2012/25

A stochastic independence approach for different measures of concentration and specialization

Christian Haedo and Michel Mouchart

CORE DISCUSSION PAPER
2012/25

A stochastic independence approach for different measures of concentration and specialization

Christian HAEDO [1] and Michel MOUCHART[2]

May 2012

Abstract

From data in the form of a two-way contingency table "Regions × Sectors", the concepts of specialization and concentration, built from the analysis of conditional distributions or profiles, is based on discrepancies 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 paper provides an extensive numerical analysis of measures derived from this approach and from other approaches used in the literature and shows that while the different measures under consideration display rather similar numerical behaviours, differences of ranking call for a particular care when interpreting the numerical results.

Keywords: absolute, relative and global specialization, industrial concentration, polarization, localization, stochastic independence, contingency tables.

[1] CIDETI, University of Bologna, Argentina.
[2] Université catholique de Louvain, CORE and ISBA, B-1348 Louvain-la-Neuve, Belgium.
E-mail: michel.mouchart@uclouvain.be

Michel Mouchart gratefully acknowledges financial support from IAP research network grant *nrP6/*03 of the Belgian government (Belgian Science Policy). Both authors gratefully acknowledge the financial support of CIDETI, that promoted intercontinental cooperation. Vicente Donato and Dominique Peeters provided a wealth of insightful comments at different stages of the paper's preparation and deserve the authors' deepest gratitude.

1 Introduction

The literature on measuring specialization and concentration in economic geography and spatial economics has been dramatically expanding

during the last twenty years, putting forth the obvious relevance of the topic. This paper aims at providing a fresh look at the measurement of concentration and specialization using the perspective of stochastic independence in the analysis of contingency tables (for the sake of brevity, we use interchangeably the framework of stochastic independence or of contingency tables). What is our contribution? A unique and simple principle: analyze the spatial structure of a country by comparing distributions, or profiles, characterizing sectors or regions, taking due account that the variables “region” or “sector” are categorical variables, *i.e.* not numerical and not even ordered.

The interest for the contingency table approach mainly lies on two issues. Firstly, it provides a coherent framework to analyze concentration and specialization at three different levels; from a decisional point of view, these three levels correspond to three different economic policy issues. Secondly, this approach provides access to significant literature open to a wide range of application fields encompassing virtually all social sciences. While the approach of contingency table does not cover all the indices proposed so far in the New Economic Geography (henceforth, NEG) literature, relying on a coherent framework is helpful to better appreciate criticisms raised against indices coming from other families, such as those derived from the Lorenz curve and its associated Gini’s index, or those associated to the works of Krugman. The practitioner may particularly appreciate an increased coherence in the treatment of issues such as the decomposition or the aggregation of sectors or regions, the ranking of different countries or different periods for the same country.

This paper has quite a pragmatic motivation as it aims at suggesting practitioners whatever is involved in selecting a particular measure of concentration or specialization. Hence, we want to compare several alternative measures for a given concept with the following questioning: (i) how to evaluate the kind of information provided by competing measures? (ii) how to evaluate the numerical behaviour of these measures under different circumstances, in particular when grouping regions or sectors, or when ordering the degree of specialization of regions or the degree of concentration of sectors? This case provides the base for a discussion of the relationship between a concept and its measure, a crucial methodological issue in social sciences. On this topic, the interested reader may like to have a serious look at Sheldon and Moore (1972) or Zeller and Carmines (1980). Accordingly, we wish to compare numerically measures of relative and global specialization, some within the framework of the stochastic independence approach, others in different frameworks. The underlying question for these comparisons is to evaluate to what extent are these measures mutually coherent and quantify the same concept. We also check whether these measures operate in the same ranking of specializations among sectors, regions or countries. A heuristic conclusion of our analysis is that the different measures proposed in the stochastic independence approach are reasonably coherent, but ranking differences require cautiousness at the interpretation stage. Finally, this paper is focused on descriptive measures of global specialization but does not consider sampling or asymptotic properties in view of an eventual statistical inference as in the works of Brülhart and Traeger (2005), Mori, Nishikimi and Smith (2005), and Mulligan and Schmidt (2005).

The paper is divided in two main parts. The first part is included in the following section, that puts forth the stochastic independence approach and explains how different measures of concentration or of specialization can be built and adapted to different levels of analysis. The second part evaluates the stochastic independence approach according to the criteria developed in the next three sections. Section 3 includes an overview of the literature in economic geography and spatial economics, while it examines the potential contribution of the stochastic independence approach. In section 4 we confront the stochastic independence approach with Argentina-related data and evaluate the numerical behaviour of the proposed measures and of other measures based on the framework of Gini and Krugman indices. Section 5 confronts the stochastic independence approach with challenges raised by grouping regions or sectors. The paper concludes with a short summary of the stochastic independence approach and with some final remarks.

2 The approach of stochastic independence

The approach called "*a stochastic independence approach*" stands for the idea that the spatial structure of an economy is analyzed in terms of distributions, and comparisons of distributions, and that regional specialization and industrial concentration are viewed as a distributional issue of statistical association between "region" and "sector".

2.1 The structure of the data

We start by describing the data we propose to handle. Indeed, the available data determine which measures can be operationalized. By the same token we introduce the notation to be used.

For a given country, let us consider regions labeled $i \in \mathcal{I} = \{1, ..., I\}$, and sectors labeled $j \in \mathcal{J} = \{1, ..., J\}$. For each pair $(i, j) \in \mathcal{I} \times \mathcal{J}$, we observe the number of primary units, 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:

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

Equivalently, the data may be represented by the complete sample size, $N_{\cdot\cdot}$, and the relative frequencies :

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

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

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

  $$p_{\vec{j}|i} = (p_{1|i}, \cdots, p_{j|i}, \cdots, p_{J|i}) \tag{3}$$

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

  $$p_{\cdot\vec{j}} = (p_{\cdot 1}, \cdots, p_{\cdot j}, \cdots, p_{\cdot J}) \tag{4}$$

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

  $$p_{\vec{i}|j} = (p_{1|j}, \cdots, p_{i|j}, \cdots, p_{I|j}) \tag{5}$$

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

  $$p_{\vec{i}\cdot} = (p_{1\cdot}, \cdots, p_{i\cdot}, \cdots, p_{I\cdot}) \tag{6}$$

Accordingly, this paper handles issues related to a discrete space, *i.e.* a space partitioned into a finite number of regions. Moreover, label i in 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. Thus the data N_{ij} provides no information about the localization of primary units *within* a region. Problems of agglomeration, or spatial dependence among regions, can therefore not be suitably handled through these data: these problems would require *additional* data related to the distance, or contiguity, between the regions. When the country is treated as a unique continuous space, the basic data refer to the localization of points in the country and the interest is focused on designing stochastic processes, such as a marked point process, in order to represent locally diffusion issues. See for instance Duranton and Overman (2005) for an analysis of localization through point processes. Through this approach, motivation is more oriented to modelling and explaining an observed spatial structure. The continuous approach cannot be developed with the data under consideration in this paper; however, readers who are interested in modelling continuous spaces may fruitfully read Barff (1987), Arbia (2001), Marcon and Puech (2003), Arbia, Espa and Quah (2007), Haedo (2009, Chapter 4), Kosfeld, Eckey and Lauridsen (2011).

2.2 A preliminary: comparaison of distributions

Since the measures of specialization and concentration are obtained from confronting distributions, first we consider the general topic of comparing two distributions of a categorical variable on the same universe $\{1, \cdots, i, \cdots, I\}$; let, more specifically, two distributions:

$$q_{\vec{i}} = (q_1, \cdots, q_i, \cdots, q_I) \tag{7}$$

$$r_{\vec{i}} = (r_1, \cdots, r_i, \cdots, r_I) \tag{8}$$

Two standard families of tools are available for comparing these two distributions: either a distance or a divergence. We use the term discrepancy to designate either one or the other and write $d(q_{\vec{i}} \mid r_{\vec{i}})$, with a subscript to identify particular specifications. In the present case, discrepancy functions are non-negative functions defined across all possible distributions, actually the $(I-1)-$dimensional simplex, taking the 0 value on the identity: $d(q_{\vec{i}} \mid q_{\vec{i}}) = 0$. The distance function is a symmetric function: $d(q_{\vec{i}} \mid r_{\vec{i}}) = d(r_{\vec{i}} \mid q_{\vec{i}})$ and satisfies the triangular inequality. The divergence function is not a symmetric function, *i.e.* $d(q_{\vec{i}} \mid r_{\vec{i}}) \neq d(r_{\vec{i}} \mid q_{\vec{i}})$; for this reason, when $d(\cdot, \cdot)$ is a divergence, we read $d(q_{\vec{i}} \mid r_{\vec{i}})$ as "the divergence of $q_{\vec{i}}$ with respect to $r_{\vec{i}}$"; when used in economic geography, $r_{\vec{i}}$ is typically considered as a "benchmark" distribution. Moreover, a divergence does not satisfied the triangular inequality and its geometric properties are derived from the properties of convex functions (for more details on f-divergence: see Csiszár 1967).

