TWO-MODE CLUSTERING THROUGH PROFILES OF REGIONS AND SECTORS

Christian Haedo and Michel Mouchart

DISCUSSION PAPER | 2019 / 14

UCLouvain

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

Two-mode clustering through profiles of regions and sectors*

Christian Haedo[a] and Michel Mouchart[b]

[a]*Fundación Observatorio PyME (FOP), Argentina*
[b]*Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), UCLouvain, Belgium*

June 13, 2019

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.

Keywords: relative sectorial specialization, relative regional concentration, two-mode clustering, 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. Tel.: +32 10474318; Mob.: +32 472335927; E-mail: michel.mouchart@uclouvain.be.

*Michel Mouchart gratefully acknowledges financial support from IAP research network grant $nrP6/03$ of the Belgian government (Belgian Science Policy). Both authors gratefully acknowledge the financial support of 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.

Contents

1 Introducing the topic of this paper

It is known that the economic activity is spatially concentrated and this concentration generates agglomeration economies: localization economies, that is, the benefits that firms derive from the presence of same sector firms in a geographical area, and urbanization economies, that is, the advantages that firms obtain from large (and often economically diverse in terms of sector of activities) cities. Therefore, the industrial policy objectives can be better fulfilled if they are more sensitive to the characteristics of the regions and of the sectors in design and delivery (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 extension of the sectors and of the regions is known to have a prominent role.

The New Economic Geography (NEG) models explaining specialization mainly originated in trade theory, while other models explaining concentration mainly came from location theory. These models combine the insights of traditional regional science with those of modern trade theory and thus attempt to provide an integrative approach to interregional and international structure of the economic activity (Krugman, 1998; Schmutzler, 1999; Fujita, Krugman and Venables, 2001; Fujita and Thisse, 2002). The explanations about the existence and determinants of the agglomeration economies started with the pioneer works of Marshall (1890), Scitovsky (1954), Arrow (1962), Becattini (1979), and Romer (1986), regarding the localization economies; and Jacobs (1969), Henderson (1985), Lucas (1988) and Glaeser *et al.* (1992) regarding the urban economies. This literature is extensively reviewed in Duranton and Puga (2000), Henderson (2003), Baldwin and Martin (2004), Rosenthal and Strange (2004), Viladecans-Marsal (2004), Ellison, Glaeser and Kerr (2010), Puga (2010), Combes and Gobillon (2015), Cottineau *et al.* (2018). The main purpose of this paper is to regroup simultaneously regions and sectors following structural similarities in terms of the presence of agglomeration economies.

This paper is based on lattice data in the form of a two-way contingency table, namely regions $\times$ sectors[1]. These data are mainly used for the identification and analysis of sectorial specialization of regions and of regional concentration of sectors. These basic concepts make sense only under a dual approach of regions and sectors. Using the perspective of a Stochastic Independence Approach (SIA) borrowed from the analysis of contingency tables and developed in Haedo and Mouchart (2018), the 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 *overall localization* of the economic activity. This approach allows a simultaneous treatment of sectorial specialization and of regional concentration thanks to a symmetric manipulations of rows and columns, in the present case of regions and sectors. This feature enhances a global view of the relative roles of regions and sectors on the regional structure of economic activity and the analysis of lattice data by means of a two-way contingency table appears as a most natural framework.

Several challenges are at stake, three of which should be singled out. A first one 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. Moreover, this paper extends in two directions the analysis of grouping for the evaluation and characterization of overall 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 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, we look for a summary of a large regions $\times$ sectors contingency table that keeps (almost) unchanged the measure of overall localization of the economic activity.

[1]Note that regions and sectors are qualitative unstructured variables and that the data, in the cells of the table, are frequency numbers rather than quantitative variables.

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). This family 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 (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.

There is a vast litterature on biclustering. 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, Bock and Van Mechelen (2017), proposes an algorithm of maximal interaction for two-mode clustering. The innovative contribution of the present paper is to base the clustering on measures of the degrees of specialization and of concentration and of overall localization by means of discrepancies (*i.e.* distance or divergence) among distributions. This introduces a particularly fruitful flexibility in the algorithm though the (exogenous) choice of the discrepancy[2].

