IMT School for Advanced Studies, Lucca

Lucca, Italy

Production Networks, Firm Productivity and Geography

PhD Program in Economics

XXX Cycle

By

Loredana Fattorini

2018

Table of Contents

List of Figures

List of Tables

Acknowledgements

Chapter 2 is the reproduction of the paper "Measuring the Input Rank in Global Supply Networks", co-authored with Armando Rungi (*IMT Lucca*), currently submitted to *Journal of International Economics* (under review).

Chapter 3 is the reproduction of the paper "Heterogeneous Firms Meet EU Cohesion Policy", co-authored with Mahdi Ghodsi (*wiiw Vienna*) and Armando Rungi (*IMT Lucca*), currently submitted to *Journal of Common Market Studies* (under review).

Chapter 4 is based on a work in progress titled "Spatial Clustering and Survival: Evidence from Firms in Umbria and Puglia".

Vita

April 10, 1990 Born, Barga (Lucca), Italy.

Education

2014 - Ph.D. Candidate in Economics XXX cycle, *IMT School of Advanced Studies*, Lucca, Italy. Supervisor: Prof. Armando Rungi.

2012 - 2014 Master's degree in Economics, *Sant'Anna School of Advanced Studies and University of Pisa*, Pisa, Italy. Final Mark: 110/110 cum laude.

Thesis title: *"Network Flow Techniques: An Application to Integrated Energy Systems"*. Supervisor: Prof. Riccardo Cambini.

2009 - 2012 Bachelor's degree in Economics and Business, *University of Pisa*, Pisa, Italy. Final Mark: 110/110 cum laude.

Thesis title: *"Un'analisi della qualità della vita nelle città attraverso la Data Envelopment Analysis"*. Supervisor: Prof. Laura Carosi.

Additional Education

2016 Barcelona CREI Macroeconomics Summer School, *Barcelona Graduate School of Economics*, Barcelona, Spain.

Course attended: *"Firms, Networks, and Macroeconomic Fluctuations"*. Prof. Julian Di Giovanni.

Experience

2017 Visiting Ph.D Student, *wiiw - The Vienna Institute of International Economic Studies*, Vienna, Austria.

Erasmus+ Mobility Consortium for traineeships "Talent at work", 6-month grant 2016/2017.

Conferences and Workshops

Talks

1. Fattorini, L., Ghodsi, M. and Rungi, A. "Cohesion Policy Meets Heterogeneous Firms", at *European Trade Study Group (ETSG)*, Florence, Italy, 2017.
2. Fattorini, L., Rungi, A. and Zhu, Z. "The Organization of Global Supply Networks", at *European Trade Study Group (ETSG)*, Helsinki, Finland, 2016.

Posters

1. Fattorini, L., Rungi, A. and Zhu, Z. "The Organization of Global Supply Networks", at *American Economic Association (ASSA) Annual Meeting 2018*, Philadelphia (PA), U.S., 2018.
2. Fattorini, L., Ghodsi, M. and Rungi, A. "Heterogeneous Firms Meet EU Cohesion Policy", at *IMT Research Symposium 2018*, Lucca, Italy, 2018.

Participation

1. wiiw Spring Seminar "Europe at a Crossroads", at *Oesterreichische Nationalbank*, Vienna, Austria, 2017.
2. "International Trade, Production Networks, and Microfoundations of Aggregate Fluctuations", at *Italian Trade Study Group (ITSG)*, Lucca, Italy, 2016.
3. IX Grass "Gruppo di Ricerca per l'Analisi delle Scelte Sociali", Lucca, Italy, 2015.

Abstract

Modern economies are organised as webs of interconnected agents. Among others, companies represent the principal actors.

In the recent decades, the emergence of supply networks, that is the organisation of technical processes in production stages involving specialised suppliers located in different countries, brings about an increasing complexity in the worldwide economic system.

Moreover, empirical studies from firm-level data provide evidence of heterogeneous distribution in companies performance also within industries, regions and countries. However, when discussing policy-making in Europe this aspect is still neglected by many, and the impact of regional or industrial policies is evaluated at the macro level.

Recently, geo-coded information on where firms run their activities is becoming available to the researchers, offering a good opportunity to improve the empirical design of specific research questions.

In this thesis, we investigate these aspects in detail, showing the utility of adapting and implementing analytical tools from Network Science and Machine Learning along with ad hoc econometric techniques. Moreover, we contribute to the literature on international economics, industrial organisation and economic geography.

With the first work, we propose an empirical tool, the *Input Rank*, adapted from the notorious *PageRank*, which is originally designed for social networks and search engines, and we test its empirical validity for choices of vertical integration.

Then, we ambitiously test the effect of the EU Cohesion Policy on an own-built dataset with firm-level total factor productivities of European manufacturing companies.

Finally, thanks to the implementation of the *DBSCAN*, a challenging density-based spatial clustering algorithm, we exploit firms' location information to identify industrial clusters and examine the likelihood of firms' survival according to their location in industrial clusters or more isolated areas.

