Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

Measure of Global Specialization and Spatial Clustering for the Identification of "Specialized" Agglomeration

Christian Haedo

Dipartimento di Scienze Statistiche "P. Fortunati"

Marzo 2009

Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

Measure of Global Specialization and Spatial Clustering for the Identification of “Specialized” Agglomeration

Christian Haedo

Coordinator
Prof. Daniela Cocchi

Tutor
Prof. Sergio Brasini

Co-tutors
Emeritus Prof. Michel Mouchart
Andres Farall

Settore Disciplinare
SECS-S/03

Dipartimento di Scienze Statistiche “P. Fortunati”
Marzo 2009

A mis padres, Nina y Nicolás

Abstract

The intensity of regional specialization in specific activities, and conversely, the level of industrial concentration in specific locations, has been used as a complementary evidence for the existence and significance of externalities. Additionally, economists have mainly focused the debate on disentangling the sources of specialization and concentration processes according to three vectors: natural advantages, internal, and external scale economies. The arbitrariness of partitions plays a key role in capturing these effects, while the selection of the partition would have to reflect the actual characteristics of the economy. Thus, the identification of spatial boundaries to measure specialization becomes critical, since most likely the model will be adapted to different scales of distance, and be influenced by different types of externalities or economies of agglomeration, which are based on the mechanisms of interaction with particular requirements of spatial proximity. This work is based on the analysis of the spatial aspect of economic specialization supported by the manufacturing industry case. The main objective is to propose, for discrete and continuous space: i) a measure of global specialization; ii) a local disaggregation of the global measure; and iii) a spatial clustering method for the identification of specialized agglomerations.

Acknowledgements

The subject of the thesis gradually came to fruition after I started working in the Research Center of the Bologna University at Buenos Aires in 2000, when I became responsible for collecting the available sources of Argentina's territorial economic statistics, and building up the data bases required to develop a research about the territorial distribution of bank credit and its impact on SMEs territorial development.

Meeting Prof. Camilo Dagum of UNIBO's Dipartimento di Scienze Statistiche, who at the time was a visiting professor in UNIBO at Buenos Aires, and occupied an office in the Research Center, was a pivotal point in my decision. I am deeply thankful to him. In his free time, he would imbue his generous and fully vital personality to my tedious work of building up the data bases, with anecdotes and stories of Argentina and other countries about the importance of economic statistics as political commitment tool to work towards a better world. In those days, the words and observations about my daily work helped me become acquainted with the methodological rigor and social commitment through which it is possible to work in a research program on applied statistics in economics. Our discussions made me feel deeply gratified, and thanks to him I started to feel almost obliged to enrol in a PHD thesis to enlarge on my scarce knowledge of statistics, acquired through the academic training as Major in Business Administration at the University of Buenos Aires.

Further on, I had the opportunity to be part of a project developed by the Research Center in 2002, aimed at putting forth the basic features of Argentina's industrial geography and the territorial distribution of SMEs across different sectors of the manufacturing industry, with an emphasis on identifying "specialized agglomerations" (Manchones territoriales-sectoriales), that could be analogous to the well-known Italian industry districts. I could then understand the underlying methodological and conceptual challenges of the goal they had set.

Soon it became evident that the technical resources available in the young UNIBO at Buenos Aires, would not be enough to reach the adequate and exact results necessary to support the technical aspects of the public action in favor of a greater economic development of my country.

From the onset, Vicente Donato, the Director of the Research Center no only supported my PHD project, but also helped me deal with it through the continuous review of my research work.

Thus, the need to engage in a PHD study to complete my scientific training finally materialized, and in 2005 I decided to apply to the PHD in Statistical Methodology for Scientific Research from the Bologna University, where I submitted my project draft thesis about the identification of specialized agglomerations. The incentive from the authorities of UNIBO at Buenos Aires and of the UNIBO School of Statistical Science, Professor Giorgio Alberti and Professor Paula Monari, respectively, was also decisive. I would like to deeply thank them for their trust in my work and commitment to the development of UNIBO's office at Buenos Aires. I would like to express my gratitude to Prof. Estela Bee Dagum, an argentine herself, who back in those days was working at the Dipartimento di Scienze Statistiche of UNIBO, for her initial scientific guidance and her assistance in promoting my social insertion in the department. Also, my deep appreciation to Prof. Daniela Cocchi, the PHD Director, for her guidance during the first months of my administrative and environment "confusion" in Bologna, and her significant advice to enhance the development of my curriculum.

I am also full of gratitude towards Prof. Furio Camillo, with whom I had the pleasure to work in the organization of the Market Research and Data Mining Master at Buenos Aires, for his methodological suggestions to analyze large databases.

In addition, my thanks to Professors Fabrizia Mealli, Giuseppe Arbia and Bruno Bracalente, jury of my thesis, for their comments and contributions to the future development of my work.

I also want to thank especially Prof. Sergio Brasini, my coach and tutor, for his full-time availability to assist me to complete my PHD thesis at UNIBO, and Andres Farall for his technical assistance in R programming, that helped me explore the operative aspects of my work hypothesis. Andres Farall's contributions through his method to obtain the best collapsed table, and his suggestion to use the bootstrap hypothesis testing method as an alternative to the significance features commonly used for image analysis, also deserve my gratitude.

Finally, my special thanks to Prof. Michel Mouchart, who from the moment I took the survival analysis course he gave during my first year, became my most important intellectual and scientific mentor. I wish to express my deepest appreciation to him, for helping me solve every concept and methodological core faced during the development of my thesis.

From the outset, Prof. Mouchart opened the doors of the Institut de Statistique and the CORE of Université catholique de Louvain, where I was able to discuss the building blocks of my thesis, and to receive the comments of Prof. Dominique Peeters and other Professors, to whom I am also thankful. As early as my first visit to the Institute, Prof. Mouchart introduced me to the study of Poisson processes, that later on became the methodological basis of the last chapter referred to the identification of specialized agglomerations in continuous space.

But above all, I am particularly full of gratitude to Prof. Mouchart for his continuous dedication to transmit not only scientific knowledge but also the teachings and attitudes of a full life devoted to the research in statistical and mathematical science.

Contents

List of Figures

List of Tables

Chapter 1

Presentation of the study

1.1 Introduction

If we observe a map with the location of economic activities, the following may be noticed: i) there are clusters of points in a certain region, i.e. the spatial distribution is not uniform; and ii) some clusters show a wide variety of activities with a high proportion of a uniform or a small number of activities.

The results of these observations may be summarized in three basics concepts: concentration, specialization and agglomeration. These concepts are reviewed in the following section.

One of the goals of economic geography is precisely to explain the location of economic activities vis-à-vis the intensity of above-mentioned concepts. Section 1.3 provides an approach to this perspective. Finally, section 1.4 puts forth the goal of this thesis.

1.2 Three basic concepts: concentration, specialization and agglomeration

Prior to introducing the causes and theoretical underpinnings of economic geography that explain such phenomena, this section seeks to clarify the differences and existing relations among such concepts from a technical and conceptual perspective.

1.2.1 Specialization vs. concentration

The concentration of production, one of the most striking feature of the geography of economic activities, is probably also the most direct evidence of the pervasive need of firms to draw benefits from the presence of externalities. The concentration of an economic activity is a distribution measure of its country share, and an activity is said to be regionally concentrated if a few regions have a large share.

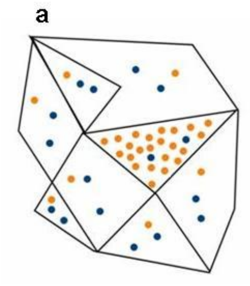
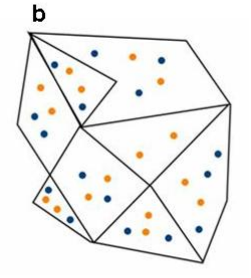
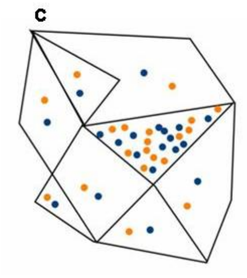
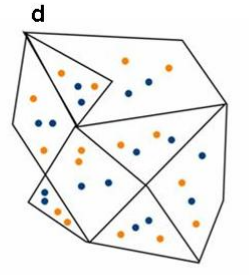
On the other hand, specialization or relative concentration of an economic activity becomes readily apparent when part of the territory has a greater proportion of elements (most often uses employment as the gauge) from a particular activity than the proportion of such activity in the whole territory. In other words, specialization compares an area's share of a particular activity with the area's share of an aggregate phenomena. Simply put, specialization identifies the way in which local activities are packed up with respect to national average. Benchmarking the degree of relative concentration of an activity in the analysis of area localization, has received considerable attention in the geographic and economic literature.

Overall specialization (e.g. of US or Europe) is then a weighted or unweighted average over the regions and overall concentration over industries. However, the specialization of countries in a particular activity and the concentration of industries in regions or countries it not identical. However, empirical studies often focus either on specialization or concentration, sometimes assuming that these would develop in parallel (Aiginger and Rossi-Hansberg 2006 discusses the basic setup of the model in Rossi-Hansberg 2005, and the implication 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). The intensity of regional specialization in specific activities, and conversely, the level of industrial concentration in specific locations, has been used as a complementary evidence for the existence and significance of externalities. Besides, economists have focused the debate mostly on disentangling the sources of specialization and concentration processes according to three vectors: natural advantages, internal and external scale economies.

Table 1.1 shows four possible combinations of these two concepts. In this example, the total geographical area could represent a country, polygonal subdivisions to regions, points to firms, and the orange and blue colors to two different economic activities.

Should we focus on the triangular central region and orange activity, it is easy to note that such location or region is specialized when its share of firms in the activity

Table 1.1: *Specialization vs. concentration*

Specialization	Concentration	
	Yes	No
Yes		
No		

exceeds its national share (Fig. *a* and *b* in Table 1.1). By opposition, the triangular central region in Fig. *c* isn't specialized because it shows the same proportion of orange and blue points than the rest of the regions.

1.2.2 Agglomeration vs. just concentration

Although a large number of indices are used by economists and geographers to inquire about the patterns of concentration, the key problem is that they do not take into account anything that is truly spatial, i.e. they only consider the global distribution of counted data, without considering the relative position of geographical units in space. In fact any statistical measure of variation or concentration satisfies the condition of anonymity with respect to the individuals, that is the property of being insensitive to any spatial permutation of individual orderings. However, this is absolutely not a desirable property for the spatial inequality measure.

On the other hand, agglomeration considers the space interdependencies between the geographical units in question, that have been used as complementary evidence of the existence and significance of externalities. Agglomeration is used to refer to the degree of spatial correlation among observations which may be distributed in a two-dimensional space in order to form some specific distance-based pattern, i.e. as a synonym of positive spatial auto-correlation. The spatial correlation at a given geographic scale translates, to a certain extent, into concentration at a more aggregated level. A-spatial concentration measures are constant under spatial permutations, and therefore they represent only one aspect of spatial concentration.

Concentration can be measured with the standard locational Gini coefficient (Krugman 1991b) or the more sophisticated γEG (Ellison and Glaeser 1997) and γMS and γUW (proposed by Maurel and Sédillot 1999) indexes. By contrast, agglomeration can be measured with the spatial auto-correlation coefficient I (Moran 1950) and the GO statistic based on the local indicator of spatial association (Getis and Ord 1992), which consider the space interdependencies between the geographical units in question (for further details about spatial processes and measures of association, see Cliff and Ord 1981, and Anselin 1995).

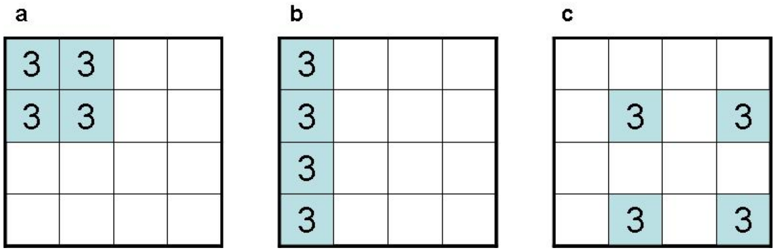
Following Arbia (2001b) let us consider a hypothetical example in which we have 12 firms located in a study area exhaustively partitioned into 16 squared cells (subregions) arranged in a 4-by-4 regular lattice grid.

Fig. 1.1 shows three very different location situations. Agglomeration (or polarization in terms of Arbia) is higher in case a than in case c. However, concentration remains unchanged in the three cases because it is essentially an a-spatial measure that remains constant under permutation, and not distinguished between the inequality of the a-spatial distribution, while it provides a permutation invariant quantification of how much variable is a phenomenon with respect to the same average (note that the three cases in Fig. 1.1 show 4 cells with 3 firms each one and 12 empty cells).[1]

The case of positive spatial auto-correlation is thus associated to a high degree

[1] Using a simple rook's definition of neighbours, in case a of Fig. 1.1 (high polarization), Moran's I assumes the $I = 0.227$ value; in the intermediate case b assumes $I = 0.222$; and in the case of high dispersion c assumes a negative value $I = -0.183$. Similarly the GO statistic assumes values of 0.67, 0.5 and 0 respectively in the three cases. However, the Gini coefficient remains unchanged in the three cases (0.75).

Figure 1.1: *Agglomeration vs. concentration*



of polarization, whereas the case of low spatial correlation indices is associated to low degrees of polarization when data are spatially dispersed in the area studied. However, a spatial correlation coefficient is not a good measure of spatial concentration since for instance, I Moran spatial auto-correlation coefficient and GO statistic measure the level of polarization and do not take into account the variability of the phenomenon. To deal with this problem, Arbia (2001b) propose a summary index that measures simultaneously a-spatial concentration and polarization, that combines the three measures: Gini coefficient, I Moran coefficient and GO statistic.

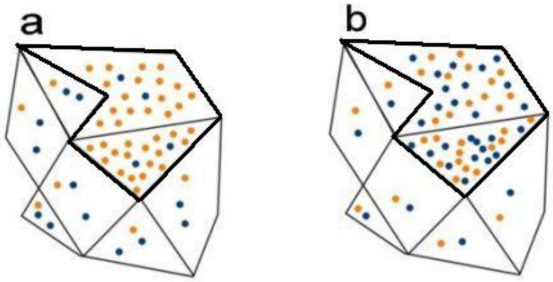
In case c of Fig. 1.1, firms are evenly distributed among locations, as if there was a completely random process that associates the 4 given clusters of the 3 plants to the 16 available locations. In this situation, although there is some concentration, spatial agglomeration is clearly not the source of such variability. By contrast, in case a, firms are distributed in such a way that we can clearly identify an agglomeration of firms in the upper-left corner of our grid, that leaves an empty space of abandoned locations. Location is here highly informative because high (low) values are surrounded by high (low) values following a pattern that is known as positive spatial correlation (Lafourcade and Mion 2003).

Finally, in a more complex space, with a distance diffusion process agglomeration is identified because it implies a distance decay pattern witch is not obviously related to concentration. Hence, there might be agglomeration without concentration, and concentration without agglomeration.

1.2.3 Specialized agglomeration vs. just agglomeration

Just agglomeration (spatial concentration) doesn't mean specialized agglomeration (spatial specialization). Spatial specialization of regions in a particular activity and

Figure 1.2: *Specialized Agglomeration vs. just agglomeration*



spatial concentration of industries in regions it not an identical phenomena, but empirical studies often focus either on the spatial analysis of specialization or of concentration, sometimes assuming that these would develop in the same way. This two different spatial configurations are closely related to a very different combination of centripetal and centrifugal forces and internal-external increasing returns of scale (for more details see Section 1.3).

Fig. 1.2 shows a spatial clusters or agglomeration, as a synonym for positive spatial auto-correlation formed for two contiguous polygons or regions (to refer to distance-based location patterns): central triangular and upper right corner. Case *a* in Fig. 1.2 shows spatial specialization, while case *b* shows spatial concentration.

Cases *a* and *b* in Fig. 1.2 have a close spatial cluster relationship with the figures *a* and *c* shown in Table 1.1. In a certain way, spatial specialization should occur when the location share of firms in the activity exceeds its national share (the case *a* of Fig. 1.2). By opposition, the spatial cluster of case *b* in Fig. 1.2 is not specialized because it has the same proportion of orange and blue points than the remaining regions.

1.2.4 The MAUP problem

Specialization, concentration and agglomeration may appear in different geographical scales and may involve different disaggregation levels in an industry, and consequently, a certain space scale is not necessarily equivalent to another. The reason for this difference probably lies in the nature and balance of the centrifugal and centripetal force systems acting in different geographical scales, which explains the different levels of analysis raised (see Krugman 1991b and Anas, Arnott and Small 1998). Hence, identifying the space limits of specialization, concentration or agglomeration becomes essential. In ecological studies, this problem is known as the "Ecological Fallacy" and lies in thinking that the relationships observed between the groups will necessarily apply to individuals. In other words, the inferences about the nature of individuals are based solely upon aggregate statistics collected for the group to which those individuals belong (Robinson 1950). In spacial statistics, this problem is known as the "Modifiable Areal Unit Problem" (MAUP), which refers to the arbitrariness of the geographical partition used (for more details, see Yule and Kendall 1950; Openshaw 1984; Arbia 1989; Amrhein 1995 and Unwin 1996). The arbitrariness of geographical boundaries gives rise to two different manifestations, namely aggregation and scale, and any statistical measure based on spatial aggregates is sensitive to the scale and aggregation problems.

Following Arbia (2001a) the case a in Fig. 1.3 poses an obvious situation involving a strong geographic concentration at the core of the study area (spatial point pattern-continuous space). Suppose we would want to measure the concentration by regional aggregates, and that we would superimpose -or use a previously defined-grid of quadrates (lattice data-discrete space) as in case b of Fig. 1.3. Each point represents a firm. In this situation, any concentration measure would identify the absence of concentration. However, if we use the same grid, but shift the origin in the northwest direction as in case c of Fig. 1.3, we would reach the opposite conclusion, since any concentration index would identify a maximum level of concentration. There lies the description of the aggregation problem. Conversely, by examining the case en which a finer grid of quadrates (one-fourth the size of the previous one) is superimposed onto the same set of data, a concentration index will take an intermediate value between case b and case c of Fig. 1.3. This is the description of the scale problem.

One can easily imagine that the situation is even worse in real cases, where the spatial units are irregular in size and shape, thus yielding an even higher degree of arbitrariness.

This example shows that, by observing any geographical distribution through regional aggregates, we would be in fact observing two separate phenomena which are matched in an unpredictable way with respect to: i) the actual distribution of objects in the space, and ii) the partition considered.

Likewise, and to illustrate the effects of the partition considered for specialization, Fig. 1.4 shows that with the same distribution of firms in the space, it is possible to find specialization and the absence of specialization, respectively.

Arbia (1989) shows that the distortions due to scale and aggregation are min-

Figure 1.3: *A continuous space distribution of firms (**a**) and three discretised versions of it. Figures (**b**) and (**c**) illustrate the aggregation problem. Figures (**b**) and (**d**) illustrate the scale problem*

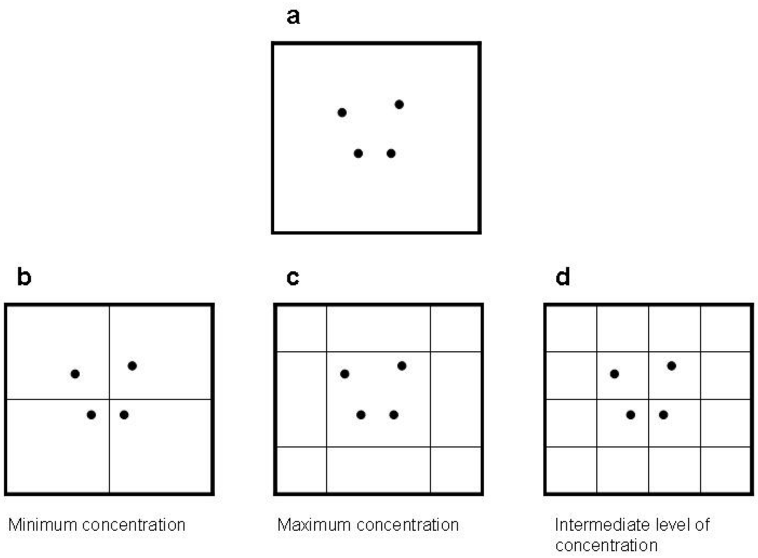
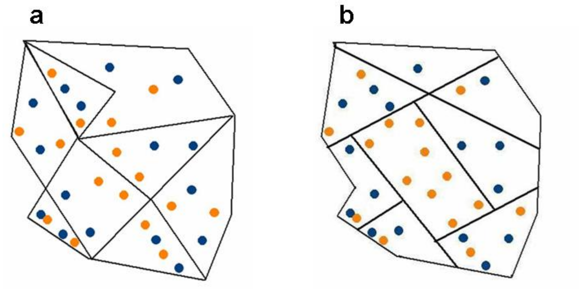


Figure 1.4: *Effects of the spatial partition considered on specialization*



imized (but never eliminated) under some very restrictive conditions involving the identity of the sub-areas considered (in terms of size, shape and neighboring structure), and the absence of spatial interdependence. Such conditions are never realized in economic geography -where data are observed within administrative regions that are unequal in size, shape and neighborhood- and where, typically, neighboring regions usually resemble more between each other than regions that are far apart. In particular, Arbia and Espa (1996) demonstrate analytically that by aggregating spatial units we can observe a general decrease of variance and an increase in the correlation between pairs of variables.

Therefore, it is important to note that the arbitrariness of partitions plays a key role in capturing the effects mentioned previously, and becomes potentially more dangerous the more unequal become the elements of it in terms of area. In this sense, to minimize MAUP, the selection of the partition would have to reflect the actual characteristics of the economy. For example, the Italian statistical institute developed a partition called Local Labor Systems (LLS), which covers urban and rural areas, while minimizing the commuting patterns and maximizing the overlap between the working and residential areas. The LLS gray thug is not dissimilar to the U.S. partition into zip-code units. In general, available data refer to a discrete space (lattice data), geographically aggregated at different levels: states, regions, cities, districts, etc. Therefore, above mentioned effects coupled with the effect of crowding can appear in different geographical scales and involve several levels of sectorial breakdown, that is, a certain spatial scale is not necessarily similar to another. The identification of spatial boundaries to measure specialization, concentration or agglomeration becomes critical, since most likely the model will be adapted to different scales of distance and be influenced by different types of externalities or economies of agglomeration, which are based on the mechanisms of interaction with particular requirements of spatial proximity.

The data available refer to a discrete space, and our limitations push us to discretise the phenomena in some way (and subsequently distort it by reducing the quantity of information). The availability of statistical data at an individual level has increased considerably in recent years as well as the GIS technologies to deal with them, and the methods for analyzing spatial data in a continuous space now form a well-consolidated methodological body (see Ripley 1981; Cressie 1993; Arbia and Espa 1996; Lawson and Denison 2002; Diggle 2003; Møller 2003; and Møller and Waagepetersen 2004). Therefore, currently there seem to be no technical obstacles to a micro approach to regional problems, i.e. to move from a discrete to a continuum space. Also, the boundaries cannot be completely ignored, given that economic conditions can change abruptly due to such changes in the tax system, in transportation costs, or to the impact of public policies at regional and sectoral levels.

By starting from these considerations, in the first stage of an exploratory analysis one can simply ignore such boundaries, i.e. one think that the shift of emphasis from a meso-to a micro-level is likely to bear interesting results. In fact, Krugman (1991a) has observed that if we wanted to understand the differences in the rate of national growth is necessary to begin by examining the differences in regional growth. Therefore, a good way to understand regional economics is to begin by examining the micro behaviour of economic agents in the space economy, and then to explore the micro foundations of regional economics (see Barff 1987; Duranton and Puga 2004; Duranton and Overman 2005 and 2006; and Arbia, Espa and Quah 2007). After a model has been identified at the micro spatial level, we could certainly superimpose an administrative grid and examine the implied meso-scenario.

1.3 An introduction of the Economic Geography

During the last 20 years, we have witnessed reinvigorated discussions about the role of the geographical space in the economic growth. In order to explain the differences in the observed evolution of the nations and regions wealth, the "geographical hypothesis" was raised to the explanatory category, that runs counter to the "institutional hypothesis" and to the "cultural hypothesis".[2] The most substantial arguments in favor of the geographical hypothesis are based on the reconsideration of the role played by the increasing scale returns of firms, transportation costs, and the mobility of production factors, especially work. Other type of literature, inspired by A. Marshall, has also revalued the role of increasing firm-external scale returns (a kind of "industrial environment" that prevails in certain territories, promoting the rapid dissemination of technical knowledge and human behavior, in favor of production).Using these new concepts, it has been demonstrated that "geography counts" (Krugman 1991a).

In addition, the new economic growth models, called "endogenous growth" models (Romer 1986 and 1990, Lucas 1988), by introducing the human capital concept and the concept involving the externalities created by the dissemination of technical

[2] For a discussion of these three optional hypotheses and the different approaches, the most comprehensive collection of writings can be found in Aghion, P., and Durlauf, S. (eds) (2005). *Handbook of Economic Growth.*

knowledge, have concluded that the economic growth process is subject to increasing scale returns, and that it consequently tends to accelerate in the long term in countries and regions with higher availability of human capital that deliberately promote their production. Due to the introduction of the concept related to the increasing scale returns both by firms and societies, seen from the perspective of the new growth models and the new economic geography models, we may conclude that the convergence between the profit level of rich and poor countries is no longer the natural result of the growth process, as has been the hypotheses posed by neoclassical growth models such as Solow's (1956). Additionally, international trade is no longer a mechanism to balance the economic development of the different world countries and regions. In fact, some countries and regions are growing more rapid than others. Space counts. Localization counts. The operation of the mechanisms of increasing scale returns by firms and societies is not evenly distributed across all territories worldwide, or across the different regions in a country.

Today we know that economic growth is shown through the spatial concentration and de-concentration of the economic activity. The rapid growth of East Asia in but short decades, or specifically of Japan, a country that with only 3.5 percent of the total surface of the region and 8 percent of its population, accounts for 72 percent of GDP and 67 percent of manufacturing GDP in its macro-region. In addition, there are strong regional disparities implied by the existence of space agglomerations of different scale firms in the same country, such as Seoul in Korea, Paris in France, Sao Paulo in Brazil or Buenos Aires in Argentina, that concentrate more than 30 percent of the respective country's GDP.

The wealth or poverty of nations seems to be increasingly related to development, and to the presence of competitive clusters of specific industries and of extended and diversified metropolitan areas. Hence, the reports released by multilateral credit entities such as the World Bank (2000), highlight the importance of agglomeration economies and of the cities, arguing that the world's liberalization and the effect of regional trade treaties reduce the power of national governments, while increasing the power of regions and cities.

The spatial concentration of the industrial activities phenomena can be very diverse.

Large metropolis such as New, Tokyo or Buenos Aires are highly diversified in many industries that are not directly related (Fujita and Tabuchi 1997), while on the contrary, many cities can be specialized in a small number of activities by sector (Henderson 1997a). Many well-known agglomerations by territory and sector are used as examples by this type of literature. For example, the high-tech sectors, such as Silicon Valley (California), Route 128 (Boston), Cambridge (UK) and Sophia Antipolis (France); the automotive sector, Baden Württemberg (Germany) and Detroit (USA); the tile and premium clothing sectors in the regions of Third Italy (Italy); and the financial services sector, Wall Street (New York) and London City (UK).

Firms and activities of different sizes may be organized differently both at an urban and at a regional level. Agglomeration can take the form of large business districts in the interior of the same city, such as Soho in London, Ginza in Tokyo, etc. At a smaller geographical scale, we may find restaurants, movie theaters, performance halls or shops grouped in the same neighborhood, street or shopping mall, that sell very similar or strictly complementary products.

Urbanization economies (Hoover 1937) refers to the advantages of the size and density of local economies. It includes the variety of specialized services implied in having different industries located close to each other. Consumers may benefit from similar advantages. Storper (1995) recognized the concept of "untraded dependencies", arising, not only from input-output linkages, but also from the conventions, rules, practices and institutions involved. Following Jacobs (1969), more diverse environments, such as those existing in cities, provide a better breeding-ground for new ideas as a result of the cross-fertilization of ideas from different areas. The diversity of ideas is used differently in the new spatial agglomeration theory (Fujita et al. 2001 and Fujita and Thisse 2002). Together, the stage of the product life-cycle and the size of the urban settlement play a crucial role for the location of production, and as suggested by Jacobs, for the development of ideas. While diversity/specialization arguments emphasize agglomeration in general, Jacobs's discussion always emphasizes the development of ideas, resulting from diversity (Ejermo 2005).

Firm groupings may appear in different geographical scales and involve different disaggregation levels by sector, and consequently, a certain space scale is not necessarily equivalent to another (see Section 1.2.4). The reason for such difference probably lies in the nature and balance of the centrifugal and centripetal force systems acting in the different geographical scales, and therefore, the different levels of analysis subsequently raised. Anas, Arnott and Small (1998) argue that the model required to account for the different distance scales is likely to be affected by different types of agglomeration economies, that are based on the interaction of mechanisms with specific space closeness requirements. Hence, identifying the space limits of agglomerations becomes essential. O'Donoghue and Gleave (2004) and Duranton and Overman (2005) put forth to initially provide a measure for the agglomeration concept, and to define clearly its boundaries before making the empirical analysis of the causes that define it, while Rosenthal and Strange (2001) and Fujita and Thisse

(2002) state that the type of agglomeration can be specified on the basis of the issue under analysis.

Thus, based on the new significance given to the “geographical hypothesis”, exploring the statistical indicator that may contribute to measure more precisely the distribution of economic phenomena in the geographical space, has regained a more vigorous lead in the last years (Ellison and Gleaser 1997, Maurel and Sédillot 1999, Mori and Smith 2005).

However, in view that in general, available data about the territorial economic activity are geographically aggregated at different levels, such as states, regions, cities, districts, etc., in other words, on a discreet space, and not always with the same degree of uniformity and continuity across the various countries and regions, the difficulties to rigorously identify the functional boundaries are not minor.

This study is based on the analysis of this perspective, and is focused on proposing a new methodology to measure some of these phenomena including the agglomeration and the specialization of the economic activity.

1.3.1 How to explain the location trends of the economic activity?

The body of theories known as “The New Economic Geography” (NEG) tries to account for the large number of economic agglomeration typologies at different geographical levels, i.e. the economic mechanisms that cause the spatial concentration of certain economic activities in certain territories.

The differences between the NEG’s modern approach and the traditional approach based on the theories of localization and economic geography, are: a) the general balance model approach; b) increasing returns or technology indivisibleness at the firm level; c) the imperfect competitiveness resulting from prior item; d) transportation costs; and e) the considerations about the factors’ mobility. Based on the new models, the goal would be to explain how the geographic structure of an economy is determined by the conflict between the “centripetal” forces that concentrate the activity in terms of territory, and the “centrifugal” forces that disperse it.

Fujita (1988), Krugman (1991a, b; 1995) and Venables (1996) can be considered among the key forerunners of the NEG approach. They acknowledge that their analytic approaches originate in the models based on general balances with monopolistic

Table 1.2: *Centripetal and centrifugal forces*

Centripetal forces	Centrifugal forces
-linkages	-immobility factors
-thick market	-land rent/commuting
-knowledge spillover and other pure economic externalities	-congestion and other pure diseconomies

competitiveness.[3] Krugman (1995) demonstrates that the constant returns-perfect competition paradigm proves unable to account for the rising and growth of large territorial economic concentrations. The presence of scale economies is essential to explain the geographical distribution of economic activities, and provide countries the incentive for specialization and trading, even if there are no differences in terms of technology or accounting factors. In particular, the trade -off between increasing returns from productive activities and transportation costs is core to understand the geographical distribution of economic activities. The first formal model acknowledged by this literature is the Krugman (1991a) regional model, that explains the rising of a core-periphery structure within a country, or within any other boundaries where there is free work mobility. Krugman demonstrates that in the presence of increasing returns, work mobility and transportation costs, the centripetal forces and forward-backward links create a trend towards agglomeration (territorial concentration) of firms and workers.

The group of external economies mentioned above as those accounting for agglomeration correspond only partially to the "Marshallian economies". The analysis of external economies dates back to more than a century, when the British economist Alfred Marshall (1890) was shocked to observe the existence of "industrial districts" that had spontaneously concentrated firms and workers from the same sector in a certain territory: cutlery manufacturing in Sheffield, socks manufacturing in Northampton, etc. The existence of these industrial districts could not be accounted for by the mere proximity of the sources of raw material and of specific natural resources. In more up-to-date terms, Marshall puts forth three reasons (see the picture above) why a group of firms can be productively more efficient than an individual firm working in isolation: the forward-backward linkages associated to large local markets that translate in the group's ability to support specialized

[3] The first operative non-competitive general balance model was developed by Dixit and Stiglitz (1977).

providers; the existence of a specialized labor market; and the profits resulting from speed of technology dissemination. Marshallian economies consider the presence and the nature of firm-external economies though territory-internal economies, and the advantages of work specialization/division that take place in the firms of a system. To build an efficient production system, facing the transaction costs that are the base of the vertical integration in large firms, such de-centralized systems basically depend on local characteristics. These characteristics create an "industrial atmosphere" defined in socio-economic terms, i.e. a group of contiguous localized firms related systemically, based on the characteristics of the local society.

Consequently, the externalities related to human capital, the information flow, the technology innovation and dissemination processes, and finally, the relationships between customers and providers in the endogenous growth models (Romer 1986, 1990 and Lucas 1988), provide the adequate theoretical and analytical framework to develop applied studies about the presence and nature of firm-external economies though territory-internal economies.

The NEG considers only the first of these marshallian externalities -the forward-backward linkages among the firms-, which may be argumentatively less significant in practice, but which, according to Fujita and Krugman themselves, is easier to formalize in models than the rest of the variables identified by Marshall.

1.3.2 The three NEG's basic models

Fujita, Krugman and Venables (2001) present three different types of models to account for firm localization, all of them based on the theory of monopolistic competitiveness: regional models, urban system models, and international models. These models are based on the same concept architecture, since as they relate to urban economy, location theory or international trade, it is only about where and why the economic activity takes place. In addition, as it has been previously noted, these authors put special emphasis on the centripetal force triggered by the links between the production and transaction of goods and services.

Due to the highly non-linear nature of the geographic phenomenon, building valid geographical balance models becomes very difficult. Currently, other general balance models that are wider than the monopolistic competitiveness are under development. These models include, for example, imperfect competitive markets in space, such as those approached by Ottaviano, Tabuchi and Thisse (2002), who consider the core-periphery "linear models". Additionally, other empirical works such as those by Dumais, Ellison and Glaeser (1997) that analyze the significance of a specific centripetal force are emerging, and also other works based on Krugman's labor pooling (1991b), that explain the specific impact on the industry and the firm through the mobility of workers' cost among the industries, and the endogenous impacts resulting from the risks investments made by the firms to increase productivity.[4]

Core-periphery regional model

This model developed by Krugman (1991a) is focused on illustrating the mechanisms through which a space economic structure may emerge or change based on the interaction among the increasing scale returns of firms, transportation costs, and factor mobility.

The model's assumptions are: there are two regions -A and B-, two production industries -agriculture and manufacturing-; and two work types -farmers and workers. The manufacturing industry produces a continuum of a variety of products that differentiate horizontally, while each product variety is produced by a different company subject to scale economies and using the work as the only input. The agricultural sector produces a homogeneous good according to the continuous return method and uses farmers as the only input. Workers can move freely across regions A and B, while farmers are non-movable and are equally distributed across both regions. Inter-regional manufacturing trade requires a positive transportation cost (in the form of an inverted U), while the agriculture good is transported at zero cost between both regions.

The farmer's immobility acts as a centrifugal force, since farmers consume both types of goods. If a larger number of companies localize in region A, the variety of horizontally differentiated products to be manufactured there will be higher. Therefore, the workers in that region who are also consumers, will be able to access more easily a wider range of products than those in region B. If the rest maintains ceteris paribus, the workers in region A will earn better salaries, attracting more workers to emigrate to the region.[5] In addition, due to the scale economies, the increase in the number of workers, and consequently of consumers, creates a broader market than that in region B. Therefore, there is an incentive to concentrate production in region A, because transportation costs and the market size make it more productive to produce in region A and transport products from there to region B. In a nutshell, the centripetal force is created through the la Myrdal (1957) circular causality of

[4] See Stahl and Walz (2001) and Gerlach, Rønde and Stahl (2001).

[5] For more, see Razin and Sadka (1997).

forward linkages -the workers' incentive to be close to the producers of consumer goods-, and backward linkages -the producers' incentive to concentrate in larger markets. When these linkages are strong enough to exceed the centrifugal force built out of the farmers' immobility and sufficiently low transportation costs, the economy is likely to follow a core-periphery pattern whereby the full manufacturing production is concentrated in a single region. The core-periphery pattern may occur when: a) transportation costs are sufficiently low; b) the diversity of goods is sufficiently differentiated; c) manufacturing expenses are sufficiently high.

Eventually, the agglomeration may not emerge. However, a small change in the critical parameters may cause an economy with two symmetrical and equal regions, to be transformed into another economy, in which small initial advantages are accumulated and end up by transforming one of the two regions in the industrial core, and the other in a de-industrialized periphery. The model's dynamic is subject to "catastrophic bifurcations", i.e. sudden changes in the trend.

