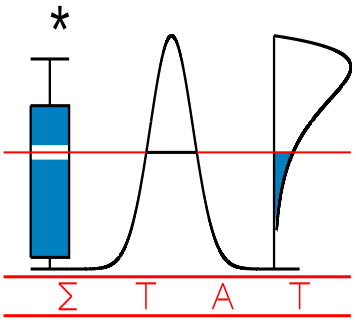
T E C H N I C A L

R E P O R T

**11008**

# A stochastic independence approach for different measures of global specialization

HAEDO, C. and M. MOUCHART




I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

# A Stochastic Independence Approach for different Measures of Global Specialization *

Christian HAEDO[a] and Michel MOUCHART[b]

[a] *University of Bologna, sede de Buenos Aires, Argentina*

[b] *ISBA, UCLouvain, Belgium*

February 28, 2011

**Abstract**

Based on data in the form of a two-way contingency table "Regions × Activities", the concepts of specialization and of concentration are naturally based on the analysis of the conditional distributions, or profiles. The natural tool for measuring the degrees of specializations are provided by discrepancies, more precisely distances or divergences, among distributions: between profiles and a uniform distribution for absolute concepts, between profiles and the corresponding marginal distribution for the relative concepts or between the joint distribution and the product of the marginal distributions for the global concept. This is the approach of stochastic independence that conducts the analysis in terms of stochastic independence between activities and regions and the global discrepancy is viewed as a measure of row-column association. This paper presents the results of an extensive analysis of the numerical values of measures derived from this approach and from other approaches widely used in the literature. A main conclusion of this analysis is that although the different measures under consideration display rather similar numerical behavior, differences of ranking about the degree of specialization among activities, among regions or among countries call for a particular care when interpreting the numerical results.

*Keywords:* absolute, relative and global specialization, industrial concentration, stochastic independence.

*Corresponding Author*: Michel MOUCHART, Institut de statistique, biostatistique et sciences actuarielles (ISBA), 20 voie du Roman Pays, B-1348 Louvain-la-Neuve(B).

E-mail: michel.mouchart@uclouvain.be

*Michel Mouchart gratefully acknowledges financial support from IAP research network grant $nrP6/03$ of the Belgian government (Belgian Science Policy). Dominique Peeters provided a wealth of insightful comments at several stages of elaboration of this work and deserves a deep gratitude from the authors.

# Contents

# 1 Introduction

For a given country, let us consider regions labeled $i \in \mathcal{I} = \{1, ..., I\}$, and activities labeled $j \in \mathcal{J} = \{1, ..., J\}$. For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, we observe the number of employees, let $N_{ij}$. Thus we obtain a two-way $I \times J$ contingency table $\mathbf{N} = [N_{ij}]$ that in turn also produces row, column and table totals:

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

Two types of issues are considered in this paper, namely the concentration of the activities within regions and the specialization of the regions in term of the activities. Thus, the contingency table $\mathbf{N} = [N_{ij}]$ is to be analyzed in terms of profiles, or relative frequencies, characterizing regions and activities, namely:

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

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

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

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

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

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

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

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

Three types of analysis should be distinguished:

- analysis of the spread of the activity specific ($p_{\vec{i}|j}$) and of the region specific ($p_{\vec{j}|i}$) profiles. For categorical variables, the spread of the frequency distribution may be characterized, among others, through entropy or through average absolute deviations. Note that the entropy may be viewed as a divergence (for more details on $f$-divergence: see Csiszár 1967) with respect to the uniform distribution;

- analysis of the divergence or distance between the activity specific profile and the marginal (or country) region profiles, $d(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ or between the region specific profile and the marginal activity profile, $d(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$. Here $\chi^2$ and Kullback-Leibler divergences or Hellinger distance are used as tools for evaluating a discrepancy between two distributions;

- a global analysis of the country in terms of a divergence, or a distance, between the actual distribution and the closest distribution reflecting independence, namely $p_{i\cdot}\, p_{\cdot j}$, *i.e.* the global analysis focuses the attention on $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$, where $p_{ij} = N_{ij}/N_{\cdot\cdot}$

The economic geography literature on the industrial concentration of activities and on regional specialization is vast and the use of words is not completely standardized. In Table 1 we adopt some conventions as close as possible to largely spread uses.

Table 1: *Some conventional definitions*

| **Technique** | **Measured concept** |
|---|---|
| Spread of $p_{\vec{j}\mid i}$ | Regional specialization |
| Spread of $p_{\vec{i}\mid j}$ | Localization or industrial concentration |
| $d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$ | Relative regional specialization |
| $d(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | Relative localization or relative industrial concentration |
| $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$ | Global specialization |

Heuristically, regional *specialization*, is a feature of the activities distribution in a region ($p_{\vec{j}\mid i}$), and a region is said to be specialized if a few activities have a large share. This may be the case, for instance, when an activity is considerably larger than other ones. *Relative regional specialization* of a specific region shows up when an area has a greater proportion of a particular activity than the proportion of that activity in the whole territory. In other words, relative regional specialization compares an area share of a particular activity with the activity share at the country level, and is accordingly measured through a discrepancy $d(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$, thus relatively to the marginal distribution $p_{\cdot\vec{j}}$. The same presentation can also be made for industrial concentration and for relative industrial concentration.

In order to introduce the concept of *global specialization*, imagine the following (artificial) experiment. Draw randomly one employee from the $N_{\cdot\cdot}$ ones in activity and classify the drawn employee into region and activity. The probability of drawing an employee from the cell $(i, j)$ is evidently $p_{ij}$. In this framework, the absence of global specialization may be approached as that of stochastic independence between the row and the column criteria. This suggests to measure the degree of global specialization through a statistic that might be used for testing independence in a contingency table. Thus, global specialization may be viewed as a particular form of association between the region and the activity variables. This approach may be called "*a stochastic independence model approach*". This is the object of this paper.

We focus attention on the global specialization within a discrete space, *i.e.* a space partitioned into a finite number of regions. In the framework of the contingency table $\mathbf{N} = [N_{ij}]$, the label $i$

of the regions is arbitrary and reflects neither spatial contiguity nor distance among regions. In a sense, this analysis is "spaceless" and motivated by policy making rather than by spatial diffusion issues. When the country is treated as a unique continuous space, the basic data refer to points in the country and the interest is focused on designing a stochastic process, such as a marked point process, in order to represent locally diffusion issues. With this last approach, motivation is more oriented toward modeling and explaining the observed spatial structure. The continuous approach is not developed in this paper.

The object of this paper is to compare numerically measures of relative and global specialization, some in the framework of the stochastic independence model, others in different frameworks. The underlying question of these comparisons is to evaluate how far these measures are mutually coherent and quantify a same concept. We also check whether these measures operate a same ranking of specializations among activities, among regions or among countries. A heuristic conclusion of our analysis is that the investigated measures are reasonably coherent but that the possible differences of ranking require some care at the stage of interpretation.

The order of presentation is a follows. We introduce global measures of specialization in next Section. The general presentation is followed by to numerical applications, one on argentinean data and another one on a comparison between Argentina, Brazil and Chile. Section 3 consider the issue of grouping regions or activities in the framework of Modifiable Areal Unit Problem (MAUP). This Section is completed with a numerical application that analyzes the grouping of argentinean regions and activities. Section 4 presents some concluding remarks about the stochastic independence approach and the other approaches, and conclusions.

# 2 Global measures of specialization

## 2.1 An overview

When defining degrees of global specialization, one possible strategy consists in first defining a regional index, characteristic of a region, and thereafter aggregate the regional indices into a global one, characteristic of the country. Conversely, one may start by first defining a global index of the country and thereafter decomposing it into regional components. This distinction is useful when trying to give structure to a set of measures of global specialization. Note also that from a stochastic independence point of view, which is essentially symmetric, the role of the regions and of the activities may be permuted, leading by so-doing to different classes of problems of interest.

The well established *Local Quotient* is defined for each cell $(i, j)$ as follows:

$$LQ_{ij} = \frac{N_{ij}/N_{i\cdot}}{N_{\cdot j}/N_{\cdot\cdot}} = \frac{N_{ij}/N_{\cdot j}}{N_{i\cdot}/N_{\cdot\cdot}} = \frac{N_{ij}N_{\cdot\cdot}}{N_{i\cdot}N_{\cdot j}} = \frac{p_{ij}}{p_{i\cdot}\ p_{\cdot j}} = \frac{p_{j|i}}{p_{\cdot j}} = \frac{p_{i|j}}{p_{i\cdot}} \tag{6}$$

The second and the third terms of (6) correspond to "relative risk" or "excess risk" in epidemiology, while the fourth term corresponds to the usual "cross-product ratio" of the $2\times2$ sub-table constructed around $N_{ij}$. The last three terms express the same concepts through proportions, *i.e.* independently of $N_{..}$ representing the size of the country. This local quotient reveals the following feature of activity $j$ in region $i$:

$$\begin{aligned} LQ_{ij} &= 1 \quad \text{no specialization} \\ &> 1 \quad \text{over-specialization} \\ &< 1 \quad \text{under-specialization} \end{aligned} \tag{7}$$

where the "no-specialization" corresponds to the row-column independence in the contingency table $\mathbf{N} = [N_{ij}]$. The last two equalities in (6) stresses the point that the specialization is an issue concerning the global structure at the country level: thus the absence of specialization of a cell $(i, j)$ means that, relatively to the distribution in the country, activity $j$ is not over-(nor under-) represented in region $i$ *and* that region $i$ is not over-(nor under-) represented for the activity $j$. Thus, "local" points to the fact that $LQ$ is localized in a cell $(i, j)$.

Among the most often used measures of independence between rows and columns of the contingency table $\mathbf{N}$, to be used as measures of global specialization, we shall focus the attention on the following three:

$$\begin{aligned} d_{\chi^2}(\mathbf{N}) &= \sum_i \sum_j p_{i\cdot} p_{\cdot j} (LQ_{ij} - 1)^2 \qquad \chi^2 - \text{divergence, or inertia} \qquad (8) \\ &= \sum_i \sum_j \frac{p_{i\cdot}(p_{j|i} - p_{\cdot j})^2}{p_{\cdot j}} = \sum_i \sum_j \frac{p_{\cdot j}(p_{i|j} - p_{i\cdot})^2}{p_{i\cdot}} \\ d_{KL}(\mathbf{N}) &= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, LQ_{ij} \log(LQ_{ij}) \qquad \text{Kullback-Leibler divergence} \qquad (9) \\ &= \sum_i \sum_j p_{i\cdot}\, p_{j|i} \log\left(\frac{p_{j|i}}{p_{\cdot j}}\right) = \sum_i \sum_j p_{\cdot j}\, p_{i|j} \log\left(\frac{p_{i|j}}{p_{i\cdot}}\right) \\ d_H(\mathbf{N}) &= \frac{1}{2} \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, (\sqrt{LQ_{ij}} - 1)^2 \qquad \text{Hellinger-distance} \qquad (10) \\ &= \frac{1}{2} \sum_i \sum_j (\sqrt{p_{i\cdot}\, p_{j|i}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2 = \frac{1}{2} \sum_i \sum_j (\sqrt{p_{\cdot j}\, p_{i|j}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2 \end{aligned}$$

These measures deserve some comments:

- As should be expected, these formulas display symmetry between regions and activities, as is the concept of stochastic independence.
- Among different equivalent forms, displaying the role of the local quotients $LQ_{ij}$ has been first privileged. These measures may therefore be viewed as a global measure of the discrepancy between the different $LQ_{ij}$ and 1. Alternatively, these global measures may also be viewed

as distances or divergences between the actual distribution embodied in the contingency table $\mathbf{N}$ and the closest distribution reflecting stochastic independence, namely the product of the marginal distributions, as is typically done when testing for stochastic independence. Equivalently they aggregate distances or divergences between the different conditional distributions and the corresponding marginal distributions.

- Hellinger's measure is symmetric between the two measures, corresponding to the actual contingency table and to the relative independent measure, and is accordingly a distance whereas the $\chi^2$ and $KL$ measures are (non symmetric) f-divergences.

