TWO-MODE CLUSTERING THROUGH PROFILES OF REGIONS AND SECTORS

Christian Haedo, Michel Mouchart

REPRINT | 2022 / 15

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Empirical Economics
https://doi.org/10.1007/s00181-022-02201-z

Two-mode clustering through profiles of regions and sectors

Christian Haedo[1] · **Michel Mouchart**[2]

Received: 24 February 2021 / Accepted: 3 January 2022

Abstract

This paper is concerned with simultaneously regrouping regions and sectors when analyzing the relative sectorial specialization of regions and the relative regional concentration of sectors. An automatic two-mode clustering algorithm is proposed with a view toward a concept of overall localization, corresponding to a discrepancy between an actual two-way contingency table (regions × sectors) and an hypothetical table reflecting independence between regions and sectors. This procedure identifies similar regions (respectively sectors) according to the relative sectorial (respectively regional) structure. This algorithm significantly reduces the size of the original table and obtain an optimal collapsed table with low level of information loss vis-à-vis the degree of overall localization. The properties and results of the algorithm are discussed through two applications, namely Argentina and Brazil.

Michel Mouchart gratefully acknowledges financial support from IAP research network grant *nr P*6/03 of the Belgian government (Belgian Science Policy). Both authors gratefully acknowledge the financial support of FOP, 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. Dominique Peeters also deserves a particular gratitude for a series of comments that lead to substantial improvements of a previous version of this paper. The help of Fernando Valli has been instrumental in shaping the algorithm developed in this paper and is deeply acknowledged. Highly appreciated and gratefully acknowledged are also comments given by technical analysts participating to seminars where this paper has been presented, namely at the Competitiveness and Innovation Division of the Inter-American Development Bank (IADB) in Madrid (Sp) and Washington (USA), at the Structural Policies and Innovation Unit of the OECD Development Centre in Paris (Fr) and at the Farnesina of the Ministero degli Affari Esteri e della Cooperazione Internazionale (MAECI) in Roma (It).

✉ Christian Haedo
chaedo@observatoriopyme.org.ar

1 Fundación Observatorio PyME (FOP), Ciudad Autónoma de Buenos Aires, Argentina

2 Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium

Published online: 10 March 2022

Keywords Relative sectorial specialization · Relative regional concentration · Overall localization · Two-mode clustering · Biclustering · Hierarchical clustering · Correspondence analysis · Large two-way contingency tables · Permutation bootstrap

1 Introducing the topic of this paper

The main purpose of this paper is to regroup simultaneously regions and sectors according to structural similarities in terms of sectorial specialization of regions and regional concentration of sectors. The economic activity is known to be spatially concentrated, generating mainly two types of agglomeration economies: localization economies, generated by the benefits derived from the presence of same sector firms in a geographical area (Fujita et al. 2001; Baldwin and Martin 2004; Combes and Gobillon 2015; among many others), and urbanization economies, generated by the advantages that firms obtain from large (and often economically diverse in terms of sector of activities) cities (Duranton and Puga 2000; Fujita and Thisse 2002; Viladecans-Marsal 2004; Cottineau et al. 2018). This paper concentrates the attention on the presence of economy of agglomerations without an analysis of their causal determinants, as done, among others, in Rosenthal and Strange (2004), Ellison et al. (2010) or Puga (2010). The industrial policy objectives can be better fulfilled if their decentralized design is more sensitive to the characteristics of the regions and of the sectors (Donato 2002; Nathan and Overman 2013). More explicitly, in the statistical analysis of specialization and concentration for understanding the sectorial and regional pattern of the economic activity, the simultaneous function of sectors and regions is known to have a prominent role.

Following the perspective of the stochastic independence approach (SIA) borrowed from the analysis of two-way contingency tables, namely regions $\times$ sectors, and developed in Haedo and Mouchart (2018), the relative measures of specialization and of concentration are based on the discrepancy (i.e. divergence or distance) between 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 *overall localization* of the economic activity (see Alonso-Villar and del Río 2013; Haedo and Mouchart 2018). This approach allows a simultaneous treatment of sectorial specialization and of regional concentration thanks to a symmetric manipulations of rows and columns. This feature enhances a global view of the relative roles of regions and sectors on the regional structure of economic activity.

This paper extends the established analysis of the overall localization in two directions. Firstly, simultaneous groupings of regions and sectors, rather than separate ones are examined. Secondly, the proposed procedure 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 sectorial 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, the objective is to look for a summary of large regions $\times$ sectors contingency tables that keeps

(almost) unchanged the measure of overall localization of the economic activity. This is obtained through an innovative extension of a biclustering algorithm.

There is a vast literature on biclustering, particularly in the fields of DNA microarray data (Ben Saber and Elloumi 2015), 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 et al. 2004; Jagalur et al. 2007). Biclustering methods are aimed at designing in a same exercise a clustering of the rows and the columns of a large array of data, an approach dating back to Hartigan (1972), Braverman et al. (1974), Govaert (1977), Bock (1979), Gilula (1986). These methods are expected to be useful to summarize large data sets by substantially smaller data sets with a similar structure. In general, the proposed algorithms are based on a concept of interaction between rows and columns, with data in the form of quantitative variables. This concept in the spirit of analysis of variance is based on local averages, often with an underlying hypothesis of normal distribution; for an historic view, see for instance Denis and Vincourt (1982) or Corsten and Denis (1990). A particularly interesting paper, Schepers et al. (2017), proposes an algorithm of maximal interaction for two-mode clustering.

The two-mode clustering algorithm developed in this paper is also in the family of so-called deterministic procedures, that operate in the spirit of descriptive statistical methods, see for instance, Duffy and Quiroz (1991), Lebart and Mirkin (1993), Govaert (1995), Mirkin (1996), Tibshirani et al. (1999), Cheng and Church (2000), Tang et al. (2001), Ciampi et al. (2005), Banerjee et al. (2007), Busygin et al. (2008), Charrad et al. (2009), Caldas and Kaski (2011), Benabdeslem and Allab (2013), Liu et al. (2018), Orzechowski et al. (2018). Most of these works are based on 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 grouping elements with similar profiles. The present algorithm is based on a technique of hierarchical clustering (HC), according to a dendrogram approach, combined with a correspondence analysis (CA). This HCCA procedure was introduced by Greenacre (1988) in a non-automatic computational program based on the cells. The present algorithm is based on distributions (marginal and conditional), rather than on cells, in an automatic procedure of grouping the cells.

For large tables, trying all possibilities of grouping is computationally expensive and may be not feasible. Therefore, a greedy algorithm that only ensures a local optimum is preferable, as already mentioned in Gan and Jiang (1999). Quite a good overview is given by Ben Saber and Elloumi (2015), with an emphasis, among others, on the interactive, exhaustive and greedy biclustering algorithms. Here, for each level of the dendrograms are built K collapsed tables, working in accordance with the dual processes just mentioned. Finally, at each level of the row and column dendrograms, permutation bootstrapping is used as a test that the envisaged regrouping performs better, see also Haedo (2009), Park et al. (2009) and Greenacre (2011).

The proposed algorithm looks for collapsing simultaneously regions and sectors with no restriction about the regions or the sectors to be clustered. Thus, for the

regions, no restriction of contiguity or of some distance-based pattern is operating. The non-inclusion of such restriction represents a "spaceless" framework motivated by multilevel policy-making rather than by spatial diffusion issues. Therefore, the clusters are kept constant under spatial permutations because of their structural nature: this is the case for clusters of regions with a similar *relative* sectorial specialization pattern, or regions with 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 *relative* regional concentration of sectors is at stake.

This paper is an extension of part of chapter 2 of Haedo (2009), with the following order of exposition. 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 two applications, namely Argentina and Brazil in the fourth section. The last section gathers some concluding remarks. These ones discuss some relationships between the related biclustering algorithms and the proposed algorithm, with particular emphasis on its innovative aspects making use of the SIA point of view.

2 The conceptual background of the algorithm

As a starting point, consider *the finite framework* for the analysis of specialization and concentration. For a given country, a finite set of disjoint regions $i \in \mathcal{I} = \{1, \ldots, I\}$, and a finite set of sectors $j \in \mathcal{J} = \{1, \ldots, J\}$ are considered, where a sector is the basic structure of a classification of economic activities.

