PhD in Economics and Management

Economics Curriculum

32st Cycle

Department of Economics and Management,

University of Padova

Two Essays about Agglomeration Dynamics and Firm Economic Performance

Head of the PhD Program: Prof. Antonio Nicolo

Supervisor: Prof. Giulio Cainelli

PhD Candidate: Yuting Jiang

Content

Acknowledgment

This thesis involves three years dedicated and hard work, which would not possible be accomplished without the help, support, trust and encouragement from several and special people I will always be grateful of.

First of all, I would like to thank my supervisor Professor Giulio Cainelli, who introduces me to the regional economics. He has always believed in me and trust in my abilities, giving me lots of opportunities to practice and improve my skills. His patience and encouragement inspired me to explore and persue more in academics.

My thanks also go to my co-superviser, Professor Li Yuan in Guangzhou University, for hosting me during my visiting period in China, financing me the dataset for my second chapter and introducing me participating in the courses and conferences related to my research.

I thank all the Professors of the Department of Economics and Management "Marco Fanno" of the University of Padova, who have ever given me the lectures or inspired me through any other kinds of learning channels, such as seminars, workshops, conferences and so on.

I thank another co-author of first chapter, and enjoy working with: Roberto Ganau, who devoted a lot in our joint work. He has always been so generous, giving me much substantial help during my 3 years PHD research.

Last but not the least, special thanks with all my heart to my family: my mom Yu Yun and my dad Jiang Xingshun, who have always supporting me, contributing whatever they have and all their love to me; my husband Lin Manjia, who has always respected my decisions, supported me along this long way, gave me much more courage defeating difficulties.

Introduction

There is a long history of research about agglomeration economies in economic geography and regional economics. Researchers have tried and still trying to answer to questions such as does spatial clustering still matter today? how it evolves? which are their determinants? how agglomeration externalities affect the economic performance of regions and firms?

The empirical literature on agglomeration economies focus on two main topics. The first is about the sources of geographic concentration of economic activities. The second is related to the effects of spatial agglomeration on firm economic performance.

Research about the determinants of agglomeration economies can be date back to Marshall (1920), which identified three different sources: input sharing, labor market pooling and knowledge spillovers. Apparel manufacturer in New York is an example of input sharing, since firms can purchase a variety of relatively cheap buttons from nearby button manufacturing firms. A software company in Silicon Valley can quickly hire one skilled programmer. Meanwhile, a skilled programmer living in Silicon Valley can easily find a new job in this cluster without moving to another place. This is a good example of labor market pooling, which reduces the searching costs for both employees and employers, as well as improves the matching quality. An example of knowledge spillovers can be the random interaction between people working in similar fields who exchange tacit knowledge with each other.

Research about the effects of spatial agglomeration on firm economic performance is more recent. Generally it refers to the effects of spatial agglomeration and thus of different types of local externalities on firms' economic performance, that is whether location within an agglomerated area generates positive returns on the economic performance of firms and, consequently, of the economic dynamisms and growth of regions.

This thesis intends to move along this line of research, specifically, try to contribute to this debate in two directions: [1] investigating the temporal dynamics of spatial agglomeration in the Italian manufacturing industry; [2] analysing the relationship between related variety and firm economic performance in China. In general, the thesis is a collection of two empirical studies dealing with spatial agglomeration from two different perspectives.

The first chapter of this thesis, "Agglomeration over time", which is co-authored with Giulio Cainelli (University of Padova) and Roberto Ganau (University of Padova and LSE), is aimed to investigate the space-time agglomeration dynamics that characterised the manufacturing industry during the recent period of the Great Recession. Specifically, the analysis uses a large sample of geo-referenced single-plant manufacturing firms observed over the period 2007-2012 and located in the Italian continental territory to explore the spatial and temporal dimensions of clustering processes, as well as their potential interaction. The empirical analysis is carried out by adopting three different statistical approaches. First, the index of industrial geographic concentration proposed by Ellison and Glaeser (1997). Second, the spatial *K*-function, originally proposed by Ripley (1976) in the context of spatial points pattern analyses. Third, the space-time *K*-function, that has been proposed by Diggle *et al.* (1995) as an extension of the univariate spatial *K*-function in order to analyse simultaneously the spatial and temporal dimensions of spatial points processes, as well as their potential interaction. The analysis based on EG index highlights the existence of heterogeneity in spatial agglomeration between different industries, but this region-based measure suffers from MAUP problem. To correct the MAUP, we introduce spatial point process method-K function,as well as M-function, which relying on micro-geographic data, rather than pre-defined spatial area, to test firm location patter against Completely Spatial Randomness (CSR). To address the dynamic process that evolve both over space and time, we apply space-time K-function, and some statistical diagnostics, to test the potential interaction between these two dimensions. By space-time analysis, we empirically confirm that, different space-time processes can lead to the spatial patterns which look the same. No significant interaction between spatial and temporal processes, which could be the short period we observe.

The second chapter of the thesis “Related Variety and Economic Growth at Firm Level in China”, which is a single-authored paper, aims at investigating the effect of related and unrelated variety on firm level economic growth in China. As empirical results of MAR externalities and Jacobs externalities impact on economic growth are various and inconclusive. Related variety and unrelated variety, a new entropy method proposed by Frenken et.al.(2007), which focuses on the structure inside industry, was applied in this chapter. Basically, firm economic proportional growth specification-Gibrat’s Law, is extended including these two agglomeration externalities-which sectoral diversity is split into related and unrelated variety for distinguishing between sectors with cognitive or technology proximity, with a sample of 84,868 Chinese firms operating in manufactory industry observed during the period 2006-2013. Recent studies about related variety and economic growth, which indeed is the main reason for regional growth, most empirical papers are about developed countries, studies about developing countries are rare. This chapter contributes an empirical study about this debate in a typical developing country, and to our knowledge, it’s the first paper analysis Chinese firms economic growth within related variety framework; besides it’s a firm level empirical research with historical data during 2006-2013, a transformation period for China, with rapid economic development and technological innovation. The results show that, correcting only for sample-selection, unrelated variety has a negative and statistically significant impact. Accounting also for the endogeneity of the two main explanatory variables – related and unrelated variety –the negative effect of unrelated variety becomes insignificant. A positive effect for related variety and negative for unrelated variety is detected only when we consider high-developed Chinese regions. Finally, a positive effect of related variety is identified for large firms.

Chapter 1

Agglomeration over time

Giulio Cainelli

Department of Economics and Management “Marco Fanno”, University of Padova

Via del Santo 33, 35123 Padova, Italy

Tel: (0039) 0498274227 − Fax: (0039) 0498274221 − E-mail: giulio.cainelli@unipd.it

Roberto Ganau

Department of Economics and Management “Marco Fanno”, University of Padova

Via del Santo 33, 35123 Padova, Italy

E-mail: roberto.ganau@unipd.it

Yuting Jiang

Department of Economics and Management “Marco Fanno”, University of Padova

Via del Santo 33, 35123 Padova, Italy

E-mail: yuting.jiang@studenti.unipd.it

Agglomeration over time

Abstract: This paper analyses the spatial and temporal dynamics of agglomeration occurred in the Italian manufacturing industry during the period of the Great Recession. The analysis relies on three different statistical methods – the Ellison and Glaeser's (1997) index of industrial geographic concentration, the spatial *K*-function, and the space-time *K*-function –, and uses a large sample of geo-referenced, single-plant manufacturing firms observed over the period 2007-2012. Our results unveil three interesting insights. First, we demonstrate that different statistical techniques can lead to different results. Second, we find that the majority of Italian manufacturing two-digit sectors experienced a process of spatial dispersion during the period of the Great Recession. Finally, even if we do not detect any statistical evidences of space-time interaction, we show that the process of spatial dispersion has been more intensive at short distances and during the first year of the Great Recession – namely, the year 2008.

Keywords: Agglomeration dynamics; Spatial methods; Manufacturing industry; Italy.

JEL Codes: C3; D24; R12.

1. INTRODUCTION

The spatial agglomeration of economic activities is a key feature of the economic geography of many countries, regions, and clusters (Porter, 1990). Well-known examples of such a phenomenon are the Silicon Valley (Saxenian, 1994), the carpet manufacturing industry in Dalton (Krugman, 1991), and the Italian industrial districts (Brusco, 1982; Becattini, 1990).

The most recent literature on spatial agglomeration has focused on two main research topics. The first one refers to the empirical determinants of the geographic concentration of production activities. In other words, it deals with the reasons why some geographic areas are more agglomerated than others. According to the Marshallian tradition (Marshall, 1920), these factors are generally identified in knowledge spillovers, input sharing, and labour market pooling (Strange and Rosenthal, 2001). The second stream of literature refers to the effects of spatial agglomeration, and, precisely, of different types of local externalities, on firms' economic performance. Thus, this second research stream aims at understanding whether location within an agglomerated area generates positive returns on the economic performance of firms (Henderson 2003; Martin *et al.*, 2011), and, consequently, of the economic dynamisms and growth of territories as a whole (Glaeser *et al.*, 1992; Henderson *et al.*, 1995; Combes, 2000; Paci and Usai, 2008).

A commonality characterising the majority of these studies is that spatial agglomeration is treated as a static – i.e. time-invariant – phenomenon. To the best of our knowledge, only few contributions have investigated empirically the temporal dynamics of spatial agglomeration processes, that is whether and how the spatial agglomeration of firms changes over time (e.g. Arbia *et al.*, 2010; Kang, 2010; Arbia *et al.*, 2014), and how these changes influence its firm-level economic returns (e.g. Martin *et al.*, 2011; Cainelli *et al.*, 2016). Indeed, the temporal dynamics is a key dimension of spatial agglomeration, which is a complex process changing continuously over both space and time. In fact, some industries experience processes of geographic clustering/concentration, while others spread over space. Moreover, the process of geographic concentration/dispersion can accelerate during some years (i.e. time clustering), while it can reduce during others (i.e. time dispersion). Agglomeration may accelerate over time during the initial stage of the product life-cycle of a local industry/cluster, when a rapid increase in the number of new firms is generally observed (Klepper, 1996; Klepper and Graddy, 1990; Kenneth, 1993; Kang, 2010), in particular years characterised by 'sudden changes' associated with historic accidents, technological revolutions or discoveries, structural transformations (e.g. industrialisation), external shocks (e.g. natural disasters, economic or financial crisis), and, finally, during some specific periods of the business cycle (Arbia *et al.*, 2010; Kang, 2010; Henderson *et al.*, 2018). On the contrary, spatial agglomeration may decelerate in years characterised by the mature stage of the product life-cycle of a local industry/cluster, or during periods without significant technological advances or structural changes. In all these cases, the underlying mechanism is the same: the entry (exit) of firms into (from) the local industry/cluster, or firms' re-location decisions, may change the agglomeration structure over both space and time. This implies that spatial agglomeration cannot be investigated as a static phenomenon, given its intrinsic dynamic nature. In other words, the temporal dimension of agglomeration – the so-called time clustering – cannot be ignored. Moreover, it should be investigated at the industry level, since sectors do not behave homogeneously over both space and time.

The aim of this paper is to investigate the agglomeration dynamics which has characterised the Italian manufacturing industry during the recent period of the Great Recession. Specifically, we employ a large sample of geo-referenced, single-plant manufacturing firms observed over the period 2007-2012, and located in the Italian continental territory, to analyse the spatial and temporal dimensions of agglomeration processes at the two-digit sector level, as well as their potential interactions. The choice of focusing on the period 2007-2012 is justified by the fact that the effects of the Great Recession have been particularly relevant for the Italian economy during these years. Indeed, during this period of time, firm demography changed significantly in many sectors, thus affecting their agglomeration structure.

We investigate these phenomena using three different statistical approaches. The first one is the Ellison and Glaeser's (1997) index of industrial geographic concentration, that is calculated for the years 2007 and 2012. We compute this index at the two-digit sector level by adopting three different spatial units of analysis: the region, that corresponds to the level 2 of the *Nomenclature des Unités Territoriales Statistiques* (NUTS) adopted by the European Union (EU); the province, corresponding to the NUTS-3 level; and the Local Labour Market (LLM), that corresponds to a functional area defined according to economic – i.e. commuting patterns of workers – rather than administrative criteria.

As it is well known, the Ellison and Glaeser's (1997) index suffers from the so-called Modifiable Areal Unit Problem (MAUP), which refers to the discretionary choice of the spatial unit used to analyse geographic-based phenomena (Arbia, 1989; Amrhein, 1995; Wong and Amrhein, 1996; Arbia, 2001). In fact, the use of pre-defined geographic units can introduce statistical biases in this kind of analysis (Arbia, 1989). As suggested by the literature, a possible solution to the MAUP consists in relying on micro-geographic data, and adopting statistical methods that treat the space as a continuum. Based on these insights, the second approach adopted in the paper consists in estimating Ripley's (1976) spatial *K*-function by relying on geographic information on the location of firms – namely, their latitude and longitude coordinates. This approach allows us to evaluate the geographic scale at which a sector shows a clustering pattern, if any, in a 'single moment of time' (Arbia and Espa, 1996; Marcon and Puech, 2003; Duranton and Overman, 2005; Arbia *et al.*, 2008; Marcon and Puech, 2010; Albert *et al.*, 2012; Scholl and Brenner, 2016). Thus, we are able to assess whether sector-specific clustering patterns took place, and how they changed between the years 2007 and 2012. This is done by comparing the values of the estimated *K*-functions in the two observational periods. In other words, these univariate functions can be used in a dynamic fashion only by analysing them separately, i.e. year by year. The limit of this method is that it does not allow us to describe the dynamics of agglomeration patterns taking place between these two years. This can be a problem, since “different space-time processes can lead to resulting spatial patterns which look the same” (Arbia *et al.*, 2010, p. 312).

The third approach adopted in the paper is aimed precisely at overcoming this limitation, and consists in estimating the space-time *K*-function (Diggle *et al.*, 1995). This function can be used to analyse simultaneously the spatial and temporal dimensions of agglomeration processes, as well as the potential existence of space-time interactions. Thus, the space-time *K*-function allows us to identify not only the existence of concentration/dispersion processes in different years – as in the case of the univariate spatial *K*-function –, but also the threshold values at which these concentration/dispersion processes occur over both space and time – for example, a rapid acceleration over time in the process of spatial clustering of a sector.

The rest of the paper is organised as follows. Section 2 describes the dataset used in the empirical analysis. Section 3 presents the statistical approaches employed. Section 4 presents and discusses the empirical evidence on the agglomeration dynamics. Section 5 draws some policy implications, and concludes the work.

2. THE DATASET

The analysis of the space-time dynamics of spatial agglomeration relies on a large sample of single-plant manufacturing joint stock companies located in the Italian continental territory. The firm-level data are drawn from the *AIDA* database (*Bureau Van Dijk*), that provides personal information and balance sheet data for Italian firms.

The original sample of 230,198 firms was first cleaned by removing multi-plant firms.[1] The choice of focusing on single-plant firms is driven by the fact that the *AIDA* database provides the

[1] Single-plant firms in the *AIDA* database have been identified using information derived from the *ASIA Archive* provided by the Italian National Institute of Statistics (Istat), that collects selected information on the entire population of firms operating in Italy.

exact address of headquarters only. Therefore, the focus on single-plant firms allows us to study the clustering dynamics of firms by considering the exact location where the economic activity takes place. Second, firms without information on the exact address were removed, given the necessity to identify the pair of geographic coordinates (i.e. latitude and longitude) for each individual observational unit. Third, firms located in the two main islands of Sicily and Sardinia, as well as firms located into smaller islands, were removed. The cleaning procedure left us with a final sample of 149,135 manufacturing firms observed over the period 2007-2012, corresponding to an unbalanced panel dataset of 614,220 observations.

The period of observation starts with the year 2007, which is generally regarded as a pre-crisis year, and ends with the year 2012, which corresponds to the first year the Italian economy entered a second wave of downturn after the recovery peak reached in 2011, as shown in Figure 1.

[--- Figure 1 near here ---]

The final sample includes firms operating in the two-digit sectors 10 to 33 of the NACE Rev. 2 classification of economic activities adopted by the EU, except for the two-digit sector "12 − Manufacture of tobacco products", which has been excluded *à priori* given the peculiarity of being a government monopoly. Appendix Table A1 reports the sample distribution by two-digit manufacturing sector, while Appendix Table A2 reports the sample distribution by two-digit sector and year, as well as the percentage change in the number of observational units between the years 2007 and 2012. Overall, the number of observational units has increased by 3.82% from the pre-crisis year 2007 to the year 2012. The majority of two-digit sectors has observed an increase in the number of firms, ranging from 0.17%, concerning the sector "32 – Other manufacturing", to a 37.45%, concerning the sector "33 − Repair and installation of machinery and equipment". On the contrary, eight out of the 23 manufacturing sectors analysed show a reduction in the number of observational units, from -0.37%, concerning the sector "18 – Printing and reproduction of recorded media", to -7.88%, concerning the sector "14 − Manufacture of wearing apparel".

Appendix Table A3 complements the previous insights by providing the sample distribution by two-digit sector and geographic area defined at the NUTS-1 level. Overall, 60.3% of sample firms is located in the Northern regions of Italy, 22.2% of sample firms is located in Central Italy; while only 17.5% of sample firms is located in the South. As shown in Appendix Table A4, it is also interesting to note that about all two-digit sectors have recorded a reduction in the average firm size − defined as the average number of employees per firm − between the years 2007 and 2012, ranging from -0.27%, concerning the sector "24 − Manufacture of basic metals", to - 20.41%, concerning the sector "23 − Manufacture of other non-metallic mineral products". On the contrary, only two sectors have shown an increase in the average firm size, namely sector "19 − Manufacture of coke and refined petroleum products" (24.34%), and sector "21 − Manufacture of basic pharmaceutical products and pharmaceutical preparations" (5.76%).

Interesting insights emerge looking at Appendix Table A5, which reports the rate of firms observed in year $t = 2007, \dots, 2011$ and survived over the subsequent periods $t + n$, with $n = 1, \dots, 5$.[2] First, the survival rate has decreased over time from $t + 1$ to $t + n$. Second, the one-year survival rate for firms observed in the year 2007 (i.e. the pre-crisis year) is higher than the corresponding value for firms observed in the subsequent years. Third, the one-year survival rate of firms observed in 2010 is slightly higher than the corresponding values for firms observed in the years 2009 and 2011. As previously underlined, this last evidence could probably depend on the fact that the Italian economy reached a recovery peak in 2011 before entering a new phase of downturn in 2012. Finally, Appendix Figure A1 maps the spatial distribution of the sample firms by two-digit manufacturing sector in the years 2007 and 2012.

[2] The survival rate over the period $t, t + n$ is defined as the share of firms observed at time t and survived at time $t + n$ over firms observed at time t.

3. STATISTICAL MODELLING

As already mentioned, the empirical analysis is carried out by adopting three different statistical approaches: the index of industrial geographic concentration (Ellison and Glaeser, 1997); the spatial *K*-function (Ripley, 1976); and the space-time *K*-function (Diggle *et al.*, 1995).

3.1. The index of industrial concentration

The first step of the empirical analysis is based on Ellison and Glaeser's (1997) index of industrial geographic concentration. This index allows for the cross-sector comparison of the degree of geographic concentration, and, specifically, it is employed to identify the clustering behaviour in terms of spatial concentration/dispersion of two-digit manufacturing sectors in the years 2007 and 2012, as well as to evaluate how agglomeration patterns potentially changed between the two years.

Three different spatial units have been considered to calculate the sector-specific concentration index. First, two administrative geographies defined at the NUTS levels 2 and 3, and corresponding to 18 regions and 90 provinces, respectively.[3] Second, the LLM, that consists of 559 functional areas.[4] Formally, the index for the two-digit manufacturing sector $s = 1, \dots, 23$ at time $t = 2007, 2012$ is computed as follows (Ellison and Glaeser, 1997):

[3] As previously specified, the NUTS-2 insular regions of Sicily and Sardinia, and the NUTS-3 regions located within these two main islands, have been excluded *à priori* from the analysis. During most of the period investigated in this paper, Italy was split into 107 NUTS-3 regions, nine (eight) of which were located in Sicily (Sardinia). Considering the abovementioned exclusion criteria, the final number of NUTS-3 regions considered is equal to 90 spatial units.

[4] Italian LLMs are defined according to the classification adopted by Istat in the 2001 Industry and Services Census, that identified 686 LLMs. The exclusion of all insular municipalities has led to drop 77 LLMs located in Sicily, 45 LLMs located in Sardinia, two LLMs located in the Ischia island (Campania region) and two LLMs located in the Elba island (Tuscany region). In addition, the cleaning procedure of the firm-level data has led to drop firms located in the municipality of San Marcello Pistoiese (Tuscany region), such that the final number of LLMs considered is equal to 559 spatial units.

$$\hat{\gamma}_{st} = \left\{ \sum_{g=1}^{G} \left(p_{sgt} - p_{gt}\right)^2 - \left[\left(1 - \sum_{g=1}^{G} p_{gt}^2\right) \sum_{k=1}^{K} z_{skt}^2 \right] \right\} \Big/ \left[\left(1 - \sum_{g=1}^{G} p_{gt}^2\right) \left(1 - \sum_{k=1}^{K} z_{skt}^2\right) \right] \qquad (1)$$

where p_{sgt} denotes the share of employment in sector s in geography $g = 1, \dots, G$ – it being the NUTS-2 region, the NUTS-3 region, or the LLM – at time t; p_{gt} denotes the share of employment in geography g at time t; and z_{skt} denotes the share of employment of firm k in sector s at time t.

The Ellison and Glaeser's (1997) index provides a simple and easy-to-understand measure to explore sector-specific clustering behaviours. Indeed, given the benchmark case of complete randomness in the location choice of firms for $E(\hat{\gamma}_s) = 0$, then positive values of the index provide evidence of concentration, while negative values provide evidence of firms locating more diffusely than expected. However, similarly to the many other region-based indexes of industrial geographic concentration – such as the Gini index used by Krugman (1991), or that proposed by Maurel and Sédillot (1999) –, it suffers from a main shortcoming, that is the MAUP.

The MAUP refers to the discretionary choice of the spatial partition used to analyse geographic-based phenomena (Arbia, 1989; Amrhein, 1995; Wong and Amrhein, 1996; Arbia, 2001). In the context of the analysis of spatial agglomeration and clustering dynamics, the MAUP emerges because neither administrative regions nor (functional) LLMs can necessarily coincide with the real economic areas where firms' location processes take place. As a consequence, the use of spatial units that differ in shape and size, and that are characterised by pre-defined geographic boundaries, can introduce statistical biases related to both the level of aggregation and the geographic scale (Arbia, 1989).

3.2. The spatial K-function

Following Arbia (2001), a possible solution to relax the MAUP consists in relying on micro-geographic data. This means that spatial agglomeration and clustering dynamics are analysed at the level of the individual firms, rather than considering spatially-aggregated patterns of industries. Specifically, the literature suggests to rely on a class of spatial points statistics, namely the *K*-functions – originally proposed by Ripley (1976), and widely employed in the contexts of ecology, molecular biology, and epidemiology, among other fields –, in order to evaluate the geographic scale at which an industry shows a clustering pattern, if any (e.g. Arbia and Espa, 1996; Marcon and Puech, 2003; Duranton and Overman, 2005; Arbia *et al.*, 2008; Marcon and Puech, 2010; Albert *et al.*, 2012; Scholl and Brenner, 2016). This type of spatial statistical approach is based on the idea of using the firm as the spatial unit of analysis, and of treating the space as a continuum rather than using pre-defined spatial areas.[5]

Therefore, the second step of the analysis of spatial agglomeration dynamics is performed by employing Ripley's (1976) *K*-function in order to evaluate whether sector-specific clustering patterns took place in – and, potentially, how they changed between – the years 2007 and 2012. Ripley's (1976) *K*-function is a distance-based method that measures the spatial concentration/dispersion of point events – in this case, firms – by counting the number of neighbouring points j occurring within a circle of radius r centred at each reference point i, with $j \neq i$, and then by comparing the observed pattern with the one that would be expected in a situation of complete spatial randomness (CSR), that is a scenario where points are located within the study region randomly and independently from each other. Formally, the estimate of the *K*-function for the two-digit sector s at time $t = 2007, 2012$ can be defined as follows:

$$\widehat{K}_{st}(r) = \frac{1}{\hat{\lambda}_{st} N_{st}} \sum_{i} \sum_{j \neq i} \frac{I\left(d_{x_{st}^{i} x_{st}^{j}}\right)}{w_{x_{st}^{i} x_{st}^{j}}} \tag{2}$$

[5] The use of micro-geographic data has been employed more recently by economists and economic geographers to analyse agglomeration-related externalities on the performance of firms (e.g. Sorenson and Audia, 2000; Rosenthal and Strange, 2003; Baldwin *et al.*, 2008; Cainelli and Lupi, 2010; Eriksson, 2011; Duschl *et al.*, 2014; Duschl *et al.*, 2015; Cainelli and Ganau, 2018), besides industrial clustering dynamics.

where the term N_{st} denotes the total number of firms operating in the two-digit sector s at time t and located in the area of the study region (W), with $\hat{\lambda}_{st} = N_{st}/Area_W$ denoting its estimated density; the term $d_{x_{st}^i x_{st}^j}$ denotes the distance in kilometres between each pair of firms i and j operating in the two-digit sector s and observed at time t – denoted by x_{st}^i and x_{st}^j as spatial points identified by their geographic coordinates; the term $I(\cdot)$ denotes an indicator function that takes value of 1 if $d_{x_{st}^i x_{st}^j} \leq r$, that is whether the distance between a pair of firms i and j ($d_{x_{st}^i x_{st}^j}$) is lower than or equal to the radius r, and value of 0 otherwise; the term $w_{x_{st}^i x_{st}^j}$ denotes the edge correction parameter, that defines the length of the overlap between the circle with radius r centred in the reference firm x_s^i and passing through the firm x_s^j which lies within the study region W (Ripley, 1977). The edge correction term avoids biased estimates of $\widehat{K}_{st}(r)$ which may occur in proximity to the boundaries of the study region W, where increases in r are not accompanied by increases in the number of firms – indeed, the number of firms can be lower in proximity to rather than at longer distances from the study region's boundaries, and there are no firms outside the study region.

Having estimated the sector- and time-specific *K*-functions, it is possible to test for the location pattern characterising the observed firms against the hypothesis of CSR. Under the null hypothesis of CSR, then $\widehat{K}_{st}(r) = \pi r^2$, such that it is possible to compute the difference between the empirical value of the *K*-function for the observed points pattern – namely, $\widehat{K}_{st}(r)$ – and the theoretical value under the hypothesis of CSR. Following Albert *et al.* (2012), this difference – defined as *M*-function – can be formalised as follows:

$$\widehat{M}_{st}^{CSR}(r) = \widehat{K}_{st}(r) - \pi r^2 \quad (3)$$

such that if $\widehat{K}_{st}(r) > \pi r^2$, then the observed firms in sector s at time t will show a clustering pattern at a certain distance r, because the observed density of firms is greater than the theoretical one; if $\widehat{K}_{st}(r) < \pi r^2$, then the observed firms will show a dispersion pattern; while if $\widehat{K}_{st}(r) = \pi r^2$, then the observed firms will show a random distribution in the space.

3.3. The space-time K-function

Although the use of micro-geographic data in the context of spatial points pattern analyses allows us to relax MAUP-related biases, a second issue emerges with respect to the temporal dimension characterising clustering/dispersion processes of firms. In fact, the spatial agglomeration of economic activities is the result of a dynamic process that evolves over both space and time.

As these processes occur over time, (sector-specific) agglomerative structures may potentially exhibit different degrees of clustering at different spatial scales along the temporal dimension. The contributions by Getis (1964) and Getis and Boots (1978) are among the first ones to highlight that time matters, besides space, in the analysis of spatial events. In particular, they suggest that it is the temporal evolutionary perspective that can help understanding the (observed) resulting spatial structure. Indeed, any spatial structure evolves over time, and, in particular, similar spatial patterns can be the result of very different space-time processes (Arbia *et al.*, 2010). This line of reasoning fits perfectly the analysis of the spatial agglomeration of economic activities. As firm-level demographic and localisation phenomena present both a spatial and a temporal feature, industrial clustering/dispersion processes are likely to evolve along both dimensions, and, as a consequence, the observed spatial configuration of firms could be the result of their interaction.

The joint analysis of the spatial and temporal processes underlying agglomerative structures has received little attention in the empirical literature. Indeed, as underlined by Arbia *et al.* (2010, pp. 311-312), "[t]ime series methods have generally disregarded the spatial dimension while spatial clustering models have been essentially static and have only analysed the outcome of the dynamic adjustments as it is observed in one single moment of time". To the best of our knowledge, only few contributions have analysed industrial clustering processes by dealing with the MAUP, and by accounting simultaneously for the spatial and temporal dimensions, as well as for their potential interaction. In particular, Arbia *et al.* (2010) focus on the long-run localisation process of firms located in Rome (Italy), and operating in the Information and Communication Technology sector. Kang (2010) studies relevant clusters of manufacturing and services industries operating within the Columbus Metropolitan Statistical Area (USA). Finally, Arbia *et al.* (2014) consider the entry and exit dynamics of pharmaceutical and medical devices firms located in the Veneto region (Italy) over the period 2004-2009. These three contributions limit their investigation to a particular industrial sector, or a particular sub-national territory – either a city or a region – of a country. On the contrary, the present study tries to provide a more general overview of spatial agglomeration processes by considering the entire manufacturing industry, and by looking at the whole Italian (continental) territory.

Following these previous contributions, the third approach adopted in our analysis consists in estimating the space-time *K*-function (Diggle *et al.*, 1995) in order to evaluate the potential interaction between spatial and temporal clustering processes. Specifically, the space-time *K*-function allows us to identify not only the existence of concentration/dispersion processes, but also the threshold values at which these behavioural patterns take place in space and time. In addition, this approach allows us to test for the potential existence of space-time interactions.

The space-time *K*-function measures the expected number of points per unitary area in the study region and per unit of time falling at a spatial distance and at a time interval equal to or lower than a radius r and a time interval t, respectively, from a reference point (French *et al.*, 2005). Drawing on Diggle *et al.* (1995), an unbiased edge-corrected estimator of the space-time *K*-function for the two-digit sector s over the observational period $t = 1, \dots, T$ can be defined as follows:

$$\widehat{K}_s(r,t) = \frac{1}{\hat{\lambda}_s^{RT} N_s} \sum_i \sum_{j \neq i} \frac{I_r\left(d_{x_s^i x_s^j}\right) I_t\left(t_{x_s^i x_s^j}\right)}{w_{x_s^i x_s^j} v_{x_s^i x_s^j}} \tag{4}$$

where the term N_s denotes the total number of firms operating in the two-digit sector s and located in the area of the study region (W), with $\hat{\lambda}_s^{RT} = N_s/(Area_W T)$ denoting the spatial and temporal joint intensity of the points process; the terms $d_{x_s^i x_s^j}$ and $t_{x_s^i x_s^j}$ denote the spatial distance and the temporal distance, respectively, between each pair of firms i and j operating in the two-digit sector s; the terms $I_r(\cdot)$ and $I_t(\cdot)$ denote indicator functions that take value of 1 if $d_{x_s^i x_s^j} \leq r$ and $t_{x_s^i x_s^j} \leq t$, respectively, and value of 0 otherwise; the terms $w_{x_s^i x_s^j}$ and $v_{x_s^i x_s^j}$ denote edge correction parameters, with $w_{x_s^i x_s^j}$ defining the length of the overlap between the circle with radius r centred in the reference firm x_s^i and passing through the firm x_s^j which lies within the study region W, and $v_{x_s^i x_s^j}$ defining the time segment of length t centred at the firm x_s^i that lies within the total observed duration time between $t = 1$ and $t = T$ (Diggle *et al.*, 1995; Arbia *et al.*, 2010).

Similar estimators can be derived for the purely spatial and temporal processes defining the space-time *K*-function, namely $\widehat{K}_s^R(r)$ and $\widehat{K}_s^T(t)$:

$$\widehat{K}_s^R(r) = \frac{1}{\hat{\lambda}_s^R N_s} \sum_i \sum_{j \neq i} \frac{I_r\left(d_{x_s^i x_s^j}\right)}{w_{x_s^i x_s^j}} \tag{5}$$

$$\widehat{K}_s^T(t) = \frac{1}{\hat{\lambda}_s^T N_s} \sum_i \sum_{j \neq i} \frac{I_t\left(t_{x_s^i x_s^j}\right)}{v_{x_s^i x_s^j}} \tag{6}$$

where the term $\hat{\lambda}_s^R = N/Area_W$ denotes the spatial intensity, and captures the number of points per unit area of the study region; the term $\hat{\lambda}_s^T = N/T$ denotes the temporal intensity, and captures the number of points per unit time; and all other terms are defined as for Equation (4).

Having estimated the sector-specific space-time *K*-functions, it is possible to test for the independence of the spatial and temporal processes under the hypothesis that $\hat{K}_s(r,t) = \hat{K}_s^R(r)\hat{K}_s^T(t)$ in the absence of space-time interaction. The baseline test statistic takes the following functional form (Gatrell *et al.*, 1996):

$$\hat{D}_s(r,t) = \hat{K}_s(r,t) - \hat{K}_s^R(r)\hat{K}_s^T(t) \tag{7}$$

that is proportional to the increased number of points within distance r and time t compared to a process with the same spatial and temporal structures, but no space-time interaction. Visual inspection through a three-dimension plot of $\hat{D}_s(r,t)$ against the spatial and temporal dimensions allows us to uncover the scale and nature of the dependence (Arbia *et al.*, 2010).

An alternative, and more intuitive, transformation of the functional defined in Equation (7) has been proposed by Diggle *et al.* (1995) and French *et al.* (2005) to consider relative quantities rather than absolute numbers. The so-called 'Diggle function' allows for a perspective plot of the $\hat{D}_s(r,t)$ surface through the following functional form:

$$\hat{D}_s^0(r,t) = \frac{\hat{D}_s(r,t)}{\hat{K}_s^R(r)\hat{K}_s^T(t)} \tag{8}$$

that is proportional to the relative increase in the number of points within distance r and time t compared to a process with the same spatial and temporal structures, but no space-time interaction. Similarly to the previous case, the function $\widehat{D}_s^0(r,t)$ can be visually inspected through a three-dimension plot against the spatial and temporal dimension in order to evaluate the existence of space-time interaction in the observed points process (Arbia *et al.*, 2010).

A third approach for the detection of space-time interactions consists in drawing inference on the empirical values of $\widehat{D}_s(r,t)$ by obtaining a significance test through a Monte Carlo approach. Specifically, Diggle *et al.* (1995) suggest to obtain m simulated spatial-temporal points patterns in order to compute m different estimates of $\widehat{D}_s(r,t)$, such that the observed variance of the m estimates, namely $\widehat{Var}_s(r,t)$, can be used as an estimator of the variance of $\widehat{D}_s(r,t)$ (Gatrell *et al.*, 1996; Arbia *et al.*, 2010). Having retrieved the estimated variance of $\widehat{D}_s(r,t)$, it is possible to compute the 'standardised residuals' as follows (Diggle *et al.*, 1995):

$$\hat{R}_s(r,t) = \frac{\widehat{D}_s(r,t)}{\sqrt{\widehat{Var}_s(r,t)}} \tag{9}$$

which gives a measure of space-time interaction representing the excess number of points of $\widehat{K}_s(r,t)$ relative to $\widehat{K}_s^R(r)\widehat{K}_s^T(t)$ (Arbia *et al.*, 2010). The advantage of plotting the 'standardised residuals' against $\widehat{K}_s^R(r)\widehat{K}_s^T(t)$ is that a two-dimension plot is easier to visualise, even though the corresponding spatial and temporal scales are not explicit. Under the hypothesis of no space-time interaction, then $E\left[\hat{R}_s(r,t)\right] = 0$ and $Var\left[\hat{R}_s(r,t)\right] = 1$, and one would expect to observe approximately 95% of the values lying in the interval $[-2,+2]$ (French *et al.*, 2005; Arbia *et al.*, 2010). Thus, substantial values of $\hat{R}_s(r,t)$ lying outside the interval $[-2,+2]$ indicate the presence of space-time interaction characterising the observed points pattern, and this space-time structure can be visually interpreted from the plots of $\widehat{D}_s(r,t)$ or $\widehat{D}_s^0(r,t)$.

However, as underlined by Arbia *et al.* (2010), the interpretation of the 'standardised residuals' plot could be misleading in the case of highly dependent residuals. Therefore, a final Monte Carlo-based test can be performed in order to assess the existence of space-time interaction. This test consists in comparing the observed sum of the functionals $\widehat{D}_s(r,t)$ over all values for r and t with the empirical distribution of the m sums of the corresponding simulated estimates of $\widehat{D}_s(r,t)$ over all r and t, such that there is evidence of overall space-time interaction if the observed sum shows a particularly high-ranked position relative to the simulated sums (Gatrell *et al.*, 1996; Arbia *et al.*, 2010).

4. EMPIRICAL EVIDENCE

4.1. Detecting spatial concentration/dispersion through the Ellison and Glaeser's (1997) index

Table 1 reports the calculated sector-specific indexes of geographic concentration for the years 2007 and 2012, as well as their percentage change, considering the three different geographic scales previously discussed – namely, the NUTS-2 region, the NUTS-3 region, and the LLM. The first interesting insight emerging from Table 1 concerns the high level of heterogeneity characterising the degree of geographic concentration in the Italian manufacturing industry. In fact, the 2007 and 2012 indexes present high levels of cross-sector variation both within the same spatial unit of analysis, and across the three geographies considered.