Many distance functions are available in the literature of probability theory. For our purposes, the most useful are:

$$d_H(q_{\vec{i}} \mid r_{\vec{i}}) = \frac{1}{2}\sum_i (\sqrt{q_i} - \sqrt{r_i})^2 \qquad \text{Hellinger-distance} \tag{9}$$

$$d^p_{L_p}(q_{\vec{i}} \mid r_{\vec{i}}) = \sum_i \mid q_i - r_i \mid^p \qquad L_p\text{-distances} \tag{10}$$

Hellinger-distance is valued in the interval $[0, 1]$ where the value 1 corresponds to mutually singular distributions (*i.e.* $q_i \; r_i = 0 \quad \forall i$); this property provides bounded measures of concentration. For the divergence functions, two cases are particularly relevant for the present field:

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

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

More information on distances and divergences between probability distributions may be found *e.g.* in Tjøstheim (1996), Gibbs and Su (2002) or Liese and Vajda (2006).

2.3 Three levels of comparison

The analysis of specialization and concentration may be operated at three different levels, namely:

1. A separate analysis of the spread of each sector specific $(p_{\vec{i}|j})$ and of each region specific $(p_{\vec{j}|i})$ profiles. For categorical variables, the spread of the frequency distribution may be viewed as a natural adaptation of the analysis of dispersion for a numerical variable, such as the

variance, to the concentration of categorical variables. A natural strategy compares the relevant distribution with a uniform distribution considered as a benchmark of minimal concentration, in the spirit of the pioneering works in information theory. In particular, the entropy may be viewed as a divergence with respect to the uniform distribution and the L_1-distance boils down to an average of absolute deviations. These analyses provide *absolute* measures in the sense of measures not depending on other regions or sectors. Note that the uniform distributions take the form $1/I$ for the regions and $1/J$ for the sectors. This benchmark for the non-concentration does not take into account the heterogeneity among regions, in terms of area or population. The same remark should be raised for the sectors, the "natural" sizes of which are typically quite different.

2. A separate analysis of each region, or each sector, comparing the relevant distribution with the corresponding marginal distribution as a benchmark of non-concentration, *i.e.* evaluating the discrepancies $d(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ and $d(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$. These comparisons take explicitly into account that the regions or the sectors are not uniformly distributed in the country under analysis and therefore provide measures that are *relative* to the overall structure of the country.

3. A *global* analysis of all the regions and sectors, by comparing the joint distribution, on regions × sectors, with the closest distribution reflecting independence, namely $p_{i\cdot}\, p_{\cdot j}$ taken as a benchmark of a completely non-concentrated, or non-specialized, country, *i.e.* the global analysis is focused on the discrepancies $d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$.

Table **??** in Appendix A summarizes these concepts. Their connections may be viewed as follows. *Absolute regional specialization* is a feature of the distribution of sectors across a region ($p_{\vec{j}|i}$), and a region is said to be absolutely specialized if a few sectors concentrate a large share of the region. This may be the case, for instance, when a sector is considerably larger than others at a country level. *Relative regional specialization* of a region shows up when an area has a greater proportion of a particular sector than the proportion of that sector in the whole territory. In other words, relative regional specialization compares an area share of a particular sector with the sector share at the country level, and is accordingly measured through a discrepancy $d(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$, thus relatively to the marginal distribution $p_{\cdot\vec{j}}$. The same comment can also be made for specific and relative industrial concentration.

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

2.4 Measures of specialization and of concentration

When defining degrees of global specialization, one possible strategy consists in defining first 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. Moreover, as the concept of stochastic independence is essentially symmetric between the two involved variables, the role of the regions and the sectors may be permuted.

A natural approach is a local one, more precisely to examine whether a cell (i,j) reveals over- or under-specialization, and aggregate over the complete table $\mathbf{N}$. The well-established *Hoover-Balassa Local Quotient* is designed to answer that question and may be equivalently defined for each cell (i,j) as follows:

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

The local quotient has been widely used in many different fields. The second and the third terms of (**??**) correspond to "relative risk" or "excess risk" in epidemiology. The fourth term corresponds to the usual "cross-product ratio" of the 2×2 sub-table constructed around N_{ij} and is a core tool in the statistical analysis of contingency tables. The last three terms express the same concepts through proportions, *i.e.* independently of $N_{\cdot\cdot}$ which represent the size of the country. The last two equalities in (**??**) emphasize that the specialization is an issue concerning the global structure at a country level: thus the absence of specialization of a cell (i,j) means that, relative to the distribution in the country, sector j is not over-(nor under-) represented in region i *and* that region i is not over-(nor under-) represented for sector j. Thus, "local" points to the fact that LQ is localized in a cell (i,j).

In the framework of stochastic independence, this local quotient reveals the following feature of sector j in region i:

$$\begin{array}{lllll} LQ_{ij} & = 1 & \text{or} & p_{ij} = p_{i\cdot}\; p_{\cdot j} & \text{no specialization} \\ & > 1 & \text{or} & p_{ij} > p_{i\cdot}\; p_{\cdot j} & \text{over-specialization} \\ & < 1 & \text{or} & p_{ij} < p_{i\cdot}\; p_{\cdot j} & \text{under-specialization} \end{array} \tag{14}$$

where "no-specialization" corresponds to the row-column independence. It should be clear from (**??**) that a discrepancy between the distributions $[p_{ij}]$ and $[p_{i\cdot}\; p_{\cdot j}]$ is equivalent to a discrepancy between the matrix $[LQ_{ij}]$ and a corresponding matrix of one's.

Among the most often used measures of independence between the rows and columns in the contingency table **N**, to be used as measures of global specialization, we shall focus on the following three:

$$\begin{aligned} d_{\chi^2}(\mathbf{N}) &= \sum_i \sum_j \frac{p_{i\cdot}(p_{j|i} - p_{\cdot j})^2}{p_{\cdot j}} = \sum_i \sum_j \frac{p_{\cdot j}(p_{i|j} - p_{i\cdot})^2}{p_{i\cdot}} \\ &= \sum_i \sum_j p_{i\cdot} p_{\cdot j}(LQ_{ij} - 1)^2 \qquad \chi^2 - \text{divergence, or inertia} \qquad (15) \\ d_{KL}(\mathbf{N}) &= \sum_i \sum_j p_{i\cdot}\, p_{j|i} \log\left(\frac{p_{j|i}}{p_{\cdot j}}\right) = \sum_i \sum_j p_{\cdot j}\, p_{i|j} \log\left(\frac{p_{i|j}}{p_{i\cdot}}\right) \\ &= \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, LQ_{ij} \log(LQ_{ij}) \qquad \text{Kullback-Leibler divergence} \qquad (16) \\ d_H(\mathbf{N}) &= \frac{1}{2} \sum_i \sum_j (\sqrt{p_{i\cdot}\, p_{j|i}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2 = \frac{1}{2} \sum_i \sum_j (\sqrt{p_{\cdot j}\, p_{i|j}} - \sqrt{p_{i\cdot} p_{\cdot j}})^2 \\ &= \frac{1}{2} \sum_i \sum_j p_{i\cdot}\, p_{\cdot j}\, (\sqrt{LQ_{ij}} - 1)^2 \qquad \text{Hellinger-distance} \qquad (17) \end{aligned}$$

Brülhart and Traeger (2005) call $d_{KL}(\mathbf{N})$ the "relative Theil index" (Theil 1967) and also propose generalizations by supplementing the original form with a so-called sensitivity parameter α. Aiginger and Davies (2004), Mulligan and Schmidt (2005), Bickenbach and Bode (2006 and 2008), Cutrini (2009), Alonso-Villar and del Río (2011) and many others have made use of the case $\alpha = 1$.

These measures deserve the following comments:

- As should be expected, these formulas display interchangeability between regions and sectors, congruently with the concept of stochastic independence.