Furthermore, 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), Tibshirani *et al.* (1999), Cheng and Church (2000), Tang *et al.* (2001), Ciampi, González Marcos and Castejón Lima (2005), Banerjee *et al.* (2007), Busygin, Prokopyev and Pardalos (2008), Charrad (2009), Caldas and Kaski (2011), Benabdeslem and Allab (2013), Liu, Zou and Ravishanker (2018), Orzechowski *et al.* (2018). 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 grouping elements with similar profiles.

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. 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 clustered. 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. Therefore, they are constant under spatial permutations and not distinguished for the inequality of the spatial distribution. The clusters to be elicited are of a structural nature, *i.e.* 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.

[2]In particular, Haedo and Mouchart (2018) proposes the use of one distance, Hellinger, and two divergences, Kullback-Leibler and χ^2 (Inertia).

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 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:

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

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.

Following the Stochastic Independence Approach (SIA), as developed in Haedo and Mouchart (2018), 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 evaluated.

More specifically, the relative sectorial 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)[3] 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 regional 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-specialized, or non-concentrated, economic structure. This measure represents the (global) information provided by the contingency table $\mathbf{N}$ about the overall localization[4] of the economy and allows one to analyze the contribution of each cell (i, j). This analysis may be conducted by means of the so-called Location Quotient[5] LQ_{ij}. This quotient, well-known in the literature on relative sectorial specialization and on relative regional concentration, may be written as follows:

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

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[6]:

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

When $p_{ij} = 0, LQ_{ij} = 0$, that is an indicator of maxima under-specialization. For the measures of relative sectorial specialization, of relative regional concentration and of overall localization, three different

[3] 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.

[4] For overall localization, Bickenbach and Bode use the term polarization in 2006 but localization in 2008. Haedo (2009) used global specialization. In 2010, Bickenbach, Bode and Krieger-Boden 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.

[5] 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 (2018).

[6] where "non-specialization" stands for: no sectorial specialization of region i or no regional concentration of sector j, and similarly for the others lines.

discrepancies are the object of attention, namely the χ^2-divergence or Inertia, the Kullback-Leibler divergence and the Hellinger-distance.

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, $d^2_H(\cdot \mid \cdot)$
Relative sectorial specialization of region i, $d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$	$= \sum_j \frac{(p_{j\mid i} - p_{\cdot j})^2}{p_{\cdot j}}$ $= \sum_j p_{\cdot j} \, (LQ_{ij} - 1)^2$	$= \sum_j p_{j\mid i} \log\left(\frac{p_{j\mid i}}{p_{\cdot j}}\right)$ $= \sum_j p_{\cdot j} \, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2} \sum_j (\sqrt{p_{j\mid i}} - \sqrt{p_{\cdot j}})^2$ $= \frac{1}{2} \sum_j p_{\cdot j} \, (\sqrt{LQ_{ij}} - 1)^2$
Relative regional concentration of sector j, $d(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$	$= \sum_i \frac{(p_{i\mid j} - p_{i\cdot})^2}{p_{i\cdot}}$ $= \sum_i p_{i\cdot} \, (LQ_{ij} - 1)^2$	$= \sum_i p_{i\mid j} \log\left(\frac{p_{i\mid j}}{p_{i\cdot}}\right)$ $= \sum_i p_{i\cdot} \, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2} \sum_i (\sqrt{p_{i\mid j}} - \sqrt{p_{i\cdot}})^2$ $= \frac{1}{2} \sum_i p_{i\cdot} \, (\sqrt{LQ_{ij}} - 1)^2$
Overall localization, $d([p_{ij}] \mid [p_{i\cdot} \, p_{\cdot j}])$	$= \sum_i \sum_j \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}$ $= \sum_i \sum_j \frac{p_{i\cdot}(p_{j\mid i} - p_{\cdot j})^2}{p_{\cdot j}}$ $= \sum_i \sum_j \frac{p_{\cdot j}(p_{i\mid j} - p_{i\cdot})^2}{p_{i\cdot}}$ $= \sum_i \sum_j p_{i\cdot} \, p_{\cdot j} (LQ_{ij} - 1)^2$	$= \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}\right)$ $= \sum_i \sum_j p_{i\cdot} \, p_{j\mid i} \log\left(\frac{p_{j\mid i}}{p_{\cdot j}}\right)$ $= \sum_i \sum_j p_{\cdot j} \, p_{i\mid j} \log\left(\frac{p_{i\mid j}}{p_{i\cdot}}\right)$ $= \sum_i \sum_j p_{i\cdot} \, p_{\cdot j} \, LQ_{ij} \log(LQ_{ij})$	$= \frac{1}{2} \sum_i \sum_j (\sqrt{p_{ij}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$ $= \frac{1}{2} \sum_i \sum_j (\sqrt{p_{i\cdot} \, p_{j\mid i}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$ $= \frac{1}{2} \sum_i \sum_j (\sqrt{p_{\cdot j} \, p_{i\mid j}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2$ $= \frac{1}{2} \sum_i \sum_j p_{i\cdot} \, p_{\cdot j} \, (\sqrt{LQ_{ij}} - 1)^2$