Urban systems models

These models are supported by the core-periphery model that has been described above, and are based on the following assumptions: instead of two regions -A and B-, the localization space is now described according to the line uniformly dividing the geographical areas, while workers are identical and free to choose their localization and occupation. The agricultural good is now produced using both factors, i.e. the workers and the land. Finally, there are positive transportation costs for both goods -agricultural and manufacturing. Consequently, the only centrifugal force in this model is the land, for it is the only immobile factor.

Under this model's assumptions (Fujita and Krugman 1995, Fujita and Mori 1997, Fujita, Krugman and Venables 2001), it is possible to find a well-defined balance model according to which the core city derives the effects of forward-backward linkages instead of being a mere assumption of the model, which translate into a gradual increase of the economy population as a whole.

This balance is based on Von Thünen's (1826) "isolated states" approach: a city, defined as a manufacturing concentration, surrounded by an agricultural belt. At a certain point in time, a new city may emerge, because the ending boundaries of the agricultural belt are already too far from the core to justify the relocation of some firms to better supply the market, therefore causing the emergence of a new city. Future population growths promote the creation of more cities. In this context, it is critical to acknowledge that for the manufacturing industry, the attraction of a localization lies in the size of the "underlying market potential" (Krugman 1993). As a result, the economy change process may be considered as a certain co-evolution: the market's potentiality determines where the economic activity will be localized, while the activity location change redefines the market's potential map. Despite the potential existence of many possible balances, the models predict some predictable regularities in the space structure. The relative force of the centripetal and centrifugal forces will determine that once the number of cities is large enough, their size and the distance separating them will tend to stabilize at a constant level.

It is generally acknowledged that the role played by the natural geography and the vast and varied number of historical circumstances and casual arbitrary acts that determine the current economic geography, was very significant. The new and more sophisticated NEG models, known as urban systems, capable of accounting for the emergence of a system of cities, are designed to consider all of these, i.e. the relationships between natural geography, historical circumstances and the force of economic geography.

According to Fujita and Krugman (2004):

> The aspects that are favorable to a certain location, such as the existence of an adequate port, usually play a catalytic role: it works so that when a new core emerges, it will locate there, instead of in a new and close location. But when a new core is already established, a new loop process is started that can reach a point in growth at which the initial localization benefits are irrelevant compared to the advantages of the self-sufficient process triggered by the agglomeration. Too uncommonly, it could be said that the natural geography is so important precisely due to the self-organizational nature of the space economy.

International models: can the means used by the NEG explain the concentration of activities by territory and sector?

The two previous NEG models provide an explanation for the emergence of spatial economic concentrations, and for the emergence of cities. However, the forces promoting concentration are not always capable of explaining the appearance of sector-specific productive activities. There is a large number of cities specialized in a small number of activities or sectors, such as Detroit (cars) and Hollywood (film) in the US or Sassuolo (tile) or Carpi (textile) in Italy.

The previous NEG models cannot properly account for this phenomenon of specialization by territory and sector, without the introduction of some changes to the prior models. Since in previous models, only two goods and two regions were considered, the spatial concentration processes imply that all the firms and consequently all the manufactured goods concentrate in a single region. All the firms are concentrated in the manufacturing region, each producing a different good, and therefore no trend towards specialization by sector may be observed.

To explain the emergence of the spatial concentration of specific activities by sector, Fujita, Krugman and Venables (2001) introduce a model that assumes two countries, one of them including two regions. In the country with two regions, working assumptions are those of the core-periphery model. In other words, these new models now have two sectors and three regions. When the trade costs between the countries plunge, the localization patterns also change, allowing the deployment of two different processes that do not necessarily go in the same direction: concentration and specialization. The space economic theory that best embraces these differences is the Rossi-Hansberg (2005) model, whereby lower transport cost increase the specialization of regions or countries and decrease the (regional) concentration of industries. The driving force for the first effect is that lower transport costs allow firms to take greater benefits from sector-specific production externalities. The driving force for the second effect is that lower transport costs shift production to regions far away from main markets, since exporting to distant locations is less costly.

For high transport costs, the incentives to specialize are low. A firm moving to an area that does not produce the manufactured good would profit from increases in the demand for its product from local consumers, but will lose from paying higher wages (agents have to import all manufactured goods). The gain from the increase in sales (home market effect) outweighs the loss from higher wages (wage effect), because agents consume agricultural goods as well. Hence, regions do not specialize. As transport costs decrease, both the home market and wage effect decrease. However, the home market effect decreases faster than the wage effect. The reason is that agents substitute local manufactured goods for foreign manufactured goods, so the value of local sales decreases as transport costs decrease. Local wages decrease at a lower rate, since part of the agents consumption is in agricultural goods. This implies that as transport costs decrease, the incentives to move to the agricultural region decrease. Eventually, it becomes unprofitable for firms to deviate, and specialization becomes in equilibrium. If transport costs are even lower, the loss in higher wages becomes less and less important, as does the gain from higher sales.

However, for low-enough transport costs, the home market effect will decrease slower than the wage effect. The reason is that if regions do not specialize, the location to which the firm is deviating will become almost as large as the original region, i.e. local markets will increase and so the benefits from relocating there will increase. Eventually, when transport costs are zero, the wage and market effects will cancel out, and there will be no incentives to deviate. This means that there will be no specialization. Aiginger and Rossi-Hansberg (2006) use two data sets for the manufacturing activity, one for the US states and ten industries, the other for EU member countries and 23 industries. In both data sets specialization and concentration do not develop in parallel, and the kind of divergence is roughly in line with the model prediction. Specialization of industries is indeed increasing over the past years in Europe and the US, and regional concentration of industries is decreasing in both areas (in Europe to a less degree, starting from a much lower level).

When the trade costs between the countries plunge, agents can consume less local goods and consequently, there is a decrease in the benefits of territorial concentration. At the same time, the decrease in trade costs between the countries increases the incentives to create clusters by sectors in particular regions, since now sales depend less on the local market, and countries can be supplied even from more distant locations, thus increasing the benefits of specialization by territory and sector.

In these NEG's models, the selection of the specific production by sector to be located in the region, from where the other internal region and the foreign country will be served, depends on the existence of some firm-external marshallian economies, that the authors of the model restrict to "cash" ones, or in other words, those derived from the intensity of the forward-backward linkages trade connections of firms located in the same territory.

The key is to abandon the emphasis put on the assumption that firms produce a complete range of horizontally differentiated goods, to emphasize the notion that now firms choose to produce only the horizontally differentiated goods that integrate vertically in the productive structure (Venables, 1996). One or more industry sectors in the upper layer of the production structure produce inputs for one or more sectors located in the lower layer. All firms located in the upper and lower layers are subject to increasing scale returns and transportation costs. Thus, the vertical relationships between the companies become apparent and encourage them to locate in the same territory. The producer of intermediate goods have incentives to locate in areas with higher market possibilities, i.e. where the industry demanding its products is located. Likewise, the producers of final goods have incentives to locate where their providers are located.

The dynamics of the third model highlights the significant role played by externalities, i.e. the firm-external economies but internal to their localization territories, in the creation and maintenance of concentrations by territory and sector in the real world.

Positive externalities derived from forward-backward linkages associated to large local markets, that translate in the group's abilities to support specialized providers, have been covered a long time ago by Hirschman (1958) and Perroux (1955). However, the means provided by the NEG today allow us to analyze these externalities as derived from the model and not as exogenous data.

Based on this model, some authors (Puga and Venables, 1996) have built an input-output matrix model in which a sector in the upper layer of the production structure produces inputs for several sector of the concentration by territory and sector, and the specialization sequences followed by the regions that become industrialized as markets expand.

By introducing the firm-external economies in this NEG's model, i.e. the mechanism of increasing returns for the production activity as a whole (and not only at the level of individual firms or establishments), it may be claimed that some countries may specialize in the production of certain consumer goods, giving an advantage to some countries to the detriment of others (Krugman 1987). Hence, some explanation may be given to the Italian advantage in tile production or to the British dominance of the financial service world. However, not only this third model with its derived external economies are capable of accounting for the specific localizations in the countries (Porter, 1990).

1.3.3 From the spatial concentration to the spatial-sectorial specialization: the introduction of Marshallian Externalities

Whether increasing returns are internal or external to firms, the logical consequence of location-specific externalities is the geographic concentration of the economic activity in a small number of locations. To counterbalance, the degree of space agglomerations depends on the strength of scale economies, the scope of transport costs, and the importance of congestion costs (derived from the presence of immobile factors or non-traded goods).

If average production costs decline as the scale of production, at the firm, industry, or regional level, rises, to concentrate production in a particular location will be beneficial.

As we have seen, the NEG model implied a change of perspective from the non-differentiated spatial economic concentration towards the specialized sectorial territory concentration. The emphasis was on the role of external economies, but only on the pecuniary economies derived from the large size of markets resulting from the territorial concentration of a large number of vertically related firms.

However, the model predictions can be subsequently reinforced if not only pecuniary economies are considered, but also non-pecuniary economies derived from spillovers between workers and learning across firms. The source of these externalities is not made explicit, but in general it is possible to imagine that the dense concentration of firms promotes learning and the exchange of knowledge, as in Marshall.

Lucas (1988) proposes a dynamic version of these externalities, in which workers learn from one another. If a worker becomes more productive through education or training, all workers in a location will also become a bit more productive. In the Lucas model, each region is a closed economy, so there is no space agglomeration per se. In a bibliographical review on the works of some economists about geographic concentration, Hanson (2000) informs that Black and Henderson (1999), building on the work by Eaton and Eckstein (1997), present a dynamic model of city formations, which combine the agglomeration economies in Henderson (1974) with the localized human capital spillovers of Lucas. This body of theories suggest that if there are spillovers in the accumulation of human capital, a worker will be more productive as the workers with whom he or she shares a given location become more educated.

External economies have a complex nature that is difficult to reduce to a single dimension. We can speak of economies with a local or international scope, with a technological or pecuniary technology, that are static or dynamic in nature, and finally, that they may have an inter- or intra-industrial scope. Hence, the empirical work about space agglomerations has to face an identification problem. The externalities that contribute to space agglomeration, such as spillovers between workers, learning across firm, or cost and demand linkages between local industries, are difficult to observe. We are left to infer their presence from the covariance of observed variables, such a wages, employment and output (Dumais, Ellison and Glaeser 1997 and Hanson 1998).

Marshallian externalities

Among the externalities linked to a territory and, through it, to a certain regional production structure, are static localization economies related to the access to certain production resources and low costs to access markets, and urbanization economies, i.e. those linked to the demand of intermediate goods and services provided to firms. Simultaneously, dynamic intra-industrial economies -within the same activity-, or inter-industrial -among different productive activities-, reflect the presence of external effects of a technology and/or pecuniary nature. All of these forces affect the territories, and thus the effectiveness of resident establishments and the firms' ability to growth. It must be noted that when these external scale economies are significant, a country that started to product a certain good before others, can consolidate its international industrial presence, although other country may be in a condition to produce the same goods at a lower cost. Seen from this perspective, "history counts".

In current terms, Fujita and Thisse (2002) summarize the relevant externalities that contribute to cluster formation:

- Mass production (the internal economies that are identical to scale economies at the firm's level);
- The availability of specialized input services;
- The formation of a highly specialized labor force and the production of new ideas, both based on the accumulation of human capital and face-to-face communications; and
- The existence of a modern infrastructure.

As mentioned in section 1.3.1, the NEG considers only the first of these marshallian externalities -the forward-backward linkages among the firms.

Scitovsky (1954) considered two categories of external economies: "technological externalities" (spillover) and "pecuniary externalities". The technological externalities refer to the effects of non-market interactions produced through processes that directly affect the utility of an individual or a firm's production function, i.e. aimed at capturing the critical role of non-market institutions, which significance and role has always been strongly emphasized by geographers and space analysts (Saxenian 1994). On the contrary, pecuniary externalities stem from the interactions measured by the market, affecting firms and workers (or consumers) according to their involvement in the interchange measured by price mechanisms.

The distinctive feature of external economies is that they only affect the agents in the same geographical area, and their impact on other distant regions may be insignificant. The advantages of the production proximity are based on human beings' tendency to interact with other human beings, whereby distance can be an impediment. This need for interaction is gravitational, and its intensity may probably increase according to the number and type of agents in each territory, and decrease with the distance between two territories. The need for interaction acts as a centripetal force and the competition for the land as a centrifugal force (Fujita and Thisse 2002).

Technology externalities are essential to account for geographical clusters of a certain limited space dimension as highly-specialized scientific cities and districts. Jaffe, Trajtember and Henderson (1993) argue that in the United States, the geographic concentration of some sectors affect the localization of patent use. Mention to patents is more frequent in local environments, and usually come from the same statistical metropolitan area or state, therefore suggesting that knowledge dissemination is concentrated in terms of space. Ausdretsch and Feldman (1996) observe that external spillovers are probably more linked within a region where the new knowledge was built. Face-to-face communication within agglomerations encourages the continuous intercommunication of ideas, and therefore are a substantial part of innovations. Most likely, these spillovers owe their existence to face-to-face contacts. For example, Saxenian (1994) emphasizes the significance of this factor that is capable of transforming Silicon Valley into an effective productive system, arguing that informal talks are omnipresent and serve as a major source of up-to-date information about competitors, customers, markets and technology. In an industry marked by increasingly rapid technological changes and a strong competition, such informal communications offer a higher value than conventional business fora.

The singularity of the market structure arising out of external economies producing an interdependence between an industry firms and workers, is compatible with the presence of a large number of small and medium firms, as has been observed by Beccattini and Brusco in Italy, and does not necessarily imply the presence of large industrial corporations. The modelization of "industrial districts" marked by the presence of the marshallian "industrial atmosphere" is compatible with the competitive paradigm (Anas, Arnott and Small 1998). The works from Becattini (1979, 1990 and 2004), and Becattini and Musotti (2003), have made several important contributions to our knowledge about industrial districts based on this Marshallian concept applied to an Italian context. Becattini emphasizes the role of the cultural and historical background of the districts, and extended Marshall's analysis of the purely economic effects of agglomeration to a broader perspective, to include the social, cultural and institutional foundations of local industrial growth.

Specialization

A homogeneous space with mechanisms of competition is not compatible with the existence of economic agglomerations such as cities or specialized areas, which are explained by economies of scale and importance of forces such as diversity of intermediate goods and effects of matching process on the labor market that lead to the division of labor and increasing returns (Fujita and Thisse 2002). The advantages of specialization shown as increasing returns are likely to arise in the final goods sector when the intermediate goods sector is described by monopolistic competition model, while imperfect competition of the major labor markets, as in big cities, reduced the matching cost average. Both models show how the urban population growth has allowed the profits that are generated by specialization and matching.

Dynamic economies of intra-industry are characterized by the relative importance of externalities, both technological and pecuniary nature, linked to the geographical setting and the production structure of the regions. When externalities are large enough in relation to business costs, the agglomeration of each type of x firms within the same region is in balance. Because of these externalities, the agglomeration of firms in the same industry generates endogenous spatial inhomogeneities that allow the existence of a balance when such inhomogeneities are strong enough, and they generate new forces that are capable of overcoming the instability generated by local competitive prices mechanisms.

The evidence shows that both types of externalities: localization economies, defined as the profits generated by the proximity of firms that produce similar products; and the urbanization economies, defined by all the benefits associated with the level of total activity prevailing in a particular area, are the source of high levels of expertise and prosperous areas, and as demonstrated by Porter (1998), the main reason for the success of the industrial cluster in the global economy is the presence of strong localization economies. The ideas of the economies of location also explains the growth and success of industrial districts, ie regions that are home to many small companies that produce similar products which are undertaken by the localized accumulation of skills associated with workers residing in these places (Becattini 1990). Both in Porter's and in Becattini's approach, the advantages of concentration at a local level lie in work division through the fragmentation of the productive process among the firms (horizontal competition) and its territorial restructuring, i.e. the vertical cooperation (input-output linkage among firms over subsequent stages).

Some industrial districts are developing high-tech activities, while others operate labor-intense activities (Third Italy: Sassuolo specializes in ceramic tiles, textiles in Prato, Montegranaro in shoes, etc.). In each case, the combination of several factors are essential for localized accumulation of various types of knowledge within a region. In this sense, Prescot (1998), emphasizes that each region is characterized by a "social capital" that affects the total productivity and this capital can vary in each region. Recent works like the Fujita and Thisse (2002) emphasize the origin of such external economies and explore the implications of marshallian externalities in the spatial distribution of production activities, regarding them as a general technological externalities without specifying their specificity. In addition, they demonstrate that a certain externality can govern the expansion of wealth in a region with such a head start at the expense of another.

From a microeconomic Chipman (1970) assumes that firms in the same sector or industry will benefit from the high productivity if they are found together. Such externalities interact with the centrifugal forces generated by the market competition within the global economy and lead to the formation of various clusters. The localization economies are the agglomeration forces and at the same time the geographic proximity generates strong price competition by encouraging firms to be located separately from others to enjoy local market power. Also, when price competition is lower because of product differentiation, firms prefer to isolate themselves when are prices high transport. Because supposedly the spatial distribution of demand is not affected by the location and size of the clusters, reducing costs associated with the agglomeration offset to a greater extent the decline in exports (by contrast, firms could enjoy higher profits by being local monopolists). Consequently, transport costs tend to be low when firms are agglomerated. In other words, firms should be able to serve virtually all markets (globalization) to maintain the benefits associated with the formation of a cluster (localization). Silicon Valley is an example: the agglomeration occurs because firms can take advantage of the high levels of localization economies while they are able to continue selling a substantial part of its output in distant markets. The high degree of product differentiation allows firms did not enter a price competition by allowing the existence of any force of agglomeration that dominates the dispersion force. These forces are crucial in many different spatial configurations and probably stronger in modern economies. When the attraction by the output of an industry grows, more firms will tend to be located within the same cluster, whose relative size will grow at the expense of other firms. As a result, economic growth, measured by the relative importance of differentiated goods tends to encourage the geographic concentration of production. Consequently, the formation of a cluster would seem to depend on the relative strength of three different forces: the size of the localization economies, the intensity of price competition and the level of transportation costs, which give rise to such structures market very different: competition within a small group and within a large group of firms. An example might be the first of the German chemical cluster, in which a small number of large companies are located together not only to reduce production costs but also as a strategy to their rivals. In relation to the second, might be some of the clusters in Italy, where companies have a negligible impact on other location choice but this affects production costs because this factor depends on where your competitors are located. In this sense, Fujita and Thisse (2002) conclude:

> Small initial advantages may lead to the emergence of a strongly polarized space once we explicitly account the existence of localized production externalities, natural amenities, or both. This effect is magnified when the mobility of factors or the transportability of products are high, or both. Indeed, either of the two possibilities allows the localized externalities to display their full impact.

In the real world, some territories traded goods instead of producing them. Agglomeration and trade are not inconsistent because the economies of scale are external to firms but internal at the industry, therefore, some firms producing the same goods by the advantages of being located together. On the other hand, add a new employee in such areas leads to higher average cost per person commuting. Accordingly, when external economies are present in an industry, specialization is the better exploitation of scale effects.

The degree of increasing returns vary with the good produced, so territories specialized in the production of different products have different sizes. By varying the economies of scale, the territories will have different levels of congestion and commuting cost. Henderson (1997) demonstrated empirically that small and medium cities tend to specialization. Therefore, the trade-off between increasing returns and commuting cost between territories is as follows:

- when increasing returns are strong in the production of an asset, the relative amount of regions that produce it decreases when they increase their size;
- when increasing returns grow, increasing the relative number of specialized areas, where these are small.

On the other hand, Duranton and Puga (2000) emphasize that some territories specialize in the production of a few products or services while others are more diversified. Historically, territories are diversified when commercial costs between them are truly low. One reason for diversification is the ability to exploit economies of scale associated with the large variety of intermediate goods and public services exist (Goldstein and Gronberg 1984). If the demand for final products of a territory-sector is not perfectly elastic, is not profitable for the industry to grow to a certain limit. Two different industries located in the same territory, which has a strong intermediate sector and a large quantity of public services, every industry becomes more productive. On the other hand, this co-localization requires more workers in the same city, hence longer commuting. Therefore, both industries in final products must pay higher wages. In this case, the balance between the gain in productivity of the two industries is given by the existence of a large intermediate sector and the commuting cost between the diversified territories (Adbel-Rahman 2000). A second reason is that the territories diversified smoothing the random shocks affecting specific industries. In this case, a territory is seen as a portfolio of activities. When an activity is negatively affected, workers have the option of moving to work in other sectors. Expected wages are higher than in a specialized area (Krugman 1991b). Duranton and Puga (2001) argue that when firms manage to be experts in what they produce will be relocated to those areas more specialized.

1.3.4 Relationship between specialization, growth and policy

The main effect of Inter-regional and international integration is the increase of economic efficiency in the space economy. However, based on the static models in which the number of firms and goods determined by the economy parameters are constant, Fujita and Thisse (2002) conclude that:

> market expansion may well be accompanied by the development of some core regions whose wealth is, in part, obtained at the expense of peripheral regions -the average welfare in the region accommodating the modern sector rises, but it decreases in the other region.

In order to know how increase and location affect each ones, we need to know first whether regional discrepancies increase or decrease over time, as well as the reasons for such potential convergence or divergence. Space and time are intrinsically mixed in the economic development process, while agglomeration and growth are a complex phenomena in themselves. Some examples of an integrated analysis of such concepts are the works by Fujita and Thisse (2002), that set forth two models based on the mobility of qualified workers, and the intensity of the spillover effects among regions. The works by Martin and Ottaviano (1999 and 2001) that studied the reciprocal influences between growth and location should also be mentioned. In addition, other works such as Donato and Haedo's (2002) and those by Donato, Haedo, Reynolds and Rocha (2008), empirically demonstrate that during periods of strong macro-economic crisis, firms closedown and the fall in manufacturing employment is lower in specialized regions.

Small cores specialized in a certain activity enable access to intra-sectorial economies, deriving significant implications for regional development policies. If external intra-sectorial economies are too strong, environments specialized in a certain activity will obtain the highest benefits from the creation of external effects. If on the contrary, intra-sectorial externalities are too weak, the diversified areas will create the greatest crossed external effects. Generally speaking, intra-sectorial economies are key for many industries, and determine the specific localization pattern.

The connection between growth and geography gets stronger inasmuch as the regional specialization in innovative activities is considered the result of the combination of abilities and skills developed in these regions. In this connection, Hirschman (1958) argues that growth is localized because social and technological innovations tend to be spatially clustered where dissemination between places is low. Feldman and Florida (1994) observe that in the past century, innovations were geographically clustered in areas where research and development-oriented firms and universities were established. This approach suggests that the development process is similar to that accounted for in the creation of regional agglomerations. Hence, agglomeration may be considered the territorial counterpart of economic growth. Fujita and Thisse (2002) conclude that the RD sector appears as a strong centripetal force at a multi-regional level, amplifying the circular causation at the core of the core-periphery model, i.e. agglomeration and development go hand in hand, putting this debate at the core of the economic policy of industrialized countries.

When the economy moves from dispersion to agglomeration, innovation develops at a faster pace. In this connection, Fujita and Thisse (2002) demonstrate that:

> even those who stay put in the periphery are better off than under dispersion, provided that the growth effect triggered by the agglomeration is strong enoughIn other words, agglomeration gives rise to regressive income distribution effects.

Additionally, when transport costs are low enough, modern and innovative sectors tend to concentrate in the same region, while other regions specialize in the production of traditional goods. This is so even when the number of firms that operate in these modern sectors may increase over time, regardless of whether or not technologies are transferable among the different regions. Fujita and Thisse (2002) demonstrate that agglomeration and growth are reinforced, thus confirming the results obtained in different contexts by Martin and Ottaviano (2001). An interesting implication of these conclusions is that the policies that encourage dispersion may potentially disrupt the global economic growth.

The rationale supporting the role of the cities in the economic growth is that these are considered the main social institutions in which technological and social innovations are developed through market and non-market institutions. In addition, the specialization of the cities changes over time creating diverse geographical patterns of economic development. For this reason, cities are the railway engines of growth.

Regional discrepancies are considered socially undesirable, and are critical from a political point of view. Further, the growth of regional disparities does not necessarily imply the impoverishment of peripheral regions. If there is proof of the persistence of these disparities, or even worse, that agglomerations give rise to peripheries in worse conditions, governments and international bodies should work on the design of active policies to encourage a more equitable distribution of wealth among nations and regions. Urban externalities are not necessarily negative, while increasing profits can be a strong force in favor of geographical concentration. Therefore, there is a non-presumption related to the direction in which governments should go with their regional and urban policies.

The NEG models should be a priority object for government intervention, since they suggest that under certain circumstances, the intervention of small-scale policies may have a strong effect, probably on a permanent basis. The lack of formal NEG models about the potential implications for public policies lies in the difficulty of going from small indicative models to models with an empirical base, that can be used to evaluate specific policies.

The coexistence of centripetal and centrifugal forces is generally relevant, and both create external effects. In addition, there are market failures for both types of agglomerations, either if they are too large (congestion and contamination), or too small (links and positive externalities derived from greater activity). Yet, geography is a critical issue for development, and undoubtedly major impacts of this type may affect policy-making. Model optimization is closely linked -although not identically- to the impact of public policies. It could be argued that considering a model's efficiency and optimization, and comparing it to the equilibrium conditions, provide a better understanding of the model's properties, even disregarding whether the results reached should or shouldn't affect the implementation of public policies (Fujita and Krugman 2004). To summarize, understanding the regional and urban growth based on an efficiency, equitable and optimum agglomeration, is critical to improve our knowledge about how a modern economy can develop.

To conclude, do economies grow faster if they are concentrated in the space?

In the police debate, increasing specialization has been welcomed, for example in the European or North American integration process, since it increases productivity. Rising concentration on the other hand, specifically the concentration (agglomeration) of economic activities in the core part or in the north, has been more controversial as it may aggravate asymmetries or differences in per-capita income.

This recent theoretical work generally supports the view that spatial proximity encourages economic growth. Martin y Ottaviano (1999) state that agglomeration and growth are mutually reinforcing processes. According to Fujita and Thisse (2002), growth and agglomeration go hand-in-hand. Baldwind and Martin (2004) claim that given localized spillovers, spatial agglomeration is conducive to growth.

As claimed by Martin (1999), the complementarities between growth and spatial concentration has a remarkable practical significance for the public policy, since it may imply supporting the economic growth at a national level at the expense of supporting the most underdeveloped regions in the country.

However, this complementarities can be non linear. Indeed, according to some authors agglomeration promotes growth at the early stages of development, but has not, or is even detrimental, in economies that have reached a certain income level (Williamson 1965). Williamson suggests that agglomeration matters most at the early stages of development. When transport and communication infrastructure is scarce and the reach of capital markets is limited, efficiency can be enhanced by concentrating the production in space; but as infrastructure improves and market expands, congestion externalities may favor a more dispersed economic geography. The dynamic game of agglomeration has to be weighed against the cost of static congestion diseconomies. The relative importance of these two effects changes over the different stages of development.

There are a relatively small number of empirical studies about the causal relationships between agglomeration and growth. Some of the most significant studies that analysis the positive relationships between urbanization, spatial inequality, industrialization and economic development, are those by Bairoch 1993, Hohenberg and Lees 1985, and Hohenberg, 2004.

The econometrical studies about the impact of agglomerations on growth are even scarcer. But based on this perspective, the most recent and significant work we are interested in is the work by Brülhart and Sbergami (2008), which explores the causal link running from agglomeration to growth, mediated by stage of development and external openness. These authors empirically investigate the impact of the within-country spatial concentration of the economic activity (agglomeration) on country level-growth. They assemble the most comprehensive data base used for this purpose to date, combining cross-section and panel data analysis of a large country level dataset with panel analysis of sectorially and regionally disaggregated data for Europe. Since agglomeration on the growth effects across sector, they investigate the impact of agglomeration of the growth of individual sectors in addition to study aggregate economic growth.

Brülhart y Sbergami (2008) have found evidence supporting the “Williamson hypothesis”: agglomeration boosts GDP growth only up to a certain level of economic development. The critical level is estimated at around 10,000 US dollars in 2006 prices, corresponding roughly to the current development level of Brazil or Bulgaria. This implies that the benefits of agglomeration become increasingly unimportant, and that the trade-off between national growth and inter-regional equity may gradually lose its relevance as the world’s economy continues to growth. Conversely, it also means that it is in the poorest countries where policies aimed at inhibiting spatial economic concentration are most damaging in terms of foregone growth.

1.4 The purpose of this research

The intensity of regional specialization in specific activities, and conversely, the level of industrial concentration in specific locations, has been used as a complementary evidence for the existence and significance of externalities. Besides, economists have focused the debate mostly on disentangling the sources of specialization and concentration processes according to three vectors: natural advantages, internal and external scale economies.

The arbitrariness of partitions plays a key role in capturing these effects and becomes potentially more dangerous the more unequal become the elements of it in terms of area. In this sense, as we mention in Section 1.2.4, the selection of the partition would have to reflect the actual characteristics of the economy. Therefore, above mentioned effects coupled with the effect of crowding can appear in different geographical scales and involve several levels of sectorial breakdown, that is, a certain spatial scale is not necessarily similar to another.

Thus, the identification of spatial boundaries to measure specialization becomes critical, since most likely the model will be adapted to different scales of distance and be influenced by different types of externalities or economies of agglomeration, which are based on the mechanisms of interaction with particular requirements of spatial proximity.

This work is based on the analysis of the spatial aspect of economic specialization supported by the manufacturing industry case. The main objective is propose for discrete and continuous space: i) a measure of global specialization; ii) a local disaggregation of the global measure; and iii) a spatial clustering method for the identification of specialized agglomerations.

In Chapter 2 specialization is approached in terms of stochastic independence: non specialization is viewed as the case where the joint proportion of employees of a region in an specific activity is equal to the product of marginal proportions of this region and activity; equivalently, the activity distribution within this region is the same as the global distribution at the country level. Hence, we propose a series of non parametric dependence measures, derived from the goodness-to-fit of the above hypothesis of non specialization, as natural measures for international comparisons of the global specialization level in addition to the basic weighted or unweighted indexes based on the Lorenz curve commonly used. On the other hand, the appropriate grouping of rows and columns of a two-way contingency table can often simplify the analysis of association between two categorical random variables. Hence, rows and column groupings have received considerable attention and have been driven by: i) decomposing global measures of non independency; ii) a focus on, and accordingly a better understanding of the sources of non independency; and iii) avoiding tables with too many cells, a larger proportion of which would be empty or nearly empty. Similar motivations are present when grouping activities and regions for a specialization analysis. Thus, we propose an automatic grouping procedure of regions and activities based on hierarchical clustering and correspondence analysis (HCCA), defining a goodness of association measure for a given collapsed table. The goal is to measure the effectiveness of the HCCA to preserve association while reducing the table dimension. The proposed quantity measures the gain in association produced by the HCCA method compared to the association that would be expected under a random grouping strategy, that enabled us to i) significantly reduce the size of the original table and obtain a collapsed table with low level of information loss vis-à-vis the degree of original specialization; and ii) identify the homogeneous regions according to the industrial structure in terms of sub, and over specialization activities in large two-way contingency tables. The study cases of this chapter refer to the cases of Argentina, Brazil and Chile, based on employment data of manufacturing industry at two digits of ISIC Rev.3. This chapter focuses on the detection of specialization through the pattern of activity concentration among regions, without considering the distance among them, in which regions are defined according to process-exogenous criteria, namely administrative entities.

In the same way as Chapter 2, Chapter 3 is based on the administrative entities but now we introduce contiguity among regions as distance criteria for the identification of specialized agglomerations in discrete space. In addition, the externalities arising from the proximity among firms, i.e. externalities and location, are concerned with firm interaction in a certain region. The spatial externalities are significant for the plant location distribution, i.e. the outcome of firms location choice. Consequently, should we measure the strength of these spillover effects, the unit of analysis would be in favor of firms. Simultaneously, dynamic intra-industrial economies -within the same activity, or inter-industrial -among different productive activities, show the presence of external effects of a spillover and/or pecuniary nature. These forces affect the territories, and thus the effectiveness of resident firms, and the firms' ability to growth. We define a probability model for the location of firm which helps us identify spatial clusters of specialized industrial allocation for a given specific manufacture sector. The approach is closely related to the cluster-identification methods proposed by Besag and Newell (1991), Kuldorff and Nagarwalla (1995), and Kuldorff (1997), that have been used to detect disease cluster in epidemiology. The basic idea is to develop a probability model of multiple clusters, called cluster schemes. Simply put, a cluster scheme is a space partition through which it is postulated that firms are more likely to locate in cluster partition than elsewhere. Thus, in this partition the model is equivalent to a multinominal sampling model. The method starts by postulating a null hypothesis of no specialized agglomeration, i.e. no clustering in terms of the uniform distribution of industrial locations across regions. Then, it continues testing this hypothesis on each activity by finding a single most significant contiguous cluster of regions with respect to this hypothesis. In other words, on a first stage we start with an individual region, and then the algorithm proceeds by adding contiguous regions to find the most significant clusters. The study cases of this chapter refer to the cases of manufacture industry of Chile, based on establishment data at two digits of ISIC Rev.3. At the end of this Chapter we show some remarks about the co-localization of firms in specialized agglomerations.

Now, in Chapter 4, the space is a unique continuum. The spatial process approach is based on geocodified data, and aims at assessing the power of attraction from a local space perspective. From the point of view of a non-homogeneous Poisson process, firm localization points are randomly distributed, and disjoint area counts are mutually independent, each based on Poisson's distribution according to which the intensity parameter forms a finite measure of the reference space, in this case a bi-dimensional space. This measure may be interpreted as the representation of the differentiated power of attraction of the space and, in the case of a specific area, the expected value of the number of locations in such area. As we mention in Section 1.2.4, the availability of geocodified data unleashes the need to quantify the specialization level of a certain activity in a particular point of the space. With this purpose, we present the methodology that uses kernel density estimators as a key tool to define a local specialization measurement for a point x as an extension of the well-known local quotient measurement to the continuous space. In addition, we propose a possible Average Specialization Measure (ASM) for continuous space. For the identification of specialized agglomerations in continuous space, we use the identification of statistical significance or significant feature of the specialization level, based on the method of bootstrap hypothesis testing proposed by Efron and Tibshirani (1993), to approximate the distribution of the local quotient under the non-specialization hypothesis. The study cases of this Chapter first analyze simulated data and next use geocodified data of firms of manufacture sector of Buenos Aires City at four digits of ISIC Rev.3.

Finally, Chapter 5 presents the summary and overall conclusions, and the potential directions to continue with this research project.

Chapter 2

Stochastic independence model approach for the measurement of global specialization

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

$$N_{i\cdot} = \sum\nolimits_{j=1}^{J} N_{ij} \tag{2.1}$$

column totals

$$N_{\cdot j} = \sum\nolimits_{i=1}^{I} N_{ij} \tag{2.2}$$

and the table total $N_{\cdot\cdot}$

$$N_{\cdot\cdot} = \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} N_{ij} = \sum\nolimits_{j=1}^{J} N_{\cdot j} = \sum\nolimits_{i=1}^{I} N_{i\cdot} \tag{2.3}$$

This chapter is focused on the detection of specialization through the pattern of activity concentration among regions, without considering the distance among them.

The underlying idea is to compare the distribution corresponding corresponding to each file i, and consequently to the regions, with the aggregate distribution corresponding to the column totals. For this purpose, we switch from the count N_{ij} to the file proportions $c_{ij} = N_{ij}/N_{i\cdot}$. Thus, for each $i = 1, ..., I$, the vector $(c_{i1}, ..., c_{ij}, ..., c_{iJ})$ corresponds to the regional distribution of activities (in particular $c_{ij} \geq 0$ and $c_{i\cdot} = \sum_j c_{ij} = 1$), whereas vector $(c_{\cdot 1}, ..., c_{\cdot j}, ..., c_{\cdot J})$, with $c_{\cdot j} = \sum_i c_{ij}$, correspond to the activity distribution aggregated at a country level, particularly $I^{-1} \sum_j c_{\cdot j} = 1$.

2.1 Basic indexes to measure local specialization

The indexes commonly used throughout the economic literature to describe the phenomenon of local specialization, i.e. the analysis of the regional industry structure or concentration of the regional activity, are based on the Lorenz curve. Concentration measures, such a those of Lorenz or Gini, have been developed for numerical variables, endowed with the natural order ($a < a + 1$). Here, activities are considered as the distributed variable, each one ordered according to their relative importance at the country level. Thus, the activities are labeled in such a way that $c_{\cdot j} \leq c_{\cdot (j+1)}$. This ordering allows us to construct a cumulative distribution of the activities, distributions (C_{ij}) and $(C_{\cdot j})$, i.e. to construct cumulative distribution functions $C_{ij} = \sum_{k=1}^{J} c_{ik}$ and $C_{\cdot j} = \sum_{k=1}^{J} c_{\cdot k}$.

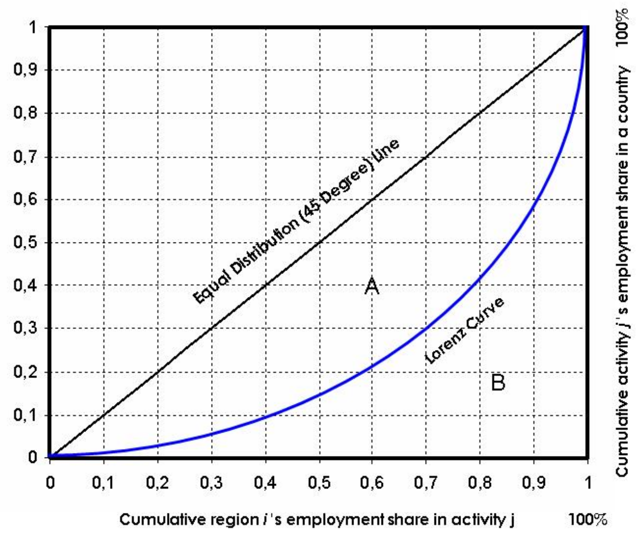
In order to compare the activity distribution of region i to that of the country, we build in the spirit of lorenz curve a bivariate graph with coordinates (C_{ij}) and $(C_{\cdot j})$, where the main diagonal would correspond to a region i with the same activity distribution as that of the country. Sorting the observations in increasing order by gradient $c_{ij}/c_{\cdot j}$, the relationship between C_{ij} and $C_{\cdot j}$ is the Lorentz curve. Fig. 2.1 shows the Lorenz curve.