Chapter 1

Introduction

Over the past years, research in international trade, both theoretical and empirical, has shifted its focus from industries and countries to firms and products. Firms through their daily activity of investing, producing, selling, and exporting, are the main drivers of the aggregate dynamics in trade, innovation and competitiveness. Empirical studies on firm-level data show how no 'average' firm within an industry, a region or a country can represent the aggregates (among others, see for example (Mayer and Ottaviano, 2008)). Therefore, we believe that understanding the firm-level facts is essential for policy making.

In this thesis, we explore some unique features about the organisation of production, the productivity and the localisation of firms, contributing to the literature on international trade, industrial organisation and economic geography. First, in the work presented in Chapter 2, we look at the organisation of production processes on a multi-dimensional scale. Adopting a network perspective, we propose an empirical tool, the *Input Rank*, and we test it on vertical integration choices. Second, in the work illustrated in Chapter 3, we carefully investigate the heterogeneity of firms in the

European Union by estimating their total factor productivity. Taking into account the high dispersion, we assess the impact of some EU Cohesion Policy tools on firms' inefficiencies. Third, in the work in Chapter 4, we explore the tendency of firms in some industries to cluster in space, even though more often they run their activity in a global context. Hence, we test the likelihood of firms' survival according to their location in industrial clusters or more isolated areas.

Overall, in the works collected in this thesis, we exploit in different ways the potential of large firm-level datasets, especially when combined with regional and industrial data. Besides, we show how adapting and implementing tools from Network Science and Machine Learning is promising in economics studies.

1.1 Production Networks

> *"The goods we buy are the end result of an elaborately choreographed transnational odyssey. These objects are part of an economy whose tendrils reach over further outward, linking, integrating, and transforming both far-flung and nearby places."*
>
> Kennedy and Florida (2004)

Production processes are more than ever spatially fragmented since a process of unbundling has occurred globally, separating the production of intermediates and final goods across national borders (Hummels et al., 2001; Baldwin, 2006; Baldwin and Lopez-Gonzalez, 2015). In fact, from a trade perspective, it is remarkable that the export of intermediate goods and services has increased sharply relative to the export of final and capital goods and services.[1] Lower trade barriers, lower transport costs, a better organisation in logistic, progress in information and communication

[1] According to OECD (2013) more than half of the world's manufacturing imports are intermediate goods (primary goods, parts and components, and semi-finished products), and more than 70% of the world's services imports are intermediate services, such as business services.

technologies have made slicing up the production process a competitive alternative to the traditional mode of production. The typical 'Made in' labels of manufacturing have become a symbol of an old era, and nowadays it is more appropriate to label products as 'Made in the World' (Antràs, 2015). In this new scenario where firms distribute their production stages around the globe, managers in charge have to face complex decisions. Among other aspects, they have to take into account labour costs, transportation costs, distances to potential markets, coordination costs, productivity and all the external risks possibly causing disruption and losses along the production process, like the recent past shows.

This new economic system involving firms' interconnections through buyer-seller relationships has been conceptualised in the early 2000s as production networks (Dicken et al., 2001; Coe et al., 2008). The latter notion captures the simple idea of firms or industries (nodes) trading tasks and outsourcing stages of production (directed edges). However, the international orientation and integration of firms lead to the emergence of sophisticated global production networks, often driven by world-leading companies (Coe et al., 2008) and also involving 'non-firm' actors.

In the next Section, we show how the network approach seeks to move beyond the analytical limitations of the traditional supply chain notion.

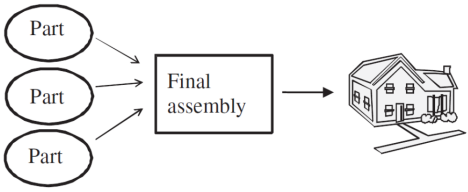
1.1.1 From Supply Chains to Networks

> *"[...] economic processes must be conceptualized in terms of a complex circuitry with a multiplicity of linkages and feedback loops rather than just 'simple' circuits or, even worse, linear flows"*
>
> Hudson (2004)

The literature on the determinants and consequences of the global supply chains has studied the phenomenon of unbundling of production assuming that an ordered and linear sequence exists, from the conception of a product

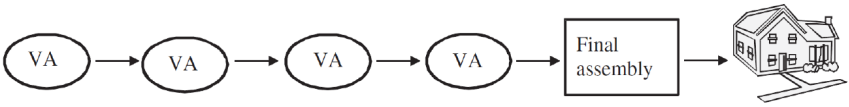
to its distribution and final use.[2] The engineering of the manufacturing process, typical for each industry, dictates the way in which different stages of production links (Baldwin and Venables, 2013). There are two extreme simplified configurations commonly used to describe the supply chain: a spider and a snake. A spider is a production process in which different components come together to be assembled in the final output (Figure 1.1). The snake structure requires the process to be a sequence of stages until the final assembly (Figure 1.2). The latter relies on the predetermined order in which operations are performed.

Figure 1.1: A spider global value chain



Source: Baldwin and Venables (2013).
Note: Each cell is a part, component or final product itself. Each arrow is the physical movement of parts, components or the good itself. Movements can be within a plant in a country, or between plants located in different countries.