- Because the stochastic independence approach operates with discrepancies among distributions, the induced measures takes the form of a double sum and may accordingly be decomposed as an average of the discrepancies between the conditional distributions and the corresponding marginal distributions. More specifically:

$$d_{\chi^2}(\mathbf{N}) = \sum_i p_{i\cdot} \left[\sum_j \frac{(p_{j|i} - p_{\cdot j})^2}{p_{\cdot j}}\right] = \sum_j p_{\cdot j} \left[\sum_i \frac{(p_{i|j} - p_{i\cdot})^2}{p_{i\cdot}}\right] \qquad (18)$$

$$d_{KL}(\mathbf{N}) = \sum_i p_{i\cdot} \left[\sum_j p_{j|i} \log\left(\frac{p_{j|i}}{p_{\cdot j}}\right)\right] = \sum_j p_{\cdot j} \left[\sum_i p_{i|j} \log\left(\frac{p_{i|j}}{p_{i\cdot}}\right)\right] \qquad (19)$$

$$d_H(\mathbf{N}) = \frac{1}{2} \sum_i p_{i\cdot} \left[\sum_j (\sqrt{p_{j|i}} - \sqrt{p_{\cdot j}})^2\right] = \frac{1}{2} \sum_j p_{\cdot j} \left[\sum_i (\sqrt{p_{i|j}} - \sqrt{p_{i\cdot}})^2\right] \qquad (20)$$

Thus these three measures of specialization accept a similar decomposition:

$$d_\omega(\mathbf{N}) = \sum_i p_{i\cdot}\, d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}) = \sum_j p_{\cdot j}\, d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}) \qquad \omega \in \{\chi^2, KL, H\} \qquad (21)$$

In other words, each of these global measures appears as an average of the relative regional specializations $d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$, or the relative localizations $d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$. Conversely, the properly weighted average of the relative regional specialization or the relative industrial concentration provide the *same* measure of global specialization. This is not due to the concepts of relative regional specialization or relative industrial concentration that, supposedly, should always evolve on the same track but rather to the structure of the measurement devices used, as remarked in Cutrini (2009). The realization of this fact has induced Bickenbach and Bode to call the global measures of specialization $d_\omega(\mathbf{N})$ measures of polarization in 2006 and of localization in 2008 and 2010.

Note. We use a slightly incoherent notation: $d_\omega(\mathbf{N})$ is a short-hand notation for $d_\omega([p_{ij}] \mid [p_{i\cdot}p_{\cdot j}])$ 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.

The discrepancies d_{χ^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:

$$d_{L_1}(\mathbf{N}) = \frac{1}{2}\sum_i\sum_j |p_{ij} - p_{i\cdot}\, p_{\cdot j}| = \frac{1}{2}\sum_i\sum_j p_{i\cdot}p_{\cdot j}\,|LQ_{ij} - 1| \tag{22}$$

$$= \frac{1}{2}\sum_i p_{i\cdot}\left[\sum_j |p_{j|i} - p_{\cdot j}|\right] = \frac{1}{2}\sum_j p_{\cdot j}\left[\sum_i |p_{i|j} - p_{i\cdot}|\right] \tag{23}$$

This distance is 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 (see for *e.g.* Krugman 1991a, Hallet 2000, Midelfart-Knarvik, Overman, Redding and Venables 2000, Mulligan and Schmidt 2005, *et al.*). Moreover, $d_{L_1}(\mathbf{N})$ shows the same representations, in terms of local quotients, and the same decompositions as in (**??**). 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). Since, similarly to d_H, its range of variation is also bounded by 1, it has not been added to our numerical evaluations in order to control the length of this paper.

3 An overview of the literature at the light of the stochastic independence approach

3.1 On the concepts and their measures

The NEG has proposed a wealth of measures of industrial concentration, regional specialization or global specialization, not infrequently out of the present proposal. Moreover, such literature is vast and the use of words is not completely standardized. One reason for this state of affairs is related to the nature of the primary units underlying the N_{ij}'s. According to the main interest of an investigation, these may be the number of employees, firms or establishments. It should therefore be expected that different interests lead to adopt different wordings even if the formal problem is identical. Thus, the conventional aspect of Table **??** is to present a convention that links different fields of interests using a unified methodology. This section is not aimed at producing an exhaustive glossary of terms used in a rapidly growing literature but rather at pinpointing major developments of the basic concepts in the field of industrial concentration and regional specialization.

The diversity of terms also relates to the multifaceted nature of the concepts in use. The NEG models explaining specialization originated mainly in trade theory, while models explaining concentration came from location theory. Here the distinction between absolute and relative concepts becomes particularly relevant. Thus, when Cutrini (2009) asserts that industrial concentration and regional specialization are "the two sides of the same medal" she refers to the global concepts; indeed (**??**) shows that the measure of the global specialization is equivalently a (suitably) weighted average of either the relative industrial concentrations or of the relative regional specializations. However, using concepts not based on the stochastic independence approach Krugman (1991a, b) develops a simulation model that produces a U-shaped relationship between changes in transport costs and specialization or concentration; Aiginger and Rossi-Hansberg (2006) discusses the basic setup of the model developed in Rossi-Hansberg (2005) and finds that specialization and concentration in fact go in opposite directions when transport costs change, in particular, lower transport costs imply higher specialization and lower concentration. Aiginger and Davies (2004) and Mulligan and Schmidt (2005) also find that absolute concepts may produce diverging evolution of industrial concentration and regional specialization.

"Polarization", used by Perroux (1950) among others (see *e.g.* Bickenbach and Bode 2006 and 2008), instead of our global specialization, has also been used for the analysis of agglomeration. Starting from a location model, Ellison and Glaeser (1997) define a concept of agglomeration not relying on spatial autocorrelation of the regions and which is viewed as a concept of relative industrial concentration; this work measures the extent of relative industrial concentration once the size of establishments (based on the Herfindahl index) and the inherent randomness in the concentration of firms are accounted for. Also, Maurel and Sédillot (1999), Devereux, Griffith and Simpson (2004), Guimarães, Figueiredo and Woodward (2007) developed new indexes following this approach.

Mori, Nishikimi and Smith (2005) have proposed a relative measure of regional specialization, based on KL-divergence, where the benchmark is given by the area of the regions rather than by the corresponding marginal distribution; these authors call it a "D-index" whereas Brülhart and Traeger (2005) call a similar index "topographic Theil index", and mention that their "relative Theil index" and their "topographic Theil index" are equivalent to the distinction in spatial statistics between "heterogeneous" and "homogenous" space as in Marcon and Puech (2003).

3.2 On the family of Gini and Krugman coefficients

A large class of the proposals in the NEG literature are based on Lorenz curves (Lorenz 1905) and Gini indices (see for *e.g.* Krugman 1991a, Kim 1995, Amiti 1999, 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, and many others). In Appendix C, details are given on a Gini index of relative regional specialization GI_i and of relative industrial concentration GI^j. Appendix C also includes details on another class of indices based on absolute deviations, as 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:

$$GI_{reg} = \sum_i p_{i\cdot}\, GI_i \qquad\qquad SK_{reg} = \sum_i p_{i\cdot}\, SK_i \tag{24}$$

or on the relative industrial concentration

$$GI^{sec} = \sum_j p_{\cdot j}\, GI^j \qquad\qquad SK^{sec} = \sum_j p_{\cdot j}\, SK^j \tag{25}$$

The Lorenz curve is a graphical representation of the spread of a distribution derived from cumulative distribution functions (see Appendix C) and raise several difficulties when used to represent industrial concentration or regional specialization, in particular: i) the Lorenz curve concern univariate distributions of a numerical, or at least ordered, variables whereas the problems of industrial concentration and of regional specialization concern a two-way contingency table, *i.e.* bivariate categorical variables. The adaptation of the Lorenz curve, and Gini's index, to the case of categorical variables, such as sector 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 each sector; these different orderings make the interpretation of the average Gini's coefficient difficult; ii) the global index GI_{reg} has been called a specialization coefficient where GI^{sec} has been called a coefficient of industrial concentration. Because these measures are not developed in the symmetric framework of stochastic independence, the global measures based on regions or on sectors do not coincide:

$$GI_{reg} \neq GI^{sec} \qquad\qquad SK_{reg} \neq SK^{sec} \tag{26}$$

The fact that in general $GI_{reg} \neq GI^{sec}$ (for a numerical illustration, see subsections **??** and **??**) raises an issue of interpretation,whereas the approach of stochastic independence considers the regions and the sectors interchangeably. iii) 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 Lerman and Yitzhaki (1989).

4 Application to Argentine data

4.1 Scope of this application

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

We want to examine different issues. First, when considering the profiles of the sectors, or of the regions, relative to their corresponding marginal (country-wide) distribution, to what extent are associated the measures of relative concentration, or relative specialization? This question may be answered through a graphic representation of these measures or through the evaluation of the correlations among them. And this question raises another one. These measures are subject to different ranges of variation: the unit interval for d_H or bounded intervals for d_{χ^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 linear or non-linear associations.