The Lorenz curve always starts at (0,0) and ends at (1,1) and is completed by linear interpolation among the ordered points corresponding to the activities j.

Gini specialization coefficient GI_i

Many indexes try to summarize the graphic information provided by the Lorenz curve in a quantitative measure that shows the difference between that curve and the situation of perfect equality. The most popular of these measures is the Gini coefficient GI_i (Gini 1912). That is, GI_i can be geometrically defined, as in Fig. 2.1, as the ratio of two geometrical areas in the unit box: (a) the area between the

Figure 2.1: *Lorenz Curve*



line of perfect equality (45-degree line in the unit box) and the Lorenz curve, which is called area A and (b) the area under the 45-degree line, or areas $A + B$. Since areas $A + B$ represent the half of the unit box, that is, $A + B = \frac{1}{2}$, the GI_i can be written as $\frac{A}{A+B} = 2A = 1 - 2B$.

For a discrete distribution, we can compute $(C_{\cdot j})$ and (C_{ij}) and then the area below Lorenz curve [1]

$$B = \frac{1}{2} \sum_{j=1}^{J} (C_{\cdot j} - C_{\cdot j-1})(C_{ij} + C_{ij-1}) \tag{2.4}$$

$$\Rightarrow GI_i = 1 - \sum_{j=1}^{J} (C_{\cdot j} - C_{\cdot j-1})(C_{ij} + C_{ij-1}) \tag{2.5}$$

There are various expressions of this definition. For example, Yao (1999) adopted a spread sheet approach using this method. Osberg and Xu (2000) modified the definition to accommodate the complex sampling survey data.

[1] Optionally, and also from a geometric point of view, we can define GI_i as two times less the area under the curve of Lorenz (Rao 1969).

In this work, we calculate GI_i from Brown's formula

$$\Rightarrow GI_i = \left| 1 - \sum_{j=1}^{J-1} \left(C_{\cdot j} - C_{\cdot j-1} \right) \left(C_{ij} + C_{ij-1} \right) \right| \tag{2.6}$$

GI_i is a measure of specialization based on the variability in the industrial structure of a region compared to the industrial structure of the whole country. GI_i takes values in the range [0;1], i.e. between 0 (perfect equality) and 1 (maximum inequality), and will be greater the further away it will be from the line of perfect equality. A value 0 means that a region has the same activity j's employment share with respect to the whole country, while on the contrary, a value 1 denotes extreme inequality, i.e that the industrial structure of a region is completely different from or unequal to the whole country.

From the above expression, it is possible to obtain an aggregate measure of specialization for the whole country

$$GI = \sum_{i=1}^{I} w_i \; GI_i \tag{2.7}$$

where w_i is the weighting assigned to region i according to its economic or demographic size, with $\sum_{i=1}^{I} w_i = 1$.

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 about the Gini coefficient see Atkinson (1983) and Lernan and Yitzhaki (1989).

Krugman index SK_i

The index SK_i proposed by Krugman (1991a) is a measure of relative specialization, based on GI_i expressed as half of the average relative difference (Kendall and Stuart 1963). This index captures the gap between the activity structure of region i and the average of the activity j structure of the other regions. It is defined as:

$$SK_i = \frac{1}{2} \sum_{j=1}^{J} abs\left(c_{ij} - \overline{c}_{ij}\right) \tag{2.8}$$

where

$$\overline{c}_{ij} = \frac{\sum_{k \neq i}^{I} N_{kj}}{\sum_{k \neq i}^{I} \sum_{j=1}^{J} N_{kj}} \tag{2.9}$$

The SK_i index takes a zero value if the activity structure of region i is identical to the average of the other regions. Given the normalization used here, the maximum value of SK_i is equal to 1 when the activity structure of one region differs completely from the rest of the country. Likewise, from above expression it is possible to obtain an aggregate measure of specialization for the whole country

$$SK = \sum_{i=1}^{I} w_i \; SK_i \tag{2.10}$$

where w_i is the weighting assigned to region i according to its economic or demographic size, with $\sum_{i=1}^{I} w_i = 1$.

The LQ_{ij} gradient

Gradient $c_{ij}/c_{\cdot j}$ is the Hoover-Balassa Local Quotient coefficient LQ_{ij}. The LQ_{ij} gradient is an index to compare the characteristics or activities of a local area across a larger system (see Gibson et al. 1991; Beyene et al. 2003; Beyene and Moineddin 2005; and Brenden et al. 2008), and is commonly used by geographers, health professionals, and economists to quantify and compare local conditions (e.g. industry share) to an overall, aggregate condition.

LQ_{ij} is often used in economic-based analysis as the initial step to begin to understand what sectors are driving a region's economy. It is a measure of the relative concentration or specialization of the economic activity (i.e. a comparative approach), and frequently uses employment as gauge.

A particular location or region i is defined as specialized in a single activity j if the employees' share in the region exceeds the national share, i.e. a region is said to be specialized in an single activity if it has an over-representation in terms of employment. A region with a high LQ_{ij} may mean that the local economy is specialized: firms take the profits of the firm-internal economies of scale, and the firm-external scale economies derived from the spatial concentration of employees.

Thus, the LQ_{ij} index for region i and activity j is defined by

$$LQ_{ij} = \frac{c_{ij}}{c_{\cdot j}} = \frac{N_{ij}N_{i\cdot}^{-1}}{N_{\cdot j}N_{\cdot\cdot}^{-1}} \tag{2.11}$$

where the numerator denotes activity j's share of employment in region i and the denominator denotes its share in the whole country.

Complementary to the previous index, we can also define the employee "Industry Quotient" IQ_{ij} along with the idea that an industry should be concentrated in a particular region if its share of employees in the region exceeds the corresponding national share. Indeed, we have

$$LQ_{ij} = \frac{N_{ij}N_{i\cdot}^{-1}}{N_{\cdot j}N_{\cdot\cdot}^{-1}} = IQ_{ij} = \frac{N_{ij}N_{\cdot j}^{-1}}{N_{i\cdot}N_{\cdot\cdot}^{-1}} \tag{2.12}$$

Values above 1 mean that the region (activity) is relatively specialized (concentrated) in the activity (location), as it has a relatively lower number of employees than it would be predicted, based on its aggregate employee's share, and values approximately equal to 1.0 indicate that a region (activity) has a number of employees compatible with the national average of this activity (region). A series of rules of thumb for the determination of cut-off values have been suggested by some authors because a region with very few employees can also show very high values of LQ_{ij}. In this sense, Botham et al. (2001) use a cut-off value 1.25 and Isaksen (1996) and Malmberg and Maskell (2002) prefer a more restrictive definition and use a cut-off of 3, while Donato and Haedo (2002) use a LQ_{ij} weighted for N_{ij}.

However, one commonly observed limitation of LQ_{ij} is its widespread use as only a point estimate without an accompanying confidence interval. Although the calculation of LQ_{ij} is straightforward, constructing confidence intervals to assess uncertainty in the LQ_{ij} estimates is difficult because the index is a ratio (for more details, see Beyene et al. 2003 and Beyene and Moineddin 2005). Closed-form solutions for constructing confidence intervals based on approximation methods are available. Gustafson (1988) proposed a method for constructing confidence intervals for Proportional Size Distribution (PSD) indices based upon a normal approximation to the binomial distribution. However, Beyene and Moineddin (2005) found that profile likelihood confidence intervals were narrower than approximation confidence intervals when sample sizes were small and when P-values were extreme; these situations often occur for PSD indices. Thus, the profile likelihood method may be the best method for constructing LQ_{ij} confidence intervals for PSD indices.

2.2 Stochastic independence model

For international comparisons, and other similar purposes, it is necessary to summarize the level or the degree of specialization of a country using the above indexes GI_i and SK_i. For this purpose, a simple weighted or unweighted average of values at regional level is usually used. For more details, see Amiti (1999), and Midelfart-Knarvik, Overman, Redding and Venables (2000).

Understanding the structure of specialization independently of its spatial pattern may be illuminated by imagining a random experiment where an employed is randomly allocated independently to a region i and activity j. Therefore, each employee generates a bivariate random variable $\mathbf{X}=(I,J)$ with values in the finite set of pairs $(i,j), i \in R = \{1,...,r\}$, $j \in A = \{1,...,a\}$ and probabilities[2]

$$P(X=(i,j)) = p_{ij} \quad , \quad X \sim \mathcal{B}er(\mathbf{P}) \quad , \quad \mathbf{P} = [p_{ij}]_{i,j \in R \times A} \tag{2.13}$$

where $\mathcal{B}er(\mathbf{P})$ is to be intended as a generalized Bernoulli.

These probabilities give a joint distribution of (I,J), with marginal distributions

$$P(I=i) = \sum_{j \in A} p_{ij} = p_{i\cdot} \quad , \quad P(J=j) = \sum_{i \in R} p_{ij} = p_{\cdot j} \tag{2.14}$$

Thus, the N observed employed X_l, $l=1,...,N$, give rise to N i.i.d observed bivariate vectors, $\mathbf{X}_1,...,\mathbf{X}_N$, $X_l \sim \mathcal{B}er(\mathbf{P})$, $l=1,...,N$.

Define an *indicator matrix* $\mathbf{Y}_l$ associated to observation $\mathbf{X}_l = (I_l, J_l)$: $\mathbf{Y}_l$ is a $\{r \times a\}$-matrix so that

$$\mathbf{Y}_l = (Y_{ij}) \quad , i \in R \ , \ j \in A \tag{2.15}$$

[2] A measure P is a function defined on a σ-algebra Σ over a set $\mathbf{X}$ and taking values in the extended interval $[0,\infty]$ such that the following properties are satisfied: 1) the empty set has a zero measure: $P(\varnothing)=0$, and 2) countable additivity or σ-additivity: if E_i, $i=1,...,\infty$, is a countable sequence of pairwise disjoint sets in Σ, the measure of the union of all the E_i is equal to the sum of the measures of each E_i: $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

where

$$Y_{ij} = \begin{cases} 1 & \text{if } (I_l, J_l) = (i, j) \\ 0 & \text{otherwise} \end{cases}$$

Thus, we will say that $\mathbf{Y}_l$ has the multinomial distribution

$$\mathbf{Y}_l \sim \mathcal{M}_{r\times a}(1, \mathbf{P}) \tag{2.16}$$

where $\mathbf{P}$ is the $\{r \times a\}$-matrix of cell probabilities p_{ij}. Our final matrix from the observed count is thus

$$\mathbf{N} = \sum\nolimits_{l=1}^{N} \mathbf{Y}_l \sim \mathcal{M}_{r\times a}(N, \mathbf{P}) \tag{2.17}$$

Then the table $\mathbf{N} = [N_{ij}]$ may be viewed as a two-way contingency table, where $N_{ij} = \sum_{l=1}^{N} 1\!\!1_{\{X_l=(i,j)\}}$ is the number of employees in cell (i, j).

From above formula (2.11) for LQ_{ij}, the employment in region i of activity j, N_{ij}, provides 3 cases:

$$i)\ \ \frac{N_{ij}}{N_{\cdot\cdot}} = \frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}} \quad \text{“}no\ specialization''\text{} \tag{2.18}$$

$$ii)\ \ \frac{N_{ij}}{N_{\cdot\cdot}} > \frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}} \quad \text{“}over\ specialization'' \tag{2.19}$$

$$iii)\ \ \frac{N_{ij}}{N_{\cdot\cdot}} < \frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}} \quad \text{“}sub\ specialization'' \tag{2.20}$$

Proportions $\left[\frac{N_{ij}}{N_{\cdot\cdot}}\right]$ and $\left[\frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}}\right]$ may be viewed as two (empirical) distributions on a bi-dimensional discrete variable.

“No specialization” implies that the joint proportion of employees of region i in activity j is equal to the product of marginal proportions of region i and activity j, i.e. the connection of specialization with stochastic independence is apparent. No specialization may be viewed as a null hypothesis of stochastic independence.

It is evident that specialization arises from the interaction between regions and activities. Searching for a measure of specialization is to search for an association between the variables I and J. More formally, the null hypothesis of non specialization could be written

$$H_0 : \; P(I = i|J = j) = P(I = i) \quad \forall i \; , \; \forall j \tag{2.21}$$

The null hypothesis would be that I, J are independent, i.e. the joint distribution is the product of its marginals

$$P(I = i, J = j) = p_{i\cdot}p_{j\cdot} = p_{ij} \quad , \quad i \in R \quad , \quad j \in A \tag{2.22}$$

In order words, the non specialization hypothesis is equivalent to the independence hypothesis between the variables I and J. A natural goal is to find the regions (i.e. region i^*) and activities (i.e. activity j^*) for which $P(I = i^*|J = j^*) > P(I = i^*)$, for some $i^* \in R$ and $j^* \in A$. In this case, we will say that the activity i^* is specialized in region j^*.

Under H_0 we know only the form $p_{ij} = p_{i\cdot}p_{\cdot j}$ but we do not know the marginal probability $p_{i\cdot}$, $p_{\cdot j}$. A reasonable procedure is to estimate them under the hypothesis of independence. These estimates can be obtained as follows: we would estimate the marginal probability $p_{i\cdot}$ of an employed falling into the i-th category (for the first categorical variable) in the usual way, by the sample proportion. The sample proportion in this case is

$$\widehat{p}_{i\cdot} = N_{\cdot\cdot}^{-1} N_{i\cdot} \tag{2.23}$$

since $N_{i\cdot}$ is the count of all objets falling in the i-th category. Analogously

$$\widehat{p}_{\cdot j} = N_{\cdot\cdot}^{-1} N_{\cdot j} \tag{2.24}$$

This yield an estimated $\widehat{N}_{ij}$ value

$$\widehat{N}_{ij} = N_{\cdot\cdot} \; \widehat{p}_{i\cdot} \; \widehat{p}_{\cdot j} = \frac{N_{i\cdot} \; N_{\cdot j}}{N_{\cdot\cdot}} \tag{2.25}$$

2.3 Global measure of specialization

A measure of the goodness-of-fit of the no specialization hypothesis may be

- based on any divergence or distance between these two distributions; and
- interpreted as an empirical measure of the degree of specialization vs. non specialization.

The desirable properties of such measure include:

- indicating the degree of fit along a continuum bounded by values such as 0 and 1, where 0 represents a complete lack of fit and 1 reflects a perfect fit;
- be independent from sample size (higher or lower values would not be obtained simply because the sample size is large or small);
- have known distributional characteristics to assist interpretation and enable the construction of a confidence interval;
- a partition showing that an association that was significant for the overall table primarily reflects differences between some categories and/or groups of categories; and
- for international comparisons, the measure should not be affected by the number of categories.

Unfortunately, no index has been able to satisfy these conditions acceptably; further, not all researchers would even agree with all of these criteria (Bollen and Long 1993).

Examples of “well-informed” measures include the moment generating and characteristic functions, as well as many entropy functionals developed in information theory. The robustness of nonparametric implementation of entropy indices is one of the main reasons for the recent surge in their popularity. The interested reader is directed to Tjøstheim’s (1996) survey on the subject and Aparicio (1998), who report superior performance for nonparametric entropy measures of dependence over the traditional measures.

Entropies are defined over the distribution space which form the bases of independence/dependence concepts in both continuous and discrete cases. Entropy is also “dimension-less” as it applies seamlessly to univariate and multivariate contexts. For these reasons, Shannon’s mutual information function has been increasingly utilized in the literature (see Joe 1989 and Robinson 1991).

However, Shannon’s relative entropy and almost all other entropies fail to be “metric”, as they violate either symmetry, or the triangularity rule, or both. This means that they are measures of divergence, not distance. This is not a problem for testing purposes, but if someone is interested in comparing distances with other distances (e.g. for cluster analysis), then the triangle inequality is essential (for more details see Maasoumi and Racine 2002). In general, a divergence measure might serve just as well as a distance and/or as a basis for constructing a test for independence.

Dependence measures are based on the divergences between the m-dimensional density p and its counterpart under the null hypothesis, q. Divergences are functionals of density pairs which, like distances, are equal to zero whenever $p = q$, and strictly positive otherwise.

The concept of statistical independence is well defined in terms of the joint distribution of variables and of examples of criteria that incorporate the divergence of joint distributions from the product of their marginals (see Gibbs and Su 2002 and Tjøstheim 1996 who use the Hellinger and several other measures).

In this section we consider tests for independence based on various dependence measures. Typically, the tests obtained through this approach are not invariant.

Pearson and likelihood ratio chi-squared tests of independence

For large samples, to test the hypothesis that two variables are statistically independent, or to test the joint distribution from the product of their marginals, the Pearson chi-squared test statistic, $\mathcal{X}^2$, and the likelihood ratio chi-squared statistics, G^2 are given by

$$\mathcal{X}^2(p \mid q) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij} - q_{ij})^2}{q_{ij}} \tag{2.26}$$

$$G^2(p \mid q) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}\log\left(\frac{N_{ij}}{q_{ij}}\right) \tag{2.27}$$

where N_{ij}, $i = 1, ..., I$, $j = 1, ..., J$, are the cell count in an $I \times J$ contingency table, and q_{ij} are the estimated expected frequencies under the independence hypothesis.

When the null hypothesis is true, $\mathcal{X}^2$ and G^2 have asymptotic chi-squared distribution with degrees of freedom, df, $(I-1)\times(J-1)$ as $n \to \infty$, with mean df and variance $2df$, and when $df \to \infty$, $\mathcal{X}^2_{df} \to \mathcal{N}$ (see for more Lancaster 1979).

These are the most popular tests of independence in contingency tables. However, the adequacy of asymptotic distribution depends both on the sample size N and the number of cells $C = I \times J$. For $\mathcal{X}^2$ test, Cochran (1954) suggests that a minimum expected value of 1 is permissible as long as no more than about 20% of the 8 cells have expected values below 5. For G^2 test, Koehler (1986) showed that the chi-square approximation is poor when N/C is less than 5. Goodman (1971), Bollen and Long (1993), and more recently Jackson, Gray and Fienberg (2008) gave grounds for using G^2/df in the model selection process. See Agresti (2002) for further discussions on the adequacy of chi-square approximation for sparse contingency tables.

Partitioning chisquared statistics helps to show that an association that was significant for the overall table primarily reflects differences between some categories and/or groups of categories. The sum of independent chi-squared statistics are themselves chi-squared statistics with degrees of freedom equal to the sum of the degrees of freedom for the individual statistics. Following Agresti (2002), for partitioning to lead to a full decomposition of G^2, the following are necessary conditions

- the degrees of freedom for the sub-tables must sum to the degrees of freedom for the original table;
- each cell count in the original table must be a cell in one and only one sub-table; and
- each marginal total of the original table must be a marginal total of one and only one sub-table.

With respect to the partition of chi-squared, Lancaster (1951) and Gilula and Haberman (1998) are based on the decomposition of treatment sums of squares in a one-way analysis of variance, while Hirschfeld (1935) provides an alternative based on canonical correlations.

Taking into account the previous Section, it seems only natural to propose as a global measure of specialization the P-value associated to the chi squared statistic of the contingency table, that is

$$P(\mathcal{X}^2_{(I-1)(J-1)} > \mathcal{X}^2) = 1 - F_{\mathcal{X}^2_{(I-1)(J-1)}}(\mathcal{X}^2) \tag{2.28}$$

However, this approach is not very useful, based on the following drawbacks:

- the number of employees N is quite large for most countries, in the order of millions;
- specialization is known to exist in most countries, so we know in advance that we are working under an alternative hypothesis; and
- the Chi squared statistic is a consistent statistic for the null hypothesis of independence.

As a result of these features, in most cases the P-value is zero. This trivial result does not allow us to use the P-value as a sensible measure of global specialization. Any comparison between countries would yield a dumb comparison between zero and zero. To solve this problem there are at least two approaches:

- use a test statistic that does not leave a fixed significance level, thus avoiding the fast convergence of the P-value to zero; and
- use ceratin characteristics of the data (contingency table) that are non-sensitive to the number of employees.

The $\mathcal{X}^2$ and Kullback-Leibler divergences

To test the null hypothesis of independency as opposed to the alternative hypothesis of specialization, it is possible to chose between the following information-type measures of difference of probability distribution known as f-divergences (see Csiszár 1967). For any convex function f, with $f(1) = 0$, we could define $d_f\ (p \mid q) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} f\left(\frac{p_{ij}}{q_{ij}}\right)$, whereas p represents the data and q represents a theoretical model of p.

The $\mathcal{X}^2$-divergence by

$$d_{\mathcal{X}^2}(p \mid q) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - q_{ij})^2}{q_{ij}} \tag{2.29}$$

have asymptotic chi-squared distributions with degrees of freedom $(I-1) \times (J-1)$ as $n \to \infty$. The adequacy of asymptotic distribution depends on the number of cells but not on the sample size.

In correspondence analysis this divergence is also known as total inertia $\mathcal{X}^2/n$, and can viewed as a measure of the magnitude of the total row (regions) squared deviations, or in an equivalent way, of the magnitude of the column (activities) squared deviations. Hence, total inertia is the sum of the squared standardized residuals $\mathbf{S}$

$$s_{ij} = \frac{(p_{ij} - q_{ij})}{\sqrt{q_{ij}}} \tag{2.30}$$

The total inertia can be expressed as the sum of row or column inertias, and this decomposition measures the partial contribution of the region i or of the activity j to the total inertia (or global measure of specialization), respectively (see for more details Jobson 1992). The next Section shows the decomposition equation.

The relative entropy or Kullback-Leibler divergence (Shannon's relative entropy) by

$$d_{KL}(p \mid q) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) \tag{2.31}$$

The relative entropy take values in $[0, \infty]$, and $d_{KL}(p \mid q) = 0$ if $p_{ij} = q_{ij}$ for all $i = 1, ..., I$ and $j = 1, ..., J$. This is not a metric, since it is not symmetric and does not satisfy the triangle inequality, $d_{KL}(p \mid q) \neq d_{KL}(q \mid p)$. However, it has many useful properties, such as being additive for independent processes. For more details, see Kullback and Leibler (1951) and Ali and Silvey (1966).

The probability P_{ij} that the employment pattern N_{ij} is realized under any distribution p_{ij} is given by the multinomial probability

$$P_{ij}(N_{ij}) = \left(\frac{N!}{\prod_{i=1}^{I} \prod_{j=1}^{J} N_{ij}!} \right) \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{N_{ij}} \tag{2.32}$$

Hence, the G^2 of the hypothesized distribution, q_{ij}, given N_{ij} is $\prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{q_{ij}}{p_{ij}} \right)^{N_{ij}}$. Thus the likelihood ratio chi-squared statistics is just $2n$ times the KullbackLeibler divergence between p and q

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \Rightarrow G^2 = 2n\ d_{KL}(p \mid q) \tag{2.33}$$

i.e. if E denotes the expectation with respect to the distribution p, $d_{KL}(p \mid q)$ may also be written as

$$d_{KL}(p \mid q) = E\left[\log\frac{p_{ij}}{q_{ij}}\right] \tag{2.34}$$

Thus, $d_{KL}(p \mid q)$, the same as in $d_{\mathcal{X}^2}(p \mid q)$, are independent from the sample size. This property is highly relevant for a comparison of the degree of specialization across regions and activities with large and small sizes, respectively.

In addition, and the same as in $d_{\mathcal{X}^2}$, d_{KL} may be defined with respect to any finite (measurable) partition of the sample space. Moreover, it is well known that there exists a powerful decomposition relation between the values of such indices for nested partitions of the sample space. For more details, see Kullback (1959), and Cover and Thomas (1991). This technique is particularly useful to identify the geographic structure of specialization, and for studying how this structure changes over time.

Let's suppose that the set of regions i, is partitioned into $M(< i)$ bundles of regions, where mth bundle, i_m, is composed of i_m regions ($\sum_{m=1}^{M} i_m = i$), so

$$\begin{aligned}
d_{KL}(p \mid q) &= \sum_{m=1}^{M}\sum_{i\in I_m}\sum_{j=1}^{J} p_{mj}p_{ij|m}\log\left(\frac{p_{mj}p_{ij|m}}{q_{mj}q_{ij|m}}\right) &(2.35)\\
&= \sum_{m=1}^{M}\sum_{j=1}^{J} p_{mj}\log\left(\frac{p_{mj}}{q_{mj}}\right) + \sum_{m=1}^{M}\sum_{j=1}^{J} p_{mj}\left\{\sum_{i\in I_m} p_{ij|m}\log\left(\frac{p_{ij|m}}{q_{ij|m}}\right)\right\} &(2.36)\\
&= d_{KL}(p_{mj} \mid q_{mj}) + \sum_{m=1}^{M}\sum_{j=1}^{J} p_{mj}\ d_{KL}(p_{ij|m} \mid q_{ij|m}) &(2.37)
\end{aligned}$$

The first term in the right hand side shows the d_{KL} among the regional bundles while the second term represents the weighted average of d_{KL} within each regional bundle. In other words, the d_{KL} for all regions can be decomposed into those representing the specialization among and within regional bundles.

Theorem. *The relative entropy d_{KL} and $d_{\mathcal{X}^2}$ divergence satisfy*

$$d_{KL} \leq \log(1 + d_{\mathcal{X}^2}) \tag{2.38}$$

In particular, $d_{KL} \leq d_{\mathcal{X}^2}$.

Proof. Following Gibbs and Su (2002), since log is a concave function, Jensen's inequality yields

$$d_{KL}(p \mid q) \leq \log\left(\int (f/g)\ f\ d\lambda\right) \leq \log(1 + d_{\mathcal{X}^2}(p \mid q)) \leq d_{\mathcal{X}^2}(p \mid q) \tag{2.39}$$

where the second inequality is obtained by noting that

$$\int \frac{(f-g)^2}{g}\ d\lambda = \int \left(\frac{f^2}{g} - 2f + g\right)\ d\lambda = \int \frac{f^2}{g}\ d\lambda - 1 \tag{2.40}$$

The Hellinger distance

We consider a normalization of the Bhattacharya-Matusita-Hellinger measure of dependence given by

$$d_H^2(p \mid q) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sqrt{p_{ij}} - \sqrt{q_{ij}}\right)^2 \tag{2.41}$$

The Hellinger distance assumes values in [0;1] and is symmetric, since $d_H^2(p \mid q) = d_H^2(q \mid p)$, and thus it is a distance contrary to other divergences. This index is a similarity coefficient indicating the correlation between two statistical distributions: the closer to zero is the index, the more the distributions are similar. By applying of the square root of the frequency, the index becomes quite robust to extreme values. Since it works with frequencies, the index is independent from the absolute quantities distributed. The squared power allows for a quantification of the absolute difference.

The Hellinger distance is not a metric. However, it has a useful property: it can be "factored" in terms of its marginals (Zolotarev 1983), enabling the representation of the distance between the distribution of vectors with independent components in terms of component-wise distances..

Reiss (1989) shows the relationship of d_H^2 among $d_{\mathcal{X}^2}$ and d_{KL} is as follows: $d_H^4 \leq 4d_{\mathcal{X}^2}$ and $d_H^2 \leq d_{KL}$.

2.4 Automatic grouping of regions and activities

The appropriate grouping of rows and columns of a two-way contingency table can often simplify the analysis of association between two categorical random variables.

Hence, rows and column groupings have received considerable attention and have been driven by:

- decomposing global measures of non independency;
- focalizing, and accordingly understanding better, the sources of non independency;
- avoiding tables with too many cells, a larger proportion of which would be empty or nearly empty.

Similar motivations are present when grouping activities and regions for a specialization analysis.

The purpose of this Section is to find regions with the same industrial structure in terms of sub, and over specialization activities in large two-way contingency tables

Finding parsimonious summaries of data sets and contingency tables created from them has been a long-term objective of statisticians. The traditional method of analysis using hierarchical log-linear models (HLLMs) as described in Bishop, Fienberg and Holland (1975), Haberman (1978), Fienberg (1980), and Whittaker (1990) structures the analysis based on the interaction terms between the variables.

Goodman (1981) proposed homogeneity and structure criteria in association models, allowing us to determine if certain rows or columns in a contingency table should be grouped. In later works, he showed the relationship between canonical scores and that corresponding to association models. Gilula (1986) developed grouping results suggested by the canonical scores in a contingency table under a canonical correlation model of Fisher (or saturated model RC). On the other hand, correspondence analysis can be seen like a re-parametrization of the canonical correlation model model, based on the results shown by Goodman (1986) and van der Heijden et al. (1994).

In statistical literature, the simplification is often called collapsing or coarsening (see Lauritzen 1996 and the references contained therein), while global recoding in the confidentiality literature (see Willenborg and de Waal 2000 who use the concept of minimizing information loss to consider alternatives).

The paired category collapsing process constructs a partition or coarsening of the categories for each variable. Such coarsening is typically not coarsening at random (see Heitjan and Rubin 1991 and Jaeger 2005) and thus there is a loss of information with respect to the original category sets. The papers by Lancaster (1949 and 1951), Goodman (1968 and 1970), Kreiner (2003) and many others, have explored alternative methods of partitioning the data in tables mainly for significance testing. Gokhale and Kullback (1978) provide illustrations, applications and extensions of theses ideas. Important results of this literature are summarized in Gilula and Haberman (1998) and recently, Jackson, Gray and Fienberg (2008) propose an approach by finding members of a class of restricted log-linear models which maximize the likelihood of the data and use it to find a parsimonious means of representing the table. In contrast with more standard approaches for model search in HLLM, this procedure systematically reduces the number of categories of the variables.

2.4.1 Hierarchical Clustering based on Correspondence Analysis (HCCA)

Our purpose is to summarize the original information, i.e. the complete contingency table $\mathbf{N} = [N_{ij}]$, to extract the most relevant patterns of specialization in the data. We must not lose sight of the problem scope. In the case of Argentina there are $I = 523$ regions. Using just the first 2 digits of the International Standard Industrial Classification of manufacturing activities (ISIC-Rev.3) there are $J = 23$ activities. The total number of employees is $N = 1.083.928$. Thus the contingency table is a 523×23 matrix of 1.083.928 employees spread in more than 12000 cells. Collapsing tables mean building tables of smaller dimension through aggregated regions (rows) and/or activities (columns). We are looking for the smallest collapsed table that preserves the observed overall level of specialization as much as possible. The total number M of possible collapsed tables for the $I \times J$ matrix $\mathbf{N}$ is

$$M = \sum\nolimits_{(m_1 \ldots m_i \ldots m_l)} \binom{I}{m_1 \ldots m_i \ldots m_l} \times \sum\nolimits_{(n_1 \ldots n_i \ldots n_k)} \binom{J}{n_1 \ldots n_j \ldots n_k} \qquad (2.42)$$

where $l \leqslant I-1$, $k \leqslant J-1$, $m_1 + \ldots + m_i + \ldots + m_l < I$ and $n_1 + \ldots + n_j + \ldots + n_k < J$.

For I and J large, as in the present case, M is huge and trying all possibilities is not feasible. This problem was already treated by Gilula (1986). He used canonical analysis to obtain scores for both rows and columns of the contingency table and then grouped the rows and columns with similar scores. Certain issues remained open as he does not provide a criteria for deciding when two or more scores are similar enough to group them.

Our approach consists of applying the Correspondence Analysis technique (Marinelli and Winzer 2003 show that it yields equal results as canonical analysis) and provides an alternative to avoid the calculation of all M possible collapsed tables. It does so by giving for every possible number of row groups and column groups the "best" grouping of rows and columns. Thus the total number of collapsed tables needed for calculation is at most $I \times J$. As a first step we conduct a Correspondence analysis of the original contingency table $\mathbf{N}$. Let $\mathbf{P} = \frac{\mathbf{N}}{N_{..}}$ be the probability matrix of $\mathbf{N}$. Let $\mathbf{r} = \{p_{i\cdot}\}$ the vector of row marginals and $\mathbf{c} = \{p_{\cdot j}\}$ the vector formed with the column marginals. Finally let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices formed with the row marginals and column marginals, respectively.

Through Singular Value Decomposition (SVD) of the matrix $(\mathbf{P} - \mathbf{rc}')$, obtain $(\mathbf{P} - \mathbf{rc}') = A\mathbf{D}_\lambda B'$, where $A'\mathbf{D}_r^{-1}A = I = B'\mathbf{D}_c^{-1}B$.

The standardized residuals matrix $\mathbf{S} = \{s_{ij}\}$ can be constructed from

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}A\mathbf{D}_\lambda B'\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}^T, \tag{2.43}$$

where $\mathbf{U} = \mathbf{D}_r^{-1/2}A$, $\mathbf{V} = \mathbf{D}_c^{-1/2}B$, and $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

In this decomposition, the dimension of $\mathbf{U}$ is $J \times K$, of $\mathbf{V}$ is $I \times K$ and of $\mathbf{D}_\lambda$ is $K \times K$, with $K \leq min(I-1, J-1)$. The diagonal elements $\lambda_1, \lambda_2, ..., \lambda_k$ of $\mathbf{D}_\lambda$ are the singular values of $(\mathbf{P} - \mathbf{rc}')$.

The principal coordinates for the rows are

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\lambda \tag{2.44}$$

and the principal coordinates for the columns are

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\lambda \tag{2.45}$$

The coordinates for $\mathbf{r}$ row deviation are given by de elements of f_{ik}, $i = 1, 2, ..., r$, $k = 1, 2, ..., k$, of $\mathbf{F}$. Similarly the coordinates for de $\mathbf{c}$ column deviations are given by the elements g_{kj}, $k = 1, 2, ..., k$, $j = 1, 2, ..., c$, of $\mathbf{G}$. The elements f_{ik} and g_{kj} are the kth scores for the row and column of the cell (i, j).

Each row of $\mathbf{F}$ provides the coordinates for a row deviation with respect to the K principal axes given by the columns of $\mathbf{U}$. Each column of $\mathbf{F}$ provides the coordinates for the r deviations with respect to a particular principal axis or column of $\mathbf{V}$. For more details, see Jobson 1992 and Mardia et al. (1979).

The divergence $d_{\mathcal{X}^2}(p \mid q)$ (Section 2.3) or total inertia $I(p \mid q)$ of the matrix of standardized residual is

$$d_{\mathcal{X}^2}(p \mid q) = I(p \mid q) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - q_{ij})^2}{q_{ij}} = \sum_{k=1}^{K} \lambda_k^2 \tag{2.46}$$

where$\lambda_1, \lambda_2 \ldots \lambda_{K-1}$ are the eigenvalues from $\mathbf{D}_\lambda$.

The total inertia can be expressed as the sum of the row inertias as follows (partition)

$$I(p \mid q) = \sum_{k=1}^{K} \sum_{i=1}^{I} \mathbf{P}_{i \cdot} f_{ik}^2 = \sum_{i=1}^{I} \left(\sum_{k=1}^{K} \mathbf{P}_{i \cdot} f_{ik}^2 \right) = \sum_{i=1}^{I} I_i \tag{2.47}$$

The row Inertia I_i measures the partial contribution of the region i to the global measure of specialization. Symmetrically, the total inertia can be expressed as the sum of the column inertias as follows

$$I(p \mid q) = \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbf{P}_{\cdot j} g_{kj}^2 = \sum_{j=1}^{J} \left(\sum_{k=1}^{K} \mathbf{P}_{j \cdot} g_{kj}^2 \right) = \sum_{j=1}^{J} I_j \tag{2.48}$$

We keep the first k coordinates of the row and columns scores. The choice of k, as will be shown later, will be an outcome of the method. In most cases k will end up being such that the first k eigenvalues will concentrate a relevant proportion of their global accumulated value. With the fixed scores we achieve an agglomerative hierarchical clustering of both row scores and column scores. For a given number of row groups we choose the grouping structure yielded by the single linkage agglomerative row clustering, that is equivalent to cut the dendrogram to such level that the number of groups be the one desired. The grouping thus created has the following property:

- The grouped rows (regions), produced using the first k scores, are the most homogeneous in terms of column (activity) profiles.
- The grouped columns (activities), produced using the first k scores, are the most homogeneous in terms of row (region) profiles.

Having done this for a fixed k, we repeat the procedure for every possible $k \in \{1 \dots K\}$, being $K = min(I-1, J-1)$. Thus we have a three dimensional array of collapsed tables $A = \{T_{ij}^k\}$ where each element T_{ij}^k is a collapsed table, produced using the first k scores, with i rows and j columns from the original table.

2.4.2 The "best" collapsed table

Having built the array A of collapsed tables, our final question is: which of the $I \times J \times K$ collapsed tables is better? We look for the smallest table that preserves the highest association possible. Two extreme and not useful cases are:

- The table T_{IJ}^k (the original one), has the maximum association with the minimum of information reduction.
- The table $T_{11}^k = N$ (the total number of cases), has the minimum association with the maximum of information reduction.