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 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 explicitly used in this paper.

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 = 5,138 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 5,138 × 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 of overall 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[7] for the $I \times J$ matrix $\mathbf{N}$ is:

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

[7]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).

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 computationally expensive. Therefore, a greedy algorithm that only ensures a local optimum is preferred.

3.2 Singular Value Decomposition and Correspondence Analysis

Let

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

be the probability matrix corresponding to $\mathbf{N}$, $\mathbf{r} = [p_{i.}]$ the vector of row marginals, $\mathbf{c} = [p_{.j}]$ 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 $\mathbf{R} = [p_{ij} - p_{i.}\, p_{.j}]$ be the matrix of residuals between an observed p_{ij} and an expected one, under an hypothesis of independence, $p_{i.}\, p_{.j}$. 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:

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

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

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

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 also be viewed as an homothetic transformation of the location quotient, by a factor $\sqrt{p_{i.}p_{.j}}$. 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, 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 $K = min(I-1, J-1)$[8], 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*. The χ^2-divergence between $[p_{ij}]$ and $[p_{i.}p_{.j}]$, or Total Inertia, may now be written as[9]:

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

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, Morineau and Warwick (1984), Greenacre (1984, 1993, 2007). More explicitly, the principal coordinates for the rows (regions) are defined as:

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

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:

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

[8] with, evidently, $\lambda_{K+1} = \cdots = \lambda_{max(I,J)} = 0$

[9] Remember that, in general, when a matrix $A = [a_{ij}]$, one has: $tr(A'A) = \sum_i\sum_j a_{ij}^2$

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 overall 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 sectorial specialization of region i and of relative regional 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 overall 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 (2018).

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:

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

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

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

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

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 (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:

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

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

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 $\mathbf{F}$, *i.e.* let the $I \times m$ matrix $\mathbf{F}_{(m)} = (f_{\vec{i}1}, \ldots, f_{\vec{i}l}, \ldots, f_{\vec{i}m})$, 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 $I+1$ 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 $\mathbf{G}$, with the $J \times m$ matrix $\mathbf{G}_{(m)} = (g_{\vec{j}1}, \ldots, g_{\vec{j}l}, \ldots, g_{\vec{j}m})$ 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 $J+1$ 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 K collapsed tables, each of dimension $(I-1)(J-1)$, and repeat the operation for each m, obtaining accordingly 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 overall 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 overall 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 overall 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

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, evidently, to let the algorithm work on real data of several countries. The purpose of these applications is to analyze the degree of overall localization using the optimal collapsed table algorithm and to evaluate how far the functioning of the algorithm provides additional information on the regional structure of the economic activity. Two of these applications are presented for different aspects of the results of the algorithm for Argentina and for Brazil.

These results are not presented with an exhaustive approach but rather with the objective of expliciting the characteristics of the proposed methodology with some insights about how these results 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.

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 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.
37	Recycling

The optimal collapsed tables of Argentina and Brazil are found using B=1,000 permutation bootstraps and the first eigenvalues selected in the step of the algorithm where the optimal collapsed tables were found: the first 15 out of 22 for Argentina and the first 15 out of 21 for Brazil.