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

It should be emphasized but these issues basically refer to the interpretation of the numerical values of these measures and their comparability among different sectors, different regions or different countries. Moreover, we also want to compare the numerical behaviors of d_H, d_{χ^2} and d_{KL} with those of Gini and of Krugman indices.

4.2 The data

The original data is concerned with the employees in the manufacturing sector and are obtained from of the Economic Census performed 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 with no employees in the manufacturing sector).

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

The final data used in this application are obtained after regrouping the 22 sectors into 17 and the 462 regions into 35. Regrouping was made from an automatic grouping procedure on large two-way contingency tables based on hierarchical clustering and correspondence analysis (HCCA), aimed at obtaining 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).

4.3 Findings

Tables **??** and **??** show the 35×17 contingency table $\mathbf{N}$ of the data along with the row and column totals $N_{i\cdot}$, $N_{\cdot j}$ with their proportions. We complete the table by providing region and sector measures of relative regional specialization $d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$, and relative industrial concentration $d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$, and finish the table with the global measures of specialization $d_\omega(\mathbf{N})$, where $\omega \in \{\chi^2, KL, H\}$.

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

$$d_{\chi^2}(\mathbf{N}) = 1.6532; \qquad d_{KL}(\mathbf{N}) = 0.3176; \qquad d_H(\mathbf{N}) = 0.0713. \tag{27}$$

As they are measured on different scales, their numerical values are difficult to interpret except for 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:

$$GI_{reg} = 0.3262; \qquad GI^{sec} = 0.3495; \qquad SK_{reg} = 0.2963; \qquad SK^{sec} = 0.3041. \tag{28}$$

As is confirmed in the sequel, d_H systematically gives a lower value of global specialization. Also, region-based and sector-based numerical values are but slightly different. Moreover, Gini's and Krugman's coefficients also take different though very similar values.

In order to compare the numerical values of all indices, a possible solution could be to 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 $1 - \textit{critical alpha}$ as a comparable measure of association.

We do not follow this path because we deem it inappropriate for the later developments, and rather take alternative ways. 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 close to that of d_{KL}. Transformed measures then become

$$\log(1 + d_{\chi^2}(\mathbf{N})) = 0.4238; \qquad d_{KL}(\mathbf{N}) = 0.3176; \qquad 4d_H(\mathbf{N}) = 0.2852. \tag{29}$$

These transformed measures bear close but not identical values and suggest a low level of specialization in Argentina, in view of the value of d_H. In subsection **??** we discuss the relative position of Argentina with respect to other countries. The transformation (**??**) ensures a similar range, namely around the interval [0 4] for the three measures; but this interval is only approximately true. In particular, it is known that the maximum value of d_{χ^2} depends on both I and J. Cramer (1946) shows that the maximum possible for d_{χ^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, consists on normalizing d_{χ^2} and d_{KL} to the interval [0 1], just as d_H. Any strictly increasing function $\mathbf{R}_+ \to [0\ 1]$ may do the job, but the simplest one might be:

$$Nd_{\chi^2} = \frac{d_{\chi^2}}{d_{\chi^2} + 1}; \qquad Nd_{KL} = \frac{d_{KL}}{d_{KL} + 1}. \tag{30}$$

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

Let us have a closer look at the decomposition of the global measure into sector-specific and region-specific measures according to (**??**), as given in Tables **??** and **??**. In Figure **??** and Figure **??** respectively (in Appendix B), we have ranked the 17 sectors, and the 35 regions, in ascending order of d_H, and plotted together the three transformed measures.

Two features should be noticed:

- the numerical values of the three modified measures display low dispersion for values under 1 but higher dispersion otherwise, for the region-specific as well as for the sector relative measures;
- the ranking between regions, or sectors, is modified each time one of the curves displays a descending piece; clearly the three rankings are similar although some discrepancies are noticeable. These discrepancies show low as well as high values for the measures. Further on we return to the issue of the ranking stability.

Let us have a look at the graphic behaviour of the normalized measures Nd_{χ^2} and Nd_{KL} compared to GI and SK, relative to d_H. They are shown in Figure **??** to illustrate the relative industrial concentration and in Figure **??** for the relative regional specialization. These curves take their values in the unit interval. These figures correspond to Figures **??** and **??** that are related to the three transformed measures. Now we notice that these five measures show roughly a similar behaviour although Nd_{χ^2} is the least similar. Moreover, the curves relative to the industrial concentrations in Figure **??** show more coherence than those related to the regional specializations in Figure **??**.

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_{χ^2}, d_{KL} and d_H), the transformed $log(1+d_{\chi^2})$, the normalized version Nd_{χ^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 **??**) and pairwise scatter diagrams (Figure **??**). Next we perform a similar analysis on the ranks in Table **??** and **??**. These tables and figures provide the results on regional specialization under the main diagonal and the results on industrial concentration above the main diagonal.

For each instance, we also provide the correlations 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 Table **??** and Figure **??** however, show that in absolute values their association is the weakest one for d_{χ^2} but the strongest one for Nd_{χ^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, however, show that their association is globally weak, in particular because the largest regions and the largest sectors are essentially outlying data for this association.

Let us now examine 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_{χ^2} or Nd_{KL} tend to substantially increase those correlations although 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}.

It should also be noticed 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 **??**, 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_{χ^2} with GI is noteworthy.

One last aspect should also be checked, namely the stability of the ranking. This aspect may be viewed as a non-parametric approach (see also Slottje 1990). This is examined in Table **??** by means of the Spearman's rank coefficient and in Figure **??** by means of scatter diagrams among ranks. Here, the rows and columns relative to $log(1+d_{\chi^2})$ are redundant compared to those related to d_{χ^2}. The same redundancy is also true for normalized versions of d_{χ^2} and d_{KL}. Again, the correlations in Table **??** are uniformly high and the correlations related to the marginal profiles are higher than in Table **??**. But now the behaviour among the ranks corresponding to the sectors (above the main diagonal) have less associations than those related 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 case viewed as objective and final measures of specialization. Finally, some peculiarities of d_{χ^2} might be attributed to the fact that d_{χ^2} is based on squared differences that tend to overweight extreme cases, while this feature is mitigated by the log transformation.

4.4 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 from the local government entities at a 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.

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 with no employees in the manufacturing sector) are 462, 5,138 and 249 for Argentina, Brazil and Chile, respectively. It should be noted that both regions of Brazil and Chile refer to the local government entity while those of Argentina refer to the catastral divisions. Thus, from an administrative point of view, Argentina's divisions cannot be directly compared to those of Brazil and Chile, although in some cases their boundaries match those of the municipalities.

The data related to employment in the manufacturing sector were obtained from of the National 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 for Chile refer to firms with 5 or more employees. As in Section **??**, sector classifications refer to the first 2 digits of the International Standard Industrial Classification (ISIC Rev.3.1) of manufacturing sectors (22 sectors after grouping the divisions 36 and 37).

Table **??** shows a summary of the results obtained from the proposed measures of global specialization and the number of cells of each contingency table. 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: this ranking (derived from a visual examination, since the correlation coefficients for the data of the 3 countries do not make sense) does not depend on the selected measure of global specialization nor on the number of cells. There is extensive literature on the comparison of contingency tables with different sizes (see *e.g.* Lauritzen 1989, van der Heijden, Mooijaart and Takane 1994, and Agresti

2002), while present results are found to be 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^{sec}, and SK^{sec} operate on a same range of variation, namely the unit interval, we systematically observe the same order, namely $d_H < SK_{reg} < SK^{sec} < GI_{reg} < GI^{sec}$, with rather substantial differences within these measures related to each country. It should also be noticed that the ranking of the three countries for each measure *and* the ranking among the five measures for each country, remain exactly the same. Comparing these results with those of subsection **??**, we observe that Gini's coefficients are systematically higher than Krugman's coefficients and that in both cases sector-based coefficients are higher than, but close to, region-based coefficients. It must also be noted 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 **??** is a collapsed table of that used in this application, implying a loss of information that will be considered in the next section.