We need a quantity that balance the trade-off between the association degree and table dimension. We begin by defining a "goodness of association" measure for a given collapsed table. The idea is to measure the effectiveness of the HCCA to preserve association while reducing the table dimension. The proposed quantity measures the gain in association produced by the HCCA method compared to the association that would be expected under a random grouping strategy

$$g(\mathbf{T}_{ij}^k) = \mathcal{X}^2(\mathbf{T}_{ij}^k) - E^*(\mathcal{X}^2(\mathbf{T}_{ij}^k)) \tag{2.49}$$

where $E^*(\mathcal{X}^2(\mathbf{T}_{i,j}^k))$ is the expected association measured with the chi-squared statistic.

This expectation arises from the finite but large population of tables with i rows and j columns that can be obtained collapsing randomly the original table using Monte Carlo with a B number of replications. This expectation is conditional to the original table and to the fixed dimensions i and j. We estimate this value with the sample mean

$$E^*(\mathcal{X}^2(\mathbf{T}_{ij}^k)) \approx \sum_{\mathbf{T}_{ij}^*} \mathcal{X}^2(\mathbf{T}_{ij}^*) \tag{2.50}$$

where $\mathbf{T}_{ij}^*$ is obtained collapsing rows and columns randomly without labels following the agglomerative hierarchical clustering schemes of the HCCA procedure, i.e

- the random grouping strategy is based on these schemes as a griding algorithm;
- hence, the number of $\mathbf{T}_{ij}^*$ is at most equal to $(I \times J \times K) \times B$.

Then we propose that the best collapsed table is that with the maximum "goodness of association" measure, i.e

$$\mathbf{T}^* = \max_{ijk} g(\mathbf{T}_{ij}^k) \tag{2.51}$$

Next we show that, as long as there is a minimum association in the data, this maximum is obtained in none of the two extreme cases mentioned before.

Theorem. *Given a contingency table* $\boldsymbol{T}$ *such that* $\mathcal{X}^2(\boldsymbol{T}) > 0$*, then* $g(T_{11}) = 0$*,* $g(T_{IJ}) = 0$ *and* $\exists i \in \{1 \dots I\}$*,* $\exists j \in \{1 \dots J\}$ *such that* $g(T_{ij}) > 0$.

2.5 Example

We apply the methodology to a classical example from Srole et al. (1962), which was analyzed by Haberman (1974, 1979), Goodman (1985) and Gilula (1986), among others. The data of Table 2.1 consists of 1660 subjects in midtown Manhattan cross-classified by mental health status and parental socioeconomic status (A being the highest; F the lowest).

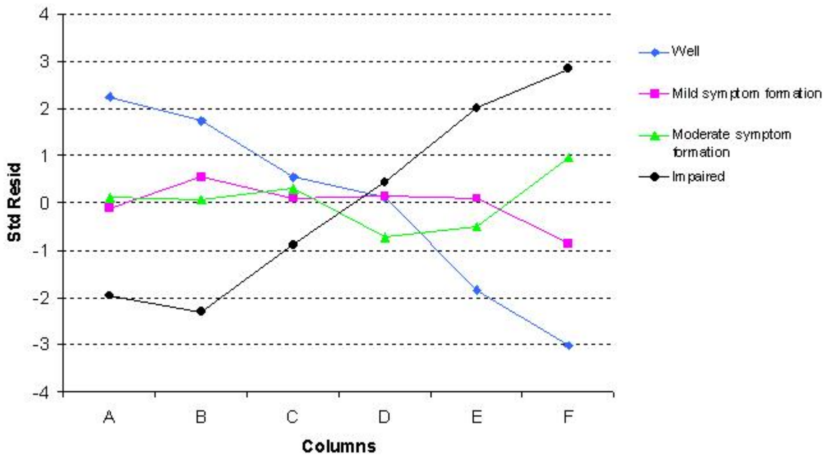
Following Gilula (1986), when the independency model is considered for this table, we found that $\mathcal{X}^2 = 45.99$, $G^2 = 47.42$, $d_{\mathcal{X}^2} = 0.02770$, $d_{KL} = 0.01428$, and $d_H^2 = 0.00366$ on 15 df. Fig. 2.2 shows the standardized residuals of Table 2.1.

Gilula (1986) obtained the maximum likelihood estimates of the canonical correlation for the above Table 2.1 and its corresponding canonical scores, and suggests

Table 2.1: *Subjects cross-classified by mental health status and parental socioeconomic status*

Mental health category	A	B	C	D	E	F
Well	64	57	57	72	36	21
Mild symptom formation	94	94	105	141	97	71
Moderate symptom formation	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

Figure 2.2: *Standardized residuals of Table 2.1*

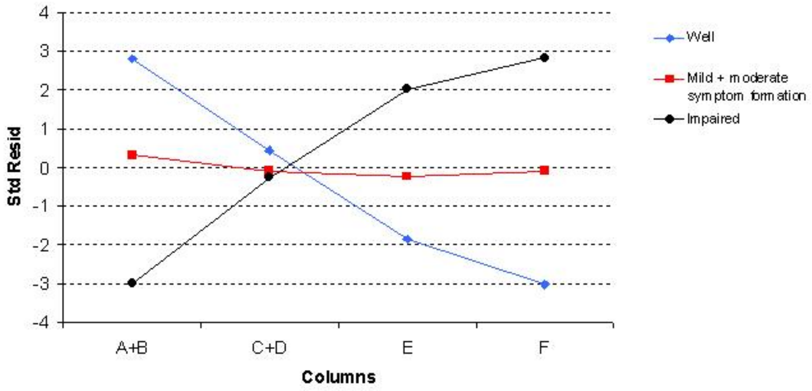


that rows 2 and 3 are homogeneous (Well and Mild symptom formation), as are columns A and B and columns C and D. Combining these rows and columns, the following is obtained: a 3 × 4 table with $\mathcal{X}^2 = 42.04$, $G^2 = 43.44$, $d_{\mathcal{X}^2} = 0.02532$, $d_{KL} = 0.01308$, and $d^2_H = 0.00336$ on 6 df. Comparing these results with the values of the original Table 2.1, there is reasonable evidence that the suggested grouping is justified. Fig 2.3 shows the standardized residuals of the collapsed table obtained by Gilula.

From Fig. 2.3 also arises that it is possible to group columns E and F.

Then we developed an R function that automatically computes the collapsed "best" table. For this example, the B number of Monte Carlo replications of $\mathbf{T}^*_{ij}$ is

Figure 2.3: *Standardized residuals of collapsed table obtained by Gilula (1986)*



equal to 100. Based on a Correspondence Analysis of Table 2.1., Tables 2.2, 2.3 and 2.4 show the principal inertias (eigenvalues), and row and column scores for the two first eigenvalues, respectively.

Table 2.2: *Principal inertias (eigenvalues) based on Correspondence Analysis of Table 2.1*

dim	value	%	cum%	scree plot
1	0.026025	93.9	93.9	*************************
2	0.001379	5.0	98.9	*
3	0.000298	1.1	100.0	
4				
Total	0.027702	100.0		

The function also returns a scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition (Fig. 2.4).

The scatter plot of the standardized residuals and the LQ_{ij} values in Fig. 2.4 serve to illustrate the following relationship: positive standardized residuals correspond to LQ_{ij} values greater than 1, i.e. "over-specialization", while negative

Table 2.3: *Rows scores for the two first eigenvalues*

Mental health category	k=1	k=2
Well	-260	12
Mild symptom formation	-30	24
Moderate symptom formation	14	-70
Impaired	237	19

Table 2.4: *Columns scores for the two first eigenvalues*

Parental socio-economic status	k=1	k=2
A	-181	-19
B	-185	-12
C	-59	-22
D	9	42
E	165	44
F	288	-62

standardized residuals are synonymous to "sub-specialization" ($LQ_{ij} < 1$). In addition to the perfectly linear correlation between residuals and LQ_{ij} values given the characteristics of this classic example, the graph shows that the best collapsed table is obtained by grouping rows and columns so as to make the most extreme possible standardized residuals or LQ_{ij} values.

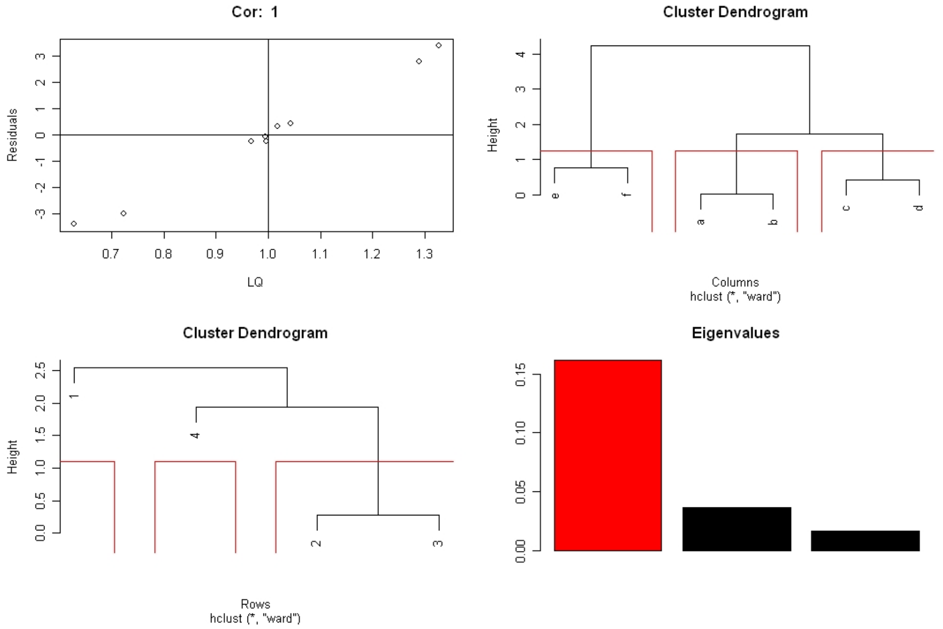
The red lines in the row and column dendrogram in Fig 2.4 show the best cutting branches selected automatically by our method, while the bar graph shows the values of the eigenvalues and the red bar indicates the eigenvalue selected.

The 3 dimensional plot (Fig 2.5) shows $g(\mathbf{T}_{ij}^k)$ for the first eigenvalues and for each level of the grouping structure of the row and column dendrogram. It is important to note that there are 3 row levels and 5 column levels (Fig. 2.4), so the total number of M collapsed tables is 15. The red point indicates the maximum goodness of association $\mathbf{T}^*$.

Our method obtains automatically a 3×3 collapsed table (red point in Fig. 2.5) with $\mathcal{X}^2 = 40.49$, $G^2 = 41.45$, $d_{\mathcal{X}^2} = 0.02439$, $d_{KL} = 0.01249$, and $d_H^2 = 0.00318$ on 4 *df*, with the advantage (as opposed to the subjective selection of the homogeneous row and column scores) that the best cutting branches of the dendrogram and the selection of the eigenvalues (the first in this case) is automatic.

Table 2.5 shows rows 2 and 3 (Well and Mild symptom formation), columns A and B, columns C and D, and columns E and F grouped respectively, with the standardized residuals and column cell proportions between parenthesis. As in Gilula's, we see that *well-being* has a decreasing prevalence inasmuch as the status gets lower as opposed to *impaired-being*, which is generally more frequent in lower statuses

Figure 2.4: *Scatter plot of standardized residuals and LQ_{ij} values, dendrogram for rows and columns based on the HCCA method, and the eigenvalues of the SVD decomposition*



than in upper statuses. "Mild" and "moderate" symptoms seem to have a similar prevalence across statuses.

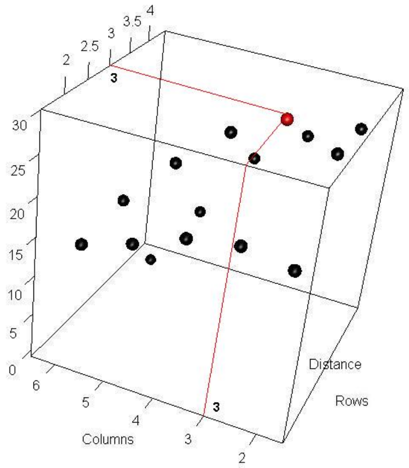
Fig. 2.6 shows the standardized residuals of the table obtained with our method and serves to illustrate that the best collapsed table is obtained by grouping rows and columns so as to make the most extreme possible standardized residuals values.

2.6 Application: Argentina, Brazil and Chile

The purpose of this section is to compare the overall degree of specialization of Argentina, Brazil and Chile using the measures described above, and obtain the best collapsed table for each of them to identify the regions with similar industrial manufacturing structures. As mentioned at the beginning of this Chapter, this methodology does not consider the distance among regions.

The spatial units are the lower level political-administrative jurisdictions called

Figure 2.5: *3 dimensional plot of* $g(\boldsymbol{T}_{ij}^{k})$ *for the first eigenvalues*



departments (523), municipalities (5,138) and communes (342) of Argentina, Brazil and Chile respectively. The final spatial units (after eliminating those without employees) are 462, 5,138 and 249 for Argentina, Brazil and Chile, respectively.

The data related to the employees in the manufacturing sector were obtained from of the Nationals Institute of Statistics and Censuses of Argentina (INDEC-1994: 1,083,928 employees), Brazil (IBGE-1998: 6,018,445 employees), and Chile[3](INE-2005: 446,613 employees) respectively. The activity classifications in Table 2.6 refers to the first 2 digits of the International Standard Industrial Classification (ISIC-Rev.3) of manufacturing activities (22 activities).

Table 2.7 shows a summary of the results obtained from the proposed measures of global specialization for the original and collapsed tables, the number of cells, and the resulting loss of information about the level of specialization.

While the absolute values of these measures are not comparable because they use different scales, the global measures of specialization in the original tables show that Chile has a higher level of specialization, followed by Brazil and Argentina, respectively (with the exception of unweighted GI and SK indexes whose values indicate that these differences are not an effect that depends on the number of cells: Chile $<$ Argentina $<$ Brazil).

[3] The data refer to the firms with 5 or more employees. This means that Chile is not directly comparable to other countries and the results are for illustration purposes only.

Table 2.5: *Data in Table 2.1 grouped with our method of best collapsed table, with standardized residuals and columns cell proportions between parenthesis*

Mental health category	A+B	C+D	E+F
Well	121	129	57
std residuals	2.813	0.440	-3.404
column cell proportions	(0.239)	(0.192)	(0.118)
Mild + moderate symptom	300	388	276
std residuals	0.325	-0.084	-0.234
column cell proportions	(0.592)	(0.578)	(0.573)
Impaired	86	154	149
std residuals	-3.010	-0.258	3.392
column cell proportions	(0.170)	(0.230)	(0.309)

The global measures of specialization on the best collapsed tables also show the same order in the level of specialization of these countries, although in this case the different measures show a decline in its absolute values as a result of the loss of information arising from grouping regions and activities.

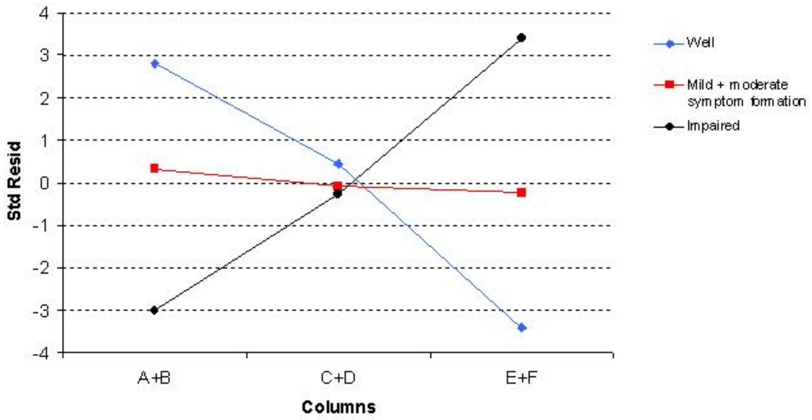
The loss in the level of specialization is quite similar for Argentina and Brazil, while Chile clearly shows a minor loss of information. The unweighted indexes show a minor loss of information with respect to other measures. The loss of information seems very reasonable vis-à-vis a substantial reduction in the size of the tables (more than 90%).

2.6.1 Best collapsed table for Argentina

Fig. 2.7 shows a scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition respectively. The red lines in the row and column dendrogram show the best cutting branches selected automatically, while the bar graph shows the values of the eigenvalues and the red bars indicates the eigenvalues selected (the first 12 out of 21).

The 3 dimensional plot (Fig. 2.8) shows $g(\mathbf{T}_{ij}^{k})$ for fixed eigenvalues and for each level of the grouping structure of the row and column dendrogram, and the maximum goodness of association $\mathbf{T}^{*}$(red point). The B number of Monte Carlo

Figure 2.6: *Standardized residuals of the best collapsed table*



replications of $\mathbf{T}^*_{ij}$ is equal to 1,000. The best collapsed table for Argentina has 35 rows (grouped regions) and 17 columns (grouped activities) reducing by 94% the number of cells of the original table (from 10,164 to 595 cells). Table 2.7 shows that the loss of information of specialization using $d_{\mathcal{X}^2}$ is 23%.

Table 2.8 and 2.9 show the number of employees and the standardized residuals by activity for the selected grouped regions (GRegion) of Argentina. These GRegions show the highest levels of specialization at a national level for the activities described below. 467 regions are grouped in 35 GRegions, while the 22 activities are brought into 17 groups. Grouped sectors (GS) are 2: 25, 28, 29, 31 and 36 (GS^1); and 30 and 33 (GS^2).

GRegions 2 and 15, formed by 7 and 10 regions respectively, are "over-specialized" in activity 18. The GRegion 2 has 6,779 employees and the GRegion 15, 3,255 (red values on Table 2.8). In the aggregate, these GRegions add 22 percent of the employees in the whole sector at a national level. The difference between these two GRegions is that the GRegion 2 is over-specialized (lower level) in activity 17, 22, 24 and grouped activities GS^2, while GRegion 15 is over-specialized in activity 19. Consequently, these GRegions are "sub-specialized", at different levels- in the rest of the activities.

The GRegion 16, formed by 4 regions, is over-specialized in activity 27. This GRegion has 13,022 employees which accounts for 35 percent of the employees in the whole sector at a national level. GRegion 16 is also over-specialized in lower

Table 2.6: *Division ISIC-Rev.3 for the manufacturing industry*

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

levels in sectors 23 and 32. Finally, GRegion 35 is formed by 2 regions and add 3,379 employees (32 percent of employees in the whole sector at a national level). This GRegion is sub-specialized for the rest of the activities.

Figures 2.9, 2.10 and 2.11 show the location of the selected GRegions for Argentina.

2.6.2 Best collapsed table for Brazil

Like in the case of Argentina, Fig. 2.12 shows a scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition respectively. The red lines in the row and column dendrogram show the best cutting branches selected automatically, while the bar graph shows the values of the eigenvalues and the red bars indicates the eigenvalues selected (the first 15 out of 21).

The 3 dimensional plot (Fig. 2.13) shows $g(\mathbf{T}_{ij}^k)$ for fixed eigenvalues and for each level of the grouping structure of the row and column dendrogram, and the

Table 2.7: *Summary of the results*

Measure	Original table			Collapsed table			Lost level of specialization (%)		
	Argentina	Brazil	Chile	Argentina	Brazil	Chile	Argentina	Brazil	Chile
$d_{\mathcal{X}^2}$	2.1580	3.1345	3.4363	1.6532	2.4162	2.8584	23.39	22.91	16.82
d_{KL}	0.5049	0.7420	0.8870	0.3176	0.4928	0.6759	37.10	33.58	23.80
d^2_H	0.1300	0.1894	0.2600	0.0713	0.1073	0.1773	45.17	43.39	31.80
GI(average)	0.6352	0.7448	0.6991	0.5704	0.6012	0.6562	10.21	19.28	6.14
GI(average)[1]	0.4621	0.5595	0.6017	0.3388	0.4135	0.5065	26.69	26.09	15.83
SK(average)	0.5338	0.6545	0.6374	0.4760	0.5165	0.5695	10.83	21.08	10.65
SK(average)[1]	0.3625	0.4521	0.5079	0.2963	0.3555	0.4500	18.26	21.38	11.40
# of cells	10,164 (462x22)	113,036 (5,138x22)	5,478 (249x22)	595 (35x17)	884 (52x17)	450 (30x15)			
Reduction # of cells				94.15%	99.22%	91.79%			

[1] Weighted for $N_{i.}/N_{..}$

Table 2.8: *Number of employees in the selected GRegions for Argentina*

GRegion	Division ISIC-Rev.3																
	15	16	17	18	19	20	21	22	23	24	GS[1]	26	27	GS[2]	32	34	35
2	3,819	6	1,496	6,779	424	258	479	1,750	4	2,147	3,300	384	68	281	114	587	21
15	851	0	193	3,255	289	149	1	151	0	4	827	61	21	8	26	81	0
16	1,409	0	732	140	59	120	462	200	784	639	2,125	454	13,022	5	281	1,328	60
35	485	0	224	24	1	164	0	76	0	85	827	61	0	0	3,379	215	0

[1] Divisions grouped: 25, 28, 29, 31 and 36.

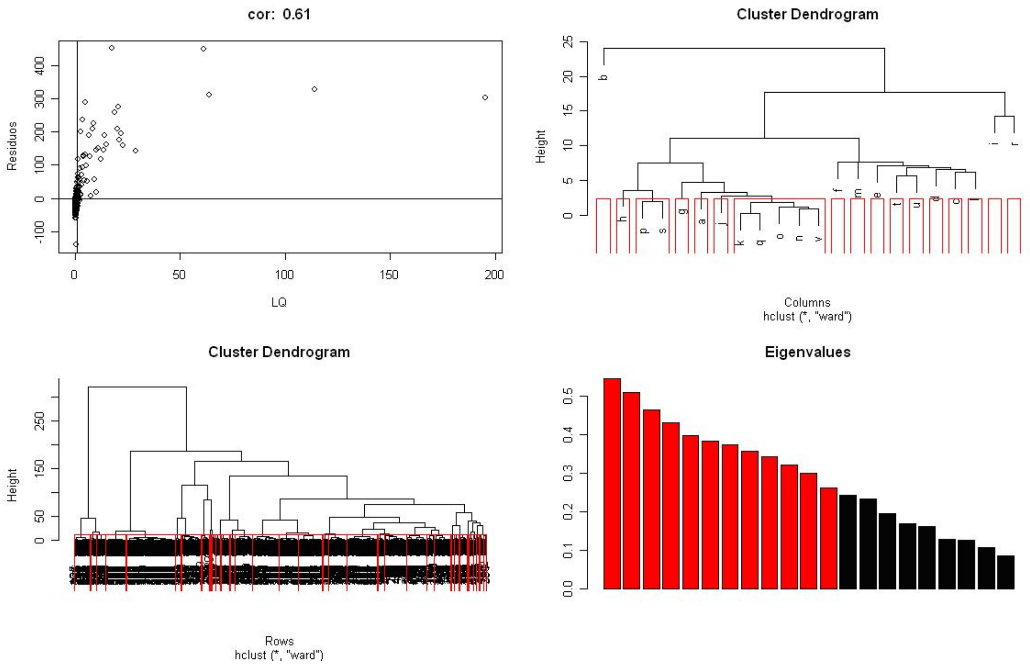
[2] Divisions grouped: 30 and 33.

maximum goodness of association $\mathbf{T}^*$ (red point). The B number of Monte Carlo replications of $\mathbf{T}^*_{ij}$ is equal to 1,000. The best collapsed table for Brazil has 52 rows (grouped regions) and 17 columns (grouped activities) reducing by 99% the number of cells of the original table (from 113,036 to 884 cells). The above Table 2.7 shows that the loss information of specialization using $d_{\mathcal{X}^2}$ is 23%.

Fig. 2.14 shows the standardized residuals for each cell of the original table for Brazil. It must be noted that most of the standardized residuals are around zero. Therefore, it is expected that most of the cells are not “over-specialized” or “sub-specialized”. The original table can then be summarized as showing almost no difference between the observed and expected values, and the degree of specialization can be explained with much less information.

As in the previous example (Section 2.5), Fig. 2.15 with the standardized residuals for each cell of the best collapsed table for Brazil shows that it is obtained by grouping rows and columns so as to make the most extreme possible standard-

Figure 2.7: *Argentina: scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition*



ized residual values, i.e. to make the "over-specialized" and the "sub-specialized" phenomenon more evident.

Tables 2.10 and 2.11, show the number of employees and the standardized residuals by activity for the selected grouped regions of Brazil. As for Argentina, these GRegions show the highest levels of specialization at a national level for the activities described below. The 5,138 regions are grouped in 52 GRegions, while the 22 activities are brought into 17 groups. There are 2 grouped sectors (GS): 22, 25, 28, 31 and 33 (GS^1); and 30 and 32 (GS^2).

GRegion 27, formed by 41 regions, is over-specialized in the grouped sectors GS^2. This GRegion has 22,070 employees (red values on Table 2.10) that account for 23 percent of employees in the whole grouped sector at a national level. GRegion 27 is also over-specialized with lower levels in the grouped sector GS^1 and in sector 35.

Figure 2.8: *Argentina: 3 dimensional plot of $g(\boldsymbol{T}_{ij}^{k})$ for fixed eigenvalues (the first 12 out of 21)*

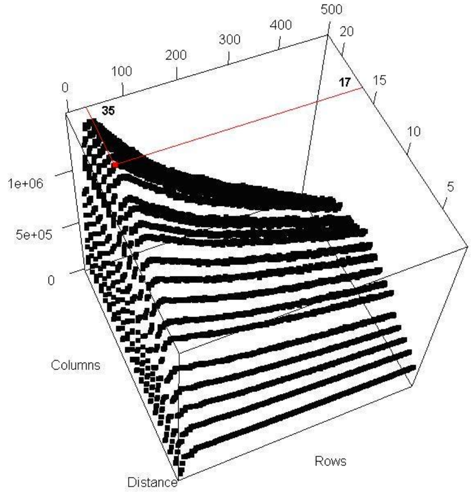


Table 2.9: *Standardized residuals in the selected GRegions for Argentina*

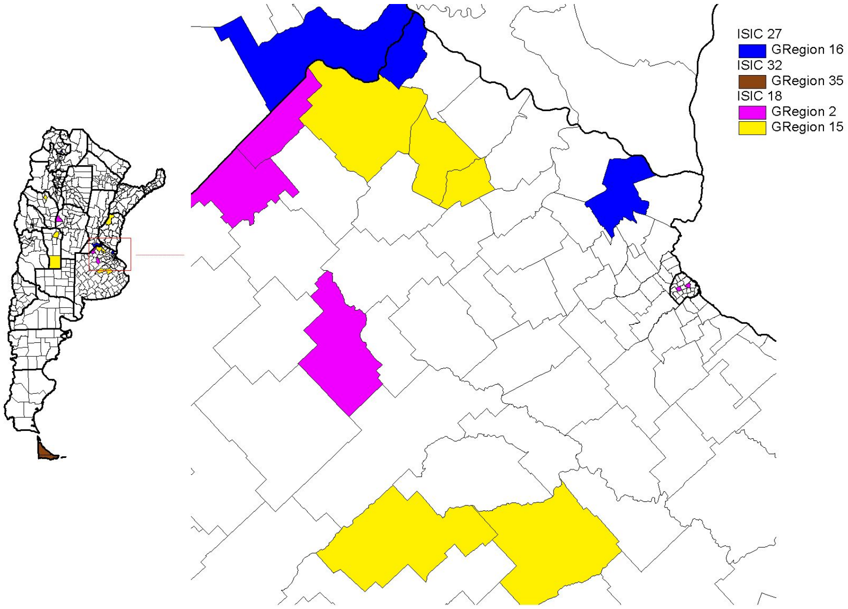
GRegion	Division ISIC-Rev.3																
	15	16	17	18	19	20	21	22	23	24	GS^1	26	27	GS^2	32	34	35
2	-27	-9	9	191	-15	-15	-2	24	-11	20	-28	-20	-25	5	-7	-23	-12
15	-18	-5	-7	189	4	-2	-12	-7	-6	-19	-16	-13	-13	-7	-4	-16	-7
16	-58	-10	-13	-26	-27	-20	-3	-25	55	-20	-44	-18	452	-14	4	-44	-9
35	-26	-5	-5	-14	-15	0	-12	-11	-6	-14	-14	-12	-14	-7	449	-8	-7

The GRegions 46 and 49, formed by 118 and 73 regions respectively, are over-specialized in activity 19. The GRegion 46 has 164,653 employees and GRegion 49 has 91,625 (red values on Table 2.10). Together, these GRegions add 70 percent of employees in the whole sector at a national level. The difference between these two GRegions is found on sub-specialization levels for the rest of the activities.

Finally, GRegion 51, formed by 48 regions, is over-specialized in activity 34. This GRegion has 122,629 employees which accounts for 43 percent of employees in the whole sector at a national level. GRegion 51 is also over-specialized with lower levels in sector 29.

Figures 2.16 and 2.17 show the location of the selected GRegions for Brazil.

Figure 2.9: *Map 1 for Argentina: location of the selected GRegions*



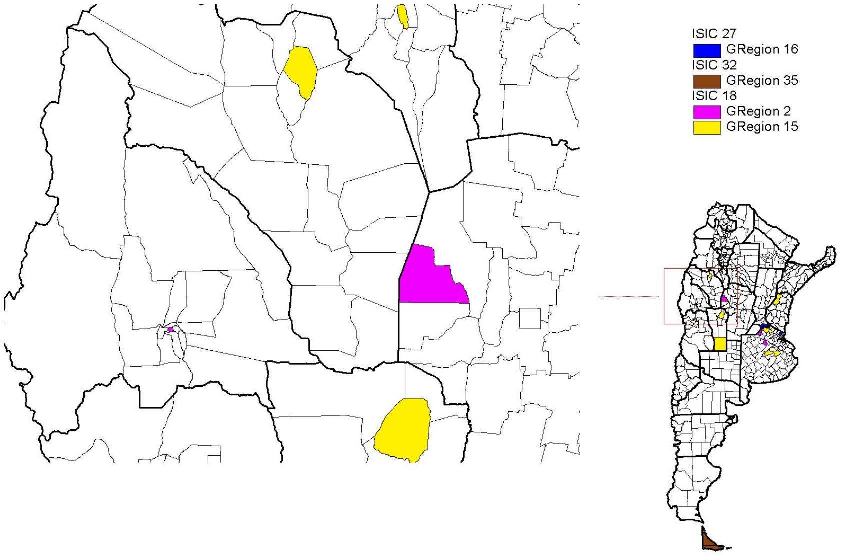
2.6.3 Best collapsed table for Chile

As in the case of Argentina and Brazil, Fig. 2.18 shows a scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition respectively. The red lines in the row and column dendrogram show the best cutting branches selected automatically, while the bar graph shows the values of the eigenvalues, and the red bars indicates the eigenvalues selected (the first 12 out of 21).

The 3 dimensional plot (Fig. 2.19) shows $g(\mathbf{T}_{ij}^{k})$ for fixed eigenvalues and for each level of the grouping structure of the row and column dendrogram, and the maximum goodness of association $\mathbf{T}^{*}$(red point). The B number of Monte Carlo replications of $\mathbf{T}_{ij}^{*}$ is equal to 1,000. The best collapsed table for Chile has 30 rows (grouped regions) and 15 columns (grouped activities) reducing by 92% the number of cells of the original table (from 5,478 to 450 cells). Above Table 2.7 shows that the loss of information of specialization using $d_{\mathcal{X}^2}$ is 17%.

Tables 2.12 and 2.13 show the number of employees and the standardized residuals by activity for the selected grouped regions (GRegion) of Chile. As in the case of Argentina and Brazil, these GRegions show the highest levels of specialization at

Figure 2.10: *Map 2 for Argentina: location of the selected GRegions*



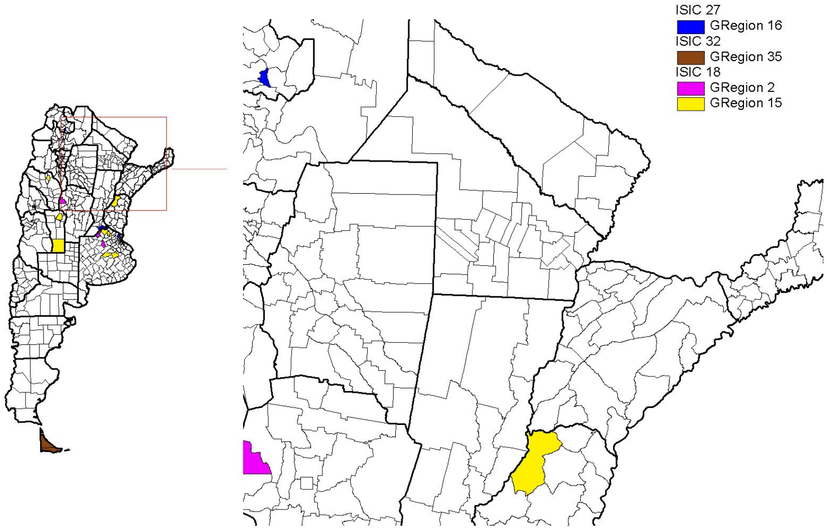
a national level in the activities described below. The 249 regions are grouped in 30 GRegions, while the 22 activities are brought into 15 groups. Grouped sectors (GS) are 3: 18 and 33 (GS^1); 22, 25, 28, 29, 31 and 34 (GS^2); and 26 and 30 (GS^3).

The GRegions 8 and 15, formed by 7 and 2 regions respectively, are over-specialized in activity 19. GRegion 8 has 3,864 employees and GRegion 15 has 2,080 employees (red values on Table 2.12). Together, these GRegions add 67 percent of employees in the whole sector at a national level. GRegion 8 is also over-specialized with lower levels in grouped sectors GS^1 and GS^2, and in sectors 17, 24, 32 and 36. Instead, GRegion 8 is over-specialized with lower levels in grouped sectors GS^2, and in sectors 17, 35 and 36.

GRegions 3 and 13, formed by 4 and 5 regions respectively, are over-specialized in activity 24. GRegion 3 has 2,574 employees and GRegion 13 has 6,754 (red values on Table 2.12). Together, these GRegions add 28 percent of employees in the whole sector at a national level. GRegion 13 is also over-specialized with lower levels in grouped sectors GS^3.

Finally, GRegions 4 and 6, formed by 9 and 10 regions respectively, are over-specialized in activity 27. GRegion 4 has 11,894 employees and the GRegion 6 has

Figure 2.11: *Map 3 for Argentina: location of the selected GRegions*



7,220 employees. Together, these GRegions add 60 percent of employees in the whole sector at a national level. GRegion 4 is also over-specialized with lower levels in grouped sectors GS^3 and in activity 24.

Figures 2.20 and 2.21 show the location of the selected GRegions for Chile.

Figure 2.12: *Brazil: scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition*

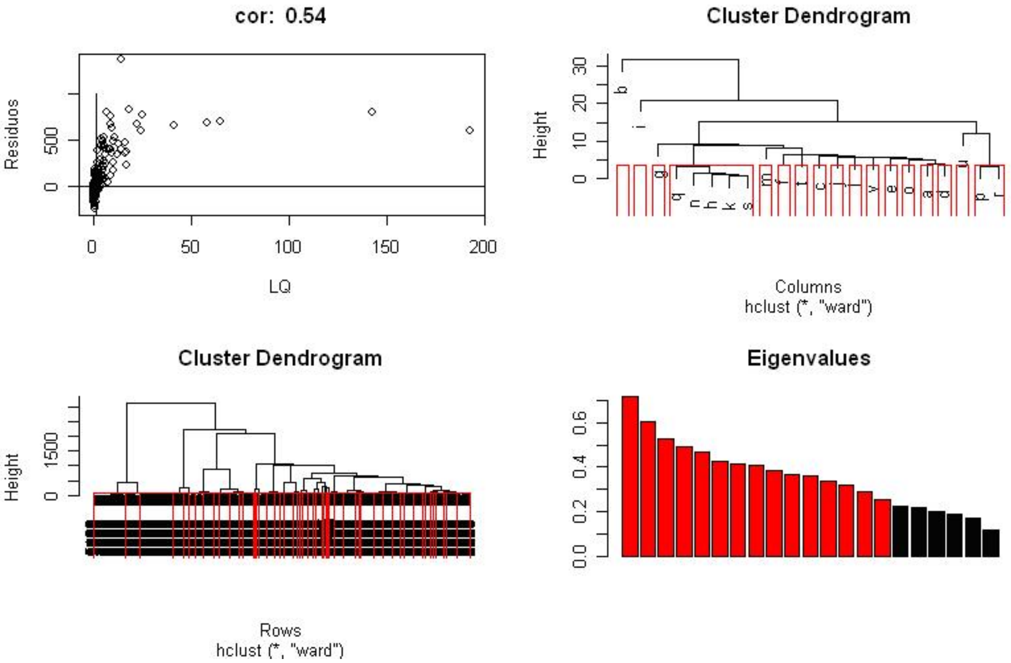


Table 2.10: *Number of employees in the selected GRegions for Brazil*

GRegion	Division ISIC-Rev.3																
	15	16	17	18	19	20	21	GS[1]	23	24	26	27	29	GS[2]	34	35	36
27	9,515	17	3,179	2,309	120	1,448	1,379	29,664	42	2,655	3,298	925	3,160	22,070	2,681	6,411	2,921
46	4,347	6	1,773	2,368	164,653	1,279	2,877	10,934	2	2,436	2,233	945	2,314	30	228	215	4,663
49	27,479	49	4,560	8,415	91,625	4,179	3,648	17,523	86	2,626	6,848	962	6,752	101	860	273	6,744
51	24,738	25	8,702	12,400	1,753	5,276	6,093	75,143	1,410	19,449	15,397	9,363	29,945	4,345	122,629	1,407	16,976

[1] Divisions grouped: 22, 25, 28, 31 and 33.