4.1 Application for Argentina

The spatial units are the lower level political-administrative jurisdictions called departments (511) of Argentina. After eliminating those without employees, 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 the results obtained from the three measures of overall localization proposed in Table 1 for the original and for the optimal collapsed table, namely the number of cells, and the resulting loss of information about the level of overall localization.

Table 3: *Summary of the 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 (491x23)	578 (34x17)	
Reduction # of cells		94.9%	

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 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 one, in the present case, to summarize and visualize large tables by representing each cell by one color: white for cells non-specializaed 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, Figure 2, for the optimal collapsed table- thus 34 rows, corresponding to g-regions, and 17 columns, corresponding to g-sectors. In Figure 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 may provide underlying information on the process leading to these specialization and concentration 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 an ampliation 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[10]. 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%, respectively, of manufacturing occupation at national level.

4.2 Application for Brazil

The spatial units are the lower level political-administrative jurisdictions called municipalities (5,138) 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 the results obtained from the three measures of overall localization proposed in Table 1.

Table 4: *Summary of the 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 (5,138x22*)	884 (52x17)	
Reduction # of cells		99.2%	

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

It may be noticed that the (5,138 × 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 $d_{\mathcal{X}^2}$.

[10]The First Law of Geography, according to Waldo R. Tobler (1970), is: "Everything is related to everything else, but near things are more related than distant things".

Figure 1: *Heatmap of the original table for overall localization of Argentina*

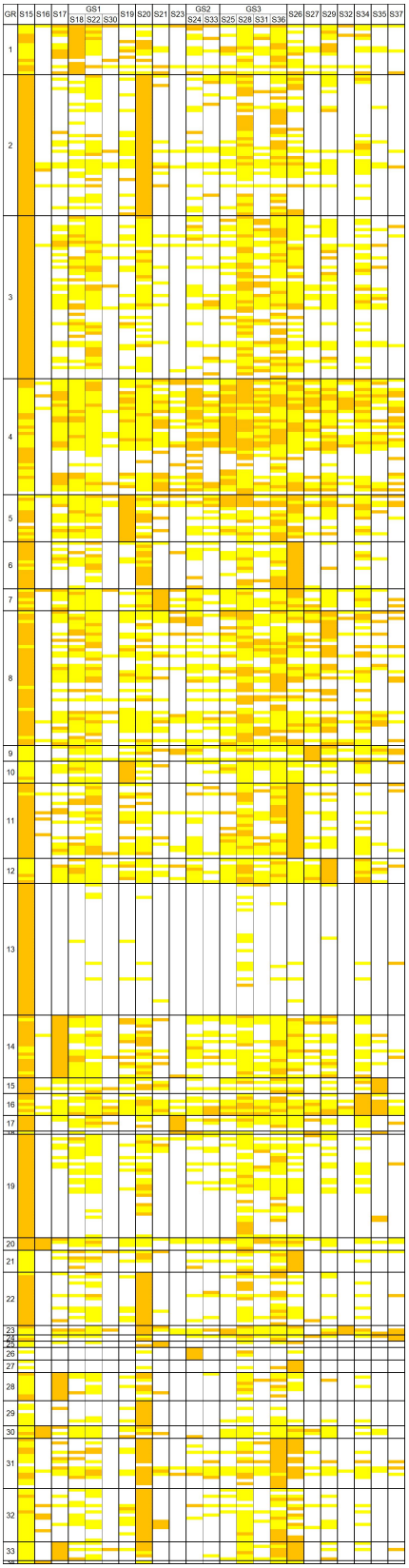


Figure 2: *Heatmap of the optimal collapsed table for overall localization of Argentina*