The comparison of the sector-specific indexes suggests clearly that MAUP-related biases are likely to affect the evaluation of the extent of geographic concentration. In particular, following the classification proposed by Ellison and Glaeser (1997) which identifies sectors as 'lowly' concentrated if $\hat{\gamma}_{st} < 0.02$, 'moderately' concentrated if $0.02 \leq \hat{\gamma}_{st} \leq 0.05$, and 'highly' concentrated if $\hat{\gamma}_{st} >$

0.05, it emerges how individual sectors exhibit a different degree of concentration with respect to the different spatial units. For example, the Italian manufacturing industry in 2007 consisted of four ‘highly’ concentrated sectors – namely, sectors “13 − Manufacture of textiles”, “15 − Manufacture of leather and related products”, “21 − Manufacture of basic pharmaceutical products and pharmaceutical preparations” and “31 − Manufacture of furniture” – when considering NUTS-2 regions, but only the sector “21 − Manufacture of basic pharmaceutical products and pharmaceutical preparations” presented a value of the concentration index higher than 0.05 when considering the NUTS-3 and LLM level geographies – see also Appendix Figure A2, which plots the degree of geographic concentration by two-digit sector in the years 2007 and 2012 calculated for the three different types of spatial units.

In addition, it emerges that the calculation based on different geographic scales leads to different sector-specific temporal dynamics of concentration. As an example, the sector “14 – Manufacture of wearing apparel” exhibits an increase in the degree of geographic concentration between 2007 and 2012 when considering the NUTS-2 geographic level, while a reduction in concentration when considering the NUTS-3 and LLM geographic levels. The opposite dynamics characterises, for example, the sector “20 – Manufacture of chemicals and chemical products” – see also Appendix Figure A3, which plots the percentage change in the degree of geographic concentration over the period 2007-2012 with respect to the different reference spatial units.

[--- Table 1 near here ---]

4.2. Dealing with the MAUP: evidence based on the spatial K-function

The analysis based on Ellison and Glaeser’s (1997) index has highlighted not only the existence of heterogeneity in the spatial agglomeration dynamics of firms operating in different sectors, but also that region-based measures of clustering suffer from MAUP-related biases that may lead to (very) imprecise conclusions. Drawing on this last insight, in this section we present and discuss the results obtained by employing the spatial *K*-function, that is by relying on a micro-geographic approach.

Table 2 summarises the key insights from the analysis, while Appendix Figure A4 reports the plots of the estimated values of the spatial *K*- and *M*-functions by two-digit manufacturing sector for the years 2007 and 2012. At a first look, the information reported in Table 2 suggests that all two-digit manufacturing sectors were characterised by a spatial concentration pattern occurring at all distances both in the year 2007 and in the year 2012. Moreover, this spatial concentration pattern appears to be statistically significant, as $\widehat{K}_s(r) > \pi r^2$ for all sectors in both years.

However, a deeper analysis highlights some interesting insights. In fact, the comparison of the estimated empirical values of *K* in the years 2007 and 2012 suggests that 14 out of 23 two-digit sectors experienced a reduction in the degree of spatial concentration between the years 2007 and 2012 at all distances. This means that these sectors, although characterised by significant spatial concentration in both years, have experienced a slight process of spatial dispersion. The opposite dynamics has characterised the sector "15 – Manufacture of leather and related products", that experienced an increase in the degree of spatial concentration at all distances in the 2012 with respect to the year 2007. The remaining sectors present different dynamics: some of them show an increase (decrease) in the degree of spatial concentration between the years 2007 and 2012 at short distances, while a decrease (increase) at longer distances, while others exhibit more complicated dynamics characterised by alternating increases and reductions in the degree of spatial concentration between the years 2007 and 2012 over different distance intervals.[6]

[6] The sector-specific comparisons of the estimated empirical *K* values between the years 2007 and 2012 are evaluated considering sector- and year-specific mean values of the estimated empirical *K* values averaged over distance bands of 10 kilometres each in the interval [0, 250] km.

[--- Table 2 near here ---]

4.3. The space-time agglomeration dynamics

As previously underlined, agglomeration tends to evolve not only over space, but also over time. Indeed, firms' agglomeration dynamics is not a static phenomenon, such that its analysis requires to evaluate its spatial and temporal dimensions simultaneously. Moving from these premises, and drawing on the insights highlighted in Table 2, this sub-section presents the results concerning the agglomeration dynamics obtained by employing the space-time *K*-function approach.

Table 3 and 4 provide a summary of the key insights emerging from the analysis, while Appendix Figure A5 reports the plots of the estimated values of the $\widehat{D}_s(r,t)$ functional, the $\widehat{D}_s^0(r,t)$ functional, the 'standardised residuals' $\widehat{R}_s(r,t)$ versus $\widehat{K}_s^R(r)\widehat{K}_s^T(t)$, and the empirical frequency distribution of the sum of the differences between the space-time *K*-function and the product of the separate space and time *K*-functions in 99 simulations resulted from the Monte Carlo test.

The analysis of the estimated $\widehat{D}_s^0(r,t)$ functionals suggests that all Italian manufacturing two-digit sectors recorded a negative pick value at short distances and at one year lag. This finding can be interpreted as the presence of a spatial dispersion process that occurred almost constant over the entire period 2007-2012, but was more intense at the short spatial distances and during the year 2008. In other words, this dynamic process hit more intensively firms located within denser agglomerated areas, and during the first year of the Great Recession. The only exception is represented by the sector "25 – Manufacture of fabricated metal products", as it seems that the process of spatial dispersion increased over time.

However, our evidence does not point to the existence of statistically significant space-time interaction, as suggested by the analysis of the 'standardised residuals' – see the last column of Table 3 –, and the Monte Carlo-based tests – see Table 4.

[--- Table 3 near here ---]

[--- Table 4 near here ---]

5. CONCLUSIONS

The spatial and temporal evolution of firms' conglomerates is one of the key phenomenon occurring in many industrialised and emerging countries. Despite its relevance, only few studies have investigated the dynamics of agglomeration over both space and time. To the best of our knowledge, this is one of first papers that attempts to investigate empirically these processes by employing a battery of statistical methods, and using a large sample of geo-referenced, single-plant firms covering almost the entire manufacturing industry.

In particular, our contribution to this literature is threefold. First, we show that different statistical techniques, generally employed in economic geography and regional economics for studying spatial agglomeration phenomena, can lead to different results. This depends on two main issues, namely the MAUP, and the fact that the temporal dimension characterising agglomeration processes is not always taken into account adequately. Second, we find that the majority of Italian manufacturing sectors experienced a process of spatial dispersion during the period of the Great Recession. Finally, although our analysis did not detect any statistical evidences of space-time interaction, we observe that the process of spatial dispersion was more intense at the short spatial distances and during the first year of the Great Recession. In other words, we have been able to identify the role of the Great Recession in affecting the temporal profile of the spatial agglomeration processes.

Our analysis, and, particularly, the identification of space-time processes in agglomeration patterns, has relevant implications in terms of generation and diffusion of local externalities (Kang, 2010). In fact, the nature and the intensity of local externalities are likely to change as a consequence of spatial and temporal changes in the agglomeration dynamics. This is an aspect that should be taken into account seriously in the analysis of the effects of the agglomerative forces on firms' economic performance.

Of course, our analysis comes with some limitations. First, the period of analysis – i.e. only six years – is not long enough to justify the presence of statistically significant space-time interactions. Second, although our intent was to provide a broader picture of the agglomeration dynamics occurred in the Italian manufacturing industry, the use of industrial sectors at the two-digit level could not be appropriate for analysing these phenomena. All these limitations will be addressed in future developments of this research.

REFERENCES

Albert J. M., Casanova M. R. and Orts V. (2012) Spatial location patterns of Spanish manufacturing firms. *Papers in Regional Science* 91(1), 107–136.

Amrhein C. G. (1995) Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment and Planning A* 27(1), 105−119.

Arbia G. (1989) *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht: Kluwer.

Arbia G. (2001) Modelling the geography of economic activities on a continuous space. *Papers in Regional Science* 80(4), 411−424.

Arbia G. and Espa G. (1996) *Statistica economica territoriale*. Padua: CEDAM.

Arbia G., Espa G. and Quah D. (2008) A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics* 34(1), 81–103.

Arbia G., Espa G., Giuliani D. and Dickson M. M. (2014) Spatio-temporal clustering in the pharmaceutical and medical device manufacturing industry: A geographical micro-level analysis. *Regional Science and Urban Economics* 49(C), 298–304.

Arbia G., Espa G., Giuliani D. and Mazzitelli A. (2010) Detecting the existence of space-time clustering of firms. *Regional Science and Urban Economics* 40(5), 311–323.

Baldwin J. R., Beckstead D., Brown W. M. and Rigby D. L. (2008) Agglomeration and the geography of localization economies in Canada. *Regional Studies* 42(1), 117−132.

Becattini G. (1990) The Marshallian industrial district as a socioeconomic notion. In F. Pyke, G. Becattini and W. Sengenberger (Eds.), *Industrial districts and inter-firm cooperation in Italy* (pp. 37–51). Geneva: International Institute for Labour Studies (ILO).

Brusco S. (1982) The Emilian model: Productive decentralisation and social integration. *Cambridge Journal of Economics* 6(2), 167–184.

Cainelli G. and Ganau R. (2018) Distance-based agglomeration externalities and neighbouring firms' characteristics. *Regional Studies* 52(7), 922−933.

Cainelli G. and Lupi C. (2010) Does spatial proximity matter? Micro-evidence from Italy. In N. De Liso and R. Leoncini (Eds.), *Internationalization, technological change and the theory of the firm* (pp. 163−186). London: Routledge.

Cainelli G., Ganau R. and Iacobucci D. (2016) Do geographic concentration and vertically-related variety foster firm productivity? Micro-evidence from Italy. *Growth and Change* 47(2), 197–217.

Combes P. P. (2000) Economic structure and local growth: France, 1984-1993. *Journal of Urban Economics* 47, 329–355.

Diggle P. J., Chetwynd A. G., Haggkvist R. and Morris S. E. (1995) Second-order analysis of space time-clustering. *Statistical Methods in Medical Research* 42(2), 124–136.

Duranton G. and Overman H. G. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies* 72(4), 1077−1106.

Duschl M., Schimke A., Brenner T. and Luxen D. (2014) Firm growth and the spatial impact of geolocated external factors. *Journal of Economics and Statistics (Jahrbücher fuer Nationalökonomie und Statistik)* 234(2–3), 234−256.

Duschl M., Scholl T., Brenner T., Luxen D. and Raschke F. (2015) Industry-specific firm growth and agglomeration *Regional Studies* 49(11), 1822−1839.

Ellison G. and Glaeser E. L. (1997) Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy* 105(5), 889−927.

Eriksson R. H. (2011) Localized spillovers and knowledge flows: How does proximity influence the performance of plants?. *Economic Geography* 87(2), 127−152.

French N. P., McCarthy H. E., Diggle P. J. and Proudman C. J. (2005) Clustering of equine grass sickness cases in the United Kingdom: A study considering the effect of position-dependent reporting on the space-time K-function. *Epidemiology and Infection* 133(2), 343–348.

Gatrell A. C., Bailey T. C., Diggle P. J. and Rowlingson B. S. (1996) Spatial point pattern analysis and its applications in geographical epidemiology. *Transactions of the Institute of British Geographers* 21(1), 256–274.

Getis A. (1964) Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. *Annals of the Association of American Geographers* 54(3), 391–399.

Getis A. and Boots B. (1978) *Models of spatial processes*. Cambridge: Cambridge University Press.

Glaeser E., Kallal H. D., Scheinkman J. A. and Shleifer A. (1992) Growth in cities. *Journal of Political Economy* 100(6), 1126–1152.

Henderson J. V. (2003) Marshall's scale economies. *Journal of Urban Economics* 53(1), 1–28.

Henderson J. V., Squires T., Storeygard A. and Weil D. (2018) The global distribution of economic activity: Nature, history, and the role of trade. *The Quarterly Journal of Economics* 133(1), 357–406.

Henderson, V., Kuncoro, A. and Turner, M. (1995) Industrial Development in Cities. *Journal of Political Economy* 103, 1067–1090.

Kang H. (2010) Detecting agglomeration processes using space-time clustering analyses. *The Annals of Regional Science* 45(2), 291–311.

Klepper S. (1996) Entry, exit growth, and innovation over the product life-cycle. *The American Economic Review* 86(3), 562–583.

Klepper S. and Graddy E. (1990) The evolution of new industries and the determinants of market structure. *Rand Journal of Economics* 21(1), 27–44.

Krugman P. R. (1991) *Geography and trade*. Cambridge: The MIT Press.

Marcon E. and Puech F. (2003) Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography* 3(4),409–428.

Marcon E. and Puech F. (2010) Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography* 10(5), 745−762.

Marshall A. (1920) *Principles of economics*. London: Palgrave Classics in Economics.

Martin P., Mayer T. and Mayneris F. (2011) Spatial concentration and plant-level productivity in France. *Journal of Urban Economics* 69, 182–195.

Maurel F. and Sédillot B. (1999) A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics* 29(5), 575– 604.

Paci R. and Usai S. (2008) Agglomeration economies, spatial dependence and local industry growth. *Revue de Economie Industrielle* 123, 87–109.

Porter M. E. (1990) *The competitive advantage of nations*. New York: Free Press.

Ripley B. D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability* 13(2), 255–266.

Ripley B. D. (1977) Modelling spatial patterns. *Journal of the Royal Statistical Society, Series B* 39(2), 172–192.

Rosenthal S. S. and Strange W. C. (2003) Geography, industrial organization, and agglomeration. *The Review of Economics and Statistics* 85(2), 377−393.

Saxenian A. (1994) *Regional advantage. Culture and competition in Silicon Valley and Route 128*. Cambridge: Harvard University Press.

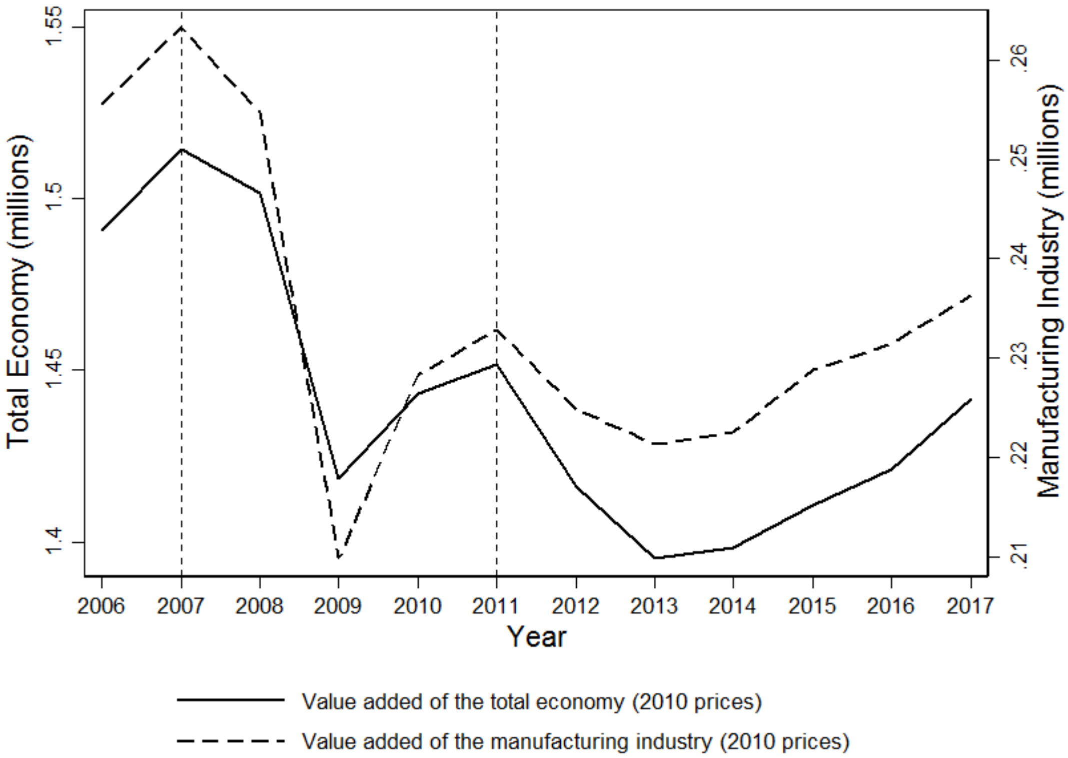
Scholl T. and Brenner T. (2016) Detecting spatial clustering using a firm-level cluster index. *Regional Studies* 50(6), 1054−1068.

Sorenson O. and Audia P. G. (2000) The social structure of entrepreneurial activity: Geographic concentration of footwear production in the United States, 1940-1989. *American Journal of Sociology* 106(2), 424−462.

Strange S. S. and Rosenthal W. (2001) The determinants of agglomeration. *Journal of Urban Economics* 50, 191-229.

Wong D. and Amrhein C. G. (1996) Research on the MAUP: Old wine in a new bottle or real breakthrough?. *Journal of Geographical Systems: Geographical Information, Analysis, Theory and Decision* 3(2), 73−76.

Figure 1: Temporal dynamics of Italian manufacturing value added.



Notes: Authors' elaboration on Istat data. Values are expressed in 2010 prices, and defined in millions of Euro.

Table 1: Ellison and Glaeser's (1997) index of geographic concentration in 2007 and 2012.

NACE Rev. 2	NUTS-2 Region			NUTS-3 Region			LLM		
	$\hat{\gamma}_s$		$\Delta\hat{\gamma}_s^{2007-2012}$	$\hat{\gamma}_s$		$\Delta\hat{\gamma}_s^{2007-2012}$	$\hat{\gamma}_s$		$\Delta\hat{\gamma}_s^{2007-2012}$
	2007	2012		2007	2012		2007	2012	
10	0.0233	0.0232	-0.10	0.0094	0.0076	-18.78	0.0041	0.0032	-20.83
11	0.0367	0.0367	0.22	0.0170	0.0178	5.18	0.0055	0.0057	3.09
13	0.0617	0.0637	3.26	0.0400	0.0401	0.13	0.0389	0.0390	0.23
14	0.0205	0.0241	17.56	0.0099	0.0098	-0.71	0.0071	0.0066	-6.58
15	0.1287	0.1275	-0.93	0.0472	0.0452	-4.15	0.0371	0.0347	-6.52
16	0.0137	0.0143	4.20	0.0100	0.0089	-11.28	0.0050	0.0045	-10.38
17	0.0027	0.0015	-45.57	0.0039	0.0029	-24.36	0.0027	0.0025	-7.62
18	0.0061	0.0026	-56.64	0.0092	0.0060	-34.30	0.0092	0.0061	-33.68
19	0.0281	0.0482	71.53	0.0073	0.0075	2.85	-0.0005	0.0036	0.41
20	0.0319	0.0296	-7.22	0.0188	0.0213	12.98	0.0162	0.0195	20.47
21	0.0857	0.0978	14.07	0.0748	0.0764	2.20	0.0851	0.0898	5.52
22	0.0078	0.0089	14.48	0.0036	0.0033	-9.16	0.0027	0.0028	3.17
23	0.0209	0.0179	-14.33	0.0094	0.0069	-27.25	0.0075	0.0061	-18.55
24	0.0344	0.0330	-4.27	0.0153	0.0177	15.80	0.0062	0.0074	20.98
25	0.0025	0.0027	5.88	0.0021	0.0021	-2.21	0.0012	0.0011	-6.31
26	0.0166	0.0142	-14.42	0.0239	0.0176	-26.25	0.0225	0.0166	-26.16
27	0.0090	0.0053	-41.21	0.0050	0.0036	-26.94	0.0046	0.0026	-44.69
28	0.0137	0.0137	-0.29	0.0032	0.0033	1.86	0.0020	0.0020	0.89
29	0.0465	0.0378	-18.82	0.0244	0.0140	-42.61	0.0147	0.0116	-20.88
30	0.0329	0.0367	11.66	0.0168	0.0233	38.58	0.0114	0.0157	37.93
31	0.0640	0.0552	-13.73	0.0310	0.0308	-0.76	0.0219	0.0218	-0.40
32	0.0263	0.0219	-16.63	0.0204	0.0191	-6.36	0.0172	0.0161	-6.24
33	0.0092	0.0110	19.16	0.0058	0.0053	-8.48	0.0040	0.0033	-16.27

Notes: Authors' elaboration on *AIDA* data. The reference for the calculation of the Ellison and Glaeser' (1997) index is the *AIDA* sample of firms. Variations of the sector-specific indexes between the years 2007 and 2012 are defined in percentage terms.

Table 2: Summary results of the estimated spatial *K*- and *M*-functions for the years 2007 and 2012.

Functional	$\widehat{K}_s(r)$				
Year	2007		2012		$\widehat{K}_s^{2007}$ vs. $\widehat{K}_s^{2012}$
Pattern	Concentration / Dispersion	$\widehat{K}_s(r) > \pi r^2$	Concentration / Dispersion	$\widehat{K}_s(r) > \pi r^2$	
NACE Rev. 2					
10	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$ up to 40 km, then $\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$
11	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
13	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances, but in [190, 200] and [220, 230] km
14	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
15	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$ at all distances
16	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
17	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ up to 40 km, then $\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$
18	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
19	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ up to 70 km and in [130, 190] km; $\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$ in [70, 130] and [190-240] km
20	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
21	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
22	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
23	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
24	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$ in [0, 20], [40, 80], [100, 110] km; $\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ in [20, 40], [80, 100], from 110 km
25	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
26	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
27	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
28	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
29	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ up to 160 km, then $\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$
30	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} < \widehat{K}_s^{2012}$ at all distances, but in [0, 10] and [20, 30] km
31	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances, but in [0, 30] and [40, 50] km
32	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances
33	Concentration at all distances	Yes	Concentration at all distances	Yes	$\widehat{K}_s^{2007} > \widehat{K}_s^{2012}$ at all distances

Notes: Authors' elaboration on *AIDA* data. Sector-specific comparisons of the estimated empirical *K* values between the years 2007 and 2012 are evaluated considering sector- and year-specific mean values of the estimated empirical *K* values averaged over distance bands of 10 kilometres each in the interval [0, 250] km. Plots of the estimated spatial *K*- and *M*-functions are reported by two-digit manufacturing sector and year in Appendix Figure A4.

Table 3: Summary results of the estimated space-time *K*-function over the period 2007-2012.

Functional / Test	Interpretation	$\widehat{D}_s^0(r,t)$	$\widehat{R}_s(r,t)$ vs. $\widehat{K}_s^R(r)\widehat{K}_s^T(t)$
NACE Rev. 2			
10	A	Negative peak at short distance and one year lag	No significant interaction
11	A	Negative peak at short distance and one year lag	No significant interaction
13	A	Negative peak at short distance and one year lag	No significant interaction
14	A	Negative peak at short distance and one year lag	No significant interaction
15	A	Negative peak at short distance and one year lag	No significant interaction
16	A	Negative peak at short distance and one year lag	No significant interaction
17	A	Negative peak at short distance and one year lag	No significant interaction
18	A	Negative peak at short distance and one year lag	No significant interaction
19	A	Negative peak at short distance and one year lag	No significant interaction
20	A	Negative peak at short distance and one year lag	No significant interaction
21	A	Negative peak at short distance and one year lag	No significant interaction
22	A	Negative peak at short distance and one year lag	No significant interaction
23	A	Negative peak at short distance and one year lag	No significant interaction
24	A	Negative peak at short distance and one year lag	No significant interaction
25	B	Negative peak at short distance and one year lag	No significant interaction
26	A	Negative peak at short distance and one year lag	No significant interaction
27	A	Negative peak at short distance and one year lag	No significant interaction
28	A	Negative peak at short distance and one year lag	No significant interaction
29	A	Negative peak at short distance and one year lag	No significant interaction
30	A	Negative peak at short distance and one year lag	No significant interaction
31	A	Negative peak at short distance and one year lag	No significant interaction
32	A	Negative peak at short distance and one year lag	No significant interaction
33	A	Negative peak at short distance and one year lag	No significant interaction

Notes: Authors' elaboration on *AIDA* data. "A" denotes a process of spatial dispersion that has been almost constant over the entire period 2007-2012, but more intense at short spatial distance and during the first year of the Great Recession (i.e. the year 2008). "B" denotes a process of spatial dispersion that has increased during the period 2007-2012. Plots of the estimated $\widehat{D}_s^0(r,t)$ functionals, and of the 'standardised residuals' are reported by two-digit manufacturing sector in Appendix Figure A5.

Table 4: Monte Carlo test of space-time interaction.

NACE Rev. 2	No. of Observations	Monte Carlo Test	No. of Simulations	P-value
10	40,355	33	99	0.67
11	5,843	54	99	0.46
13	25,397	30	99	0.70
14	37,173	7	99	0.93
15	24,967	24	99	0.76
16	21,139	52	99	0.48
17	9,809	26	99	0.74
18	24,519	80	99	0.20
19	1,002	41	99	0.59
20	14,339	34	99	0.66
21	2,266	49	99	0.51
22	26,663	53	99	0.47
23	31,198	65	99	0.35
24	9,221	33	99	0.67
25	127,249	81	99	0.19
26	22,328	41	99	0.59
27	25,684	36	99	0.64
28	69,472	37	99	0.63
29	7,606	28	99	0.72
30	9,623	53	99	0.47
31	29,932	36	99	0.64
32	24,345	38	99	0.62
33	24,090	6	99	0.94

Notes: Authors' elaboration on *AIDA* data. The plots of the empirical frequency distribution of the sum of the differences between the space-time *K*-function and the product of the separate space and time *K*-functions resulted from the Monte Carlo tests are reported by two-digit manufacturing sector in Appendix Figure A5.

APPENDIX

Table A1: Sample distribution by two-digit manufacturing sector.

NACE Rev. 2 Classification		Firms	
Code	Description	No.	%
10	Manufacture of food products	10,318	6.92
11	Manufacture of beverages	1,420	0.95
13	Manufacture of textiles	6,100	4.09
14	Manufacture of wearing apparel	9,960	6.68
15	Manufacture of leather and related products	6,202	4.16
16	Manufacture of wood, wood and cook products (except furniture), straw articles, plaiting materials	5,008	3.36
17	Manufacture of paper and paper products	2,291	1.54
18	Printing and reproduction of recorded media	5,783	3.88
19	Manufacture of coke and refined petroleum products	242	0.16
20	Manufacture of chemicals and chemical products	3,428	2.30
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	547	0.37
22	Manufacture of rubber and plastic products	6,261	4.20
23	Manufacture of other non-metallic mineral products	7,482	5.02
24	Manufacture of basic metals	2,113	1.42
25	Manufacture of fabricated metal products, except machinery and equipment	29,904	20.05
26	Manufacture of computer, electronic and optical products	5,437	3.65
27	Manufacture of electrical equipment	6,308	4.23
28	Manufacture of machinery and equipment N.E.C.	16,398	11.00
29	Manufacture of motor vehicles, trailers and semi-trailers	1,896	1.27
30	Manufacture of other transport equipment	2,619	1.76
31	Manufacture of furniture	7,342	4.92
32	Other manufacturing	5,771	3.87
33	Repair and installation of machinery and equipment	6,305	4.23
Total		149,135	100.00

Notes: Authors' elaboration on *AIDA* data.

Table A2: Sample distribution by two-digit manufacturing sector and year.

NACE Rev. 2	Number of Firms						Percentage Change 2007-2012
	2007	2008	2009	2010	2011	2012	
10	6,390	6,371	6,458	6,630	7,253	7,253	13.51
11	951	930	941	973	1,023	1,025	7.78
13	4,329	4,330	4,221	4,175	4,254	4,088	-5.57
14	6,371	6,405	6,246	6,084	6,198	5,869	-7.88
15	4,105	4,150	4,134	4,117	4,297	4,164	1.44
16	3,345	3,482	3,476	3,503	3,700	3,633	8.61
17	1,622	1,611	1,611	1,638	1,695	1,632	0.62
18	4,104	4,135	4,033	3,992	4,166	4,089	-0.37
19	167	163	160	172	178	162	-2.99
20	2,344	2,345	2,343	2,391	2,480	2,436	3.92
21	390	374	371	373	375	383	-1.79
22	4,358	4,418	4,402	4,424	4,606	4,455	2.23
23	5,101	5,119	5,105	5,192	5,436	5,245	2.82
24	1,569	1,534	1,503	1,516	1,575	1,524	-2.87
25	20,261	20,885	21,067	21,225	22,151	21,660	6.90
26	3,815	3,775	3,671	3,688	3,768	3,611	-5.35
27	4,106	4,196	4,208	4,305	4,511	4,358	6.14
28	11,386	11,455	11,497	11,555	11,983	11,596	1.84
29	1,268	1,278	1,243	1,264	1,308	1,245	-1.81
30	1,514	1,632	1,586	1,651	1,682	1,558	2.91
31	4,898	4,959	5,001	4,955	5,158	4,961	1.29
32	4,028	4,079	4,061	4,015	4,127	4,035	0.17
33	3,327	3,540	3,874	4,187	4,589	4,573	37.45
Total	99,749	101,166	101,212	102,025	106,513	103,555	3.82

Notes: Authors' elaboration on *AIDA* data.

Table A3: Sample distribution by two-digit manufacturing sector and NUTS-1 geographic area.

NACE Rev. 2	North West		North East		Centre		South	
	No.	%	No.	%	No.	%	No.	%
10	2,263	4.55	2,356	5.86	2,259	6.83	3,440	13.15
11	335	0.67	388	0.97	214	0.65	483	1.85
13	2,599	5.23	882	2.19	1,956	5.91	663	2.54
14	2,002	4.03	2,340	5.82	2,714	8.20	2,904	11.10
15	575	1.16	1,267	3.15	2,852	8.62	1,508	5.77
16	1,201	2.42	1,547	3.85	1,145	3.46	1,115	4.26
17	740	1.49	556	1.38	607	1.83	388	1.48
18	2,068	4.16	1,337	3.33	1,493	4.51	885	3.38
19	67	0.13	31	0.08	55	0.17	89	0.34
20	1,513	3.04	807	2.01	645	1.95	463	1.77
21	264	0.53	65	0.16	156	0.47	62	0.24
22	2,734	5.50	1,606	4.00	1,041	3.15	880	3.37
23	1,611	3.24	2,094	5.21	1,911	5.77	1,866	7.14
24	1,104	2.22	450	1.12	299	0.90	260	0.99
25	11,656	23.45	8,759	21.79	5,051	15.26	4,438	16.97
26	2,288	4.60	1,211	3.01	1,291	3.90	647	2.47
27	2,634	5.30	1,893	4.71	1,041	3.15	740	2.83
28	6,969	14.02	5,851	14.56	2,091	6.32	1,487	5.69
29	797	1.60	499	1.24	300	0.91	300	1.15
30	774	1.56	547	1.36	833	2.52	465	1.78
31	1,559	3.14	2,586	6.43	1,985	6.00	1,212	4.63
32	1,756	3.53	1,568	3.90	1,701	5.14	746	2.85
33	2,191	4.41	1,553	3.86	1,451	4.38	1,110	4.24
Total	49,700	100.00	40,193	100.00	33,091	100.00	26,151	100.00

Notes: Authors' elaboration on *AIDA* data. Percentage values are defined on column totals.

Table A4: Average size of firms in 2007 and 2012.

NACE Rev. 2	Average Size of Firms		Percentage Change 2007-2012
	2007	2012	
10	14.20	13.12	-7.62
11	10.34	8.96	-13.29
13	16.95	14.83	-12.52
14	11.50	10.34	-10.12
15	14.92	14.11	-5.40
16	12.35	10.19	-17.47
17	18.32	17.96	-1.97
18	10.25	8.90	-13.11
19	11.61	14.43	24.34
20	16.63	16.46	-1.02
21	48.01	50.77	5.76
22	18.07	17.18	-4.90
23	14.02	11.16	-20.41
24	30.08	30.00	-0.27
25	14.38	12.76	-11.27
26	12.22	12.06	-1.32
27	17.05	15.08	-11.56
28	17.98	16.78	-6.67
29	41.49	36.07	-13.06
30	15.57	12.72	-18.34
31	14.48	12.16	-16.04
32	11.09	9.51	-14.31
33	9.11	7.61	-16.46

Notes: Authors' elaboration on *AIDA* data. The average firm size is defined as the number of employees per firm for each two-digit sector.

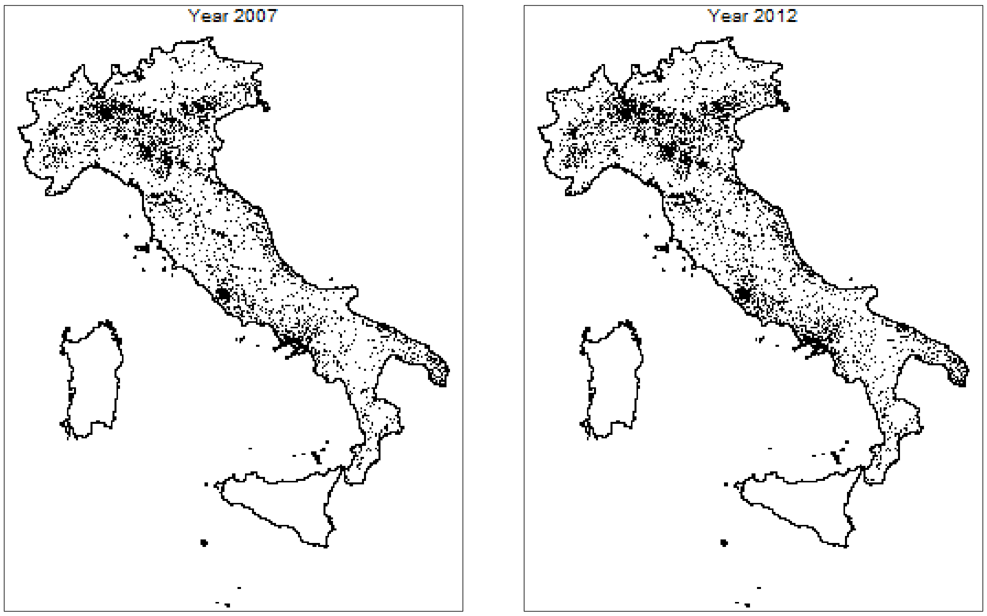
Table A5: Survival rate over the period 2007-2012.

Observed in Year	Surviving in Year				
	2008	2009	2010	2011	2012
2007	89.43	81.72	74.71	73.62	68.87
2008	...	88.08	79.44	77.65	72.31
2009	...	...	86.45	83.30	77.37
2010	...	...	...	89.00	81.91
2011	...	...	...	...	87.10

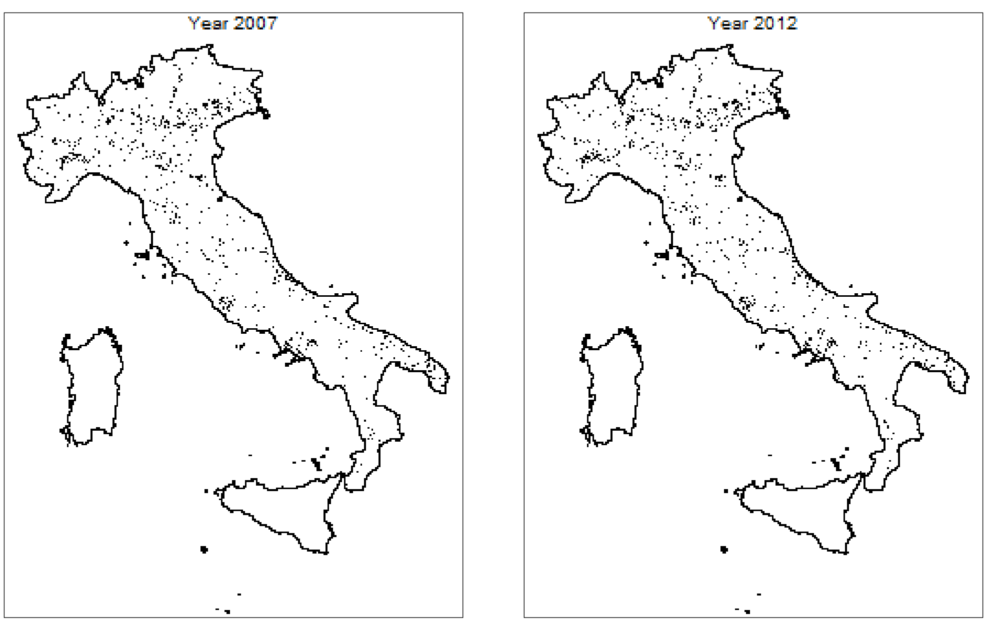
Notes: Authors' elaboration on *AIDA* data. The table reports percentage values. The survival rate over the period $t, t+n$ is calculated as the share of firms observed at time t and survived at time $t+n$ over firms observed at time t.

Figure A1: Spatial distribution of *AIDA* sample firms by two-digit manufacturing sector in 2007 and 2012.