[2] Divisions grouped: 30 and 32.

Figure 2.13: *Brazil: 3 dimensional plot of $g(\boldsymbol{T}_{ij}^{k})$ for fixed eigenvalues (the first 15 out of 21)*

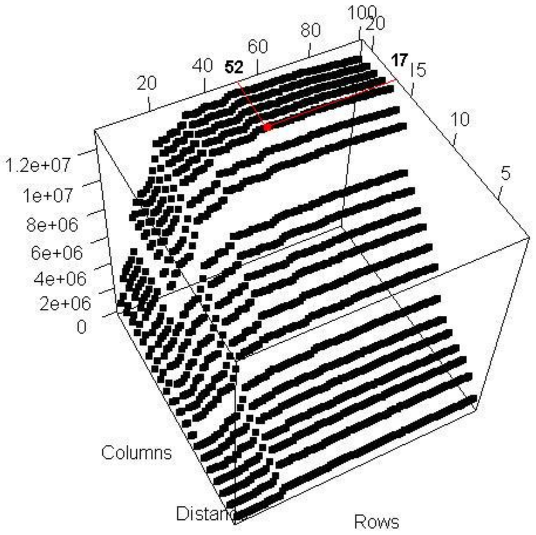


Figure 2.14: *Standardized residuals of original table of Brazil*

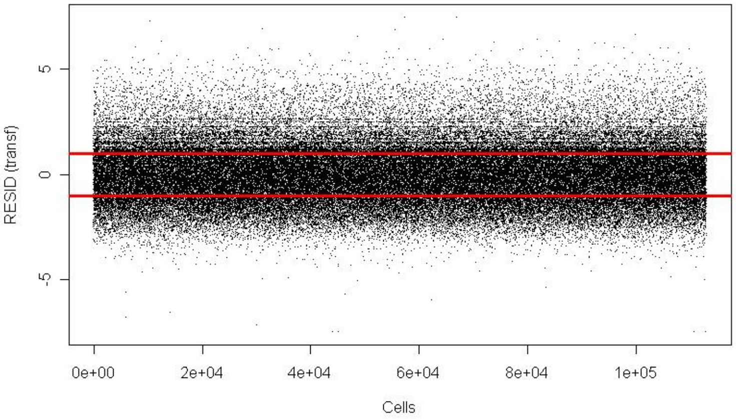


Figure 2.15: *Standardized residuals of the best collapsed table of Brazil*

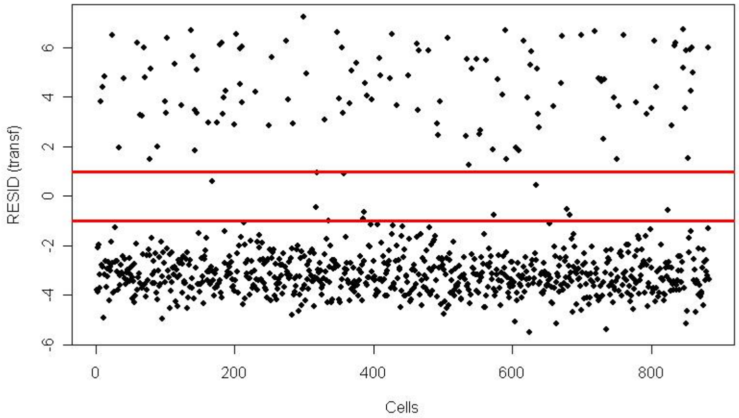


Table 2.11: *Standardized residuals in the selected GRegions for Brazil*

GRegion	Division ISIC-Rev.3																
	15	16	17	18	19	20	21	GS[1]	23	24	26	27	29	GS[2]	34	35	36
27	-58	-15	-23	-66	-73	-41	-20	91	-21	-35	-28	-33	-35	537	-26	197	-35
46	-171	-23	-85	-118	1,373	-81	-31	-141	-34	-83	-87	-64	-92	-56	-96	-37	-68
49	-36	-20	-51	-64	761	-44	-14	-94	-30	-75	-36	-59	-45	-52	-84	-33	-40
51	-161	-30	-72	-111	-135	-84	-31	27	-15	-3	-36	-8	50	-18	809	-31	-30

Table 2.12: *Number of employees in the selected GRegions for Chile*

GRegion	Division ISIC-Rev.3														
	15	16	17	GS[1]	19	20	21	GS[2]	23	24	GS[3]	27	32	35	36
3	0	0	0	0	0	0	0	0	0	2,574	0	0	0	0	0
4	1,666	0	0	36	0	103	5	3,582	0	1,953	726	11,894	0	16	22
6	96	0	0	0	0	0	0	6	0	119	9	7,220	0	0	0
8	5,667	0	2,212	3,483	3,864	899	994	9,323	0	2,937	634	1,618	80	5	1,146
13	8,050	0	552	223	236	43	522	4,417	0	6,754	1,122	647	0	13	374
15	429	0	161	0	2,080	38	100	1,700	0	90	76	0	0	192	213

[1] Divisions grouped: 18 and 33.

[2] Divisions grouped: 22, 25, 28, 29, 31 and 34.

[3] Divisions grouped: 26 and 30.

Figure 2.16: *Map 1 for Brazil: location of the selected GRegions*

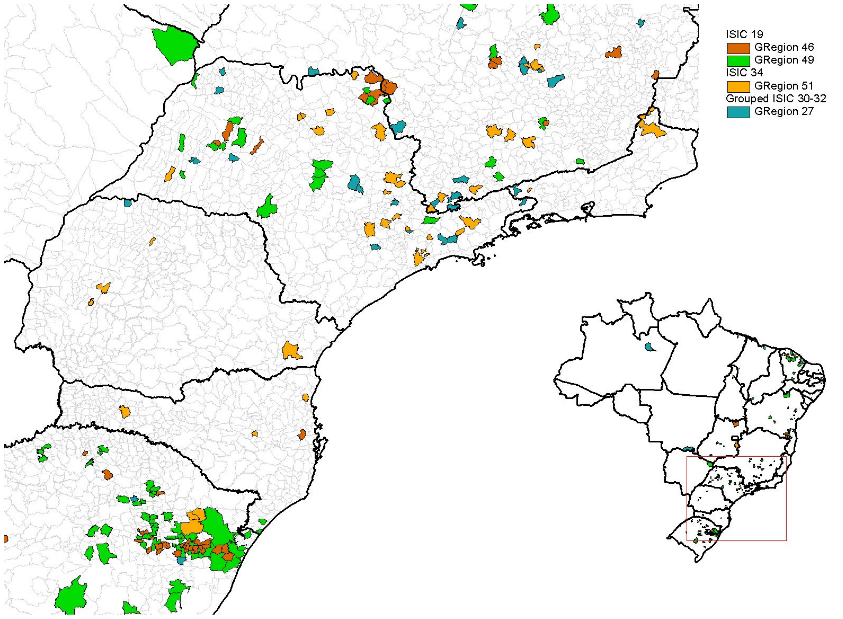


Figure 2.17: *Map 2 for Brazil: location of the selected GRegions*

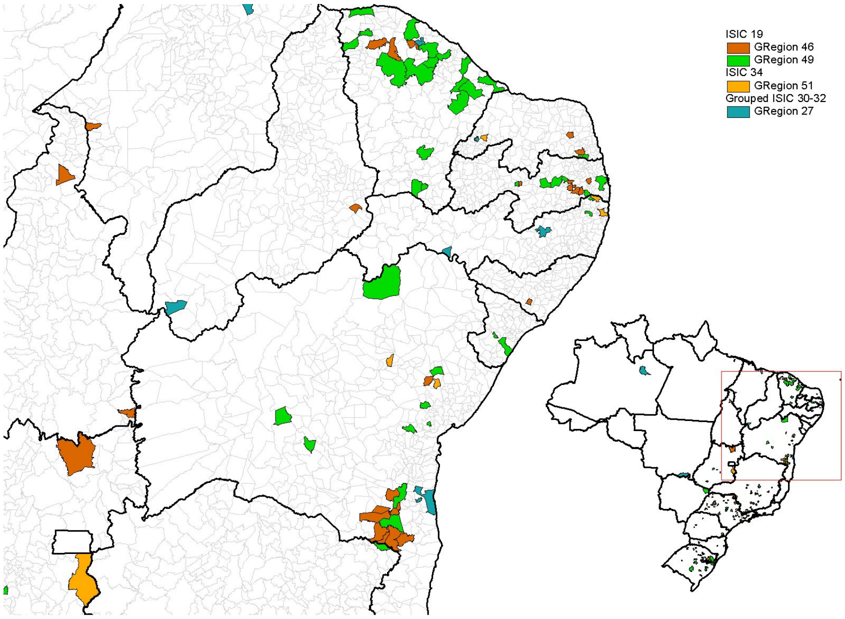


Figure 2.18: *Chile: scatter plot of standardized residuals and LQ_{ij} values, the row and column dendrogram based on the HCCA method, and the eigenvalues of the SVD decomposition*

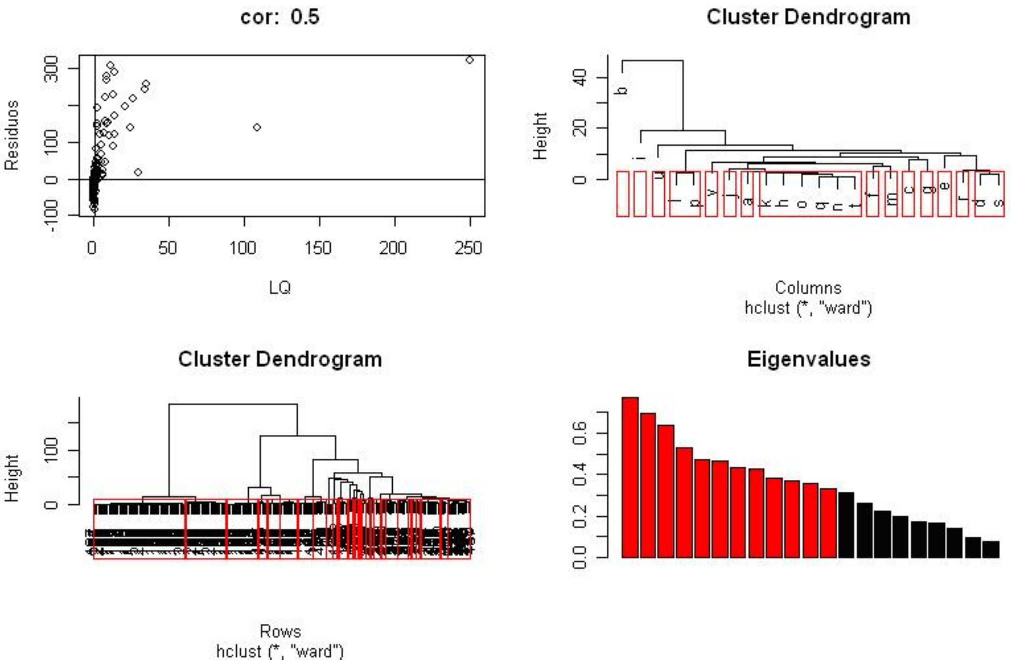


Table 2.13: *Standardized residuals in the selected GRegions for Chile*

GRegion	Division ISIC-Rev.3														
	15	16	17	GS^1	19	20	21	GS^2	23	24	GS^3	27	32	35	36
3	-30	-2	-9	-11	-7	-15	-10	-22	-4	173	-9	-14	-1	-6	-9
4	-64	-5	-25	-28	-20	-40	-27	-5	-12	12	3	278	-3	-16	-25
6	-49	-3	-15	-18	-12	-26	-16	-38	-7	-18	-15	291	-2	-10	-16
8	-55	-7	39	54	125	-37	-6	37	-15	10	-14	-15	19	-21	1
13	-1	-6	-5	-25	-10	-44	-11	-1	-13	123	13	-24	-3	-17	-14
15	-32	-3	1	-15	196	-19	-6	23	-6	-15	-7	-19	-1	15	3

Figure 2.19: *Chile: 3 dimensional plot of $g(\boldsymbol{T}_{ij}^{k})$ for fixed eigenvalues (the first 12 out of 21)*

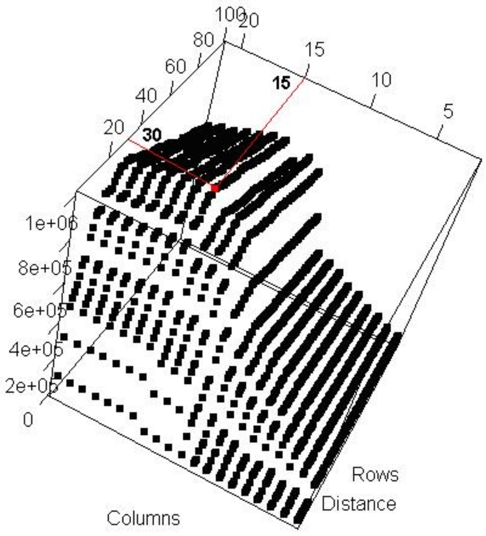


Figure 2.20: *Map 1 for Chile: location of the selected GRegions*

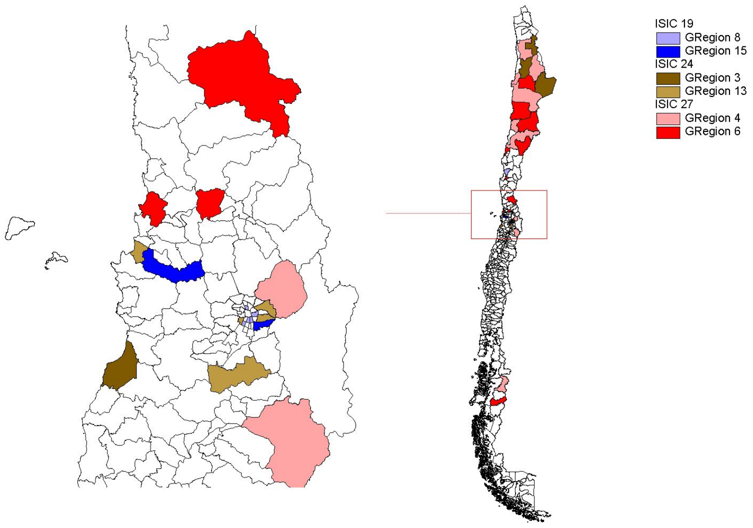
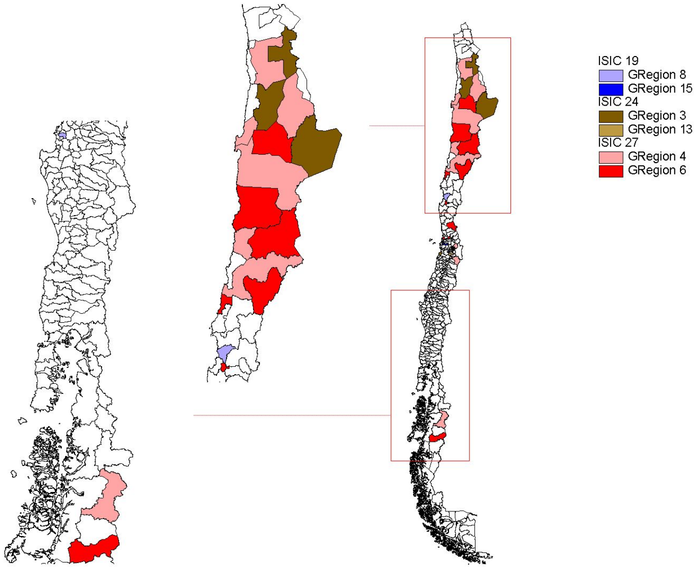


Figure 2.21: *Map 2 for Chile: location of the selected GRegions*



Chapter 3

Probability model for the identification of specialized agglomeration in discrete space

The positive spatial correlation is a key feature of New Economic Geography (NEG) models, and in particular of the so-called "market potential functions" that can be derived from them (Section 1.3). For instance, Fujita, Krugman and Venables (2001) have obtained several reduced-form equilibrium equations, in which a variable expressing the attractiveness of a location turns out to be a positive function of the level of economic activity in the surrounding regions. Fujita and Thisse (2002) show the substantial difference in the geographical scope of traditional externalities (pecuniary externalities) that are the engine of agglomeration in NEG models. Traditional spillovers or more localized externalities (Marshallian externalities, Urbanization, Porter or Dynamics externalities), require some degree of physical interaction among agents within the same place. Consequently, pecuniary externalities can be part of the forces that boost concentration, while positive spatial correlation are their distinctive feature with respect to either traditional externalities or factor endowments.

The externalities arising from the proximity among firms, i.e. externalities and location, are concerned with firm interaction in a certain region. The spatial externalities are significant for the plant location distribution, i.e. the outcome of firms' location choice (Section 1.3.3). Consequently, should we measure the strength of these spillover effects, the unit of analysis would be in favor of firms. Consider, for example, the following two polar cases. In the first one, there is only one large industry firm located in region j. In the second case, now there are many firms belonging to the industry located in region j. Clearly, the employment-based Gini coefficient would take the same high value in both cases, indicating a strong concentration pattern for the industry. However, the nature of this concentration is completely different in the two situations involved. In the first one, concentration occurs at the establishment level, resulting in a unique operating plant that hires all workers (industrial concentration). This could be reasonably associated to factors that are "internal" to a firm or an industry, such as increasing returns to scale in production, regardless of where the firm is located. Conversely, the second case is characterized by a co-location of different firms in the same place (spatial concentration), and suggests that there are some "external" elements such as localized externalities, natural resources, factor endowments or demand and input-output linkages, driving the process.

Simultaneously, dynamic intra-industrial economies -within the same activity, or inter-industrial -among different productive activities, show the presence of external effects of a spillover and/or pecuniary nature. These forces affect the territories, and thus the effectiveness of resident establishments, and the firms' ability to growth.

We will define a probability model for the location of establishment which will help us identify spatial clusters of specialized industrial allocation in discrete space for a given specific manufacture sector (activity a, $a \subseteq A$, where A is a the set of all manufacturing sectors: $A = 1, ..., k_A$).

3.1 The model

Currently, there are very few formal models of the overall spatial pattern of industrial agglomerations, and thus the majority of these models are focused on the simple "two-region case" (see Section 1.3.2). However, the extent to which such models are extendable to more complex regional systems is not yet clear (for more details see Fujita and Mori 2005a and b), while there is not even consensus as to how agglomerations should be defined in more general settings.

The purpose of this Section is to develop a probability model to identify specialized agglomerations in terms of multiple-cluster patterns, i.e. we propose an approach based on a probability model for multi-regional systems in terms of statistical cluster analysis. The basic idea is to develop a probability model of multiple clusters, called "cluster schemes". Simply put, a cluster scheme is a space partition through which it is postulated that firms are more likely to locate in cluster partition than elsewhere. Thus, in this partition the model is equivalent to a multinominal sampling model.

The method starts by postulating a null hypothesis of "no specialized agglomeration", i.e. "no clustering" in terms of the uniform distribution of industrial locations across regions. Then, it continues testing this hypothesis on each activity a by finding a single "most significant" contiguous cluster of regions with respect to this hypothesis. In other words, on a first stage we have an individual region, and then it starts adding contiguous regions to find the most significant clusters.

Methodologically, this approach is closely related to cluster-identification methods proposed by Besag and Newell (1991), Kuldorff and Nagarwalla (1995), and Kuldorff (1997), that have been used for the detection of disease cluster in epidemiology. Recently, Mori and Smith (2006) used this approach to the identification "just" industrial agglomerations (as per Section 1.2.3). The description of our model will be based on the latter.

The location behavior of individual establishments in a given activity a can be treated as independent random samples from unknown activity location probability distribution P^a. The observable location data is assumed to be only in terms of establishment counts of a set of disjoint basic regions, $\Omega_r \subseteq \Omega$, indexed by $R = \{1, ..., k_R\}$. These regions are assumed to partition Ω,

$$\bigcup\nolimits_{r=1}^{k_R} \Omega_r = \Omega \tag{3.1}$$

Hence the location probabilities of each basic region of the location probability distribution P^a:

$$P^a = [P^a(r) = P^a(\Omega_r) : r \in R] \tag{3.2}$$

To identify areas of relative intense specialization of an activity a, we now consider an approximation of P^a by probability models, $P^a_{\mathbf{c}}$. Each model is characterized by a "cluster scheme", $\mathbf{C}$, consisting of disjoint regional cluster, $C_j \subset R, j = 1, ..., k_{\mathbf{c}}$, within which specialization activity is supposedly more intense. Hence, the areal extent of cluster C_j in Ω is denoted by

$$\Omega_{C_j} = \bigcup\nolimits_{r \in C_j} \Omega_r \ , \ j = 1, ..., k_{\mathbf{c}} \tag{3.3}$$

The regions $\{\Omega_r : r \in \Omega_{C_j}\}$ in each cluster are contiguous, then Ω_{C_j} is a connected set of regions. Hence the corresponding locations probabilities are

$$p^a_{\mathbf{c}}(j) \equiv P^a_{\mathbf{c}}(\Omega_{C_j}) \ , \ j = 1, ..., k_{\mathbf{c}} \tag{3.4}$$

To complete these probability models, let the set of *remaining or residual regions* be denoted by

$$C_0 = R - \bigcup\nolimits_{j=1}^{k_{\mathbf{c}}} C_j \quad , \quad \Omega_{C_0} = \Omega - \bigcup\nolimits_{j=1}^{k_{\mathbf{c}}} \Omega_{C_j} \tag{3.5}$$

and so the corresponding location probability are

$$p^a_{\mathbf{c}}(0) = P^a_{\mathbf{c}}(\Omega_{C_0}) = 1 - \sum\nolimits_{j=1}^{k_{\mathbf{c}}} p^a_{\mathbf{c}}(j) \tag{3.6}$$

Each *cluster sheme*, $\mathbf{C} = (C_0, C_1, ..., C_{k_{\mathbf{c}}})$ is a partition of the regional index set R, and the location probabilities $[p^a_{\mathbf{c}}(j) : j = 0, 1, ..., k_{\mathbf{c}}]$ yield a probability distribution on $\mathbf{C}$.

Finally, we need to specify a probability distribution for the basic regions. That is, to make an assumption of the conditional probabilities of an individual establishment located in a basic region, $r \in C_j$ given that the establishment belongs to the cluster C_j. We will assume that this probabilities are proportional to the importance of basic region r, that is

$$P^a_{\mathbf{c}}(\Omega_r | \Omega_{C_j}) = \frac{n_r}{n_{C_j}} \quad , \quad r \in C_j \ , \ j = 0, 1, ..., k_{\mathbf{c}} \tag{3.7}$$

where

$$n_{C_j} = \sum\nolimits_{r \in C_j} n_r \tag{3.8}$$

If we consider that inside a geographical area or region, that includes the presence of services, infrastructure, transportation cost and of others factors that enable the development of the industrial activity, it could initially be stated that the industrial localization, or rather, the election of a place inside the area is based on a stochastic mechanism, i.e. completely at random. However, is important to note that the arbitrariness of partitions plays a key role in capturing above-mentioned effects, while it becomes potentially more dangerous the more unequal are the elements of it in terms of area where data are observed within administrative regions that are unequal in size, shape and neighborhood, and where neighboring regions typically resemble each other more than regions that are far apart. In this sense, to minimize MAUP (see Section 1.2.4) the election of the partition would have to reflect the actual characteristics of the economy (for example, the Local Labor Systems in Italy). Due to the unavailability of such partition, we use the proportion of firms for each administrative region as a proxy variable for the presence of factors for the industry localization to make the conditional probabilities of an individual establishment located in a basic region, $r \in C_j$.

Since $\Omega_r \in \Omega_{C_j}$ implies that

$$P^a_{\mathbf{c}}(\Omega_r|\Omega_{C_j}) = \frac{P^a_{\mathbf{c}}(\Omega_r)}{P^a_{\mathbf{c}}(\Omega_{C_j})} = \frac{P^a_{\mathbf{c}}(r)}{p^a_{\mathbf{c}}(j)} \tag{3.9}$$

and for all $r \in R$

$$P^a_{\mathbf{c}}(r) = p^a_{\mathbf{c}}(j)\frac{n_r}{n_{C_j}} \quad , \quad r \in C_j \tag{3.10}$$

Hence for each cluster scheme $\mathbf{C}$, the above formula yields a well defined cluster probability model,

$$P^a_{\mathbf{c}} = [P^a_{\mathbf{c}}(r) : r \in R] \tag{3.11}$$

that is comparable to the unknown true model (3.2). It should be emphasiezed that both P^a and $P^a_{\mathbf{c}}$ are probability models based on basic regions $r \in R$. The first one P^a is a saturated model in which no clustering is achieved. The second one is a simplified one where a specific cluster scheme ($\mathbf{C}$) is postulated. Note that for each given cluster scheme, $\mathbf{C} = (C_0, C_1, ..., C_{k_{\mathbf{c}}})$, the only unknown parameters are given by the $k_{\mathbf{c}}$-dimensional vector of cluster probabilities, $p^a_{\mathbf{c}} = [p^a_{\mathbf{c}}(j) : j = 1, ..., k_{\mathbf{c}}]$.

We will now consider a chosen sequence of n independent location for each establishment i, $i = 1, ..., n$, modeled by a random (indicator) vector, $X^{a(i)} = (X^{a(i)}_r : r \in R)$, with $X^{a(i)}_r = 1$ if i locates in region r, and $X^{a(i)}_r = 0$ otherwise. The random matrix of indicators $X = (X_r^{a(i)} : i = 1, ..., n)$ represents the set of location decisions, with the following finite set of possible realizations *(location patterns)*:

$$\Delta_R^a(n) = \left\{ x = (x_r^{a(i)} : r \in R, i = 1, ..., n) \in \{0,1\}^{n \times k_R} : \sum\nolimits_{r \in R} x_r^{a(i)} = 1, i = 1, ..., n \right\} \tag{3.12}$$

By independency, the probability distribution of the X sample under the unknown true distribution (3.2) is given for each location pattern, $x \in \Delta_R^a(n)$, by:

$$P^a(x) = \prod\nolimits_{i=1}^{n} \prod\nolimits_{r \in R} P^a(r)^{x_r^{a(i)}} \tag{3.13}$$

Likewise, the postulated distribution of X for each cluster probability model $P_{\mathbf{c}}^a$ is given for each location pattern, $x \in \Delta_R^a(n)$, by:

$$P_{\mathbf{c}}^a(x|p_{\mathbf{c}}^a) = \prod\nolimits_{i=1}^{n} \prod\nolimits_{r \in R} P_{\mathbf{c}}^a(r)^{x_r^{a(i)}} = \prod\nolimits_{i=1}^{n} \prod\nolimits_{j=0}^{k_{\mathbf{c}}} \prod\nolimits_{r \in C_j} \left(p_{\mathbf{c}}^a(j) \frac{n_r}{n_{C_j}} \right)^{x_r^{a(i)}} \tag{3.14}$$

It appears that the locational frequencies $n_{C_j}^a(x)$ are sufficient statistics

$$n_{C_j}^a(x) = \sum\nolimits_{i=1}^{n} \sum\nolimits_{r \in C_j} x_r^{a(i)} \quad , \quad j = 0, 1, ..., k_{\mathbf{c}} \tag{3.15}$$

since by definition

$$P_{\mathbf{c}}^a(x|p_{\mathbf{c}}^a) = \prod\nolimits_{j=0}^{k_{\mathbf{c}}} \left[p_{\mathbf{c}}^a(j)^{\sum_{i=1}^{n} \sum_{r \in C_j} x_r^{a(i)}} \prod\nolimits_{i=1}^{n} \prod\nolimits_{r \in C_j} \left(\frac{n_r}{n_{C_j}} \right)^{x_r^{a(i)}} \right] = b_{\mathbf{c}}^a(x) \prod\nolimits_{j=0}^{k_{\mathbf{c}}} p_{\mathbf{c}}^a(j)^{n_{C_j}^a(x)} \tag{3.16}$$

where the factor $b_{\mathbf{c}}^a(x)$

$$b_{\mathbf{c}}^a(x) = \prod\nolimits_{i=1}^{n} \prod\nolimits_{j=0}^{k_{\mathbf{c}}} \prod\nolimits_{r \in C_j} \left(\frac{n_r}{n_{C_j}} \right)^{x_r^{a(i)}} \tag{3.17}$$

is completely independent from the parameter vector $p_{\mathbf{c}}^a$.

3.2 Selection of the best cluster scheme

Since we know that specialization phenomena exits in the manufacturing sector and each probability model $P^a_{\mathbf{c}}$ represents a particular cluster scheme, we aim at to finding the model $P^a_{\mathbf{c}}$ (cluster scheme) which best captured "this specialization phenomena". For this purpose, we postulated the null hypothesis of no specialization and look for the most distant probability model from this null hypothesis. Given a statistic to test the null hypothesis of no specialization we would like to identify the probability model $P^a_{\mathbf{c}}$ which yields the strongest rejection. That is, the best cluster scheme by definition.

3.2.1 Likelihood-ratio statistic

To test the null hypothesis (H_0) of no specialization we need to define a probability model $P^a_0(r)$ that is compatible with (H_0), which in the present context amounts to the hypothesis that $\mathbf{C} = \{C_0\}$. From equation 10, we know that in the no specialization scenario the probability of one establishment of activity a located in the region r should follow n_r/n, so

$$P^a_0(r) = \frac{n_r}{n_{C_0}} = \frac{n_r}{n} \tag{3.18}$$

that is no dependent on activity a.

Hence should we allow $P^a_0 = [P^a_0(r) : r \in R]$ denote the non specialized agglomeration or non clustering, we could now consider the null hypothesis that the true distribution is equal to the non specialization model:

$$H_0 : P^a = P^a_0 \tag{3.19}$$

Thus the non specialization model, P^a_0, is nested in each cluster model, $P^a_{\mathbf{c}}$ (for more details see Kuldorff and Nagarwalla 1995). If we now rewrite (3.18) as

$$P^a_0(r) = \frac{n_r}{n} = \frac{n_{C_j}}{n}\frac{n_r}{n_{C_j}} \quad , \quad r \in R \tag{3.20}$$

then from (3.10), P^a_0 is the special case of $P^a_{\mathbf{c}}$ with

$$p^a_{\mathbf{c}}(j) = \frac{n_{C_j}}{n} \quad , \quad j = 1, ..., k_{\mathbf{c}} \tag{3.21}$$

Hence if we now denote the log likelihood of P^a_0 given x as

$$L^a_0(x) = L(P^a_0|x) = \ln\left[\prod\nolimits_{i=1}^{n}\prod\nolimits_{r\in R} P^a_0(r)^{x_r^{a(i)}}\right] \tag{3.22}$$

and note that by definition

$$\begin{aligned}
L^a_0(x) &= \ln\left[\prod\nolimits_{i=1}^{n}\prod\nolimits_{j=0}^{k_{\mathbf{c}}}\prod\nolimits_{r\in C_j}\left(\frac{n_{C_j}}{n}\frac{n_r}{n_{C_j}}\right)^{x_r^{a(i)}}\right] && (3.23)\\
&= \ln\left\{\prod\nolimits_{j=0}^{k_{\mathbf{c}}}\left[\left(\frac{n_{C_j}}{n}\right)^{\sum_{i=1}^{n}\sum_{r\in C_j} x_r^{a(i)}}\prod\nolimits_{i=1}^{n}\prod\nolimits_{r\in C_j}\left(\frac{n_r}{n_{C_j}}\right)^{x_r^{a(i)}}\right]\right\} && (3.24)\\
&= \ln\left[b^a_{\mathbf{c}}(x)\prod\nolimits_{j=0}^{k_{\mathbf{c}}}\left(\frac{n_{C_j}}{n}\right)^{n^a_{C_j}(x)}\right] && (3.25)\\
&= \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(x)\ln\left(\frac{n_{C_j}}{n}\right) + \ln b^a_{\mathbf{c}}(x) && (3.26)
\end{aligned}$$

then under H_0, a natural test of this hypothesis is the log-likelihood-ratio statistic

$$T^a_{\mathbf{c}}(X) = -2\left[L^a_0(X) - \widehat{L}^a_{\mathbf{c}}(X)\right] \tag{3.27}$$

chi-square distributed with k_c degree of freedom.

From (3.16), for any given cluster scheme $\mathbf{C}$, the log likelihood of parameter vector, $p^a_{\mathbf{c}}$, given observed locations x, is:

$$L(p^a_{\mathbf{c}}|x) = \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(x)\ln p^a_{\mathbf{c}}(j) + \ln b^a_{\mathbf{c}}(x) \tag{3.28}$$

The second term is independent of $p^a_{\mathbf{c}}$, and by differentiation, the maximum-likelihood estimate $\widehat{p}^a_{\mathbf{c}} = [\widehat{p}^a_{\mathbf{c}}(j) : j = 1, ..., k_{\mathbf{c}}]$ of $p^a_{\mathbf{c}}$ for each $j = 1, ..., k_{\mathbf{c}}$ is given by

$$\widehat{p}^a_{\mathbf{c}}(j) = \frac{n^a_{C_j}(x)}{n^a} \tag{3.29}$$

Hence, the associated estimate of the maximum-likelihood value for model $P^a_{\mathbf{c}}$ is given by

$$\widehat{L}^a_{\mathbf{c}}(x) = L(\widehat{p}^a_{\mathbf{c}}|x) = \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(x) \ln\left(\frac{n^a_{C_j}(x)}{n^a}\right) + \ln b^a_{\mathbf{c}}(x) \tag{3.30}$$

This together with (3.26) shows that

$$\begin{aligned} L^a_0(X) - \widehat{L}^a_{\mathbf{c}}(X) &= \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(\frac{n_{C_j}}{n}\right) - \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(\frac{n^a_{C_j}(X)}{n^a}\right) && (3.31)\\ &= \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(\frac{n_{C_j}/n}{n^a_{C_j}(X)/n^a}\right) && (3.32)\end{aligned}$$

and hence that

$$\begin{aligned} T^a_{\mathbf{c}}(X) &= -2\sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(\frac{n_{C_j}/n}{n^a_{C_j}(X)/n^a}\right) && (3.33)\\ &= 2\sum\nolimits_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(\frac{n^a_{C_j}(X)/n^a}{n_{C_j}/n}\right) && (3.34)\end{aligned}$$

It must be noted that the argument of the logarithm in (3.34) is the Hoover-Balassa Local Quotient coefficient (LQ_{ij}) (see Section 2.1). In addition, if we divide both sides of (3.34) for $2\sum_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X)$

$$\frac{T^a_{\mathbf{c}}(X)}{2\sum_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X)} = \frac{\sum_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X) \ln\left(LQ^a_{C_j}\right)}{\sum_{j=0}^{k_{\mathbf{c}}} n^a_{C_j}(X)} \tag{3.35}$$

$$\frac{T^a_{\mathbf{c}}(X)}{2N} = \sum\nolimits_{j=0}^{k_{\mathbf{c}}} w^a_{C_j} \ln\left(LQ^a_{C_j}\right) \tag{3.36}$$

we will obtain the similar weight $w^a_{C_j}$ used heuristically by Donato and Haedo (2002).

The asymptotic P-value for this likelihood-ratio test is given by

$$P - value = 1 - F_{k_{\mathbf{c}}}(T^a_{\mathbf{c}})$$

where $F_{k_{\mathbf{c}}}$ denotes the cumulative distribution function for the chi-square distribution with $k_{\mathbf{c}}$ degrees of freedom. We rejects the null hypothesis if the value of the likelihood-ratio test is sufficiently large, or if the corresponding P-value is sufficiently small.

Given the set of basic regions, it would of course be desirable to compare all possible cluster schemes that can be formed from these regions, and then to identify best cluster scheme. But there could be an extraordinary number of possible cluster schemes can be enormous for even modest numbers of basic regions and in addition the chi-square distribution has fatter tails for larger degrees of freedoms. As a result of these features, in most cases the P-value is zero. This trivial result does not allow us to use the P-value as a sensible measure to compare all possible cluster schemes and any comparison would yield a dumb comparison of zero with zero. The next measure, BIC, differs from this test precisely in the way it penalizes larger numbers of clusters and offers an additional model-selection criteria to resolve these potential over-fitting problems.

