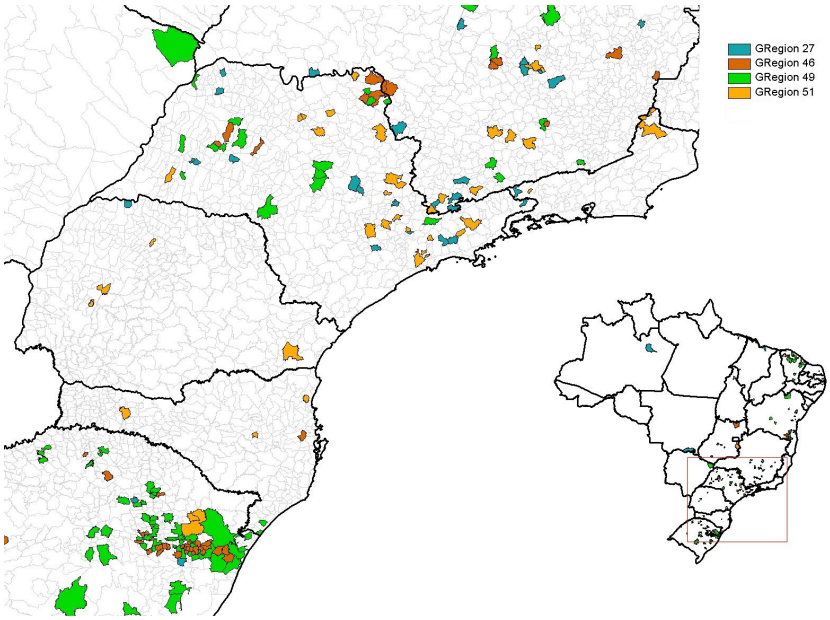
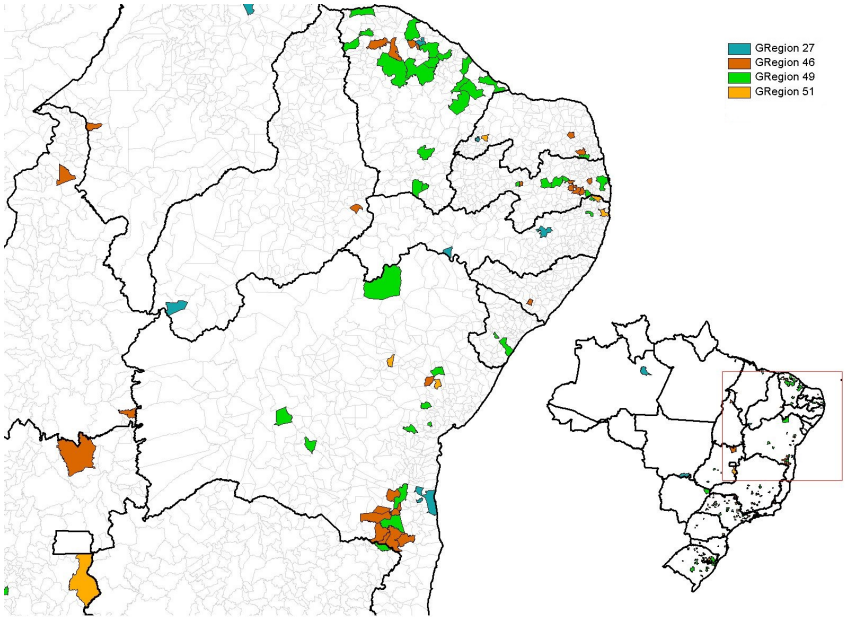


Figure 9: *Map 2 for Brazil: location of the four selected g-regions (north-east)*



5 Conclusions

The Introduction of this paper mentions three challenges the proposed algorithm should consider. These concluding remarks first report on these challenges and thereafter extend the discussion.

The three applications, in Section 4, reveal the following features:

- In this algorithm, the basic criterion for grouping regions is the similarity of the sectorial structure: this is the natural answer to the *first challenge*.