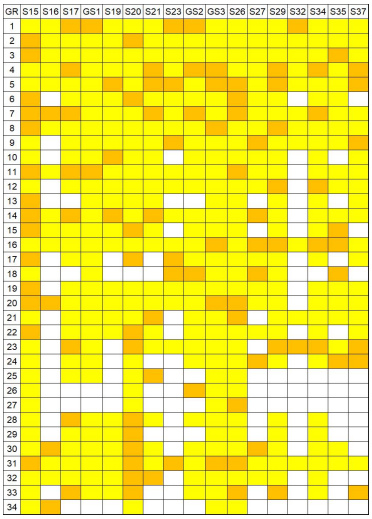
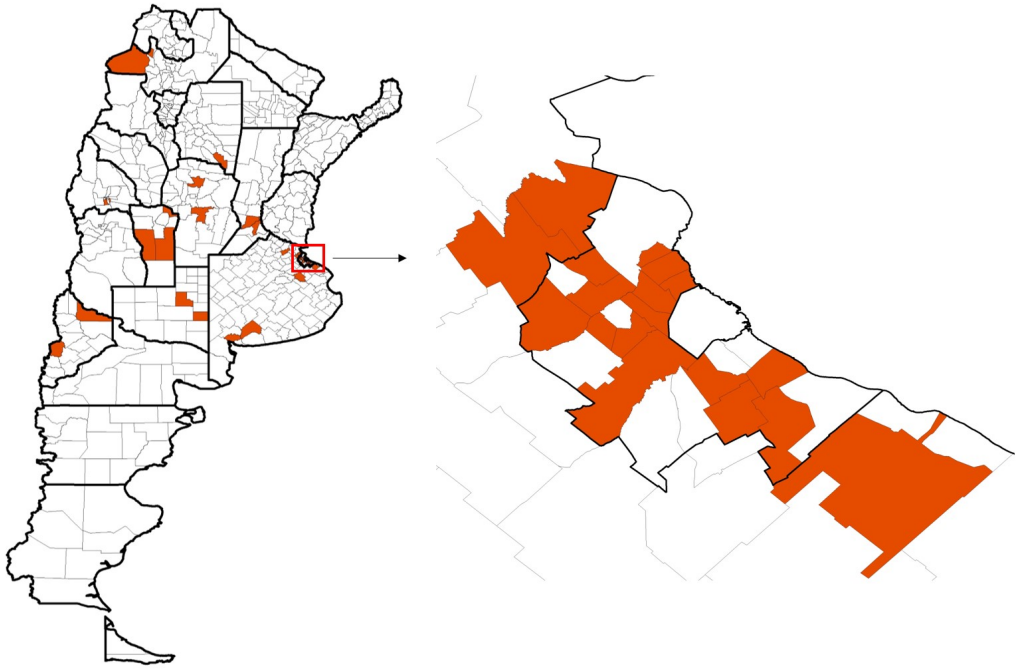


Figure 3: *GR4 of Argentina*



Similarly to Figure 2 for Argentina, Figure 4 is the heatmap of the optimal collapsed table- thus 52 rows, corresponding to GR's, and 17 columns, corresponding to GR'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 (6). 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-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 simultaneouly homogenous regions of similar relative specialization structure and homogenous sectors of similar relative concentration structure.

5 Conclusions

The results of the applications demonstrate that the innovative aspects of this algorithm are in providing a new way of understanding the similarities and dissimilarities of the processes of relative specialization and concentration between regions or sectors that are not bound to spatial or technological closeness. The connection with the regional-sectorial distribution of the employment, the presence of agglomeration economies and the tendency to co-localize employment are accordingly of a major interest.

A distinctive feature of this greedy algorithm in an unsupervised framework 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

Figure 4: *Heatmap of the optimal collapsed table for overall localization of Brazil*