For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, a number N_{ij} of primary units are observed; these could be for instance number of employees or number of establishments. Putting together these data, a two-way $I \times J$ contingency table $\mathbf{N} = [N_{ij}]$ is obtained, with the ith row for regions, jth column for sectors. The table totals are denoted as follows:

$$N_{i\cdot} = \sum_{j=1}^{J} N_{ij} \quad N_{\cdot j} = \sum_{i=1}^{I} N_{ij} \qquad N_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij} = \sum_{j=1}^{J} N_{\cdot j} = \sum_{i=1}^{I} N_{i\cdot} \tag{1}$$

These data may also be given in terms of the total number of data, $N_{\cdot\cdot}$, along with the conditional and marginal proportions:

$$p_{j|i} = \frac{N_{ij}}{N_{i\cdot}} \qquad p_{i|j} = \frac{N_{ij}}{N_{\cdot j}} \qquad p_{\cdot j} = \frac{N_{\cdot j}}{N_{\cdot\cdot}} \qquad p_{i\cdot} = \frac{N_{i\cdot}}{N_{\cdot\cdot}} \tag{2}$$

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. Following the stochastic independence approach (SIA), as developed in Haedo and Mouchart (2018), the relative measures of sectorial specialization and of regional 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 evaluated. When the components of a vector are indexed by i (regions) or by j (sectors), an arrow above the index that marks the components of the vector is used; for instance, one writes $p_{\vec{j}|i} = (p_{1|i}, \ldots, p_{j|i}, \ldots, p_{J|i})$ and for the global row profile (or marginal distribution) one writes $p_{.\vec{j}} = (p_{.1}, \ldots, p_{.j}, \ldots, p_{.J})$.

The relative sectorial specialization of region i is measured by a discrepancy $d(p_{\vec{j}|i} \mid p_{.\vec{j}})$ that operates a comparison between its profile (or conditional distribution) of the ith row, $p_{\vec{j}|i}$, and the global row profile (or marginal distribution), $p_{.\vec{j}}$, used as a benchmark of non-specialization. Similarly, the relative regional concentration of sector j is measured by a discrepancy $d(p_{\vec{i}|j} \mid p_{\vec{i}.})$ that operates the comparison between the profile (or conditional distribution) of the jth column, $p_{\vec{i}|j}$, and the global column profile (or marginal distribution), $p_{\vec{i}.}$, used as a benchmark of non-concentration.

The discrepancy $d([p_{ij}] \mid [p_{i.}\ p_{.j}])$ 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 and is taken as a benchmark of a completely non-specialized, or non-concentrated, economic structure. This measure represents the (global) information provided by the contingency table **N** about the overall localization of the economy. Note that for overall localization, Bickenbach and Bode use the term polarization in (2006) but localization in (2008); moreover Haedo (2009) used global specialization. In Bickenbach et al. (2010) use localization. Overall localization is used for instance by Alonso-Villar and del Río (2013) and is also adopted in Haedo and Mouchart (2018) and in this paper.

This analysis may be conducted by means of the so-called Location Quotient LQ_{ij} (Florence 1939), 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 (2018). This quotient, well known in the literature on relative sectorial specialization and on relative regional concentration, may be written as follows:

$$LQ_{ij} = \frac{p_{ij}}{p_{i.}\ p_{.j}} = \frac{p_{j|i}}{p_{.j}} = \frac{p_{i|j}}{p_{i.}} \tag{3}$$

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

$$\begin{array}{rll} LQ_{ij} = 1 \quad \text{or} & p_{ij} = p_{i.}\ p_{.j} & \text{non-specialization} \\ > 1 \quad \text{or} & p_{ij} > p_{i.}\ p_{.j} & \text{over-specialization} \\ < 1 \quad \text{or} & p_{ij} < p_{i.}\ p_{.j} & \text{under-specialization} \end{array} \tag{4}$$

where "non-specialization" stands for no sectorial specialization of region i or no regional concentration of sector j. Note that $p_{ij} = 0$ or $LQ_{ij} = 0$ is an indicator of maxima under-specialization.

For the measures of relative sectorial specialization, of relative regional concentration and of overall localization, three different discrepancies are the object of attention, namely the χ^2-divergence or Inertia, the Kullback–Leibler divergence and the Hellinger-distance. Once a measure of overall localization is considered as an adequate summary representation of the regional and sectorial 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 1 suggest three families of that algorithm. The $\mathcal{X}^2$ family rather speaks of Inertia and is the one explicitly used in this paper. The KL family uses the expression of "loss of information" because of its roots in information theory. Hellinger-distance is rather frequently used in mathematical statistics and presents a specific interest: it is an L_2-distance on the square root of the probabilities and is valued in the bounded interval $[0, 1]$, where the value 1 corresponds to mutually singular distributions and 0 corresponds to two identical distributions.

3 Two-mode clustering of regions and sectors

3.1 Motivation of the algorithm

A purpose of this algorithm 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 overall localization. The nature of the actual challenge should be kept in mind. In the case of Brazil, for instance, there are $I = 5138$ regions. Using just the first 2 digits of the International Standard Industrial Classification of manufacturing sectors (ISIC-Rev.3), there are $J = 22$ sectors. In 1998, for example, the total number of employees is $N = 6{,}018{,}445$. Therefore, the contingency table is a 5138×22 matrix of 6,018,445 primary units spread in 113,036 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 on the overall localization of the country. Collapsing tables means building tables of smaller dimension through aggregated regions (rows) and/or sectors (columns). In an exhaustive algorithm, the total number M of collapsed tables to be examined for the $I \times J$ matrix $\mathbf{N}$ is:

$$M = \sum\nolimits_{(m_1 \ldots m_i \ldots m_l)} \binom{I}{m_1 \ldots m_i \ldots m_l} \times \sum\nolimits_{(n_1 \ldots n_i \ldots n_k)} \binom{J}{n_1 \ldots n_j \ldots n_k} \quad (5)$$

where $l \leq I-1, k \leq J-1, m_1 + \ldots + m_i + \ldots + m_l < I$ and $n_1 + \ldots + n_j + \ldots + n_k < J$. Equation (5) 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 $l \leq n-1, m_1 + \ldots + m_i + \ldots + m_l < n$. 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 Eq. (5),

Table 1 Measures of specialization, of concentration and of overall localization

Concepts	χ^2-divergence or Inertia, $d_{\chi^2}(\cdot \mid \cdot)$	Kullback–Leibler divergence, $d_{KL}(\cdot \mid \cdot)$	Hellinger-distance, $d_H^2(\cdot \mid \cdot)$
Relative sectorial	$= \sum_j \frac{(p_{j\mid i} - p_{\cdot j})^2}{p_{\cdot j}}$	$= \sum_j p_{j\mid i} \log\left(\frac{p_{j\mid i}}{p_{\cdot j}}\right)$	$= \frac{1}{2}\sum_j (\sqrt{p_{j\mid i}} - \sqrt{p_{\cdot j}})^2$
specialization of region i,	$= \sum_j p_{\cdot j}\,(LQ_{ij} - 1)^2$	$= \sum_j p_{\cdot j}\, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2}\sum_j p_{\cdot j}\,(\sqrt{LQ_{ij}} - 1)^2$
$d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$			
Relative regional	$= \sum_i \frac{(p_{i\mid j} - p_{i\cdot})^2}{p_{i\cdot}}$	$= \sum_i p_{i\mid j} \log\left(\frac{p_{i\mid j}}{p_{i\cdot}}\right)$	$= \frac{1}{2}\sum_i (\sqrt{p_{i\mid j}} - \sqrt{p_{i\cdot}})^2$
concentration of sector j,	$= \sum_i p_{i\cdot}\,(LQ_{ij} - 1)^2$	$= \sum_i p_{i\cdot}\, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2}\sum_i p_{i\cdot}\,(\sqrt{LQ_{ij}} - 1)^2$
$d(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$			
Overall localization,	$= \sum_i \sum_j \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}$	$= \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}\right)$	$= \frac{1}{2}\sum_i \sum_j (\sqrt{p_{ij}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$
$d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$	$= \sum_i \sum_j \frac{p_{i\cdot}(p_{j\mid i} - p_{\cdot j})^2}{p_{\cdot j}}$	$= \sum_i \sum_j p_{i\cdot}\, p_{j\mid i} \log\left(\frac{p_{j\mid i}}{p_{\cdot j}}\right)$	$= \frac{1}{2}\sum_i \sum_j (\sqrt{p_{i\cdot}\, p_{j\mid i}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$
	$= \sum_i \sum_j \frac{p_{\cdot j}(p_{i\mid j} - p_{i\cdot})^2}{p_{i\cdot}}$	$= \sum_i \sum_j p_{\cdot j}\, p_{i\mid j} \log\left(\frac{p_{i\mid j}}{p_{i\cdot}}\right)$	$= \frac{1}{2}\sum_i \sum_j (\sqrt{p_{\cdot j}\, p_{i\mid j}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$
	$= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\,(LQ_{ij} - 1)^2$	$= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2}\sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\,(\sqrt{LQ_{ij}} - 1)^2$