- Distinct g-regions imply different sectorial structures between the g-regions and similar sectorial structure of the regions within each g-regions. This property of the proposed algorithm provides a useful information for the policy maker either within a country or between two neighboring countries. This feature may be viewed as a contribution to the *second challenge*. Interestingly enough, this information may also be used for extending the analysis of complexity and ubiquity as proposed in the Atlas of Economic Complexity, see Hausmann *et al.* (2015), when it is mentioned that "individual specialization begets diversity at the national and global level". This has also been a major achievement in the development of the project *Mapas Industriales de América Latina y el Caribe* (MIALC), see Haedo and Mouchart (2015b).

- When treating large contingency tables regions $\times$ sectors, it is possible to reduce substantially the number of cells, by a factor of more than 90%, along with quite a reasonable loss of the information contained in the table on the global localization, namely with a factor around 20%. This is quite an achievement regarding the *third challenge*.

- Particular features of a country appear more explicitly after collapsing the table. The high concentration of employment in the main sectors of over-specialization of those selected g-regions is in line with the observation that employment has a high tendency to co-localize in specialized regions, suggesting the presence of economies of agglomeration. For instance, g-regions 4 and 6 of Chile are both over-specialized in sector 27 and gather 60 % of the country employment of that sector (see Table 6 and Table 5). A similar phenomenon may also be seen in g-regions 2 and 15 of Argentina, both over-specialized in sector 18 (Table 7 and Table 8), and g-regions 46 and 49 of Brazil, both over-specialized in sector 19 (Table 9 and Table 10).

- The algorithm does not include any restriction of contiguity or of some distance-based pattern within the g-regions. Nevertheless, regions from different g-regions with common sector of over-specialization may display a clear spatial clustering that suggests an effect of specialized agglomeration. For instance, the south-east part of Figure 8, for Brasil, displays a spatial clustering of regions of the g-regions 46 and 49 that are both over-specialized in sector 19. Similarly, the northern part of Figure 3 and the north-west part of Figure 4, for Chile and Argentina, show a same effect of specialized agglomeration. These observations comfort some of the findings of Haedo and Mouchart (2015a).

- Each country starts with the same sectors: the 22 divisions of ISIC-Rev.3.1 and concludes the algorithm with almost the same number of grouped sectors, namely 15 (for Chile) or 17 (for

Argentina and Brazil). These g-sectors are however different for each country because they reveal different regional structures of the economic activity, see Table 11.

Table 11: *Divisions of ISIC-Rev.3.1 within the g-sectors of Chile, Argentina and Brazil*

g-sector	Chile	Argentina	Brazil
$GS1$	18, 33	25, 28, 29, 31, 36	22, 25, 28, 31, 33
$GS2$	22, 25, 28, 29, 31, 34	30, 33	30, 32
$GS3$	26, 30	-	-

Although the definition of the ISIC-divisions is the same for the three countries, the actual content for each country is different. This is likely to be due to the fact that the actual technological content of each sector may be different across the countries. Table 11 shows, for instance, that, in each of the three countries, the sectors 25, 28 and 31 are grouped in a same g-sector although the content of the corresponding g-sectors is different. This is an interesting observation the explanation of which deserves further analysis.

A distinctive feature of our algorithm is to be completely automatic in the sense that the number of clusters and the dimension of the factor space of the final solution are determined by the algorithm itself rather than by pre-specified parameters. The motivation for this choice is the development of a tool designed for deepening specific concepts of economic geography, in the framework of the SIA, at variance from other biclustering algorithms oriented toward specific characteristics of the resulting regrouping. Indeed the underlying concepts of g-regions and of g-sectors are completely specified by the algorithm itself, at variance from an algorithm requiring the a priori specification of some parameters. In this later case, such algorithms would define a family of concepts rather than a unique concept. In the present case, the concerts of g-regions and of g-sectors are defined in terms of relative concentration and specialization.