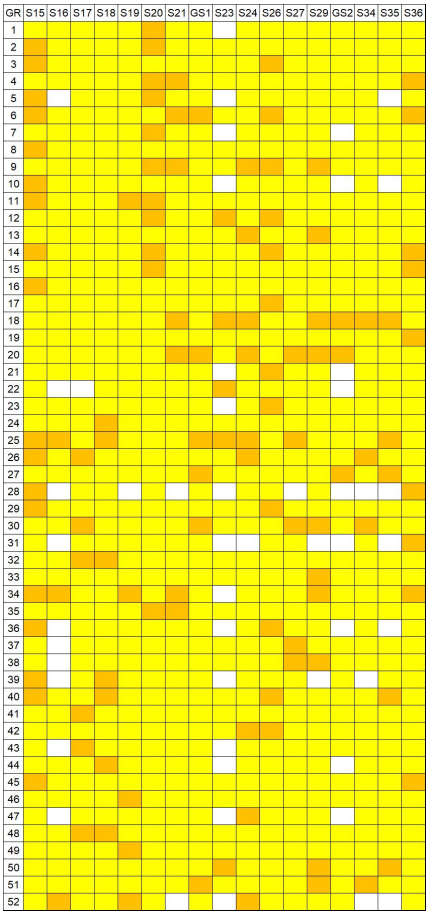


Figure 5: *Standardized residuals of the original table of Brazil*

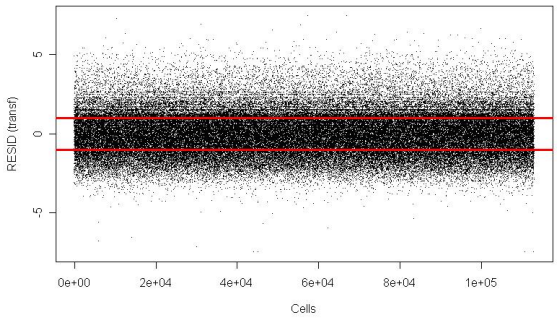
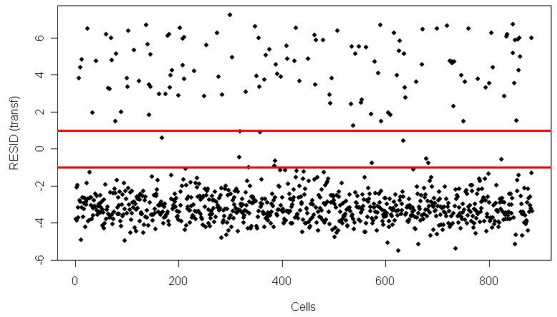


Figure 6: *Standardized residuals of the optimal collapsed table of Brazil*



resulting regrouping. Indeed 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. In this later case, such algorithms would define a family of concepts rather than a unique concept. In the present case, the concerts of GR's and of GS's are defined in terms of relative specialization and concentration respectively.

The Introduction of this paper mentions three challenges the proposed algorithm should consider:

- In this algorithm, the basic criterion for grouping regions, respectively sectors, is the similarity of the sectorial, respectively 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 provides a useful information for the policy maker either within a country or between countries. For instance, it is of a particular interest for policy cooperation to notice that the GR2 of Argentina (Figure 2) and the GR2 of Brazil (Figure 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).
- 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 overall localization, namely with a factor around 20%. This is quite an achievement regarding the *third challenge*.

Concerning the issue of local agglomeration economies, particular features of a country appear more explicitly after collapsing the original table:

- Figure 3 of the metropolitan zone of Buenos Aires, GR4, 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 and GR24 of Brazil are over-specialized in only one sector and accordingly suggest the presence of localization economies.

- The algorithm does not include any restriction of contiguity or of some distance-based pattern within the GR's. Nevertheless, in Figure 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).

Table 4 shows that, in line with the SIA, the choice of a specific metric 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.

References

Alonso-Villar, O. and del Río, C. (2013), Concentration of economic activity: an analytical framework. *Regional Studies* **47**: 756-772.

Arrow, K. (1962), The economic implications of learning by doing. *Review of Economic Studies* **29**: 155-73.

Baldwin, R.E. and Martin, P. (2004), Agglomeration and regional growth. In J.V. Henderson and J.F. Thisse (eds.), *Handbook of Urban and Regional Economics*. Amsterdam: Elsevier.

Bhattacharya, A. and Cui, Y. (2017), A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules. *Scientific Reports - Nature* **7**: 4162. doi:10.1038/s41598-017-04070-4 9

Becattini, G. (1979), Dal "settore" industriale al "distretto" industriale. Alcune considerazioni sull'unità di indagine dell'economia industriale. *Rivista di Economia e Politica Industriale* **1**: 7-21.

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.