5 Grouping of regions or sectors

5.1 Grouping of regions

Let us operate a partition of I regions into M "grouped regions", that will be called "g-regions" for the sake of of clarity. Thus:

$$\mathcal{I} \,=\, \{1,2,\cdots,I\} \,=\, \bigcup_{m=1}^{M} \mathcal{I}_m \qquad \mathcal{I}_m \cap \mathcal{I}_{m'} \,=\, \emptyset \;\; (m \neq m') \qquad \#(\mathcal{I}_m) \,=\, I_m \qquad \sum_m I_m \,=\, I \quad (31)$$

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

$$q_{m\cdot} \,=\, \sum_{i\in\mathcal{I}_m} p_{i\cdot} \qquad q_{m|j} \,=\, \sum_{i\in\mathcal{I}_m} p_{i|j} \tag{32}$$

$$q_{\vec{m}\cdot} \,=\, (q_{1\cdot},\cdots,q_{m\cdot},\cdots,q_{M\cdot}) \qquad q_{\vec{m}|j} \,=\, (q_{1|j},\cdots,q_{m|j},\cdots,q_{M|j}) \tag{33}$$

Furthermore:

$$p_{i|m} \,=\, \frac{p_{i\cdot}}{q_{m\cdot}}\, \mathbb{1}_{\{i\in\mathcal{I}_m\}} \qquad p_{i|j,m} \,=\, \frac{p_{i|j}}{q_{m|j}}\, \mathbb{1}_{\{i\in\mathcal{I}_m\}} \tag{34}$$

The KL-divergence has a characteristic feature, namely to accept a decomposition related to grouping rows or columns, which outcome is similar to a decomposition of the variance resulting from 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), Cutrini (2009), Alonso-Villar and del Río (2011), among others.

Indeed, if we start with the second term of (??), we will subsequently obtain:

$$
\begin{align}
d_{KL}(\mathbf{N}) &= \sum_j p_{\cdot j} \left[\sum_i p_{i|j} \log \left(\frac{p_{i|j}}{p_{i\cdot}} \right) \right] \tag{35} \\
&= \sum_j p_{\cdot j} \left[\sum_m q_{m|j} \log \frac{q_{m|j}}{q_{m\cdot}} \left\{ \sum_{i \in \mathcal{I}_m} p_{i|j,m} \right\} + \sum_m q_{m|j} \left\{ \sum_{i \in \mathcal{I}_m} p_{i|j,m} \log \frac{p_{i|j,m}}{p_{i|m}} \right\} \right] \notag \\
&= \sum_j p_{\cdot j} \left[d_{KL}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j} \; d_{KL}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] \tag{36}
\end{align}
$$

In (??), as a general result, the KL-measure of specialization is viewed as a weighted average of industrial concentration, namely $d_{KL}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ in (??), whereas in (??) each sector measure is decomposed with respect 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 specific sector measures of the specializations among 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 sector specific measures of the specializations among the composing regions of each g-regions;
- *Global = Between + Within.*

Two polar cases are of interest. First, let's suppose that $M = 1$, *i.e.* that all the regions in the country are grouped into a unique g-region, namely the country. 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 disapears 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 $d_{KL}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot})$ matches $d_{KL}(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$ in (??).

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) measures how much an aggregation decrease the specialization. Heuristically, the ratio (Between/Global) may be seen as a measure of association between specialization and the aggregation criterion. It must be noted that in the extreme 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 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. Let us compare the effects of two nested partitions. Thus let us consider the partition given in (**??**) along with a finer partition:

$$\mathcal{I} = \{1, 2, \cdots I\} = \bigcup_{m'=1}^{M'} \mathcal{I}_{m'} \qquad \mathcal{I}_{m'_1} \cap \mathcal{I}_{m'_2} = \emptyset \ \ (m'_1 \neq m'_2) \qquad \#(\mathcal{I}_{m'}) = I_{m'} \qquad \sum_{m'} I_{m'} = I$$
$$M < M' \quad \forall m' \, \exists m : \mathcal{I}_{m'} \subset \mathcal{I}_m \tag{37}$$

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

5.2 Grouping of sectors

The same analysis can be applied when grouping sectors instead of regions. Hence, we will now consider a partition of the sectors into L g-sectors:

$$\mathcal{J} = \{1, 2, \cdots, J\} = \bigcup_{l=1}^{L} \mathcal{J}_l \qquad \mathcal{J}_l \cap \mathcal{J}_{l'} = \emptyset \ \ (l \neq l') \qquad \#(\mathcal{J}_l) = J_l \qquad \sum_{l} J_l = J \tag{38}$$

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

$$r_{\cdot l} = \sum_{j \in \mathcal{J}_l} p_{\cdot j} \qquad r_{l|i} = \sum_{j \in \mathcal{J}_l} p_{j|i} \tag{39}$$

$$r_{\cdot \vec{l}} = (r_{\cdot 1}, \cdots, r_{\cdot l}, \cdots, r_{\cdot L}) \qquad r_{\vec{l}|i} = (r_{1|i}, \cdots, r_{l|i}, \cdots, r_{L|i}) \tag{40}$$

Furthermore:

$$p_{j|l} = \frac{p_{\cdot j}}{r_{\cdot l}} \, \mathbb{1}_{\{j \in \mathcal{J}_l\}} \qquad p_{j|i,l} = \frac{p_{j|i}}{r_{l|i}} \, \mathbb{1}_{\{j \in \mathcal{J}_l\}} \tag{41}$$

We may now repeat the decomposition of the KL-measure of specialization related to a grouping of sectors. Indeed, starting with the first term of (**??**), we successively obtain:

$$d_{KL}(\mathbf{N}) = \sum_i p_{i\cdot} \left[\sum_j p_{j|i} \log \left(\frac{p_{j|i}}{p_{\cdot j}} \right) \right] \tag{42}$$

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

$$= \sum_i p_{i\cdot} \left[d_{KL}(r_{\vec{l}|i} \mid r_{\cdot \vec{l}}) + \sum_l r_{l|i} \, d_{KL}(p_{\vec{j}|i,l} \mid p_{\vec{i}|l}) \right] \tag{43}$$

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

As a final remark, aggregating regions into larger ones, or aggregating sectors, 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 the aggregation, the lower the specialization.

Measures d_{χ^2} and d_H accept the same decomposition related to a grouping of the regions (rows) or of the sectors (columns), but unlike $d_{KL}(\mathbf{N})$ their decompositions are not exact and show residuals to be denoted as $R_{\chi^2}(\mathbf{N})$ and $R_H(\mathbf{N})$, respectively. Thus, for the decomposition related to to a grouping of regions, we obtain:

$$d_{\chi^2}(\mathbf{N}) = \sum_j p_{\cdot j} \left[d_{\chi^2}(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j} \, d_{\chi^2}(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] + R_{\chi^2}(\mathbf{N}) \tag{44}$$

$$d_H(\mathbf{N}) = \sum_j p_{\cdot j} \left[d_H(q_{\vec{m}|j} \mid q_{\vec{m}\cdot}) + \sum_m q_{m|j} \, d_H(p_{\vec{i}|j,m} \mid p_{\vec{i}|m}) \right] + R_H(\mathbf{N}) \tag{45}$$

And similarly for a sector grouping.

5.3 Grouping argentinean regions and sectors

We now examine numerically the impact of grouping regions and/or sectors. Initially, a natural question is raised: is the impact of these groupings on the degree of specialization similar for the three global measures?

We use the same data as in section **??** and analyze the impact, on global specialization, of regrouping regions or sectors, by evaluating numerically the terms of the decomposition (**??**), (**??**) and (**??**), and the corresponding terms for the sectors.

We first consider an arbitrary aggregation of regions by assembling the first 10 regions into a single 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_{χ^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_{χ^2}, d_{KL} or d_H?

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

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

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 in lines 1 and 2.

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_{χ^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_{χ^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, as shown in Table **??**, the aggregation has been operated on fairly homogenous and 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 smaller.

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 0.2262 to 0.6971. Together these 10 regions represent only 2.69% of total employment. We now observe in Table **??**, that the residual part is considerably bigger than in Table **??**, raising from 0.76% to 21.91% for d_{χ^2} and from 0.60% to 3.72% for d_H, with the same sign as in Table **??**. The share of the between term, in line 8, increases considerably for the three measures from around 70% to around 90%. Notice however that for d_{χ^2} the between term decreases but its share, taking into account the inflated residual term, has increased. These results show that aggregating small regions into a unique one mildly affects the global level of specialization, at variance from aggregating large regions.

We now consider an arbitrary aggregation of sectors by assembling the first 5 sectors into a single one (representing 40.92% of the global employment), leaving the other sectors as singletons in the aggregated partition. The results are presented in Table **??** in the same format as in Table **??**. 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_{χ^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 related to d_{KL} has an intermediate value between those related to d_{χ^2} and to d_H, once the effect of the residual term has been taken into account, *i.e.* line 8 rather than 7.

We now turn, in Table **??**, to another arbitrary partition of the sectors, by regrouping the last 5 sectors. 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 **??** and **??**. For d_{χ^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

?? and **??** show that when regrouping 5 smaller sectors, representing 13% of the total employment, the global measure of specialization is less affected than by regrouping 5 larger sectors, representing 41% of the total employment.

If we examine the 4 regrouping exercises we find that:

- the share of the residual terms are always substantially lower for d_H than for d_{χ^2}; moreover, the residuals of d_{χ^2}show a substantially higher variability than those of d_H;
- the sign of the residual terms is systematically positive for d_{χ^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 sectors, than when aggregating smaller ones.