- Each measure is given in the form of a double sum and may accordingly be decomposed as an average of the distances, or divergences, between the conditional distributions and the corresponding marginal distributions, relatively either to the regions or to the activities. More specifically we have:

$$d_{\chi^2}(\mathbf{N}) = \sum_i p_{i\cdot} \left[ \sum_j \frac{(p_{j|i} - p_{\cdot j})^2}{p_{\cdot j}} \right] = \sum_j p_{\cdot j} \left[ \sum_i \frac{(p_{i|j} - p_{i\cdot})^2}{p_{i\cdot}} \right] \tag{11}$$

$$d_{KL}(\mathbf{N}) = \sum_i p_{i\cdot} \left[ \sum_j p_{j|i} \log \left( \frac{p_{j|i}}{p_{\cdot j}} \right) \right] = \sum_j p_{\cdot j} \left[ \sum_i p_{i|j} \log \left( \frac{p_{i|j}}{p_{i\cdot}} \right) \right] \tag{12}$$

$$d_H(\mathbf{N}) = \frac{1}{2} \sum_i p_{i\cdot} \left[ \sum_j (\sqrt{p_{j|i}} - \sqrt{p_{\cdot j}})^2 \right] = \frac{1}{2} \sum_j p_{\cdot j} \left[ \sum_i (\sqrt{p_{i|j}} - \sqrt{p_{i\cdot}})^2 \right] \tag{13}$$

Thus these three measures of specialization accept a similar decomposition:

$$d_\omega(\mathbf{N}) = \sum_i p_{i\cdot}\ d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) = \sum_j p_{\cdot j}\ d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \qquad \omega \in \{\chi^2, KL, H\} \tag{14}$$

In other words, each of these global measures appears as an average of the relative regional specializations $d(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$, or of the relative localizations $d(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$.

*Note.* We use a slightly incoherent notation: $d_\omega(\mathbf{N})$ is a short-hand notation for $d([p_{ij}], [p_{i\cdot}p_{\cdot j}])$ that does not make explicit the two distributions $[p_{ij}]$ and $[p_{i\cdot}p_{\cdot j}]$ conforming the divergence, whereas for instance in $d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$ we make the relevant distributions explicit.

- In the literature, the region specific indexes $d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$ have also been called the relative specialization of region $i$ whereas the activity specific indexes $d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ have also been called relative concentration of activity $j$. Thus (14) tells us that for the three measures, the properly weighted average of the regional specialization or of the industrial concentration provide a same measure of global specialization. This feature would induce Bickenbach and Bode to consider the global measure of specialization $d_\omega(\mathbf{N})$ as measures of polarization in (2006) or of localization en (2008 and 2010).

- Bollen and Long (1993) summarize a number of desirable properties for such measures but recognize that no single measure meets them all, moreover not all researchers would even agree with all these properties.

- This paper focuses the attention on descriptive measures of global specialization but does not consider issues concerning sampling or asymptotic properties in view of an eventual statistical inference.

The literature on economic geography has proposed of wealth of measures of industrial concentration, of regional specialization or of global specialization, often not in the present framework of stochastic independence. A large class of these proposals are based on Lorenz curves and Gini indices. In Appendix A, details are given on a Gini index of relative regional specialization $GI_i$ and of relative industrial concentration $GI^j$. In Appendix A are also given details on another class of indices based on absolute deviations and due to Krugman. They provide other indices of relative regional specialization $SK_i$ or of relative industrial concentration $SK^j$. These indices may also be aggregated into measures of global specialization by means of weighted averages, either on the relative regional specialization:

$$GI_{reg} = \sum_i p_{i\cdot}\, GI_i \tag{15}$$

$$SK_{reg} = \sum_i p_{i\cdot}\, SK_i \tag{16}$$

or on the relative industrial concentration

$$GI^{act} = \sum_j p_{\cdot j}\, GI^j \tag{17}$$

$$SK^{act} = \sum_j p_{\cdot j}\, SK^j \tag{18}$$

Because these measures are not developed in the symmetric framework of stochastic independence, the global measures based on regions and activities do not coincide:

$$GI_{reg} \neq GI^{act} \tag{19}$$

$$SK_{reg} \neq SK^{act} \tag{20}$$

## 2.2 Application to argentinean data

### Scope of this application

In order to better understand which aspects of global specialization are captured by each of the three measures, we conduct a diversified investigation of the numerical behaviour of these measures evaluated in a specific case.

We want to examine different issues. Firstly, when considering the profiles of the activities, or of the regions, relatively to their corresponding marginal (country-wide) distribution, how much are associated the measures of relative concentration, or of relative specialization? This question may be answered by a graphic representation of these measures or by evaluating correlations among them. But this question also raises another one. These measures are subject to different ranges of variation: the unit interval for $d_H$ or bounded intervals for $d_{\chi^2}$ or $d_{KL}$. A comparison of their behaviour is therefore easier if they are transformed into measures with similar, or identical, range of variation. Some transformations are considered but a uniform standardization, to the unit interval for instance, is not feasible because their maximum values depend on the dimensions of the table, $I$ and $J$, or on extreme values. A graphic representation of these measures, along with some of their transformations, reveals sometime linear sometime non-linear associations.

Observing, and hopefully explaining, these differences of behaviour is one way for better interpreting these measures. Other issues are: when considering the different measures of relative concentration or of relative specialization, do these measures provide a same ordering of the activities, or of regions? When evaluating the global degree of specialization for different countries, is the ordering the same for each measure?

It should stressed but these issues basically regard the interpretation of the numerical values of these measures and their comparability among different activities, different regions or different countries. Moreover, we also want to compare the numerical behaviour of $d_H$, $d_{\chi^2}$ and $d_{KL}$ with that of Gini and of Krugman indices.

**The data**

The original data concerns the employees in the manufacturing sector and are obtained from of the Economic Census made by the National Institute of Statistic and Censuses of Argentina (INDEC-1994: 1,083,928 employees). The spatial units or regions are the political-administrative jurisdictions called departments (462 out of 523 after eliminating those without employees in the manufacturing sector).

The activity classifications refers to the first 2 digits of the International Standard Industrial Classification (ISIC Rev.3.1) of manufacturing activities (`http://unstats.un.org/unsd/cr/registry/regcs.asp?Cl=17&Lg=1&Co=D`). They are 22 activities after grouping the divisions 36 (Manufacture of furniture; manufacturing n.e.c.) and 37 (Recycling).

The final data used in this application are obtained by regrouping the 22 activities into 17 and the 462 regions into 35. The regrouping was made from an automatic grouping procedure on large two-way contingency tables based on hierarchical clustering and correspondence analysis (HCCA), aimed to obtain a "Best Collapsed Table" with low level of information loss vis-à-vis the degree of specialization in the original data (see more in Haedo 2009).

**Findings**

Appendix B shows the $35 \times 17$ contingency table $\mathbf{N}$ of the data along with the rows and columns totals $N_{i.}$, $N_{.j}$ with their proportions; we complete the table by providing the regions and activities measures of relative regional specialization $d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$, and of relative industrial concentration $d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$, and conclude the table by the global measures of specialization $d_\omega(\mathbf{N})$, where $\omega \in \{\chi^2, KL, H\}$ (Tables 9 and 10).

Let us look at the numerical values of the three measures of global specialization:

$$d_{\chi^2}(\mathbf{N}) = 1.6532; \qquad d_{KL}(\mathbf{N}) = 0.3176; \qquad d_H(\mathbf{N}) = 0.0713. \tag{21}$$

As they are measured on different scales, their numerical values are difficult to interpret except $d_H$ that takes values in the unit interval. Thus, only the numerical value of $d_H$ can be compared with Gini's and Krugman's coefficients that are also valued in the unit interval.

We obtain:

$$GI_{reg} = 0.3262; \qquad GI^{act} = 0.3495; \qquad SK_{reg} = 0.2963; \qquad SK^{act} = 0.3041. \tag{22}$$

As will be confirmed in the sequel, $d_H$ systematically gives a lower value of global specialization. Also, the numerical value of the region based and of the activity based are different by very close. Moreover, Gini's and Krugman's coefficients also take different but neighbouring values.

In order to compare the numerical values of all indices, a possible solution could be: take a statistical view to evaluate the asymptotic distribution (or an approximation of the small sample distribution by means of a resampling procedure) and compute the critical alpha corresponding to a test of independence. Each would have a same asymptotic, or approximate, distribution uniform on [0 1]. Take $1 -$ *critical alpha* as a comparable measure of association.

We do not take this way because is not appropriate for the later developments and rather take alternative ones. Gibbs and Su (2002) and Reiss (1989) have proposed the following transformations: $\log(1 + d_{\chi^2})$ and $4d_H$, respectively, in order to provide them with a range approximately close to that of $d_{KL}$. The transformed measures become

$$\log(1 + d_{\chi^2}(\mathbf{N})) = 0.4238; \qquad d_{KL}(\mathbf{N}) = 0.3176; \qquad 4d_H(\mathbf{N}) = 0.2852. \tag{23}$$

These transformed measures have close but not identical values and suggest a low level of specialization in Argentina, in view of the value of $d_H$. Later, in next subsection 2.3, we discuss the relative position of Argentina with respect to other countries. The transformation (23) ensures a similar range, namely around the interval [0 4], for the three measures; but this interval is approximately true only. In particular, it is known that the maximum value of $d_{\chi^2}$ depend on both $I$ and $J$. Cramer (1946) shows that the maximum possible for $d_{\chi^2}$ is $min\{I-1, J-1\}$ and may be obtained only if $I = J$; this issue motivated the proposition of Cramer, namely $Cd_{\chi^2} = \frac{d_{\chi^2}}{min(I-1,J-1)}$

when proposing measures of association in contingency tables; for more information see for instance, Bishop, Fienberg and Holland (1975), Everitt (1977) or Agresti (2002). Another difficulty is that there is no such range for $d_{KL}$. A simple, but not totally satisfactory proposal consist in normalizing $d_{\chi^2}$ and $d_{KL}$ to the interval [0 1], just as $d_H$. In principle, any strictly increasing function $\mathbf{R}_+ \to [0\ 1]$ may do the job but the simplest one might be:

$$Nd_{\chi^2} = \frac{d_{\chi^2}}{d_{\chi^2}+1}; \qquad Nd_{KL} = \frac{d_{KL}}{d_{KL}+1}. \tag{24}$$

The results, namely $Nd_{\chi^2}(\mathbf{N}) = 0.6231$ and $Nd_{KL}(\mathbf{N}) = 0.2410$ suggest that the transformations (24) are not satisfactory for making the values of $d_{\chi^2}$, $d_{KL}$ and $d_H$ easily comparable.

Let us now have a closer look at the decomposition of the global measure into activity specific and region specific measures according to (16), as given in Tables 9 and 10. In Figure 1, respectively Figure 2, we have ranked the 17 activities, respectively the 35 regions, in ascending order of $d_H$ and plotted together the three transformed measures.

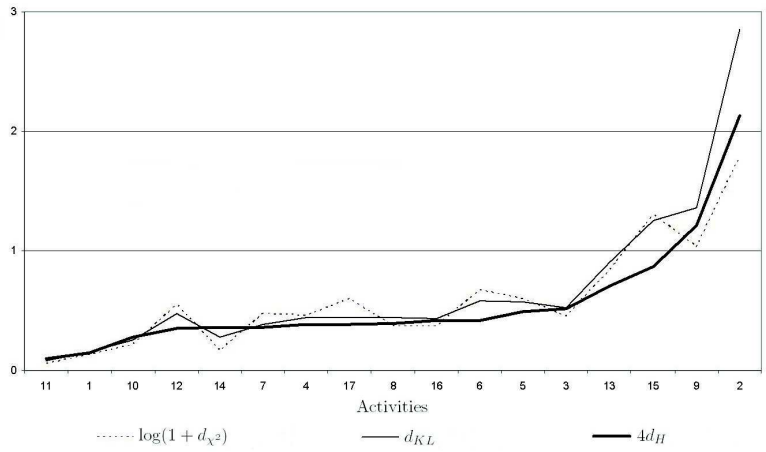



Figure 1: *Level of relative industrial concentration ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$) of transformed measures*

Two features should be noticed:

- the numerical value of the three modified measures display low dispersion for values under 1 but higher dispersion otherwise, for the region specific as well for the activity relative measures;
- the ranking between regions, or activities, is modified each time one of the curves display a descending piece; clearly the three rankings are similar but some discrepancies are noticeable. Notice that these discrepancies show up for low as well as for high values of the measures. Later on, we come back to the issue of the stability of the ranking.