3.2.2 The bayesian information criterion (BIC)

The bayesian information criterion (BIC) was introduced by Schwartz (1978). This BIC has proved to yield a consistent estimator of a true model whenever it is among the candidate models. Now the unknown parameter vector, $p^a_{\mathbf{c}}$, is a random vector with prior distribution given by density, $\psi^a_{\mathbf{c}}(\cdot)$. Hence the associated marginal event probabilities $P^a_{\mathbf{c}}$ are given by

$$P^a_{\mathbf{c}}(x) = E_{p^a_{\mathbf{c}}}[P^a_{\mathbf{c}}(x|p^a_{\mathbf{c}})] = \int_{p^a_{\mathbf{c}}} P^a_{\mathbf{c}}(x|p^a_{\mathbf{c}})\psi^a_{\mathbf{c}}(p^a_{\mathbf{c}})dp^a_{\mathbf{c}} \tag{3.37}$$

These Bayes factors or marginal probabilities are another natural criterion for model selection. Given any two candidate cluster schemes, **C** and **C'**, for data $x \in \Delta_R^a(n)$, if $P_{\mathbf{c}}^a(x) > P_{\mathbf{c'}}^a(x)$ the principle of maximum likelihood suggests that **C** should be a better data model than **C'**. This can be equally written as a likelihood ratio condition:

$$\frac{P_{\mathbf{c}}^a(x)}{P_{\mathbf{c'}}^a(x)} > 1 \tag{3.38}$$

or

$$\ln P_{\mathbf{c}}^a(x) - \ln P_{\mathbf{c'}}^a(x) > 0 \tag{3.39}$$

Hence for large sample sizes n

$$\ln P_{\mathbf{c}}^a(x) = \ln E_{p_{\mathbf{c}}^a}[P_{\mathbf{c}}^a(x|p_{\mathbf{c}}^a)] \approx \ln P_{\mathbf{c}}^a[x|\widehat{p}_{\mathbf{c}}^a(x)] - \frac{k_{\mathbf{c}}}{2}\ln(n) \tag{3.40}$$

where this restrictive formulation is completely independent from the prior density $\psi_{\mathbf{c}}^a(\cdot)$, and for large sample sizes, the posterior distribution of $p_{\mathbf{c}}^a$ given data $x \in \Delta_R^a(n)$ eventually concentrates around $\widehat{p}_{\mathbf{c}}^a(x)$, regardless of the prior distribution. If we multiply both sides by -2, then the best models of x are those with the smallest

$$\begin{aligned} BIC_{\mathbf{c}}^a(x) &= -2\ln P_{\mathbf{c}}^a[x|\widehat{p}_{\mathbf{c}}^a(x)] + k_{\mathbf{c}}\ln(n) && (3.41)\\ &= -2\widehat{L}_{\mathbf{c}}^a(x) + k_{\mathbf{c}}\ln(n) && (3.42)\end{aligned}$$

values. If we now denote the BIC value for the random benchmark model by

$$BIC_0^a(x) = -2L_0^a(x) \tag{3.43}$$

then we may reformulate this measure in terms of BIC-differences from this benchmark as follows:

$$\begin{aligned} \Delta_{\mathbf{c}}^{BIC}(x) &= BIC_0^a(x) - BIC_{\mathbf{c}}^a(x) && (3.44)\\ &= -2[L_0^a(x) - \widehat{L}_{\mathbf{c}}^a(x)] - k_{\mathbf{c}}\ln(n) && (3.45)\\ &= T_{\mathbf{c}}^a(x) - k_{\mathbf{c}}\ln(n) && (3.46)\end{aligned}$$

and hence the better cluster schemes $\mathbf{C}$ are now those with larger difference values $\Delta_{\mathbf{c}}^{BIC}$.

If we rewrite (3.34) as follows

$$T_{\mathbf{c}}^{a}(X) = 2\left\{ n \ln \frac{n}{n^a} - \sum\nolimits_{j=0}^{k_{\mathbf{c}}} n_{C_j} \ln \frac{n_{C_j}}{n_{C_j}^{a}(X)} \right\} \tag{3.47}$$

we would illustrate how the degree of freedom $k_{\mathbf{c}}$ penalizes larger number of clusters in the likelihood-ratio test. The penalty for adding parameters grows without limitations as the sample size n increases. Hence, would be reasonable to expect that models with a smaller number of parameters will be favored as n becomes larger.

Note that BIC-differences (3.46) differ from likelihood-ratio test (3.47) as much as they penalize larger number of clusters. This BIC measure will always yield a consistent selection of the true model whenever this model is one of the candidates and tend to select a more parsimonious model when sample sizes are sufficiently large.

3.3 Cluster detection procedure

Methodologically, this procedure is in contrast with those that are often adopted to detect disease cluster in epidemiology. Openshaw et al. (1988) proposed the geographical analysis machine (GAM) as an exploratory cluster detection method. Besag an Newell (1991) proposed statistically rigorous alternatives to the GAM based on circles of fixed populations radius and circles of fixed case radius, respectively. Kuldorff and Nagarwalla (1995) and Kuldorff (1997) generalized the previous procedures to arbitrary collections of clusters using likelihood ratio test. They typically consider a circular area centered at each region to be a potential cluster, and find the central region and the radius of the circle which corresponds to the highest significance level of concentration. However, such circular clustering of regions will contain many irrelevant low-density regions unless the “true” agglomeration is roughly circular. Thus, approaches tend to result in identifying unreasonable large clusters. In this sense, Mori and Smith (2006) define a minimal boundary closure basically assumes convexity of each cluster, i.e. consider that a roughly convex closure of the significant regions is the geographic coverage of the cluster.

Our approach is essentially an elaboration of the basic ideas proposed by Besag and Newell (1991) in which given the set of basic regions we could start with individual regions and then add contiguous region to find the most significant cluster, comparing all possible cluster scheme that can be formed from these regions. Hence, to identify best cluster scheme,

$$\mathbf{C}^* = \arg\max_{\mathbf{C}} \Delta_{\mathbf{c}}^{BIC} \tag{3.48}$$

But as mentioned in Section 3.2.1, the number of possible cluster schemes can be enormous for even modest quantities of basic regions and the procedure of clusterization could remain in a loop on a local maxima. Thus, it is necessary to consider limited search procedures that yield reasonable approximations to best cluster schemes.

We developed a greedy forward algorithm that uses the $\Delta_{\mathbf{c}}^{BIC}$ as a selection criteria and starts with a baseline configuration with all regions forming a unique consolidated cluster. The steps of the procedure are the follows:

1 The first step is to choose the region, which will form a separated one region cluster, that maximizes the configuration criteria based on larger difference values of $\Delta_{\mathbf{c}}^{BIC}$. There are R possible regions to choose from.

  The outcome of this first step is a two cluster configuration: one cluster formed by the chosen region and the other cluster consisting of the remaining regions.

2 The second step is to choose from the $R - 1$ not chosen regions the region which maximizes the configuration criteria.

  The outcome of the second step will depend on the cluster configuration of the previous step. At least three regions should have been formed: i) one cluster with the first chosen region; ii) another cluster with the second chosen region; and iii) a third cluster with the remaining regions.

  If the two chosen region are contiguous then the least number of clusters is two: one with the two chosen regions and the other with the remaining regions.

3 The algorithm stops when no choice of region provides an increase in the configuration criteria.

The algorithm follows in a forward manner choosing the regions, one by one, whilst the configuration criteria is increased. It must be noticed that this methodology does not provide a global maximum of the configuration criteria. It is possible that the best configuration, in terms of the selected configuration criteria, is different from the resulting configuration of the algorithm.

3.4 Application: Manufacture industry in Chile

The spatial units are the lower level political-administrative jurisdictions in Chile (communes). The firms' data for the manufacture sector with 5 or more employees have been taken from the National Institute of Statistics and Censuses of Chile (INE-2005).The activity classifications refer to the first 2 digits of the International Standard Industrial Classification (ISIC-Rev.3) for the following manufacturing sectors:

- Division 24: Manufacture of chemicals and chemical products;
- Division 20: Manufacture of wood and of products of wood and cork, except furniture; and manufacture of articles of straw and plaiting materials;
- Division 25: Manufacture of rubber and plastics products;
- Division 28: Manufacture of fabricated metal products, except machinery and equipment; and
- Division 29: Manufacture of machinery and equipment n.e.c.

3.4.1 Specialized agglomeration of division ISIC 24

Division 24 has 33,078 employees in 304 firms distributed across 69 regions. Table 3.1 shows the best cluster scheme formed by 5 clusters: two high-specialization clusters (1 and 2), two low specialization clusters (N1 and N2), and the remaining regions. It must be noted that non specialization implies (by Section 2.2) that the joint proportion of firms in region i in activity j is equal to the product of marginal proportions in region i and activity j. Consequently, the best cluster scheme is made of "over-specialized" (clusters 1 and 2) and "sub-specialized" regions (clusters N1 and N2), following the trade-off in order between the number of firms, contiguity or number of clusters, and very high and low specialization levels. In this respect, the column of "% of firms" in Table 3.1 shows in each cluster how the model-selection criteria to obtain the best cluster scheme prioritizes the number of firms in the first place.

Table 3.1 shows that the "over-specialized" clusters 1 and 2, each formed by 3 contiguous regions (in the middle and upper portion of Fig. 3.1), have 12% and 17% of firms and 3% and 6% of employees, respectively (37 firms with 5,706 employees and 8 firms with 2,046 employees, respectively). Together, these clusters add 15 percent and 24 percent of the firms and employees in the whole sector at a national level, respectively.

The "sub-specialized" cluster N1 is formed by 10 contiguous regions (in the center of Fig. 3.1) and has 19% of firms and employees in the whole sector at a national level (59 firms with 6,405 employees). The "sub-specialized" cluster N2, shown in Fig. 3.2, is formed by only one region that represents the largest number of firms (equal to 2) between those with a lower specialization level or sub-specialized regions ($LQ_{ij} = 0.45$).

Finally, the remaining regions are 52 and have 65% and 57% of firms and employees in the whole sector at a national level, respectively (197 firms and 18,721 employees).

Fig. 3.1 and 3.2 show the location of the regions with the best cluster scheme for sector 24.

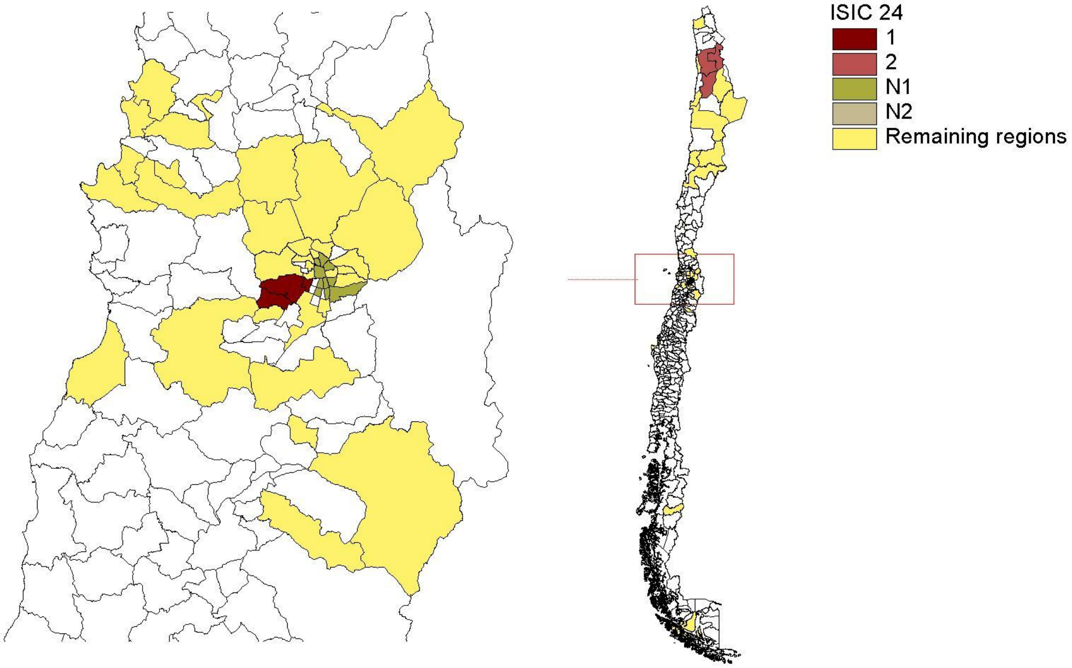
Table 3.1: *Best cluster scheme of division ISIC 24: Manufacture of chemicals and chemical products*

Cluster	# of regions	% of firms[1]	% of employees[2]	LQ$_{ij}$			
				Mean	Variance	Min	Max
Over-specialized 1	3	12.2	17.3	3.19	0.33	2.39	3.63
Over-specialized 2	3	2.6	6.2	14.10	15.91	8.46	16.92
Sub-specialized N1	10	19.4	19.4	0.74	0.04	0.41	1.15
Sub-specialized N2	1	1.0	0.6	0.45			
Remaining regions	52	64.8	56.5	2.87	10.54	0.37	16.92

[1] Percentage of firms in the whole sector at a national level.

[2] Percentage of employees in the whole sector at a national level.

Figure 3.1: *Map 1 for specialized agglomeration of division ISIC 24*

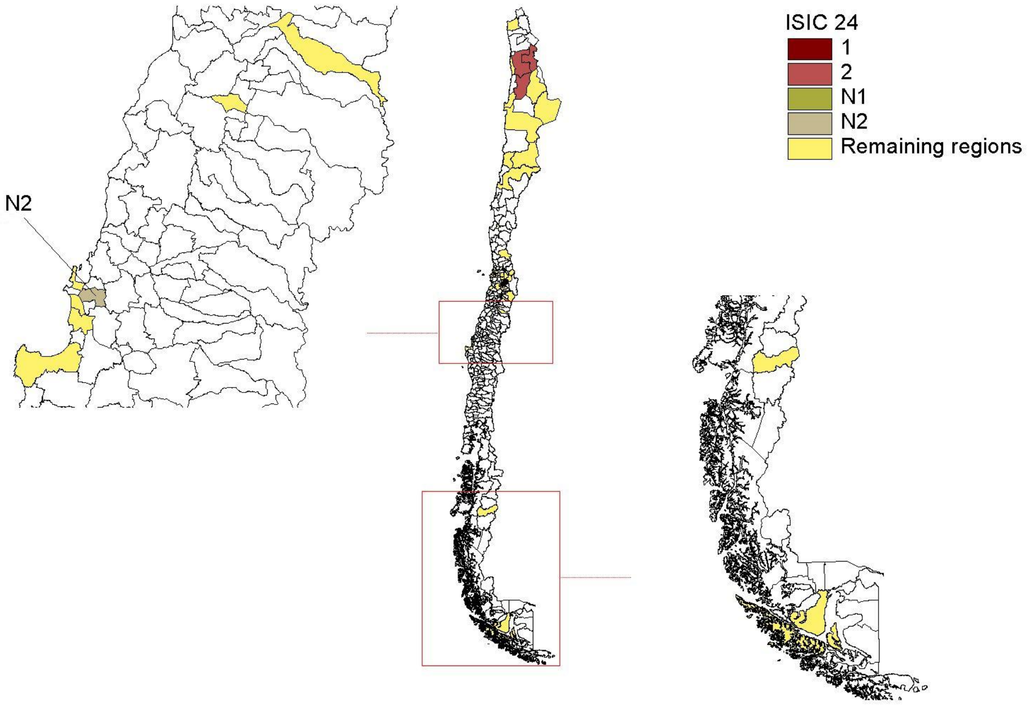


3.4.2 Specialized agglomeration of division ISIC 20

Division 20 has 39,745 employees in 337 firms distributed across 121 regions. The best cluster scheme for this sector is formed by 8 clusters: 4 over-specialized, 3 sub-specialized and the remaining regions.

Table 3.2 shows that over-specialized agglomeration 1, formed by 20 contiguous regions, has 21% of firms and 45% of employees in the whole sector at a national level (70 firms with 17,698 employees). Over-specialized clusters 2 and 4 are formed one by only 1 region each, while the over-specialized cluster 3 is formed by 4 regions with de same value of $LQ_{ij} = 15.22$. With respect to cluster 1, these last three clusters show a greater proportion of firms and employees. Together, these 4 over-specialized clusters add 32 percent and 56 percent of the firms and employees in the

Figure 3.2: *Map 2 for specialized agglomeration of division ISIC 24*



whole sector at a national level, respectively. These results show that the firms in this manufacturing sector have a relatively higher tendency to co-localize in specialized agglomerations with respect to sector 24.

The "sub-specialized" cluster N1 is formed by 29 contiguous regions and has 18% of firms and 6% of employees in the whole sector at a national level. Although the firms' data for the manufacturing sector in Chile is related to firms with 5 or more employees, this cluster is formed by smaller firms, as clusters N2 and N3. Together, these 3 sub-specialized clusters add 23 percent and 7 percent of the firms and employees in the whole sector at a national level, respectively.

Finally, the remaining regions totalize 61 and have 45% and 37% of firms and employees in the whole sector at a national level, respectively.

Fig. 3.3 shows the location of the regions in the best "over-specialized" cluster scheme formed by 4 clusters.

Table 3.2: *Best cluster scheme of division ISIC 20: Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials*

Cluster	# of regions	% of firms[1]	% of employees[2]	LQ_{ij}			
				Mean	Variance	Min	Max
Over-specialized 1	20	20.7	44.5	7.63	13.57	2.17	15.22
Over-specialized 2	1	6.5	4.4	13.39			
Over-specialized 3	4	3.3	4.5	15.22	0		
Over-specialized 4	1	1.5	2.2	9.51			
Sub-specialized N1	29	18.3	5.8	0.54	0.12	0.08	1.25
Sub-specialized N2	4	4.1	1.4	1.15	0.13	0.81	1.69
Sub-specialized N3	1	0.3	0.1	0.18			
Remaining regions	61	45.3	37.1	4.77	19.39	0.42	15.22

[1] Percentage of firms in the whole sector at a national level.

[2] Percentage of employees in the whole sector at a national level.

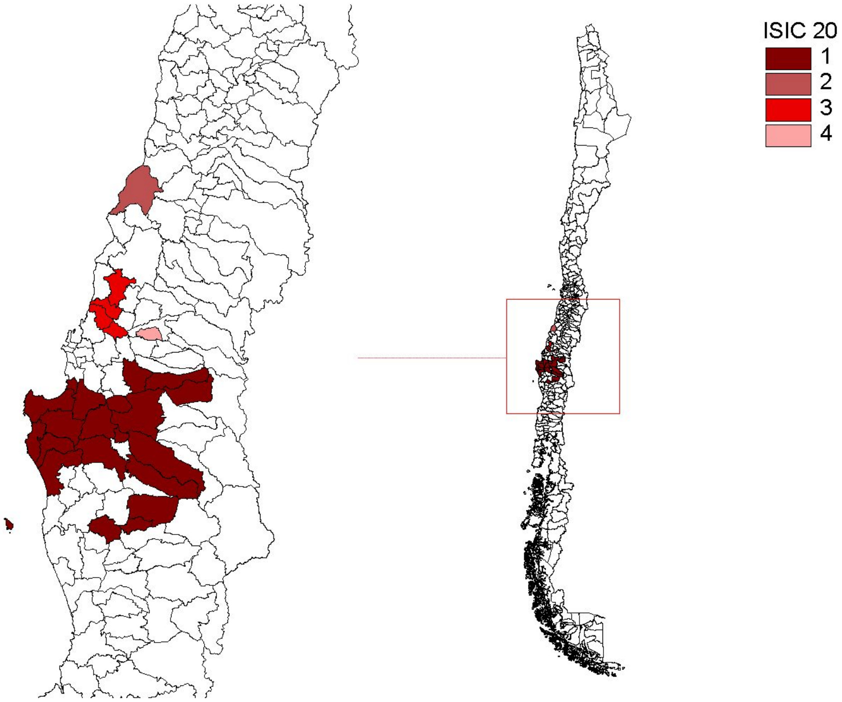
3.4.3 Specialized agglomeration of division ISIC 25

Division 25 has 22,929 employees in 326 firms distributed across 66 regions. The best cluster scheme for this sector is formed by only 2 clusters: 1 over-specialized and the remaining regions.

Table 3.3 shows that over-specialized agglomeration, formed by 18 contiguous regions, has 67% of firms and employees in the whole sector at a national (217 firms with 15,359 employees). These results show that the firms in this manufacturing sector have a high tendency to co-localize in specialized agglomerations.

Fig. 3.4 shows the location of the 18 contiguous regions in the over-specialized cluster.

Figure 3.3: *Over-specialized agglomeration of division ISIC 20*



3.4.4 Specialized agglomeration of division ISIC 28

Division 28 has 24,156 employees in 393 firms distributed across 76 regions. As with division 25, the best cluster scheme of this sector is formed by only 2 clusters: 1 over-specialized and the remaining regions.

Table 3.4 shows that over-specialized agglomeration, formed by 20 contiguous regions, has 51% of firms and 56% of employees in the whole sector at a national level (201 firms with 13,504 employees). These results show also that the firms in this manufacturing sector have a high tendency to co-localize in specialized agglomerations.

Fig. 3.5 shows the location of the 20 contiguous regions in the over-specialized

Table 3.3: *Best cluster scheme of division ISIC 25: Manufacture of rubber and plastics products*

Cluster	# of regions	% of firms[1]	% of employees[2]	LQ_{ij}			
				Mean	Variance	Min	Max
Over-specialized 1	18	33.4	33.0	1.93	0.43	0.99	3.36
Remaining regions	48	66.6	67.0	1.12	0.83	0.25	5.26

[1] Percentage of firms in the whole sector at a national level.

[2] Percentage of employees in the whole sector at a national level.

cluster.

Table 3.4: *Best cluster scheme of division ISIC 28: Manufacture of fabricated metal products, except machinery and equipment*

Cluster	# of regions	% of firms[1]	% of employees[2]	LQ_{ij}			
				Mean	Variance	Min	Max
Over-specialized 1	20	51.1	55.9	2.75	5.93	1.58	13.09
Remaining regions	56	48.9	44.1	1.31	3.44	0.16	13.09

[1] Percentage of firms in the whole sector at a national level.

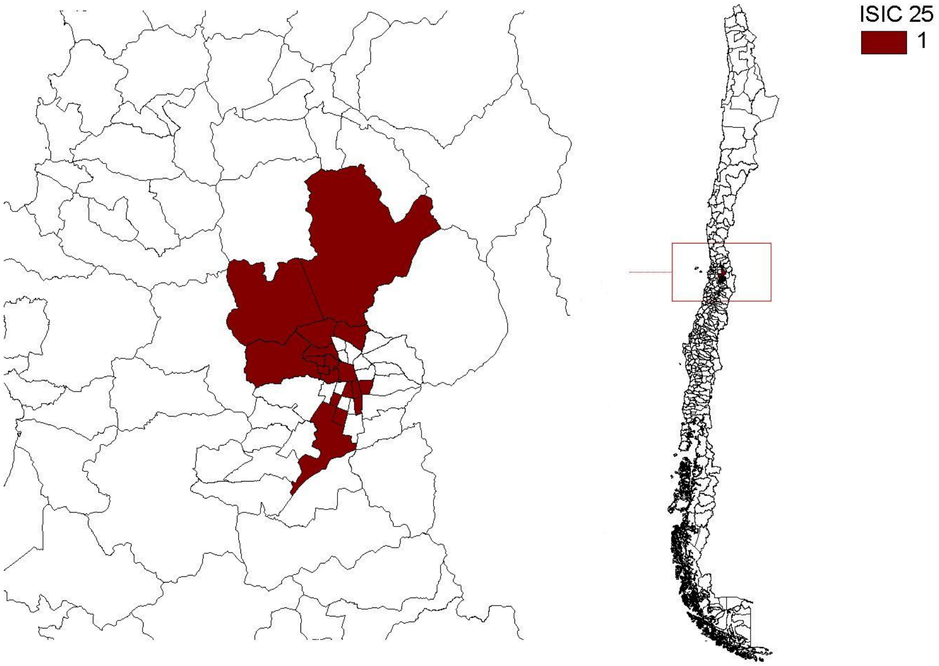
[2] Percentage of employees in the whole sector at a national level.

3.4.5 Specialized agglomeration of division ISIC 29

Division 29 has 19,140 employees in 308 firms distributed across 76 regions. The best cluster scheme for this sector is formed by 3 clusters: 2 over-specialized and the remaining regions.

Table 3.5 shows that over-specialized cluster 1, formed by 15 contiguous regions, has 35% of firms and 41% of employees in the whole sector at a national level (107 firms with 7,871 employees). The over-specialized cluster 2 is formed by 2 contiguous regions and has 10% and 7% of the firms and employees in the whole sector at a national level, respectively (31 firms with 1,334 employees). Together, these clusters add 45 percent and 48 percent of the firms and employees of the whole sector at a national level, respectively.

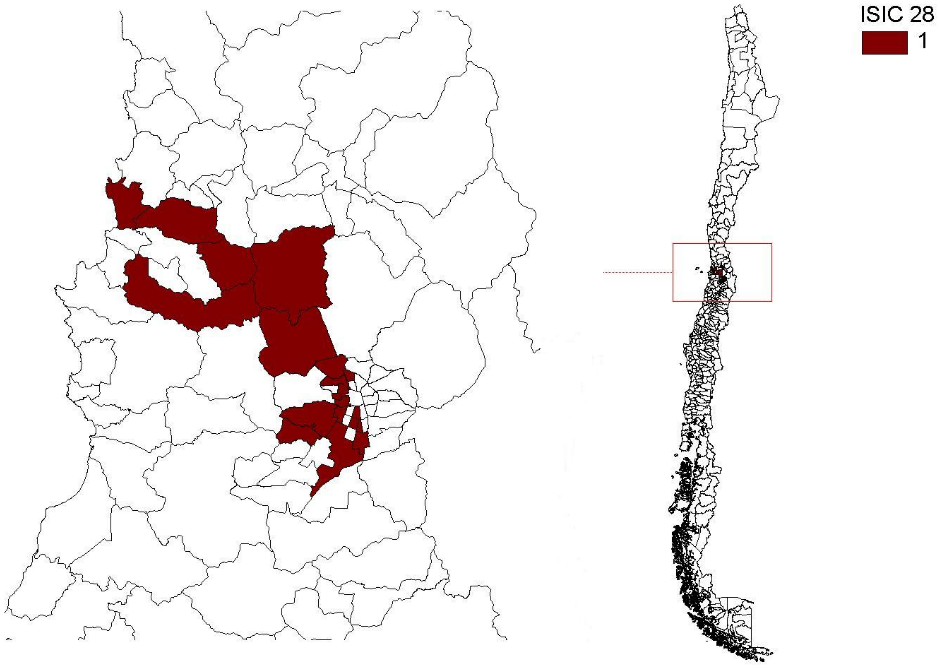
Figure 3.4: *Over-specialized agglomeration of division ISIC 25*



As with divisions 25 and 28, the firms of this manufacturing sector show a high tendency to co-localize in specialized agglomerations.

Fig. 3.6 shows the location of the contiguous regions in over-specialized clusters (18 regions in cluster 1 and 2 regions in cluster 2).

Figure 3.5: *Over-specialized agglomeration of division ISIC 28*



3.4.6 Some remarks about the co-localization of firms in specialized agglomerations

The purpose of this section is not to provide a thorough analysis of the phenomenon of the co-localization of firms in specialized agglomerations, but rather to highlight some relevant features that can drive the analysis and its comparison with similar studies.

It must be noted that the firms' data for the Chile manufacturing sector is related to establishments with 5 or more employees. The firm and employee rates are calculated on the basis of the whole sector at a national level.

To summarize the prior sections, if we analyzed Fig 3.7 it becomes readily ap-

Table 3.5: *Best cluster scheme of division ISIC 29: Manufacture of machinery and equipment n.e.c.*

Cluster	# of regions	% of firms[1]	% of employees[2]	LQ_{ij}			
				Mean	Variance	Min	Max
Over-specialized 1	15	34.7	41.1	1.24	0.15	0.52	2.03
Over-specialized 2	2	10.1	7.0	2.35	0.19	1.92	2.78
Remaining regions	59	55.2	51.9	1.53	1.45	0.33	8.35

[1] Percentage of firms in the whole sector at a national level.

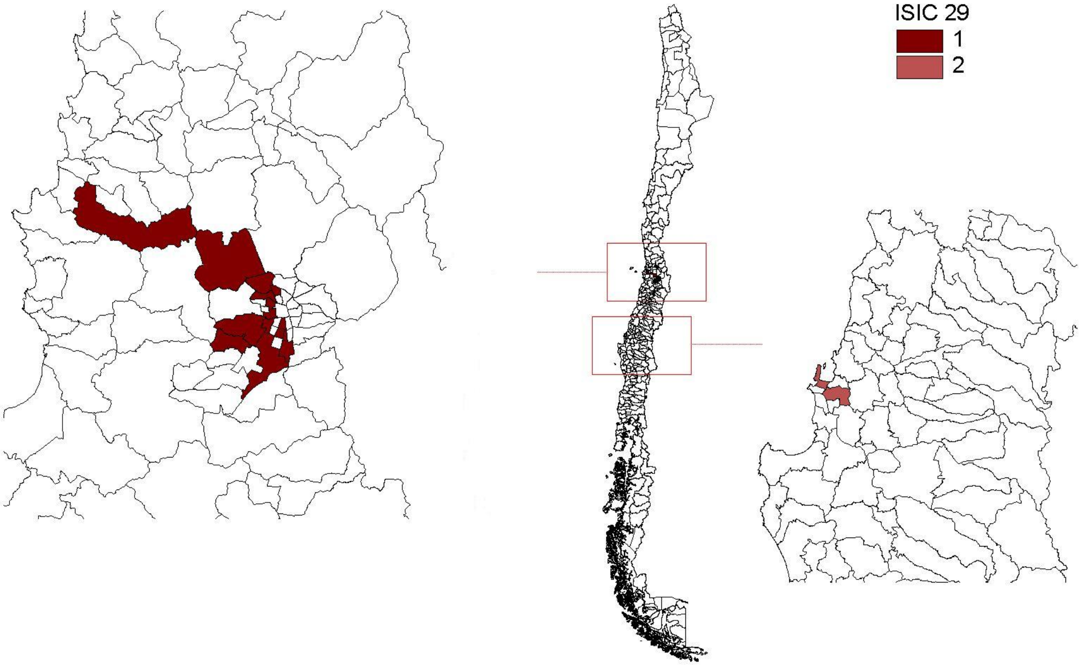
[2] Percentage of employees in the whole sector at a national level.

parent that the firms of ISIC 25, 28 and 29 are prone to co-locate in over-specialized agglomerations, while it becomes more evident in division 25. This can be supported in two features. Firstly, over-specialized agglomerations are formed by a single cluster of contiguous regions that in the aggregate represent less than a third of the total number of regions in the country with a sector firm (see Figures 3.4, 3.5 and 3.6, respectively). Secondly, the non-existence of significant sub-specialized agglomerations denotes that these sector firms are either located homogeneously across the remaining regions without affecting their structure sector configuration, or are located in over-specialized agglomerations. Additionally, it must be noted that these sector firms have similar mean sizes, both within over-specialized agglomerations and in the remaining regions, while they are hardly greater than the national average in the over-specialized clusters of divisions 28 and 29 (Fig. 3.8).

Although the firms in division 20 show a lower tendency to co-locate in over-specialized agglomeration vis-à-vis the previous sectors (Fig 3.7), nonetheless Fig. 3.8 shows that the mean size of the firms located there is greater than that of the firms located in the remaining clusters (sub-specialized and remaining regions) and that of the national average. This could be supported in the direct relationship between the sector and the presence of natural resources.

Contrary to the preceding examples, the firms in division 24 show a low tendency to co-locate in specialized regions, and practically do not affect the sector structure configuration of the regions where the firms are located. However, as shown in Fig. 3.8, the size of the firms located in over-specialized agglomerations is greater compared to the remaining clusters and the national average, a situation that could be also compared to the case of the German chemical cluster, in which a small number of large companies are located together, not only to reduce production costs but also as a strategy towards their competitors. The formation of this type of

Figure 3.6: *Over-specialized agglomeration of division ISIC 29*



clusters would seem to depend on the relative strength of the size of the localization economies (for more details, see Section 1.3.3.).

Figure 3.7: *Percentage of firms and employees in the best cluster schemes*

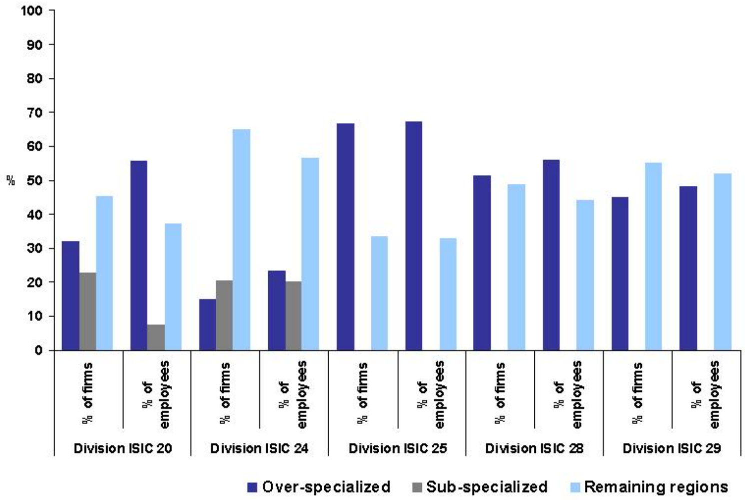
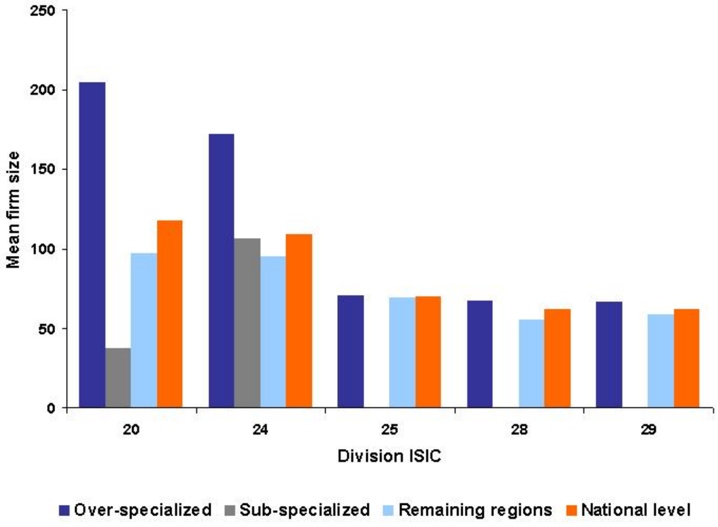


Figure 3.8: *Mean firm size in the best cluster schemes*



Chapter 4

Spatial point patterns clustering for the identification of specialized agglomeration

Unlike in Chapters 2 and 3, in which regions were defined according to process-exogenous criteria, namely administrative entities, in this Chapter the space is a unique continuum. The spatial process approach is based on geocodified data, and aims at assessing the power of attraction from a local space perspective. As we mentioned in chapter 1 (MAUP problem), the availability of geocodified data allow for quantifying the specialization level of a certain activity at a particular point in space. From the point of view of a non-homogeneous Poisson process, firm localization points are randomly distributed, and disjoint area counts are mutually independent, each based on Poisson's distribution according to which the intensity parameter forms a finite measure of the reference space, in this case a bi-dimensional space. This measure may be interpreted as the representation of the differentiated power of attraction of the space and, in the case of a specific area, the expected value of the number of locations in such area.

The approach of spatial point pattern analysis using de geocodified data already has known considerable developments in economic geography. Barff (1987) uses a point pattern analysis of manufacturing in Cincinnati (Ohio) and focuses on the importance of production technology in understanding the degree of urban industrial clustering. Duranton and Overman (2005) study the detailed location pattern of industries, and particulary the tendency for industries to cluster relatively to overall manufacturing through the develop of distance-based tests of localization. Recently, Arbia et al. (2007) suggest an adequate model of spatial coagglomeration of industries and describe a class of spatial statistical methods to be used in the empirical analysis of spatial clusters. They use a set of European Patent Office (EPO) data and produce a series of empirical evidences referred to as the pairwise intrasectoral spatial distribution of patents in Italy in the nineties. In this analysis they are able to identify some distinctive joint patterns of location between patents of different sectors and to propose some possible economic interpretations. These works focus on the co-agglomerations or clustering tendency of manufacturing sectors, i.e. on the spatial concentration or just agglomeration (in the sense of Section 1.2.3) but do not pay attention to the identification of these spatial clusters. By contrast, this chapter concentrate the attention on the *identification* of *specialized* agglomerations.

4.1 Spatial point patterns

A spatial point pattern is a set of locations irregularly distributed within a region of interest M. We refer to this set of points as events, to distinguish them from arbitrary points of the region in question.

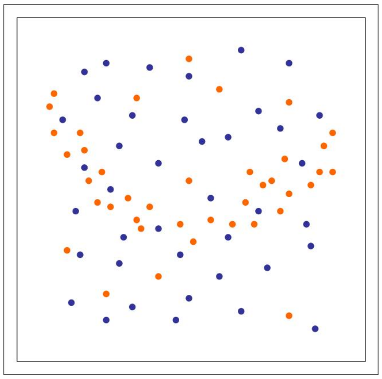
Throughout this chapter, bold face will be used for points in a two-dimensional space. The observed point pattern $\mathbf{x}$ will be treated as a realization of a random point process $\mathbf{X}$. A point process is a stochastic process generating a random set of points; the number of points is random, as well as the locations of the points.

In the case of a Poisson processes, the intensity of the processes represents expected number of points per area. The intensity may be constant (uniform or homogeneous) or may vary from location to location (non-homogeneous).

Fig 4.1 introduces the idea of multivariate point pattern. In this fictitious example, the points represents firms of two different types (hence, bivariate) in manufacture sectors. The data consist of the location of 80 firms, amongst which 40 are of orange activity, whilst the remaining 40 blue points are the firms in the rest of activities. Fig. 4.1, might also represent the position of two types of trees in an area of woodland or cells in the retina.

For more details about spatial point patterns see Diggle (2003) and Møller and Waagepetersen (2004).

Figure 4.1: *Bivariate point pattern*



4.1.1 Poisson processes

A Poisson process is a counting process and its characteristic feature is a property of statistical independence. More specifically:

Definition 1. Let Λ be a Borel measure on $\mathbb{R}^2$. A Poisson process with parameter Λ is a point processes with the following properties:

- (PP1): for any bounded Borel $A \subseteq \mathbb{R}^2$, $N(A)$ has a Poisson distribution with intensity parameter $\Lambda(A)$. Thus $E[N(A)] = \Lambda(A)$;
- (PP2): for any bounded disjoint Borel sets A and B, $N(A)$ and $N(B)$ are stochastically independent.