6 Discussions and conclusions

6.1 The stochastic independence approach in a nutshell

Based on data in the form of a two-way contingency table "Regions × Sectors", 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 to measure the degree of specializations are provided by discrepancies $d_\omega(\cdot \mid \cdot)$, more precisely distances or divergences, among distributions: between profiles and a uniform distribution for absolute concepts $(d_\omega(p_{\vec{j}|i} \mid [a_i = I^{-1}])$ or $d_\omega(p_{\vec{i}|j} \mid [b_j = J^{-1}]))$ that represent the spread of a distribution on categorical variables, between profiles and the corresponding marginal distribution $(d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$ or $d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}))$ for the relative concepts, or between the joint distribution and the product of the marginal distributions $d_\omega([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$ for the global concept. This is the approach of stochastic independence that governs the analysis in terms of stochastic independence between sectors and regions, while the global discrepancy is viewed as a measure of row-column association.

The relative and 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 sectors 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 sector specific or region specific measures of specializations, to eventually obtain different global measures of specialization. Such is the case for Gini's and Krugman's indices. The KL-measures enjoy a suitable decomposition with respect to regrouping. The residual terms of the other measures make more difficult an evaluation of the impact of regrouping, particularly in the case where the residual term is substantial .

6.2 Final remarks

This paper advocates the reference to a single framework for the study of regional specialization and of industrial concentration, namely a systematic reliance on discrepancies among frequency distributions. This integrating framework is particularly relevant when trying to cope with the substantial variety of approaches and of choices of words in spatial economy or economic geography.

The choice of a particular family of discrepancies is not the main object of this paper, but as a hint for the practitioner three of them, namely χ^2, KL and H, have been explored through numerical illustrations, and prove to be reasonably coherent in terms of the substantial conclusions to be drawn when they are simultaneously evaluated on several data set. However, it should be pointed out that they have not always provided the same ranking among countries nor the same impact of groupings.

The analysis of regional specialization and industrial concentration quite often involve contingency tables with an extreme heterogeneity of sizes of cells (i, j) or of 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. This heterogeneity raises issues related to the robustness of the measurement and the interpretation of international comparisons. The illustrations of subsection **??** and the Section **??** on the impact of grouping provide hints on questions that quite clearly deserve further attention.

Instead of developing a single integrating framework, an alternative road, not pursued in this paper, would be to describe general properties potentially attractive for measurements, possibly by axiomatizing these properties. Bollen and Long (1993) summarizes a number of desirable properties for discrepancy measures but recognizes that no single measure meets them all; moreover not all researchers would even agree with all of these properties. In their search for desirable properties, Combes and Overman (2004) emphasizes the analysis of the deviation from a benchmark distribution. Alonso-Villar and del Río (2011) mentioned some of the main properties we also meet when using discrepancies measures.

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 a log-linear model for a 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-sector association.

The systematically negative correlation between the size of the region, and of the sector, and the measures of specialization has been noticed in the application of Section **??**. This observation provides a first hint on the impact of grouping on specialization and concentration. As already noticed in 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), and refers to the role of the geographical partition in use (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. The same issue is also raised in the case of sector aggregation. Therefore, the arbitrariness of partitions plays a key role in capturing the effects mentioned previously, and becomes potentially more dangerous the more unequal become its elements. 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. The analysis of groupings in Section **??** maybe viewed as a first hint for obtaining a better grasp of the consequences of grouping regions or sectors in the MAUP problem.

References

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

Aiginger, K., and Davies, S. (2004), Industrial specialization and geographic concentration: two sides of the same coin? Not for the European Union. *Journal of Applied Economics* **7**: 231-248.

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

Alonso-Villar, O., and del Río, C. (2011), Concentration of economic activity: an analytical framework. *Regional Studies.* DOI: 10.1080/00343404.2011.587796.

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 sectors on continuous space. *Papers in Regional Science* **80**: 411-424.

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

Arbia, G., Espa, G., and Quah, D. (2007), A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. Department of Economics, University of Trento, Working Paper 5.

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

Barff, R. (1987), Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio. *Annals of the Association of American Geographers* **77**: 89-103.

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

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

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

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.

Combes, P., and Overman, H. (2004), The spatial distribution of economic activities in the EU. In J.V. Henderson and J.-F. Thisse (eds.), *Handbook of Regional and Urban Economics 4*, Elsevier-North Holland.

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.

Cutrini, E. (2009), Using entropy measures to disentangle regional from national localization patterns. *Regional Science and Urban Economics* **39**: 243-250.

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.

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

Devereux, M., Griffith, R., and Simpson, H. (2004), The geographic distribution of production activity in the UK. *Regional Science and Urban Economics* **34**: 533-564.

Ellison, G., and Glaeser, E. (1997), Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economic* **105**: 889-939.

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.

Guimarães, P., Figueiredo, O., and Woodward, D. (2007), Measuring the localization of economic activity: a parametric approach. *Journal of Regional Science* **47**: 753-774.

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 141.

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 sectors: the trends in US regional manufacturing structure, 1860-1987. *The Quarterly Journal of Economics* **110**: 881-908.

Kosfeld, R., Eckey, H., and Lauridsen, J. (2011), Spatial point pattern analysis and industry concentration. *The Annals of Regional Science* **2**: 311-328. DOI: 10.1007/s00168-010-0385-5.

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 91.

Lauritzen, S. (1989), *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.

Liese, F., and Vajda, I. (2006), On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory* **52**: 4394-4412. DOI:10.1109/TIT.2006.881731.

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

Marcon, E., and Puech, F. (2003), Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography* **3**: 409-428.

Maurel, F., and Sédillot, B. (1999), A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics* **29**: 575-604.

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

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

Mulligan, G., and Schmidt, C. (2005), A note on localization and specialization. *Growth and Change* **36**: 565-576.

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.

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

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.

Sheldon, E.H.B., and Moore, W.E. (eds.) (3rd ed.:1972), *Indicators of Social Change: Concepts and Measurements.* New York: Russel Sage Foundation.

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. non-parametric 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.

ZELLER, R., AND CARMINES, E. (1980), *Measurement in the Social Sciences. The Link between Theory and Data.* New York: Cambridge University Press.

Appendix A: Tables

Table 1: *Some conventional definitions*

Technique	**Measured concept**
$d(p_{\vec{j}\|i} \mid [1/J])$	Absolute regional specialization
$d(p_{\vec{i}\|j} \mid [1/I])$	Absolute industrial concentration
$d(p_{\vec{j}\|i} \mid p_{\cdot\vec{j}})$	Relative regional specialization
$d(p_{\vec{i}\|j} \mid p_{\vec{i}\cdot})$	Relative industrial concentration
$d([p_{ij}] \mid [p_{i\cdot}\, p_{\cdot j}])$	Global specialization

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

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

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

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

[3] Not significant

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

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

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

[3] Not significant

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^{sec}	0.4880	0.5925	0.6358
SK_{reg}	0.3625	0.4521	0.5079
SK^{sec}	0.3980	0.4856	0.5897
#of cells	10,164 (462x22)	113,036 (5,138x22)	5,478 (249x22)

Table 5: *Arbitrary grouping of 10 first regions*

N°	Item	d_{χ^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

Table 6: *Arbitrary grouping of 10 last regions*

N°	Item	d_{χ^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

Table 7: *Arbitrary grouping of 5 first sectors*

N°	Item	d_{χ^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

Table 8: *Arbitrary grouping of the 5 last sectors*

N°	Item	d_{χ^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

Table 9: *Argentine data (1)*

sector Region	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: *Argentine data (2)*

sector Region	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

Appendix B: Figures

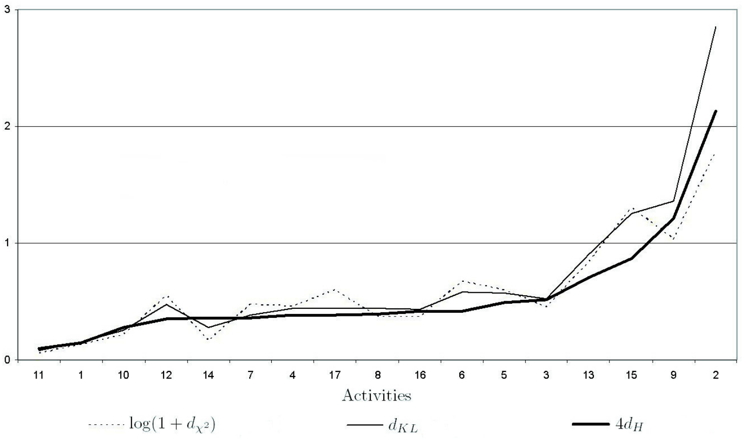

Figure 1: *Level of relative industrial concentration* $(d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot}))$ *of transformed measures*