Let us have a look on the graphic behaviour of the normalized measures $Nd_{\chi^2}$ and $Nd_{KL}$ compared with $GI$ and $SK$, relatively to $d_H$, in Figure 3 for the relative industrial concentration

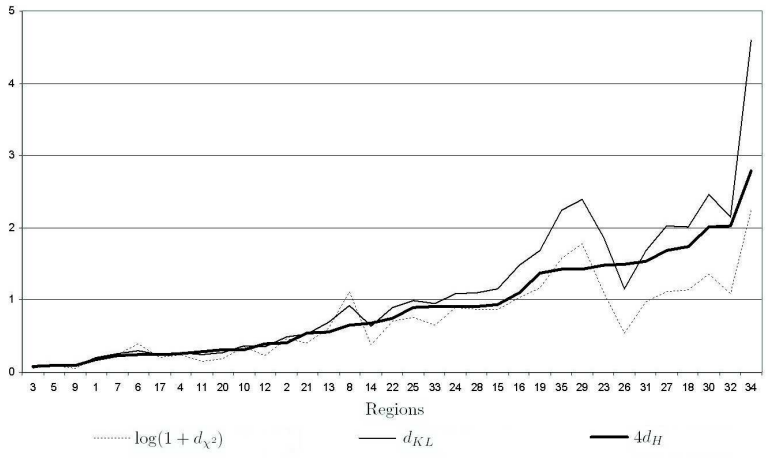



Figure 2: *Level of relative regional specialization* ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$) *of transformed measures*

and in Figure 4 for the relative regional specialization. All these curves take values in the unit interval. These figures correspond to Figures 1 and 2 that were concerned with the three transformed measures. We now notice that the five measures have roughly a similar behaviour although $Nd_{\chi^2}$ is the least similar. Moreover, the curves relative to the industrial concentrations, in Figure 3, display more coherence than those relative to regional specializations in Figure 4.

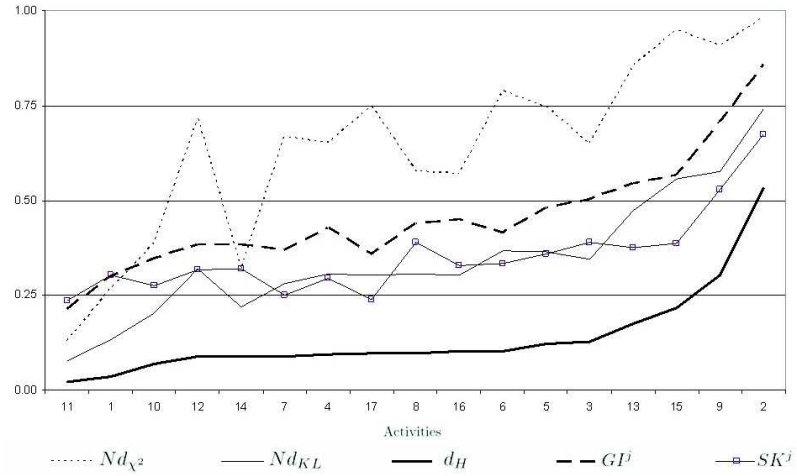



Figure 3: *Level of relative industrial concentration* ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$) *of normalized measures*

In order to get a deeper insight into the meaning of these measures, we examine the joint behaviour of 8 measures: the first 3 measures ($d_{\chi^2}$, $d_{KL}$ and $d_H$), the transformed $log(1+d_{\chi^2})$, the normalized version $Nd_{\chi^2}$ and $Nd_{KL}$, and the Gini and Krugman coefficient $GI$ and $SK$. We first examine their numerical values by means of (pairwise) correlations (Table 2) and of pairwise scatter diagrams (Figure 5). Next we perform a similar analysis on the ranks in Table 3 and 6. These tables

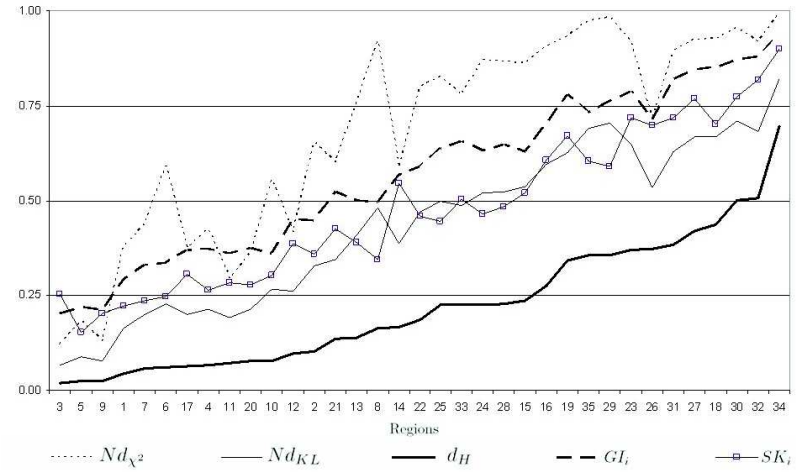



Figure 4: *Level of relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$) of normalized measures*

and figures provide the results on regional specialization under the main diagonal and the results on industrial concentration above the main diagonal.

Table 2: *Correlations between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$-above the main diagonal) measures (I=35, J=17)*

| **Item** | $p_{\vec{i}.}$ | $p_{.\vec{j}}$ | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ | $log(1+d_{\chi^2})$ | $Nd_{\chi^2}$ | $Nd_{KL}$ | $GI^j$ | $SK^j$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $p_{\vec{i}.}$ | - | - | - | - | - | - | - | - | - | - |
| $p_{.\vec{j}}$ | - | - | $-.293^3$ | $-.437^3$ | $-.460^3$ | $-.546^2$ | $-.748^1$ | $-.646^1$ | $-.569^2$ | $-.367^3$ |
| $d_{\chi^2}$ | $-.179^3$ | - | - | $.955^1$ | $.930^1$ | $.867^1$ | $.548^2$ | $.789^1$ | $.801^1$ | $.828^1$ |
| $d_{KL}$ | $-.406^2$ | - | $.810^1$ | - | $.992^1$ | $.947^1$ | $.714^1$ | $.924^1$ | $.931^1$ | $.917^1$ |
| $d_H$ | $-.455^1$ | - | $.649^1$ | $.960^1$ | - | $.920^1$ | $.696^1$ | $.919^1$ | $.952^1$ | $.938^1$ |
| $log(1+d_{\chi^2})$ | $-.460^1$ | - | $.741^1$ | $.953^1$ | $.889^1$ | - | $.867^1$ | $.968^1$ | $.897^1$ | $.810^1$ |
| $Nd_{\chi^2}$ | $-.636^1$ | - | $.407^2$ | $.784^1$ | $.813^1$ | $.877^1$ | - | $.905^1$ | $.786^1$ | $.608^1$ |
| $Nd_{KL}$ | $-.574^1$ | - | $.527^1$ | $.902^1$ | $.936^1$ | $.928^1$ | $.956^1$ | - | $.960^1$ | $.862^1$ |
| $GI_i$ | $-.575^1$ | - | $.480^1$ | $.882^1$ | $.952^1$ | $.863^1$ | $.907^1$ | $.980^1$ | - | $.946^1$ |
| $SK_i$ | $-.450^1$ | - | $.510^1$ | $.888^1$ | $.967^1$ | $.817^1$ | $.818^1$ | $.935^1$ | $.972^1$ | - |

[1]Significant at level 0.01 (two-sided)

[2]Significant at level 0.05 (two-sided)

[3]Not significant

Each time we also consider the association with the relevant marginal profiles $p_{i.}$ (first column) and $p_{.j}$ (first row) and notice a systematic negative association between the marginal profiles and the relative measures. Both the Table 2 and the Figure 5 however show that in absolute values their association is the weakest one for $d_{\chi^2}$ but the strongest one for $Nd_{\chi^2}$. This systematically negative association shows that smaller sectors or smaller regions are expected to be relatively more

specialized, as an effect of size. The scatter diagrams and the absolute values of the correlation show however that their association is globally weak, in particular because the largest regions and the largest sectors are essentially outlying data for this association.

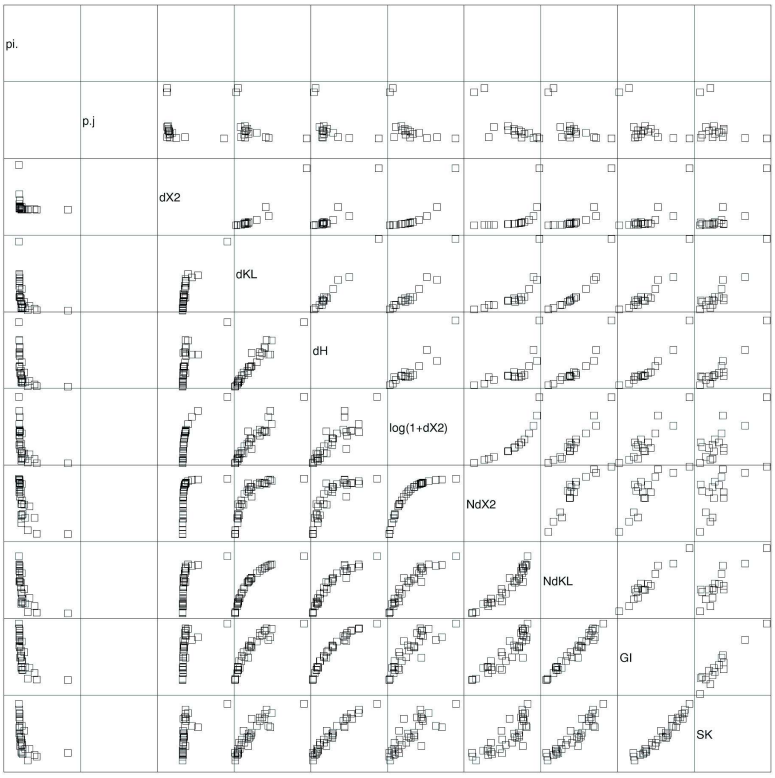



Figure 5: *Dispersion between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.}$)-above the main diagonal) measures (I=35, J=17)*

Let us now turn to the associations among the 8 measures. All pairwise correlations are positive and significantly high. There is no clear indication that the transformed version $log(1 + d_{\chi^2})$ or the normalized version $Nd_{\chi^2}$ or $Nd_{KL}$ tend to substantially increase those correlations but some are surprisingly high: most with $d_H$ and with $Nd_{KL}$, particulary between $d_H$ and $d_{KL}$, and also between $GI$ and $d_{KL}$.

Notice also that the correlations among the measures of relative industrial concentration behave in an essentially similar way as those of relative regional specializations. The scatter diagrams, in Figure 5, show however that most of these associations are non-linear, calling for more care when interpreting coefficients of linear association. But the linearity of the relationships of $SK$ with $d_H$, $Nd_{KL}$ and $GI$, and of $Nd_{\chi^2}$ with $GI$ is noteworthy.

Table 3: *Ranking correlations between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$-above the main diagonal) measures (I=35, J=17)*

| **Item** | $p_{\vec{i}.}$ | $p_{.\vec{j}}$ | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ | $GI^j$ | $SK^j$ |
|---|---|---|---|---|---|---|---|
| $p_{\vec{i}.}$ | - | - | - | - | - | - | - |
| $p_{.\vec{j}}$ | - | - | $-.755^1$ | $-.618^1$ | $-.608^1$ | $-.512^2$ | $-.373^3$ |
| $d_{\chi^2}$ | $-.795^1$ | - | - | $.917^1$ | $.824^1$ | $.718^1$ | $.554^2$ |
| $d_{KL}$ | $-.846^1$ | - | $.959^1$ | - | $.917^1$ | $.887^1$ | $.789^1$ |
| $d_H$ | $-.851^1$ | - | $.854^1$ | $.972^1$ | - | $.951^1$ | $.838^1$ |
| $GI_i$ | $-.861^1$ | - | $.893^1$ | $.964^1$ | $.987^1$ | - | $.895^1$ |
| $SK_i$ | $-.786^1$ | - | $.842^1$ | $.938^1$ | $.977^1$ | $.975^1$ | - |

[1]Significant at level 0.01 (two-sided)

[2]Significant at level 0.05 (two-sided)

[3]Not significant

A last aspect should also be checked, namely the stability of the ranking. This aspect may be viewed as non-parametric approach (see also Slottje 1990). This is examined in Table 3 by means of Spearman's rank coefficient and in Figure 6 by means of scatter diagrams among ranks. Here, the rows and columns relative to $log(1 + d_{\chi^2})$ are redundant with those relative to $d_{\chi^2}$. The same redundancy is also true for normalized versions of $d_{\chi^2}$ and $d_{KL}$. Again the correlations, in Table 3, are uniformly high and the correlations with respect to the marginal profiles are higher than in Table 2. But now the behaviour among the ranks relative to the activities (above the main diagonal) have less associations than those relative to the regions (under the main diagonal), comforting what was previously noticed.