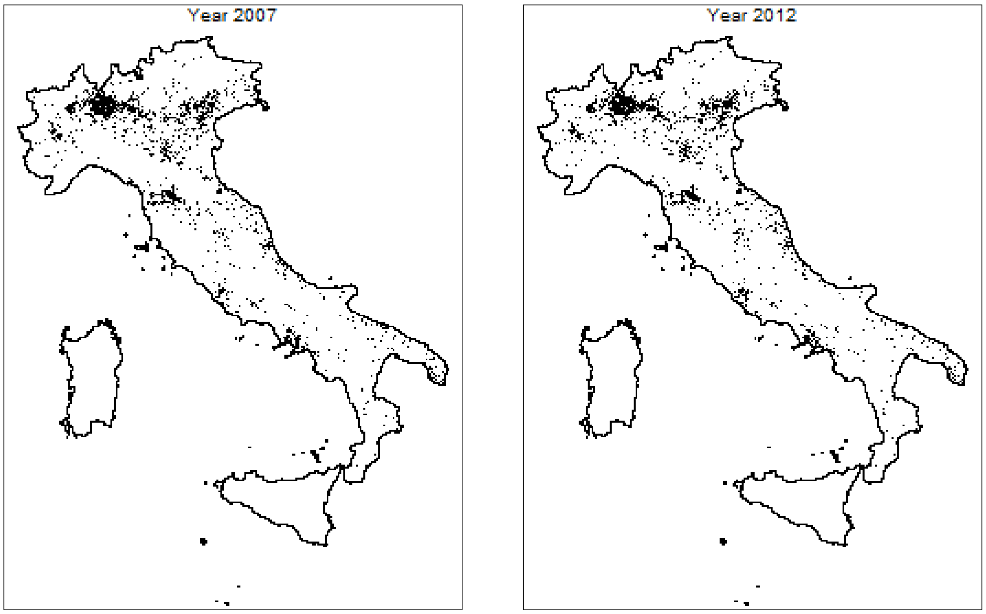
Sector 10 – Manufacture of food products



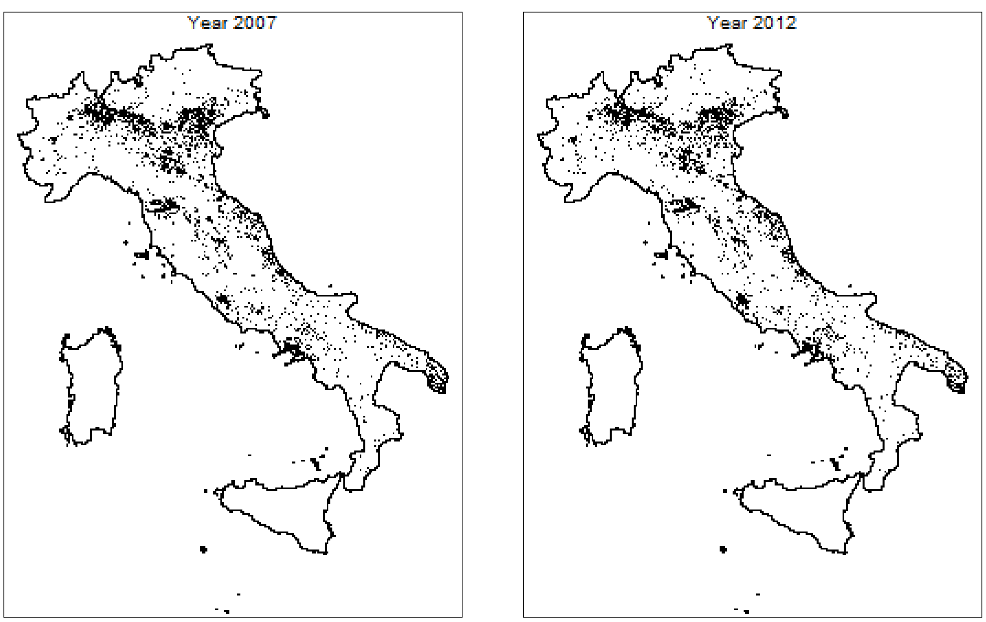
Sector 11 – Manufacture of beverages



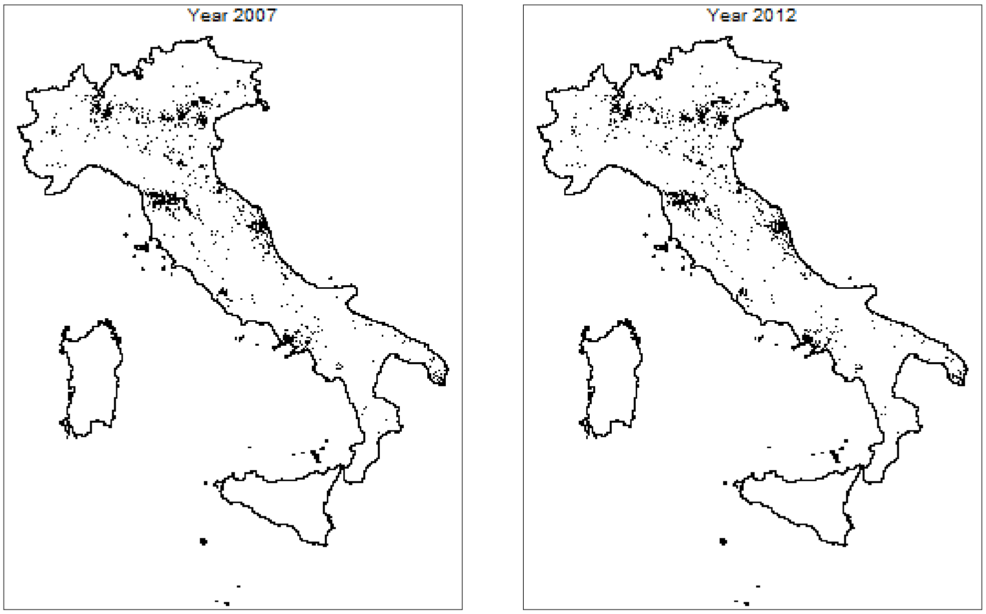
Sector 13 – Manufacture of textiles



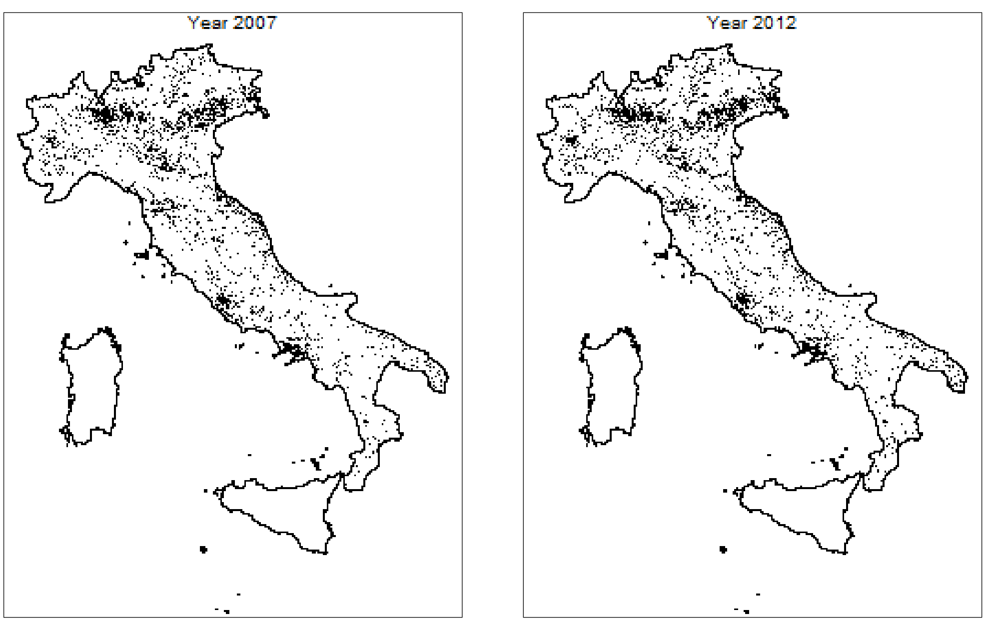
Sector 14 – Manufacture of wearing apparel



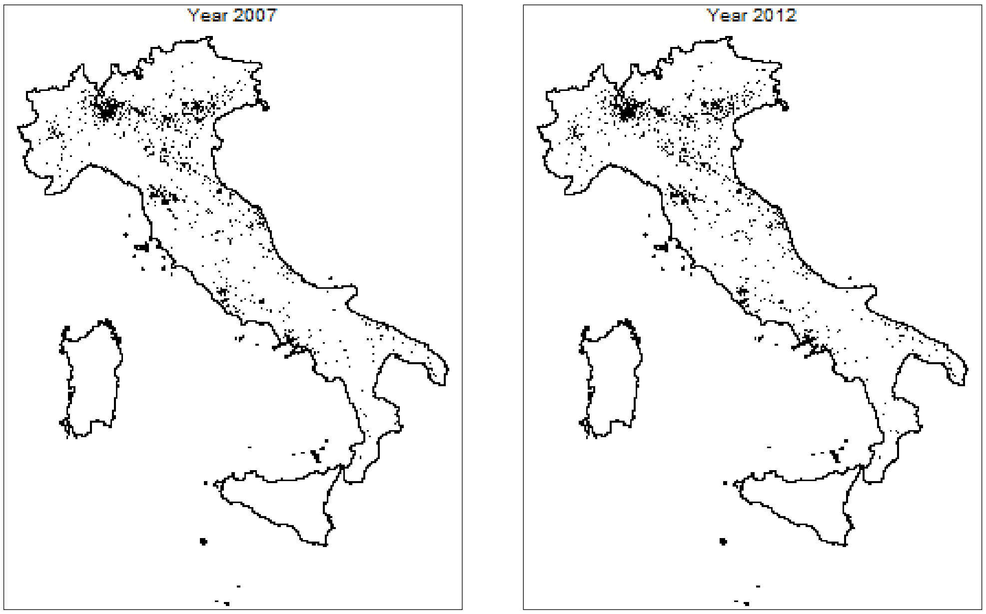
Sector 15 – Manufacture of leather and related products



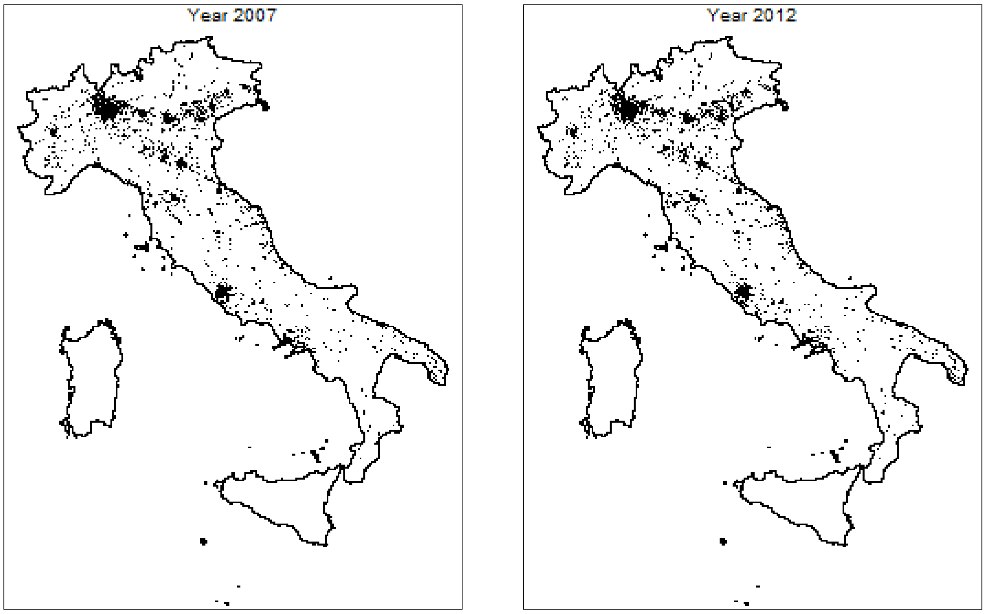
Sector 16 – Manufacture of wood, wood and cook products (except furniture), straw articles, plaiting materials



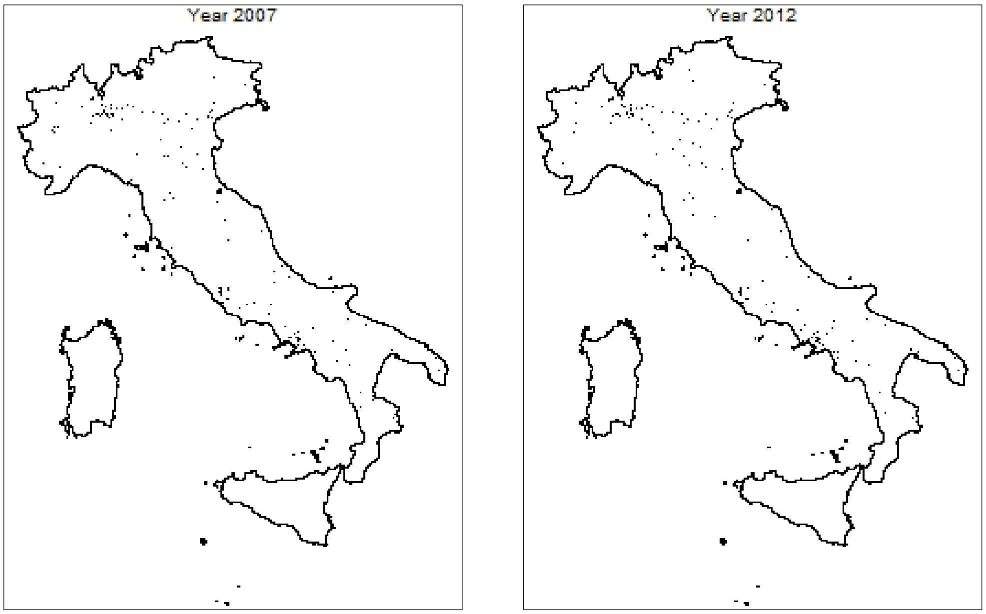
Sector 17 – Manufacture of paper and paper products



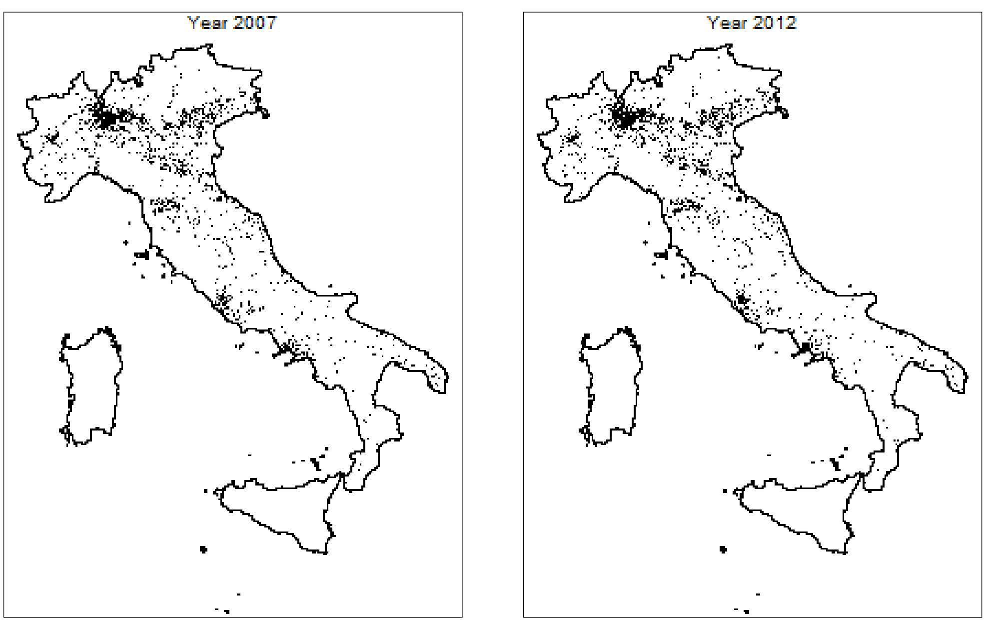
Sector 18 – Printing and reproduction of recorded media



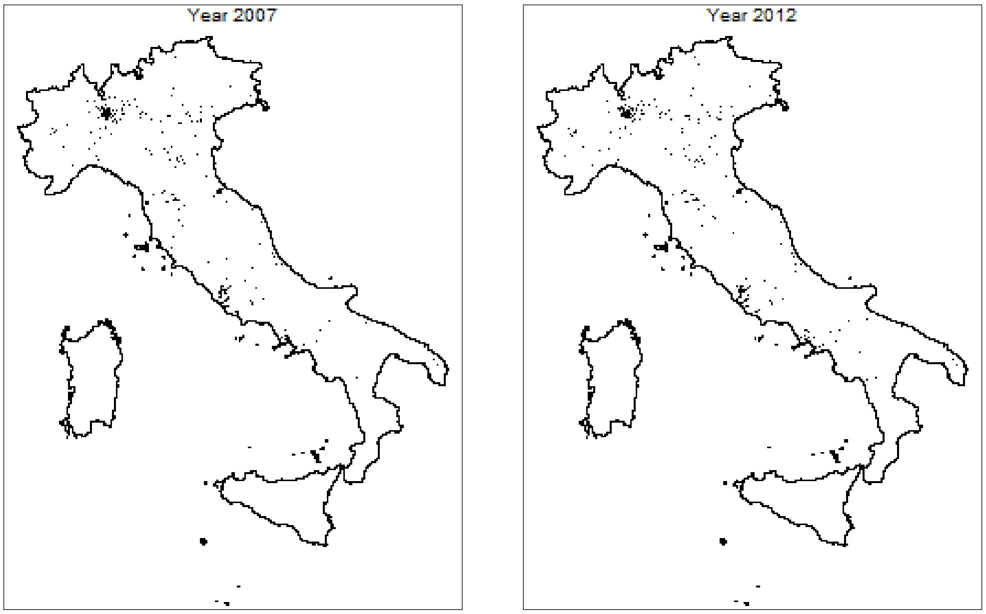
Sector 19 – Manufacture of coke and refined petroleum products



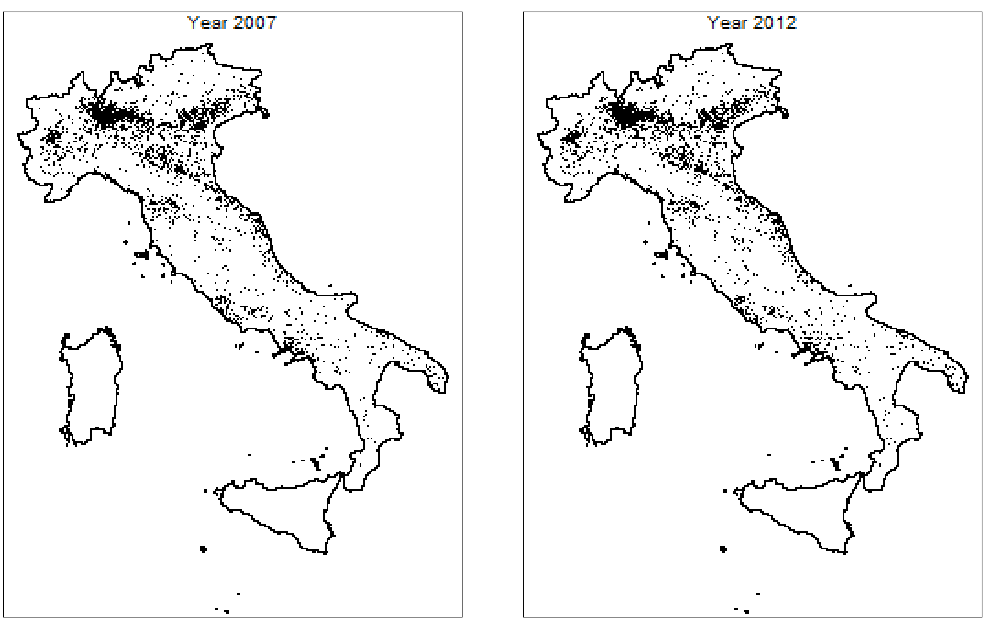
Sector 20 – Manufacture of chemicals and chemical products



Sector 21 – Manufacture of basic pharmaceutical products and pharmaceutical preparations

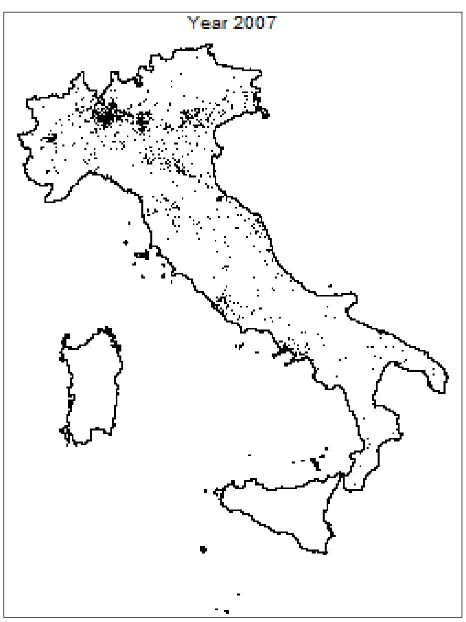


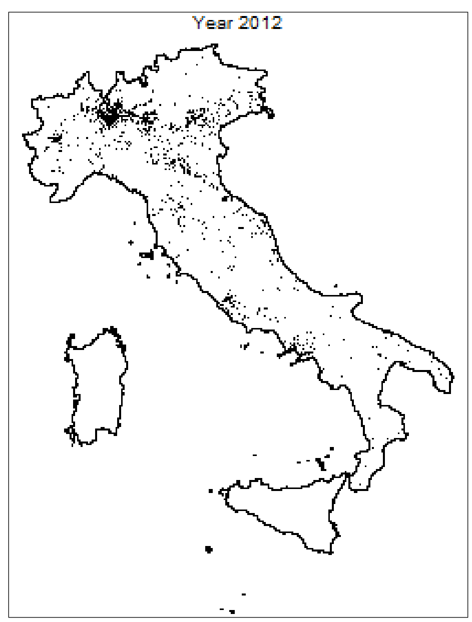
Sector 22 – Manufacture of rubber and plastic products



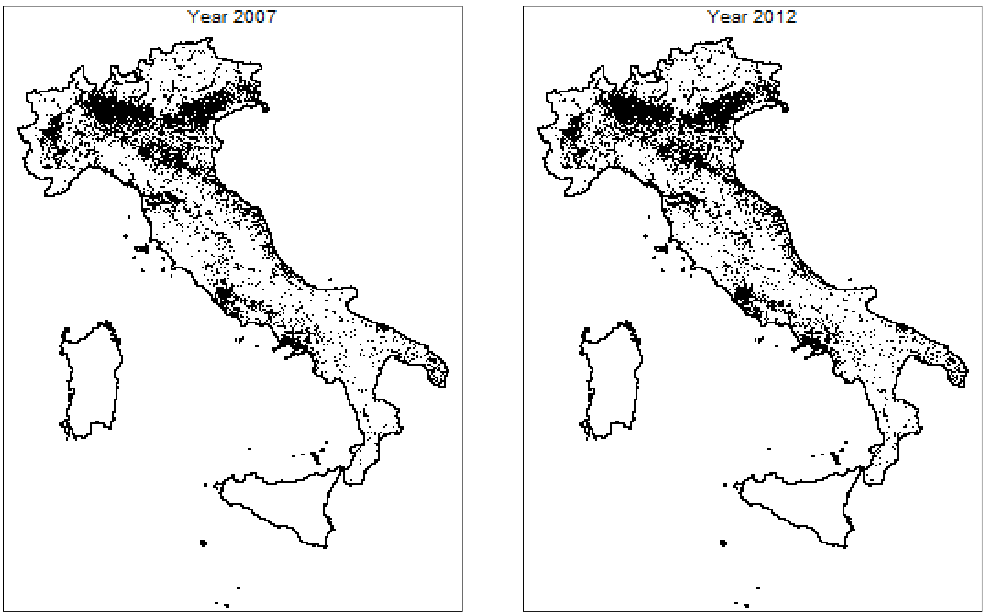
Sector 23 – Manufacture of other non-metallic mineral products

Sector 24 – Manufacture of basic metals

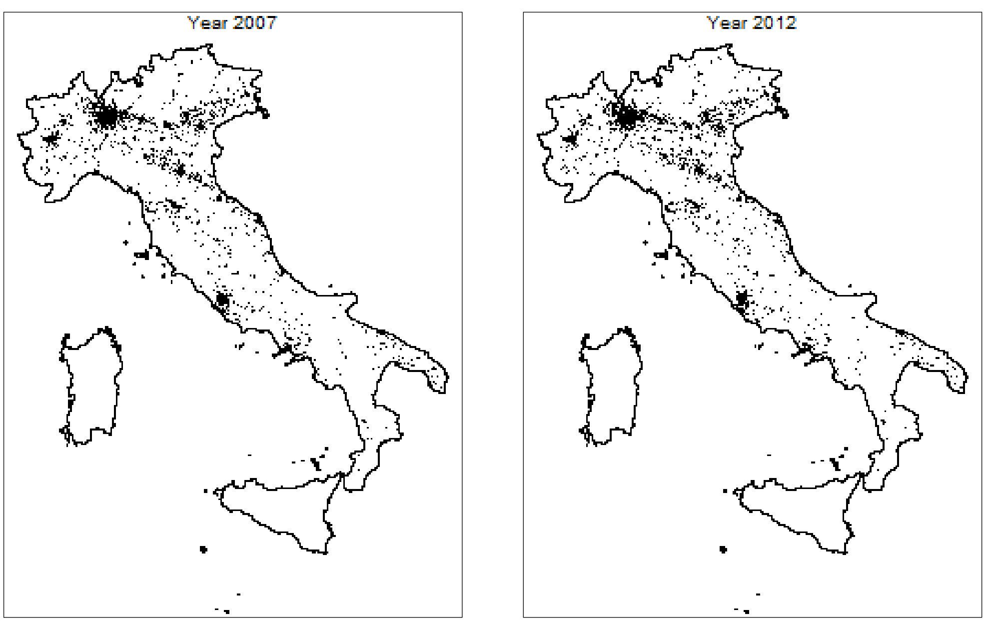




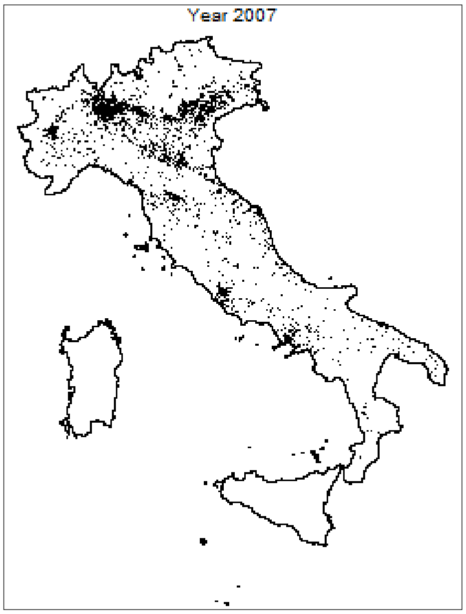
Sector 25 – Manufacture of fabricated metal products, except machinery and equipment

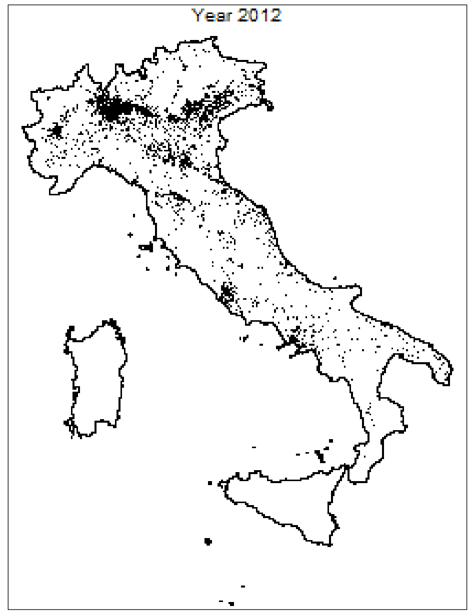


Sector 26 – Manufacture of computer, electronic and optical products



Sector 27 – Manufacture of electrical equipment

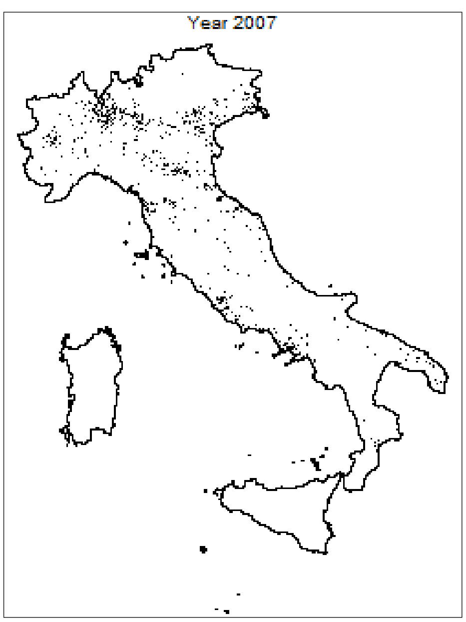


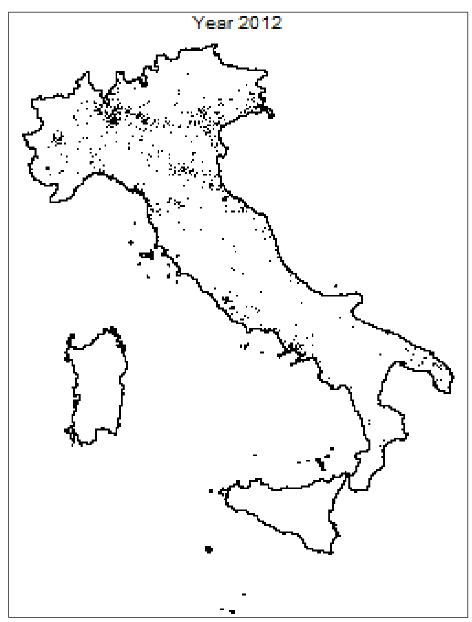


Sector 28 – Manufacture of machinery and equipment N.E.C.

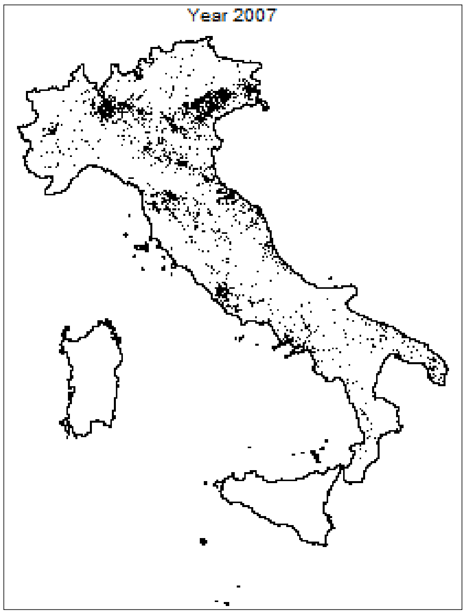
Sector 29 – Manufacture of motor vehicles, trailers and semi-trailers

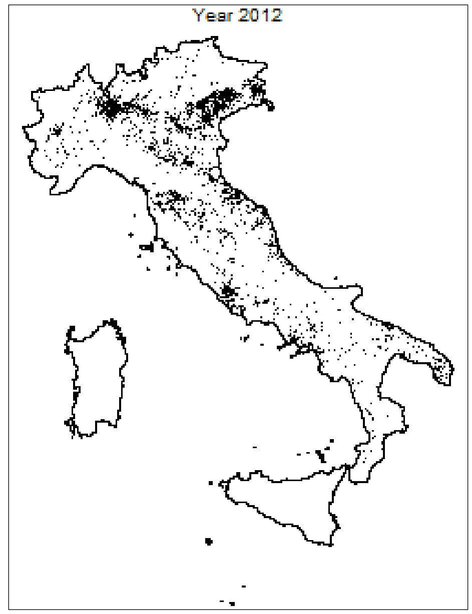
Sector 30 – Manufacture of other transport equipment



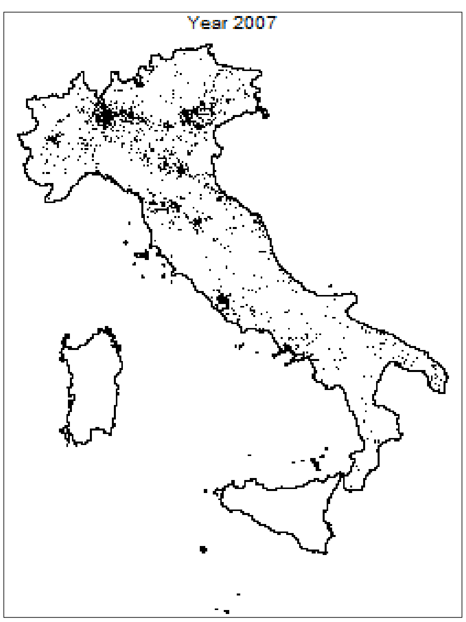


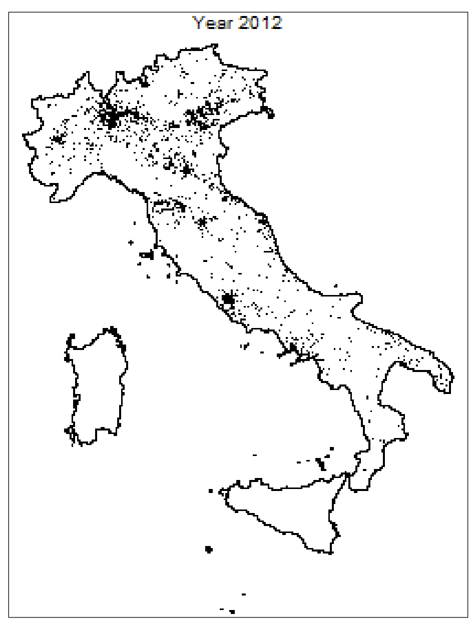
Sector 31 – Manufacture of furniture



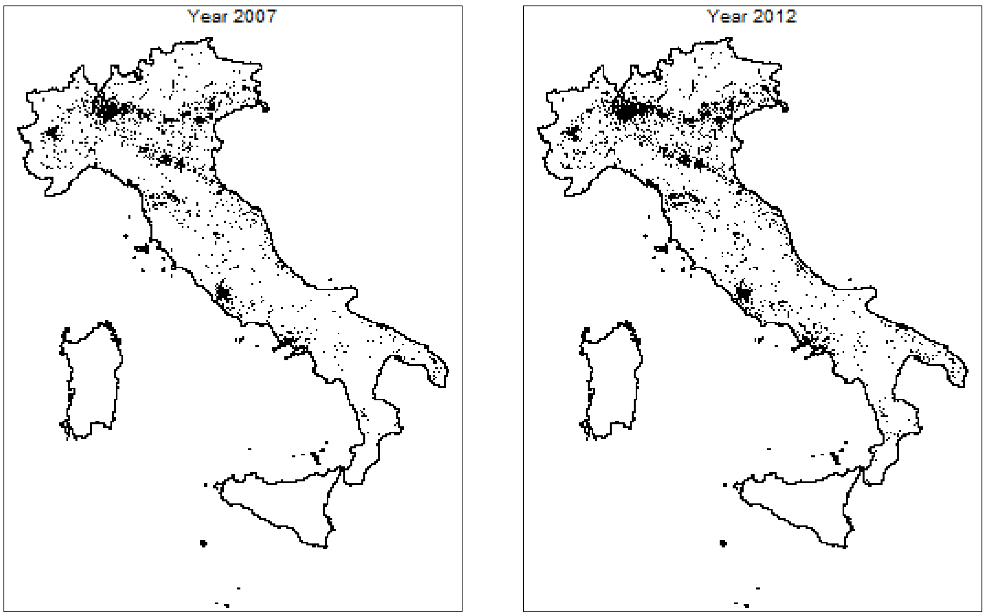


Sector 32 – Other manufacturing



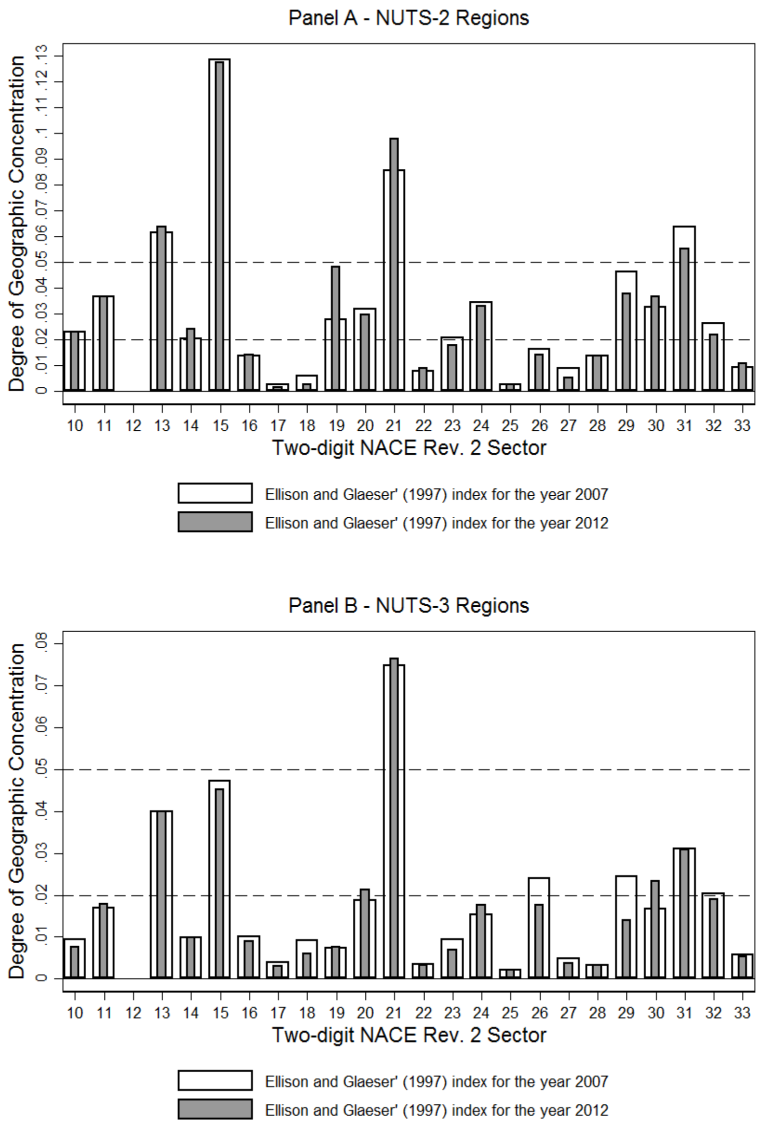


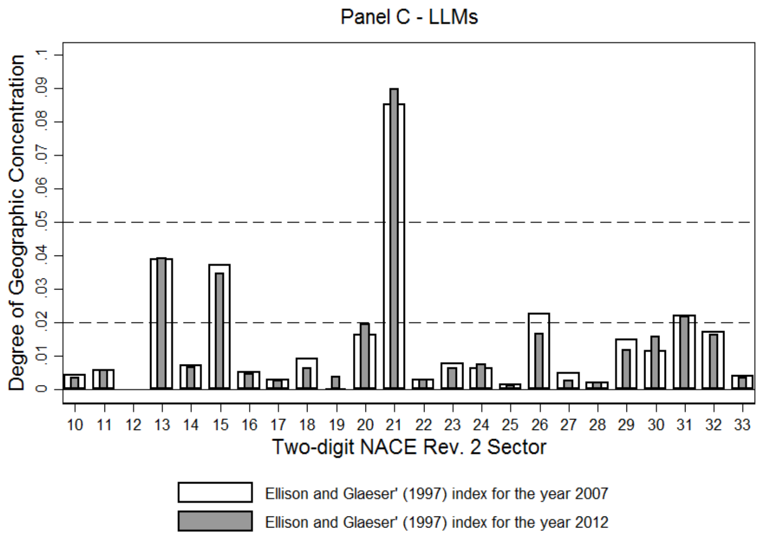
Sector 33 – Repair and installation of machinery and equipment



Notes: Authors' elaboration on *AIDA* data. The maps report the spatial distribution of the *AIDA* sample firms by two-digit manufacturing sector in the years 2007 and 2012.