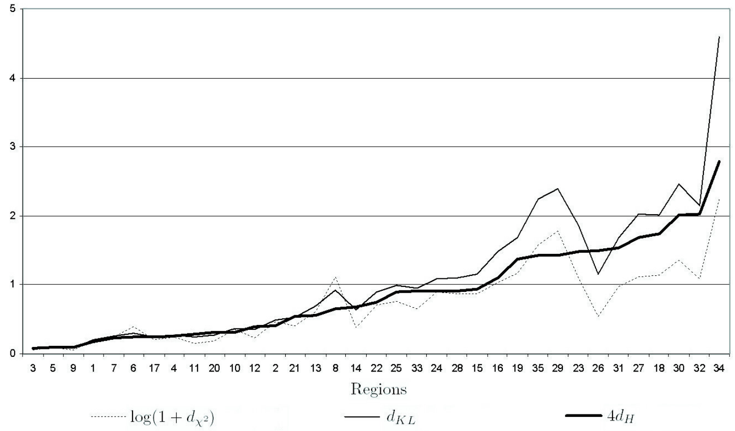

Figure 2: *Level of relative regional specialization* $(d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}}))$ *of transformed measures*

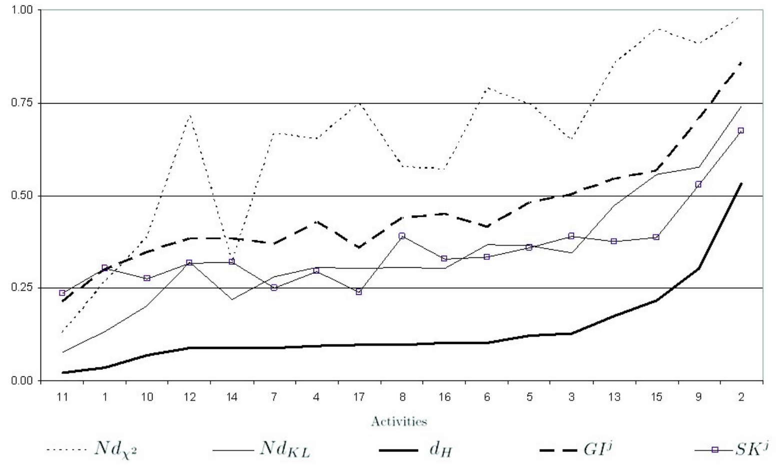

Figure 3: *Level of relative industrial concentration* ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}.})$) *of normalized measures*

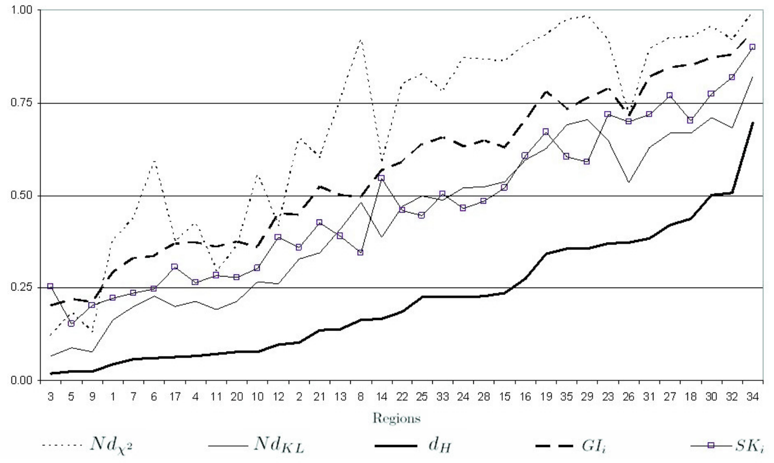

Figure 4: *Level of relative regional specialization* ($d_\omega(p_{\vec{j}|i} \mid p_{.\vec{j}})$) *of normalized measures*

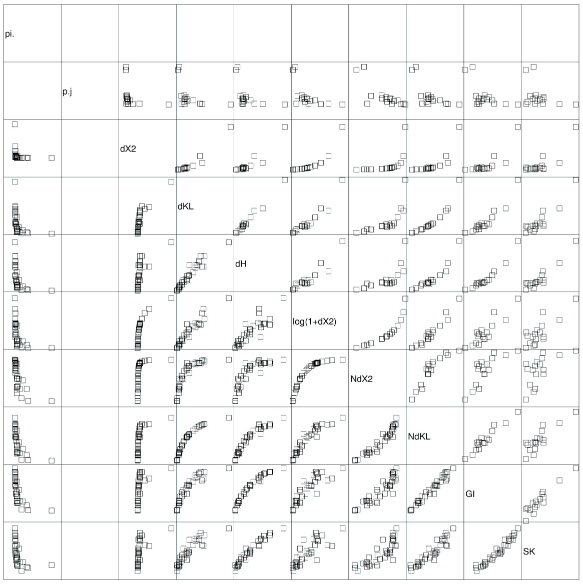

Figure 5: *Dispersion between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$-above the main diagonal) measures (I=35, J=17)*

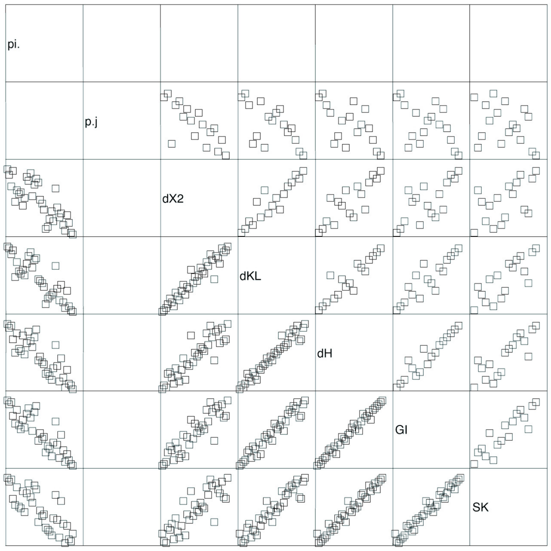

Figure 6: *Ranking dispersion between relative regional specialization ($d_\omega(p_{\vec{j}|i} \mid p_{\cdot\vec{j}})$-under the main diagonal) and between relative industrial concentrations ($d_\omega(p_{\vec{i}|j} \mid p_{\vec{i}\cdot})$-above the main diagonal) measures (I=35, J=17)*

Appendix C: On 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 of x such that the proportion of individuals with a value of X lower or equal to x is the same as that 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 the Gini index (Gini 1912) for the characterization of the relative regional specialization: the simultaneity of two dimensions, namely region and sector, 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 sectors may be arranged according to the increasing order of the local quotient:

$$LQ_{i,j_i(1)} < LQ_{i,j_i(2)} < ... < LQ_{i,j_i(k)} < ... < LQ_{i,j_i(J)} \tag{46}$$

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 sectors that, *in region* i, have a local quotient lower or equal to that of the k-th sector upon the ordering given in (??), and $P^{(i)}_{j_i(k)|i}$ represents the similar proportion, now relatively to the region i only. We now construct a curve for region i, 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 sector 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 (??). The actual curve may accordingly 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 (the 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. ?? where area $\alpha = [P^{(i)}_{\cdot k} - P^{(i)}_{\cdot k-1}] \times \frac{1}{2}[P^{(i)}_{k|i} + P^{(i)}_{k-1|i}]$).

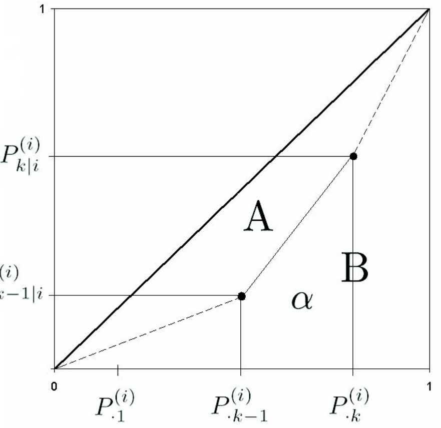

Figure 7: *Lorenz curve for specialization*

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

$$GI_i = 1 - \sum_{1 \leq k \leq J} \left(P^{(i)}_{\cdot k} - P^{(i)}_{\cdot k-1}\right)\left(P^{(i)}_{k|i} + P^{(i)}_{k-1|i}\right) \tag{47}$$

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 sector 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 sector, the share of which is infinitely small in the country.