This algorithm looks for collapsing simultaneously regions and sectors, with no restriction about the regions or the sectors to be clustered. A possible prospect for future research would consist in the introduction of contiguity- or distance-based pattern - among regions and among sectors. These measures of neighborhood might be used as a system of penalizations when forming g-regions or g-sectors and should be specified according to objectives of an economic policy.

The results of the applications demostrate that the innovative aspects of this algorithm are in providing a new way of understanding the similarities and dissimilarities of the processes of industrial concentration and regional specialization between sectors or regions that are not bound to technological or spatial closeness. The connection with the regional-sectorial distribution of the employment and the tendency to co-localize employment in specialized regions is accordingly of a major interest. From this geographical point of view, the formal characteristics of this algorithm do not constitute the main interest.

Table 4 shows that, in line with the SIA approach (Haedo and Mouchart, 2012), the choice of a specific metric for measuring the global localization may be crucial in the paths of the clustering procedures. This paper is based on the χ^2-divergence. Thus a promising avenue for future research might be to develop similar algorithms based on other metrics such as Kullback-Leibler divergence (KL) or Hellinger-distance (H) and compare the results with the present algorithm. In this direction, Rao (1995) and Marinelli and Winzer (2004) provide useful starting points.

Acknowledgements: 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. A special thank is due to Vicente N. Donato for the impetus he gave to the development of the topic of this paper.

References

Benabdeslem, K. and Allab, K. (2013), Bi-clustering continuous data with self-organizing map. *Neural Computing and Applications* **22**: 1551-1562.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., and Modha, D.S. (2007), A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Maching Learning Research* **8**: 1919-1986.

Benzécri, J.P. (1973), *Analyse des Données*. París: Dunod.

Benzécri, J.P. (1992), *Correspondence Analysis Handbook*. New York: Dekker.

Bickenbach, F., and Bode, E. (2006), Disproportionality measures of concentration, specialization and polarization. Kiel Institute for the World Economy, Working Paper 1276.

Bickenbach, F., and Bode, E. (2008), Disproportionality measures of concentration, specialization and localization. *International Regional Science Review* **31**: 359-388.

Bickenbach, F., Bode, E., and Krieger-Boden, C. (2010), Closing the gap between absolute and relative measures of localization, concentration or specialization. Kiel Institute for the World Economy, Working Paper 1660.

Bock, H.H. (1979), Simultaneous clustering of objects and variables. In *INRIA*: 187-203.

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

Braverman, E.M., Kiseleva, N.E., Muchnik, I.B., and Novikov, S.G. (1974), Linguistic approach to the problem of processing large bodies of data. *Automation and Remote Control* **35**: 1768-1788.

Caldas, J. and Kaski, S. (2011), Hierarchical generative biclustering for microrna expression analysis. *Journal of Computational Biology* **18**: 251-261.

Cazes, P. (1986), Correspondance entre deux ensembles et partition de ces deux ensembles. *Les Cahiers de l'Analyse des Données* **11**: 335-340.

Charrad, M., Lechevallier, Y., Saporta, G. and Ben Ahmed, M. (2009), Détermination du nombre des classes dans l'algorithme croki de classification croisée. In *EGC*: 447-448.

Cheng, Y. and Church, G.M. (2000), Biclustering of expression data. In *ISMB*: 93-103.

Ciampi, A., González Marcos, A. and Castejón Limas, M. (2005), Correspondence analysis and two-way clustering. *Statistics and Operations Research Transactions* **29**: 27-42.

Duffy, D.E. and Quiroz, A.J. (1991), A permutation-based algorithm for block clustering. *Journal of Classification* **8**: 65-91.

Escofier, B. (1978), Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistique Appliquée* **16**: 29-37.

Eckart, C. and Young, G. (1936), The approximation of one matrix by another of lower rank. *Psychometrika* **1**: 211-218. doi:10.1007/BF02288367.

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

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

Gilula, Z. (1986), Grouping and associations in contingency tables: an exploratory canonical correlation approach. *Journal of American Statistical Association* **81**: 773-779.