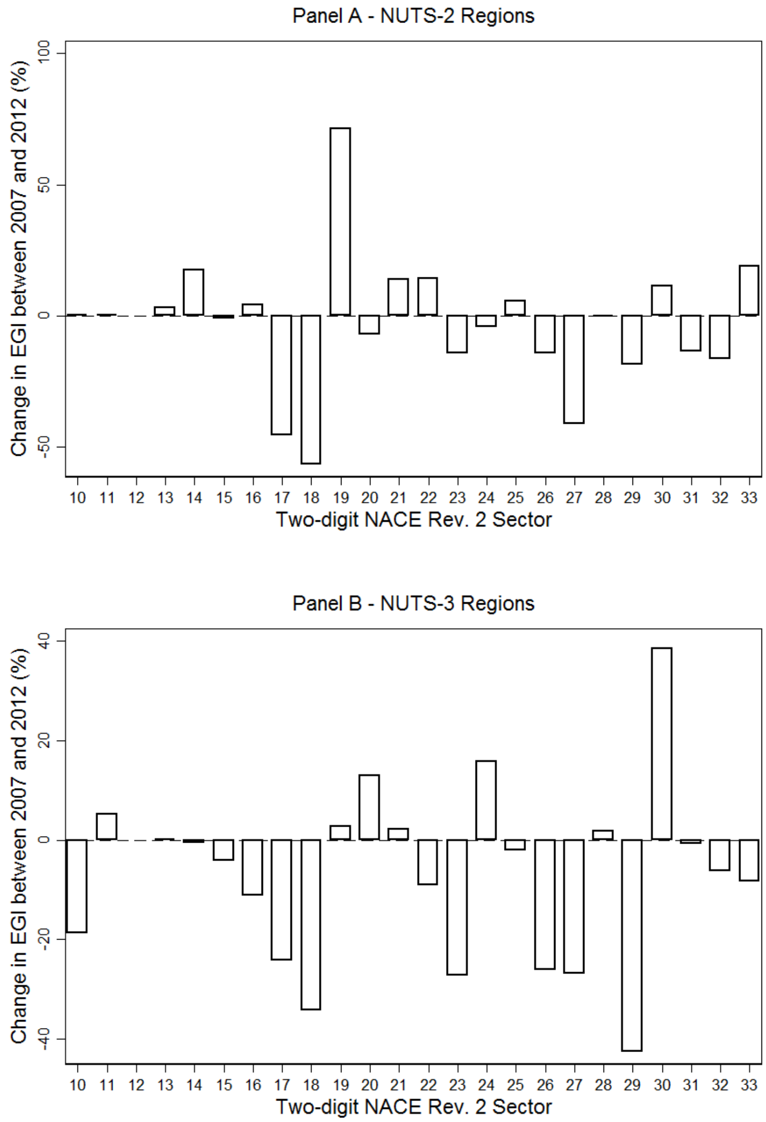
Figure A2: Ellison and Glaeser's (1997) index of geographic concentration in 2007 and 2012.

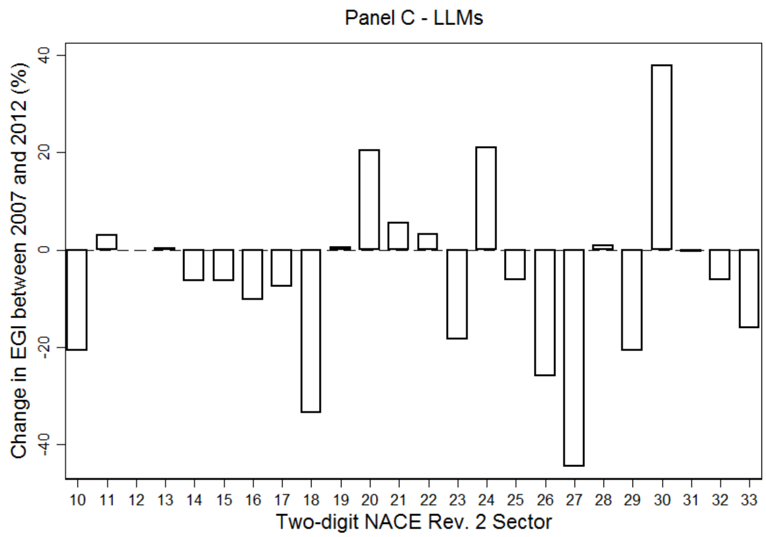




Notes: Authors' elaboration on *AIDA* data. Values refer to those reported in Table 1. Panel A plots the Ellison and Glaeser's (1997) index defined at the geographic NUTS-2 level. Panel B plots the Ellison and Glaeser's (1997) index defined at the geographic NUTS-3 level. Panel C plots the Ellison and Glaeser's (1997) index defined at the geographic level of the LLM. The dashed reference lines correspond to the threshold values of 0.02 and 0.05 identified by Ellison and Glaeser (1997) to classify the degree of geographic concentration of industries.

Figure A3: Variation of the Ellison and Glaeser's (1997) index of geographic concentration.



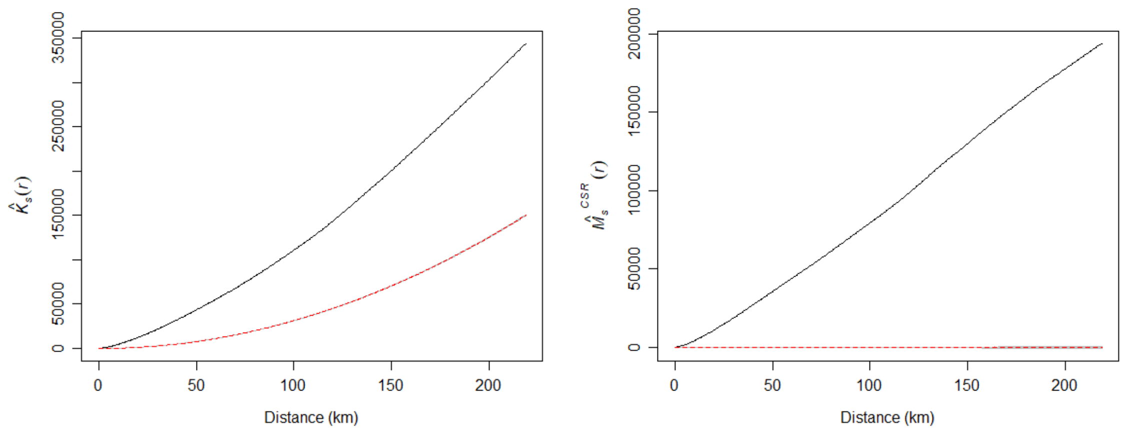


Notes: Authors' elaboration on *AIDA* data. Values refer to those reported in Table 1. Panel A plots the percentage variations between the years 2007 and 2012 of the Ellison and Glaeser's (1997) index defined at the geographic NUTS-2 level. Panel B plots the percentage variations between the years 2007 and 2012 of the Ellison and Glaeser's (1997) index defined at the geographic NUTS-3 level. Panel C plots the percentage variations between the years 2007 and 2012 of the Ellison and Glaeser's (1997) index defined at the geographic level of the LLM.

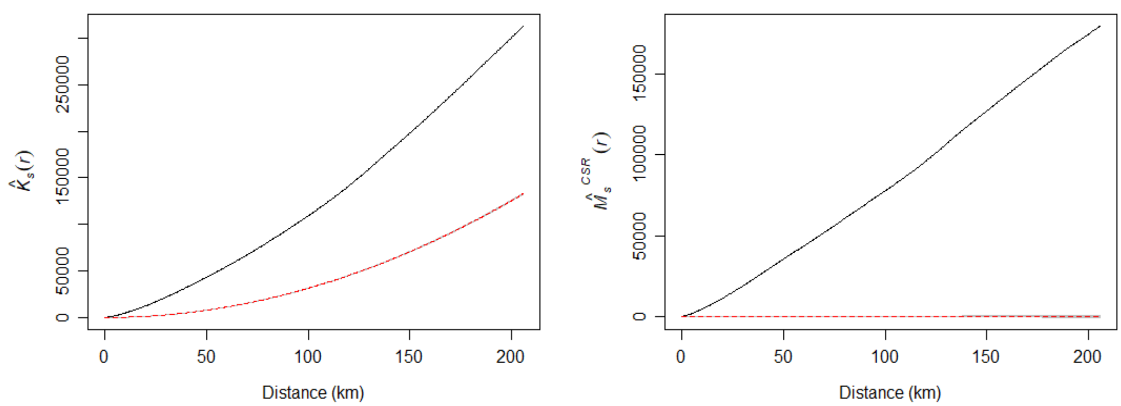
Figure A4: Results of the spatial *K*-function by two-digit manufacturing sector in 2007 and 2012.

Sector 10 – Manufacture of food products

Year 2007

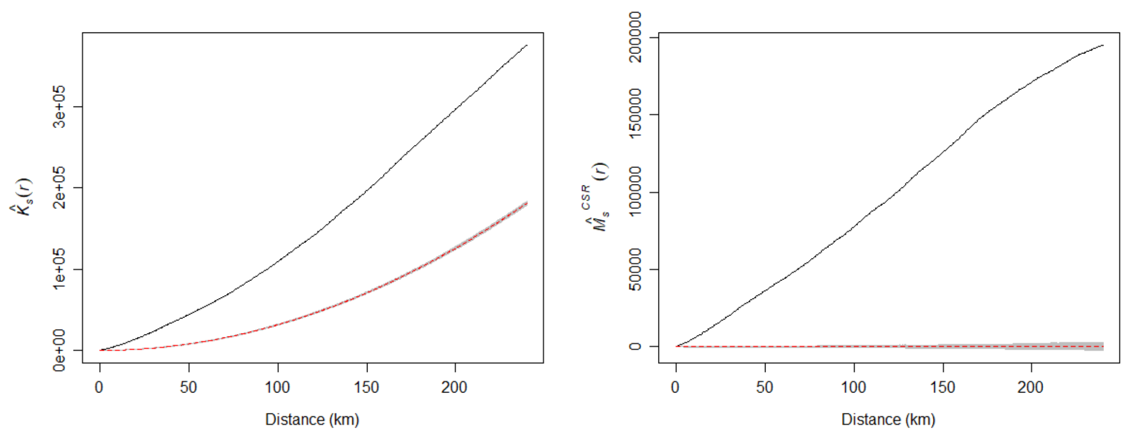


Year 2012

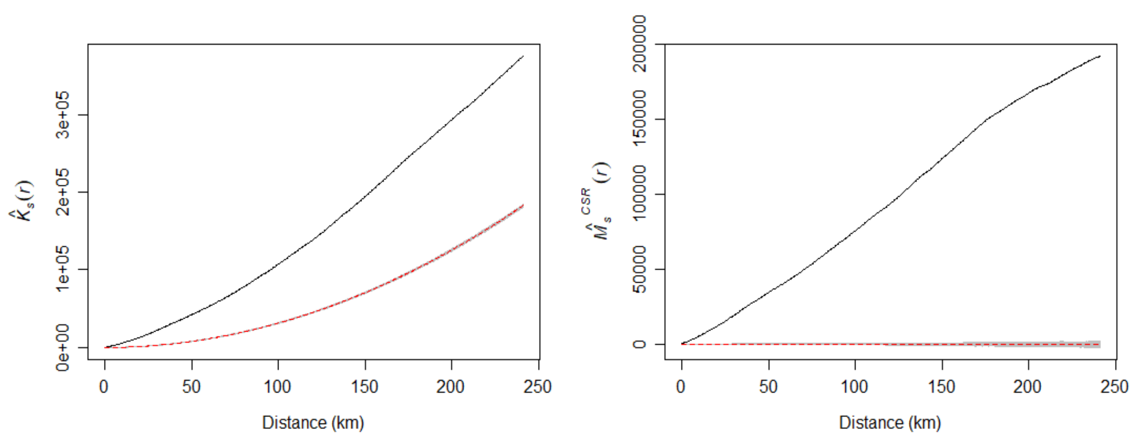


Sector 11 – Manufacture of beverages

Year 2007

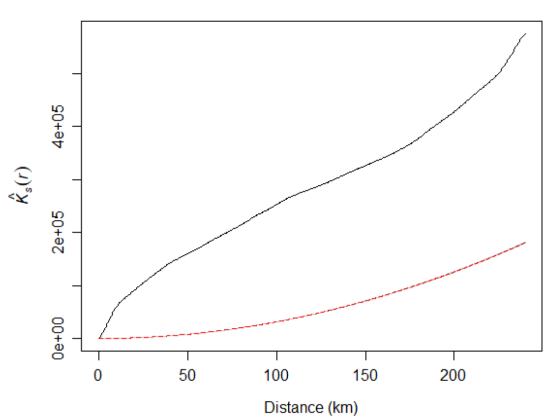


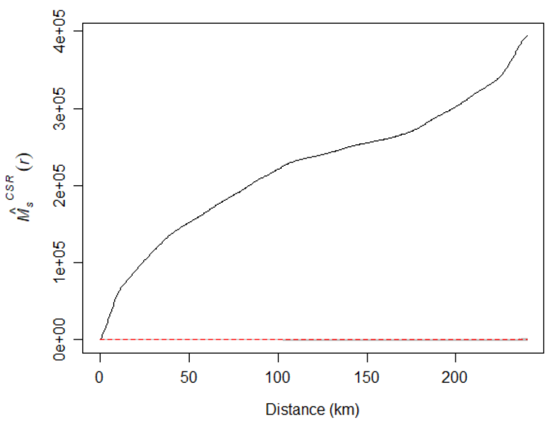
Year 2012



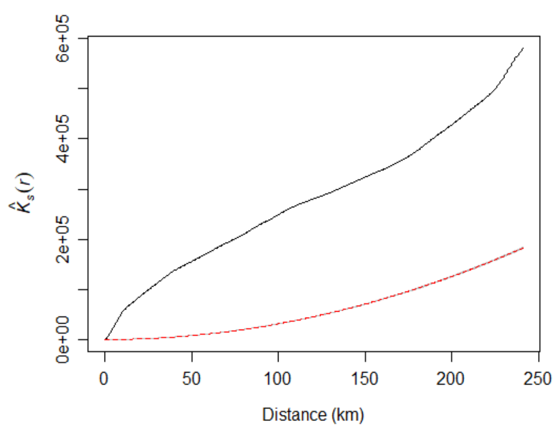
Sector 13 – Manufacture of textiles

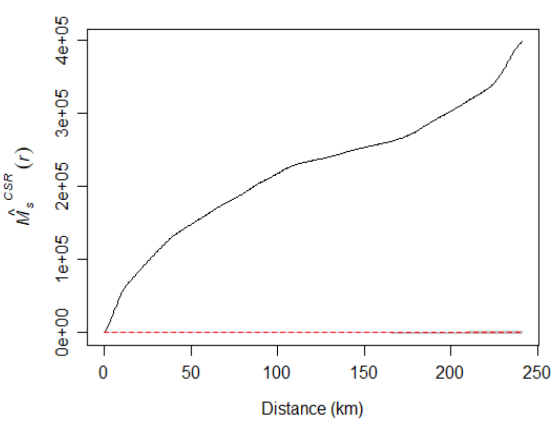
Year 2007





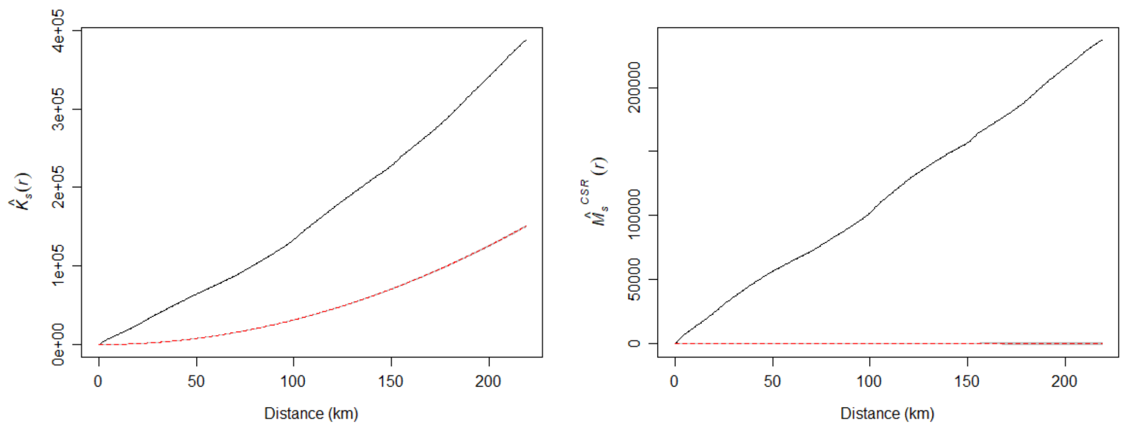
Year 2012



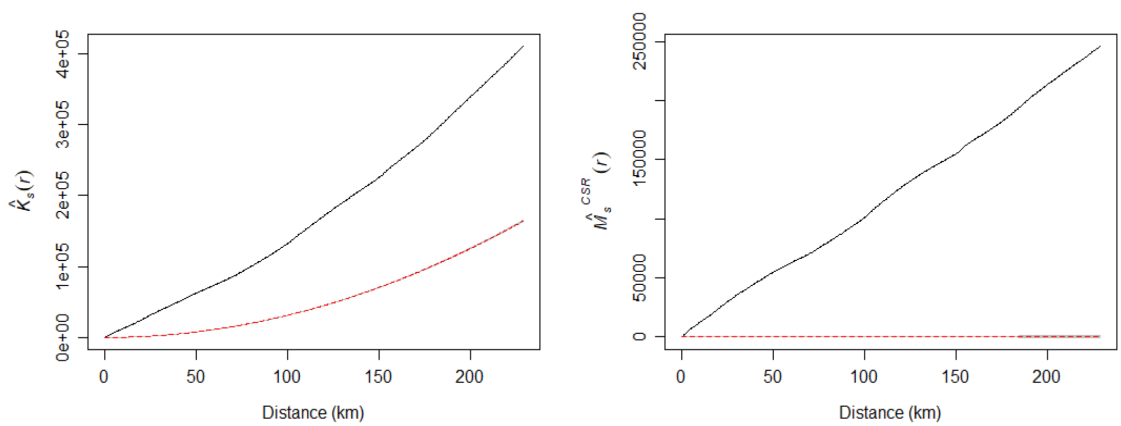


Sector 14 – Manufacture of wearing apparel

Year 2007

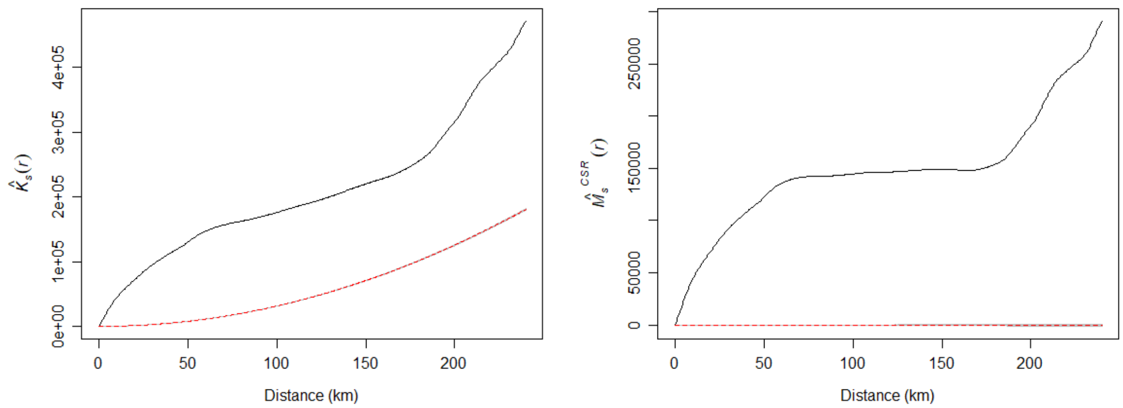


Year 2012

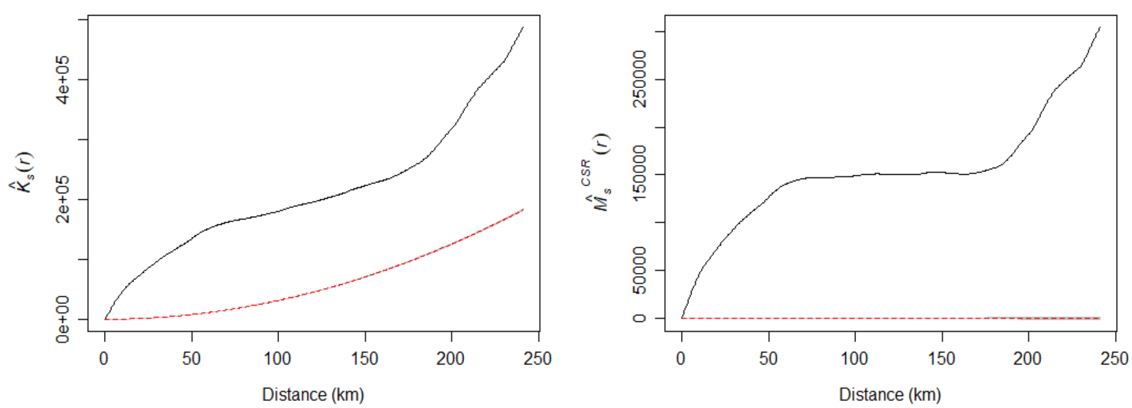


Sector 15 – Manufacture of leather and related products

Year 2007

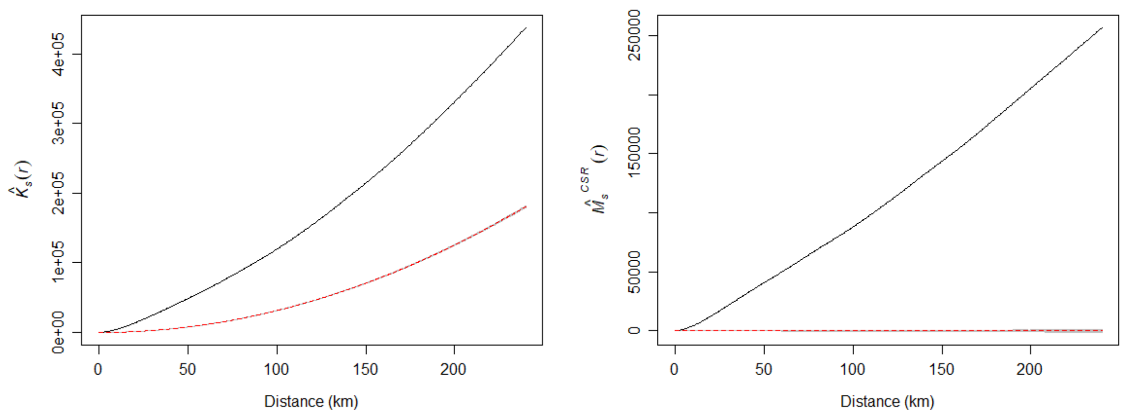


Year 2012

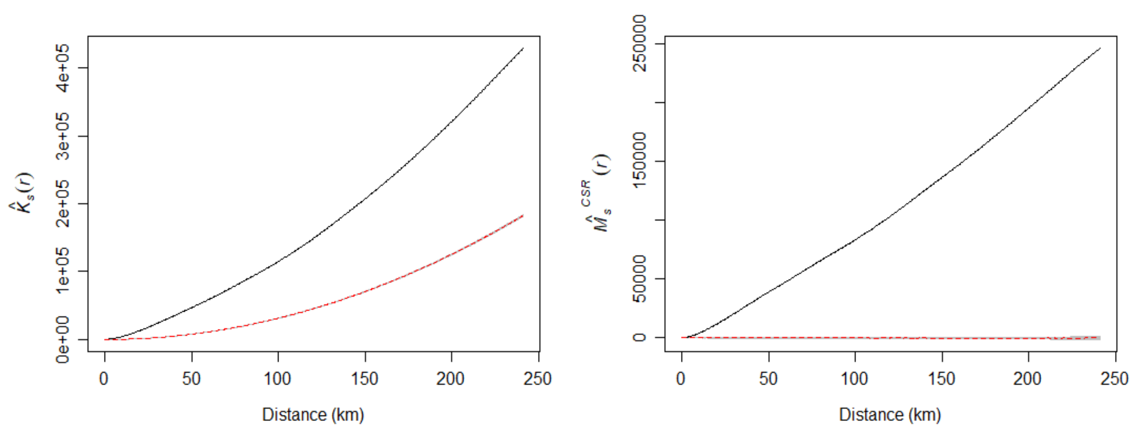


Sector 16 – Manufacture of wood, wood and cook products (except furniture), straw articles, plaiting materials

Year 2007

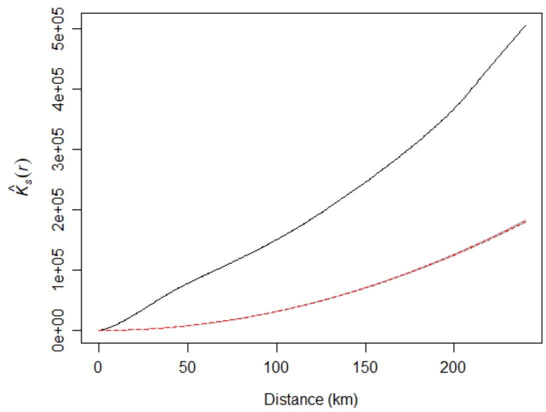


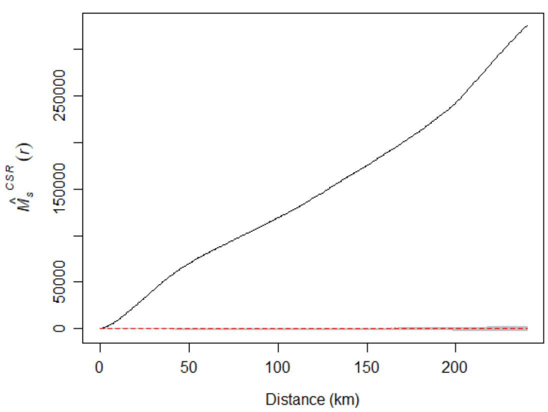
Year 2012



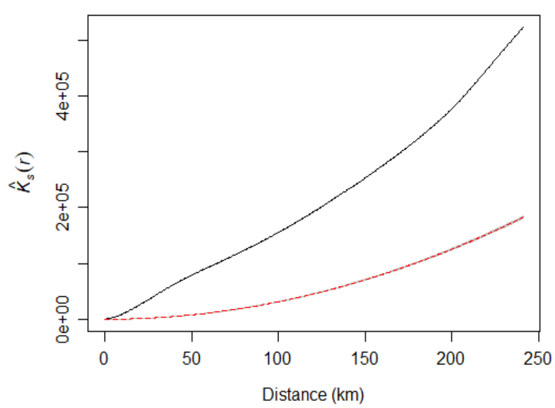
Sector 17 – Manufacture of paper and paper products

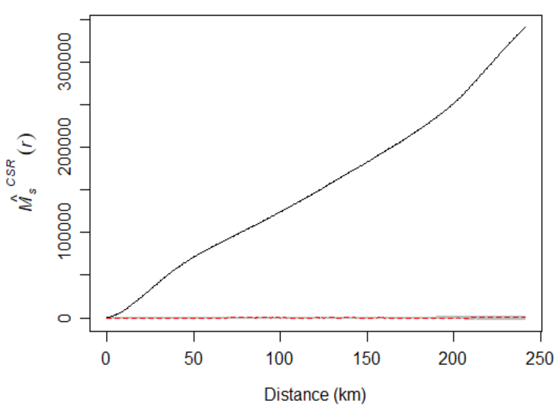
Year 2007





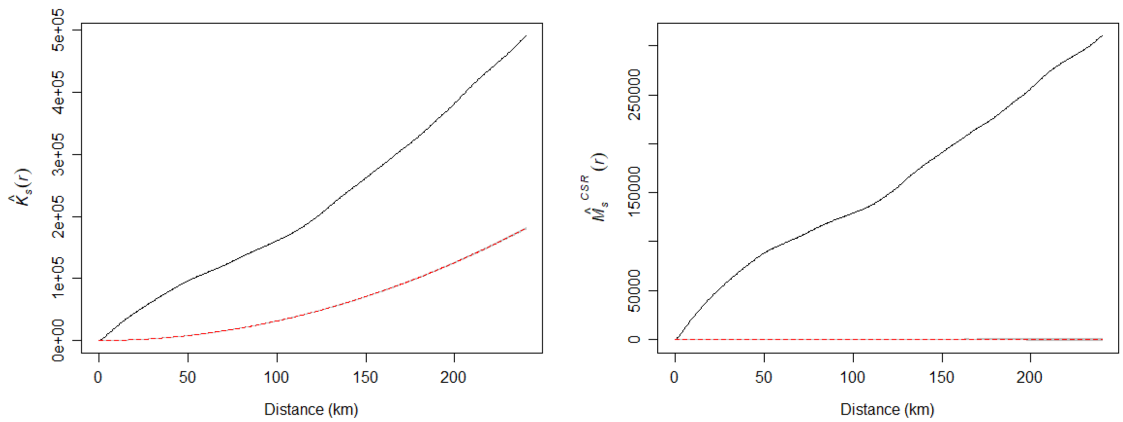
Year 2012



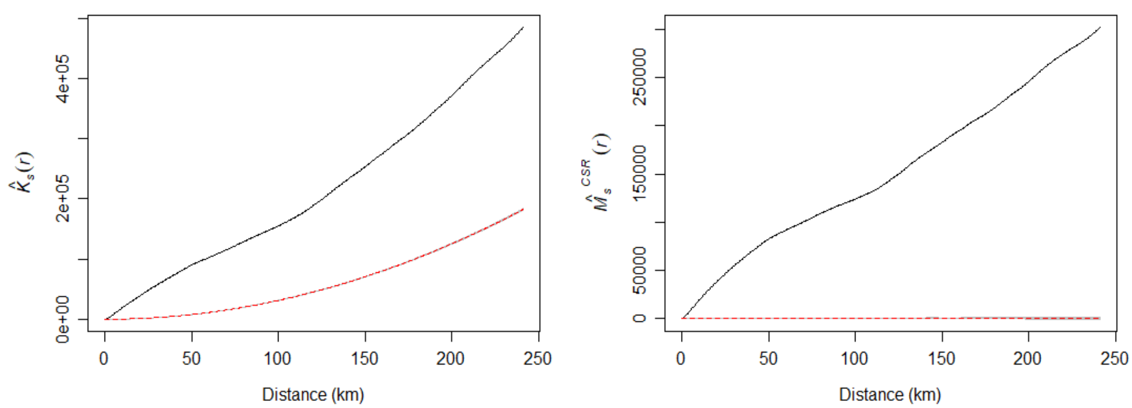


Sector 18 – Printing and reproduction of recorded media

Year 2007

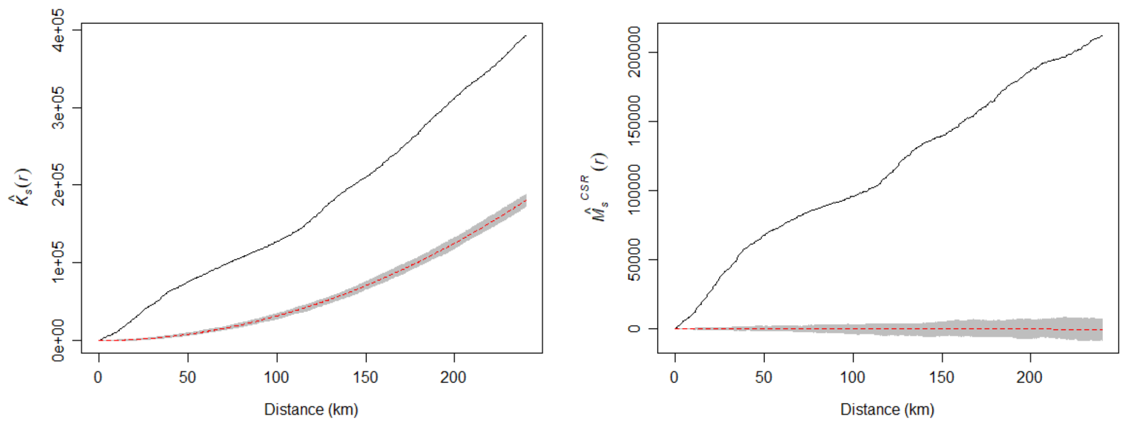


Year 2012

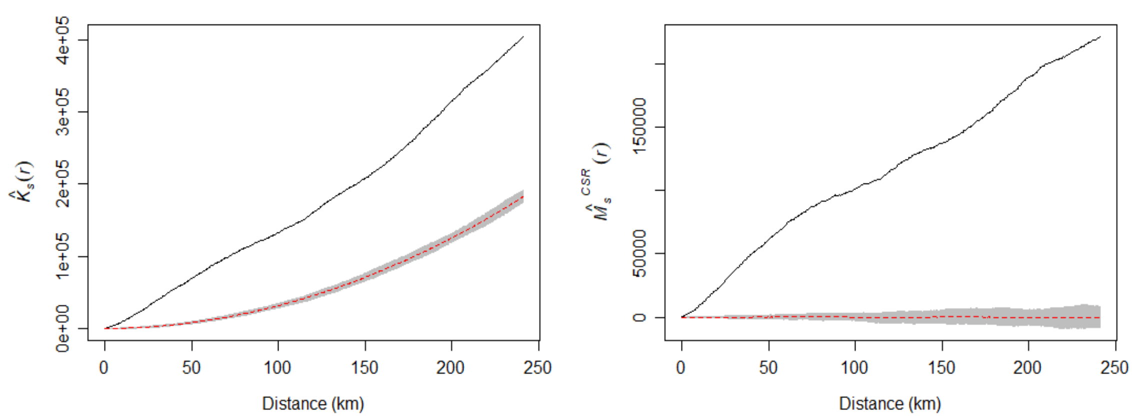


Sector 19 – Manufacture of coke and refined petroleum products

Year 2007

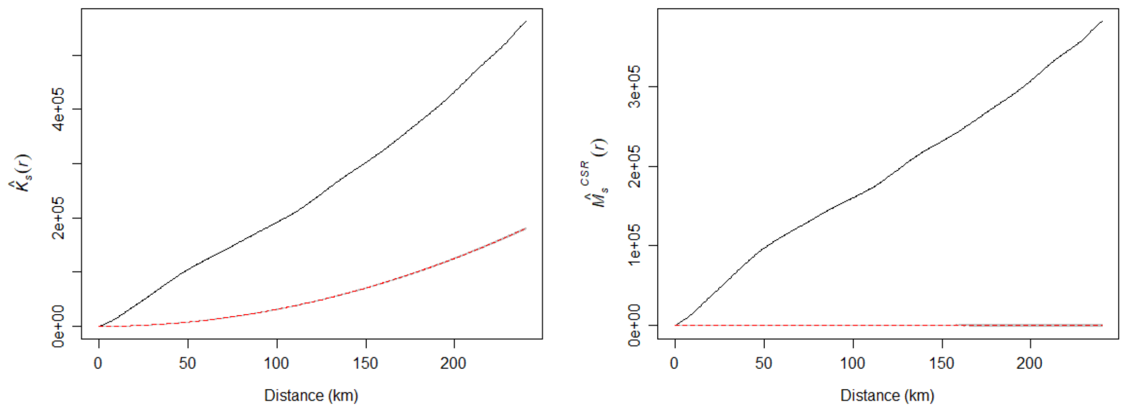


Year 2012

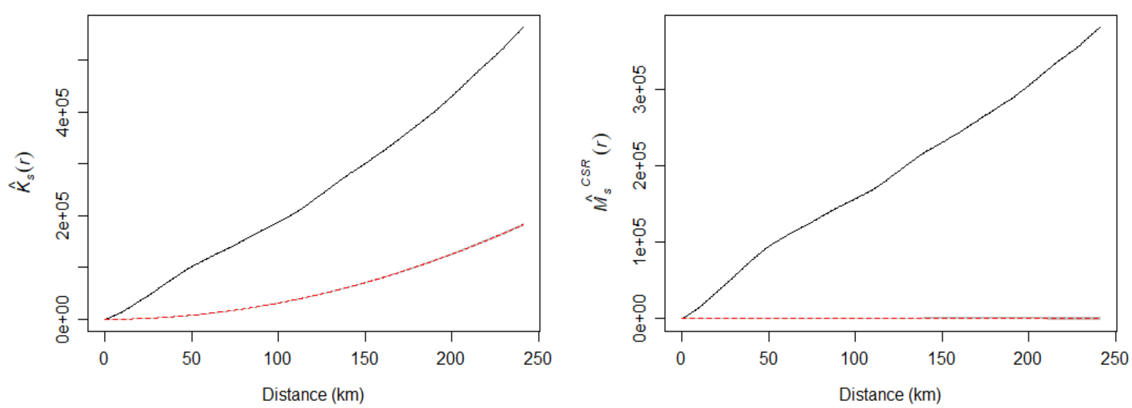


Sector 20 – Manufacture of chemicals and chemical products

Year 2007

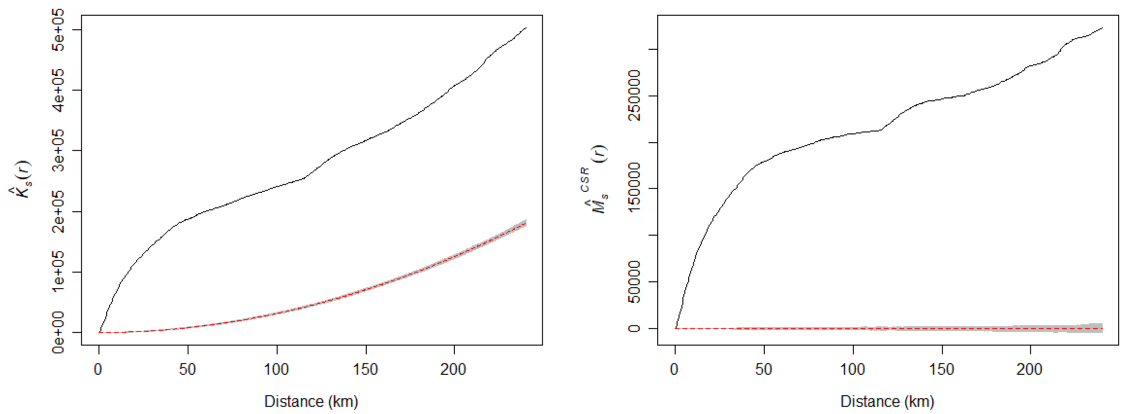


Year 2012

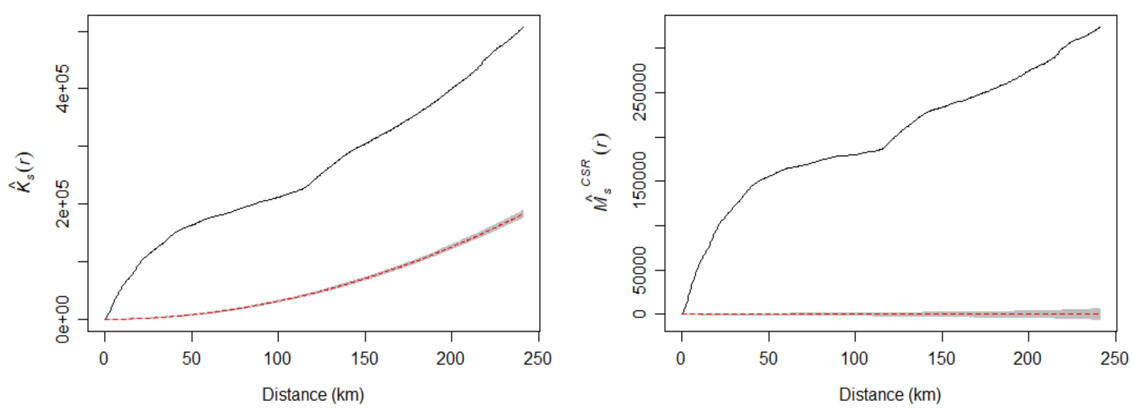


Sector 21 – Manufacture of basic pharmaceutical products and pharmaceutical preparations