As a first conclusion, the high rank correlation among all the measures considered so far comfort the overall coherence of these measures but the possible modifications among the ranking should be considered as a signal that these measures should be interpreted with care and, in no cases, should be viewed as objective and definitive measures of specialization. Finally, some peculiarities of $d_{\chi^2}$ might be attributed to the fact that $d_{\chi^2}$ is based on squared differences that tend to overweight extreme cases and this feature is sweetened by the log transformation.

## 2.3 Comparison between Argentina, Brazil and Chile

The aim of this subsection is to compare the overall degree of specialization of Argentina, Brazil and Chile using the measures described above, based on employment data of the local government entities of lower level. We analyze the evaluated measures with a particular attention to the dramatically different dimensions of the contingency tables of each country, due to the difference on the number of regions.

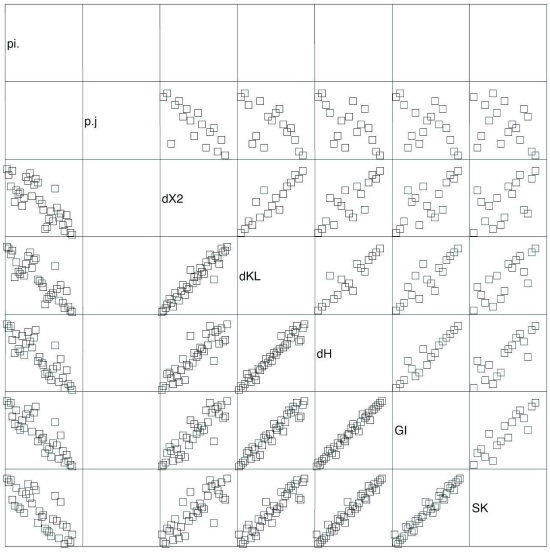



Figure 6: *Ranking dispersion between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$-above the main diagonal) measures (I=35, J=17)*

The regional units are the political-administrative jurisdictions called departments (#523), municipalities (#5,138) and communes (#342) for Argentina, Brazil and Chile respectively. The final number of regional units (after eliminating those without employees in the manufacturing sector) are 462, 5,138 and 249 for Argentina, Brazil and Chile, respectively. It is noteworthy that both regions of Brazil and Chile refer to the local government entity while those of Argentina refer to the cadastral divisions. Thus, from an administrative point of view, Argentina's divisions are not directly comparable with those of Brazil and Chile, although in some cases their boundaries coincide with those of the municipalities.

The data related to the employees in the manufacturing sector were obtained from of the Nationals Economic Census made by the National Institutes of Statistics and Censuses of Argentina (INDEC-1994: 1,083,928 employees), Brazil (IBGE-1998: 6,018,445 employees), and Chile (INE-2005: 446,613 employees), respectively. The data of Chile refer to the firms with 5 or more employees.

As in Section 2.2, the activity classifications refers to the first 2 digits of the International Standard Industrial Classification (ISIC Rev.3.1) of manufacturing activities (22 activities after grouping the divisions 36 and 37).

Table 4 shows a summary of the results obtained from the proposed measures of global specialization and the number of cells of each contingency table.

Table 4: *Summary of the results*

| **Measure** | **Argentina** | **Brazil** | **Chile** |
|---|---|---|---|
| $d_{\chi^2}(\mathbf{N})$ | 2.1580 | 3.1345 | 3.4363 |
| $d_{KL}(\mathbf{N})$ | 0.5049 | 0.7420 | 0.8870 |
| $d_H(\mathbf{N})$ | 0.1300 | 0.1894 | 0.2600 |
| $GI_{reg}$ | 0.4621 | 0.5595 | 0.6017 |
| $GI^{act}$ | 0.4880 | 0.5925 | 0.6358 |
| $SK_{reg}$ | 0.3625 | 0.4521 | 0.5079 |
| $SK^{act}$ | 0.3980 | 0.4856 | 0.5897 |
| #of cells | 10,164 (462x22) | 113,036 (5,138x22) | 5,478 (249x22) |

While the absolute values of these measures lie on different scales, the global measures of specialization show that Chile has a higher level of specialization, followed by Brazil and Argentina, respectively for all proposed global measure. Thus, this ranking does not depend on the selected measure of global specialization and does not depend on the number of cells. There is an extensive literature on the comparison of contingency tables with different sizes (see for *e.g.* van der Heijden *et al.* 1996, Lauritzen 2002, and Agresti 2002), but the present results are found relatively stable among these measures. Moreover, a similar stability is revealed in different simulations developed for this purpose (not shown in this paper) following extreme scenarios, not only referring to the dimension of the contingency tables but also to different levels of global specialization.

Once again, although $d_H$, $GI_{reg}$, $SK_{reg}$, $GI_{act}$, and $SK_{act}$ operate on a same range of variation, namely the unit interval, we systematically observe a same ranking, namely $d_H < SK_{reg} < SK^{act} < GI_{reg} < GI^{act}$, with rather substantial differences among these three measures. Also to be noticed, the fact that the ranking among the three countries for each measures *and* the ranking among the five measures for each country remain exactly the same. Comparing these results with those of subsection 2.2, we observe that Gini's coefficient are systematically higher than Krugman's coefficients and that in both cases activity based coefficients are higher than, but close to, region based coefficients. Note also that, in the case of Argentina, all coefficients are lower than in this application. This is due to the fact that, as already mentioned, the contingency table used in subsection 2.2 is a collapsed table of that used in this application, implying a loss information to be considered in next section.

# 3 Grouping of Regions or Activities

## 3.1 The MAUP problem

Different levels of aggregation, of either region or activities, typically imply different measures of specialization. As already noticed for the case of concentration (see *e.g.* Krugman 1991b and Anas, Arnott and Small 1998), the reason for these differences lies in the nature and balance of the centrifugal and centripetal force systems acting in different geographical scales. This problem is known as the "Modifiable Areal Unit Problem" (MAUP), which refers to the role of the geographical partition used (for more details, see Yule and Kendall 1950; Openshaw 1984; Arbia 1989; Amrhein 1995 and Unwin 1996). The arbitrariness of geographical boundaries gives rise to two different manifestations, namely aggregation and scale, and any statistical measure based on spatial aggregates is sensitive to the scale and aggregation problems. As same issue is also raised for the aggregation of activities.

The Figure 7 illustrates these problems. This example shows that, by observing a geographical distribution through regional aggregates, we would be in fact observing two separate phenomena which are matched in an unpredictable way with respect to: i) the actual distribution of objects in the space, and ii) the partition considered (for a formal argument, see *e.g.* Arbia 2001).

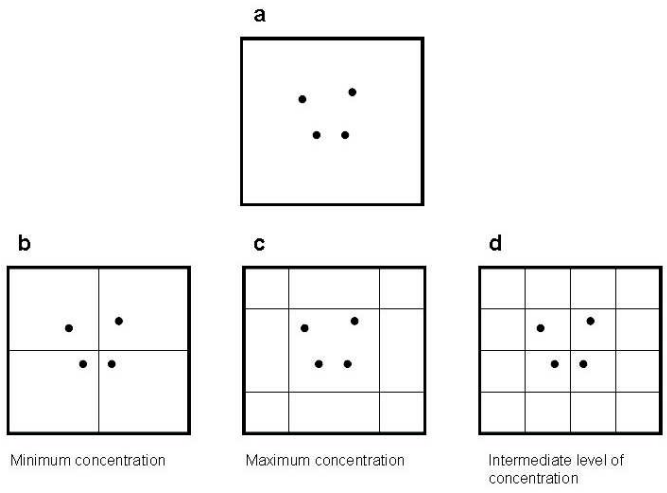



Figure 7: *A continuous space distribution of firms (**a**) and three discretized versions of it. Figures (**b**) and (**c**) illustrate the aggregation problem. Figures (**b**) and (**d**) illustrate the scale problem*

Likewise, and to illustrate the effects of the partition for specialization, Figure 8 shows that with a same distribution of firms in the space, it is possible to find specialization and the absence of specialization, respectively. In this example, the total geographical area could represent a country, polygonal subdivisions would correspond to regions, points to firms, and the orange and blue colors to two different economic activities.

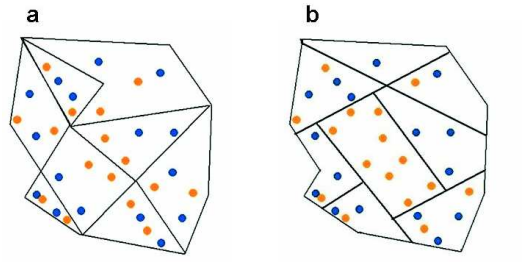



Figure 8: *Effects of the spatial partition on specialization. Figure (**a**) reveals a low (or not) level of specialization and Figure (**b**) a high level*

Therefore, it is important to note that the arbitrariness of partitions plays a key role in capturing the effects mentioned previously, and becomes potentially more dangerous the more unequal become the elements of it in terms of area. Arbia (1989) and Arbia and Espa (1996) discuss the distortions due to scale and aggregation and the possibilities of constructing optimal partitions of the space. One debatable issue is whether the boundaries between the discrete spatial (or sectoral) units should be such that they conform geographical areas (or sectoral activities) that are homogeneous in terms of their characteristics of interest, or such that the spatial (sectoral) units define areas (or activities) with distinctively different characteristics.

In the sequel, we focus the attention on evaluating the impact of grouping regions and/or activities on the measures of specialization.

## 3.2 Groupings

One way of dealing with the MAUP problem is to obtain a better grasp of the consequences of grouping regions and/or activities. Such is the object of this section.

Let us operate a partition of the $I$ regions into $M$ "grouped regions", to be called "g-regions" for the ease of exposition. Thus:

$$\mathcal{I} = \{1, 2, \cdots, I\} = \bigcup_{m=1}^{M} \mathcal{I}_m \qquad \mathcal{I}_m \cap \mathcal{I}_{m'} = \emptyset \ \ (m \neq m') \qquad \#(\mathcal{I}_m) = I_m \qquad \sum_m I_m = I \quad (25)$$

Using $q$ to denote probabilities on the space of the g-regions, we define:

$$q_{m\cdot} = \sum_{i \in \mathcal{I}_m} p_{i\cdot} \qquad q_{m|j} = \sum_{i \in \mathcal{I}_m} p_{i|j} \quad (26)$$

$$q_{\vec{m}\cdot} = (q_{1\cdot}, \cdots, q_{m\cdot}, \cdots, q_{M\cdot}) \qquad q_{\vec{m}|j} = (q_{1|j}, \cdots, q_{m|j}, \cdots, q_{M|j}) \quad (27)$$

Furthermore:

$$p_{i|m} = \frac{p_{i\cdot}}{q_{m\cdot}} \mathbb{1}_{\{i \in \mathcal{I}_m\}} \qquad p_{i|j,m} = \frac{p_{i|j}}{q_{m|j}} \mathbb{1}_{\{i \in \mathcal{I}_m\}} \quad (28)$$

The KL-divergence enjoys of a characteristic feature, namely to accept a decomposition relative to a grouping of the rows, or of the columns, providing a result similar to a decomposition of the variance made of the sum of a "within" term and a "between" term. This decomposition is well-known in the literature on information theory and has been widely used in spatial economics, see for instance Shorrocks (1980, 1982 and 1984), Mori, Nishikimi and Smith (2005), Brülhart and Traeger (2005), among others.