Goodman, L. (1981), Criteria for determining whether certain categories in a cross-classification table should be combined with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology* **87**: 612-650.

Goodman, L. (1985), The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13**: 10-69.

Govaert, G. (1977), Algorithme de classification d'un tableau de contingence. In *INRIA*: 487-500.

Govaert, G. (1995), Simultaneous clustering of rows and columns. *Control and Cybernetics* **24**, No. 4.

Govaert, G. and Nadif, M. (2008), Block clustering with bernoulli mixture models: comparison of different approaches. *Computational Statistics and Data Analysis* **52**: 3233-3245.

Govaert, G. and Nadif, M. (2010), Latent block model for contingency tables. *Communications in Statistics, Theory and Methods* **3**: 416-425.

Govaert, G. and Nadif, M. (2013), *Co-Clustering*. Hoboken: John Wiley & Sons.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M.J. (1988), Clustering the rows and columns of a contingency table. *Journal of Classification* **5**: 39-51.

Greenacre, M.J. (1993), Multivariate generalizations of correspondence analysis. In C.M. Cuadras and C.R. Rao (eds.), *Multivariate Analysis: Future Directions 2*. Amsterdam: North-Holland.

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

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 (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`

Haedo, C. and Mouchart, M. (2015a), Specialized agglomerations with lattice data: model and detection. *Spatial Statistics* **11**: 113-131.

Haedo, C. and Mouchart, M. (2015b), Methodological framework for the analysis of industrial geographical data. CIDETI Working Paper 2015/08, Università di Bologna, Representación en Argentina. Appendix for "La Geografía Industrial de América Latina y el Caribe" - MIALC, `http://lac.geoind.info`

Hartigan, J.A. (1972), Direct clustering of a data matrix. *Journal of the American Statistical Association* **67**: 123-129.

Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A.R. and Yildirim, M.A. (2015), *Atlas of Economic Complexity: Mapping Paths to Prosperity*. Cambridge, MA: MIT Press. `http://atlas.cid.harvard.edu/media/atlas/pdf/HarvardMIT_AtlasOfEconomicComplexity.pdf`

Hirotsu, C. (1983), Defining the pattern of association in two-way contingency tables. *Biometrika* **70**: 579-589.

Jagalur, M., Pal, C., Learned-Miller, E., Zoeller, R.T., and Kulp, D. (2007), Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics* **8**: S5.

Jambu, M. (1978), *Classification Automatique pour l'Analyse des Données, I- Méthodes et Algorithms*. Paris: Dunod.

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

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2015), Estimation and selection for the latent block model on categorical data. *Statistics and Computing* **25**: 1201-1216.

Lebart, L. and Mirkin, B.G. (1993), Correspondence analysis and classification. In C.M. Cuadras and C.R. Rao (eds.), *Multivariate Analysis: Future Directions*. Amsterdam: North-Holland.

Lebart, L., Morineau, A. and Warwick, K.H. (1984), *Multivariate Descriptive Statistical Analysis*. New York: John Wiley & Sons.

Madeira, S.C. and Oliveira, A.L. (2004), Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**: 24-45.

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

Marinelli, C. and Winzer, N. (2004), Agrupamiento de filas y columnas homogéneas en modelos de correspondencia. *Revista de Matemática: Teoría y Aplicaciones* **11**: 59-68.

Mirkin, B. (1996), *Mathematical Classification and Clustering*. Dordrecht: Kluwer.

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

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

Rao, C.R. (1995), A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *QÜESTIIÓ* **19**: 23-63.

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 M. Hazewinkel (ed.), *Encyclopedia of Mathematics*. New York: Springer.

Tang, C., Zhang, L., Zhang, A. and Ramanathan, M. (2001), Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In *BIBE*: 41-48.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D. and Brown, P. (1999), Clustering methods for the analysis of dna microarray data. Technical report, Department of Statistics, Stanford University.

van Mechelen, I., Bock, H.H., de Boeck, P. (2004), Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research* **13**: 363-394.

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