Figure 1.2: A snake global value chain



Source: Baldwin and Venables (2013).
Note: Each cell is a production stage at which value is added to a product for final consumption. Each arrow is the physical movement of parts, components or the good itself. Movements can be within a plant in a country, or between plants located in different countries.

[2]For example Antràs and Chor (2013) in their studies on the organization of global value chains, collapse to the [0,1] range the world technological process of production, where 0 indicates the beginning of the production line and 1 the proximity to the final demand.

It is understandable that the chain structure captures the sequential transformation of inputs into outputs, emphasising the value added at each stage of the production process of goods and services. However, all production processes are better thought of as a combination of the snake and spider structure.[3]

In fact, each stage of the production chain may be embedded in a much wider set of non-linear/horizontal relationships (Coe et al., 2008). Besides, the same input can be used at several stages of production, not in linear progression, before reaching the final demand, and the sourcing of an industry can assume different topological configurations. Therefore, albeit an advancement with respect to the previous literature, where each stage of production was considered separately, a linear organisation of production is not realistic, and the network approach allows catching the complexity of actual sourcing strategies.

More in general, production networks highlight the real-world connectivity of modern economies among different firms, industries, and countries. An important aspect that can be captured from this structure is the differential power of relations and actors. In other words, one can investigate who controls key resources (e.g., inputs of production) and possibly disrupts the production network.[4] Indeed, each producer plugged into a production network must decide whether to *make or buy* each input of production given the limited information on indirect transactions and scarce time to outreach the intricate web of direct and indirect suppliers.

1.1.2 Firm Boundaries: *Make or Buy?*

Since the 1980s, many studies have attempted to model the choice of *make or buy* an input of production (vertical integration) that leads to the formation of firm boundaries, based on the degree of contractibility

[3]In Figure 2.1 of Chapter 2, we plot the production network for the U.S. economy using data from U.S. BEA 2002 input-output tables, where each node corresponds an industry and links represent industry-pair transactions.

[4]In the network science, most commonly, it is suggested that power in a network is a function of *positionality* within the network (for example centrality). Several measures of centrality, catching the absolute position of a node in the network, are well-defined and implemented. See Katz (1953), Freeman et al. (1979), Brin and Page (1998) and Kleinberg (1999) for the most used.

of that input and on the institutional environment of the market where companies operate.

The pioneering work of Coase (1937) helped to build a theory of the firm. In fact, new approaches to understanding what determines a firm's boundaries borrow from the theoretical literature on incomplete contracts (Williamson, 1971, 1975, 1979; Grossman and Hart, 1986) and incorporate these frameworks in general equilibrium models. Acemoglu et al. (2007) for the first time investigated how the degree of contractual incompleteness and the extent of technological complementarities between intermediate inputs affect the choice of technology by headquarters.[5] More recently, Harms et al. (2012) analysed the offshoring decision of firms whose production process is characterized by a particular sequence of steps and a non-monotonic variation of transportation costs. Costinot et al. (2012) built a theory of global supply chains, in which the key feature is that production is sequential and standardized in structure, and has offered a first look at how vertical specialization shapes the interdependence of countries. Then, Antràs and Chor (2013) developed a property-rights model of firms' boundaries choice along the value chain, and introduced two measures of industry's production line position, named *DUse_TUse* and *DownMeasure*.[6] The main prediction that emerged from the model is that the position of an input in the value chain is an important determinant of the ownership structure decisions related to that input. Hence, according to the model, final-good producers integrate production stages that are relatively more *downstream* (*upstream*) when final demand is sufficiently elastic (inelastic).

Production processes are sequential in nature, i.e. *downstream* stages cannot commence until *upstream* stages are completed, therefore the organizational decisions along what most assume a production line relate to the contract setting more appropriate to secure each stage input. Usually, firms operate in an environment of incomplete

[5]For a detailed review on trade and firms' organization strategies of multinational enterprises, see Antràs and Yeaple (2014).

[6]Several measures of positioning along supply chains have been proposed, see Fally (2012), Antràs et al. (2012), Miller and Temurshoev (2017),Wang et al. (2017), Antràs and Chor (2018), and for bilateral industry-pairs Alfaro et al. (2017).

contracts and intermediate producers along with the final producer bargain over the surplus associated with a particular stage. Owning a supplier is a source of power for the firm because it enhances its bargaining power through the residual control rights, but at the same time, it reduces the incentive of suppliers to invest in the relationship.

In the framework of production networks and firm boundaries (following Antràs and Chor (2013) theoretical model), in Chapter 2 we address the following research questions:

- How to measure the local technological relevance of any input for the completion of a network-like production process?
- Does the local technological relevance of any input in a network-like production process have a role in the choice of vertical integration of a firm?