In the smooth case, the intensity Λ may be represented by the Riemann's integral of a local intensity λ: $\Lambda(A) = \int_A \lambda(\mathbf{x})\ d\mathbf{x}$. When the intensity is a finite measure, i.e. $\Lambda(M) = m$, $f(\mathbf{x}) = \lambda(\mathbf{x})/m$ is a probability density: $\int_M f(\mathbf{x})\ d\mathbf{x} = 1$.

Definition 2. A Poisson process is stationary, or homogeneous, if the parameter Λ is proportional to the Lebesgue measure, i.e. $\Lambda(A) = m|A|$ for some finite m where $|A|$ is the Lebesgue measure (area) of A. This process is sometimes referred to *Complete Spatial Randomness* (CSR) (Kingman 1967).

Thus, in the general case the expected number of events in space varies as the intensity function does, whereas in the homogeneous case, the expected number of events is constant per unit area. In both cases, however, spatial independence is maintained. It might be emphasized that in a Poisson process approach, a significant clusterization of point is attributed to the variation of the process intensity. Moreover, this intensity is deemed to represent the only cause of clusterization because conditionally on the intensity what is left is pure randomness.

If it is suspected that the intensity may be non-homogeneous, the intensity function or intensity measure can be estimated nonparametrically by techniques similar to those used for estimating probability densities, such as quadrat counting and kernel smoothing. For more details about Poisson processes see Daley and Vere-Jones (1972), Ripley (1981), Kingman (1993) and Kutoyants (1998).

4.2 Measurement of local specialization in continuous space

Consider now a universe of reference M where M is a Borel region of $\mathbb{R}^2$. For the spatial analysis of specialization it is natural to assume that $\Lambda(M) = m$ is finite. We may accordingly decompose the Poisson process into a (marginal) process generating a number of points in M, i.e. a Poisson random variable $N(M)$ with parameters $\Lambda(M)$, and a process conditional on $N(M)$ generating the location of the $N(M)$ points. In such a case, as mentioned above, the normalized intensity $f(\mathbf{x}) = \lambda(\mathbf{x})/m$ is a probability density.

When considering a multivariate Poisson process, $N = (N^1, ..., N^a, ..., N^A)$, where N^a is a Poisson process for activity a, $N^a(M)$ represents the total number of firms for activity a in region M, the realization of witch is denoted n^a. When aggregating all activities, we define the random process $N^+ = \sum_a N^a$ and n^+, the realized value of $N^+(M)$, is the realized total number of firms for region M.

Thus, we now assume that the localization pattern takes the form of a non-homogeneous Poisson process, conditional on the total number of firms, namely n^a for activity a and n^+ for all activities aggregated. Here, a significant clustering of firms is interpreted as an effect of higher values of the intensity function of the process.

Let us adjust to the continuous case the Hoover-Balassa Local Quotient coefficient, introduced in Section 2.1, as follows:

$$LQ^a(\mathbf{x}) = \frac{\frac{\lambda^a(\mathbf{x})}{n^a}}{\frac{\lambda^+(\mathbf{x})}{n^+}} = \frac{f^a(\mathbf{x})}{f^+(\mathbf{x})} \tag{4.1}$$

where $\lambda^a(\mathbf{x})$, respectively $\lambda^+(\mathbf{x})$, is the local intensity of the process associated to activity a, respectively of the aggregated process. Two features should be noticed of this local measurement of specialization. Firstly, the normalized densities f^a and f^+ integrate to 1 and have therefore comparable values at point $\mathbf{x}$ whereas λ^a and λ^+ do not. Secondly, the equality to 1 of this ratio means that point $\mathbf{x}$ has a similar attractivity among the n^a firms of sector a and among the n^+ firms of all sectors together. Thus, this ratio close to 1 for each sector of activity means that this point reveals no specialization. A ratio higher than 1 in a substantial neighborhood of a given point $\mathbf{x}$ reveals an area particularly attractive for a specific activity and eventually a degree of specialization.

4.2.1 Average Specialization Measure (ASM)

Because in formula (4.1), no specialization around activity a at point $\mathbf{x}$ means that $\overline{LQ}^a(\mathbf{x}) = 1$, i.e. $\lambda^a(\mathbf{x})/n^a = \lambda^+(\mathbf{x})/n^+$, a possible Average Specialization Measure (ASM) for the continuous space M is

$$ASM = \frac{1}{n^+} \sum_{a=1}^{A} \sum_{i \in I^a} \left(\frac{\lambda^a(\mathbf{x}_i)}{n^a} - \frac{\lambda^+(\mathbf{x}_i)}{n^+} \right)^2 \tag{4.2}$$

where i is the index of firms, and I^a is the set of the indexes of firms in activity a; more explicitly, the firms are ordered in such a way that: $I^1 = \{1, ..., n^1\}$, $I^2 = \{n^1 + 1, ..., n^1 + n^2\}$, ..., $I^A = \{n^+ - n^a + 1, n^+ - n^a + 2, ..., n^+\}$. In the construction of ASM, averaging over the n^+ firms of each country allows for comparison between countries, even of quite different size.

4.3 Kernel method

The kernel method is a non-parametric method used for a density estimation and has been a popular technique for analyzing one and two-dimensional data; see Bowman and Azzalini (1997), Scott (1992), Simonoff (1996), Wand and Jones (1995) for examples. Scott (1992)'s book applies to multivariate density estimation. The nature of the kernel methods is first briefly presented in the framework of a non-parametric density estimation.

Consider a set of observed data points $\mathbf{x}$ assumed to be a sample from an unknown probability density function, say f. Density estimation is the construction of an estimator of the density function f from the observed data. For two-dimensional data, Rosemblatt (1956) has proposed the following kernel estimator for an unknown density f:

$$\widehat{f}(\mathbf{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{4.3}$$

where h is a so-called smoothing parameter or bandwidth and the function K, called a “kernel”, is a known function defined for two-dimensional $\mathbf{x}$, satisfying

$$\int\limits_{\mathbb{R}^2} K(\mathbf{x})d\mathbf{x} = 1$$

Usually K will be a radially symmetric unimodal probability density function, for example the standard bivariate normal density function

$$K(\mathbf{x}) = (2\pi)^{-1} \exp(-\tfrac{1}{2}\mathbf{x}^t\mathbf{x})$$

Thus, the kernel estimator depends on two parameters: the bandwidth h and the kernel density K. It is generally considered that the density kernel estimator is robust with respect to kernel choices; this eventually justifies the usual choice of a Gaussian kernel (for details, see Silverman 1992).

A natural use of density estimation is a description of some properties of a given set of data: density estimation may indeed give valuable indications on such features as skewness or multimodality (i.e. the presence of several local maxima in the density) in the data. In this application, the technique of density estimation is particulary suitable to detect local specialization.

Thus, an estimator of local measurement of specialization of point $\mathbf{x}$ in the activity a with bandwidth h_1 and h_2 is the following

$$\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x}) = \frac{\frac{\widehat{\lambda}^{a}_{h_1}(\mathbf{x})}{n^a}}{\frac{\widehat{\lambda}^{+}_{h_2}(\mathbf{x})}{n^+}} \tag{4.4}$$

$$= \frac{\widehat{f}^{a}_{h_1}(\mathbf{x})}{\widehat{f}^{+}_{h_2}(\mathbf{x})} \tag{4.5}$$

$$= \frac{\frac{1}{n^a h_1^2}\sum\limits_{i\in I^a} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_1}\right)}{\frac{1}{n^+ h_2^2}\sum\limits_{i\in I^+} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_2}\right)} \tag{4.6}$$

where I^+ is the set of all indexes of firms, and $\widehat{\lambda}^{a}_{h_1}(\mathbf{x})/n^a$ and $\widehat{\lambda}^{+}_{h_2}(\mathbf{x})/n^+$ are a kernel local density estimators with bandwidth selection h_1 and h_2 of firms in the activity a and of the aggregated process, respectively.

As the matter of fact, in this Chapter we generalize to a two-dimensional space a suggestion of Flahaut, Mouchart, San Martin and Thomas (2003) that introduced the idea of estimating the intensity of a non-homogeneous Poisson process on the real line through a technique of nonparametric density estimation after normalizing the intensity into a probability density.

Thus $\widehat{f}^{a}_{h_1}(\mathbf{x})$ tells how far companies dedicated to activity a tend to concentrate around point $\mathbf{x}$ whereas $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ tells whether this concentration around point $\mathbf{x}$ is larger than for all activities aggregated. This is precisely the issue of specialization.

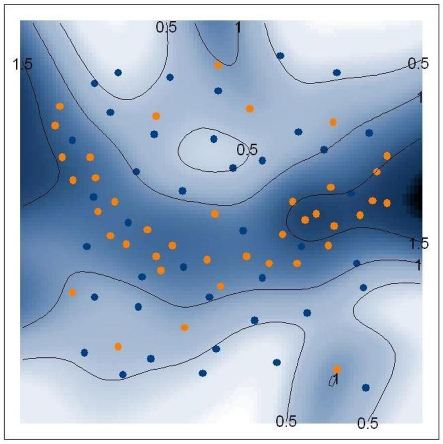
The following Fig. 4.2 illustrates a specialization measurement (using continuous $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ with spherical normal kernels with $h_1 = 0.15$ and $h_2 = 0.13$) for the above Fig. 4.1 (location of 80 firms, amongst which 40 are of orange activity, whilst the remaining 40 blue points are the firms of rest of activities).

It can be observed that the specialization pattern for the orange activity has a "U" shape in the central layer of the territory.

4.4 Bandwidth selection

As noticed in Flahaut, Mouchart, San Martin and Thomas (2003),"The kernel estimator depends on two parameters: the bandwidth h and the kernel density K.

Figure 4.2: *Specialization measurement*



It may be shown that the density kernel estimator is generally robust with respect to kernel choices; this eventually justifies the usual choice of a Gaussian kernel (for details, see Silverman (1986, Chapter 3). For a given kernel K, the kernel estimator critically depends on the choice of the smoothing parameter h. An appropriate choice of the smoothing parameter should be determined by the purpose of the estimate. Silverman (1986, Section 3.4.1) suggests a subjective choice of the smoothing parameter if the purpose of the estimation is to explore the data in order to propose possible statistical models and hypotheses.

In addition, he suggests an automatic choice of the smoothing parameter, which may be considered as a starting point for subsequent subjective adjustments (Silverman, 1986, p. 44). Indeed, an optimal smoothing parameter h_{opt} may be obtained by minimizing the approximate integrated mean square error; such an optimal bandwidth is proportional to $n^{-1/5}$, where n is the sample size. The constant of proportionality depends on the unknown density f; for computing it, iterative methods are typically used (see Silverman, 1986, p. 40). The initial iteration often makes use of a reference bandwidth h_{ref}, defined by both the kernel K and the unknown density f; when f is Gaussian with variance σ^2 the reference bandwidth is obtained by:

$$h_{ref} = 1.06\ \sigma n^{-1/5} \tag{4.7}$$

(end of citation).

An optimal bandwidth requires a performance criterion, a common choice of which is the Kullback-Leibler information

$$I(f,\widehat{f}) = \int\limits_{\mathbf{x}\in M} f(\mathbf{x}) \log \left(f(\mathbf{x})/\widehat{f}(\mathbf{x}) \right) \, d\mathbf{x} \tag{4.8}$$

which is a divergence between f and $\widehat{f}$. An alternative may be the integrated squared error

$$ISE = \int\limits_{\mathbf{x}\in M} \left(\widehat{f}(\mathbf{x}) - f(\mathbf{x}) \right)^2 \, d\mathbf{x} \tag{4.9}$$

The mean integral squared error (MISE) criteria is accordingly defined as:

$$MISE(h) = E\left(\int\limits_{\mathbf{x}\in M} \left(\widehat{f}_h(\mathbf{x}) - f_h(\mathbf{x}) \right)^2 \, d\mathbf{x} \right) \tag{4.10}$$

The asymptotic form of h which minimizes the mean integrated squared error is given by

$$h = \left[\frac{\int\limits_{\mathbf{x}\in M} K^2(\mathbf{x}) \, d\mathbf{x}}{n(\int\limits_{\mathbf{x}\in M} K(\mathbf{x})\mathbf{x}^2 \, d\mathbf{x})^2 \int\limits_{\mathbf{x}\in M} (f''(\mathbf{x}))^2 \, d\mathbf{x}} \right]^{1/5} \tag{4.11}$$

In this work we will adopt a data-based bandwidth selection approach. One alternative for h_1 and h_2 bandwidth selection is to use cross-validation of least square, according to Rudemo (1982) and Bowman's (1984) suggestions. However, as it will be noted in Section 4.6 we will make a bootstrap, we opted for bandwidths through smoothing bootstrap, following Taylor (1989).

The bootstrap version of (4.10) is:

$$MISE^*(h) = E^*\left(\int\limits_{\mathbf{x}\in M} \left(\widehat{f}^*_h(\mathbf{x}) - \widehat{f}_h(\mathbf{x}) \right)^2 \, d\mathbf{x} \right) \tag{4.12}$$

where $\widehat{f}_h^*(\mathbf{x})$ is the kernel estimate using resampled $\mathbf{x}_i^* \sim \widehat{f}_h(\mathbf{x})$. Realizations of $\mathbf{x}^*$ could be generated as follows; see, for example Silverman (1992):

- choose an integer I with equal probability from $\{1, \ldots, n\}$;
- generate a random variable $\phi \sim K(\mathbf{x})$;
- set $\mathbf{x}^* = \mathbf{x}_I + h\phi$.

However, for the Gaussian kernel, we shall see that there is no need for this approach as the bootstrap mean can be calculated without resampling. By minimizing $MISE^*(h)$ we minimize an estimate of mean integrated squared error, and hence, get an h close to the minimizing integrated squared error. And obvious concern al this stage of using $\widehat{f}_h^*(\mathbf{x})$ is that each application of the smooth bootstrap inflates the variance. Similarly, for Kullback-Leibler information we would minimize

$$E^* \left[\int\limits_{\mathbf{x} \in M} \widehat{f}_h(\mathbf{x}) \log \widehat{f}_h(\mathbf{x}) / \widehat{f}_h^*(\mathbf{x}) \; d\mathbf{x} \right] \tag{4.13}$$

where again the expectation E^* is taken with respect to the distribution of $\mathbf{x}^*$.

4.5 Problem related to the use of normal kernels

4.5.1 Problems on the boundaries

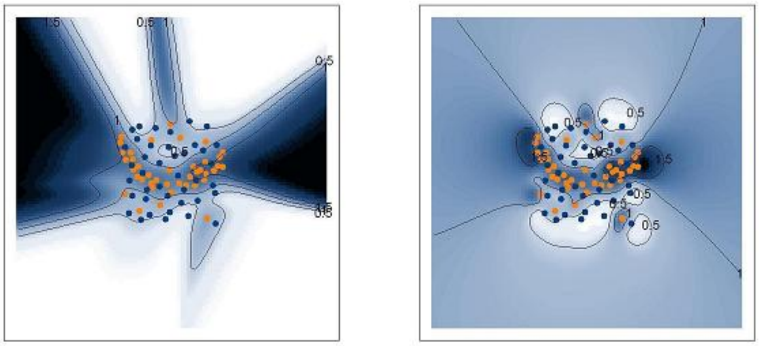
One of the issues raised by the use of normal kernels is that for regions with low density of firms, the ratio of estimated densities in $\widehat{LQ}^{a}_{h_1 h_2}(\mathbf{x})$ involves very small values for both densities (numerator and denominator). If the numerator is greater than the denominator, the ratio will tend to reciprocate the proportion of firms in the activity a of interest, namely n^+/n^a. If the denominator is greater than the numerator, the ratio will tend to zero. This is due to the thin tails typical of the normal distribution.

The next two figures display a local measurement of specialization based on a zoom-out of the Fig. 4.2. In this illustration, the orange-colored points represent firms of the activity of interest, whereas the blue-colored points represent the other activities.

The left-hand graph of Fig. 4.3 is based on normal kernels. The regions far from the center and bordered, toward the center, by blue-colored points become white regions, illustrating the fact that the measurement of specialization tends to zero. In contrast, regions far from the center and bordered, toward the center, by orange-colored points become dark-colored regions, reflecting the fact that the local measurement of specialization tends to n^+/n^a. Next subsection gives a more precise statement of this phenomenon.

The right-hand graph of Fig. 4.3 shows a local measurement of specialization with Cauchy kernels. All the zones that are far from the center show a moderate color intensity that corresponds to the 1-value of the limit in the local measurement of specialization.

Figure 4.3: *Boundary problem of kernel estimator: local measurement of specialization with normal kernels (left-hand) and with Cauchy kernels (right-hand)*



This behavioral difference in the continuous $\widehat{LQ}^a_{h_1h_2}(\mathbf{x})$ that can be observed when using Cauchy kernels instead of normal kernels derives from the following results:

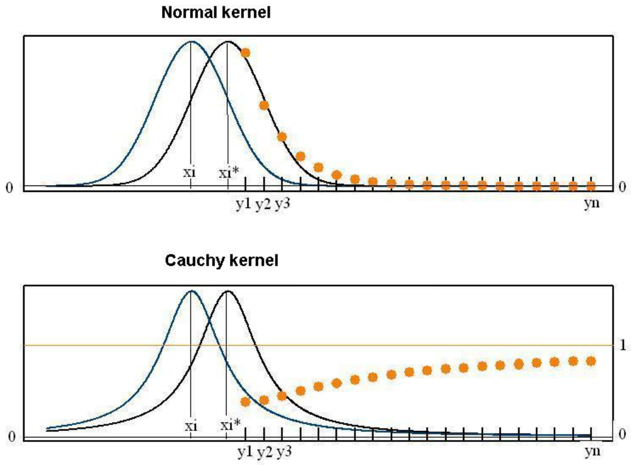
$$\lim_{\mathbf{x}\to\infty} \tfrac{f(\mathbf{x})}{f(\mathbf{x}-\Delta)} = 0 \text{ if } f \text{ is a standard normal density}$$

$$\lim_{\mathbf{x}\to\infty} \tfrac{f(\mathbf{x})}{f(\mathbf{x}-\Delta)} = 1 \text{ if } f \text{ is a standard Cauchy density}$$

when $\Delta > 0$

The upper diagram in the following Fig. 4.4 pair shows the convergence to zero of the density ratio for the normal case, while the lower diagram shows the convergence to one of the same ratio for the Cauchy case.

Figure 4.4: *Convergence to zero of the density ratio: normal case (upper diagram) and Cauchy case (lower diagram)*



The problem of boundary effects has long been recognized, starting with Gasser and Müller (1979), while more recent research of methods to handle boundary effects for kernel function estimates include those by Marron and Ruppert (1994), Jones and Foster (1996), Müller and Stadtmüller (1999), and more recently Hazelton and Marshall (2008) for bivariate density estimation. The reason for the boundary problem in the kernel estimator is that, at a boundary point, the kernel mass falls outside the support of the function to be estimated, and is therefore lost. We can consider that this problem is caused by a discontinuity in the function to be estimated across the boundary.

Although we believe it is important to show how choosing the K parameter may affect the estimation of the specialization level at the support ends, it still remains a minor problem vis-à-vis the selection of the h parameter. As we will see in the examples included at the end of this Chapter, the bootstrap method proposed in section 4.6 eliminates or correctly isolates this disadvantage to identify the statistical significance of the specialization level, even using normal kernels.

4.5.2 Formal exposition

We define the Zone of Influence of the ϵ margin for the i firm (Z_i^ϵ) as the region in the map formed by the points for which the i firm is at least less ϵ closer than any other firm, more formally

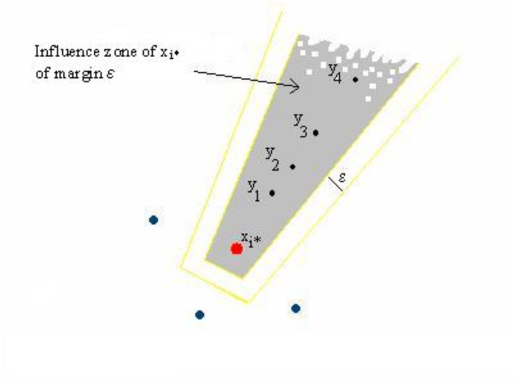
$$Z_i^\epsilon = \{\mathbf{x} \in \mathbb{R}^2 : d(\mathbf{x}_k, \mathbf{x}) < d(\mathbf{x}_i, \mathbf{x}) - \epsilon, \forall k \in I^+, k \neq i\} \tag{4.14}$$

where $d(\cdot,\cdot)$ is a Euclidean distance between points of $\mathbb{R}^2$.

These regions will be bounded for the innermost points, while they will generally be unbounded for the outermost points. It can easily be observed that if we consider $\epsilon = 0$, the zones of influence defined will match the well-known Voronoi cells (see for more de Berg et al. 1997).

This definition will be useful to create a sequence of divergent points in the zone of influence of a firm of the activity a, and calculate the LQ limit applied to such sequence. Next Fig. 4.5 shows the influence zone of x_{i^*} point of margin ϵ.

Figure 4.5: *Influence zone of x_{i^*} point of margin ϵ*



Theorem 1 (zones of high intensity).*For the unbounded zones of influence corresponding to the firms of the activity a, as we move away from such firms, the specialization measure (using a normal spherical kernel) will tend to reciprocate de proportion of firms with such activity n^+/n^a.*

In formal terms, given $\mathbf{x}_1, ..., \mathbf{x}_n$, $\epsilon > 0$, any firm i^* of the activity a ($i^* \in I^a$), $\mathbf{y}_n \in Z^{\epsilon}_{i^*} \subset \mathbb{R}^2$ a sequence of points in the map that belong to the zone of influence of margin ϵ of the firm i^* such that

$$d(\mathbf{y}_n, \mathbf{x}_{i^*}) < d(\mathbf{y}_{n+1}, \mathbf{x}_{i^*})$$

$$d(\mathbf{y}_n, \mathbf{x}_{i^*}) \overset{n\to\infty}{\longrightarrow} \infty$$

i.e. monotonically increasing sequence of distance, therefore

$$\widehat{LQ}^{a}_{h}(\mathbf{y}_n) \overset{n\to\infty}{\longrightarrow} \tfrac{n^+}{n^a}$$

Demonstration:

$$\begin{aligned} \widehat{LQ}^{a}_{h}(\mathbf{y}_n) &= \frac{\frac{\widehat{\lambda}^a_h(\mathbf{y}_n)}{n^a}}{\frac{\widehat{\lambda}^+_h(\mathbf{y}_n)}{n^+}} && (4.15) \\ &= \frac{\frac{1}{n^a h^2} \sum\limits_{i \in I^a} K\left(\frac{\mathbf{y}_n - \mathbf{x}_i}{h}\right)}{\frac{1}{n^+ h^2} \sum\limits_{i \in I^+} K\left(\frac{\mathbf{y}_n - \mathbf{x}_i}{h}\right)} && (4.16) \end{aligned}$$

If we divide the numerator and the denominator by $K(\frac{\mathbf{y}_n - \mathbf{x}_{i^*}}{h})$

$$\frac{\frac{1}{n^a h^2}\left(1 + \sum\limits_{i \in I^a, i \neq i^*} \frac{K(\frac{\mathbf{y}_n - \mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n - \mathbf{x}_{i^*}}{h})}\right)}{\frac{1}{n^+ h^2}\left(1 + \sum\limits_{i \in I^+, i \neq i^*} \frac{K(\frac{\mathbf{y}_n - \mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n - \mathbf{x}_{i^*}}{h})}\right)} \tag{4.17}$$

we should see that

$$\sum_{i \in I^+, i \neq i^*} \frac{K(\frac{\mathbf{y}_n - \mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n - \mathbf{x}_{i^*}}{h})} \overset{n\to\infty}{\longrightarrow} 0 \tag{4.18}$$

and consequently, also

$$\sum_{i\in I^a, i\neq i^*} \frac{K(\frac{\mathbf{y}_n-\mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} \overset{n\to\infty}{\longrightarrow} 0 \tag{4.19}$$

And since $I^a \subset I$

$$0 \leq \sum_{i\in I^a, i\neq i^*} \frac{K(\frac{\mathbf{y}_n-\mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} \leq \sum_{i\in I^+, i\neq i^*} \frac{K(\frac{\mathbf{y}_n-\mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} \tag{4.20}$$

it results in

$$\widehat{LQ}_h^a(\mathbf{y}_n) \overset{n\to\infty}{\longrightarrow} \frac{n^+}{n^a} \tag{4.21}$$

To demonstrate (4.18) we must note that $\forall i \in I^+, i \neq i^*$ and since $\mathbf{y}_n \in Z_{i^*}^{\epsilon}$ then

$$\begin{align} \frac{K(\frac{\mathbf{y}_n-\mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} &< \frac{K(\frac{(\mathbf{y}_n-\mathbf{x}_{i^*})(1+\epsilon)}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} \tag{4.22}\\ &= \frac{f(\frac{\|\mathbf{y}_n-\mathbf{x}_{i^*}\|(1+\epsilon)}{h})}{f(\frac{\|\mathbf{y}_n-\mathbf{x}_{i^*}\|}{h})} \tag{4.23}\\ &= \frac{\frac{e^{-\frac{1}{2}\|\mathbf{y}_n-\mathbf{x}_{i^*}\|^2(1+\epsilon)^2}}{h^2}}{\frac{e^{-\frac{1}{2}\|\mathbf{y}_n-\mathbf{x}_{i^*}\|^2}}{h^2}} \tag{4.24}\\ &= e^{-\frac{1}{2}\|\mathbf{y}_n-\mathbf{x}_{i^*}\|^2\epsilon^2} \tag{4.25} \end{align}$$

where f is the univariate normal density function. Hence

$$\sum_{i\in I^+, i\neq i^*} \frac{K(\frac{\mathbf{y}_n-\mathbf{x}_i}{h})}{K(\frac{\mathbf{y}_n-\mathbf{x}_{i^*}}{h})} < (n^+-1)\; e^{-\frac{1}{2}\|\mathbf{y}_n-\mathbf{x}_{i^*}\|^2\epsilon^2} \tag{4.26}$$

and since based on the hypothesis $d(\mathbf{y_n}, \mathbf{x_{i^*}}) \overset{n\to\infty}{\longrightarrow} \infty$, given ϵ, there is an n that makes $(n^+-1)\; e^{-\frac{1}{2}\|\mathbf{y}_n-\mathbf{x}_{i^*}\|^2\epsilon^2}$ as small as desired.

4.6 Identification of specialized agglomeration

4.6.1 Introducing the methodology

The identification of specialized agglomerations in continuous space has a close relationship with the identification of statistical significance or significant feature of the specialization level. In the analysis of three or higher-dimensional data the discovery of significant features may have more interest than the estimation of the whole data density. Feature significance is an extension of kernel density estimation which is used to establish the statistical significance of features (e.g. local modes). For one- and two-dimensional data Chaudhuri and Marron (1999) and Godtliebsen et al. (2002) looked for as significant features from different perpectives: local extrema, valleys, ridges, saddle points and steep gradients. They developed techniques for determining and visualizing these features. Chaudhuri and Marron (2000) and Hannig and Marron (2006) produced asymptotic distributional results for the one-dimensional case. As the number of dimensions increases, local maxima become the single most important features, and identification of these maxima is the goal.

In feature significance we focus on a range of bandwidths, rather than on a "best" bandwidth selection. For one-dimensional data this approach leads to the "Sizer" plots of Chaudhuri and Marron (1999). For bivariate data Godtliebsen et al. (2002) employ diagonal bandwidth matrices with the same bandwidth for both dimensions, thus reducing the two-dimensional problem to one dimension. For the single diagonal bandwidth one can then proceed as in the one-dimensional "Sizer" case.

In contrast, we propose a bootstrap methodology to evaluate the significance of the local specialization following the method of bootstrap hypothesis testing proposed by Efron and Tibshirani (1993), and Davison and Hinkley (1997). Although there is a very large literature on bootstrapping in statistic, a surprisingly small proportion of it is devoted to bootstrap testing. Instead, the focus is usually on estimating bootstrap standard errors and constructing bootstrap confidence intervals. The basic idea of any sort of hypothesis test is to compare the observed value of a test statistic with the distribution that it would follow if the null hypothesis were true. The method of bootstrap hypothesis testing generates a large number of simulated values of the test statistic and compare it with the empirical distribution function of the simulated ones. Recently, Reiczigel et al. (2005) uses this approach to test the stochastic equality of two populations and Mackinnon (2007) proposed a bootstrap and Monte Carlo methods for testing hypothesis in econometrics (test for structural change with an unknown break point).

Although the values of the $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ function greater than 1 suggest high specialization levels for the activity a, we should still find a measure of the empirical evidence to support this value, i.e. the statistical signification of the specialization. To determine the significance of the estimated specialization levels, we calculate for each point $\mathbf{x}$ of interest the distribution of the values for the $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ function, based on the non-specialization (of the activity a) hypothesis, i.e. based on the hypothesis: $\lambda^a(\mathbf{x})/n^a \leq \lambda^+(\mathbf{x})/n^+$, and consider as significant specialization points those points $\mathbf{x}$ with a value of $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ significantly higher than 1 or significantly *over-specialized* (in the sense of Chapter 2 and 3), relative to the distribution under the non-specialization (of the activity a) hypothesis. The steps are:

1. bandwidth selection of the activity a through smoothing bootstrap (h_1);
2. bandwidth selection of the total activity through smoothing bootstrap (h_2);
3. using the estimated density of the total activity, $\widehat{\lambda}^+_{h_2}(\mathbf{x})/n^+$, we create B samples of size n^a;
4. for each bootstrap sample of step 3, we estimate the $\widehat{\lambda}^{a*}_{h_1}(\mathbf{x})/n^a$ density (with the h_1 established in step 1), and together with $\widehat{\lambda}^+_{h_2}(\mathbf{x})/n^+$ density, we calculate the $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$ function for all the points of interest:

$$\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x}) = \frac{\frac{\widehat{\lambda}^{a*}_{h_1}(\mathbf{x})}{n^a}}{\frac{\widehat{\lambda}^{+}_{h_2}(\mathbf{x})}{n^+}} \tag{4.27}$$

Therefore, for each $\mathbf{x}$ point of interest, we have:

- the value of the original $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ function;
- B Monte Carlo replications of the $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$ function.

4.6.2 Identifying high-specialization regions

The significance of $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ is accordingly evaluated thought the bootstrap distribution of $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$. A point $\mathbf{x}$ in the map is detected as a high specialization (of the activity a) point, inasmuch as it meets

$$\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x}) > 1 \tag{4.28}$$

and

$$\widehat{P}^{*}\left(\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x}) > \widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) < \alpha \tag{4.29}$$

where

$$\widehat{P}^{*}\left(\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x}) > \widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) = 1 - \widehat{F}^{*}\left(\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) = \frac{1}{B}\sum_{j=1}^{B} I\left(\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x}) > \widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) \tag{4.30}$$

where $\widehat{F}^{*}$ denotes the bootstrap distribution function of the $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$, B is the number of the bootstrap samples, or simulated data sets, indexed by j, and $I(\cdot)$ denotes the indicator function, which is equal to 1 when its argument is true and 0 otherwise. Thus the bootstrap P-value is, in general, simply the proportion of the bootstrap test statistics $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$ that are more extreme than the observed test statistic $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$. Of course, rejecting the null hypothesis whenever $\widehat{P}^{*}\left(\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) < \alpha$ is equivalent to rejecting it whenever $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ exceeds the 1-α quantile of $\widehat{F}^{*}$.

Perhaps surprisingly, this procedure can actually yield an exact test in certain cases. The key requirement is that the test statistic $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ should be pivotal, which means that its distribution does not depend on anything that is unknown. This implies that $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ and the $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$ all follow the same distribution if the null is true. In addition, the number of bootstrap samples B must be such that $\alpha(B+1)$ is an integer, where α is the level of the test. If a bootstrap test satisfies these two conditions, then it is exact. This sort of test, which was originally proposed in Dwass (1957), is generally called a Monte Carlo test. For an introduction to Monte Carlo testing, see Dufour and Khalaf (2001) and Diggle (2003).

It is quite easy to see why Monte Carlo tests are exact. Imagine sorting all $B+1$ test statistics. Then rejecting the null whenever $\widehat{P}^{*}\left(\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x}) > \widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})\right) < \alpha$

implies rejecting it whenever $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ is one of the largest $\alpha(B+1)$ statistics. But, if $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ and the $\widehat{LQ}^{a*}_{h_1h_2}(\mathbf{x})$ all follow the same distribution, this happens with probability precisely α. For example, if B=999 and α=0.01, we reject the null whenever $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ is one of the 10 largest test statistics.

Since a Monte Carlo test is exact whenever $\alpha(B+1)$ is an integer, it is tempting to make B very small. In principle, it could be as small as 19 for α=0.05 and as small as 99 for α=0.01. There are two problems with this, however. The first problem is that the smaller is B the less powerful is the test. The loss of power is proportional to $1/B$; see Jöckel (1986) and Davidson and MacKinnon (2000).

The second problem is that, when B is small, the results of the test can depend nontrivially on the particular sequence of random numbers used to generate the bootstrap test statistics. Since $\widehat{P}^{*}$ is just a frequency, the standard error of $\widehat{P}^{*}$ is $P^{*}(1-P^{*})/B$. Thus, when P^{*}=0.05, the standard error of $\widehat{P}^{*}$ is 0.0219 for B=99, 0.0069 for B=999, and 0.0022 for B=9,999. This suggests that it might be dangerous to use a value of B less than 999 and it would not be unreasonable to use B=9,999.

As a conclusion, in this chapter we define a specialization measurement for a point $\mathbf{x}$. This definition is an extension of the well-known LQ measurement to the continuous space. To quantify the degree of significance of the specialization measurement, we used the bootstrap method to approximate the distribution of the function LQ under the non-specialization hypothesis. As a result, we define a specific point in the map $\mathbf{x}$ as specialized, if the LQ value for that point is larger than a pixel quintile of the bootstrap distribution for the same point.

4.7 Graphic representation of the specialized agglomerations

This section shows the results of implementing the methodology proposed in the previous section with the following changes: for simplicity and calculation speed, i) we used the asymptotic method for the bandwidth selection of h_1 and h_2; and ii) we created an equally spaced rectangular grid of points in M map, with size $n^{+} \times n^{+}$, to reduce the number of points, and consequently, the number of tests of significance (see for more Diggle 2003). Thus, for each of the $\left(n^{+}\right)^{2}$ cells we compute the estimated value $\widehat{LQ}^{a}_{h_1h_2}(\mathbf{x})$ as in (4.28) and the significance level as in $\widehat{P}^{*}$-value (4.30), and eventually construct an histogram of $\widehat{P}^{*}$-values. The B number of Monte

Carlo replications for the null hypothesis bootstrap is equal to 1,000, and the critical level α is equal to 0.05.

Table 4.1 shows the schemes of simulations for seven examples, where $I_{(2)}$ denotes the 2×2 identity matrix.

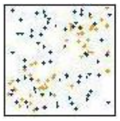
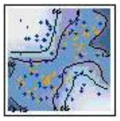
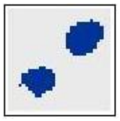
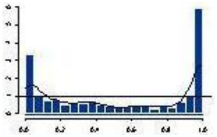
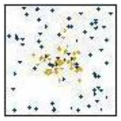
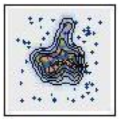
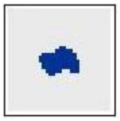
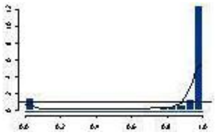
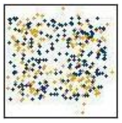
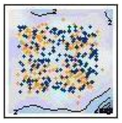
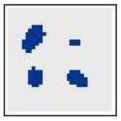
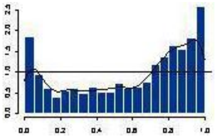
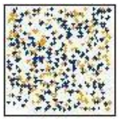
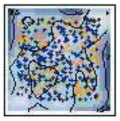
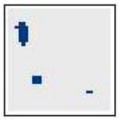
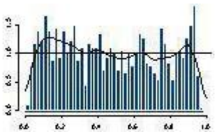
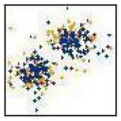
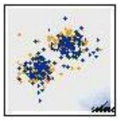
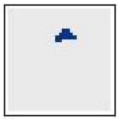
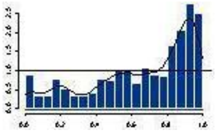
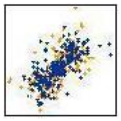
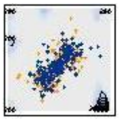
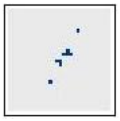
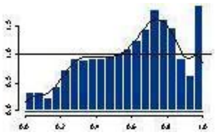
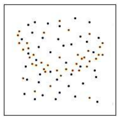
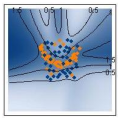
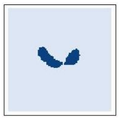
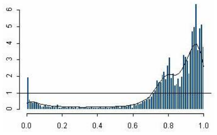
For each example, Table 4.2 gives the following diagrams: i) the locations of the simulated points in the M map (first graph); ii) the specialization level (second graph); iii) the significant specialized regions (third graph), in which the zones with a high specialization level are viewed as contiguous arrays of blue points; and iv) the p-values histogram (fourth graph), in which the solid line shows the level of non-specialization hypothesis.