Indeed, starting with the second term of (12), we successively obtain:

$$
\begin{aligned}
d_{KL}(\mathbf{N}) &= \sum_j p_{\cdot j} \left[ \sum_i p_{i|j} \log \left( \frac{p_{i|j}}{p_{i\cdot}} \right) \right] && (29) \\
&= \sum_j p_{\cdot j} \left[ \sum_m q_{m|j} \log \frac{q_{m|j}}{q_{m\cdot}} \left\{ \sum_{i \in \mathcal{I}_m} p_{i|j,m} \right\} + \sum_m q_{m|j} \left\{ \sum_{i \in \mathcal{I}_m} p_{i|j,m} \log \frac{p_{i|j,m}}{p_{i|m}} \right\} \right] \\
&= \sum_j p_{\cdot j} \left[ d_{KL}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j}\, d_{KL}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] && (30)
\end{aligned}
$$

In (29), the KL-measure of specialization is viewed, as a general result, as a weighted average of industrial concentration, namely $d_{KL}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ in (14), whereas in (30) each of the activity measures is decomposed relatively to a partition of the regions into a "Between" term and a "Within" term, namely:

- *Between*: $\sum_j p_{\cdot j}\, d_{KL}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot})$, this is a weighted average of the activity specific measures of the specializations among the g-regions;
- *Within*: $\sum_j \sum_m p_{\cdot j} q_{m|j}\, d_{KL}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m})$, this is a (doubly) weighted average of the activity specific measures of the specializations among the composing regions of each g-regions;
- *Global = Between + Within.*

Two polar cases are of interest. Suppose first that $M = 1$, *i.e.* that all the regions of the country are grouped into a unique g-region, thus the country itself. In this case, the Between g-regions term vanishes and in the Within g-regions term the weighted average has only one term with $q_{m|j} = 1$ and the sum $\sum_{i \in \mathcal{I}_m}$ is equivalent to $\sum_{1 \leq i \leq I}$. Conversely, when $M = I$, each g-region has exactly one region and the Within g-regions term vanishes because each $d_{KL}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m})$ would represent a divergence between two degenerate one-point distributions whereas in the Between g-regions term $d_{KL}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot})$ coincides with $d_{KL}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ in (14).

Similarly to the analysis of variance, the ratio (Between/Global) may be interpreted as a measure of how far an aggregation criterion maintains the Global degree of specialization, the other ratio (Within/Global), measuring how far does an aggregation decrease the specialization. Heuristically, the ratio (Between/Global) may be seen as a measure of association between specialization and the

criterion of aggregation; remind that in the limit case of aggregation into a unique region, the Between term would annihilate. But another polar case would be obtained by aggregating identical, or very similar, regions. This would produce the within term to annihilate, or to decrease substantially. Thus, the ratio Between/Global may also be interpreted as a measure of the homogeneity of the aggregated regions. This feature is a central argument for constructing the "Best Collapsed Table" in Haedo (2009).

The two polar cases suggest the following issue. Let us compare the effects of two nested partitions; thus let the partition given in (25) along with a finer partition:

$$\mathcal{I} = \{1, 2, \cdots I\} = \bigcup_{m'=1}^{M'} \mathcal{I}_{m'} \qquad \mathcal{I}_{m'_1} \cap \mathcal{I}_{m'_2} = \emptyset \ (m'_1 \neq m'_2) \qquad \#(\mathcal{I}_{m'}) = I_{m'} \qquad \sum_{m'} I_{m'} = I$$
$$M < M' \quad \forall m' \ \exists m : \ \mathcal{I}_{m'} \subset \mathcal{I}_m \tag{31}$$

We may evaluate the sign of the changes in the between-term and in the within-term by refining successively each member of the coarser partition leaving its other members unaffected. The preceding reasoning shows that this refinement increases the between term and eventually decreases the within term, the limit being obtained in the case $M = I$.

The same analysis can be repeated when grouping activities instead of regions. Thus we now consider a partition of the activities into $L$ g-activities:

$$\mathcal{J} = \{1, 2, \cdots, J\} = \bigcup_{l=1}^{L} \mathcal{J}_l \qquad \mathcal{J}_l \cap \mathcal{J}_{l'} = \emptyset \ (l \neq l') \qquad \#(\mathcal{J}_l) = J_l \qquad \sum_l J_l = J \tag{32}$$

Using $r$ to denote probabilities on the space of the g-activities, we define:

$$r_{\cdot l} = \sum_{j \in \mathcal{J}_l} p_{\cdot j} \qquad r_{l|i} = \sum_{j \in \mathcal{J}_l} p_{j|i} \tag{33}$$

$$r_{\cdot \vec{l}} = (r_{\cdot 1}, \cdots, r_{\cdot l}, \cdots, r_{\cdot L}) \qquad r_{\vec{l}|i} = (r_{1|i}, \cdots, r_{l|i}, \cdots, r_{l|i}) \tag{34}$$

Furthermore:

$$p_{j|l} = \frac{p_{\cdot j}}{r_{\cdot l}} \mathbb{1}_{\{j \in \mathcal{J}_l\}} \qquad p_{j|i,l} = \frac{p_{j|i}}{r_{l|i}} \mathbb{1}_{\{j \in \mathcal{J}_l\}} \tag{35}$$

We may now repeat the decomposition of the KL-measure of specialization relatively to a grouping of activities. Indeed, starting with the first term of (12), we successively obtain:

$$d_{KL}(\mathbf{N}) = \sum_i p_{i\cdot} \left[ \sum_j p_{j|i} \log \left( \frac{p_{j|i}}{p_{\cdot j}} \right) \right] \tag{36}$$

$$= \sum_i p_{i\cdot} \left[ \sum_l r_{l|i} \log \frac{r_{l|i}}{r_{\cdot l}} \left\{ \sum_{j \in \mathcal{J}_l} p_{j|i,l} \right\} + \sum_l r_{l|i} \left\{ \sum_{j \in \mathcal{J}_l} p_{j|i,l} \log \frac{p_{j|i,l}}{p_{j|l}} \right\} \right]$$

$$= \sum_i p_{i\cdot} \left[ d_{KL}(r_{\vec{l}|i} \mid r_{\cdot \vec{l}}) + \sum_l r_{l|i} \, d_{KL}(p_{\vec{j}|i,l} \mid p_{\vec{i}|l}) \right] \tag{37}$$

Similarly to what has been observed for the regions, the two polar cases of interest now become: aggregating all activities into only 1 (*i.e.* $L = 1$) let the between g-activities term vanish and the within g-activities term be equal to the global measure whereas the finest partition, *i.e.* $L = J$, let the within g-activities term vanish and the between g-activities term be equal to the global measure.

As a final remark, aggregating regions into large ones, or aggregating activities, for instance by using less digit classification, always decreases the global measure of specialization because it only retains the between term and neglect the within term of the global measure before aggregation. Moreover, the coarser is the aggregation, the lower is the specialization. This remark may be viewed as a formal explanation of the impact of aggregation in the discussion of the MAUP problem.

The measures $d_{\chi^2}$ and $d_H$ accept the same decomposition relative to a grouping of the regions (rows) or of the activities (columns), but at difference from $d_{KL}(\mathbf{N})$ their decompositions are not exact and have residuals to be denoted as $R_{\chi^2}(\mathbf{N})$ and $R_H(\mathbf{N})$, respectively. Thus, for the decomposition relative to a grouping of the regions, we obtain:

$$d_{\chi^2}(\mathbf{N}) = \sum_j p_{\cdot j} \left[ d_{\chi^2}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j}\, d_{\chi^2}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] + R_{\chi^2}(\mathbf{N}) \tag{38}$$

$$d_H(\mathbf{N}) = \sum_j p_{\cdot j} \left[ d_H(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j}\, d_H(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] + R_H(\mathbf{N}) \tag{39}$$

And similarly for a grouping of activities.

### 3.3 Application to grouping of argentinean regions and activities

In next section, we examine the impact of grouping regions and/or activities. Thus, a natural question is raised: is the impact of these groupings on the degree of specialization similar for the three global measures?

We use again the same data as in section 2.2 and analyze the impact, on global specialization, of regrouping regions or activities by evaluating numerically the terms of the decomposition (30), (38) and (39), and the corresponding terms for the activities.

We first consider an arbitrary aggregation of regions by assembling the first 10 regions into a unique one (representing .7520 of the global employment), leaving the other regions as singletons in the aggregated partition. We analyze the numerical results from the following perspective:
(i) when decomposing the global measure of specialization with respect to an aggregation, how important are the residual terms for $d_{\chi^2}$ and $d_H$, knowing that there is no residual for $d_{KL}$?
(ii) does the ratio (Between/Global) strongly or weakly depend on the measure $d_{\chi^2}$, $d_{KL}$ or $d_H$?

Table 5 presents the numerical results for the aggregation of regions in the following order:

line 1: the 3 global measures, as given in (21);

line 2: the sum of the between and the within term;

line 3, 4 and 5: the between, within and residual terms;

line 6: the ratio of the residual term with the global term;

line 7 and 8: the ratio of the between term with the global term as given in lines 1 and 2.

Table 5: *Arbitrary grouping of 10 first regions*

| N° | Item | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ |
|---|---|---|---|---|
| 1 | $d_w(\mathbf{N})$ | 1.6532 | 0.3176 | 0.0713 |
| 2 | $d_w(\mathbf{N})$ Grouping | 1.6406 | 0.3176 | 0.0717 |
| 3 | Between | 1.2631 | 0.2107 | 0.0468 |
| 4 | Within | 0.3775 | 0.1069 | 0.0249 |
| 5 | Residual | 0.0126 | 0.0000 | -0.0004 |
| 6 | % Residual on $d_w(\mathbf{N})$ | 0.76 | 0.00 | -0.60 |
| 7 | % Between on $d_w(\mathbf{N})$ | 76.40 | 66.35 | 65.69 |
| 8 | % Between on $d_w(\mathbf{N})$ Grouping | 76.99 | 66.35 | 65.30 |

We notice the following features. Firstly, in this application the residual terms are never substantial, namely less than 1% of the global measure (lines 5 and 6). But this residual term may be positive (for $d_{\chi^2}$) or negative (for $d_H$). Secondly, the information provided by the ratio (Between/Global), lines 7 and 8, is not identical but fairly robust with respect to the 3 measures ($d_{\chi^2}$, $d_{KL}$ or $d_H$). This may be viewed as an indication that the (arbitrary) regroupment of 35 into 26 regions modifies significantly, but not dramatically, the global degree of specialization; one reason may be that the aggregation has been operated on fairly homogenous regions and fairly large regions with a percentage of the total employment ranging from 0.77% to 32.02% and $d_H$ ranging from 0.0189 to 0.1646; as the 10 aggregated regions cover more than 75% of the total employment, the remaining 25 regions are of smaller dimension.

Let us now consider another (arbitrary) partition by regrouping the last 10 regions. These are mostly small regions (representing between 0.05% and 0.51% of the total employment) with high specialization due their small sizes with $d_H$ ranging from .2262 to .6971. Together these 10 regions represent only 2.69% of the total employment. We now observe, in Table 6, that the residual part is considerably bigger than in Table 5 from 0.76% to 21.91% for $d_{\chi^2}$ and from 0.60% to 3.72% for $d_H$, with the same sign as in Table 5. The share of the between term, in line 8, increase considerably for the three measures from around 70% to around 90%. Notice however that for $d_{\chi^2}$ the between term decreases but its share, taking into account the inflated residual term, increases. This results shows that aggregating small regions into a unique one does affect only midly the global level of specialization, at variance from aggregating large regions.

Table 6: *Arbitrary grouping of 10 last regions*

| N° | Item | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ |
|---|---|---|---|---|
| 1 | $d_w(\mathbf{N})$ | 1.6532 | 0.3176 | 0.0713 |
| 2 | $d_w(\mathbf{N})$ Grouping | 1.2909 | 0.3176 | 0.0739 |
| 3 | Between | 1.2088 | 0.2911 | 0.0663 |
| 4 | Within | 0.0821 | 0.0265 | 0.0076 |
| 5 | Residual | 0.3622 | 0.0000 | -0.0027 |
| 6 | % Residual on $d_w(\mathbf{N})$ | 21.91 | 0.00 | -3.72 |
| 7 | % Between on $d_w(\mathbf{N})$ | 73.12 | 91.65 | 93.08 |
| 8 | % Between on $d_w(\mathbf{N})$ Grouping | 93.64 | 91.65 | 89.74 |

We now consider an arbitrary aggregation of activities by assembling the first 5 activities into a unique one (representing 40.92% of the global employment), leaving the other activities as singletons in the aggregated partition. The results are presented in Table 8 in the same format as in Table 7. We notice that in this second application the residual terms are substantially higher than in the first application with 15% and 5% of the global measure and the signs are the same as in the first application, positive for $d_{\chi^2}$ and negative for $d_H$. The three ratios of the terms (Between/Global) are different in value but with a similar order of magnitude. In both applications the ratio relative to $d_{KL}$ has a value intermediary between those relative to $d_{\chi^2}$ and to $d_H$, once the effect of the residual term has been taken into account, *i.e.* line 8 rather than 7.