1.2 Firm Heterogeneity and EU Regional Policy

Firms embedded in production networks are heterogeneous along many dimensions, such as size, cost structure, profits, sales, productivity, etc.. Dispersion in firm size has been well documented. Empirically, the distribution of firm size is very skewed and has been described by a power law (Gabaix, 2009, 2016; Luttmer, 2010). Also, even within narrowly defined industries, there is evidence of large dispersion in firm outcomes such as revenue, employment, labour productivity and total factor productivity (Syverson, 2011), reflecting misallocation of resources and market distortions. In particular, the productivity[7] distribution is asymmetric with a large density of low-productive firms and few highly productive firms. Although this empirical regularity applies to all countries and industries, the shape of the distribution can differ across countries, reflecting their structural characteristics.

In the recent years, as policy-makers refocus on growth, there has been an increasing interest in assessing the competitiveness at different levels of analysis.[8] Firm-level productivity is an essential indicator of micro competitiveness. Porter (1990) defined micro competitiveness as the ability of a given firm to successfully compete in a particular business environment through innovation, i.e. new technology and the new way of doing things. Therefore, critical is the design of appropriate competitiveness-enhancing policies.

When discussing policy-making in Europe, the tendency is to refer to aggregate country-, region- or industry-level statistics as these are easier

[7] As an indicator of firm's efficiency, productivity can be measured estimating total factor productivity (TFP). In fact,TFP has been defined as the efficiency of a firm to turn inputs of production such as capital, labour, and intermediate inputs into products at the lowest possible cost. In other words, TFP is the portion of output not explained by the amount of inputs used in production (Hulten, 2001; Katayama et al., 2009; Van Beveren, 2012).

[8] In 2012, the President of the European Central Bank (ECB), Mario Draghi stated: "A competitive economy, in essence, is one in which institutional and macroeconomic conditions allow productive firms to thrive. In turn, the development of these firms supports the expansion of employment, investment and trade." (`https://www.ecb.europa.eu/press/key/date/2012/html/sp121130.en.html`)

to calculate, understand, and finally communicate (Altomonte et al., 2012). In general, the impact of regional or industrial policies is evaluated at the macro-level. Actually, there are firms that from the bottom shape the aggregate statistics through their daily activity of investing, producing, selling and exporting. The last decade of empirical studies on firm-level data shows how no 'average' firm *within* an industry, a region or a country can represent the aggregates (among others, see for example Mayer and Ottaviano (2008)). Therefore, thanks to the increasing availability of micro-data at the firm-level, it is now possible to check whether the impact of regional policies mandated at the EU level is heterogeneous *within* country, region and industry, which is often as relevant as differences *across* countries, regions and industries. In conclusion, as emphasised in Altomonte et al. (2016), similar set of policies dictated *ex-ante* may end up producing very different outcomes *ex-post*, because of the heterogeneity of firm's performance which nowadays cannot be disregarded.

Given the facts mentioned above emerging from adopting a firm-level view, in Chapter 3 we ambitiously try to answer the following research question:

- Does the provision of EU funds have an effect on the productivity of firms?

1.3 On the Geography of Firms

'Where' firms' activity takes place is extremely important in understanding their performance, dynamics of entry and exit from the market, and also their involvement in global value chains and production networks.

Nowadays, a lot of firms are globally organised in some corporate functions (e.g., finance, human resources, IT, procurement, legal, facilities management) and production stages, however especially in the initial phase of producing good and/or services their activity is highly localised in space also in proximity to similar businesses. Whereas, final goods and services are more global in their consumption.

Therefore, the idea of global production networks introduced in Section 1.1 should be understood as a spatial ensemble of different local and regionally centred economic activities (Coe et al., 2008).

Recently, a larger amount of micro-geographic data is becoming available and usable, calling for a more accurate exploration of non-random spatial patterns of firms.

1.3.1 The Use of Micro-Geographic Data

In the recent years, for researchers, it is becoming more common to get access to data sets collecting spatial information of firms. Although, in most official databases, spatial information is still only available in the form of administrative divisions, either because only a specific territorial level is recorded or because confidentiality rules forbid the dissemination of precise location information, in some, mostly commercial databases it is possible to retrieve exact localisation of firms from postal addresses. Then, these postal addresses have to be translated to geographic coordinates, i.e. latitude and longitude, through geocoding procedure which nowadays allow the treatment of large datasets.[9]

Geo-coded data offers new room for empirical investigations: it is possible to get new insights invisible with aggregated data, and therefore answer 'old' and new research questions more accurately; the higher geographical

[9]For a review on the availability and usability of large datasets containing exact firm's location, see Lennert (2015).

resolution can lead to econometric or other methodological improvements; it is possible to overcome the scaling and regional's boundary problem, known as 'Modifiable Areal Unit Problem' as first described in Openshaw (1984).