Year 2007

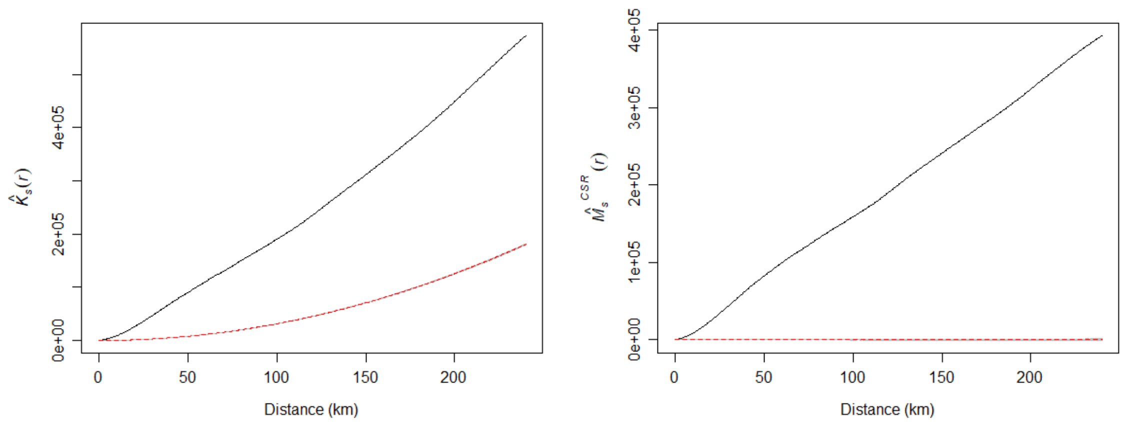


Year 2012

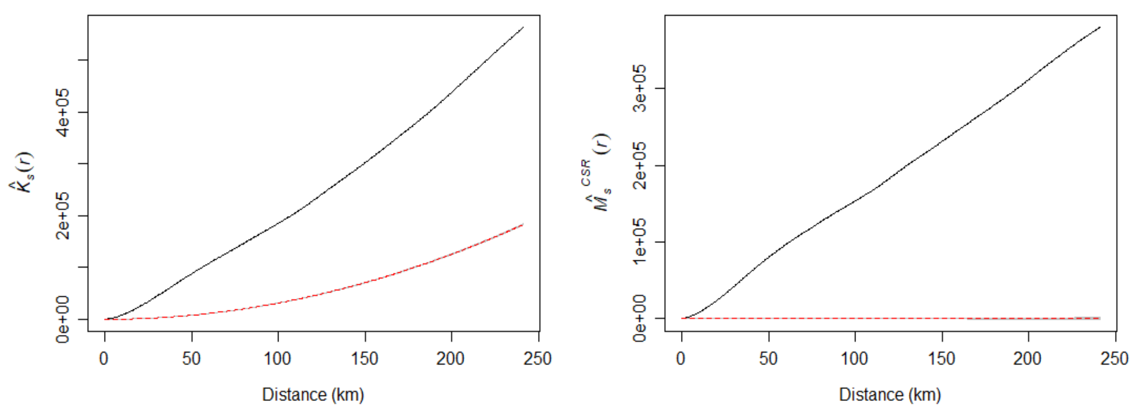


Sector 22 – Manufacture of rubber and plastic products

Year 2007

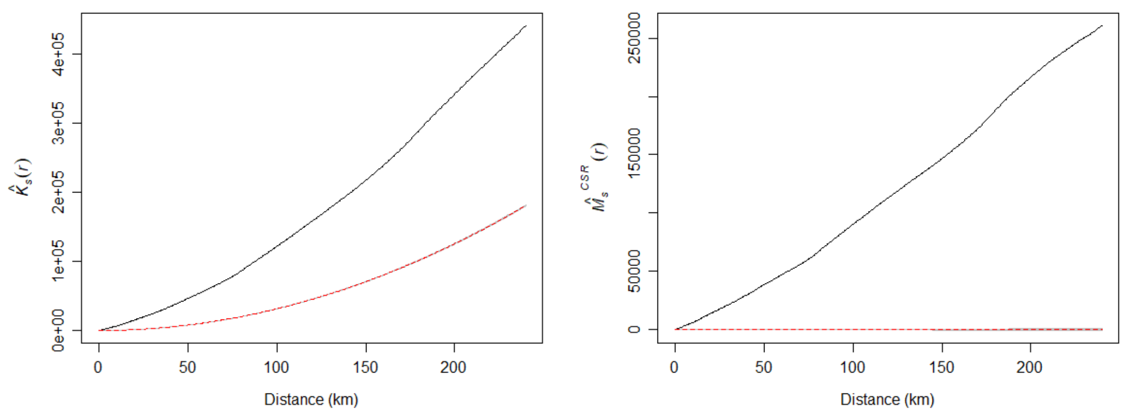


Year 2012

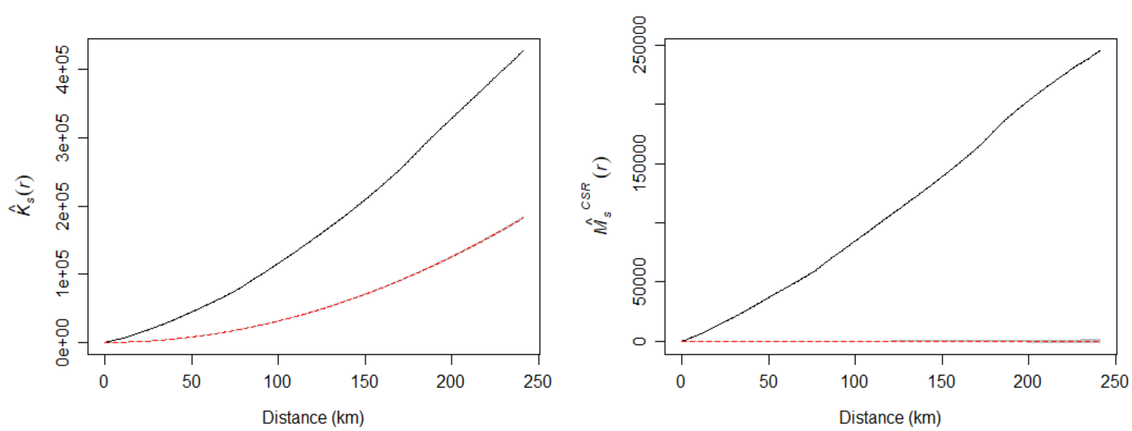


Sector 23 – Manufacture of other non-metallic mineral products

Year 2007

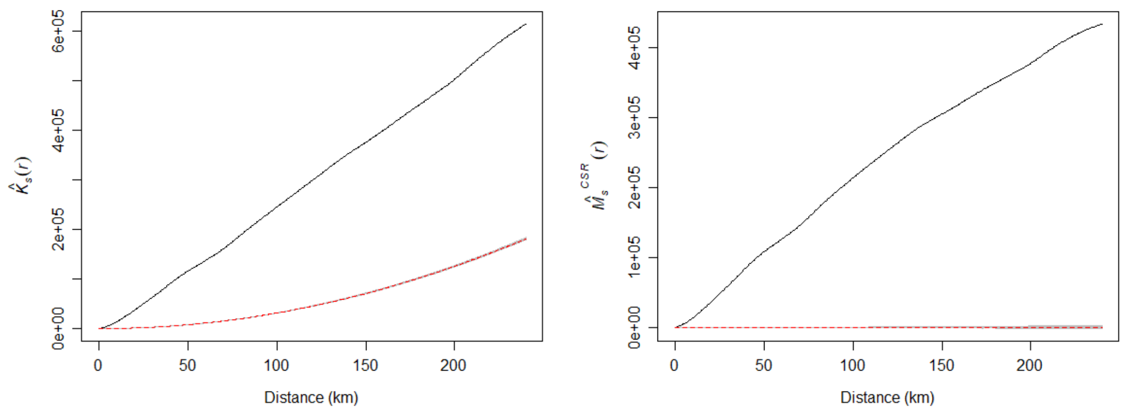


Year 2012

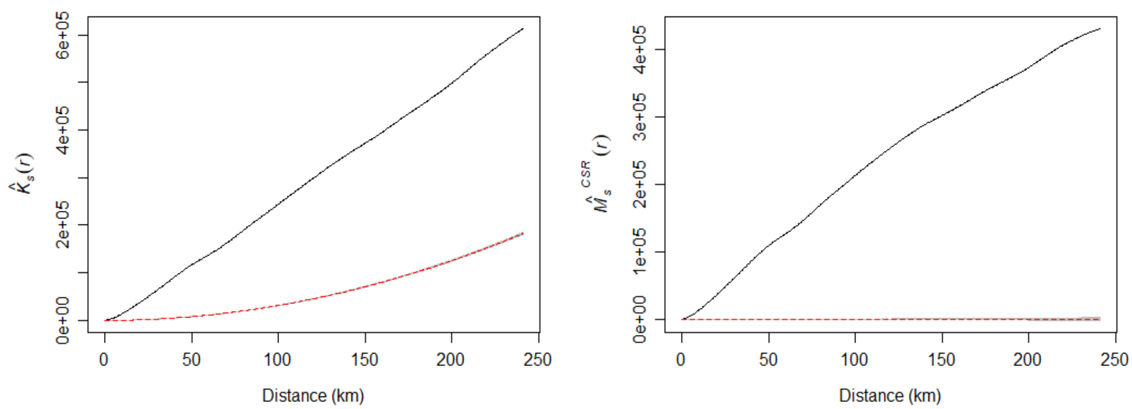


Sector 24 – Manufacture of basic metals

Year 2007

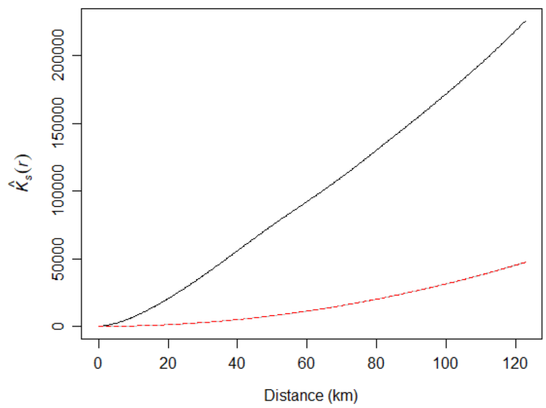


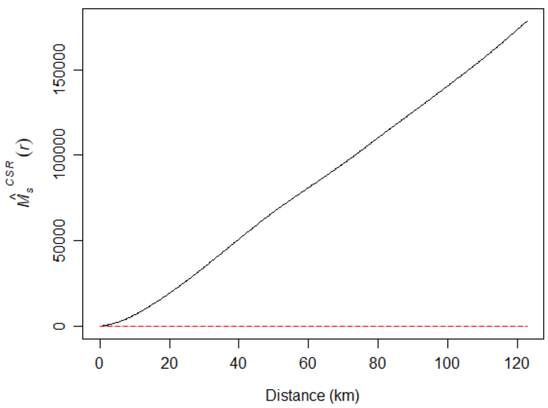
Year 2012



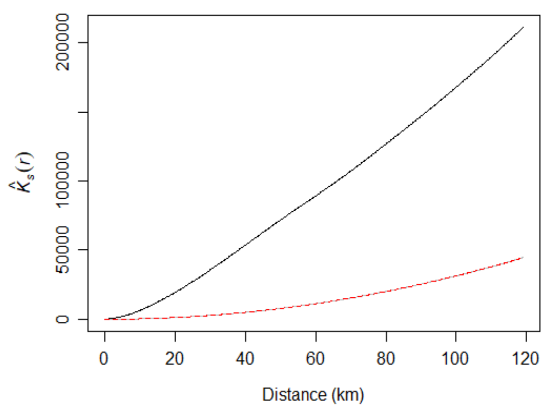
Sector 25 – Manufacture of fabricated metal products, except machinery and equipment

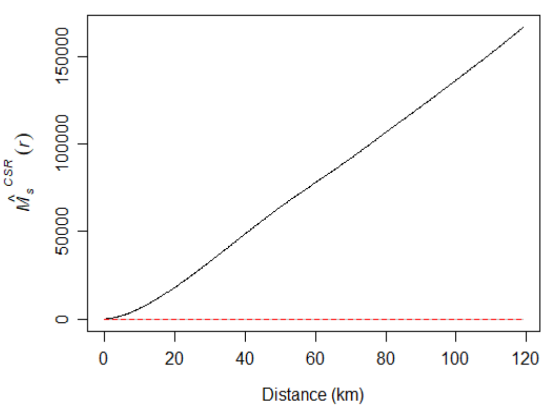
Year 2007





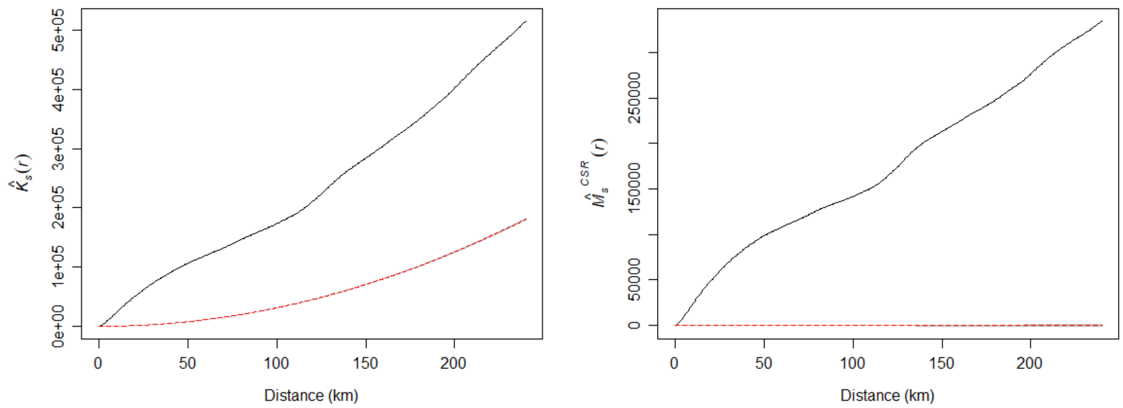
Year 2012



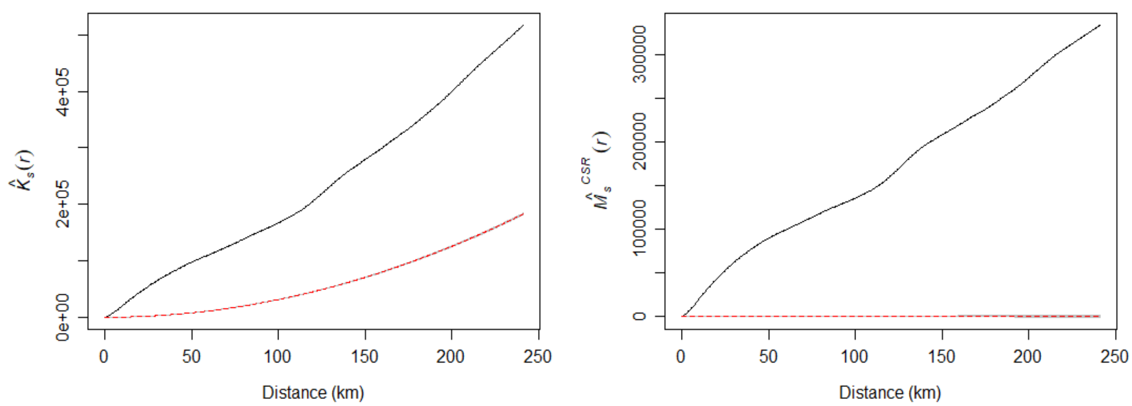


Sector 26 – Manufacture of computer, electronic and optical products

Year 2007

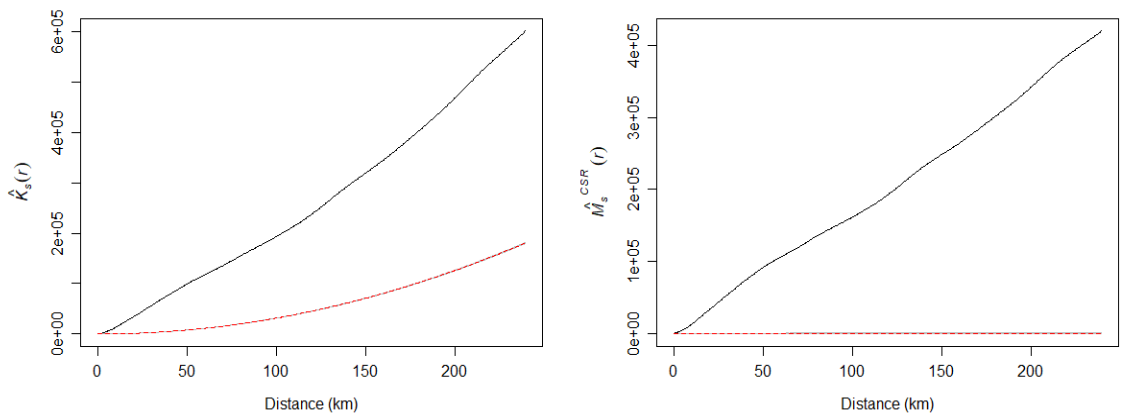


Year 2012

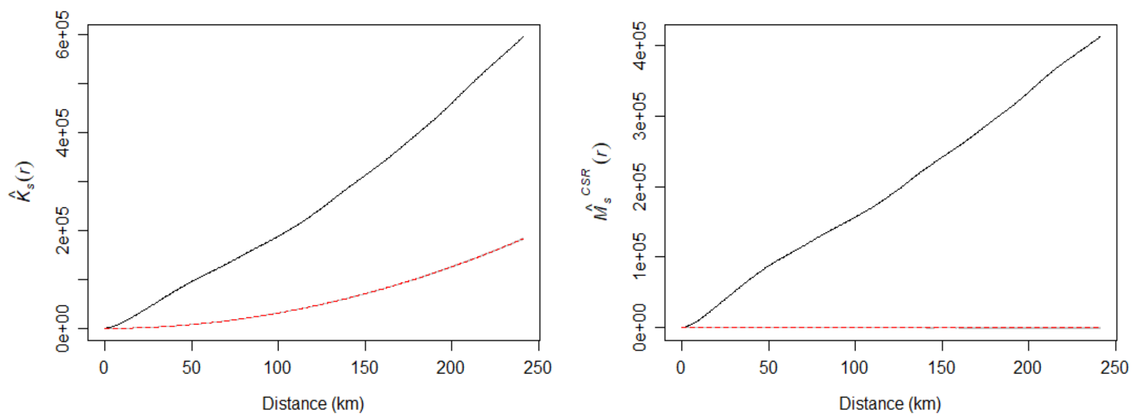


Sector 27 – Manufacture of electrical equipment

Year 2007

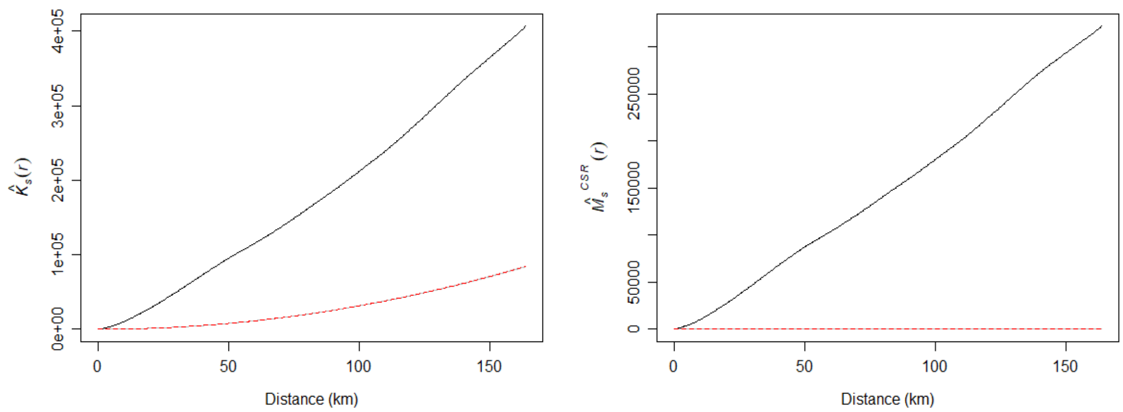


Year 2012

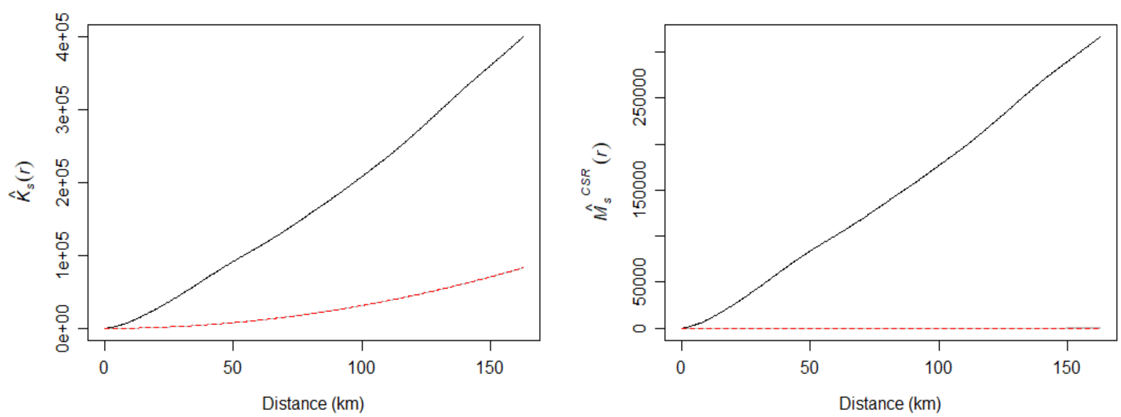


Sector 28 – Manufacture of machinery and equipment N.E.C.

Year 2007

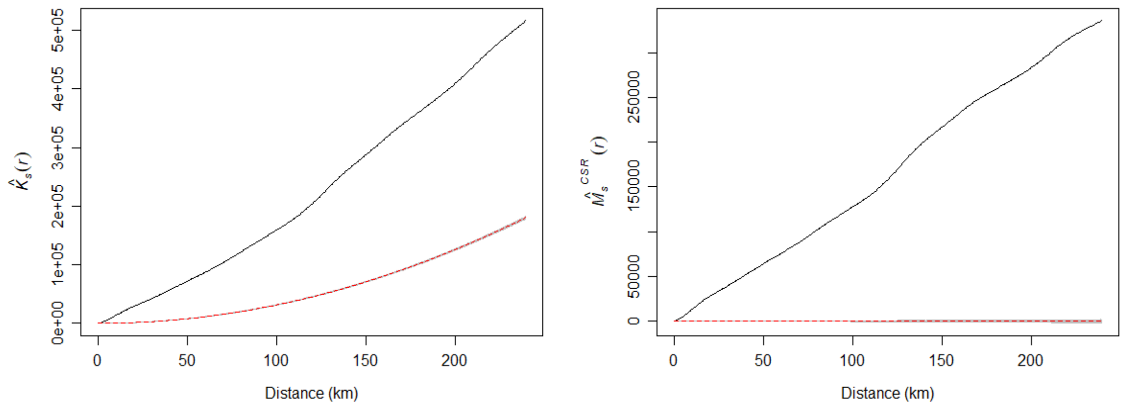


Year 2012

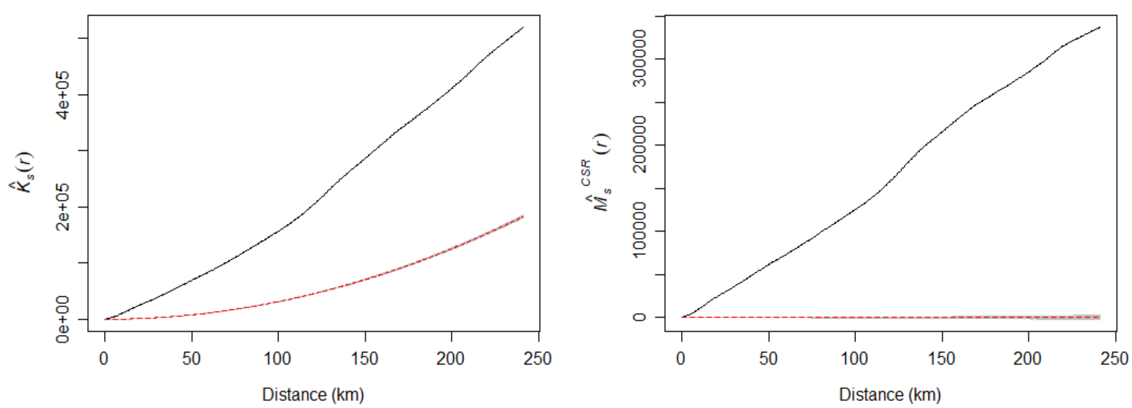


Sector 29 – Manufacture of motor vehicles, trailers and semi-trailers

Year 2007

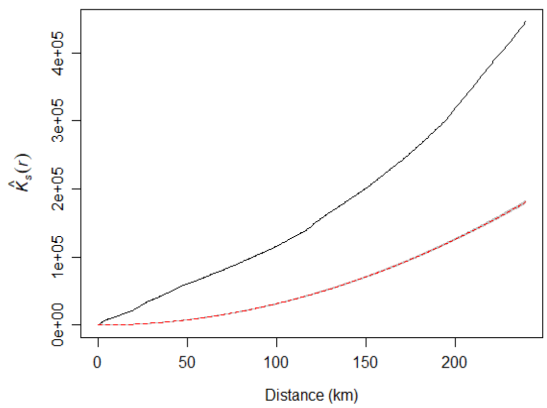


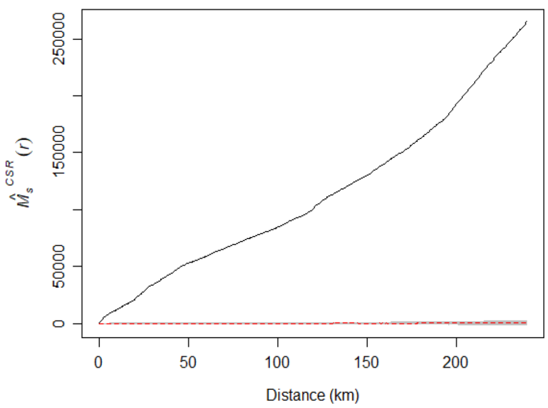
Year 2012



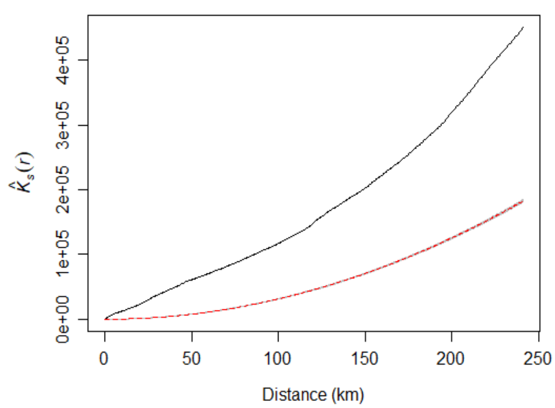
Sector 30 – Manufacture of other transport equipment