Table 7: *Arbitrary grouping of 5 first activities*

| N° | Item | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ |
|---|---|---|---|---|
| 1 | $d_w(\mathbf{N})$ | 1.6532 | 0.3176 | 0.0713 |
| 2 | $d_w(\mathbf{N})$ Grouping | 1.4005 | 0.3176 | 0.0750 |
| 3 | Between | 1.0513 | 0.2250 | 0.0509 |
| 4 | Within | 0.3492 | 0.0925 | 0.0240 |
| 5 | Residual | 0.2527 | 0.0000 | -0.0037 |
| 6 | % Residual on $d_w(\mathbf{N})$ | 15.28 | 0.00 | -5.16 |
| 7 | % Between on $d_w(\mathbf{N})$ | 63.59 | 70.86 | 71.48 |
| 8 | % Between on $d_w(\mathbf{N})$ Grouping | 75.06 | 70.86 | 67.97 |

We now turn to another arbitrary partition of the activities, by regrouping the last 5 activities. The percentages of the global employment range from 0.086% to 6.82%; together they represent 13.04% of the total employment. The values of $d_H$ range from 0.0970 to 0.2165. Let us now compare the results in Table 7 and 8. For $d_{\chi^2}$ and $d_H$, the residual share remains at a similar level with the same sign. The share of the Between term, in line 8, considerably increases. Tables 7 and 8 display

the fact that when regrouping 5 smaller activities, representing 13% of the total employment, the global measure of specialization is less affected than by regrouping 5 larger activities, representing 41% of the total employment.

Taking on overview of these 4 exercises of regrouping we notice that:

- the share of the residual terms are always substantially lower for $d_H$ than for $d_{\chi^2}$;
- the sign of the residual terms is systematically positive for $d_{\chi^2}$ and negative for $d_H$;
- the share of the Between terms, after taking into account the residual term (*i.e.* line 8 of the Table) is smaller when aggregating larger regions, or activities, than when aggregating smaller ones.

Table 8: *Arbitrary grouping of 5 last activities*

| N° | Item | $d_{\chi^2}$ | $d_{KL}$ | $d_H$ |
|---|---|---|---|---|
| 1 | $d_w(\mathbf{N})$ | 1.6532 | 0.3176 | 0.0713 |
| 2 | $d_w(\mathbf{N})$ Grouping | 1.3591 | 0.3176 | 0.0747 |
| 3 | Between | 1.2622 | 0.2825 | 0.0653 |
| 4 | Within | 0.0969 | 0.0351 | 0.0095 |
| 5 | Residual | 0.2941 | 0.0000 | -0.0035 |
| 6 | % Residual on $d_w(\mathbf{N})$ | 17.79 | 0.00 | -4.87 |
| 7 | % Between on $d_w(\mathbf{N})$ | 76.35 | 88.95 | 91.59 |
| 8 | % Between on $d_w(\mathbf{N})$ Grouping | 92.87 | 88.95 | 87.33 |

# 4 Discussions and conclusions

## 4.1 The stochastic independence approach in a nutshell

Based on data in the form of a two-way contingency table "Regions × Activities", the concepts of specialization and of concentration are naturally based on the analysis of the conditional distributions, or profiles, $p_{\vec{j}|i}$ for the regional specializations or $p_{\vec{i}|j}$ for the industrial concentrations. The natural tools for measuring the degrees of specializations are provided by discrepancies $d(\cdot \mid \cdot)$, more precisely distances or divergences, among distributions: between profiles and a uniform distribution for absolute concepts ($d(p_{\vec{j}|i} \mid [a_i = I^{-1}])$ or $d(p_{\vec{i}|j} \mid [b_j = J^{-1}])$) that represent the spread of a distribution on categorical variables, between profiles and the corresponding marginal distribution ($d(p_{\vec{j}|i} \mid p_{.\vec{j}})$ or $d(p_{\vec{i}|j} \mid p_{\vec{i}.})$) for the relative concepts or between the joint distribution and the product of the marginal distributions $d([p_{ij}] \mid [p_{i.}\, p_{.j}])$ for the global concept. This is the approach

of stochastic independence that conducts the analysis in terms of stochastic independence between activities and regions and the global discrepancy is viewed as a measure of row-column association.

The relative and the global concepts may be written in terms of the local quotients $LQ_{ij}$ only; thus the local quotient is a local indicator of association at the level of the cell $(i, j)$ in the contingency table.

As the concept of stochastic independence is naturally symmetric between the activities and the regions, the global concept of specialization is uniquely defined, at variance from concepts developed in other frameworks that construct global measures of specializations by aggregating activity specific or region specific measures of specializations and eventually obtain different global measures of specialization. So is the case for Gini's and Krugman's indices.

## 4.2 A mathematical digression

The discrepancies $d_{\chi^2}$, $d_{KL}$ and $d_H$ have been widely used in several chapters of mathematical statistics. Another distance is also widely used; this is the $L_1$-distance based on the absolute deviations among probabilities

$$\begin{aligned} d_{L_1}(\mathbf{N}) &= \sum_i \sum_j |p_{ij} - p_{i\cdot}\, p_{\cdot j}| \\ &= \sum_i \sum_j p_{i\cdot} p_{\cdot j}\, |LQ_{ij} - 1| \\ &= \sum_i p_{i\cdot} \left[ \sum_j \left| p_{j|i} - p_{\cdot j} \right| \right] \\ &= \sum_j p_{\cdot j} \left[ \sum_i \left| p_{i|j} - p_{i\cdot} \right| \right] \end{aligned} \tag{40}$$

This distance of equivalent to the distance of total variation and has been widely used in particular for the analysis of robustness in mathematical statistics. It has also been used in economic geography. Moreover, $d_{L_1}(\mathbf{N})$ enjoys the same representations, in terms of local quotients, and the same decompositions as those given in Section 2.1. Because, in the present problem, the marginal probabilities are always strictly positive, both $I$ and $J$ are always finite and the range of variation is bounded by 2, its behaviour is essentially the same as $d_H$; this is the reason for not adding $d_{L_1}$ in our numerical evaluations.

## 4.3 On other approaches

Different approaches of global specialization have been envisaged in the economic literature, developed in frameworks that do not rely on stochastic independence.

Many relative measures are also based on the well-known Hoover-Balassa Local Quotient coefficient $LQ_{ij}$, also known as Location quotient. The analysis of the discrepancy between the activity

specific profile and the marginal (or country) region profile, $d(p_{\vec{i}|j} \mid p_{\vec{i}.})$, or between the region specific profile and the marginal activity profile, $d(p_{\vec{j}|i} \mid p_{.\vec{j}})$, may be divided into those based on Gini coefficient and Krugman index (*e.g.* Krugman 1991a, Kim 1995, Amiti 1998, Duranton and Puga 2000, Hallet 2000, Brülhart 2001, Dohse, Krieger-Boden and Soltwedel 2002, Midelfart-Knarvik, Overman, Redding and Venables 2002, Lafourcade and Mion 2003, Rossi-Hansberg 2005, Aiginger and Rossi-Hansberg 2006, Bickenbach and Bode 2006 and 2008, and many others), those based on Shannon's relative entropy or Generalized entropy ($GE$), also called relative Theil index (Theil 1967), and those based on the Coefficient of Variation (CV) (*e.g.* Aiginger and Davies 2001, Brülhart and Träger 2005, Bickenbach and Bode 2006 and 2008, and many others). It is interesting to notice that $d_{L_1}$ has been attributed as a variant of Relative Mean Deviation (RMD) of Krugman index in Bickenbach and Bode 2006 and 2008.

The proposals based on the Lorenz curve (Lorenz 1905), which is a graphical representation of the spread of a distribution derived from cumulative functions (see Appendix A) raise several difficulties: i) measures of concentration concern univariate distributions whereas the problems of concentration and of specialization concern a two-way contingency table; thus a common feature of these alternative measures is to analyze specialization by aggregating univariate properties, most often by aggregating over the regions some regional index of specialization, or by aggregating over activities some index of spatial dispersion (concentration) of each activities; and ii) originally they were designed for numerical variables. The adaptation of the Lorenz curve, and Gini index, to the case of categorical variables, such as activity or region, is obtained by ordering the (arbitrary) labels according to the ascending order of the local quotient; this implies a different ordering for each region and for each activity. These different orderings make the interpretation of the average Gini's coefficient difficult. The global index $GI^{reg}$ has been called a specialization coefficient where $GI^{act}$ has been called a coefficient of industrial concentration. The fact that in general $GI^{reg} \neq GI^{act}$ (see subsections 2.2 and 2.3) raises an issue of interpretation, particularly for the approach of stochastic independence that considers the regions and the activities interchangeably.

The Gini coefficient is based on the mean of the industrial structure distribution. This means it implicitly lends greater weight to the middle structure classes, which makes it more resistant vis-à-vis the underestimation of very high and very low employment structures. For these same attributes, the Gini coefficient has been criticized as tending to underestimate the amount of inequality (owing to the lower weight of values on the edge of the distribution). For more details see Atkinson (1983) and Lernan and Yitzhaki (1989).

As mentioned before, the local quotient $LQ_{ij}$ accepts a double reading in terms of the specialization within region $i$ (row-reading) or in terms of the specialization within activities $j$ (column-reading). This fact induces a unique concept of global specialization in terms of association, symmetrically between activity and region. This unicity is indeed reflected by the measurements developed

within the stochastic independence approach. Other approaches typically produce different global values by aggregating either region specific or activity specific measures; as a matter of fact, the difference between such pairs of measurements are minor and should be considered as unnecessary noises that should be attributed to a particular device of measurement and that jeopardize a clear understanding of the global concept of specialization. Given the categorical nature of the variables activity and region, the concepts of concentration and of specialization are naturally cast in terms of discrepancy between a distribution of interest (conditional or joint) and a reference distribution. Reference to a uniform distribution characterizes absolute measures of specialization whereas reference to the relevant marginal distribution characterizes relative measures of specialization. This is natural as far as absolute and relative specialization correspond to different problems of interest, in particular from the point of view policy making. Other reference distributions may however be more difficult to interpret and be eventually questionable but debating this issue falls out of the scope of this paper.

## 4.4 Final remarks

The approach of the stochastic independence has clarified the ideas (of, at least, these authors!). Table 1 has been a substantial step in this direction. Clarifying the ideas avoids catching attention on irrelevant issues but does by no means imply solving all problems! The field of spatial specialization very often involves the analysis of contingency tables with an extreme heterogeneity of the marginals, *i.e.* the ratios $(\max_j\ p_{\cdot j})/(\min_j\ p_{\cdot j})$ or $(\max_i\ p_{i\cdot})/(\min_i\ p_{i\cdot})$ may be extremely high, or of sizes of cells $(i, j)$. This heterogeneity raises issues on the robustness of the measurement and on the interpretation of international comparisons. The application of subsection 2.3 and the Section 3 on the impact of grouping provide hints on questions that quite clearly deserve further the attention.

Recently, Tajar (2003, Chapter 6), developed a representation of a two-way contingency table by means of copula, to be called a uniform representation of a discrete bivariate distribution. Interestingly enough, the construction is based on log-linear model for bivariate discrete variable where the first order interaction is determined by the cross-product ratio, or local quotient. This analysis opens potentially interesting avenues for a different approach to specialization from the point of view of region-activity association.

# References


Agresti, A. (2002), *Categorical Data Analysis*. New York: John Wiley and Sons.

Aiginger, K., and Davies, S. (2001), Industrial specialization and geographic concentration: two sides of the same coin? University of Linz, Working Paper Nr. 23.

Aiginger, K. and Rossi-Hansberg, E. (2006), Specialization and concentration: a note on theory and evidence. *Empirica* **33**: 255-266.

Amiti, M. (1999), Specialization patterns in Europe. *Weltwirtschaftliches Archiv* **135**: 573-593.

Amrhein, C. (1995), Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and Planing* **27**: 105-119.

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

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

Arbia, G. (2001), Modelling the geography of economic activities on continuous space. *Papers in Regional Science* **80**: 411-424.

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

Atkinson, A. (1983), *The Economics of Inequality*. Oxford: Clarendon Press.

Bickenbach, F., and Bode, E. (2006), Disproportionality measures of concentration, specialization and localization. Kiel Institute for the World Economy, Working Paper Nr. 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 Nr. 1660.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bollen, K., and Long, J. (1993), *Testing structural equation models*. Beverly Hills, CA: Sage.

Brülhart, M. (2001), Evolving geographical concentration of European Union. *Weltwirtschaftliches Archiv* **137**: 215-243.