we have $n = I$ and $m = J$. More information may be found in Rota (1964), Gardner (1978), Branson (2000) or Sloane (2001). For I and J large, as in the present case, M is huge and trying all possibilities is computationally either unfeasible or very expensive. Therefore, for large tables, a greedy algorithm that ensures a suitable approximation of the global optimum is preferred, as already mentioned in Gan and Jiang (1999). For a good overview on the interactive, exhaustive and greedy biclustering algorithms see also Ben Saber and Elloumi (2015).

3.2 Singular value decomposition and correspondence analysis

Let the algorithm be presented on the basis of the stochastic independence approach (SIA). Let

$$\mathbf{P} = \frac{\mathbf{N}}{N_{..}} = \left[\frac{N_{ij}}{N_{..}}\right]$$

be the probability matrix corresponding to $\mathbf{N}$, $\mathbf{r} = [p_{i\cdot}] = p_{\vec{i}\cdot}$ the vector of row marginals, $\mathbf{c} = [p_{\cdot j}] = p_{\cdot\vec{j}}$ 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. Let $\mathbf{R} = [p_{ij} - p_{i\cdot}\, p_{\cdot j}]$ be the $I \times J$ matrix of residuals between an observed p_{ij} and an expected one, under an hypothesis of independence, $p_{i\cdot}\, p_{\cdot j}$. The matrix of the residuals may be conveniently written as: $\mathbf{R} = \mathbf{P} - \mathbf{r}\mathbf{c}'$. Later on, it will be rather worked on the matrix of the standardized residuals defined as:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\mathbf{R}\mathbf{D}_c^{-1/2} \qquad s_{ij} = \frac{p_{ij} - p_{i\cdot}.p_{\cdot j}}{\sqrt{p_{i\cdot}.p_{\cdot j}}}. \tag{6}$$

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

$$s_{ij} = \sqrt{p_{i\cdot}.p_{\cdot j}}\ \left[LQ_{ij} - 1\right] \tag{7}$$

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 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 viewed as an adjustment for the problems raised by small regions, and small sectors, in line with the works of Moineddin et al. (2003), O'Donoghue and Gleave (2004) and Guimarães, Figueiredo and Woodward (Guimarães et al. 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:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}' \tag{8}$$

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, evidently, $\lambda_{K+1} = \cdots = \lambda_{\max(I,J)} = 0$. The dimension of $\mathbf{U}$ is $I \times K$, of $\mathbf{V}$ is $J \times K$ and $\mathbf{D}_\lambda$ is accordingly $K \times K$; moreover $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = I_{(K)}$. The space $\boldsymbol{R}^K$ is called the *factor space*. Remembering that with a matrix $E = [e_{ij}]$ one has: $tr(E'E) = \sum_i \sum_j e_{ij}^2$, the χ^2-divergence between $[p_{ij}]$ and $[p_{i\cdot}\, p_{\cdot j}]$, or total inertia, may now be written as:

$$\begin{aligned} d_{\chi^2}\left([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}]\right) = \phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i\cdot}\, p_{\cdot j}\right)^2}{p_{i\cdot}\, p_{\cdot j}} = \sum_i \sum_j s_{ij}^2 = tr\mathbf{S}'\mathbf{S} \\ &= \sum_{k=1}^{K} \lambda_k^2 \end{aligned} \tag{9}$$

Thus, the 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. Indeed, CA defines a sequence of subspaces containing an increasing proportion of the total inertia; for more information see e.g. Benzécri (1973; 1992), Lebart et al. (1984), Greenacre (1984, 1993, 2007). More explicitly, the principal coordinates for the rows (regions) are defined as:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\,\mathbf{U}\,\mathbf{D}_\lambda = [\,f_{ik}] \qquad f_{ik} = p_{i\cdot}^{-1/2}\,\lambda_k\,u_{ik} \qquad (i,k) \in I \times K \tag{10}$$

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

$$\mathbf{F}'\mathbf{D}_r\mathbf{F} = \mathbf{D}_\lambda^2 \qquad i.e. \quad \lambda_k^2 = \sum_i p_{i\cdot}\, f_{ik}^2 \tag{11}$$

Thus, Eq. (11) decomposes the kth eigenvalue of $\mathbf{S}'\mathbf{S}$, also the kth component of the inertia, according to the contribution of each region i, namely $p_{i\cdot}\, f_{ik}^2$. Similarly, the principal coordinates for the columns (sectors) are defined as:

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\,\mathbf{V}\,\mathbf{D}_\lambda = [g_{jk}] \qquad g_{jk} = p_{\cdot j}^{-1/2}\,\lambda_k\,v_{jk} \qquad (j,k) \in J \times K \tag{12}$$

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

$$\mathbf{G}'\mathbf{D}_c\mathbf{G} = \mathbf{D}_\lambda^2 \qquad i.e. \quad \lambda_k^2 = \sum_j p_{\cdot j}\, g_{jk}^2 \tag{13}$$

Thus, Eq. (13) decomposes the kth eigenvalue of $\mathbf{S}'\mathbf{S}$ according to the contribution of each sector j, where $p_{\cdot j} g_{jk}^2$ 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 m first eigenvalues λ_k and $\mathbf{U}_{(m)}$ and $\mathbf{V}_{(m)}$ be the submatrices made of the m first 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 overall localization measure ϕ^2, in Eq. (9), in terms of the contributions of each factor k and of the contribution of the regions i, respectively the sectors j:

$$\phi^2 = \sum_i \sum_k p_{i\cdot}\, f_{ik}^2 = \sum_j \sum_k p_{\cdot j}\, g_{jk}^2 \tag{14}$$

For more information, see e.g. Mardia et al. (1979), Jobson (1992) and Greenacre (2007).

Let us write the rows and columns profiles as follows:

$$\mathbf{D}_r^{-1}\,\mathbf{P} = \left[\frac{p_{ij}}{p_{i\cdot}}\right] = \left[p_{j|i}\right] \qquad \mathbf{P}\,\mathbf{D}_c^{-1} = \left[\frac{p_{ij}}{p_{\cdot j}}\right] = \left[p_{i|j}\right] \tag{15}$$

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

$$d_{\chi^2}\left(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}\right) = \sum_j \frac{\left(p_{j|i} - p_{\cdot j}\right)^2}{p_{\cdot j}} = \left(p_{\vec{j}|i} - p_{\cdot\vec{j}}\right)' D_c^{-1} \left(p_{\vec{j}|i} - p_{\cdot\vec{j}}\right) = \sum_k f_{ik}^2 \tag{16}$$

$$d_{\chi^2}\left(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}\right) = \sum_i \frac{\left(p_{i|j} - p_{i\cdot}\right)^2}{p_{i\cdot}} = \left(p_{\vec{i}|j} - p_{\vec{i}\cdot}\right)' D_r^{-1} \left(p_{\vec{i}|j} - p_{\vec{i}\cdot}\right) = \sum_k g_{jk}^2 \tag{17}$$

Therefore, the decomposition of the total inertia as a measure of overall localization, in (14), may also be written in terms of average of relative sectorial specialization or of regional concentration:

$$\phi^2 = \sum_i p_{i\cdot}\, d_{\chi^2}\left(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}\right) = \sum_j p_{\cdot j}\, d_{\chi^2}\left(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}\right) \tag{18}$$

3.3 Automatic optimal collapsed table: the algorithm based on HCCA

This section gives the essentials of the algorithm, developed using the software R.