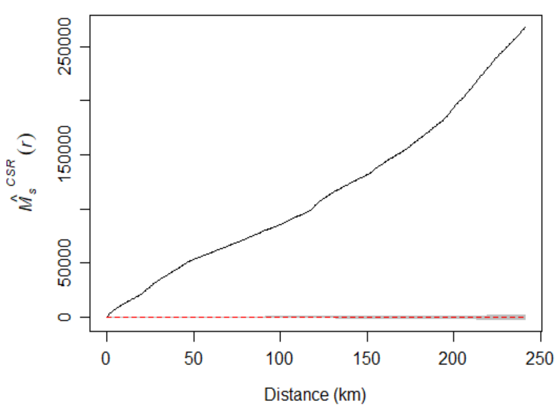
Year 2007





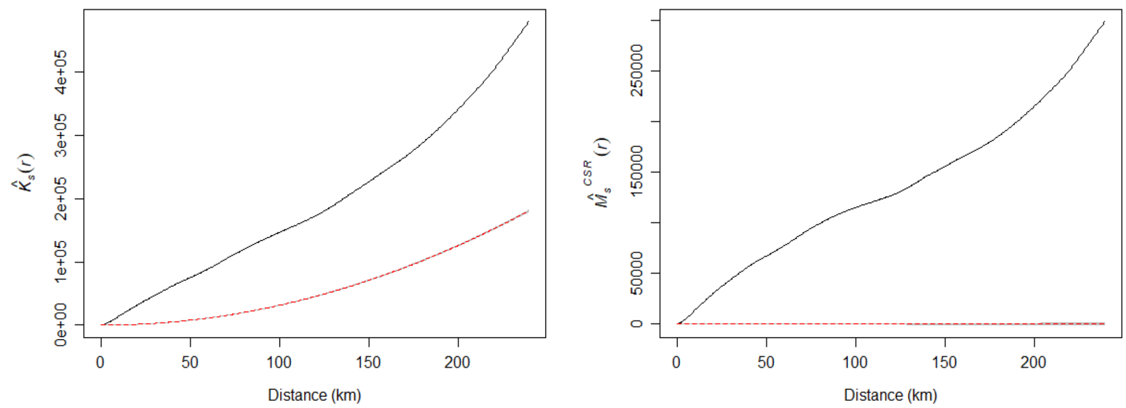
Year 2012



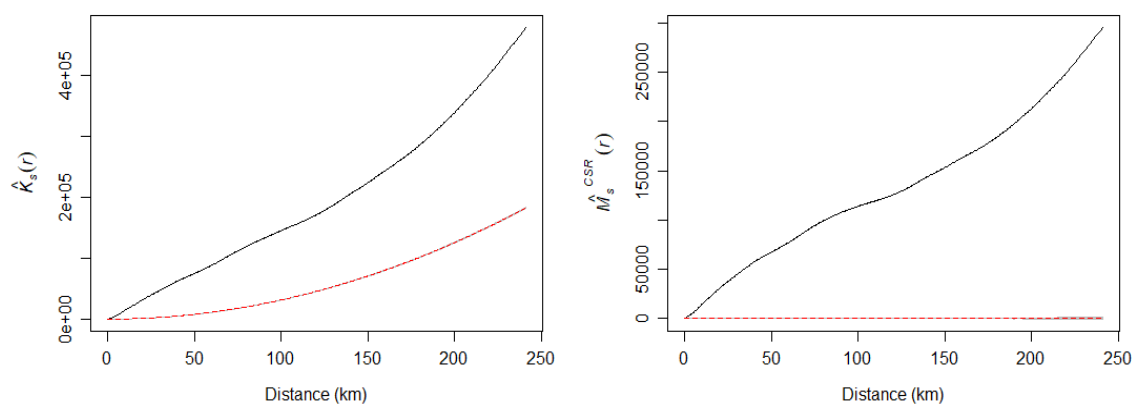


Sector 31 – Manufacture of furniture

Year 2007

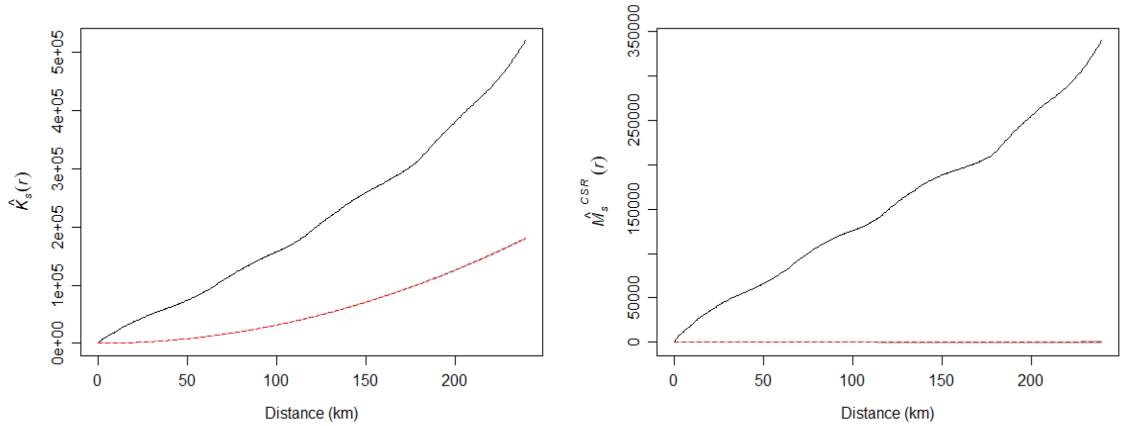


Year 2012

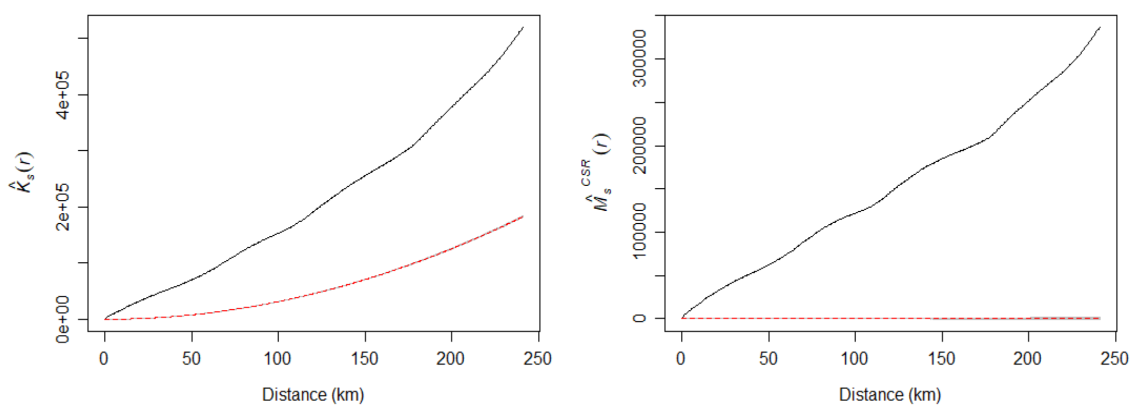


Sector 32 – Other manufacturing

Year 2007

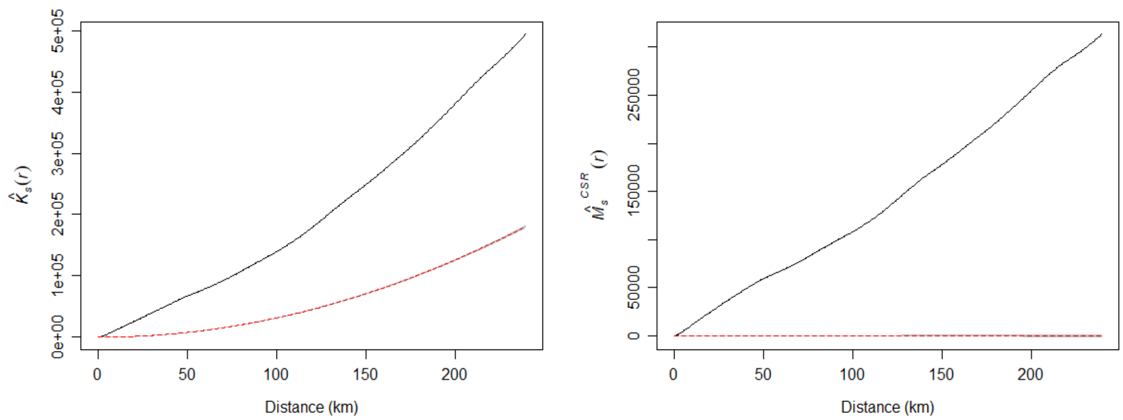


Year 2012

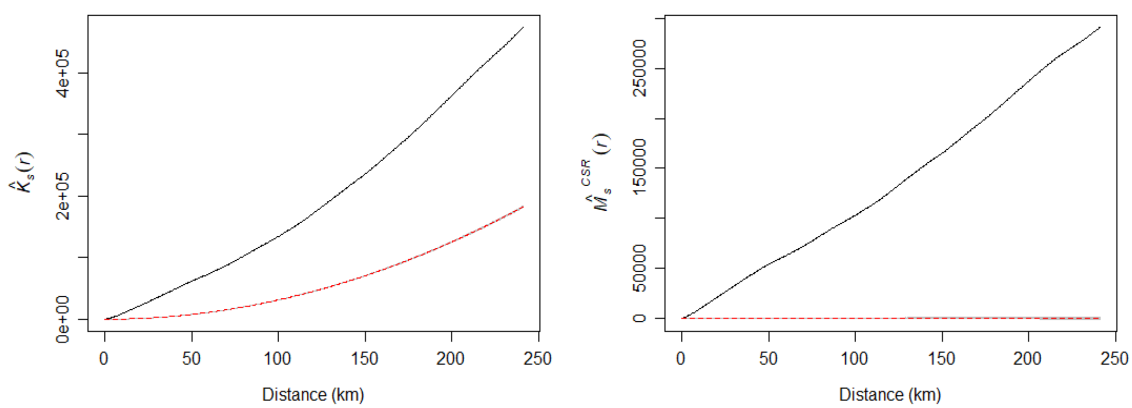


Sector 33 – Repair and installation of machinery and equipment

Year 2007



Year 2012

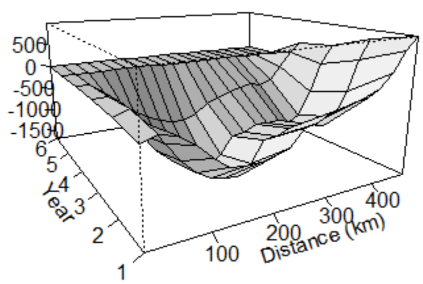


Notes: Authors' elaboration on *AIDA* data. The plots report the estimated values of the spatial *K*- and *M*-functions for the years 2007 and 2012.

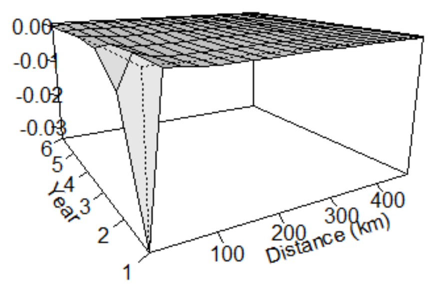
Figure A5: Results of the estimated space-time *K*-function by two-digit manufacturing sector.

Sector 10 – Manufacture of food products

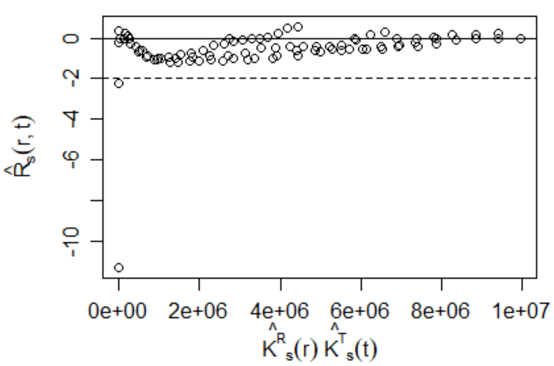
$\widehat{D}_s(r,t)$



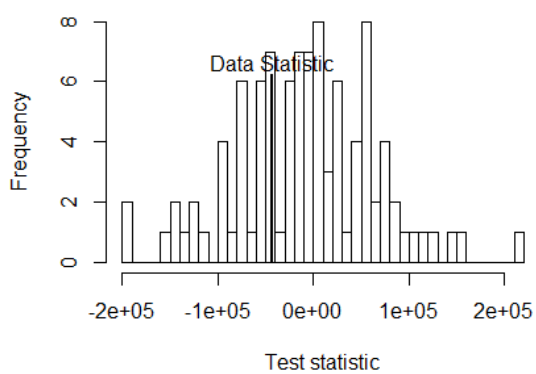
$\widehat{D}_s^0(r,t)$



Standardised Residuals

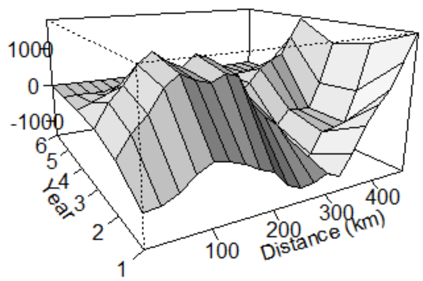


Monte Carlo test

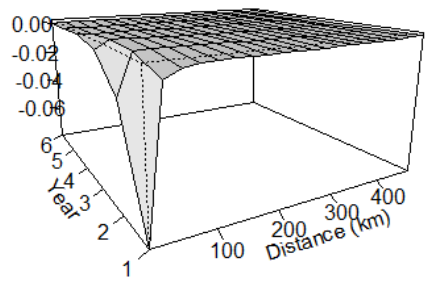


Sector 11 – Manufacture of beverages

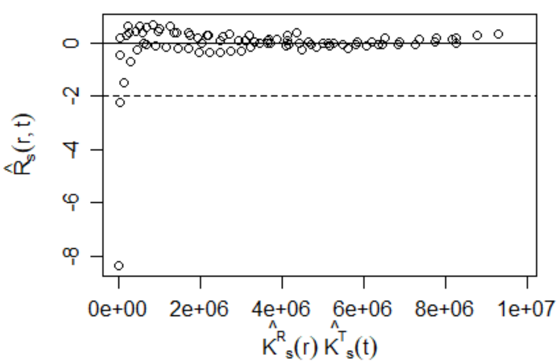
$\widehat{D}_s(r,t)$



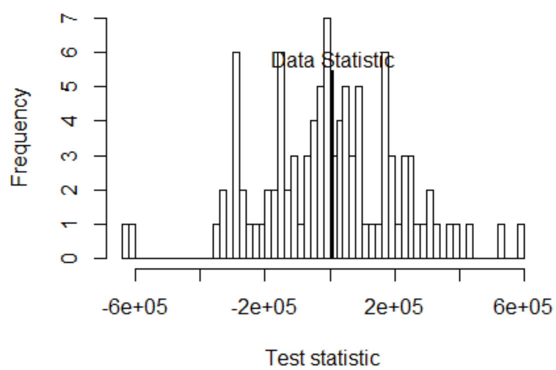
$\widehat{D}_s^0(r,t)$



Standardised Residuals

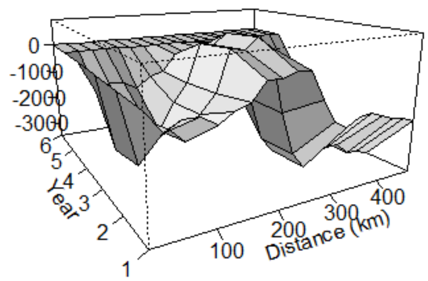


Monte Carlo test

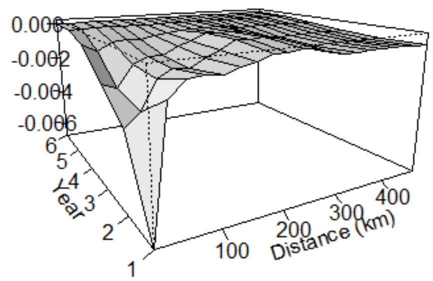


Sector 13 – Manufacture of textiles

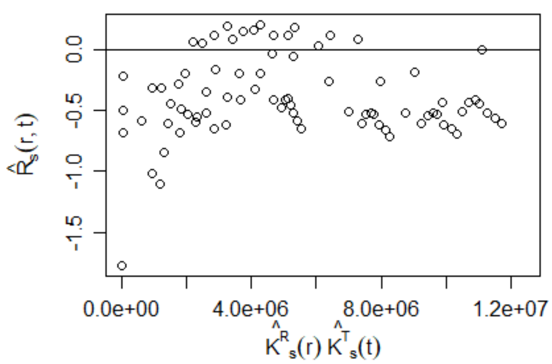
$\widehat{D}_s(r,t)$



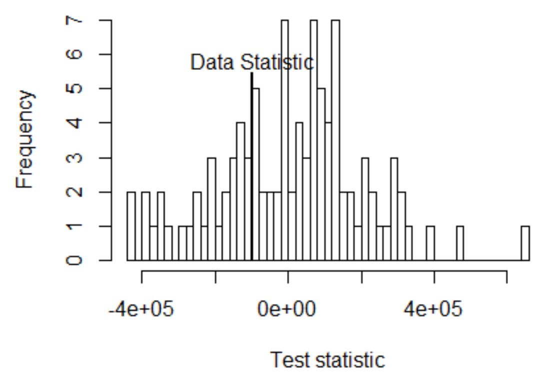
$\widehat{D}_s^0(r,t)$



Standardised Residuals

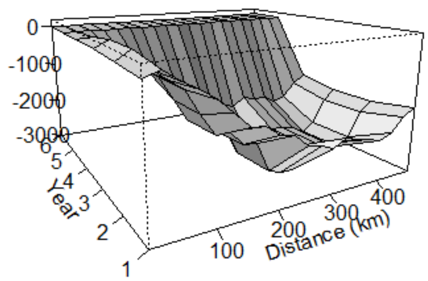


Monte Carlo test

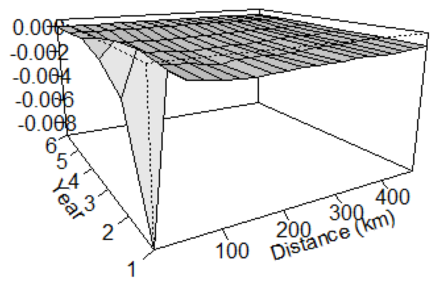


Sector 14 – Manufacture of wearing apparel

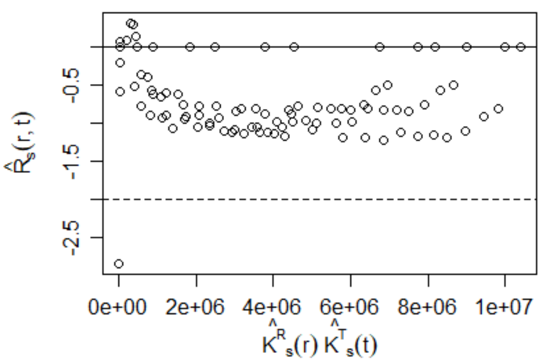
$\widehat{D}_s(r,t)$



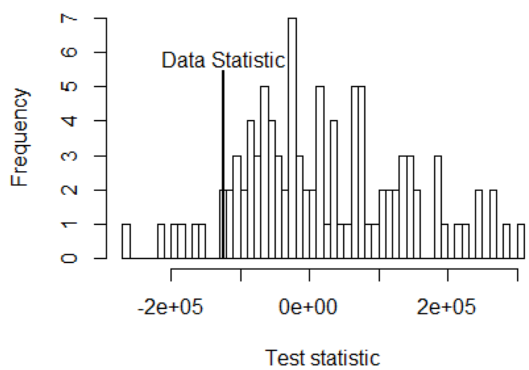
$\widehat{D}_s^0(r,t)$



Standardised Residuals

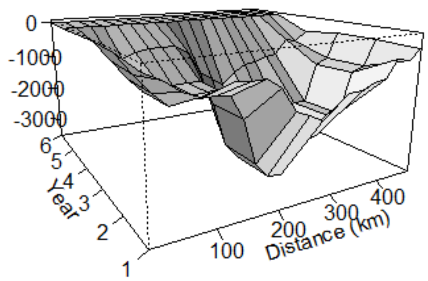


Monte Carlo test

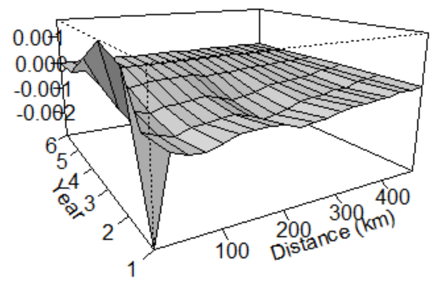


Sector 15 – Manufacture of leather and related products

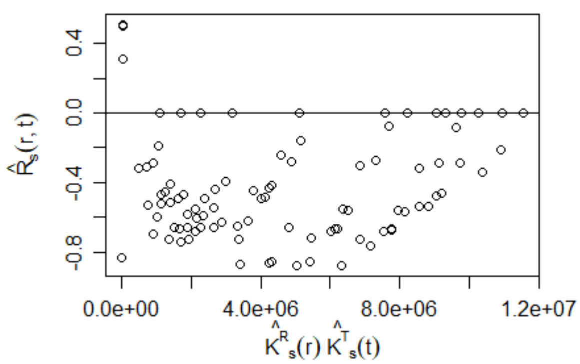
$\widehat{D}_s(r,t)$



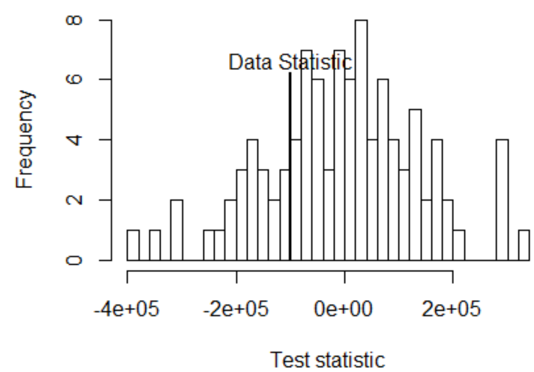
$\widehat{D}_s^0(r,t)$



Standardised Residuals

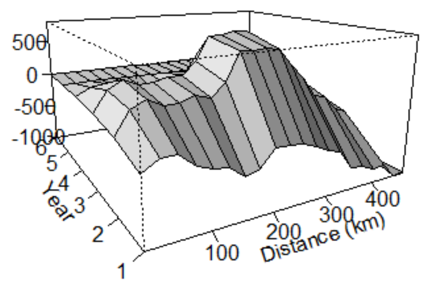


Monte Carlo test

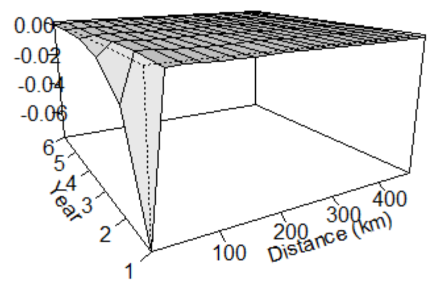


Sector 16 – Manufacture of wood, wood and cook products (except furniture), straw articles, plaiting materials

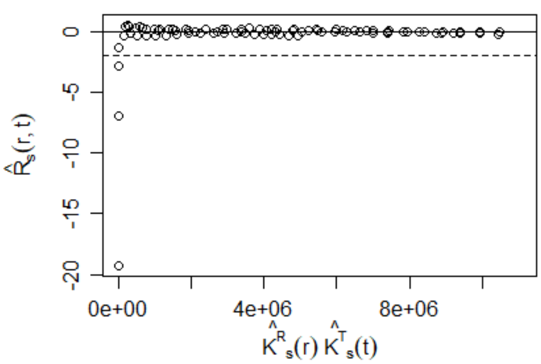
$\widehat{D}_s(r,t)$



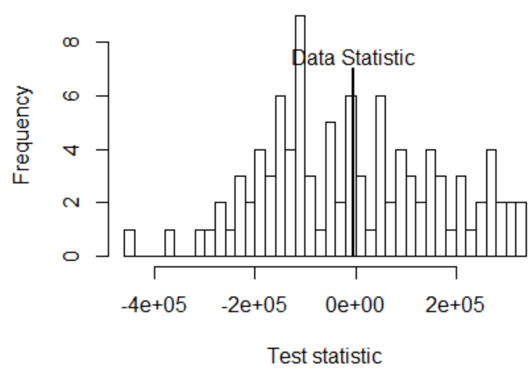
$\widehat{D}_s^0(r,t)$



Standardised Residuals

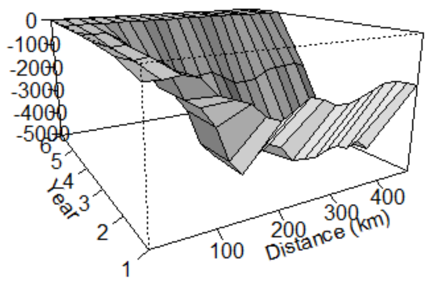


Monte Carlo test

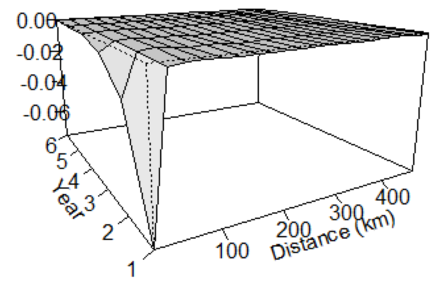


Sector 17 – Manufacture of paper and paper products

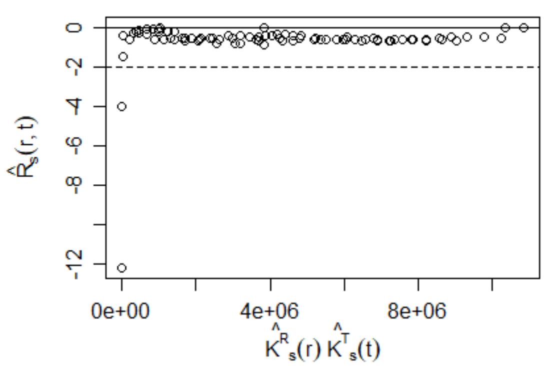
$\widehat{D}_s(r,t)$



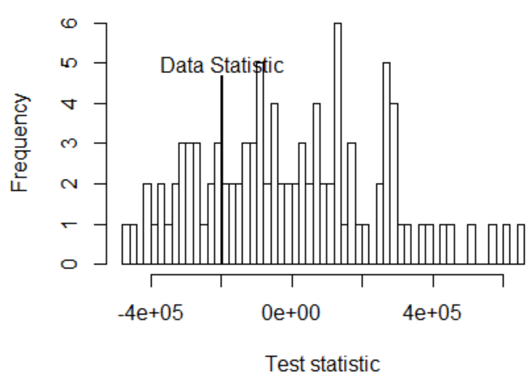
$\widehat{D}_s^0(r,t)$



Standardised Residuals

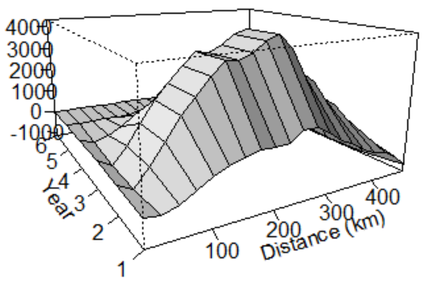


Monte Carlo test

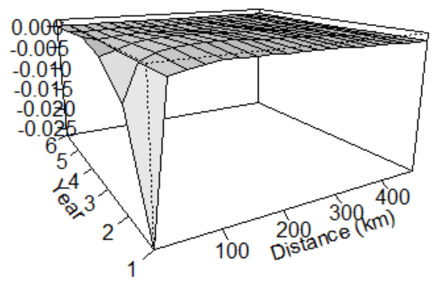


Sector 18 – Printing and reproduction of recorded media

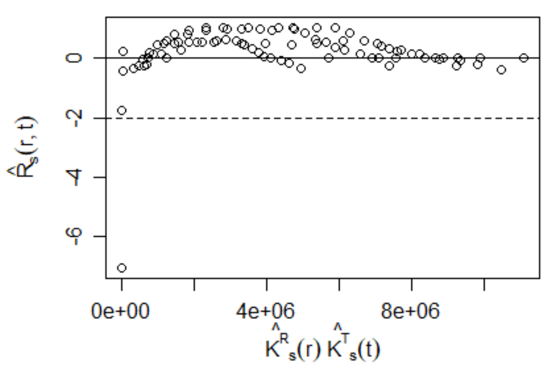
$\widehat{D}_s(r,t)$



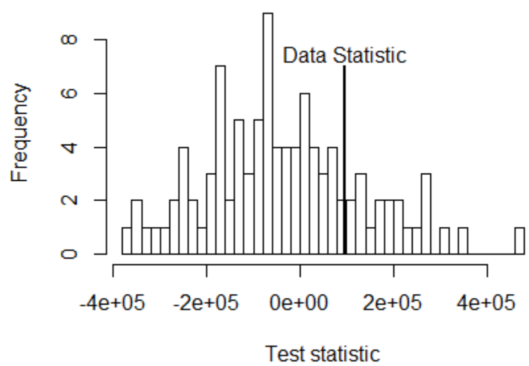
$\widehat{D}_s^0(r,t)$



Standardised Residuals

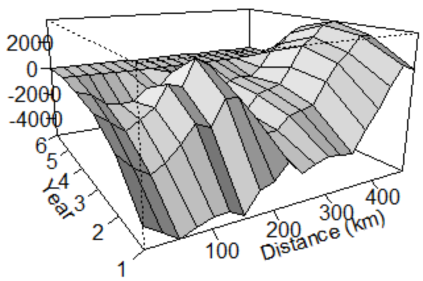


Monte Carlo test

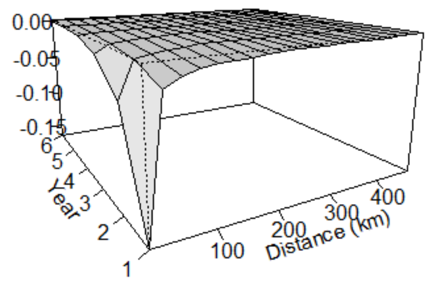


Sector 19 – Manufacture of coke and refined petroleum products

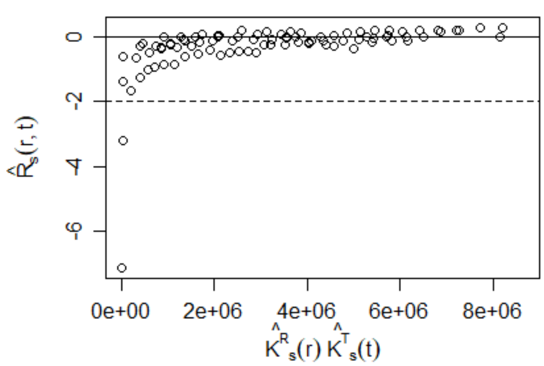
$\widehat{D}_s(r,t)$



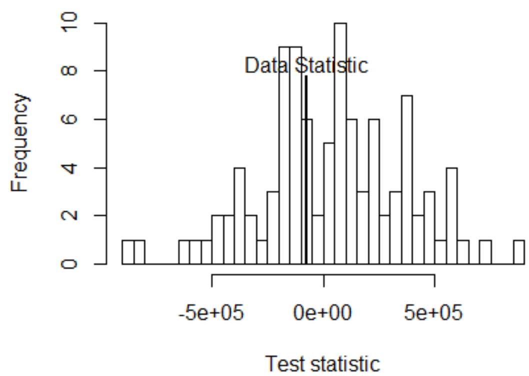
$\widehat{D}_s^0(r,t)$



Standardised Residuals

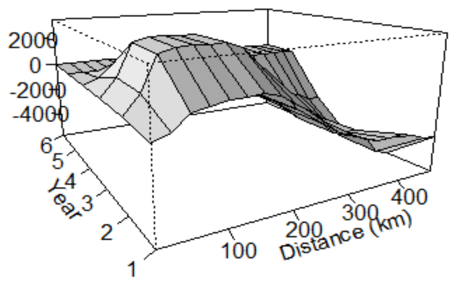


Monte Carlo test

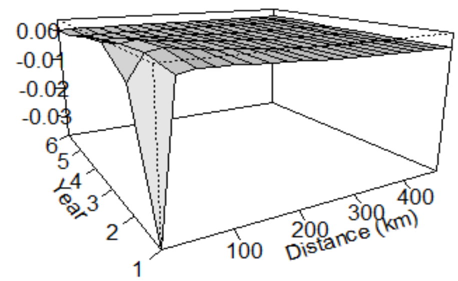


Sector 20 – Manufacture of chemicals and chemical products

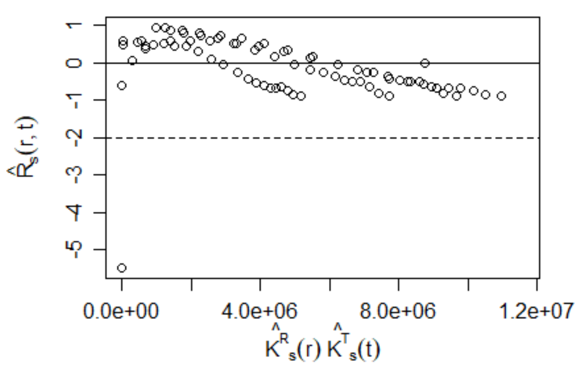
$\widehat{D}_s(r,t)$



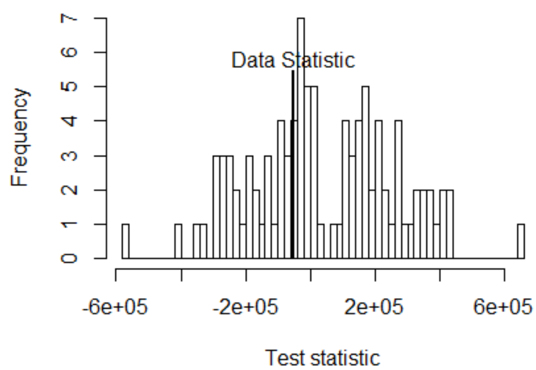
$\widehat{D}_s^0(r,t)$



Standardised Residuals

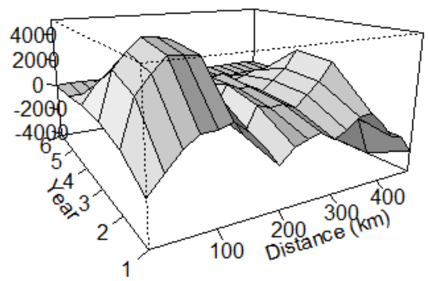


Monte Carlo test

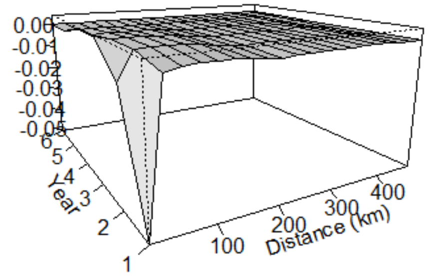


Sector 21 – Manufacture of basic pharmaceutical products and pharmaceutical preparations

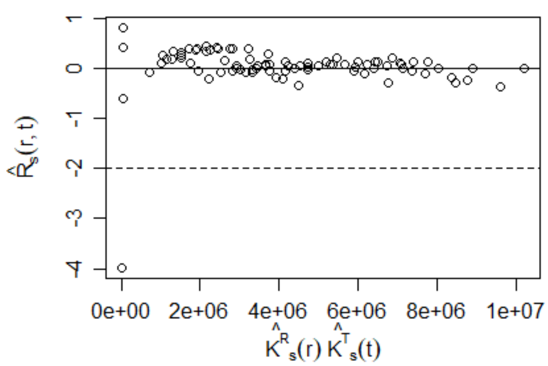
$\widehat{D}_s(r,t)$



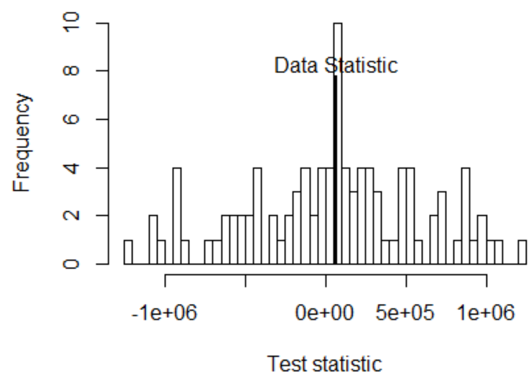
$\widehat{D}_s^0(r,t)$



Standardised Residuals

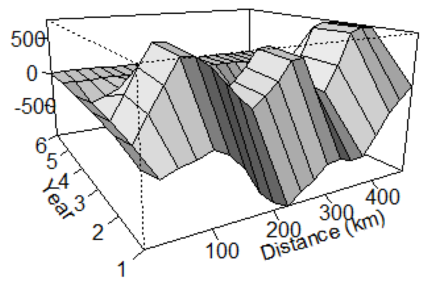


Monte Carlo test

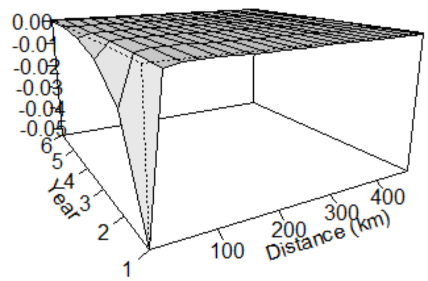


Sector 22 – Manufacture of rubber and plastic products

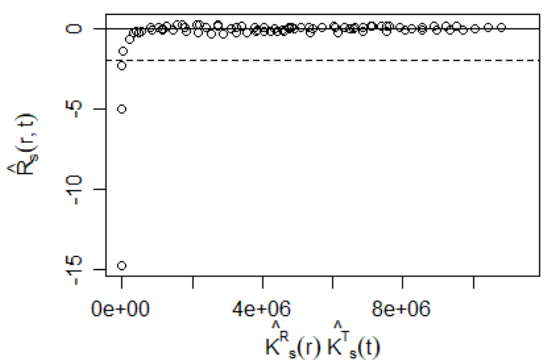
$\widehat{D}_s(r,t)$



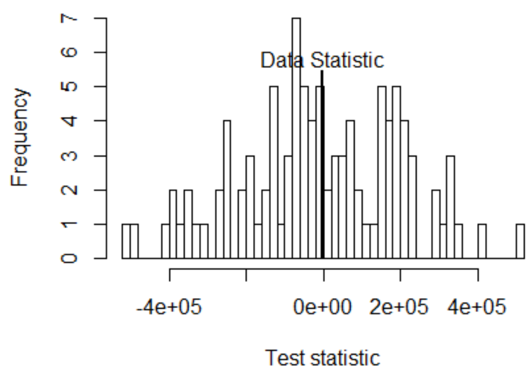
$\widehat{D}_s^0(r,t)$



Standardised Residuals

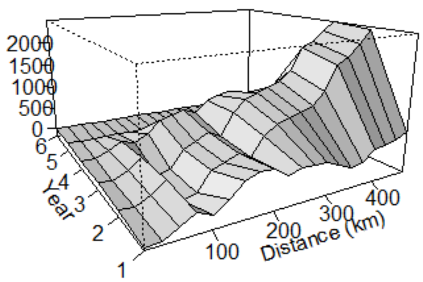


Monte Carlo test

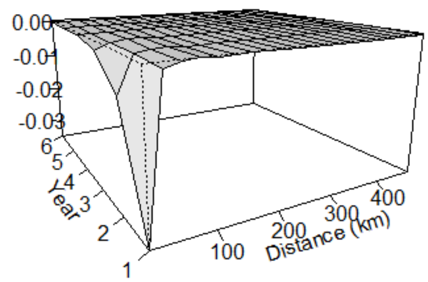


Sector 23 – Manufacture of other non-metallic mineral products

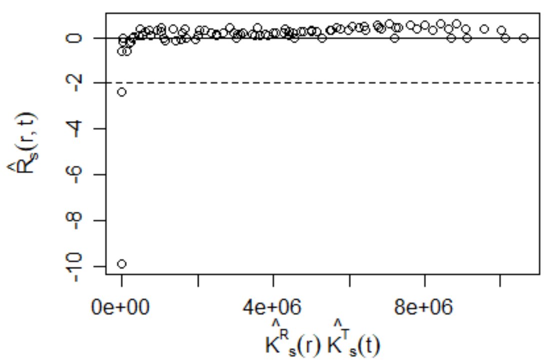
$\widehat{D}_s(r,t)$



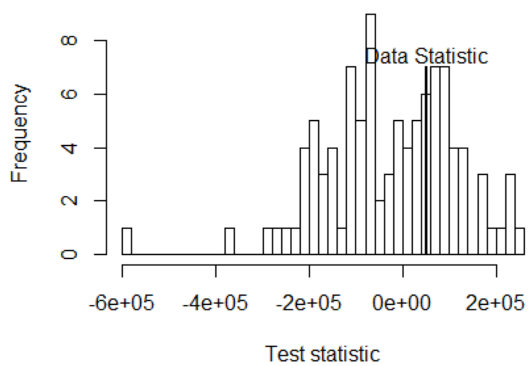
$\widehat{D}_s^0(r,t)$



Standardised Residuals

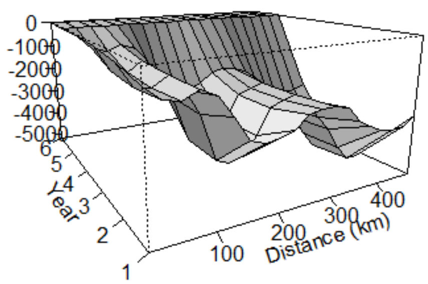


Monte Carlo test

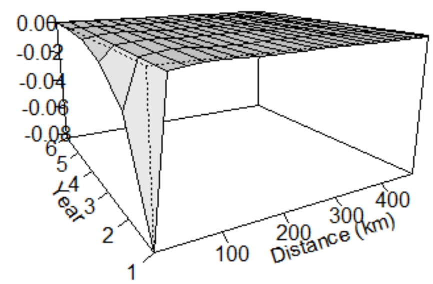


Sector 24 – Manufacture of basic metals

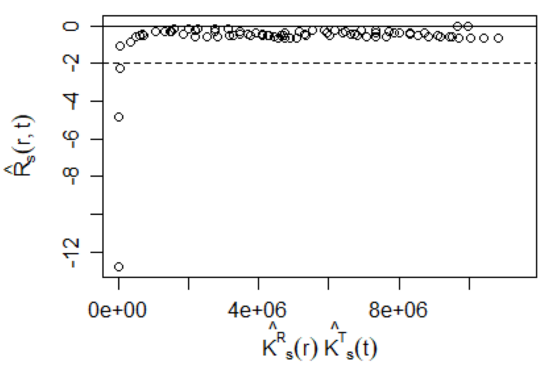
$\widehat{D}_s(r,t)$



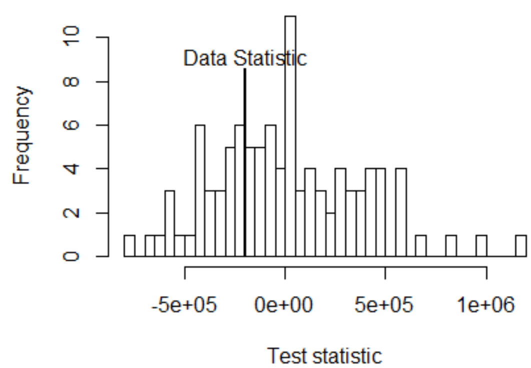
$\widehat{D}_s^0(r,t)$



Standardised Residuals

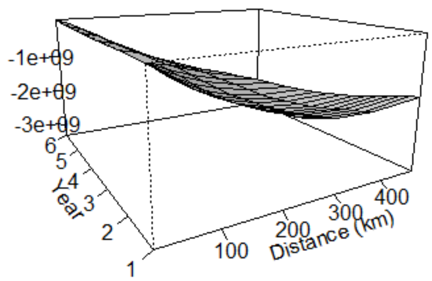


Monte Carlo test

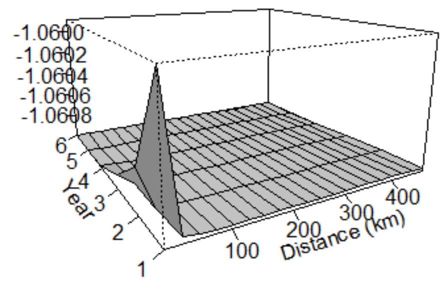


Sector 25 – Manufacture of fabricated metal products, except machinery and equipment