Busygin, S., Prokopyev, O. and Pardalos, P.M. (2008), Biclustering in data mining. *Computers and Operations Research* **35**: 2964-2987.

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.

Combes, P. and Gobillon, L. (2015), The empirics of agglomeration economies. In G. Duranton, J.V. Henderson and W.C. Strange (eds.), *Handbook of Regional and Urban Economics*. Amsterdam: Elsevier.

Corsten, L. and Denis, J. (1990), Structuring Interaction in Two-Way Tables by Clustering. *Biometrics* **46**: 207-215.

Cottineau, C., Finance, O., Hatna, E., Arcaute, E. and Batty, M. (2018), Defining urban clusters to detect agglomeration economies. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/2399808318755146

Denis, J. B. and 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. Buenos Aires: Fundación Observatorio PyME, CIDETI Working Paper 2002/01.

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

Duranton, G. and Puga, D. (2000), Diversity and specialization in cities: why, where and when does it matter? *Urban Studies* **37**: 533-555.

Ellison, G., Glaeser, E.L. and Kerr, W.R. (2010), What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review* **100**: 1195-1213.

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.

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

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

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

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.

Glaeser, E., Kallal, H., Scheinkman, J. and Shleifer, A. (1992), Growth in cities. *Journal of Political Economy* **100**: 1126-1152.

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. `http://amsdottorato.cib.unibo.it/1735/1/Christian_Haedo_tesi.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, part of the project *Mapas Industriales de América Latina y el Caribe (MIALC)*. Buenos Aires: Fundación Observatorio PyME, CIDETI Working Paper 2015/08. `https://www.geoecon.io/slides/slide/metodologia-1`
Haedo, C. and Mouchart, M. (2018), A stochastic independence approach for different measures of concentration and specialization. *Papers in Regional Science* **97**: 1151-1168.
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`
Henderson, J.V. (1985), *Economic Theory and the Cities*. Orlando: Academic Press.
Henderson, J.V. (2003), Marshall's scale economies. *Journal of Urban Economics* **53**: 1-28.
Hirotsu, C. (1983), Defining the pattern of association in two-way contingency tables. *Biometrika* **70**: 579-589.
Jacobs, J. (1969), *The Economy of Cities*. London: Jonathan Cape.
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.
Krugman, P. (1998), What's new about the new economic geography? *Oxford Review of Economic Policy* **14**: 7-17.
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.

Liu, H., Zou, J. and Ravishanker, N. (2018), Multiple day biclustering of high-frequency financial time series. *Stat* **7**: e176. https://doi.org/10.1002/sta4.176

Lucas, R. (1988), On the mechanics of economic development. *Journal of Monetary Economics* **22**: 3-42.

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.

Marshall, A. (1890), *Principles of Economics*. London: Macmillan.

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.

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

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

Orzechowski, P., Sipper, S., Huang, X. and Moore, J.H. (2018), EBIC: an evolutionary-based parallel biclustering algorithm for pattern discovery. *Bioinformatics*, bty401. https://doi.org/10.1093/bioinformatics/bty401

Puga, D. (2010), The magnitude and causes of agglomeration economies. *Journal of Regional Science* **50**: 203-219.

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

Romer, P. (1986), Increasing returns and long-run growth. *Journal of Political Economy* **64**: 1002-1037.

Rosenthal, S. and Strange, W.C. (2004), Evidence on the nature and sources of agglomeration economies. In J.V. Henderson and J.F. Thisse (eds.), *Handbook of Urban and Regional Economics*. Amsterdam: Elsevier.

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

Schepers, J., Bock, H-H. and Van Mechelen, I. (2017), Maximal interaction two-mode clustering. *Journal of Classification* **34**: 49-75.

Schmutzler, A. (1999), The new economic geography. *Journal of Economic Surveys* **13**: 355-379.

Scitovsky, T. (1954), Two concepts of external economies. *Journal of Political Economy* **62**: 143-151.

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.

Tobler, W.R. (1970), A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**: 234-240.

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

Viladecans-Marsal, E. (2004), Agglomeration economies and industrial location: city-level evidence. *Journal of Economic Geography* **4**: 565-582.

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