Brülhart, M., and Traeger, R. (2005), An account of geographic concentration patterns in Europe. *Regional Science and Urban Economics* **35**: 597-624.

Cramer, H. (1946), *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Csiszár, I. (1967), Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicrum Hungarica* **2**: 229-318.

Dohse, D., Krieger-Boden, C., and Soltwedel, R. (2002), EMU and regional labor market disparities in Euroland. In J. Cuadrado-Roura and M. Parellada (eds.). *Regional Convergence in the European Union*. Berlin: Springer.

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

Everitt, B. (1977), *The analysis of Contingency Tables*. London: Chapman and Hall.

Gibbs, A., and Su, E. (2002), On choosing and bounding probability metrics. *International Statistical Review* **70**: 419-435.

Gini, C. (1912), Variabilità e mutabilità, contributo allo studio delle distribuzioni e relazioni statisiche. *Studi Economico-Giuridici dell'Università di Cagliari* **3**: 1-158.

Haedo, C. (2009), Measure of Global Specialization and Spatial Clustering for the Identification of "Specialized" Agglomeration. Ph.D. thesis, Bologna: Dipartimento di Scienze Statistiche "P.Fortunati", Università di Bologna (I).

Hallet, M. (2000), Regional specialization and concentration in the EU. European Commission, Economic Papers Nr. 141.

Joe, H. (1989), Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84**: 157-164.

Kendall, M., and Stuart, A. (1963), *The Advanced Theory of Statistics. Volume 1: Distribution Theory*. London: Griffin.

Kim, S. (1995), Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860-1987. *The Quarterly Journal of Economics* **110**: 881-908.

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

Krugman, P. (1991b), *Geography and Trade*. Cambridge: MIT Press.

Lafourcade, M., and Mion, G. (2003), Concentration, agglomeration and the size of plants: disentangling the source of co-location externalities. Université catholique de Louvain, CORE Discussion Paper Nr. 91.

Lauritzen, S. (1982), *Lectures on Contingency Tables*. Aalborg: University of Aalborg Press.

Lerman, R., and Yitzhaki, S. (1989), Improving the accuracy of estimates of the Gini Coefficient. *Journal of Econometrics* **42**: 43-47.

Lorenz, M. (1905), Methods of measuring the concentration of wealth. *Journal of the American Statistical Association* **9**: 209-219.

Midelfart-Knarvik, K., Overman, H., Redding, S., and Venables, A. (2000), The location of European industry. European Comission, Economic Papers Nr. 142.

Mori, T., Nishikimi, K., and Smith, T. (2005), A divergence statistic for industrial localization. *Review of Economics and Statistics* **87**: 635-651.

Openshaw, S. (1984), *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Osberg, L., and Xu, K. (2000), International comparison of poverty intensity: index decomposition and bootstrap inference. *Journal of Human Resources* **35**: 51-81.

Reiss, R. (1989), *Approximate distributions of order statistics*. New York: Springer-Verlag.

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

Shorrocks, A. (1980), The class of additively decomposable inequality measures. *Econometrica* **48**: 613-625.

Shorrocks, A. (1982), Inequality decomposition by factor components. *Econometrica* **50**: 193-211.

Shorrocks, A. (1984), Inequality decomposition by population subgroups. *Econometrica* **52**: 1369-1385.

Slottje, D. (1990), Using grouped data for constructing inequality indices: parametric vs. nonparametric methods. *Economics Letters* **32**: 193-197.

Tajar, A. (2003), Measuring and modelling dependence. Ph.D. thesis, Louvain la Neuve: ISBA, Université catholique de Louvain (B).

Theil, H. (1967), *Economics and Information Theory*. Amsterdam: North-Holland.

Tjøstheim, D. (1996). Measures and tests of independence: a survey. *Statistics* **28**: 249-284.

Unwin, D. (1996), GIS, spatial analysis and spatial statistics. *Progress in Human Geography* **20**: 540-551.

van der Heijden, P., Mooijaart, A., and Takane, Y. (1994), Correspondence analysis and contingency table models in correspondence analysis in the social sciences. In M. Greenacre and J. Blasius (eds.). *Correspondence Analysis in the Social Sciences*. London: Academic Press.

Xu, Kuan (2003), How has the literature on Gini's index evolved in the past 80 years? Dalhousie University, Economics Working Paper.

Yule, U., and Kendall, M. (1950), *An Introduction to the Theory of Statistics*. London: Charles Griffin.

# Appendix A: Gini and Krugman indexes

Many indexes commonly used throughout the economic literature to describe the phenomenon of regional specialization and industrial concentration, are based on the Lorenz curve (Lorenz 1905). The Lorenz curve is a graphical representation of the spread of a distribution based on the cumulative functions. More explicitly, for a numerical variable $X$, the Lorenz curve is represented on the unit square $[0\ 1]^2$ with a coordinate system made of the functions $F_X(x)$, the cumulative distribution function, and $\mu_X(x)$, the relative mean function:

$$F_X(x) = \sum_{u_j \leq x} f_X(u_j) \quad or \quad \int_0^x f_X(u)du;$$

$$\mu_X(x) = \frac{\sum_{u_j \leq x} u_j f_X(u_j)}{\sum_0^\infty u_j f_X(u_j)} \quad or \quad \frac{\int_0^x u f_X(u)du}{\int_0^\infty u f_X(u)du}.$$

The points on the main diagonal represent individuals with a value $x$ such that the proportion of individuals with a value of $X$ lower or equal to $x$ is the same as of their corresponding proportion of the overall average. Thus, a distribution where each individual is characterized with a same value $x$ would be represented by the main diagonal. The area between the main diagonal and the Lorenz curve may accordingly be interpreted as a graphical representation of the spread of the distribution.

The Lorenz curve has been originally developed for a univariate numerical variable. Two issues are at stake in the following extension of Gini index (Gini 1912) to the characterization of the relative regional specialization: the simultaneity of two dimensions, namely region and activity, and the categorical feature of these two variables for which there is no natural order as in the case for numerical variables.

For a given region $i$, the activities may be ordered in increasing order of the local quotient:

$$LQ_{i,j_i(1)} \ < \ LQ_{i,j_i(2)} \ < ... \ < \ LQ_{i,j_i(k)} \ < ... \ < \ LQ_{i,j_i(J)} \tag{41}$$

where $j_i$ is a permutation of $\{1, ..., J\}$ different for each region $i$. Finally we construct the coordinates of the unit square through the increasing sequences of the following cumulative functions:

$$P^{(i)}_{\cdot j_i(1)} \ < \ P^{(i)}_{\cdot j_i(2)} \ < ... \ < \ P^{(i)}_{\cdot j_i(k)} \ < ... \ < \ P^{(i)}_{\cdot j_i(J)}$$

and

$$P^{(i)}_{j_i(1)|i} \ < \ P^{(i)}_{j_i(2)|i} \ < ... \ < \ P^{(i)}_{j_i(k)|i} \ < ... \ < \ P^{(i)}_{j_i(J)|i}$$

where $P^{(i)}_{\cdot k} = \sum_{a \leq k} p_{\cdot j_i(a)}$ and $P^{(i)}_{k|i} = \sum_{a \leq k} p_{j_i(a)|i}$, respectively. Thus, $P^{(i)}_{\cdot j_i(k)}$ represents the proportion of the country cumulative employment of the activities that, *in region i*, have a local quotient lower or equal to that of the $k$-th activity upon the ordering given in (41) and $P^{(i)}_{j_i(k)|i}$ represents the

similar proportion, now relatively to the region $i$ only. We now construct, for region $i$, a curve, connecting by linear interpolation the points with coordinates $P^{(i)}_{\cdot j_i(k)}$ and $P^{(i)}_{j_i(k)|i}$. A region $i$ where each activity has a unit local quotient is represented by the main diagonal. The actual curve of a region $i$ will not cross the main diagonal because of the ordering (41). The actual curve may accordingly to be considered as a Lorenz curve and the area between the curve and the main diagonal may be interpreted as a graphical representation of specialization.

The relative Gini specialization coefficient of region $i$, $GI_i$, is constructed geometrically as the ratio (area between the Lorenz curve and the main diagonal, say A/area under the main diagonal), or equivalently 1-(area under the Lorenz curve, say B/area under the main diagonal) (Fig. 9 where area $\alpha = [P^{(i)}_{\cdot k} - P^{(i)}_{\cdot k-1}] \times \frac{1}{2}[P^{(i)}_{k|i} + P^{(i)}_{k-1|i}]$).

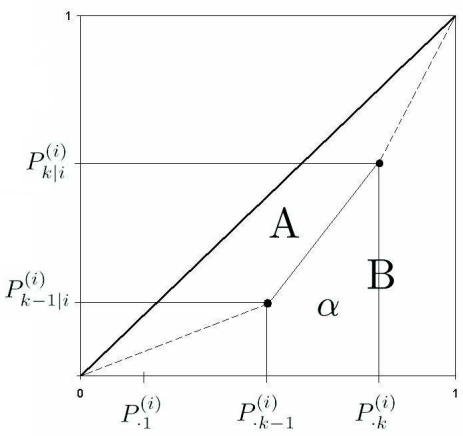



Figure 9: *Lorenz curve for specialization*

As the area under the main diagonal is equal to 1/2, we obtain:

$$GI_i = 1 - \sum_{1 \leq k \leq J} \left( P^{(i)}_{\cdot k} - P^{(i)}_{\cdot k-1} \right) \left( P^{(i)}_{k|i} + P^{(i)}_{k-1|i} \right) \tag{42}$$

where $P^{(i)}_{\cdot 0} = P^{(i)}_{0|i} = 0$. This is only a geometric presentation of Gini coefficient. Lerman and Yitzhaki (1989), Osberg and Xu (2000) and Xu (2003) provide an interesting overview of alternative presentations and their respective merits.

$GI_i$ takes values in the range [0 1], *i.e.* a value 0 means that a region has the same activity shares as those of the whole country, while a value 1 denotes the limit case of extreme relative specialization for a region with a unique activity, the share of which is infinitely small in the country.

The same construction may be considered for each activity in order to construct a relative industrial concentration coefficient

$$GI^j = 1 - \sum_{1 \leq r \leq I} \left( P^{(j)}_{r\cdot} - P^{(j)}_{r-1\cdot} \right) \left( P^{(j)}_{r|j} + P^{(j)}_{r-1|j} \right) \tag{43}$$

where $P^{(j)}_{0\cdot} = P^{(j)}_{0|j} = 0$, under an activity-specific reordering of the regions:

$$LQ_{i_j(1),j} \ < \ LQ_{i_j(2),j} \ < ... \ < \ LQ_{i_j(r),j} \ < ... \ < \ LQ_{i_j(I),j}. \tag{44}$$

The index $SK_i$ proposed by Krugman (1991a) is a measure of regional specialization or industrial concentration, expressed as half of the Relative Mean Deviation (RMD) based on the Manhattan distance (see for more details Kendall and Stuart 1963). The relative version of this index captures the gap between the activity structure of region $i$ and the average of the activity $j$ structure of the other regions. It is defined as:

$$SK_i = \frac{1}{2} \sum_j \mid p_{j|i} - \overline{p_{\cdot j}} \mid \tag{45}$$

where

$$\overline{p_{\cdot j}} = \frac{\sum_{m \neq i}^{I} N_{mj}}{\sum_{m \neq i}^{I} \sum_j N_{mj}} \tag{46}$$

The $SK_i$ index takes a zero value if the activity structure of region $i$ is identical to the average of the other regions. Given the normalization used here, the maximum value of $SK_i$ is equal to 1 when the activity structure of one region differs completely from the rest of the country.