An overview

In line with (16) and (17), 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:

$$d^2_{D_c}\left(p_{\vec{j}|i} \mid p_{\vec{j}|i'}\right) = \sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} - p_{j|i'}\right)^2 = \left(p_{\vec{j}|i} - p_{\vec{j}|i'}\right)' D_c^{-1} \left(p_{\vec{j}|i} - p_{\vec{j}|i'}\right) \tag{19}$$

$$d^2_{D_r}\left(p_{\vec{i}|j} \mid p_{\vec{i}|j'}\right) = \sum_i \frac{1}{p_{i\cdot}} \left(p_{i|j} - p_{i|j'}\right)^2 = \left(p_{\vec{i}|j} - p_{\vec{i}|j'}\right)' D_r^{-1} \left(p_{\vec{i}|j} - p_{\vec{i}|j'}\right) \tag{20}$$

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

$$\frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} + p_{i'\cdot}} \sum_j \frac{1}{p_{\cdot j}} \left(p_{j|i} - p_{j|i'}\right)^2 = \frac{p_{i\cdot}\, p_{i'\cdot}}{p_{i\cdot} + p_{i'\cdot}} (p_{\vec{j}|i} - p_{\vec{j}|i'})' D_c^{-1} \left(p_{\vec{j}|i} - p_{\vec{j}|i'}\right) \tag{21}$$

The overall 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 (19) and (20). The two rows are then merged by summing their frequencies and the profile and mass are recalculated. The same measure of difference as (21) is calculated at each stage of the clustering. Similarly, mergings of two columns are also operated. By doing so, each clustering decreases the dimension of the table.

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 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:

$$\mathcal{I}_{(0)} = \{\{1\}, \{2\}, \ldots, \{I\}\} \qquad \mathcal{J}_{(0)} = \{\{1\}, \{2\}, \ldots, \{J\}\} \tag{22}$$

$$\mathcal{I}_{(I)} = \{\{1, 2, \ldots, I\}\} \qquad \mathcal{J}_{(J)} = \{\{1, 2, \ldots, J\}\} \tag{23}$$

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

First step: building collapsed tables from HCCA

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

Consider the first m columns of $\mathbf{F}$, i.e. let the $I \times m$ matrix $\mathbf{F}_{(m)} = (f_{\vec{i}1}, \ldots, f_{\vec{i}k}, \ldots, f_{\vec{i}m})$, where $f_{\vec{i}k}$ represents the kth column of $\mathbf{F}$, and obtain a hierarchical clustering of the rows of $\mathbf{F}_{(m)}$, corresponding to the regions, or rows of $\mathbf{S}$. For $a = 0, \ldots, I$, let $\mathcal{I}_{(a,m)}$ be a partition of $\mathcal{I}$ with $I - a$ elements, i.e. $\mathcal{I}_{(a,m)} = \{\mathcal{I}_{1,(a,m)}, \ldots, \mathcal{I}_{l,(a,m)}, \ldots, \mathcal{I}_{I-a,(a,m)}\}$, with $\forall m : \mathcal{I}_{(0,m)} = \mathcal{I}_{(0)}$ and $\mathcal{I}_{(I,m)} = \mathcal{I}_{(I)}$, and each element of the partition $\mathcal{I}_{l,(a,m)}$ is a subset of regions in the first m ones. Consider now the nested sequence of $I + 1$ partitions of regions, starting with $\mathcal{I}_{(0)}$ and where each cluster is obtained as an optimized clustering scheme based on $\mathcal{I}_{(a-1,m)}$. Thus for each m, there are only $I - 1$ relevant levels of the hierarchical clustering.

2. *Work on the columns (sectors).* Again, for each m, with $m = 1, 2, \ldots, K$:

   Repeat the same with the columns of $\mathbf{G}$, with the $J \times m$ matrix $\mathbf{G}_{(m)} = (g_{\vec{j}1}, \ldots, g_{\vec{j}k}, \ldots, g_{\vec{j}m})$, where $g_{\vec{j}k}$ represents the kth column of $\mathbf{G}$, and obtain a dendrogram through an hierarchical clustering of the rows of $\mathbf{G}_{(m)}$, corresponding to the sectors, or columns of $\mathbf{S}$. Similarly, for $b = 0, \ldots, J$, let $\mathcal{I}_{(b,m)}$ be a partition of $\mathcal{J}$ with $J - b$ elements, i.e. $\mathcal{J}_{(b,m)} = \{\mathcal{J}_{1,(b,m)}, \ldots, \mathcal{J}_{l,(b,m)}, \ldots, \mathcal{J}_{J-b,(b,m)}\}$, with $\forall m : \mathcal{J}_{(0,m)} = \mathcal{J}_{(0)}$ and $\mathcal{J}_{(J,m)} = \mathcal{J}_{(J)}$, and each element of the partition $\mathcal{J}_{l,(b,m)}$ is a subset of sectors in the first m ones. Again, consider now the nested sequence of $J + 1$ partitions of sectors, starting with $\mathcal{J}_{(0)}$ and where each cluster is obtained as an optimized clustering scheme based on $\mathcal{J}_{(b-1,m)}$. Thus for each m, there are only $J - 1$ relevant levels of the hierarchical clustering.

3. *Building collapsed tables*:

   For each level of the row and column dendrograms, build K collapsed tables, each of dimension $(I - 1)(J - 1)$, repeat the operation for each m, obtaining accordingly K collapsed tables $T_{\mathcal{I}_{(a,m)} \times \mathcal{J}_{(b,m)}}$, and calculate the corresponding inertia $\phi^2\left(T_{\mathcal{I}_{(a,m)} \times \mathcal{J}_{(b,m)}}\right)$.

Second step: identifying an optimal collapsed table through bootstrapping

Up to now, the greedy algorithm has built, from all the dendrograms levels, an array A of $(I-1)(J-1)K$ collapsed tables. The final question is: which of these collapsed tables is better in the sense of an optimal compromise between a smallest table and an highest overall localization (i.e. association) possible? This control is realized by permutation bootstrapping that provides an efficient tool for achieving this trade-off. Let us consider whether a particular table $T_{\mathcal{I}_{(a,m)} \times \mathcal{J}_{(b,m)}}$ 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 (a, b, m). Let π_r be a permutation defined on $\mathcal{I}$, i.e. $\pi_r : \mathcal{I} \to \mathcal{I}$, bijective and let us write $\pi_r(\mathcal{I}_{(a,m)})$ for the image of the partition $\mathcal{I}_{(a,m)}$ transformed by π_r. Similarly, let π^c be a permutation defined on $\mathcal{J}$ and its image $\pi^c(\mathcal{J}_{(b,m)})$ of the partition $\mathcal{J}_{(b,m)}$. Given (π_r, π^c), a transformed table $T_{\pi_r(\mathcal{I}_{(a,m)}) \times \pi^c(\mathcal{J}_{(b,m)})}$ is defined following the same partition scheme as the optimized table $T_{\mathcal{I}_{(a,m)} \times \mathcal{J}_{(b,m)}}$ with shuffled labels. The corresponding inertia $\phi^2\left(T_{\pi_r(\mathcal{I}_{(a,m)}) \times \pi^c(\mathcal{J}_{(b,m)})}\right)$ is then computed.

Note that the transformed table $T_{\pi_r(\mathcal{I}_{(a,m)})\times\pi^c(\mathcal{J}_{(b,m)})}$ has a same dimension as $T_{\mathcal{I}_{(a,m)}\times\mathcal{J}_{(b,m)}}$; thus, their inertia are comparable. The difference $\phi^2\left(T_{\mathcal{I}_{(a,m)}\times\mathcal{J}_{(b,m)}}\right) - \phi^2\left(T_{\pi_r(\mathcal{I}_{(a,m)})\times\pi^c(\mathcal{J}_{(b,m)})}\right)$ is an effect of the label shuffling of rows and 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, in the array A, the differences

$$\psi(a,b,m) = \phi^2\left(T_{\mathcal{I}_{(a,m)}\times\mathcal{J}_{(b,m)}}\right) - \mathbb{E}_B\,\phi^2\left(T_{\pi_r(\mathcal{I}_{(a,m)})\times\pi^c(\mathcal{J}_{(b,m)})}\right) \tag{24}$$

represent how much the optimized table has gained, in inertia, relatively to a table with a same cluster scheme but with randomly shuffled rows and columns. The algorithm terminates defining the optimal collapsed table, $\mathcal{I}_{(a^*,m^*)} \times \mathcal{J}_{(b^*,m^*)}$, by solving the following maximization problem:

$$\left(a^*, b^*, m^*\right) = \arg\max_{a,b,m}\ \psi(a,b,m), \tag{25}$$

balancing the trade-off between the overall localization 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. □

4 Applications

Preliminary works with simulated data had provided encouraging results about the performance of this algorithm, motivating a further step with real data. The next step was to let work this algorithm on real data of several countries. The purpose of these applications is to analyze and evaluate how far the functioning of this algorithm provides suitable information on the regional–sectorial structure of the economic activity. Two of these applications, for Argentina and for Brazil, are presented for different aspects of the results of this algorithm. These results are presented with the objective of making explicit the characteristics of the proposed methodology and some insights about how these might be used for policy making.