The microeconomic approach to spatial analysis is gaining importance over the traditional approach based on country or regional aggregates. Many research questions have been already addressed using micro-geographic data. Among others, Arbia (2001) proposed a prototype spatial micro-economic model describing the birth, survival and growth of firms on a continuous surface. Boix et al. (2015) identified and provided stylized facts about clusters of creative industries (e.g., Printing, Publishing, Design, etc.) in sixteen European countries. Ruffner and Spescha (2018) investigated the impact of spatial clustering on firm innovation for a sample of Swiss firms, combining information about the innovation activities with the exact location of firms. This allows testing the effect of agglomeration externalities at a much disaggregated geographic level.[10]

1.3.2 Agglomeration and Firm Survival

In the new global scenario, local proximity of firms producing similar, competing and/or related products together with supporting institutions still matter. Externalities that spatial agglomeration creates in the short and in the long-run have an impact on firm's productivity, innovation, employment and location choice already investigated in many studies (Duranton and Puga, 2004; Rosenthal and Strange, 2004; Beaudry and Schiffauerova, 2009; Combes et al., 2012).

Only recently, the literature on firm survival has started to take into account the geography of firms by investigating the potential costs and benefits of being located in proximity. For instance, firms producing similar products within the localised cluster may face higher competition, and thus only the most productive firms are more likely to survive; or firms located in highly specialised areas, such as industrial districts, may be more vulnerable to idiosyncratic shocks (Basile et al., 2017). Also,

[10] See Blind et al. (2018) for a recent collection of papers on the use of micro-geographic data in economic research.

the related local variety (Jacobs, 1969), arising from the co-existence of diverse industries sharing common or complementary knowledge, may protect firms from idiosyncratic demand shocks and therefore increase the survival probability. Finally, urban agglomeration captured by the overall size of the economy may increase the likelihood of survival when firms face higher local demand and supply of public services, and congestion costs (e.g., land prices, crime rates) are negligible (Ciccone and Hall, 1996).

Exploiting micro-geographic data for a sample of Italian firms, in Chapter 4 we empirically answer the following research questions:

- How to identify industrial clusters on continuous space from large micro-geographic datasets?
- Where in space are industrial clusters located?
- Are firms located in industrial clusters more likely to survive than firms operating in more isolated areas?

1.4 Contributions

Here follows a detailed description of the contributions given by each chapter of the thesis.

Chapter 2: Measuring the Input Rank in Global Supply Networks

In Chapter 2, we introduce a network dimension in the study of Global Values Chains, proposing an empirical tool, the *Input Rank* adapted from the notorious *PageRank* which has been originally designed for social networks and search engines (Brin and Page, 1998) but it has rapidly spread as a network tool in so many different domains, from genetics to the engineering of road networks (see Gleich (2015)). We believe that the intuitive Markov stochastic process that we imagined behind the *Input Rank* is promising for other economic applications willing to study the technological relevance of firms or sectors in a network-like economy. Hence, in this work we contribute to the empirical literature on quantifying the position of production stages in technological networks beyond the linear technological sequence, i.e. the supply chain. Further, we also contribute to the literature of vertical integration choices by positively testing that the *make or buy* decision of a producer can be driven by the technological centrality of an input in the specific production process.

Chapter 3: Heterogeneous Firms Meet EU Cohesion Policy

The contribution of the work presented in Chapter 3 is twofold. First, we contribute to the existing literature by estimating the firm-level TFP for EU manufacturing firms within each 2-digit of the NACE classification for the period 2007-2015. In the end, we constructed a unique dataset of 542,876 firms which is already used in other papers in our research pipeline and can be exploited for future works. Second, given our findings, we believe that we may contribute to the debate of the next months, as the negotiation for the next budget of the EU Cohesion Policy has just started, where a revision of ERDF governance and its administrative procedures is envisioned (EC, 2009). We support the idea that a better targeted policy

design as in the case of research, technology, and development is preferable, because it unambiguously promotes investments in innovation in products and production processes, with an impact in the short run.

Chapter 4: Spatial Clustering and Survival: Evidence from Firms in Umbria and Puglia

In Chapter 4, we show the utility of implementing a density-based spatial clustering algorithm, the *DBSCAN*, from Machine Learning to identify industrial clusters beyond the administrative boundaries. To date, few works have considered the richness of geo-coded data for firms. Hence, we make an effort to cover this gap by showing an application on a sample of Italian firms, but which could be usefully extended to whatever country or region. Moreover, we contribute to the literature on firm survival by investigating the impact of spatial clustering on the likelihood of survival and testing the effect of agglomeration variables along with other firm and industry characteristics.

CHAPTER 2

Measuring the Input Rank in Global Supply Networks

2.1 Introduction

Modern economies are organised as webs of specialised producers. Each company can be plunged into a supply network that starts with the idea of a product in engineering, design or research labs. After that, parts and components are manufactured and assembled, then they reach producers of final products who require the services of marketing, advertising and distribution companies to get to the market.

In fact, the configuration of production processes can be much complex and recursive in nature, when the same intermediate goods and services are repeatedly needed along the supply network, at different stages of the production process. Take logistics and distribution services, which are crucial in the delivery of parts and components to other companies, as well as in the case of final goods destined to consumers. Raw materials are the basis of so many manufactured inputs and outputs. Innovation

can require the services of R&D labs, engineering and design at various stages during the production process.