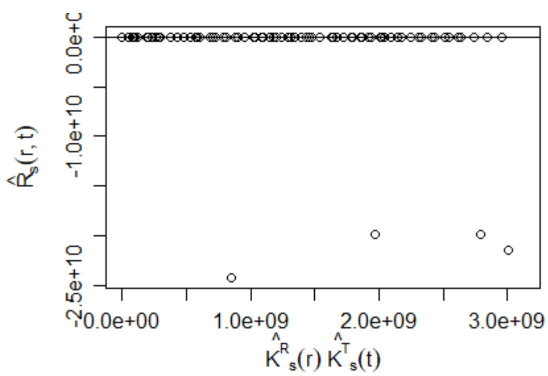
$\widehat{D}_s(r,t)$



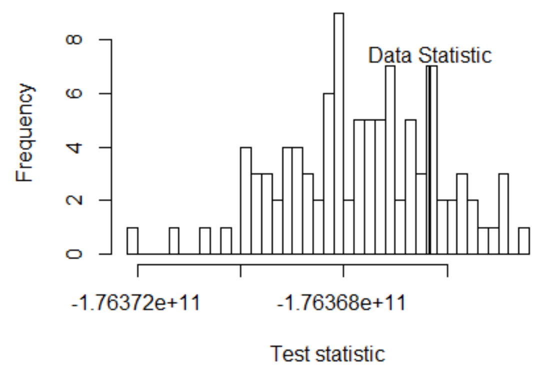
$\widehat{D}_s^0(r,t)$



Standardised Residuals

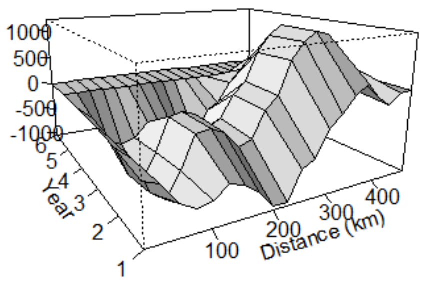


Monte Carlo test

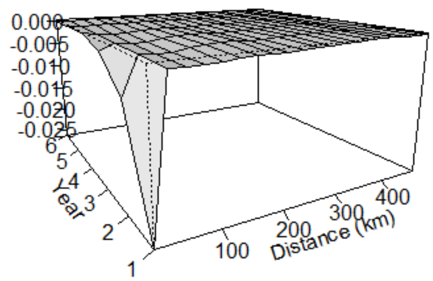


Sector 26 – Manufacture of computer, electronic and optical products

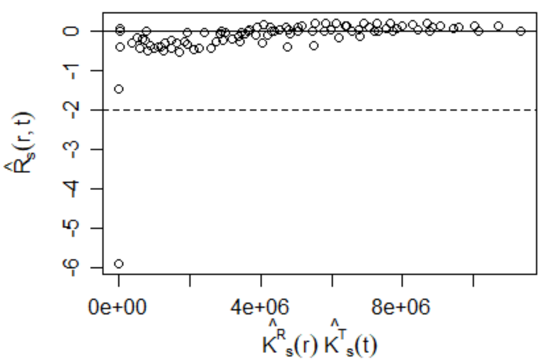
$\widehat{D}_s(r,t)$



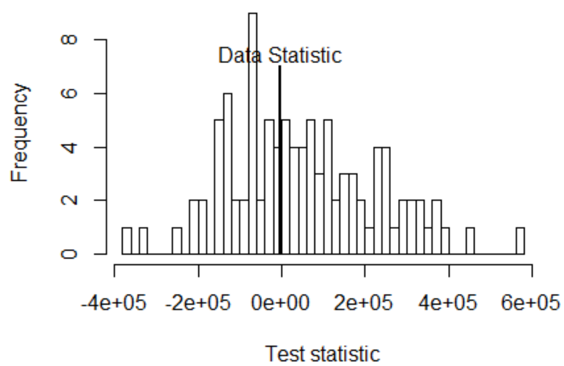
$\widehat{D}^0_s(r,t)$



Standardised Residuals

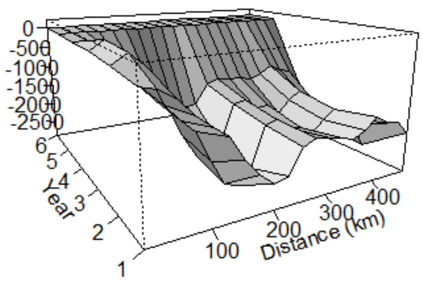


Monte Carlo test

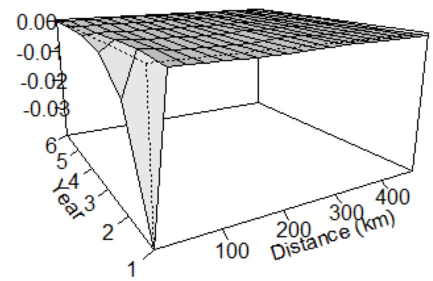


Sector 27 – Manufacture of electrical equipment

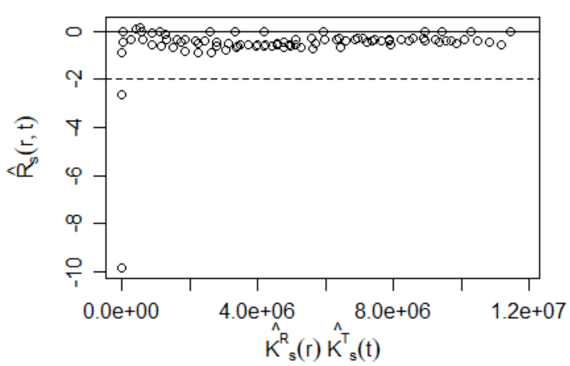
$\widehat{D}_s(r,t)$



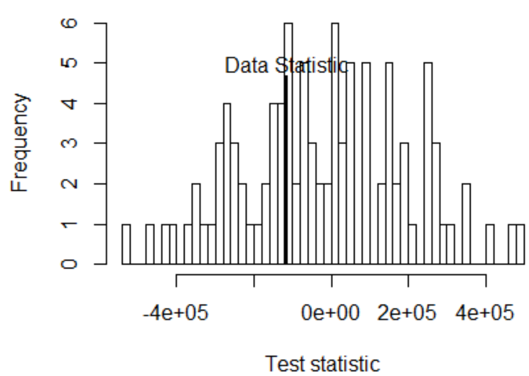
$\widehat{D}_s^0(r,t)$



Standardised Residuals

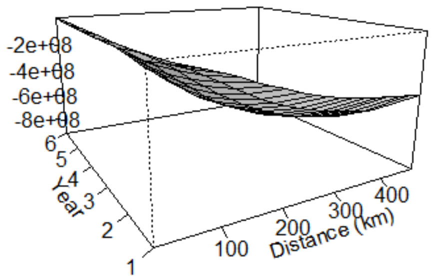


Monte Carlo test

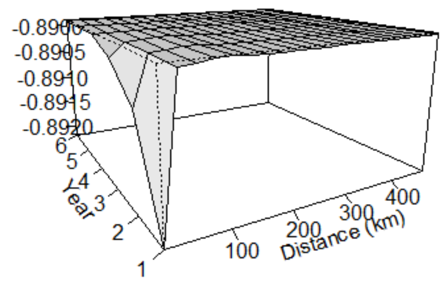


Sector 28 – Manufacture of machinery and equipment N.E.C.

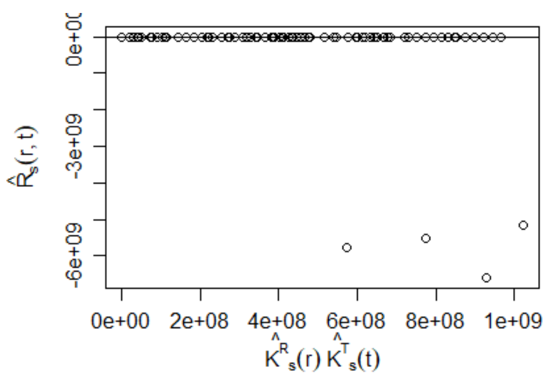
$\widehat{D}_s(r,t)$



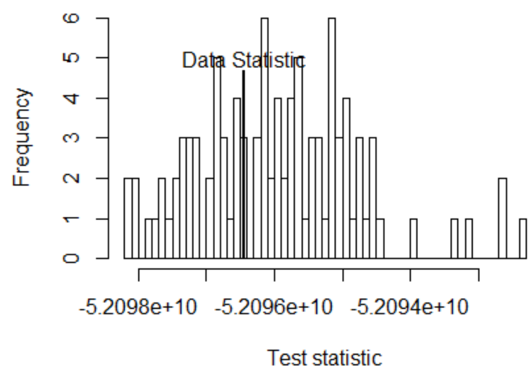
$\widehat{D}_s^0(r,t)$



Standardised Residuals

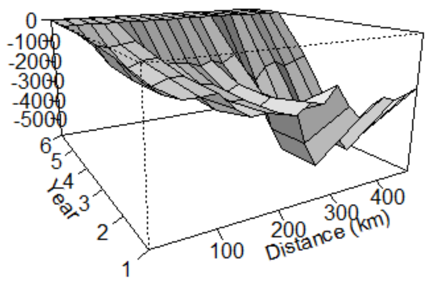


Monte Carlo test

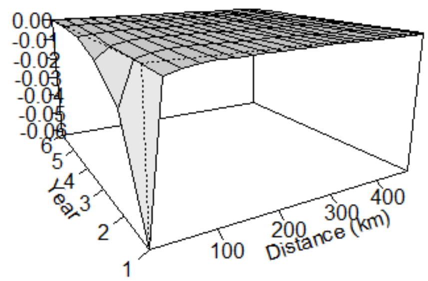


Sector 29 – Manufacture of motor vehicles, trailers and semi-trailers

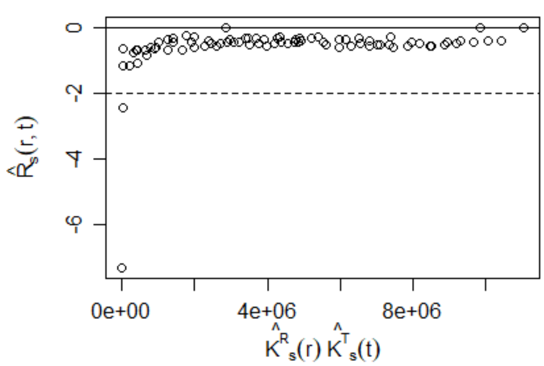
$\widehat{D}_s(r,t)$



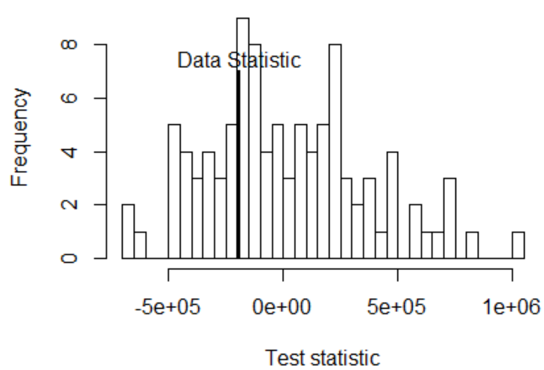
$\widehat{D}_s^0(r,t)$



Standardised Residuals

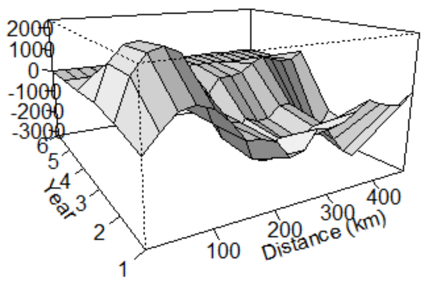


Monte Carlo test

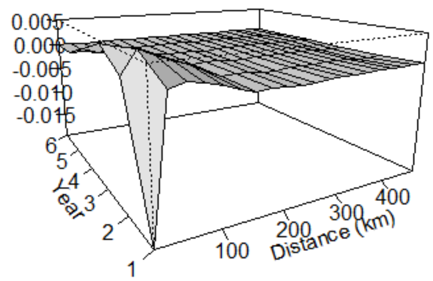


Sector 30 – Manufacture of other transport equipment

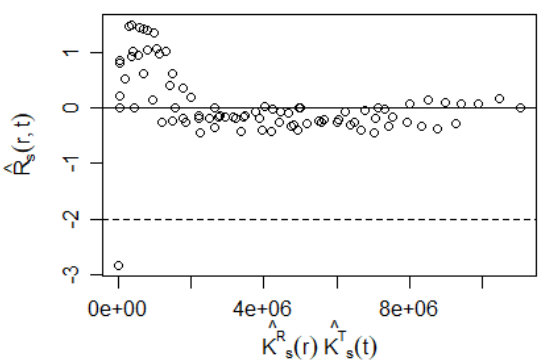
$\widehat{D}_s(r,t)$



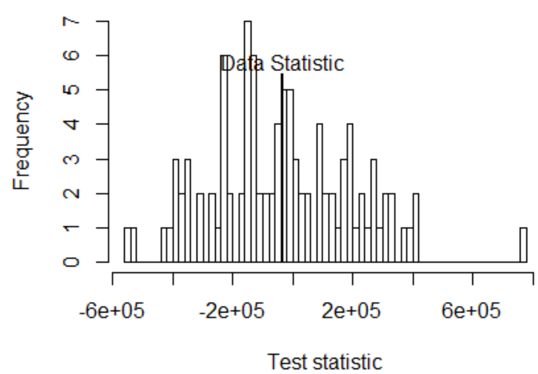
$\widehat{D}_s^0(r,t)$



Standardised Residuals

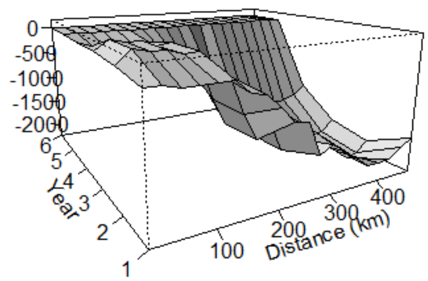


Monte Carlo test

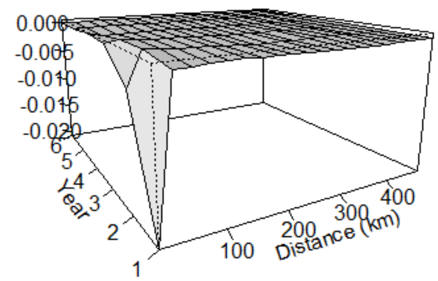


Sector 31 – Manufacture of furniture

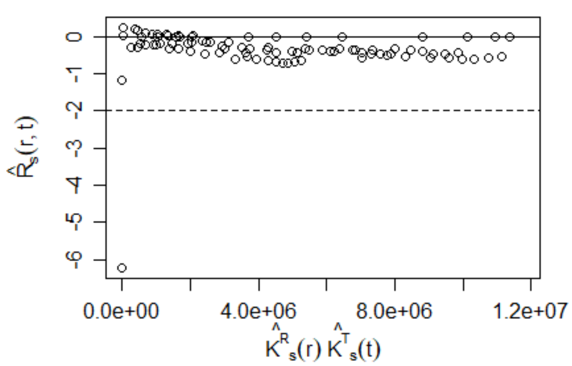
$\widehat{D}_s(r,t)$



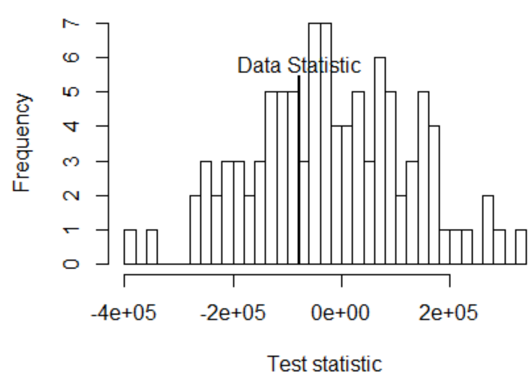
$\widehat{D}_s^0(r,t)$



Standardised Residuals

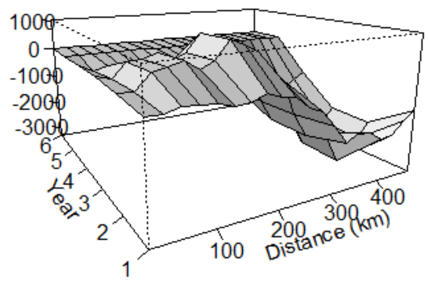


Monte Carlo test

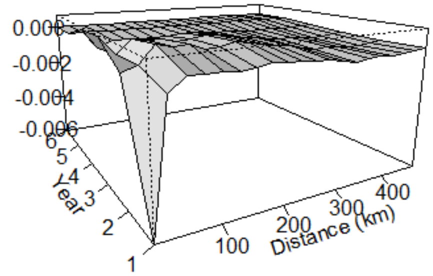


Sector 32 – Other manufacturing

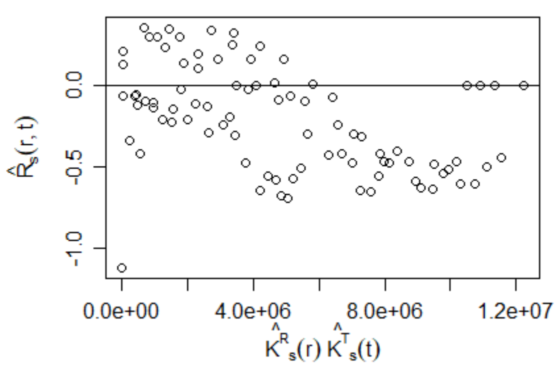
$\widehat{D}_s(r,t)$



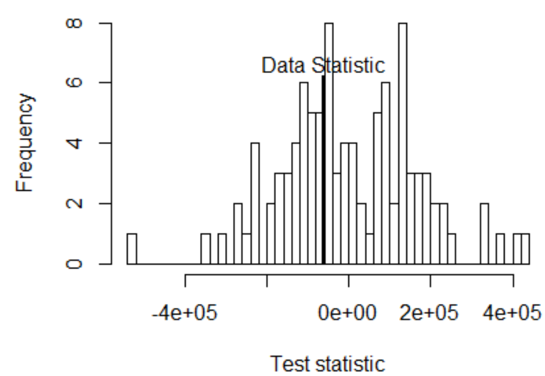
$\widehat{D}_s^0(r,t)$



Standardised Residuals

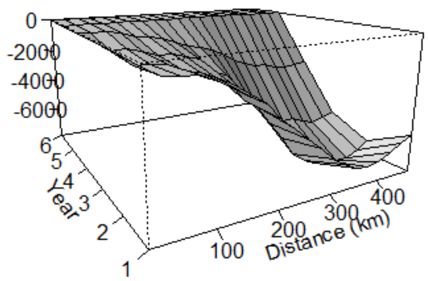


Monte Carlo test

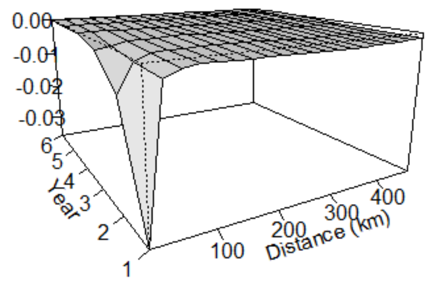


Sector 33 – Repair and installation of machinery and equipment

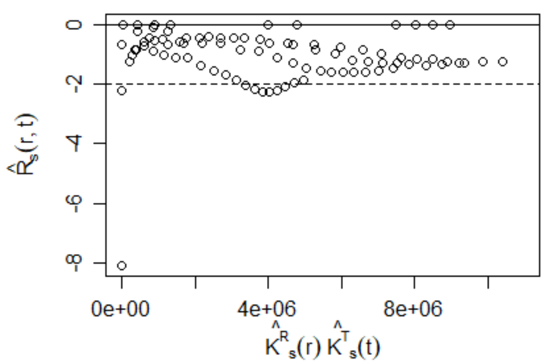
$\widehat{D}_s(r,t)$



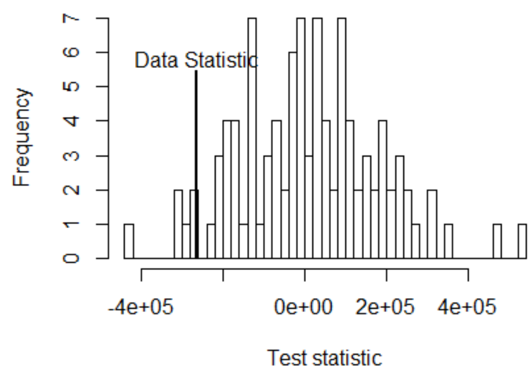
$\widehat{D}_s^0(r,t)$



Standardised Residuals



Monte Carlo test



Notes: Authors' elaboration on *AIDA* data. The plots report the estimated values of the $\widehat{D}_s(r,t)$ functional, the $\widehat{D}_s^0(r,t)$ functional, the 'standardised residuals' $\hat{R}_s(r,t)$ versus $\widehat{K}_s^R(r)\widehat{K}_s^T(t)$, and the empirical frequency distribution of the sum of the differences between the space-time *K*-function and the product of the separate space and time *K*-functions in 99 simulations resulted from the Monte Carlo test.

Chapter 2

Related variety and economic growth at firm level in China

Yuting Jiang

Department of Economics and Management "Marco Fanno",

University of Padova, Padova, Italy

Yuting Jiang

Department of Economics and Management "Marco Fanno",

University of Padova

Via del Santo 33, 35123 Padova, Italy

E-mail: yuting.jiang@studenti.unipd.it

Related variety and economic growth at firm level in China

Abstract: Using a sample of 84,868 Chinese manufacturing firms operating in the period 2006-2013, we analyze whether related and unrelated variety affect firms' economic performance. Our results show that, correcting only for sample-selection, unrelated variety has a negative and statistically significant impact. Accounting also for the endogeneity of our two main explanatory variables – related and unrelated variety – we find that these two variables become insignificant. A positive effect for related variety and negative for unrelated variety is detected only when we consider high-developed Chinese regions. Finally, a positive effect of related variety is identified for large firms.

Keywords: agglomeration; firm economic growth; related and unrelated variety;

JEL Codes: C21, C36, N65, O12, R11

1.INTRODUCTION

Spatial agglomeration is one of the main driver behind regional economic growth (Desrochers and Leppälä, 2010). The literature has identified two types of agglomeration externalities: (i.) localization and (ii.) diversification economies (Glaeser *et al.*, 1992). This distinction is commonly referred to as the Marshall-Arrow-Romer (MAR) model against the Jacobs externalities (Glaeser *et al.*, 1992). The MAR theory suggests that knowledge spillovers take place between firms belonging to the same industry (localization economies). On contrary, the Jacobs theory (Jacobs, 1969) suggests that knowledge spillovers take place between firms belonging to different industries (diversification economies).

A recent stream of literature has developed the new concept of related variety (Frenken *et.al,* 2007). According to this new perspective what really matters is not diversification *per se* but related variety: i.e., knowledge spillovers between firms operating in 'different but related' sectors. In other words, inter-industry knowledge spillovers – i.e. the cross-fertilization of ideas, knowledge and technologies across industries – take place between sectors sharing the same knowledge and technological base, made it possible that firms in the correlated industries co-located, thereby generates the industry agglomeration phenomenon, which results in the regional economic growth as well as individual economic growth.

There have been a bunch of literature about firm economic growth, Gibrat's Law is one of the applied methods. According to the Gibrat's Law, a commonly accepted interpretation is proportionate effect-that the growth rate of a given firm is independent form its beginning size. We start from this classical equation, testing if firms beginning size has effects on its economic growth. However, the Gibrat's Law is not able to capture the inside sector structure, as the firms we choosed in our research including almost all 2-digit sectors from manufactory industry, other than a specific sector. Therefore, we construct an augmented regression by adding two variables-related and unrelated variety,which are able to measure the industry structure inside, to see if these two variables have significant effects on firm level economic growth. To the best of our knowledge, which hasn't be done by previous literature. Besides, we would like to see after controlling the impact of related and unrelated variety, the Gibrat's Law holds or not.

The aim of this paper is to investigate the effect of related and unrelated variety on the economic growth of a sample of 84,868 Chinese manufacturing firms in the period 2006-2013. This is done by estimating a firm economic proportional growth equation *à la* Gibrat which also include these two measures of agglomeration.

This study intends to contribute to the debate about the role played by related variety on the economic growth at the firm level in an emerging country. In fact, during these years (2006-2013) the Chinese economy experienced a period of deep productive and technological transformation, with an acceleration of the process of economic growth.

The paper is organized as follows. In Section 2 we discuss the literature about the Gibrat's Law and the concept of related variety. In Section 3 we present the dataset and the econometric methodology adopted. Section 4 describes and discusses the main results of our analysis. Section 5 concludes the work.

2. RELATED LITERATURE

2.1. The Gibrat's law

The Gibrat's law – sometimes called as Gibrat's rule of proportionate growth – was first developed by Gibrat (1931). In its original version hypothesized that, a skewed distribution consisting in a large number of additive and independent variables could be converted into a normal distribution by transforming the initial variables with a logarithmic function. After Mansfield (1962) and Chesher (1979), the idea of the Gibrat's law evolved into a test for the proportional growth: a theory that many studies have used as starting point of their empirical analysis. To date the common interpretation of the Gibrat's law is slightly different from its original version. In fact, it states that a firm's proportional growth rate is independent of its absolute size. In other words, according to this law, small and large firms should grow at the same rate. This version of the Gibrat's law was initially tested by Mansfield (1962), who investigated three different industries. From this contribution, many studies (Wagner,1992; Geroski, 1995; Caves, 1998) have estimated this equation. In many cases the evidence does not support the law (Reid, 1995; Audretsch, 1995; Harhoff *et al.*, 1998; Weiss, 1998; Audretsch *et al.*, 1999; Almus and Nerlinger,2000; Calvo, 2006).

Some authors hypothesize that the rejection of the law is caused by the fact that small firms have generally a high probability of dying. Using quantile regression techniques, Lotti *et. al.* (2003) show that the Gibrat's Law holds for new entrants. In other words, estimates based on surviving firms could be affected by a sample selection bias, which tends to magnify the rapid growth of smaller firms.

For the sake of clarity, this paper use Gibrat's law as a starting point for investigating the firm'economic growth in China. Referring to the common understanding of Gibrat's law, we test if the firm's economic growth rate is independent of its initial income level. Generally, Gibrat's law holds if firm growth is independent respect to determinants such as firm age and size. The Gibrat's hypothesis is that the estimated coefficients of the income level at the beginning and the firms age are not different from zero (Maine *et al.*, 2010).

Although Gibrat's law provide a useful framework for testing proportionate growth, it is not able to fully capture the determinants of firm income growth. For this reason, we try to incorporate in our specifications measures for agglomeration externalities. This in order to investigate their effects on firm income growth, and to see if the Gibrat's law will hold in specifications with more variables.

2.2. Agglomeration and related variety

The analysis of agglomeration economies dates back to Marshall (1920). The literature has identified two types of agglomeration externalities: (i.) localization and (ii.) diversification economies (Glaeser *et al.*, 1992). This distinction is commonly referred to as the Marshall-Arrow-Romer (MAR) model against Jacobs externalities (1969). The MAR theory suggests that knowledge spillovers take place between firms belonging to the same industry. The geographic concentration of an industry facilitates the transmission of knowledge, information, and technologies among economic agents, thus promoting both knowledge spillovers among firms and incremental and process innovations. These local externalities are generally referred to as localization economies. On the contrary, the Jacobs theory (Jacobs, 1969) suggests that knowledge spillovers take place between firms belonging to different industries. The diversity/variety of the industrial structures at the regional level promotes the exchange and the cross-fertilization of information, ideas and technologies, which in turn promote radical and product innovations. These local externalities are generally referred to as diversification economies. There has been a long debate on which of these two agglomerative forces – localization or diversification – mostly contribute to the economic development of a region. Despite the high number of papers on this topic, the conclusions of this empirical literature are not univocal and conclusive.

Recently, a new stream of literature, which has gained more and more attention, has developed the concept of related variety (Frenken *et.al,* 2007). According to this approach what really matters is not diversification/variety per se, but related variety. This concept assumes that knowledge spillovers within a region/local system occur among firms operating in 'different but related' sectors. In fact, the differentiated industrial mix in a local system/region can improve the opportunity to interact, copy, modify, and recombine ideas, practices, and technologies across sectors. Geographic proximity among firms makes this process of recombining existing pieces of knowledge in totally new ways more likely to occur. The recombination leads to new products and services. The process occurs only if firms share the same technological and knowledge base. In fact, knowledge spreads among firms operating in different sectors only if the cognitive distance is not too large (Nooteboom, 2000). In other words, knowledge spillovers take place only if some of the sectors in a local system/region are complementary in terms of shared competences, knowledge, and technologies.

There are a bunch of empirical studies focused on the relationship between related variety and regional economic growth. Brachert *et. al.* (2011) show how related variety is one of the main sources of German regional employment growth during the period 2003-2008. In other words, they find that related sectors foster regional economic growth. Firgo and Mayerhofer (2016) study related variety and employment growth by using highly disaggregated data at sub-regional level. They find that unrelated variety positively affects employment growth in Austria. Taking into account for the sectoral heterogeneity, Mameli *et al.*, (2012) show that related variety seems to influence more intensively knowledge intensive service sectors rather than manufacturing industries. Analysing the impact of related variety in Finland, Hartog *et al.*, (2012) do not find evidence of a role of related variety in influencing the employment growth. Only after decomposing between high-tech and low/medium-tech sectors, they find a positive impact of related variety on employment growth for high-tech sectors. Bishop and Gripaios (2010) argue that looking at the effect of related variety on regional employment growth, even distinguishing between manufacturing and services might be an oversimplification. Since sectors are heterogeneous, the mechanisms behind knowledge spillovers can differ between sectors.

A recent study by Aarstad *et. al.* (2016), conducted using a multi-level analyses on Norwegian data, show that related industrial variety has a positive effect on firm innovation, while unrelated variety has the opposite effect on productivity.

Studies on China's economy which investigate the effects of related and unrelated variety on firm level performance are very few. The only exception is the work of Howell *et al.* (2018) which investigates the effects of related and unrelated variety on new entrepreneurial firms' survival chances in China. And to our knowledge, no previous studies on related and unrelated variety address the potential endogeneity of these two variables.

Considering of these, this paper wants to contribute to this stream of literature investigating the effect of related variety on the economic performance of Chinese firms. Besides, we also take into account the endogeneity of related and unrelated variety, using Bartik instrument correcting the endogeneity issue caused by these two variables.

3. THE DATASET AND THE ECONOMETRIC METHODOLOGY

3.1. The dataset

This study is focused on the firm level economic growth of a sample of Chinese manufacturing firms during the period 2006-2013. Our main data source is the non-listed enterprise database, which is a firm-level annual micro-database, established and implemented according to the standards published by the Chinese National Bureau of Statistics. It covers a wide range of economic activities such as the extractive and the electricity industry, all manufacturing sectors and the gas and water production. For all the firms belonging to these sectors are available economic and financial information such as production, sales, number of employees, geographic location and so on. Only firms with an annual income in the main business above 5 million of yuan are included in the dataset.

In order to construct our dataset we select two years: 2006 and 2013. We eliminate observations with missing information or invalid information about location (district level), industry category (2-digit level, according to the national industrial classification - 2003), production, sales and employment. Our focus is on those firms which exist both in 2006 and 2013. We drop those starting up after 2006. Firms which changed their locations were also eliminated.

This left us with a final sample of 84,868 observations, covering almost all 2-digit manufacturing industries from sector 13 to sector 43, except Tobacco (sector 16) and Oil processing, coking and nuclear fuel processing (sector 25).

3.2 Measuring related and unrelated variety

We use entropy measures to define related and unrelated variety (Frenken *et al.*, 2007). The main advantage of an entropy measure is that it can be decomposed at each digit sectoral level. The decomposable nature of entropy measure implies that variety can entry into a regression without causing collinearity (Jacquemin *et al.*, 1979; Attaran, 1986). Unrelated variety in each city is computed by the entropy at 2-digit level, related variety is computed by the weighted sum of entropy at 4-digit level within each 2-digit sector.

Be more formally, let all four-digit sectors g located in district d at time $t = 2006$, fall exclusively into a two-digit sector j, where $j = 1, \dots, J$. The two-digit shares, P_{jdt}, is the sum of four-digit shares p_{gdt}:

$$P_{jdt} = \sum_{g \in j} p_{gdt} \qquad (1)$$

The unrelated variety (UV), or the entropy at two-digit level, is given by:

$$\mathrm{UV}_{dt} = \sum_{J=1}^{J} P_{jdt} log_2 \left(\frac{1}{P_{jdt}}\right) \qquad (2)$$

Related variety (RV), as the weighted sum of entropy within each two-digit sector is given by:

$$\mathrm{RV}_{dt} = \sum_{j=1}^{J} P_{jdt} \times [\sum_{g \in j} \frac{p_{gdt}}{P_{jdt}} log_2 (\frac{1}{p_{gdt}/P_{jdt}})] \qquad (3)$$

3.3. The econometric methodology

We investigate the effect of related and unrelated variety on to the economic growth at the firm-level during the period 2006-2013 in China. To our knowledge, this is the first analysis of the Gibrat's law for the Chinese economy, which also includes agglomeration measures. The firm level economic growth equation is defined as follows:

$$\text{Income Growth}_{ijd} = \alpha + \beta \log(\text{Income}_{ijdt}) + \rho\, \text{age}_{ijdt} + \text{v}\, \text{V}D_{ijdt} + \delta \log(popdens_{ct}) + \gamma\, \text{R}V_{dt} + \delta \text{U}V_{dt} + \epsilon_p + \theta_j + \vartheta_s + \varepsilon_{ijdt} \quad (4)$$

The dependent variable is firm's income growth, which is defined (in logs) as: $(\text{T} = 2013, \text{t} = 2006)$: $\text{Income Growth}_{ijd} = \ln(\text{income}_{ijd,T}) - \ln(\text{income}_{ijd,\text{t}})$. This variable represents the economic growth of firm i operating in the 2-digit sector j and located in district d between the year $\text{t} = 2006$ and $\text{T} = 2013$. We use the main business income of each firm to proxy its income level.

In addition to related variety and unrelated variety, a set of other explanatory variables include: (i.) the income level in the initial year 2006 (Income_{ijdt}); (ii.) firms' age (age_{ijdt}) defined as 2006 minus the firm's start operation year; (iii.) the firm's vertical disintegration level ($\text{V}D_{ijdt}$) defined as the ratio between industrial intermediate input and gross output value in 2006:

$$\text{Vertical Disintegration}_{ijdt} = \frac{\text{purchased intermediate input}_{ijdt}}{gross\ output\ value_{ijdt}} \quad (5)$$

To control for geographic heterogeneity, we introduce in our main specifications a measure of population density ($\text{popden}s_{ct}$) calculated as the 2006 population in city c per square-kilometer. This variable – which is also taken in log form – is a proxy for urbanization economies.