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

$$GI^j = 1 - \sum_{1 \leq r \leq I} \left(P^{(j)}_{r\cdot} - P^{(j)}_{r-1\cdot}\right)\left(P^{(j)}_{r|j} + P^{(j)}_{r-1|j}\right) \tag{48}$$

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

$$LQ_{i_j(1),j} \; < \; LQ_{i_j(2),j} \; < ... \; < \; LQ_{i_j(r),j} \; < ... \; < \; LQ_{i_j(I),j}. \tag{49}$$

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 sector structure of region i and the average of the sector j structure of the other regions. It is defined as:

$$SK_i = \frac{1}{2} \sum_j \mid p_{j|i} - \overline{p_{\cdot j}} \mid \tag{50}$$

where

$$\overline{p_{\cdot j}} = \frac{\sum_{m \neq i}^{I} N_{mj}}{\sum_{m \neq i}^{I} \sum_j N_{mj}} \tag{51}$$

The SK_i index takes a zero value if the sector 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 sector structure of one region differs completely from the rest of the country.

The index for relative industrial concentration is constructed similarly:

$$SK^j = \frac{1}{2} \sum_i \mid p_{i|j} - \overline{p_{i\cdot}} \mid \tag{52}$$

where

$$\overline{p_{i\cdot}} = \frac{\sum_{l \neq j}^{J} N_{il}}{\sum_i \sum_{l \neq j}^{J} N_{il}} \tag{53}$$

Recent titles

CORE Discussion Papers

2011/53. Philippe DE DONDER and Pierre PESTIEAU. Private, social and self insurance for long-term care: a political economy analysis.

2011/54. Filippo L. CALCIANO. Oligopolistic competition with general complementarities.

2011/55. Luc BAUWENS, Arnaud DUFAYS and Bruno DE BACKER. Estimating and forecasting structural breaks in financial time series.

2011/56. Pau OLIVELLA and Fred SCHROYEN. Multidimensional screening in a monopolistic insurance market.

2011/57. Knud J. MUNK. Optimal taxation in the presence of a congested public good and an application to transport policy.

2011/58. Luc BAUWENS, Christian HAFNER and Sébastien LAURENT. Volatility models.

2011/59. Pierre PESTIEAU and Grégory PONTHIERE. Childbearing age, family allowances and social security.

2011/60. Julio DÁVILA. Optimal population and education.

2011/61. Luc BAUWENS and Dimitris KOROBILIS. Bayesian methods.

2011/62. Florian MAYNERIS. A new perspective on the firm size-growth relationship: shape of profits, investment and heterogeneous credit constraints.

2011/63. Florian MAYNERIS and Sandra PONCET. Entry on difficult export markets by Chinese domestic firms: the role of foreign export spillovers.

2011/64. Florian MAYNERIS and Sandra PONCET. French firms at the conquest of Asian markets: the role of export spillovers.

2011/65. Jean J. GABSZEWICZ and Ornella TAROLA. Migration, wage differentials and fiscal competition.

2011/66. Robin BOADWAY and Pierre PESTIEAU. Indirect taxes for redistribution: Should necessity goods be favored?

2011/67. Hylke VANDENBUSSCHE, Francesco DI COMITE, Laura ROVEGNO and Christian VIEGELAHN. Moving up the quality ladder? EU-China trade dynamics in clothing.

2011/68. Mathieu LEFEBVRE, Pierre PESTIEAU and Grégory PONTHIERE. Measuring poverty without the mortality paradox.

2011/69. Per J. AGRELL and Adel HATAMI-MARBINI. Frontier-based performance analysis models for supply chain management; state of the art and research directions.

2011/70. Olivier DEVOLDER. Stochastic first order methods in smooth convex optimization.

2011/71. Jens L. HOUGAARD, Juan D. MORENO-TERNERO and Lars P. ØSTERDAL. A unifying framework for the problem of adjudicating conflicting claims.

2011/72. Per J. AGRELL and Peter BOGETOFT. Smart-grid investments, regulation and organization.

2012/1. Per J. AGRELL and Axel GAUTIER. Rethinking regulatory capture.

2012/2. Yu. NESTEROV. Subgradient methods for huge-scale optimization problems.

2012/3. Jeroen K. ROMBOUTS, Lars STENTOFT and Francesco VIOLANTE. The value of multivariate model sophistication: An application to pricing Dow Jones Industrial Average options.

2012/4. Aitor CALO-BLANCO. Responsibility, freedom, and forgiveness in health care.

2012/5. Pierre PESTIEAU and Grégory PONTHIERE. The public economics of increasing longevity.

2012/6. Thierry BRECHET and Guy MEUNIER. Are clean technology and environmental quality conflicting policy goals?

2012/7. Jens L. HOUGAARD, Juan D. MORENO-TERNERO and Lars P. ØSTERDAL. A new axiomatic approach to the evaluation of population health.

2012/8. Kirill BORISSOV, Thierry BRECHET and Stéphane LAMBRECHT. Environmental maintenance in a dynamic model with heterogenous agents.

2012/9. Ken-Ichi SHIMOMURA and Jacques-François THISSE. Competition among the big and the small.

2012/10. Pierre PESTIEAU and Grégory PONTHIERE. Optimal lifecycle fertility in a Barro-Becker economy.

Recent titles

CORE Discussion Papers - continued

2012/11. Catherine KRIER, Michel MOUCHART and Abderrahim OULHAJ. Neural modelling of ranking data with an application to stated preference data.
2012/12. Matthew O. JACKSON and Dunia LOPEZ-PINTADO. Diffusion and contagion in networks with heterogeneous agents and homophily.
2012/13. Claude D'ASPREMONT, Rodolphe DOS SANTOS FERREIRA and Jacques THEPOT. Hawks and doves in segmented markets: A formal approach to competitive aggressiveness.
2012/14. Claude D'ASPREMONT and Rodolphe DOS SANTOS FERREIRA. Household behavior and individual autonomy: An extended Lindahl mechanism.
2012/15. Dirk VAN DE GAER, Joost VANDENBOSSCHE and José Luis FIGUEROA. Children's health opportunities and project evaluation: Mexico's *Oportunidades* program.
2012/16. Giacomo VALLETTA. Health, fairness and taxation.
2012/17. Chiara CANTA and Pierre PESTIEAU. Long term care insurance and family norms.
2012/18. David DE LA CROIX and Fabio MARIANI. From polygyny to serial monogamy: a unified theory of marriage institutions.
2012/19. Carl GAIGNE, Stéphane RIOU and Jacques-François THISSE. Are compact cities environmentally friendly?
2012/20. Jean-François CARPANTIER and Besik SAMKHARADZE. The asymmetric commodity inventory effect on the optimal hedge ratio.
2012/21. Concetta MENDOLICCHIO, Dimitri PAOLINI and Tito PIETRA. Asymmetric information and overeducation.
2012/22. Tom TRUYTS. Stochastic signaling: Information substitutes and complements.
2012/23. Pierre DEHEZ and Samuel FEREY. How to share joint liability: A cooperative game approach.
2012/24. Pilar GARCIA-GOMEZ, Erik SCHOKKAERT, Tom VAN OURTI and Teresa BAGO D'UVA. Inequity in the face of death.
2012/25. Christian HAEDO and Michel MOUCHART. A stochastic independence approach for different measures of concentration and specialization.

Books

M. JUNGER, Th. LIEBLING, D. NADDEF, G. NEMHAUSER, W. PULLEYBLANK, G. REINELT, G. RINALDI and L. WOLSEY (eds) (2010), *50 years of integer programming, 1958-2008: from the early years to the state-of-the-art*. Berlin Springer.

G. DURANTON, Ph. MARTIN, Th. MAYER and F. MAYNERIS (eds) (2010), *The economics of clusters – Lessons from the French experience*. Oxford University Press.

J. HINDRIKS and I. VAN DE CLOOT (eds) (2011), *Notre pension en heritage*. Itinera Institute.

M. FLEURBAEY and F. MANIQUET (eds) (2011), *A theory of fairness and social welfare*. Cambridge University Press.

V. GINSBURGH and S. WEBER (eds) (2011), *How many languages make sense? The economics of linguistic diversity*. Princeton University Press.

I. THOMAS, D. VANNESTE and X. QUERRIAU (eds) (2011), *Atlas de Belgique – Tome 4 Habitat*. Academia Press.

W. GAERTNER and E. SCHOKKAERT (eds) (2012), *Empirical social choice*. Cambridge University Press.

L. BAUWENS, Ch. HAFNER and S. LAURENT (eds) (2012), *Handbook of volatility models and their applications*. Wiley

CORE Lecture Series

R. AMIR (2002), Supermodularity and complementarity in economics.
R. WEISMANTEL (2006), Lectures on mixed nonlinear programming.
A. SHAPIRO (2010), Stochastic programming: modeling and theory.