Against this background, in recent decades, supply networks have been increasingly fragmenting on a global scale since a process of unbundling started due to the dramatic advances in transportation and communication technologies that scattered production stages across different countries (Baldwin, 2006).

Although fragmentation can originate either *spider-like* or *snake-like* configurations, depending on the technological peculiarities of the production processes (Baldwin and Venables, 2013), the international organisation of production has been mainly studied after assuming a separation of production stages along ordered and linear sequences. That is, complex production networks have been proxied as *snake-like* processes running from the conception of the product to its delivery for final usage (Costinot et al., 2012; Fally, 2012; Antràs and Chor, 2013; Alfaro et al., 2017; Antràs and de Gortari, 2017), therefore neglecting the *spider-like* nature of the organization of production for sake of simpler assumptions on theory and empirics. Specialization patterns by country along global value chains (GVC) and firm-level choices of vertical integration have been both envisaged as partitions of ideally linear segments oriented on *upstream-downstream* directions.

So far, empirical efforts have followed theory when proposing positioning metrics, e.g., the *upstreamness* or *downstreamness* of a production stage, which simulate a technological sequence constructed on input-output tables (Fally, 2012; Antràs et al., 2012; Antràs and Chor, 2013; Alfaro et al., 2017; Miller and Temurshoev, 2017; Wang et al., 2017; Antràs and Chor, 2018). Albeit an advancement for understanding the interdependence among buyers and suppliers, linear circuits certainly underestimate the relative central importance of some inputs, which can magnify or dampen a shock in the presence of technological loops, kinks and corners.

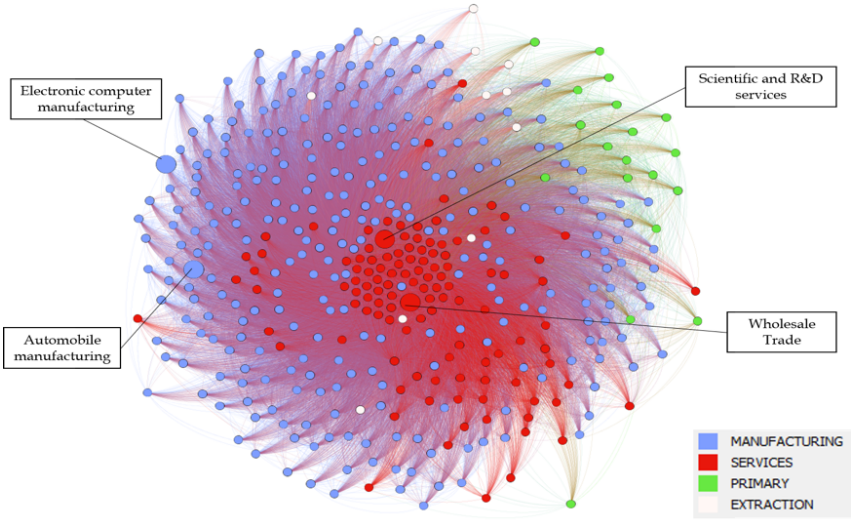
Take the case of the U.S. economy, which we plot as a production network in Figure 2.1. According to the U.S. BEA 2002 input-output tables, the U.S. economy can be represented as a collection of 425 industries (i.e., nodes) linked by 51,768 transactions (i.e., edges). In Figure 2.1, we organise

U.S. industries on a two-dimension space according to their reciprocal connectivity, following a Fruchterman and Reingold (1991) layout, which in our case posits more requested inputs at the centre stage. Interestingly, services industries make the core of the U.S. production network because they are used as direct inputs in many other manufacturing and services industries. Primary industries, like agriculture and forestry, are rather peripheral and mostly located in the north-west area of the graph. Among services, let us pick the case of R&D (code 541700) and Wholesale Trade (code 541800), which seem to be among the most connected industries. In fact, wholesalers have a prominent role in professionally distributing many intermediate inputs in different moments of the production process, whereas R&D services are pivotal in fostering innovation across most U.S. sectors. Let us consider now the case of two consumer goods industries: Electronic Computer Manufacturing (code 334111) and Automobile Manufacturing (code 336111). Although their products can be used as capital goods in some other sectors, they appear to be at the periphery of the U.S. production network, because their final usage prevails on the intermediate usage.

In fact, at first glance, Figure 2.1 shows a rather compact network with a relatively high density, i.e. the fraction of actual linkages out of all potential linkages is 0.286. The average path length connecting any two industries is just 1.7 links, pointing to a *small-world* nature of the U.S. economy. Briefly, on average, any producer sources inputs from most of the other industries, either directly or indirectly. Indeed, the network of Figure 2.1 is not separable: it is self-contained in a unique connected component, and it is always possible to run seamlessly from one node to another, just following input linkages.