The term ϵ_p denotes a set of geographic dummies defined at provincial level in order to capture systematic differences across geographic areas in terms of natural resources, public infrastructures, social capital, industrialization and policy efficiency. In fact, administrative divisions in China consists of several levels. The provincial level is the first level which includes provinces, autonomous regions, municipalities and special administrative regions. The second level is the prefectural level – city level – which includes prefecture-level cities and prefectures. In 2019 there are 34 provincial units (23 provinces, 5 autonomous regions, 4 municipalities and 2 special administrative regions) and 333 prefectural units. A city is further divided in different districts. Central and local governments have different place-based policies such as tax incentives, public subsidies. Natural resources, technology, education and the health system can vary significantly across regions.

The term θ_j denotes a set of industry dummies defined at 2-digit level. They are introduced in order to control for productive, organizational and technological differences. As the returns to scale theory points out, firm size influences its economic growth. For this reason, the term ϑ_s denotes a set of dummy variables, defined according to the firms' size (measured in terms of employees). Specifically, we consider four size dummies: (i.) small firm are those with a number of employees between 0 and 50; (ii.) medium firms are those with a number of employees between 50 and 95; (iii.) medium-large firms are those with a number of employees between 95 and 200; and finally (iv) large firms are those with a number of employees above 200.

3.4. Identification strategy

There are two main econometric problems in our estimation procedure. The first concerns the sample selection. The second is related with the endogeneity of our two main explanatory variables: related and unrelated variety.

We estimate our firm level income growth equation adopting a Heckman (1979) two-steps sample selection model. Only for the sub-sample of survived firms over the period 2006-2013 we observe the income growth. This means that if we estimate our main equation only for this sub-sample of firms using an OLS we would get biased estimates. There is a clear-cut sample-selection problem. For this reason, we adopt a two-steps sample selection method in order to account for this problem. To capture the non-random survival of firms during year 2006 and 2013, we first estimate a probit regression for firm survival, where the dependent variable is a binary variable, taking value 1 if the firm is observed both at the beginning (t=2006) and the end of the period (T=2013) and 0 otherwise. Then we estimate the augmented firm economic growth equation (4) including the Inverse Mills Ratio. The firm survival is modeled as an unknown non-linear function (Griffith *et al.*, 2009) on firm size, fixed asset, output value, and total profit (all in logs) in the beginning year 2006. These firm level characteristics are suitable excluded variables from the economic growth equation that affect the probability if firm survival. As the non-linear functional form determining a firm's exit decision is unknown, we follow Olley and Pakes (1996) and Pavcnik (2002) by adopting a semiparametric specification, which approximates the unknown function with a polynomial expansion in firm size, log fixed asset, log output value, log total profit and their interactions.

Summing up, the Heckman model is estimated as follows. First, the probit model is estimated for the whole sample; then the inverse Mills ratio (λ) obtained from the selection equation is added to the economic growth equation (4) as an additional regressor, thus correcting for the sample selection bias. Finally, the augmented version of the economic growth regression is estimated by OLS using the sub-sample of surviving firms in the 2006-2013 period. Through the Heckman two-steps method, we can obtain unbiased and consistent estimates for the growth equation (4).

Despite the correction for the sample selection, our estimates might still be biased for the (potential) presence of endogeneity of related variety and unrelated variety. There are several reasons for the endogeneity in these two variables. One is certainly reverse causality. For example, related and unrelated variety could explain firm's economic growth, but at the same time, firm's economic growth may induce leadership effects, which may attract other up-stream and down-stream firms to set up around this rapid developed firm, or the firm economic growth is due to the advantage of its location, with big potential market and less production costs, it will give rise to more firms which producing similar goods come nearby. Both will lead to clustering/agglomeration processes. Related and unrelated variety could be generated in this way. Also exogenous shocks can affect firms' economic performance and regional industrial distribution simultaneously.

To account for the endogeneity problem, we follow the strategy of Autor and Duggan (2003), a modification of Bartik's (1991) shift-share approach. The main idea of this approach is that each industrial sector would have experienced at local level (in our case the district level) the same dynamics (in terms of employment) experienced at national level over the period 2000-2005 without sector specific or local city level shocks. Put differently, the instrument variables should exclude any shock associated with the event that China join in the WTO in 2001, which are specific to both the industrial sector and local area. According to this rationale, we construct the instruments for related and unrelated variety respectively. The instrument $\mathrm{I}V_{rv}$ accounts for the sectoral variations at four-digit level within a same two-digit sector, and is defined as :

$$\mathrm{IV_{rv}} = \sum_{\substack{g=1 \\ g\in j}}^{G} \{ \left(\frac{n_{gd(t-6)}}{\sum_{\substack{g=1 \\ g\in j}}^{G} n_{gd(t-6)}} \right) [\log(n_{g(-d)(t-1)}) - \log(n_{g(-d)(t-6)})] \} \quad (6)$$

Where $n_{gd(t-6)}$ denotes the number of employees in a four-digit sector g within a two-digit sector j (g ∈ j), and located in the district d at time $\mathrm{t} - 6 = 2000$; the term $n_{g(-d)(t-1)}$ and $n_{g(-d)(t-6)}$ denote the number of employees working in a four-digit sector g at the national level excluding the district d, at time $\mathrm{t} - 1 = 2005$ and $\mathrm{t} - 6 = 2000$ respectively. We choose year 2000 (before Chinese accession to the WTO) to construct the "share" component, in order to capture original state before the shock (WTO membership). While "shift" component is defined with the period 2000-2005, considering one year of lag (at 2005) to relax endogeneity issue. The instrument for related variety specified in equation (6) calculates for each four-digit sector g falling within a two-digit sector j, the shares of employments in district d. It means that the share referring to the year 2000 of each four-digit sector at local level changes relying on the specific district considered. Then these four-digit sector shares multiplied by the change for the employment for the same four-digit sector at national level but without the reference district d, during the period 2000-2005, are summed over the corresponding 4-digit sectors. Thus, the instrumental variable captures dynamics which are specific to the four-digit sectors within each two-digit sector for each district.

The instrumental variable ($\mathrm{I}V_{uv}$) accounts for the variation at unrelated two-digit sector level is defined as follows:

$$\mathrm{I}V_{uv} = \sum_{j=1}^{J} \{ \left(\frac{n_{jd(t-6)}}{\sum_{j=1}^{J} n_{jd(t-6)}} \right) [\log(n_{j(-d)(t-1)}) - \log(n_{j(-d)(t-6)})] \} \quad (7)$$

where $j = 1,2 \dots, J$ denotes two-digit sectors. In this case, the instrumental variable is defined by calculating, for each two-digit sector j located in district d, considering the shares in term of number of employees working for that specific two-digit sector j and located in district d among all two-digit sectors' employment within district d. This share of each two-digit sector, defined for the year 2000 (t-6=2000) changes depending on the particular district d concerned. Once again, the "shift" term is calculated by the rate of change in terms of number of employees observed for the same two-digit sector and at the national level, but excluding the district d of reference, during period 2000-20005. Then the "shift" part multiplied by the "share" part we mentioned above, are summed over the corresponding 2-digit sectors, which constructs our instrument for un-related variety, capturing the dynamics which are particular to each two-digit sector and each district.

Finally, we follow the method proposed by Wooldridge (2010) to solve such estimation with sample selection and endogeneity issue addressed simultaneously. First, a reduced-form selection equation is estimated by a probit model with the set of external instrumental variables ($\mathrm{I}V_{rv}$, $\mathrm{I}V_{uv}$) and the non-linear form exclusion restriction added to the exogenous variables entering Equation (4), and excluding the two endogenous variables ($\mathrm{R}V_{jct}$, $\mathrm{U}V_{jct}$). Second, the firm economic growth equation is estimated via Two-Stage Least Square (2SLS) regression with the Inverse Mills Ratio obtained from the first-stage selection model as an additional regressor. In addition, we also estimate equation (4) with a Generalized Method of Moments (GMM) approach for a comparison. Standard errors are clustered at district level in all specifications, which allows the error term to be correlated across firms within each district. (Bertrand *et al.*, 2004)

The endogeneity of these two variables – related and unrelated variety – is tested by using the Durbin χ^2 statistics and the Cragg-Donald Wald F statistics (Wooldridge, 2010). The null hypothesis is that the variables are exogenous. This hypothesis is rejected by our tests in both specifications. This means that these two variables are indeed endogenous in the equation. According to the results for the weak identification test, the Wald F statistics are above the rule of thumb value 10 in all specifications. Thus, there is no weak instrumentation in our case.

In order to check if the results still hold for different size firms, we split the whole sample into four sub-samples according to the firm size (already defined). In addition, we focus on firms located in well-developed regions: the top 3 Chinese developed regions according to gross regional production in the beginning year 2006. Table 3 shows the results for different size firms referring to IV-TSLS estimation. Table 4 is the estimation results referring to IV-TSLS method for firms located in well-developed region.

4. EMPIRICAL RESULTS

Table 1 reports the OLS estimates of income growth equation correcting for the sample selection bias. Bootstrapped standard errors are clustered at district level in all specifications. This is done in order to allow the error term to be correlated across firms within each district. Specifically, column (1) is the results of our baseline specification without adding neither related variety nor unrelated variety. Column (2) and column (3) are results for the specifications with related variety and unrelated variety, respectively. Column (4) is the results introducing both. The inverse mills ratio (lambda) is negative and statistically significant. This means that we need to correct for sample selection. The results reported in Table 1 suggest a strong negative effect of unrelated variety on firm's income growth. This result is in line with previous studies about the role played by related and unrelated variety on regional employment growth (Frenken *et al.*, 2007; Saviotti and Frenken, 2008).

[---Table 1 near here---]

The firms' income level at the beginning negatively affects the income growth in the future as the estimated coefficients of the variable "$\log(\text{Income}_{ijdt})$" are all negative and statistically significant in all specifications, and the estimated elasticities are almost similar, around 30%. Since the estimated coefficient for the firm income level at beginning year is different from 0, we say that the Gibrat's law does not hold in our case. Smaller firms grow at a different rate respect to larger firms. Taking into account the impact of firm's age and vertical disintegration, we find that they are negative and significant. For the size dummy variables, it emerges that medium-small, medium-large and large firms tend to experience larger income change than small firms. It confirms the Gibrat's test, which indicates an un-proportional increase of firms' income. Compared to small firms, firms survived during the financial crisis, the larger they are, the worse they perform. This is shown from the estimated coefficients for medium, medium-large and large firms, which are all negative and highly statistically significant. Moreover, the magnitude increases with firms' size. This finding is not surprising. Small firms, which survived during economic crises, have experienced generally less shocks to their economic performance compared to larger firms. A possible explanation is that, small firms are usually less internationalized and thus less exposed to external shocks than larger firms which are generally more involved in international activities. For this reason, they were less hit by the financial crisis and have had a better economic performance.

[---Table 2 near here---]

Results for the IV estimation which correct for the endogeneity issue are reported in Table 2. Column (1) reports the results with IV Two-Stage-Least-Square method, column (2) results of the IV Generalized Method of Moment (GMM). Both methods get quite similar results after 500 replications of the bootstrap standard errors. They suggest that related and unrelated variety do not significantly influence firm level economic growth anymore after correcting for the endogeneity of our two main independent variables. The estimated negative and significant coefficient for the inverse Mills ratio (lambda) suggest the need for correcting sample selection. The endogeneity of variables related variety and unrelated variety is tested by employing Durbin χ^2 statistics. This test rejects the null hypothesis: variables are not exogenous. For the Cragg-Donald Wald F statistic, the values are higher than the conservative cut-off value of 10 in both specifications. The sign of the estimated coefficients for unrelated variety is still negative (as in the OLS estimates) and coincide with previous studies (Frenken *et al.*, 2007). The magnitude is larger, around 0.05 rather than 0.025 in OLS estimation. As it is negative, we can see the OLS estimation is upward biased.

The results of Table 1 and 2 allow us to distinguish between agglomeration externalities generated from different but related sectors and from totally different and not related sectors for the whole sample. In the sense that we can identify between these two agglomeration externalities, which accounts relatively more across firms within a localized industry on their income growth.

[---Table 3 near here---]

Table 3 reports the IV-TSLS estimates for different size firms. The estimated coefficients vary a lot between groups. Inverse Mills ratios are positive for small and medium small firms, but negative for medium-large and large firms. Related variety is significantly positive just for large firms. Unrelated variety is not significant in all sub-groups. This is in line with the results for the whole sample. Firm level characteristics, such as the firm income level at beginning, age and vertical disintegration are significant for medium-large and large firms.

[---Table 4 near here---]

We also run a sub-sample analysis for the top 3 developed regions according to gross regional production value in 2006-results are presented in table 4. The negative and significant coefficients for Inverse Mills Ratio indicate that the correction for sample selection is necessary. However, differently from the analysis for the whole sample, related variety has a positive and statistically significant impact on the firm-level economic growth. On the contrary, unrelated variety has a negative and significant impact on firm economic growth when we only consider firms located in high developed regions in China. Firm specific characteristics such as the income level in the beginning year 2006, age and vertical disintegration are all negative and significant.

5. CONCLUSIONS

Using a sample of 84,868 Chinese manufacturing firms during the period 2006-2013, this paper tested the Gibrat's Law jointly with two agglomeration variables, related variety and unrelated variety (Frenken *et al.*, 2007). Our results show that the Gibrat's Law is rejected by data even including agglomeration externalities. Furthermore, we analyze the roles played by these two variables on firm-level economic growth. The results show that, unrelated variety has a significant and negative effect on firm economic growth only when we correct for the sample selection bias. After correcting for endogeneity, both related and unrelated variety become insignificant. Disaggregating our dataset according to firm size and regions where they are located we find more interesting results. For large firms, related variety seem to have significant positive effect. This is true also for firms located in high-developed Chinese regions where related variety significantly and positively influence firm economic performance, while unrelated variety has a negative impact. As shown by some previous studies, local knowledge spillovers are more easily absorbed by firms operating in similar but different sectors. Our results seem to confirm that this kind of knowledge spillovers are the main driven forces to the economic growth of Chinese manufacturing firms in high-developed regions rather than in rural areas or less-industrialized regions.

Our research has some limitations. First, it is an analysis based on a cross-section. Panel datasets would allow us to conduct a more sophisticated analysis of the dynamics of firms' economic growth: for example, taking into account time lags. Second, the sample period is short, from 2006 to 2013, and the Great Recession is within this period. China experienced significant changes also during this period of time.

To conclude, we enrich the empirical literature on the relationship between agglomeration and firm economic performance referring to a developing country-China. According to our results, after correcting for sample selection bias and addressing for the endogeneity issue, the impact of related and unrelated variety on firm level economic performance is mixed. Gibrat's law do not hold even incorporating the agglomeration externalities. Previous studies found that related and unrelated variety had significant impact on economic growth at regional level. While according to our research, we do not find these effects on firm level economic growth. The firm level economic growth seems to rely mainly on firm specific characteristics such as the initial income level, age, and vertical disintegration.

REFERENCES

Aarstad, J., Kvitastein, O. A., & Jakobsen, S. E. (2016). Related and unrelated variety as regional drivers of enterprise productivity and innovation: A multilevel study. Research Policy, 45(4), 844-856.

Almus, M. (2000). Testing" Gibrat's Law" for young firms–empirical results for West Germany. Small Business Economics, 15(1), 1-12.

Attaran, M. (1986). Industrial diversity and economic performance in US areas. The Annals of Regional Science, 20(2), 44-54.

Audretsch, D. B. (1995). Innovation, growth and survival. International journal of industrial organization, 13(4), 441-457.

Audretsch, D. B., Santarelli, E., & Vivarelli, M. (1999). Start-up size and industrial dynamics: some evidence from Italian manufacturing. International Journal of Industrial Organization, 17(7), 965-983.

Autor, D. H., & Duggan, M. G. (2003). The rise in the disability rolls and the decline in unemployment. The Quarterly Journal of Economics, 118(1), 157-206.

Bartik, T. J. (1991). Who benefits from state and local economic development policies?

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? The Quarterly journal of economics, 119(1), 249-275.

Bishop, P., & Gripaios, P. (2010). Spatial externalities, relatedness and sector employment growth in Great Britain. Regional Studies, 44(4), 443-454.

Boschma, R., Minondo, A., & Navarro, M. (2012). Related variety and regional growth in Spain. Papers in Regional Science, 91(2), 241-256.

Brachert, M., Kubis, A., & Titze, M. (2011). Related variety, unrelated variety and regional functions: Identifying sources of regional employment growth in Germany from 2003 to 2008(No. 15/2011). IWH Discussion Papers.

Calvo, J. L. (2006). Testing Gibrat's law for small, young and innovating firms. Small business economics, 26(2), 117-123.

Cainelli, G., & Ganau, R. (2019). Related variety and firm heterogeneity. What really matters for short-run firm growth? Entrepreneurship & Regional Development, 1-17.

Caves, R. E. (1998). Industrial Organization and New Findings on the Turnover and. Journal of economic literature, 36(4), 1947-1982.

Chesher, A. (1979). Testing the law of proportionate effect. The Journal of industrial economics, 403-411.

Maine, E. M., Shapiro, D. M., & Vining, A. R. (2010). The role of clustering in the growth of new technology-based firms. Small Business Economics, 34(2), 127-146.

Ellison, G., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? Evidence from co-agglomeration patterns. American Economic Review, 100(3), 1195-1213.

Ellison, G., & Glaeser, E. L. (1999). The geographic concentration of industry: does natural advantage explain agglomeration? American Economic Review, 89(2), 311-316.

Firgo, M., & Mayerhofer, P. (2018). (Un) related variety and employment growth at the sub-regional level. Papers in Regional Science, 97(3), 519-547.

Frenken, K., Van Oort, F., & Verburg, T. (2007). Related variety, unrelated variety and regional economic growth. Regional studies, 41(5), 685-697.

Geroski, P. A. (1995). What do we know about entry? International Journal of Industrial Organization, 13(4), 421-440.

Glaeser E. L., Kallal H. D., Scheinkman J. A. & Shleifer A. (1992), Growth in Cities. Journal of Political Economy, 100, 6, 1126-1152.

Glaeser, E. L., & Resseger, M. G. (2010). The complementarity between cities and skills. Journal of Regional Science, 50(1), 221-244.

Glaeser, E. L. (Ed.). (2010). Agglomeration economics. University of Chicago Press.

Glaeser, E. L., & Mare, D. C. (2001). Cities and skills. Journal of labor economics, 19(2), 316-342.

Gibrat, R. (1931). Les inégalits économiques. Sirey.

Griffith, R., Redding, S., & Simpson, H. (2009). Technological catch - up and geographic proximity. Journal of Regional Science, 49(4), 689-720.

Harhoff, D., Stahl, K., & Woywode, M. (1998). Legal form, growth and exit of West German firms—empirical results for manufacturing, construction, trade and service industries. The Journal of industrial economics, 46(4), 453-488.

Hartog, M., Boschma, R., & Sotarauta, M. (2012). The impact of related variety on regional employment growth in Finland 1993–2006: high-tech versus medium/low-tech. Industry and Innovation, 19(6), 459-476.

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica: Journal of the econometric society, 153-161.

Henderson, J. V. (2003). Marshall's scale economies. Journal of urban economics, 53(1), 1-28.

Holmes, T. J., & Lee, S. (2010). Cities as six-by-six-mile squares: Zipf's law? In Agglomeration economics (pp. 105-131). University of Chicago Press.

Howells, J., & Bessant, J. (2012). Introduction: innovation and economic geography: a review and analysis. Journal of economic geography, 12(5), 929-942.

Howell, A., He, C., Yang, R., & Fan, C. C. (2018). Agglomeration,(un)-related variety and new firm survival in China: Do local subsidies matter?. Papers in Regional Science, 97(3), 485-500.

Jacobs, Jane (1969), The Economy of Cities, New York: Vintage Books.

Jacquemin, A. P., & Berry, C. H. (1979). Entropy measure of diversification and corporate growth. The journal of industrial economics, 359-369.

Lotti, F., Santarelli, E., & Vivarelli, M. (2003). Does Gibrat's Law hold among young, small firms? Journal of evolutionary economics, 13(3), 213-235.

Mameli, F., Iammarino, S., & Boschma, R. (2012). Regional variety and employment growth in Italian labour market areas: services versus manufacturing industries.

Mansfield, E. (1962). Entry, Gibrat's law, innovation, and the growth of firms. The American economic review, 52(5), 1023-1051.

Maine, E. M., Shapiro, D. M., & Vining, A. R. (2010). The role of clustering in the growth of new technology-based firms. Small Business Economics, 34(2), 127-146.

Marshall, A., & Marshall, M. P. (1920). The economics of industry. Macmillan and Company.

Olley, S. and Pakes, A.(1996). The dynamics of productivity in the telecommunications equipment industry. Econometrica, 64(6), 1263–1298.

Pavcnik, N. (2002). Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants. The Review of Economic Studies, 69(1), 245-276.

Reid, G. C. (1995). Early life-cycle behaviour of micro-firms in Scotland. Small Business Economics, 7(2), 89-95.

Rosenthal, S. S., & Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. Handbook of regional and urban economics. Elsevier, 2004((4), 2119-2171.

Saviotti, P. P., & Frenken, K. (2008). Export variety and the economic performance of countries. Journal of Evolutionary Economics, 18(2), 201-218.

Van Oort, F. G., & Atzema, O. A. (2004). On the conceptualization of agglomeration economies: The case of new firm formation in the Dutch ICT sector. The Annals of Regional Science, 38(2), 263-290.

Wagner, J. (1992). Firm size, firm growth, and persistence of chance: Testing GIBRAT's law with establishment data from Lower Saxony, 1978–1989. Small Business Economics, 4(2), 125-131.

Weiss, C. R. (1998). Size, growth, and survival in the upper Austrian farm sector. Small Business Economics, 10(4), 305-312.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Table 1 Sectoral Variety and Firm Economic Growth---OLS Method

Estimation Method	OLS	OLS	OLS	OLS
income growth_{ijd}	(1)	(2)	(3)	(4)
$\log(\text{income}_{ijdt})$	-0.347***	-0.325***	-0.324***	-0.323***
	(0.011)	(0.011)	(0.011)	(0.011)
age_{ijdt}	-0.007***	-0.007***	-0.007***	-0.007***
	(0.001)	(0.001)	(0.001)	(0.001)
$\text{V}D_{ijdt}$	-0.283***	-0.262***	-0.275***	-0.270***
	(0.044)	(0.041)	(0.040)	(0.040)
$\log(\text{popdes}_{ct})$	-0.132***	0.027	0.077*	0.082*
	(0.037)	(0.038)	(0.039)	(0.040)
Small size	Ref.	Ref.	Ref.	Ref.
Medium small size	-0.156***	-0.195***	-0.194***	-0.193***
	(0.019)	(0.019)	(0.019)	(0.018)
Medium large size	-0.243***	-0.294***	-0.288***	-0.288***
	(0.022)	(0.022)	(0.022)	(0.021)
Large size	-0.237***	-0.302***	-0.294***	-0.294***
	(0.027)	(0.026)	(0.026)	(0.025)
Geographic dummy	Yes	Yes	Yes	Yes
Industrial dummy	Yes	Yes	Yes	Yes
$\text{R}V_{dt}$		-0.394***		-0.002
		(0.038)		(0.056)
$\text{U}V_{dt}$			-0.026***	-0.025***
			(0.002)	(0.004)
lambda	-7.402***	-10.417***	-10.438***	-10.438***
	(0.713)	(0.729)	(0.739)	(0.720)
No. of Obs.	83,067	83,067	83,067	83,067
R-Squared	0.498	0.503	0.507	0.507
Log Likelihood	-119115	-118684	-118369	-118334
Selection Equation				
No. of Obs.	84,868	84,868	84,868	84,868
Pseudo-R-Squared	0.007	0.015	0.015	0.015
Log Pseudo-Likelihood	-8594.431	-8529.757	-8526.469	-8523.460
Wald Chi Square[p-value]	128.94[0.000]	195.29[0.000]	192.40[0.000]	192.36[0.000]

Note: * p<0.1, ** p<0.05, *** p<0.01. Bootstrapped Standard errors clustered at district level are shown in parentheses. All specifications include provincial geographic dummies and 2-digit industrial dummies and a constant term. lambda denotes the Inverse Mills Ratio from the selection equations.

Table 2 Sectoral Variety and Firm Economic Growth---IV Method

Estimation Method	IV-TSLS	IV-GMM
income growth$_{ijd}$	(1)	(2)
$\log(\text{income}_{\text{ijd}t})$	-0.346***	-0.346***
	(0.104)	(0.104)
$age_{\text{ijd}t}$	-0.007***	-0.007***
	(0.012)	(0.012)
$\text{V}D_{ijdt}$	-0.307	-0.307
	(1.395)	(1.395)
$\log(\text{popdes}_{ct})$	-0.045	-0.045
	(0.754)	(0.754)
Small size	Ref.	Ref.
Medium small size	-0.148***	-0.148***
	(0.327)	(0.327)
Medium large size	-0.226***	-0.226***
	(0.555)	(0.555)
Large size	-0.216***	-0.216***
	(0.619)	(0.619)
Geographic dummy	Yes	Yes
Industrial dummy	Yes	Yes
$\text{R}V_{dt}$	0.894	0.894
	(6.210)	(6.210)
$\text{U}V_{dt}$	-0.052	-0.052
	(1.154)	(1.154)
lambda	-7.582***	-7.582***
	(2.029)	(2.029)
Endogeneity Test(chi-square[p-value])	7.129[0.028]	7.129[0.028]
Cragg-Donald Wald F statistics	894.789	894.789
No. of Obs.	83,067	83,067
R-Squared	0.448	0.448
Log Likelihood	-120235	-120235
Chi Square	117954.10	117954.10
Selection Equation		
No. of Obs.	84,868	84,868
Pseudo-R-Squared	0.015	0.007
Log Pseudo-Likelihood	-8523.46	-8593.29
Wald Chi Square[p-value]	192.36[0.000]	131.40[0.000]

Note: * p<0.1, ** p<0.05, *** p<0.01. Bootstrapped standard errors clustered at district level show are in parentheses. All specifications include provincial geographic dummies and 2-digit industrial dummies and a constant term. Lambda denotes the Inverse Mills Ratio from the selection equations.

Table 3 Sub-sample Estimation for Firm Size Heterogeneity (IV-TSLS)

Estimation Method	TSLS	TSLS	TSLS	TSLS
income growth$_{ijd}$	Small Size (1)	Medium Small (2)	Medium Large (3)	Large Size (4)
$\log(\text{income}_{ijdt})$	-0.723***	-0.588***	-0.431***	-0.241***
	(0.050)	(0.020)	(0.020)	(0.010)
age_{ijdt}	-0.013***	-0.013***	-0.016***	0.004***
	(0.000)	(0.000)	(0.000)	(0.000)
VD_{ijdt}	-0.129	-0.081	-0.306***	-0.197**
	(0.100)	(0.060)	(0.090)	(0.080)
$\log(\text{popdes}_{ct})$	-0.162	-0.075	0.000	-0.252***
	(0.260)	(0.110)	(0.110)	(0.070)
Geographic dummy	Yes	Yes	Yes	Yes
Industrial dummy	Yes	Yes	Yes	Yes
RV_{dt}	0.806	0.9	0.791	0.619*
	(1.390)	(0.620)	(0.790)	(0.360)
UV_{dt}	-0.025	-0.04	-0.05	-0.024
	(0.100)	(0.040)	(0.050)	(0.020)
Lambda	4.631***	4.999***	-3.955***	-17.696***
	(1.560)	(1.480)	(1.340)	(2.320)
Endogeneity Test (chi-square [p-value])	7.393[0.025]	9.362[0.009]	3.652[0.161]	18.283[0.000]
Cragg-Donald Wald F statistics	27.033	180.622	158.923	453.833
No. of Obs.	14,376	19,765	21,665	27,261
Model F Statistic [p-value]	208.57[0.000]	274.61[0.000]	290.66[0.000]	423.25[0.000]
R-Squared	0.447	0.44	0.441	0.437
Log Likelihood	-20181.1	-28007.3	-30679.9	-39552.3

Note: * p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at district level are shown in parentheses. Industry dummy are constructed at 2-digit industrial level, Geographic dummy are constructed at province level.

Table 4 Sub-sample Estimation for Top 3 Developed Regions (IV-TSLS)

Estimation Method	IV-TSLS
income growth$_{ijd}$	(1)
$\log(\text{income}_{ijdt})$	-0.344***
	(0.030)
age_{ijdt}	-0.009***
	(0.000)
VD_{ijdt}	-0.182*
	(0.100)
$\log(\text{popdes}_{ct})$	-0.068
	(0.150)
Small size	Ref.
Medium small size	-0.126***
	(0.040)
Medium large size	-0.203***
	(0.040)
Large size	-0.238***
	(0.060)
Geographic dummy	Yes
Industrial dummy	Yes
RV_{dt}	1.241*
	(0.710)
UV_{dt}	-0.060*
	(0.030)
lambda	-5.775***
	(1.580)
Endogeneity Test(chi-square[p-value])	8.424[0.015]
Cragg-Donald Wald F statistics	769.899
No. of Obs.	33,669
R-Squared	0.427
F statistics[p-value]	323.848[0.000]
Log Likelihood	-50320.77

Note: * p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at district level are shown in parentheses. Industry dummy are constructed at 2-digit industrial level, Geographic dummy are constructed at province level.

APPENDIX

Table A1 National industry classification and code (GB/T 4754—2002)

13	Agricultural and sideline food processing industry	28	Chemical fiber manufacturing industry
14	Food manufacturing industry	29	Rubber products industry
15	Wine, beverages and refined tea manufacturing	30	Plastic products industry
16	Tobacco Manufacturing industry	31	Non-metallic mineral products industry
17	Textile industry	32	Ferrous metal smelting and rolling processing industry
18	Textile and apparel industry	33	Non-ferrous metal smelting and rolling processing industry
19	Leather, fur, feathers and their products and footwear industry	34	Metal products industry
20	Wood processing and wood, bamboo, rattan, brown, grass product	35	General equipment manufacturing industry
21	Furniture manufacturing industry	36	Special equipment manufacturing industry
22	Paper and paper products	37	Automobile Manufacturing
23	Printing and recording media reproduction industry	39	Electrical machinery and equipment manufacturing
24	Culture, education, industry and art, sports and entertainment	40	Computer, communications and other electronic equipment manufa
25	Oil processing, coking and nuclear fuel processing	41	Instrumentation manufacturing industry
26	Chemical raw materials and chemical products manufacturing	42	Other manufacturing
27	Pharmaceutical manufacturing industry	43	Comprehensive utilization of waste resources

Table A2 Firms and employees distribution at provencial level

provcd	Num.Firms	Percent (%)	Cum. (%)	provcd	Num. Employees	Percent (%)	Cum. (%)
130000	3,882	4.01	4.01	130000	1,316,570	4.46	4.46
140000	777	0.8	4.82	140000	545,453	1.85	6.30
210000	6,449	6.67	11.49	210000	1,600,686	5.42	11.72
230000	999	1.03	12.52	230000	388,082	1.31	13.04
320000	14,693	15.19	27.72	320000	4,238,566	14.35	27.39
330000	17,049	17.63	45.35	330000	3,957,860	13.40	40.79
340000	2,517	2.6	47.95	340000	640,113	2.17	42.95
350000	6,709	6.94	54.89	350000	2,128,293	7.21	50.16
360000	2,181	2.26	57.14	360000	612,010	2.07	52.23
370000	14,625	15.12	72.27	370000	4,102,365	13.89	66.12
410000	4,105	4.25	76.51	410000	1,319,083	4.47	70.59
420000	2,619	2.71	79.22	420000	922,493	3.12	73.71
430000	4,142	4.28	83.51	430000	875,858	2.97	76.68
440000	9,071	9.38	92.89	440000	4,257,888	14.42	91.09
450000	1,090	1.13	94.01	450000	319,555	1.08	92.17
510000	3,020	3.12	97.14	510000	967,435	3.28	95.45
520000	406	0.42	97.56	520000	224,270	0.76	96.21
530000	610	0.63	98.19	530000	202,947	0.69	96.90
610000	1,077	1.11	99.3	610000	544,508	1.84	98.74
620000	279	0.29	99.59	620000	216,444	0.73	99.47
630000	72	0.07	99.66	630000	35,317	0.12	99.59
640000	169	0.17	99.84	640000	49,158	0.17	99.76
650000	156	0.16	100	650000	71,604	0.24	100.00
Total	96,697	100		Total	29,536,558	100.00	

Note: This is firm and employee distribution for original dataset.

Table A3 Firms and employees distribution at 2-digit level

indcd2_2003	Num.Firms	Percent(%)	Cum.(%)	indcd2_2003	Num. Employees	Percent(%)	Cum.(%)
13	6,397	6.62	6.62	13	1,167,541	3.95	3.95
14	2,254	2.33	8.95	14	606,306	2.05	6.01
15	1,482	1.53	10.48	15	469,599	1.59	7.60
16	69	0.07	10.55	16	67,821	0.23	7.83
17	8,985	9.29	19.84	17	2,823,227	9.56	17.38
18	3,644	3.77	23.61	18	1,432,771	4.85	22.23
19	2,288	2.37	25.98	19	1,155,023	3.91	26.14
20	2,058	2.13	28.11	20	350,344	1.19	27.33
21	1,041	1.08	29.18	21	314,869	1.07	28.40
22	2,519	2.61	31.79	22	583,204	1.97	30.37
23	1,305	1.35	33.14	23	270,618	0.92	31.29
24	1,036	1.07	34.21	24	407,719	1.38	32.67
25	695	0.72	34.93	25	369,000	1.25	33.92
26	7,574	7.83	42.76	26	1,630,965	5.52	39.44
27	2,309	2.39	45.15	27	705,169	2.39	41.83
28	605	0.63	45.77	28	207,121	0.70	42.53
29	1,311	1.36	47.13	29	411,227	1.39	43.92
30	4,215	4.36	51.49	30	844,833	2.86	46.78
31	7,884	8.15	59.64	31	1,818,766	6.16	52.94
32	2,266	2.34	61.98	32	1,631,851	5.52	58.46
33	2,021	2.09	64.07	33	686,406	2.32	60.79
34	4,711	4.87	68.95	34	1,044,448	3.54	64.32
35	8,780	9.08	78.03	35	1,903,840	6.45	70.77
36	4,205	4.35	82.37	36	1,149,000	3.89	74.66
37	4,459	4.61	86.99	37	1,846,327	6.25	80.91
39	5,967	6.17	93.16	39	1,854,056	6.28	87.19
40	3,326	3.44	96.6	40	2,674,834	9.06	96.24
41	1,356	1.4	98	41	478,999	1.62	97.86
42	1,778	1.84	99.84	42	610,677	2.07	99.93
43	157	0.16	100	43	19,997	0.07	100.00
Total	96,697	100		Total	29,536,558	100	

Table A4 Definition for main variables

Main varieble	Definition
$\log(\text{income}_{ijdt})$	$\text{Income Growth}_{ijd} = \ln(\text{income}_{ijd,T}) - \ln(\text{income}_{ijd,t})$
age_{ijdt}	2006 minus the firm's start operation year
VD_{ijdt}	$\text{Vertical Disintegration}_{ijdt} = \frac{\text{purchased intermediate input}_{ijdt}}{\text{gross output value}_{ijdt}}$
$\log(\text{popdes}_{ct})$	the 2006 population in city c per square-kilometer
RV_{dt}	$RV_{dt} = \sum_{j=1}^{J} P_{jdt} \times [\sum_{g\in j} \frac{p_{gdt}}{P_{jdt}} log_2(\frac{1}{p_{gdt}/P_{jdt}})]$
UV_{dt}	$UV_{dt} = \sum_{J=1}^{J} P_{jdt} \log_2\left(\frac{1}{P_{jdt}}\right)$

Table A5 Correlation matrix of selected explanatory variables

		[1]	[2]	[3]	[4]	[5]	[6]
$\log(\text{income}_{ijdt})$	[1]	1					
age_{ijdt}	[2]	0.2251	1				
$\text{V}D_{ijdt}$	[3]	0.0427	-0.0024	1			
$\log(\text{popdes}_{ct})$	[4]	0.0953	0.0264	0.0492	1		
$\text{R}V_{dt}$	[5]	0.1133	0.016	0.1042	0.4238	1	
$\text{U}V_{dt}$	[6]	0.1124	0.0163	0.0961	0.437	0.8966	1

Table A6 Descriptive statistics of dependent and continuous explanatory variables

Variable	Mean	Std. Dev.	Min	Max
income $growth_{ijd}$	0.171	1.432	-12.622	8.039
$\log(\text{income}_{\text{ijd}t})$	10.599	1.270	5.328	18.872
age_{ijdt}	7.806	8.880	0.000	406.000
$\text{V}D_{ijdt}$	0.745	0.160	0.000	10.373
$\log(\text{popdes}_{ct})$	6.290	0.506	3.228	7.783
$\text{R}V_{dt}$	1.301	0.628	0.000	2.898
$\text{U}V_{dt}$	19.402	11.784	0.000	51.214
$\log(\text{fixasset}_{\text{ijd}t})$	10.063	1.437	5.257	18.390
$\log\left(\text{profit}_{\text{ijd}t}\right)$	7.113	1.900	0.000	16.217
$\log\left(\text{output}_{\text{ijd}t}\right)$	10.637	1.266	5.328	18.878