The sector classification used in both applications refers to the first 2 digits (divisions) of the International Standard Industrial Classification (ISIC-Rev.3.1) of manufacturing industry (23 divisions) and is given in Table 2.

The optimal collapsed tables of Argentina and Brazil are found using $B = 1000$ permutation bootstraps. The first eigenvalues selected in the steps of the algorithm where the optimal collapsed tables were found are: the first 15 out of 22 for Argentina and the first 15 out of 21 for Brazil.

Table 2 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 nonmetallic 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.
37	Recycling

4.1 Application for Argentina

Here, the regions are the lower level political-administrative jurisdictions called departments (511) of Argentina. After eliminating those without employees in the manufacturing industry, there are remaining 491. The data, related to the number of employees in the manufacturing industry, were obtained from of the National Institute of Statistics and Censuses of Argentina (INDEC-2004: 941,337 employees).

Table 3 shows a summary of results in terms of the three measures of overall localization proposed in Table 1, concerning the original and the optimal collapsed table, namely the number of cells and the resulting loss of information about the level of overall localization.

The loss of information, between 18.9% and 34.5%, seems quite attractive vis-à-vis a substantial reduction in the size of the table (94.9%). The percentages of the loss of information clearly depend on the underlying divergences. That $d_{\mathcal{X}^2}$ be associated

Table 3 Summary results of Argentina

Measure	Original table	Optimal collapsed table	Lost level of overall localization (%)
$d_{\mathcal{X}^2}$	1.9519	1.5833	18.9
d_{KL}	0.2060	0.1350	34.5
d_H	0.0949	0.0671	29.4
# of cells	11,293	578	
	(491 × 23)	(34 × 17)	
Reduction # of cells		94.9%	

with the smallest loss of information may be connected to the fact that this version of the algorithm is based on optimizing the loss in terms of $d_{\mathcal{X}^2}$. Figures 1 and 2 are heatmaps that allows to summarize and visualize large tables by representing each cell by one color: white for cells non-specialized or non-concentrated, yellow for cells under-specialized or under-concentrated and orange for cells over-specialized or over-concentrated. These figures provide the contribution to overall localization of each pair (region, sector): Figure 1, for the original data—thus 491 rows and 23 columns—and, Fig. 2, for the optimal collapsed table- thus 34 rows, corresponding to g-regions, and 17 columns, corresponding to g-sectors. In Fig. 1, the g-regions are separated by an horizontal black line and numerated in the first column (under GR). Similarly for the g-sectors, in the first row, in case of two lines, the first one gives the number of the g-sector (GS) and the second one gives the original labels of the sector.

These results provide a deeper understanding about the realization of specialization and concentration in economy. Thus, it may be clearly viewed that every GR has a specific characterization in terms of sectorial specialization and every GS in terms of regional concentration. For instance, the GR1 is over-specialized in 5 sectors and GS1 is over-concentrated in 2 GR's. These facts suggest that the results of the two-mode clusterization provides underlying information on the process leading to these specializations and concentrations for specific GR's and GS's.

Let us now have a spatial view on the results of the algorithm. Figure 3 gives a map of the GR4 where the right-hand side part gives a zoom in of the metropolitan zone of Buenos Aires. As mentioned in the Introduction, this algorithm does not impose restrictions of contiguity or of some distance-based pattern on the regions to be clustered. Nevertheless, it should be noted that in the metropolitan zone of Buenos Aires and in the rest of Argentina several neighboring original regions are clustered within the GR4. The GR4 clusters 37 regions with 26.3% for the manufacturing occupation at the national level and is specialized in the 7 g-sectors: 17, 21, GS2, GS3, 29, 34 and 37 with 30.5%, 34.3%, 41.5%, 39.8%, 28.7%, 28.1% and 33.6%, respectively, of manufacturing occupation at national level. In the metropolitan zone of Buenos Aires, 18 regions are contiguous representing 83.8% of the occupation in GR4 and 22.0% at the national level. These 18 regions represent, on the specialized sectors in GR4, a manufacturing occupation of 88.5%, 77.6%, 81.7%, 87.1%, 83.9%, 93.0% and 60.2%, respectively; and 27.0%, 26.6%, 33.9%, 34.7%, 24.1%, 26.2% and 20.3%,

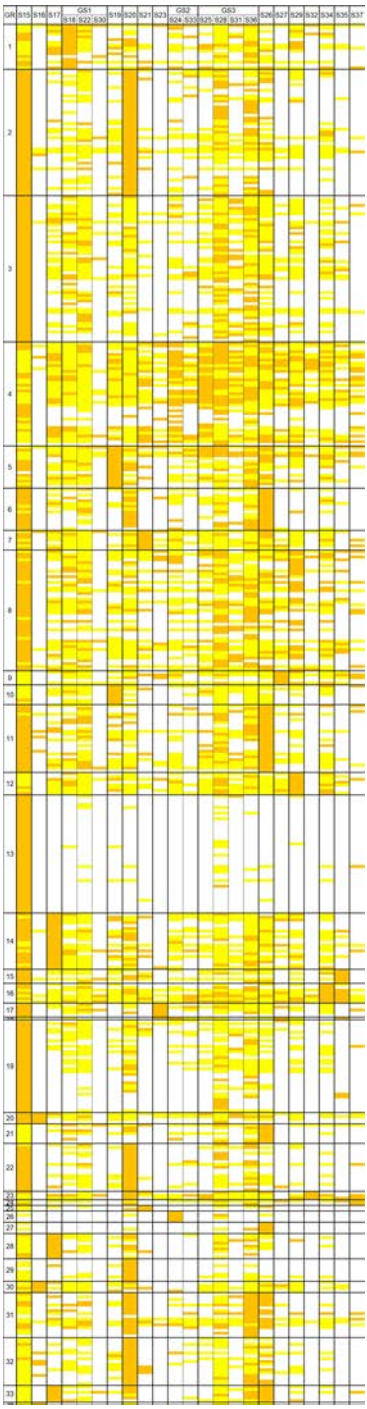

Fig. 1 Heatmap of the original table of Argentina

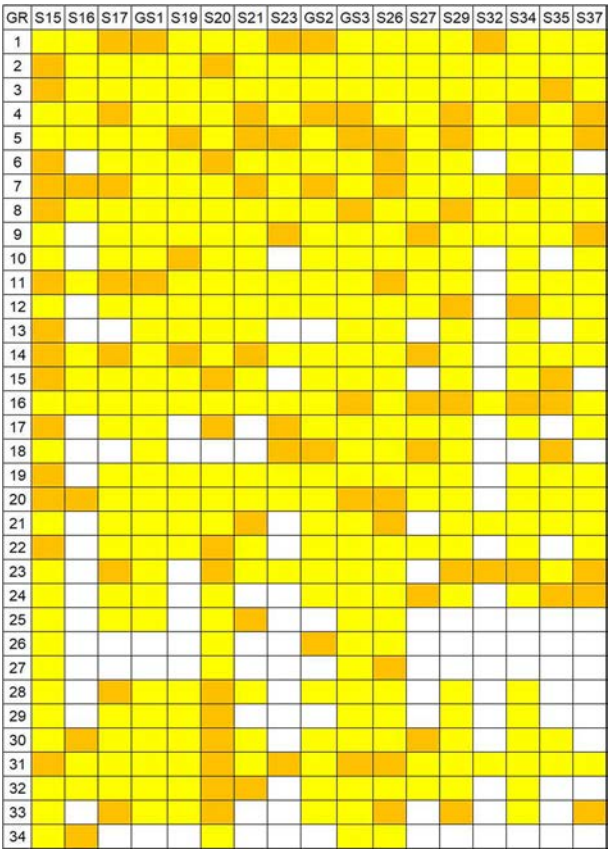

Fig. 2 Heatmap of the optimal collapsed table of Argentina

respectively, of manufacturing occupation at national level. These results provide an information in a form particularly relevant for developing economic spatial and sectorial multilevel policies.