The index for relative industrial concentration is constructed similarly:

$$SK^j = \frac{1}{2} \sum_i \mid p_{i|j} - \overline{p_{i\cdot}} \mid \tag{47}$$

where

$$\overline{p_{i\cdot}} = \frac{\sum_{l \neq j}^{J} N_{il}}{\sum_i \sum_{l \neq j}^{J} N_{il}} \tag{48}$$

# Appendix B: Tables of argentinean data

Table 9: *Argentinean data (1)*

| Region \ Activity | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 28,919 | 272 | 4,238 | 7,104 | 2,106 | 1,977 | 2,577 | 22,108 | 601 | 9,003 | 21,385 | 3,299 |
| 2 | 3,819 | 6 | 1,496 | 6,779 | 424 | 258 | 479 | 1,750 | 4 | 2,147 | 3,300 | 384 |
| 3 | 50,279 | 1,280 | 25,655 | 14,639 | 17,799 | 5,731 | 9,574 | 9,431 | 1,348 | 29,783 | 119,628 | 10,942 |
| 4 | 3,825 | 0 | 279 | 613 | 157 | 199 | 846 | 521 | 170 | 4,463 | 5,127 | 352 |
| 5 | 16,261 | 0 | 1,818 | 1,377 | 1,152 | 1,494 | 2,709 | 1,458 | 511 | 4,054 | 17,944 | 7,958 |
| 6 | 6,157 | 0 | 809 | 317 | 118 | 442 | 219 | 514 | 1,778 | 2,991 | 3,012 | 804 |
| 7 | 8,487 | 0 | 1,336 | 675 | 5,311 | 658 | 878 | 419 | 26 | 2,025 | 4,789 | 885 |
| 8 | 2,791 | 1,977 | 55 | 54 | 18 | 208 | 61 | 80 | 2 | 68 | 2,436 | 356 |
| 9 | 47,042 | 0 | 2,053 | 3,096 | 2,932 | 2,072 | 1,960 | 3,397 | 65 | 6,113 | 39,813 | 4,345 |
| 10 | 15,456 | 0 | 978 | 1,644 | 2,470 | 2,368 | 925 | 2,559 | 327 | 1,199 | 14,123 | 2,491 |
| 11 | 8,323 | 0 | 64 | 225 | 77 | 602 | 491 | 1,364 | 0 | 651 | 3,526 | 725 |
| 12 | 12,516 | 0 | 164 | 253 | 290 | 1,630 | 202 | 494 | 39 | 248 | 3,645 | 1,307 |
| 13 | 3,188 | 11 | 332 | 571 | 481 | 673 | 185 | 300 | 0 | 519 | 2,328 | 6,243 |
| 14 | 31,462 | 0 | 128 | 458 | 417 | 874 | 61 | 606 | 0 | 411 | 3,804 | 528 |
| 15 | 851 | 0 | 193 | 3,255 | 289 | 149 | 1 | 151 | 0 | 4 | 827 | 61 |
| 16 | 1,409 | 0 | 732 | 140 | 59 | 120 | 462 | 200 | 784 | 639 | 2,125 | 454 |
| 17 | 18,929 | 12 | 8,856 | 1,338 | 856 | 930 | 594 | 1,799 | 0 | 814 | 6,340 | 1,842 |
| 18 | 1,375 | 0 | 6 | 34 | 4,568 | 23 | 0 | 40 | 0 | 5 | 117 | 30 |
| 19 | 388 | 0 | 2 | 63 | 5 | 248 | 0 | 11 | 1,118 | 786 | 436 | 44 |
| 20 | 6,737 | 0 | 72 | 2,654 | 334 | 298 | 89 | 474 | 0 | 165 | 4,655 | 497 |
| 21 | 8,372 | 0 | 29 | 73 | 49 | 3,195 | 223 | 536 | 0 | 322 | 2,158 | 591 |
| 22 | 2,509 | 0 | 6,917 | 518 | 540 | 427 | 30 | 318 | 1 | 28 | 1,291 | 581 |
| 23 | 476 | 0 | 2 | 88 | 5 | 32 | 136 | 131 | 0 | 25 | 468 | 4,668 |
| 24 | 1,069 | 3 | 54 | 16 | 0 | 171 | 2,011 | 59 | 0 | 613 | 597 | 152 |
| 25 | 1,411 | 0 | 99 | 177 | 2,239 | 109 | 0 | 205 | 0 | 10 | 328 | 261 |
| 26 | 4,657 | 0 | 0 | 0 | 1 | 24 | 0 | 20 | 0 | 3 | 75 | 48 |
| 27 | 408 | 0 | 19 | 9 | 5 | 2,715 | 1,563 | 24 | 0 | 49 | 211 | 215 |
| 28 | 1,426 | 11 | 11 | 27 | 23 | 1,907 | 0 | 88 | 0 | 189 | 462 | 46 |
| 29 | 332 | 961 | 35 | 6 | 178 | 40 | 1 | 40 | 0 | 127 | 59 | 18 |
| 30 | 108 | 0 | 0 | 20 | 0 | 798 | 0 | 2 | 0 | 0 | 36 | 27 |
| 31 | 180 | 0 | 1,913 | 7 | 7 | 80 | 0 | 10 | 0 | 345 | 103 | 62 |
| 32 | 85 | 0 | 0 | 0 | 0 | 20 | 0 | 2 | 0 | 1,147 | 22 | 25 |
| 33 | 415 | 3 | 632 | 4 | 5 | 481 | 0 | 50 | 0 | 16 | 198 | 83 |
| 34 | 24 | 477 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 20 | 4 |
| 35 | 485 | 0 | 224 | 24 | 1 | 164 | 0 | 76 | 0 | 85 | 827 | 61 |
| $N_{\cdot j}$ | 290,171 | 5,013 | 59,201 | 46,258 | 42,916 | 31,120 | 26,277 | 49,237 | 6,774 | 69,047 | 266,215 | 50,389 |
| $p_{\cdot \vec{j}}$ | 0.2677 | 0.0046 | 0.0546 | 0.0427 | 0.0396 | 0.0287 | 0.0242 | 0.0454 | 0.0062 | 0.0637 | 0.2456 | 0.0465 |
| $d_{\chi^2}(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 0.3625 | 59.8952 | 1.8482 | 1.8675 | 2.9312 | 3.7377 | 2.0062 | 1.3557 | 9.7915 | 0.6328 | 0.1499 | 2.5095 |
| $d_{KL}(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 0.1546 | 2.8511 | 0.5262 | 0.4437 | 0.5727 | 0.5833 | 0.3892 | 0.4407 | 1.3626 | 0.2533 | 0.0851 | 0.4716 |
| $d_H(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 0.0376 | 0.5331 | 0.1283 | 0.0953 | 0.1233 | 0.1043 | 0.0897 | 0.0978 | 0.3039 | 0.0699 | 0.0236 | 0.0889 |

Table 10: *Argentinean data (2)*

| Region \ Activity | 13 | 14 | 15 | 16 | 17 | $N_{i\cdot}$ | $p_{\vec{i}\cdot}$ | $d_{\chi^2}(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$ | $d_{KL}(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$ | $d_H(p_{\vec{j}\mid i} \mid p_{\cdot\vec{j}})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,747 | 2,271 | 1,318 | 2,557 | 872 | 112,354 | 0.1037 | 0.6112 | 0.1982 | 0.0433 |
| 2 | 68 | 281 | 114 | 587 | 21 | 21,917 | 0.0202 | 1.8948 | 0.4870 | 0.1025 |
| 3 | 12,232 | 5,490 | 3,892 | 25,779 | 3,556 | 347,038 | 0.3202 | 0.1380 | 0.0727 | 0.0189 |
| 4 | 550 | 142 | 53 | 282 | 55 | 17,634 | 0.0163 | 0.7436 | 0.2753 | 0.0661 |
| 5 | 1,297 | 233 | 456 | 3,642 | 206 | 62,570 | 0.0577 | 0.2266 | 0.0978 | 0.0246 |
| 6 | 658 | 74 | 221 | 791 | 419 | 19,324 | 0.0178 | 1.4528 | 0.2960 | 0.0610 |
| 7 | 244 | 134 | 12 | 543 | 261 | 26,683 | 0.0246 | 0.7784 | 0.2522 | 0.0584 |
| 8 | 39 | 8 | 57 | 167 | 4 | 8,381 | 0.0077 | 11.8685 | 0.9283 | 0.1646 |
| 9 | 2,899 | 931 | 180 | 7,550 | 1,119 | 125,567 | 0.1158 | 0.1490 | 0.0857 | 0.0246 |
| 10 | 1,875 | 490 | 404 | 25,526 | 799 | 73,634 | 0.0679 | 1.2604 | 0.3673 | 0.0787 |
| 11 | 109 | 27 | 16 | 491 | 10 | 16,701 | 0.0154 | 0.4207 | 0.2396 | 0.0716 |
| 12 | 40 | 18 | 20 | 468 | 60 | 21,394 | 0.0197 | 0.7144 | 0.3571 | 0.0981 |
| 13 | 42 | 7 | 22 | 164 | 22 | 15,088 | 0.0139 | 3.1200 | 0.6900 | 0.1403 |
| 14 | 423 | 28 | 10 | 426 | 50 | 39,686 | 0.0366 | 1.4241 | 0.6328 | 0.1683 |
| 15 | 21 | 8 | 26 | 81 | 0 | 5,917 | 0.0055 | 6.3716 | 1.1579 | 0.2359 |
| 16 | 13,022 | 5 | 281 | 1,328 | 60 | 21,820 | 0.0201 | 9.9012 | 1.4830 | 0.2749 |
| 17 | 175 | 295 | 210 | 1,832 | 484 | 45,306 | 0.0418 | 0.6022 | 0.2491 | 0.0627 |
| 18 | 4 | 1 | 0 | 1 | 0 | 6,204 | 0.0057 | 12.8803 | 2.0079 | 0.4362 |
| 19 | 810 | 6 | 4 | 1 | 1,093 | 5,015 | 0.0046 | 13.7846 | 1.6876 | 0.3436 |
| 20 | 111 | 29 | 4 | 653 | 127 | 16,899 | 0.0156 | 0.5702 | 0.2708 | 0.0768 |
| 21 | 334 | 8 | 2 | 263 | 19 | 16,174 | 0.0149 | 1.5169 | 0.5287 | 0.1365 |
| 22 | 47 | 47 | 5 | 366 | 40 | 13,665 | 0.0126 | 4.0248 | 0.8928 | 0.1873 |
| 23 | 4 | 5 | 4 | 65 | 17 | 6,126 | 0.0057 | 11.5757 | 1.8579 | 0.3713 |
| 24 | 0 | 41 | 1 | 28 | 2 | 4,817 | 0.0044 | 6.7694 | 1.0841 | 0.2266 |
| 25 | 2 | 0 | 0 | 54 | 1 | 4,896 | 0.0045 | 4.7676 | 0.9956 | 0.2243 |
| 26 | 0 | 0 | 0 | 0 | 0 | 4,828 | 0.0045 | 2.4799 | 1.1556 | 0.3738 |
| 27 | 1 | 1 | 1 | 11 | 0 | 5,232 | 0.0048 | 12.1284 | 2.0210 | 0.4202 |
| 28 | 1 | 1 | 12 | 65 | 2 | 4,271 | 0.0039 | 6.4580 | 1.1002 | 0.2286 |
| 29 | 0 | 0 | 0 | 15 | 2 | 1,814 | 0.0017 | 60.1717 | 2.3845 | 0.3575 |
| 30 | 0 | 0 | 1 | 0 | 2 | 994 | 0.0009 | 21.5243 | 2.4656 | 0.5016 |
| 31 | 0 | 1 | 0 | 0 | 2 | 2,710 | 0.0025 | 8.4426 | 1.6892 | 0.3839 |
| 32 | 0 | 0 | 0 | 0 | 0 | 1,301 | 0.0012 | 11.2352 | 2.1474 | 0.5071 |
| 33 | 3 | 1 | 0 | 8 | 0 | 1,899 | 0.0018 | 3.5439 | 0.9519 | 0.2262 |
| 34 | 0 | 0 | 0 | 0 | 0 | 528 | 0.0005 | 175.4862 | 4.5909 | 0.6971 |
| 35 | 0 | 0 | 3,379 | 215 | 0 | 5,541 | 0.0051 | 36.8669 | 2.2439 | 0.3552 |
| $N_{\cdot j}$ | 36,758 | 10,583 | 10,705 | 73,959 | 9,305 | 1,083,928 | | | | |
| $p_{\cdot\vec{j}}$ | 0.0339 | 0.0098 | 0.0099 | 0.0682 | 0.0086 | | | | | |
| $d_{\chi^2}(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 5.8657 | 0.4719 | 19.1844 | 1.3368 | 2.9976 | | | $d_{\chi^2}(\mathbf{N})$ =1.6532 | | |
| $d_{KL}(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 0.8976 | 0.2827 | 1.2530 | 0.4345 | 0.4386 | | | | $d_{KL}(\mathbf{N})$ =0.3176 | |
| $d_H(p_{\vec{i}\mid j} \mid p_{\vec{i}\cdot})$ | 0.1766 | 0.0896 | 0.2165 | 0.1035 | 0.0970 | | | | | $d_H(\mathbf{N})$ =0.0713 |