These graphs provide some hint on the actual working of the proposed methodology. The first column "Simulated points" make clear the difference among the 7 examples. In particular, examples 5 and 6 display substantial clustering whereas examples 1 to 4 and 7 provide locations more spread on the map. The second column "Specialization level" is based on the values of the estimated local quotient $\widehat{LQ}^{a}_{h_1 h_2}(\mathbf{x})$. Two features should be noticed. These graphs illustrate clearly the connection between the locations (first column) and the value of the local quotient. Secondly, the second column also illustrate the frontier problem typical of the kernel estimators. The third column "Significant specialized regions" also deserves two remarks. Firstly, checking for significance of the value of the local quotient, through bootstrapping, provide a more synthetic view than the second column: even though the level alpha is arbitrary, the resulting view better corresponds to the needs of spatial analysis by explicitly bordering specialized regions. Secondly, bootstrapping naturally eliminates the frontier problems. Finally, the last column "P-values histogram" provide a synthetic view of the method through two interesting features. Firstly, the different pattern of each example results in different pattern of clusterization, i.e. the proportion of significant cells is clearly different through the 7 examples. Secondly, the method illustrates explicitly a compensation effect, i.e. that to over-specialized regions correspond, by logical necessity, sub-specialized regions even though this compensation effect is different among the 7 examples.

Table 4.1: *Schemes of simulated random points*

Example	n^+	Blue points $(n^+ - n^a)$	Orange points (n^a)	Bandwidths	
				h_1	h_2
1	130	$70 : \mathcal{U}(0,1)^2$	$30 : \mathcal{N}\left(\begin{bmatrix}0.25\\0.25\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$ $30 : \mathcal{N}\left(\begin{bmatrix}0.75\\0.75\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$	0.13	0.11
2	120	$70 : \mathcal{U}(0,1)^2$	$50 : \mathcal{N}\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$	0.06	0.09
3	400	$200 : \mathcal{U}(0,1)^2$	$50 : \mathcal{N}\left(\begin{bmatrix}0.25\\0.25\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$ $50 : \mathcal{N}\left(\begin{bmatrix}0.25\\0.75\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$ $50 : \mathcal{N}\left(\begin{bmatrix}0.75\\0.25\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$ $50 : \mathcal{N}\left(\begin{bmatrix}0.75\\0.75\end{bmatrix}, 0.0156\ \mathrm{I}_{(2)}\right)$	0.10	0.09
4	600	$300 : \mathcal{U}(0,1)^2$	$300 : \mathcal{U}(0,1)^2$	0.10	0.09
5	400	$100 : t\left(\begin{bmatrix}2\\2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$ $100 : t\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$	$100 : t\left(\begin{bmatrix}2\\2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$ $100 : t\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$	0.53	0.48
6	600	$100 : \mathcal{U}(-1,1)^2$ $100 : t\left(\begin{bmatrix}2\\2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$ $100 : t\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$	$100 : \mathcal{U}(-1,1)^2$ $100 : t\left(\begin{bmatrix}2\\2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$ $100 : t\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \mathrm{I}_{(2)}, 5\right)$	0.66	0.59
7	80	40: exactly those of fig. 4.1	40: exactly those of fig. 4.1	0.15	0.13

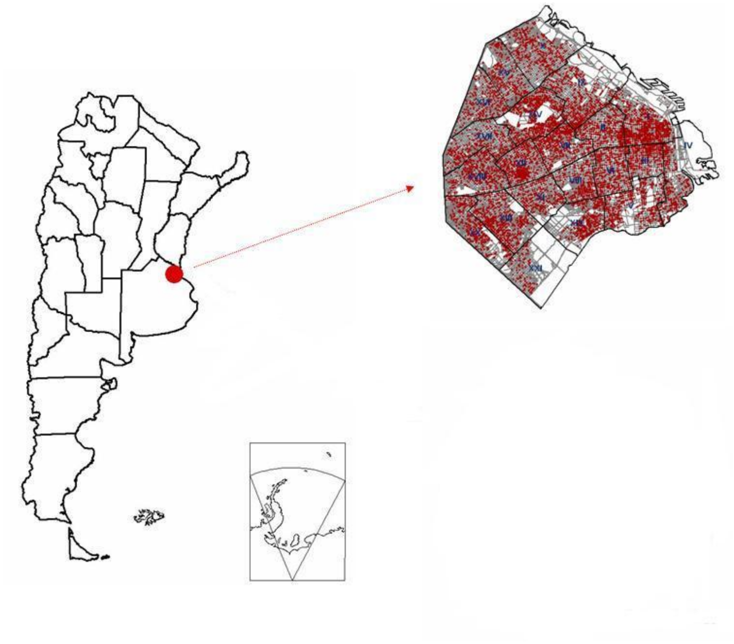
Table 4.2: *Graphic representation of the specialized agglomerations*

Example	Simulated points	Specialization level	Significant specialized regions	P-values histogram
1				
2				
3				
4				
5				
6				
7				

4.8 Application: Manufacture sector data of Buenos Aires City

For this application, we have geocodified 10,657 firms of the manufacturing sector in Buenos Aires City (Fig. 4.6). First, we had to homogenize the name of the streets declared by the firms on the basis of the name of the georeferenced streets map of Buenos Aires City, according to a record linkage algorithm developed by the Research Center of the Università di Bologna at Buenos Aires. Finally, the firms were geocodified using the ArcGis software.

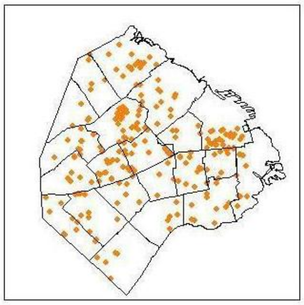
Figure 4.6: *Locations of 10,657 geocodified firms of the manufacturing sector in Buenos Aires City*



4.8.1 Specialized agglomerations of pharmaceutical and medical equipment sectors

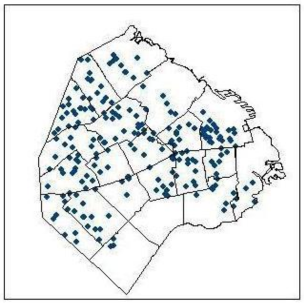
The following Fig. 4.7 shows the locations of 248 geocodified firms of the Pharmaceutical sector (class ISIC Rev. 3: 2423)

Figure 4.7: *Locations of 248 geocodified firms of the Pharmaceutical sector in Buenos Aires City*



Next Fig. 4.8 shows the locations of 223 geocodified firms of the Medical equipment sector (class ISIC Rev. 3: 3311)

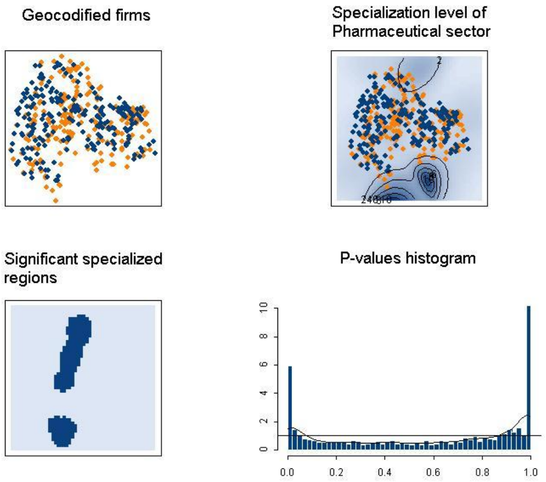
Figure 4.8: *Locations of 223 geocodified firms of the Medical equipment sector in Buenos Aires City*



In the same way as that of the above examples, the next figures shows: i) the locations in the M map in which each $\mathbf{x}$ point of interest (n^a) is assigned an orange color, while the rest of points $(n^+ - n^a)$ are in blue (upper left corner); ii) the specialization level (upper right corner); iii) the significant specialized regions (lower left corner), in which the zones with a high specialization level are viewed as contiguous arrays of blue points; and iv) the p-values histogram (lower right corner), in which the solid line shows the level of non-specialization hypothesis. The B number of Monte Carlo replications for the null hypothesis is equal to 1,000.

For expository purpose, we only consider two sectors (Fig. 4.9): Pharmaceutical, as a sector of interest, and Medical equipment, deemed to represent "the other sectors". The critical level α is equal to 0.01 and the reference bandwidths are (in meters): $h_1 = 1312.50$ and $h_2 = 1220.13$.

Figure 4.9: *Specialized agglomerations of Pharmaceutical sector*



Observing the left-hand graph of significant specialized regions in Fig. 4.10, one may wonder why the region m_1 appears as significant while region m_2 does not? Using the R function "Jittering" for separating points for plotting (this technique add a random noise to separate points with identical values) makes to note that the region m_1 had 11 nearest orange points whereas the region m_2 had 9 and more separate orange points (right-hand graph).

The next Fig. 4.11 shows how far the choice of the critical level alpha affects the identification of specialized agglomerations. For instance, using a critical level α of 0.05, the region B of the above diagram appears now as significant specialized region with a new small specialized region in the downtown.

In Fig. 4.12 the point of interest (orange color) are the Medical equipment firms while the rest of points (blue color) are the Pharmaceutical firms. The critical level α is equal to 0.01. The references bandwidths are (in meters): $h_1 = 1293.13$ and $h_2 = 1220.13$. Notice the difference of specialized agglomerations in Fig. 4.9 and in Fig. 4.12.

Figure 4.10: *Specialized agglomerations of Pharmaceutical sector: original locations (left-hand) and jitter locations (right-hand)*

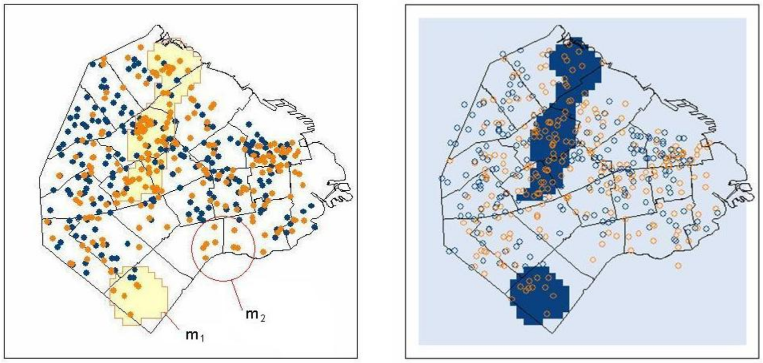
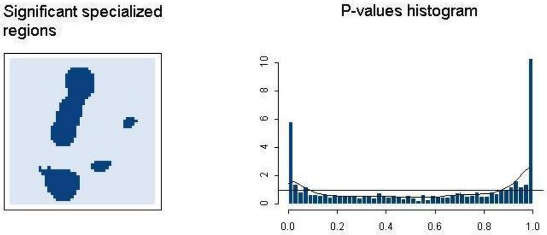


Figure 4.11: *Specialized agglomerations of Pharmaceutical sector with critical level α equal to 0.05*



Finally, in Fig. 4.13 graphs shown de geocodified firms of Pharmaceutical and Medical equipment sectors respectively and the specialized agglomerations at critical level α to 0.01. The significant specialized regions of Pharmaceutical had 68 firms (27%) while the significant specialized regions of Medical equipment had 45 firms (20%). In the sense of Chapter 3, the Pharmaceutical firms had a higher tendency to co-localize in specialized agglomerations to the Medical equipment firms.

Figure 4.12: *Specialized agglomerations of Medical equipment sector*

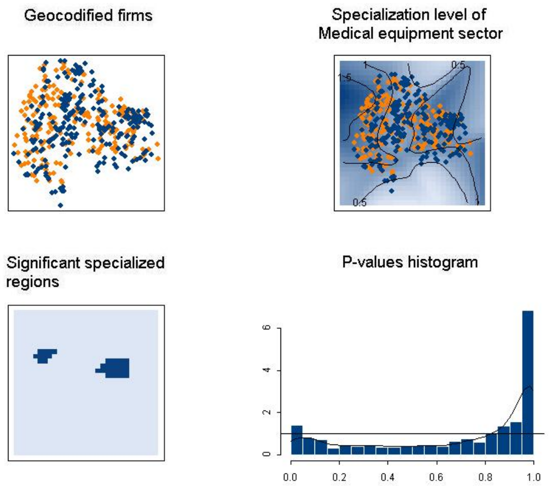
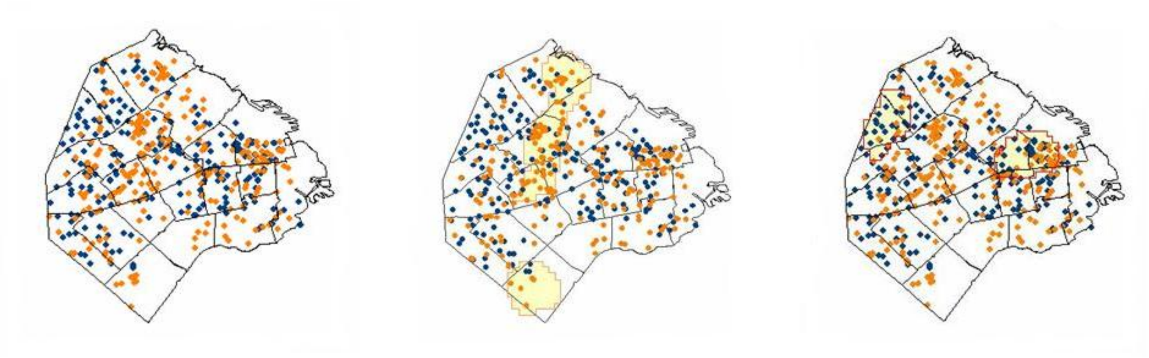


Figure 4.13: *Geocodified firms (left-hand graph), and Specialized agglomerations of Pharmaceutical (center graph) and Medical equipment (right-hand graph) sectors*



Chapter 5

Summary and overall conclusions

5.1 An overview of this thesis

If we observe a map with the location of economic activities, three clearly defined phenomena become readily apparent: concentration, specialization and agglomeration. While Economic Geography studies the underlying causes on the basis of different theoretical models, there are only a few studies that cover the aspects developed in this thesis that refer to the identification of specialized agglomerations, as well as to their global quantification.

In this research two general approaches are put in contrast. Chapters 2 and 3 take a discrete approach with exogenously defined regions, namely administrative entities. The data are provided, for each region, by the number of employees and firms categorized into sectors of activities and the labels of the region are arbitrary. In contrast, Chapter 4 take a continuous approach of a unique universe of reference and the data are provided by the geocodification of the firms with no reference to region.

Our major original contributions to this research were:

- provide a graphical and conceptual explanation of the differences between concentration, specialization, agglomeration itself and specialized agglomerations (Chapter 1);
- analyze specialization in terms of stochastic independence: non specialization is viewed as the case where the joint proportion of employees of region i in activity j is equal to the product of marginal proportions of region i and

activity j; equivalently, the distribution of activities within region i is the same as the global distribution at the country level (Chapter 2);

- the subsequent use of non parametric dependence measures as natural measures for the global specialization level (Chapter 2);
- the development of an automatic grouping procedure of regions and activities based on hierarchical clustering and correspondence analysis (HCCA), defining a goodness of association measure for a given collapsed table, that enabled us to i) significantly reduce the size of the original table and obtain a best collapsed table with low level of information loss vis-à-vis the degree of original specialization (e.g. in the case of Brazil, the number of cells of the original table is reduced by 99%, namely from 113,036 to 884 cells, while the lost specialization information using $d_{\mathcal{X}^2}$ was 23%); and ii) identify the homogeneous regions according to the industrial structure in terms of sub, and over specialization activities in large two-way contingency tables;
- the adaptation of the cluster-identification methods proposed by Besag and Newell (1991) to identify specialized agglomeration in discrete space (Chapter 3);
- the measurement of firm tendency to co-localize in specialized agglomerations according to the industrial activity (Chapter 3 and Chapter 4);
- define a local specialization measurement for a point $\mathbf{x}$ as an extension of the well-known LQ measurement to the continuous space (Chapter 4);
- a possible Average Specialization Measure (ASM) in continuous space (Chapter 4); and
- the relationship between the identification of specialized agglomerations in continuous space and the identification of statistical significance or significant feature of the specialization level based on the method of bootstrap hypothesis testing proposed by Efron and Tibshirani (1993) to approximate the distribution of the LQ under the non-specialization hypothesis (Chapter 4).

The application field of the methods developed in this work is sizable, and may encompass applications for genetic studies, epidemiology sciences, and social and politics topics, to mention just a few.

5.2 Some possible extentions

The future methodological challenges of this one probably more related to the topics put forth in Chapters 2 and 4. For example, to include certain restrictions for the region and activity grouping method of Chapter 2, related to the distance among regions and to the activities connection observed in the input-output matrix, respectively.

A drawback of the identification of the specialized agglomeration method in continuous space (Chapter 4) is that for a given $n^+ \times n^+$ grid, we are achieving a $(n^+)^2$ hypothesis test. Hence, the contiguous arrays of blue points in the non-significant specialized regions occur as a result of the lack of a global significance level, i.e. in the $(n^+)^2$ hypothesis test it is expected that certain number of test results are significant when they are not. More formally, the false discovery rate (FDR) is the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses. A simple Bonferroni correction or the Benjamini and Hochberg (2005) method are widely accepted and commonly used. These methods can provide corrected p-values or estimates of the FDR for a given threshold p-value. Storey and Tibshirani (2003) introduces a measure of statistical significance called the q-value associated with each tested feature in addition to the traditional p-value. Their approach avoids a flood of false positive results. Perone Pacifico et al. (2004) extends FDR to random fields, for which there are uncountably many hypothesis tests. They develop a method for finding regions in the fields domain where there is a significant signal while controlling either the proportion of area or the proportion of clusters in which false rejections occur. The method produces confidence envelopes for the proportion of false discoveries as a function of the rejection threshold. Moreover, powerful resampling approaches (like that used to estimate the FDR in this correspondence) are precise and increasingly common with the rise in available computing power. Regardless of the method used, the objective is to determine a statistical cutoff that results in a reasonable number of false positives. An acceptable adjusted p-value or FDR is somewhat arbitrary, but for the latter metric a value lower than $10 - 20\%$ is commonly cited.

To reduce the number of tests, one option is to implement the binning method, also called “nearest neighbor binning” (see Wand and Jones 1995), moving each data point to the grid point that is its nearest neighbor. Then the mapped points are counted to give a matrix of bin counts.

Following Wang et al. (2006), an alternative for the kernel estimator of specialization measurement is a nearest neighbor approach to estimate divergence between the density of activity a and the density of the whole of activities for each point $\mathbf{x}$ in the map, and to measure the specialization level by the Kullback-Leibler divergence. One advantage of this approach is that it eliminates naturally the boundary problem of the kernel estimator, i.e. when the kernel mass falls outside the support of the function to be estimated caused by a discontinuity in this function across the boundary.

Finally, an alternative to the bootstrap hypothesis testing is to approximate the statistical properties of the statistic under the actual distributions of both activity a and the whole of activities through a re-sampling scheme in order to replace n^a points from the activity a of the original sample and n^+ points now from the original sample of the whole of activities.

References

Adbel-Rahman, H. (2000). City systems: general equilibrium approaches. In J-M. Huriot and J-F. Thisse (eds.). *Economics of Cities. Theoretical Perspectives.* Cambridge: Cambridge University Press.

Aghion, P., and Durlauf, S. (eds.) (2005). *Handbook of Economic Growth.* Amsterdam: Elsevier-North Holland.

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

Ali, M., and Silvey, D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society* **28**: 131-140.

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

Amiti, M. (1999). Specialization patterns in Europe. *Weltwirtschaftliches Archiv* **135**: 1-21.

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.

Anselin, L. (1995). Local indicator of spatial association - LISA. *Geographical Analysis* **27**: 93-115.

Aparicio, F. (1998). Testing independence by nonparametric kernel method. *Statistics and Probability Letters* **34**: 201-210.

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

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

Arbia, G. (2001b). The role of spatial effects in the empirical analysis of regional concentration. *Journal of Geographical Systems* **3**: 271-281.

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.

Ausdretsch, D., and Feldman, M. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review* **86**: 630-640.

Bairoch, P. (1993). *Economics and World History: Myths and Paradoxes.* Chicago: Chicago University Press.

Baldwin R., and Martin, P. (2004). Agglomeration and regional growth. In J.V. Henderson and J-F. Thisse (eds.). *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography.* Amsterdam: Elsevier North-Holland.

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.

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

Becattini, G. (1990). The Marshallian industrial district as a socio-economic notion. In: G. Becattini, F. Pyke, and W. Sengenberger (eds.). *Industrial Districts and Inter-firm Cooperation in Italy.* Geneva: International Institute for Labour Studies.

Becattini, G. (2004). *Industrial Districts: A New Approach to Industrial Change*. Cheltenham: Edward Elgar.

Becattini, G., and Musotti, F. (2003). Measuring the district effect. Reflections on the literature. *Banca Nazionale del Lavoro Quarterly Review* **56**: 259-290.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**: 289-300.

Besag, J. y Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society* **154**: 143-155.

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

Beyene, J., and Moineddin, R. (2005). Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology* **5**: 32.

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

Black, D., and Henderson, J.V. (1999). The theory of urban growth. *Journal of Political Economy* **107**: 252-284.

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

Botham, R., Gibson, H., Martin, R., Miller, P., and Moore, B. (2001). Business clusters in the UK: A first assessment. Report for the Department of Trade and Industry by a consortium led by Trends Business Research.

Brülhart, M., and Sbergamiz, F. (2008). Agglomeration and growth: cross-country evidence. *Micro-Dyn.* Working paper 14.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**: 353-360.

Bowman, A.W., and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford: Oxford University Press.

Brenden, T., Murphy, B., and Wagner, T. (2008). Novel tools for analyzing proportional size distribution index data. *North American Journal of Fisheries Management* **28**: 1233-1242.

Chaudhuri, P., and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**: 807-823.

Chaudhuri, P., and Marron, J. S. (2000). Scale space view of curve estimation. *The Annals of Statistics* **28**: 408-428.

Chipman, J. (1970). External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* **85**: 347-385.

Cliff, A., and Ord, J. (1973). *Spatial Autocorrelation.* London: Pion.

Cochran, W. (1954). Some methods of strengthening the common chi-squared tests. *Biometrics* **10**: 417-451.

Cover, T., and Thomas, J. (1991). *Elements of Information Theory.* New York: John Wiley and Sons.

Cressie, N. (1993). *Statistics for Spatial Data.* Chichester: John Wiley and Sons.

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

Daley, D.J., and Vere-Jones, D. (1972). A summary of the theory of point processes. In P. Lewis (ed.). *Stochastic Point Processes: Statistical Analysis, Theory and applications.* New York: John Wiley and Sons.

Davidson, R., and MacKinnon, J. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews* **19**: 55-68.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

de Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. (1997). *Computational Geometry: Algorithms and Applications.* Berlin: Springer-Verlag.

Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns.* London: Arnold.

Dixit, A., and Stiglitz, J. (1977). Monopolistic competition and optimum product diversity. *American Economic Review* **67**: 297-308.

Donato, V., and Haedo, C. (2002). *La Nueva Geografía Industrial Argentina.* Buenos Aires: Unión Industrial Argentina and Università di Bologna at Buenos Aires.

Donato, V., Haedo, C., Reynolds, P. and Rocha, H. (2008). Local production systems, entrepreneurship and regional development: theoretical arguments and empirical evidence from Argentine. In G. Becattini and F. Sforzi (eds.). *Sviluppo Locale. Capitale Imprenditoriale.* Torino: Rosenberg and Sellier.

Dufour, J-M., and Khalaf, L. (2001). Monte Carlo test methods in econometrics. In B. Baltagi (ed.). *A Companion to Econometric Theory.* Oxford: Blackwell Publishers.

Dumais, G., Ellison, G., and Glaeser, E. (1997). Industrial concentration as a dynamic process. *National Bureau of Economic Research, Cambridge.* Working Paper 6270.

Duong, T., Cowling, A., Koch, I., and Wand, M. (2007). Feature significance for multivariate kernel density estimation. Submitted.

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 Puga, D. (2001). Nursery cities: urban diversity, process innovation, and the life-cycle of products. *American Economic Review* **91**: 1454-1477.

Duranton, G., and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In J.V. Henderson and J-F. Thisse (eds.). *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography.* Amsterdam: Elsevier North-Holland.

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

Duranton, G., and Overman, H. (2006). Exploring the detailed location patterns of UK manufacturing industries using micro-geographic data. *CEPR Discussion Papers* 5858.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28**: 181-187.

Eaton, J., and Eckstein, Z. (1997). Cities and growth: theory and evidence from France and Japan. *Regional Science and Urban Economics* **27**: 443-474.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Ejermo, O. (2005). Technological diversity and Jacobs' externality hypothesis revisited. *Growth and Change* **36**: 167-195.

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

Feldman, M., and Florida, R. (1994). The geographic sources of innovation: technological infrastructure and product innovation in the United States. *Annals of the Association of American Geographers* **84**: 210-229.

Fienberg, S. (1980). *The Analysis of Cross Classified Categorical Data.* Cambridge, MA: MIT Press.

Flahaut, B., Mouchart, M., San Martin, E., and Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach. *Accident Analysis and Prevention* **35**: 991-1004.

Fujita, M. (1988). A monopolistic competition model of spatial agglomeration: A differentiated product approach. *Regional Science and Urban Economics* **18**: 87-124.

Fujita, M. and P. Krugman (1995). When is the economy monocentric? Von Thünen and Chamberlin unified. *Regional Science and Urban Economics* **25**: 505-528.

Fujita, M., and Krugman, P. (2004). The new economic geography: past, present, and future. *Regional Science* **83**: 139-164.

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

Fujita, M., and Mori, T. (1997). Structural stability and evolution of urban systems. *Regional Science and Urban Economics* **27**: 399-442.

Fujita, M., and Mori, T. (2005a). Frontiers of the new economic geography. *Papers in Regional Science* **84**: 377-405.

Fujita, M., and Mori, T. (2005b). Transport development and the evolution of economic geography. *Portuguese Economic Journal* **4**: 129-156.

Fujita, M., and Tabuchi, T. (1997). Regional growth in postwar Japan. *Regional Science and Urban Economics* **27**: 643-70.

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

Gasser, T., and Müller, H.G. (1979). Kernel estimation of regression functions. Smoothing techniques for curve estimation. In T. Gasser and M. Rosenblatt (eds.). *Lecture Notes in Mathematics 757.* London: Springer-Verlag.

Gerlach, K., Rønde, T., and Stahl, K. (2001). *Firms come and go, labor stays: Agglomeration in high-tech industries.* University of Mannheim.

Getis, A., and Ord, J. (1992). The analysis of spatial association by use of distance statistic. *Geographical Analysis* **24**: 189-206.

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

Gibson, L., Miller, M., and Wright, N. (1991). Location quotient: a basic tool for economic development analysis. *Economic Development Review* **9**: 65-68.

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

Gilula, Z. and Haberman, S. (1998). Chi-square, partition of. In P. Armitage and T. Colton (eds.). *Encyclopedia of Biostatistics.* Chichester: Wiley.

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

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

Godtliebsen, F., Marron, J., and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* **11**: 1-21.

Gokhale, D., and Kullback, S. (1978). *The Information in Contingency Tables.* New York: Dekker.

Goldstein, G., and Gronberg, T. (1984). Economies of scope and economic of agglomeration. *Journal of Urban Economics* **16**: 91-104.

Goodman, L. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interaction in contingency tables, with or without missing entries. *Journal of American Statistical Association* **63**: 1091-1131.

Goodman, L. (1970). How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology* **75**: 1-40.

Goodman, L. (1971). The analysis of multi-dimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**: 33-61.

Goodman, L. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology* **87**: 612-50.

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

Goodman, L. (1986). Some useful extensions of usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* **54**: 243-309.

Gustafson, K. A. (1988). Approximating confidence intervals for indices of fish population size structure. *North American Journal of Fisheries Management* **8**: 139-141.

Haberman, S. (1974). *The Analysis of Frequency Data.* Chicago: University of Chicago Press.

Haberman, S. (1978). *Analysis of Qualitative Data: Introductory Topics.* New York: Academic Press.

Haberman, S. (1979). *Analysis of Qualitative Data, Volume II: New Developments.* New York: Academic Press.

Hannig, J., and J. S. Marron (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association* **101**: 484-499.

Hanson, G. (1998). Market potential, increasing returns, and geographic concentration. *NBER Working Paper* 6429.

Hanson, G. (2000). Scale economy and the geographic concentration of industry. *NBER Working Paper Series* 8013.

Hazelton, M.L., and Marshall, J.C. (2008). Linear boundary kernels for bivariate density estimation. *Statistics and Probability Letters.* In press.

Heitjan, D., and Rubin, D. (1991). Ignorability and categorical data. *Annals of Statistics* **19**: 2244-2253.

Henderson, J.V. (1974). The sizes and types of cities. *American Economic Review* **64**: 670-656.

Henderson, J.V. (1997). Medium size cities. *Regional Science and Urban Economics* **27**: 583-612.

Hirschfeld, H. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* **31**: 520-524.

Hirschman, A. (1958). *The Estrategy of Development*. New Heaven: Yale University Press.

Hohenberg, P. (2004). The historical geography of european cities: an interpretative essay. In J.V. Henderson and J-F. Thisse (eds.). *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*. Amsterdam: Elsevier North-Holland.

Hohenberg, P., and Lees, L.H. (1985). *The Making of Urban Europe, 1000-1950*. Cambridge: Harvard University Press.

Hoover, E. M. (1937). *Location Theory and the Shoe and Leather Industries*. Cambridge, MA: Harvard University Press.

Isaksen, A. (1996). Towards increased regional specialization? The quantitative importance of new industrial spaces in Norway, 1970-1990. *Norsk Geografisk Tidsskrift* **50**: 113-123.

Jacobs, J. (1969). *The Economy of Cities*. New York: Vintage.

Jaffe, A., Trajtenber, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evident by patent citations. *Quarterly Journal of Economics* **108**: 577-598.

Jackson, L., Gray, A., and Fienberg, S. (2008). Sequential category aggregation and partitioning approaches for multi-way contingency tables based on survey and census data. *Annals of Applied Statistics* **2**: 955-981.

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

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

Jöckel, K-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics* **14**: 336-347.

Jones, M., and Foster, P. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica* **6**: 1005-1013.

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

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics* **21**: 59-78.

Kingman, J. (1993). *Poisson Processes.* Oxford: Clarendon Press.

Koehler, K. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of American Statistical Association* **81**: 483-493.

Kreiner, S. (2003). Introduction to DIGRAM. Technical Report. Dept. Biostatistics, University of Copenhagen, Denmark.

Krugman, P. (1987). The narrow moving band, the Dutch disease, and the competitive consequences of Mrs. Thatcher. *Journal of Development Economics* **27**: 41-55.

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

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

Krugman, P. (1993). On the number and location of cities. *European Economic Review* **37**: 293-298.

Krugman, P. (1995). *Development, Geography, and Economic Theory.* Cambridge: MIT Press.

Kullback, S. (1959). *Information Theory and Statistics.* New York: John Wiley and Sons.

Kullback, S., and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematics and Statistics* **22**: 79-86.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods* **26**: 1481-1496.

Kulldorff, M., and Nagarwalla, N. (1995). Spatial disease cluster: detection and inference. *Statistics in Medicine* **14**: 799-810.

Kutoyants, Y. (1998). *Statistical Inference for Spatial Poisson Processes.* New York: Springer-Verlag.

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

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

Lancaster, H. (1949). The derivation and partition of chi-squared in certain discrete distributions. *Biometrika* **36**: 117-129.

Lancaster, H. (1951). Complex contingency tables treated by the partition of chi-squared. *Journal of Royal Statistical Society* **13**: 242-249.

Lancaster, A. (1979). *The Chi-Squared Distributions.* New York: John Wiley and Sons.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Oxford University Press.

Lawson, A., and Denison, D.(eds.) (2002). *Spatial Cluster Modelling.* London: Champan and Hall/CRC.

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

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

Maasoumi, E., and Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics* **107**: 291-312.

MacKinnon, J. (2007). Bootstrap hypothesis testing. *Queen's Economic Department.* Working paper 1127.

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

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

Marron, J.S., Ruppert, D. (1994). Transformation to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society* **56**: 653-671.

Marshall, A. (1890). *Principles of Economics.* London: Macmillan. 8th edition published in 1920.

Martin, P. (1999). Public policies, regional inequalities and growth. *Journal of Public Economics* **73**: 85-105.

Martin, P., and Ottaviano, G. (1999). Growing locations: Industry location in a model of endogenous growth. *European Economic Review* **43**: 281-302.

Martin, P., and Ottaviano, G. (2001). Grow and agglomeration. *International Economic Review* **42**: 947-968.

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.

Møller, J. (ed.) (2003). *Spatial Statistics and Computational Methods*. New York: Springer-Verlag.

Møller, J., and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. London: Champan and Hall/CRC.

Moran, P.(1950). Notes on continuous stochastic phenomena. *Biometrika* **37**: 17-23.

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

Mori, T., and Smith, T. (2006). A probabilistic modeling approach to the detection of industrial agglomerations. In progress.

Müller, H.G., and Stadtmüller, U. (1999). Multivariate boundary kernels and a continuous least squares principle. *Journal of the Royal Statistical Society* **61**: 439-458.

Myrdal, G. (1957). *Economic Theory and Underdeveloped Regions*. London: Duckworth.

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

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

Openshaw, S., Craft, A.W., Charlton, M., and Birch, J.M. (1988). Investigation of leukaemia cluster by use of a geographical analysis machine. *Lancet* **1**: 272-273.

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

Ottaviano, G., Tabuchi, T., and Thisse, J-F. (2002). Agglomeration and trade revisted. *International Economic Review* **43**: 409-436.

Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association* **99**: 1002-1014.

Perroux, F. (1955). Note sur la notion de pôle de croissance. *Economique Appliquée* **7**: 307-320.

Porter, M. (1990). *The Competitive Advantages of Nations.* New york: Macmillian.

Porter, M. (1998). *On Competition.* Cambridge: A Harvard Business Review Book.

Prescott, E. (1998). Nedded: a theory of total factor productivity. *International Economic Review* **39**: 525-551.

Puga, D., and Venables, A. (1996). The spread of industry. Spatial agglomeration and economic development. *Journal of the Japanese and International Economics* **10**: 440-464.

Rao, C. (1969). Two descompositions of concentration ratio. *Journal of Royal Statistical Society* **132**: 418-425.

Razin, A., and Sadka, E. (1997). International migration and international trade. In M.R. Rosenzweig and O. Stark (eds.). *Handbook of Population and Family Economics.* Amsterdam: North-Holland.

Reiczigel, J., Rózsa, L., and Zakariás, I. (2005). A bootstrap test of stochastic equality of two populations. *The American Statistician* **59**: 156-161.

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

Ripley, B. (1981). *Spatial Statistics.* Chichester: John Wiley and Sons.

Robinson, P. (1991). Consistent nonparametric entropy-based testing. *Review of Economic Studies* **58**: 437-453.

Robinson, W. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review* **15**: 351-357.

Romer, P. (1986). Increasing returns and long run growth. *Journal of Political Economy* **94**: 1002-1037.

Romer, P. (1990). Endogenous technological change. *Journal of Political Economy* **98**: 71-101.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**: 832-837.

Rosenthal, S., and Strange, W. (2001). The determinant of agglomerations. *Journal of Urban Economics* **50**: 191-229.

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

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**: 65-78.

Saxenian, A. (1994). *Regional Advantage: Culture and Competition in Silicon Valley and Route 128.* Cambridge: Harvard University Press.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**: 461-464

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

Scott, D.W. (1992). Multivariate Density Estimation: Theory, Practice and Visualization. New York: John Wiley and Sons.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* New York: Springer-Verlag.

Silverman, B.W. (1992). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Solow, R. (1956). A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics* **70**: 65-94.

Srole, L., and Langer, T. (1962). *Mental Health in The Metropolis: The Midtown Manhattan Study.* New York: McGraw-Hill.

Stahl, K., and Walz, U. (2001). Will there be a concentration of alikes? The impact of labor market structure on industry mix in the presence of product market shocks. *Hamburg Institute of International Economics.* Working Paper 140.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences* **100**: 9440-9445.

Storper, M. (1995). The resurgence of regional economies ten years later: the region as a nexus of untraded interdependencies. *European Urban and Regional Studies* **2**: 191-221.

Taylor, C.C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76**: 705-712.

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

Thünen, J.H. von (1826). Der Isoliere Staat in Beziehung auf Landtschaft und Nationalökonomie. Hamburg: Perthes. English translation: *The Isolated State.* Oxford: Pergamon Press (1966).

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.

Venables, A. (1996). Equilibrium locations of vertically linked industries. *International Economic Review* **37**: 341-359.

Wand, M.P., and Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Wang, Q., Kulkarni, S., and Verdú, S. (2006). A nearest-neighbor approach to estimating divergence between continuous random vectors. *2006 IEEE International Symposium on Information Theory.* Seatle, USA.

Williamson, J.(1965). Regional inequality and the process of national development. *Economic Development and Cultural Change* **13**: 3-45.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* New York: John Wiley and Sons.

Willenborg, L., and de Waal, T. (2000). *Elements of Statistical Disclosure Control. Lecture Notes in Statistics 155.* New York: Springer-Verlag.

World Bank (2000). *Entering the 21st Century. World Development Report 1999/2000.* Oxford: Oxford University Press.

Yao, S. (1999). On the decomposition of Gini coefficients by population class and income source: a spreadsheet approach and application. *Applied Economics* **31**: 1249-1264.

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

Zolotarev, V. (1983). Probability metrics. *Theory of Probability and Its Applications* **28**: 278-302.