4.2 Application for Brazil

Again, the regions are the lower level political-administrative jurisdictions called municipalities (5138) of Brazil. The data, related to the number of employees in the manufacturing industry, were obtained from of the National Institute of Statistics and Censuses of Brazil (IBGE-1998: 6,018,445 employees).

Similarly to Table 3 for Argentina, Table 4 shows a summary of results obtained from the three measures of overall localization proposed in Table 1.

It may be noticed that the (5138×22) original table is now reduced to a (52×17) optimal collapsed table, i.e. a reduction of 99.2% of the number of cells but only a loss of information of 21.1% in terms of the $d_{\mathcal{X}^2}$.

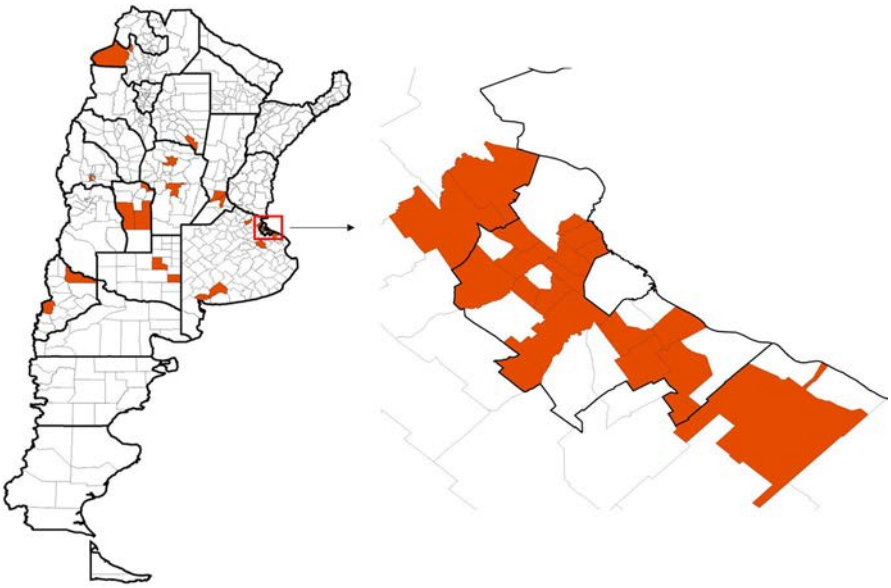

Fig. 3 GR4 of Argentina

Table 4 Summary results of Brazil

Measure	Original table	Optimal collapsed table	Lost level of overall localization (%)
$d_{\mathcal{X}^2}$	3.0604	2.4151	21.1
d_{KL}	0.3222	0.2140	33.6
d_H	0.1524	0.1067	30.0
# of cells	113,036	884	
	(5138 × 22*)	(52 × 17)	
Reduction # of cells		99.2%	

*For Brazil, sector 37 is included in sector 36 (ISIC-Rev.3)

Similarly to Fig. 2 for Argentina, Fig. 4 is the heatmap of the optimal collapsed table, thus 52 rows, corresponding to GR's, and 17 columns, corresponding to GS's. Again, it may be observed that every GR has a specific characterization in terms of sectorial specialization and every GS in terms of regional concentration. For instance, the GR2 is over-specialized in 2 sectors and GS2 is over-concentrated in 3 GR's.

The results of the algorithm for Brazil are now used to comment on the standardized residuals, introduced in (7). As mentioned before, these residuals may be viewed as an homothetic transformation of the location quotient, thus as a measure of local relative specialization and concentration, in connection to the concept of overall localization. Figure 5 depicts the standardized residuals of the 113,036 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- or under-specialized or over- or under-

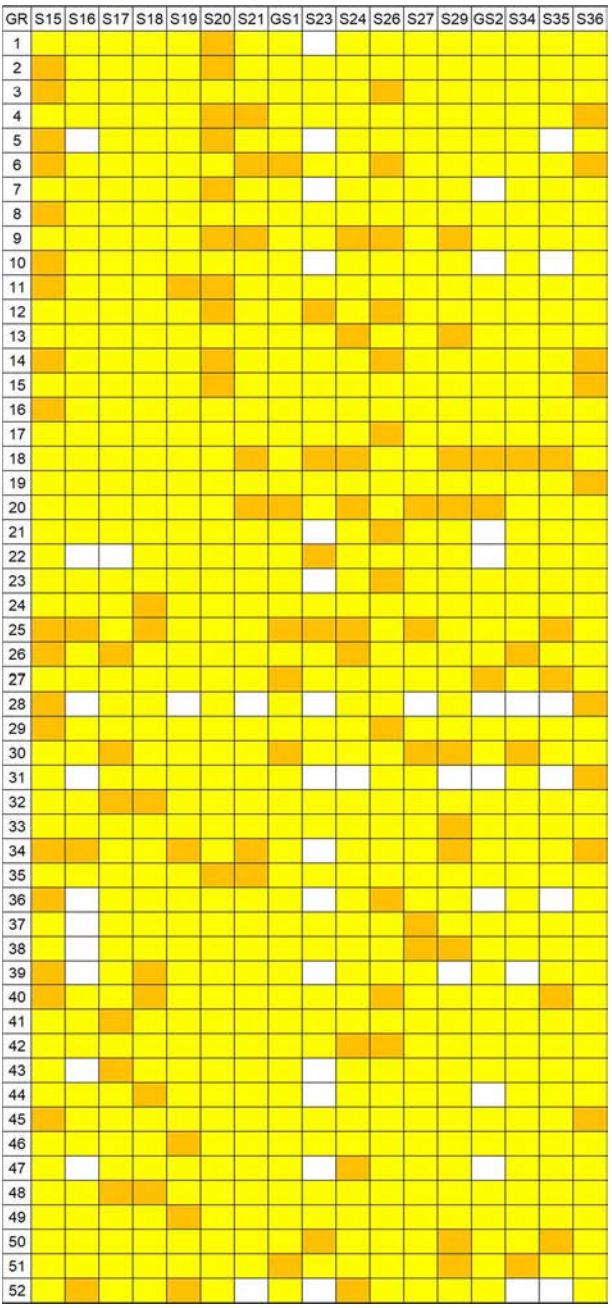

Fig. 4 Heatmap of the optimal collapsed table of Brazil

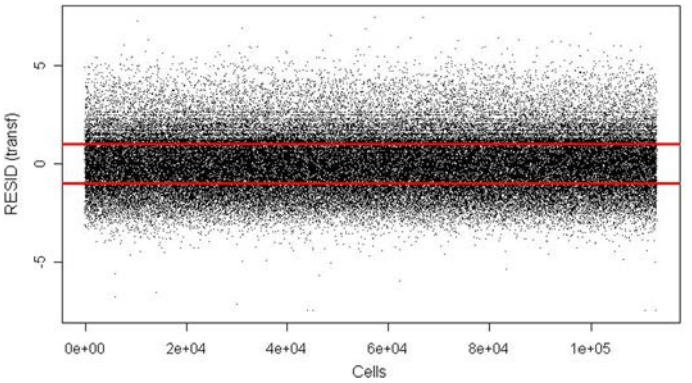

Fig. 5 Standardized residuals of the original table of Brazil

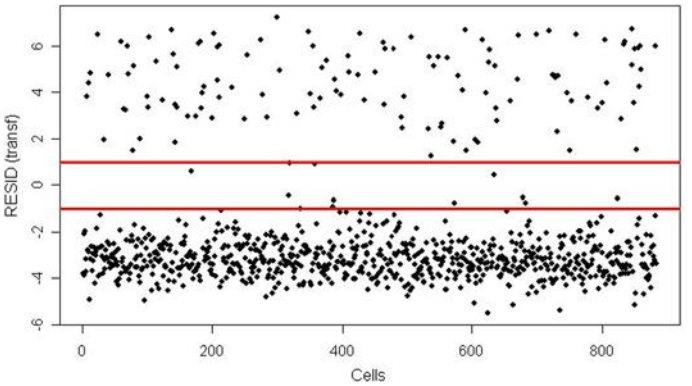

Fig. 6 Standardized residuals of the optimal collapsed table of Brazil

concentrated, suggesting that the degree of overall localization can be explained with much less information. Figure 6 shows the standardized residuals for the 884 cells of the optimal collapsed table obtained by grouping regions and sectors. The strong decrease of the zero and close to zero residuals make the overall localization phenomenon more apparent. As a matter of fact, grouping into g-regions and g-sectors succeeds in better identifying simultaneously homogenous regions of similar relative specialization structure and homogenous sectors of similar relative concentration structure.