Once we compare the positions of selected industries in the production network with their position on the *downstreamness* segment (Antràs and Chor, 2013), in Figure 2.2, we curiously find that both R&D and Wholesale Trade are in the middle of the ideally linear supply chain. This is in contrast with the stylized chain we may have in mind, where a representative business line would start with R&D services and ends with distributions services. In fact, when we review computation methodologies, we find

Figure 2.1: Input-Output network from U.S. BEA 2002 I-O tables, and selected industries

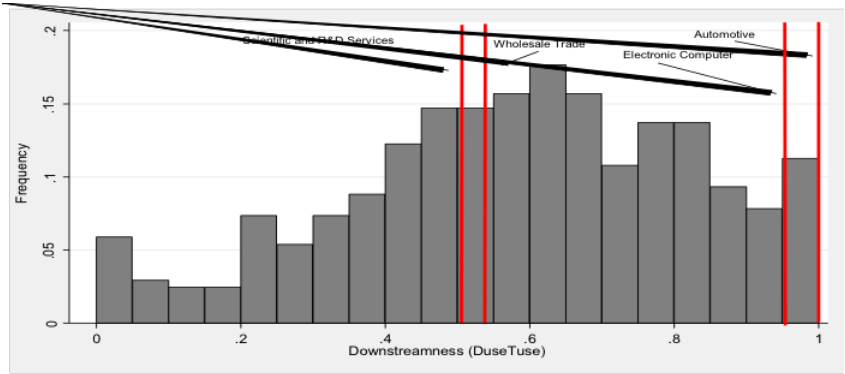


Source: Own elaboration.
Note: Nodes represent 425 6-digit NAICS industries from the U.S. Bureau of Economic Analysis 2002 Input-Output tables. Edges represent 51,768 industry-pair transactions. The graph is visualised using a Fruchterman and Reingold (1991) layout in the GEPHI software. More connected industries (weighted out-degree) at the centre of the graph. Selected industries in evidence.

that *downstreamness* segments are essentially derived from the weighted relative usage of the tasks, intermediate vis à vis final, in one or more industries, therefore confounding the distance from the final demand and the central role they may have across different production processes, when the actual production network is collapsed on a segment.[1]

[1]More recently, Alfaro et al. (2017) compute a *Relative upstreamness* to consider the heterogeneity of input positions oriented towards different outputs. However, also in this case, the position of R&D services is on average located in the middle of the output-specific technological sequences, i.e., the average *upstreamness* value is 3.044 for an indicator that originally ranges approximately from 1 to 8.9.

Figure 2.2: *Downstreamness* of selected industries from U.S. BEA 2002 I-O tables



Source: Own elaboration.
Note: *Downstreamness* (*DUse_Tuse*) sourced from Antràs and Chor (2013). Frequency indicates how many industries out of total 425 from U.S. input-output tables are found in that position. Selected industries: Scientific Research and Development Services (code 541700, value 0.504); Wholesale Trade (code 541800, value 0.666); Electronic Computer Manufacturing (code 334111, value 0.959); Automobile Manufacturing (code 336111, value 0.999).

Instead, we argue that the mutually interactive and recursive nature of production processes is better understood when we consider not only how inputs enter in a different order (*downstream* vs *upstream*), but also how central they are because they are required more than once along the same production processes (as inputs of inputs).

In the end, a bird's eye view of the U.S. production network represented in Figure 2.1 returns an idea of a 'global' centrality for each industry, for example rendering the crucial role of the R&D services for the entire U.S. economy. However, what we are actually interested in is a measure of the 'local' technological relevance of any (direct or indirect) input for the completion of a specific target output.

In this respect, we introduce the *Input Rank* as a solution to a Markov stochastic process that ranks direct and indirect inputs oriented towards a target output, once assuming that producers have limited information

on indirect transactions and scarce time to outreach the intricate web of direct and indirect suppliers. For this, we get our inspiration from the *PageRank* centrality, which is a measure originally used in social networks and search engines to assess the relevant content of information (Brin and Page, 1998). The tool has by now spread to many different domains[2], from biology and genetics to financial debts, bibliometrics and engineering of road networks (Gleich, 2015).

In our perspective, the *Input Rank* can be seen as the result of a Markov stochastic process started by a producer that is embedded in a supply network made of direct and indirect suppliers. We can easily assume that a representative producer does not know the details of indirect transactions, further *upstream*, although they can be much relevant for the completion of her output. Therefore, she starts navigating her web of suppliers (e.g., making phone calls to her direct or indirect suppliers) to collect information on the quantity and quality of these transactions. In her *random walks*, she can encounter some resistance, for example, due to a reduced contractibility of some indirect inputs, which means that they are not quoted at any exchange nor are they referenced priced, possibly tailored for the specific need of their buyers. When resistance is encountered, because information is difficult to collect, then it is more likely that the producer stops her exploration along that path, goes back to headquarters and starts following a different trail. However, at the end of each exploration, she can update her ranking, considering more relevant the inputs that are encountered more often and that are required more. The idea is straightforward:

- A (direct or indirect) input that is also relatively more requested to produce other (direct or indirect) inputs must rank relatively higher;

- A (direct or indirect) input that is relatively more requested to produce other highly-requested (direct or indirect) inputs is relatively

[2]For a previous adaptation of a *PageRank* centrality in the economics domain, see the *DebtRank* by Battiston et al. (2012), where connectivity among financial institutions and debt exposures are considered to determine the systemic importance of a node in a financial network.

more relevant than a (direct or indirect) input that links with less-requested inputs.