5 Conclusions

Section 3 describes in a systematic way the steps of the algorithm, based on the HCCA procedure that has been widely used in combination with different types of algorithms. The innovative aspects of the present procedure are now made explicit. A distinctive feature of this greedy algorithm is to be completely automatic in the sense that, in the final solution, the number of clusters and the dimension of the factor

space m are determined by the algorithm itself rather than by pre-specified parameters. Therefore, the underlying concepts of GR's and of GS's are completely specified by the algorithm itself, at variance from an algorithm requiring the a priori specification of some parameters. This algorithm produces an automatic computational program based on the distributions (marginal and conditional) rather than on the cells.

Through the approach of SIA, this algorithm is providing an understanding of the similarities and dissimilarities of the simultaneous dual processes generating a family of concepts, namely sectorial specialization, regional concentration and overall localization. This innovative aspect has been illustrated on the heatmap of Fig. 1 for the application of Argentina.

This greedy algorithm has been developed as a trade-off between the computational time and the approximation of a global optimum available from an exhaustive algorithm. In an exhaustive algorithm, the total number of collapsed tables to be examined is given by equation (5) for each m, to be identified from formulae (10) and (12), and consequently the duration of computing is increasing exponentially with the size of the table.

The proximity of the final result of this greedy algorithm with respect to the result of an exhaustive algorithm is an important issue. The performances given in Tables 3 and 4 indicate that the result of an exhaustive algorithm is not better than the result of the present greedy algorithm, in particular because the greedy algorithm achieves a diminution of the number of cells of the table by more than 95%. Interestingly enough, the greedy algorithm solution of Argentina has been confronted to a further clustering of rows and columns separately according to the algorithm of Hahsler et al. (2019) that revealed no further clustering. These positive performances are due to the use of permutation bootstrap for controlling the efficiency of each step, as mentioned in the second step of Sect. 3.3 and just after Table 2. Indeed, Gan and Jiang (1999) had concluded that the bootstrapping techniques are a best way to progressively achieve the global optimum on large tables.

Here, for each level of the row and column dendrograms are built K collapsed tables, each of dimension $(I-1)(J-1)$; this follows the strategy of working in accordance with the dual processes just mentioned. The operation is repeated for each m, obtaining accordingly K collapsed tables $T_{\mathcal{I}_{(a,m)}\times\mathcal{J}_{(b,m)}}$ and their bootstrap versions $T_{\pi_r(\mathcal{I}_{(a,m)})\times\pi^c(\mathcal{J}_{(b,m)})}$, with the corresponding inertia $\phi^2\left(T_{\mathcal{I}_{(a,m)}\times\mathcal{J}_{(b,m)}}\right)$ and their bootstrap versions $\phi^2\left(T_{\pi_r(\mathcal{I}_{(a,m)})\times\pi^c(\mathcal{J}_{(b,m)})}\right)$. So, the final cutoff level of the row and column dendrograms is obtained from the optimal collapsed table, rather from the permutation bootstrap made separately for the rows and the columns. This may be viewed as a contribution to the framework derived from SIA that simultaneously considers a family of concepts rather than a succession of separated concepts.

Several challenges are at stake, three of which should be singled out. A *first challenge* consists in identifying regions, respectively sectors, with identical, or similar, structure. These regions, respectively sectors, might be associated in view of developing a joint policy of between- and within-sectorial and regional cooperation. At contrast, a *second challenge* consists in identifying regions, respectively sectors, with complementary sectorial, respectively regional, structure. 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 controlled loss of information while the reduction of the overall dimension is substantial. This is indeed a crucial challenge with large tables.

- The basic criterion for grouping regions and sectors is the similarity of the dual sectorial–regional structure: this is the natural answer to the *first challenge*.
- Distinct GR's imply different sectorial structures between the GR's and similar sectorial structure of the regions within each GR. This property of the proposed algorithm also provides a useful information for a multilevel policy maker, either within a country or between several countries. For instance, it is of a particular interest for policy cooperation to notice that the GR2 of Argentina (Fig. 2) and the GR2 of Brazil (Fig. 4) are over- and under-specialized in the same sectors. 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).
- Tables 3 and 4 indicate that the optimal collapsed tables of this greedy algorithm have diminished the number of cells by around 95% and looses between 20 and 35% of the information, according to the used discrepancy. This is quite an achievement regarding the *third challenge*. These results also demonstrate that one of the innovative aspects of this algorithm is in providing flexibility for interpreting the final results, through different discrepancies.

The results of the applications provide a new way for simultaneously understanding the similarities and dissimilarities of the dual process of sectorial specialization and regional concentration, without spatial or technological closeness, respectively. The non-inclusion of restrictions of contiguity or of distance-based patterns within the GR's represents a "spaceless" framework motivated by multilevel policy-making rather than by spatial diffusion issues. Concerning the issue of local agglomeration economies, particular features of a country appear nevertheless more explicitly after collapsing the original table:

- The metropolitan zone of Buenos Aires in GR4 of Fig. 3 is over-specialized in various sectors and suggests the presence of urban economies (large cities and economically diverse in terms of over-specialized sectors), while the GR10 of Argentina, in Fig. 2, and GR24 of Brazil, in Fig. 4, are over-specialized in only one sector and accordingly suggest the presence of localization economies.
- Moreover, in Fig. 3, it should be noticed that within GR4, 18 regions of the metropolitan zone of Buenos Aires are contiguous, displaying a clear spatial clustering that suggests an effect of specialized agglomeration; this fact comforts some of the findings of Haedo and Mouchart (2015a). This agrees with the First Law of Geography, according to Tobler (1970), saying: "Everything is related to everything else, but near things are more related than distant things." This is a desirable property due to the fact that in this spaceless framework contiguity is possibly

observable and evaluated a posteriori in the final result rather than promoted a priori in the algorithm.

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

In line with the biclustering algorithms of Bhattacharya and Cui (2017) and Orzechowski et al. (2018) based on parallel computing platform, the extension of the present algorithm to n-dimensional contingency tables is under development. Although the object of the present paper was oriented toward tables of big size and was inspired by the conclusions of Ben Saber and Elloumi (2015), another objective of future research might be to look for a boundary of the dimension of the table allowing an exhaustive algorithm and, in such a case, to evaluate the economy of time through a greedy algorithm.

Funding Michel Mouchart gratefully acknowledges financial support from IAP research network Grant No. P6/03 of the Belgian government (Belgian Science Policy). Both authors gratefully acknowledge the financial support of FOP, that promoted intercontinental cooperation.

Declarations

Conflict of interest The authors declares that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alonso-Villar O, del Río C (2013) Concentration of economic activity: an analytical framework. Reg Stud 47:756–772

Baldwin RE, Martin P (2004) Agglomeration and regional growth. In: Henderson JV, Thisse JF (eds) Handbook of urban and regional economics. Elsevier, Amsterdam

Banerjee A, Dhillon I, Ghosh J, Merugu S, Modha DS (2007) A generalized maximum entropy approach to Bregman co-clustering and matrix approximation

Ben Saber H, Elloumi M (2015) DNA microarray data analysis: a new survey on biclustering. Int J Comput Biol 4:21–37

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

Benzécri JP (1973) Analyse des Données. Dunod, París

Benzécri JP (1992) Correspondence analysis handbook. Dekker, New York

Bhattacharya A, Cui Y (2017) A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules. Sci Rep 7:4162. https://doi.org/10.1038/s41598-017-04070-4

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

Bickenbach F, Bode E (2008) Disproportionality measures of concentration, specialization and localization. Int Reg Sci Rev 31:359–388
Bickenbach F, Bode E, 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 HH (1979) Simultaneous clustering of objects and variables. In: INRIA, pp 187–203
Branson D (2000) Stirling numbers and Bell numbers: their role in combinatorics and probability. Math Sci 25:1–31
Braverman EM, Kiseleva NE, Muchnik IB, Novikov SG (1974) Linguistic approach to the problem of processing large bodies of data. Autom Remote Control 35:1768–1788
Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. Comput Oper Res 35:2964–2987
Caldas J, Kaski S (2011) Hierarchical generative biclustering for microrna expression analysis. J Comput Biol 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, Ben Ahmed M (2009) Détermination du nombre des classes dans l'algorithme croki de classification croisée. In: EGC, pp 447–448
Cheng Y, Church GM (2000) Biclustering of expression data. In: ISMB, pp 93–103
Ciampi A, González Marcos A, Castejón Limas M (2005) Correspondence analysis and two-way clustering. Stat Oper Res Trans 29:27–42
Combes P, Gobillon L (2015) The empirics of agglomeration economies. In: Duranton G, Henderson JV, Strange WC (eds) Handbook of regional and urban economics. Elsevier, Amsterdam
Corsten L, Denis J (1990) Structuring interaction in two-way tables by clustering. Biometrics 46:207–215
Cottineau C, Finance O, Hatna E, Arcaute E, Batty M (2018) Defining urban clusters to detect agglomeration economies. Environ Plan B: Urban Anal City Sci. https://doi.org/10.1177/2399808318755146
Denis JB, Vincourt P (1982) Panorama des méthodes statistiques d'analyse des interactions genotype × milieu. Agronomie 2:219–230
Donato V (2002) Políticas públicas y localización industrial en Argentina. Fundación Observatorio PyME, Buenos Aires, CIDETI working paper 2002/01
Duffy DE, Quiroz AJ (1991) A permutation-based algorithm for block clustering. J Classif 8:65–91
Duranton G, Puga D (2000) Diversity and specialization in cities: why, where and when does it matter? Urban Stud 37:533–555
Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1:211–218
Ellison G, Glaeser EL, Kerr WR (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. Am Econ Rev 100:1195–1213
Escofier B (1978) Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Rev Stat Appl 16:29–37
Florence P (1939) Report of the location of industry. Political and Economic Planning, London
Fujita M, Krugman P, Venables A (2001) The spatial economy. Cities, regions, and international trade. MIT Press, Cambridge
Fujita M, Thisse J-F (2002) Economics of agglomeration. Cities, industrial location, and regional growth. Cambridge University Press, Cambridge
Gan L, Jiang J (1999) A test for global maximum. J Am Stat Assoc 94:847–854
Gardner M (1978) The Bells: versatile numbers that can count partitions of a set, primes and even rhymes. Sci Am 238:24–30
Gilula Z (1986) Grouping and associations in contingency tables: an exploratory canonical correlation approach. J Am Stat Assoc 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. Am J Sociol 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. Ann Stat 13:10–69
Govaert G (1977) Algorithme de classification d'un tableau de contingence. In: INRIA, pp 487–500
Govaert G (1995) Simultaneous clustering of rows and columns. Control Cybern 24(4):437–458

Govaert G, Nadif M (2008) Block clustering with Bernoulli mixture models: comparison of different approaches. Comput Stat Data Anal 52:3233–3245
Govaert G, Nadif M (2010) Latent block model for contingency tables. Commun Stat Theory Methods 3:416–425
Govaert G, Nadif M (2013) Co-clustering. Wiley, Hoboken
Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic Press, London
Greenacre MJ (1988) Clustering the rows and columns of a contingency table. J Classif 5:39–51
Greenacre MJ (1993) Multivariate generalizations of correspondence analysis. In: Cuadras CM, Rao CR (eds) Multivariate analysis: future directions 2. North-Holland, Amsterdam
Greenacre MJ (2007) Correspondence analysis in practice. Chapman & Hall/CRC, Boca Raton
Greenacre MJ (2011) A simple permutation test for clusteredness. Barcelona GSE working paper 555
Guimarães P, Figueiredo O, Woodward D (2003) A tractable approach to the firm location decision problem. Rev Econ Stat 84:201–204
Guimarães P, Figueiredo O, Woodward D (2009) Dartboard tests for the location quotient. Reg Sci Urban Econ 39:360–364
Haedo C (2009) Measure of global specialization and spatial clustering for the identification of "Specialized" Agglomeration. Ph.D. thesis, Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Bologna. http://amsdottorato.cib.unibo.it/1735/1/Christian_Haedo_tesi.pdf
Haedo C, Mouchart M (2015a) Specialized agglomerations with lattice data: model and detection. Spatial Stat 11:113–131
Haedo C, Mouchart M (2015b) Methodological framework for the analysis of industrial geographical data, part of the project Mapas Industriales de América Latina y el Caribe (MIALC). Fundación Observatorio PyME, Buenos Aires. https://www.geoecon.info/slides/slide/metodologia-1
Haedo C, Mouchart M (2018) A stochastic independence approach for different measures of concentration and specialization. Pap Reg Sci 97:1151–1168
Hahsler M, Piekenbrock M, Doran D (2019) dbscan: Fast density-based clustering with R. J Stat Softw 91:1–30. https://doi.org/10.18637/jss.v091.i01
Hartigan JA (1972) Direct clustering of a data matrix. J Am Stat Assoc 67:123–129
Hausmann R, Hidalgo CA, Bustos S, Coscia M, Chung S, Jimenez J, Simoes AR, Yildirim MA (2015) Atlas of economic complexity: mapping paths to prosperity. MIT Press, Cambridge. 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 RT, Kulp D (2007) Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. BMC Bioinform 8:S5
Jambu M (1978) Classification Automatique pour l'Analyse des Données, I- Méthodes et Algorithms. Dunod, Paris
Jobson J (1992) Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer, New York
Keribin C, Brault V, Celeux G, Govaert G (2015) Estimation and selection for the latent block model on categorical data. Stat Comput 25:1201–1216
Lebart L, Mirkin BG (1993) Correspondence analysis and classification. In: Cuadras CM, Rao CR (eds) Multivariate analysis: future directions. North-Holland, Amsterdam
Lebart L, Morineau A, Warwick KH (1984) Multivariate descriptive statistical analysis. Wiley, New York
Liu H, Zou J, Ravishanker N (2018) Multiple day biclustering of high-frequency financial time series. Stat 7:e176. https://doi.org/10.1002/sta4.176
Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinf 1:24–45
Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London
Marinelli C, 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. Kluwer, Dordrecht
Moineddin R, Beyene J, Boyle E (2003) On the location quotient confidence interval. Geogr Anal 35:249–256
Nathan M, Overman H (2013) Agglomeration, clusters, and industrial policy. Oxf Rev Econ Policy 29:383–404
O'Donoghue D, Gleave B (2004) A note on methods for measuring industrial agglomeration. Reg Stud 38:419–427

Orzechowski P, Sipper S, Huang X, Moore JH (2018) EBIC: an evolutionary-based parallel biclustering algorithm for pattern discovery. Bioinformatics 34:3719–3726. https://doi.org/10.1093/bioinformatics/bty401

Park PJ, Manjourides J, Bonetti M, Paganob M (2009) A permutation test for determining significance of clusters with applications to spatial and gene expression data. Comput Stat Data Anal 53:4290–4300

Puga D (2010) The magnitude and causes of agglomeration economies. J Reg Sci 50:203–219

Rao CR (1995) A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. QÜESTIIÓ 19:23–63

Rosenthal S, Strange WC (2004) Evidence on the nature and sources of agglomeration economies. In: Henderson JV, Thisse JF (eds) Handbook of urban and regional economics. Elsevier, Amsterdam

Rota G-C (1964) The number of partitions of a set. Am Math Mon 71:498–504

Schepers J, Bock H-H, Van Mechelen I (2017) Maximal interaction two-mode clustering. J Classif 34:49–75

Sloane NJA (2001) Bell numbers. In: Hazewinkel M (ed) Encyclopedia of mathematics. Springer, New York

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

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

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240

Van Mechelen I, Bock H-H, De Boeck P (2004) Two-mode clustering methods: a structured overview. Stat Methods Med Res 13:363–394

Viladecans-Marsal E (2004) Agglomeration economies and industrial location: city-level evidence. J Econ Geogr 4:565–582

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244