In principle, our *Input Rank* can fit the analysis of complex webs of firm-to-firm transactions, as well as the study of more aggregate buyer-supplier linkages recorded in input-output tables. For the sake of comparison with previous studies on the global organisation of production, we compute our *Input Rank* exploiting U.S. input-output tables sourced from the Bureau of Economic Analysis (BEA). After that, we test its empirical validity as a determinant of vertical integration choices in the fashion of Antràs and Chor (2013), Alfaro et al. (2017), and Del Prete and Rungi (2017), on a sample of 20,489 U.S. parent companies controlling 154,836 affiliates worldwide.

We find that a higher *Input Rank* is positively associated to higher odds that a (direct or indirect) input is vertically integrated within a firm boundary, relatively more when the demand faced by the *root* producer is more elastic. With a general reference to the contract theory of the firm, we argue that a choice of vertical integration is a way for the parent company to prevent that a central (direct or indirect) supplier reneges on her commitment, therefore endangering the functioning of the entire supply network. Our findings are robust to different sample compositions, to changes in parameters that measure the ability to outreach by a producer on the complex network structure, and to several empirical strategies. Interestingly, our findings are also robust to the inclusion of *downstreamness*/*upstreamness* metrics, which show some ambivalence in the case of midstream parents, in line with what previously found in Del Prete and Rungi (2017). Therefore, we discuss how the role of the elasticity of substitution is not clear when a producer of intermediate inputs starts vertical integration.

The rest of the chapter is organized as follows. The next section positions our contribution with respect to related literature. Section 2.3 introduces the *Input Rank* and its properties. In Section 2.4, we compute the *Input Rank* on U.S. input-output tables and describe some preliminary evidence. In Section 2.5, we test the role of the *Input Rank* in firm-level

choices of vertical integration. Concluding remarks are offered in Section 2.6.

2.2 Related Literature

A flourishing literature is emerging to study how the network dimensions in the organisation of production can contribute to explain the response of aggregate fluctuations to microeconomic shocks (Acemoglu et al., 2012; Carvalho, 2014; Acemoglu et al., 2016). On the other hand, the shape of a production network is increasingly seen as the result of endogenous collective choices by buyers and suppliers, who establish reciprocal input linkages hence shaping both individual and aggregate productivities (Oberfield, 2018).

Yet, the fragmentation of global value chains (GVC) is still modelled and tested as on a linear sequence made of producers, who decide whether to *make or buy* an input, even though the existence of *spider-like* vs *snake-like* production has been acknowledged as depending on engineering details (Baldwin and Venables, 2013). To date, few works have considered the richness of buyer-supplier networks from an international perspective[3], and there is much to do for understanding the implications of network structures on the global organisation of production. We make an effort to cover this gap starting with the introduction of the *Input Rank* as a measure of the technological relevance of an input, which takes into account the recursive nature of real-world webs of suppliers and buyers.

Realistically, we assume that each producer is plunged into a production network made of both direct and indirect suppliers, where inputs may be recursively used at different production stages. Therefore, we assume that a representative producer may have direct knowledge of the transactions in which she is a contracting party, but she has only limited knowledge of the transactions occurring among suppliers of suppliers. Nonetheless, what happens in the *upstream* technological and contractual space has consequences on her ability to deliver an output. Thus, she starts *random walks*

[3]For a first review of the first attempts made until now, see Bernard and Moxnes (2018).

(say, random phone calls) on her web of suppliers to acquire knowledge about indirect input transactions. We represent such a process as a Markov discrete chain in the spirit of the *PageRank* problem, which originally ranks the consumption of information and regulates the workings of many social networks (Brin and Page, 1998). More specifically, we get inspired by the 'personalised' version of the *PageRank* proposed by Haveliwala (2002) and White and Smyth (2003). In fact, the *PageRank* has become a more general tool of the network theory and it is adapted to many diverse scientific domains (Gleich, 2015), from biology and bibliometrics, to neuroscience and engineering of road networks.

To show a practical empirical usage of the *Input Rank*, we test its role for the firm-level choices of vertical integration. Since the 1980s, several attempts have been made to model the *make or buy* decision based on the relative degree of contractibility between a buyer and a supplier[4]. Acemoglu et al. (2007) are the first to study a theoretical framework where unique headquarters commit to contracts with multiple suppliers. More recently, Harms et al. (2012) analyse the offshoring decision of firms whose production process is characterised by a sequence of steps and a non-monotonic variation of transportation costs. Costinot et al. (2012) derive a sequential multi-country model in which mistakes can occur with a given probability along a sequence. Hence, countries performing more knowledge-intensive tasks are better situated relatively more *upstream*, participating to a larger share in world income distribution. Interestingly, Fally and Hillberry (2015) include Coasian transaction costs to explain the length of a supply chain and the cross-country variation in gross output-to-value added ratios.

In each of the previous works, the notion of a sequence assumes different shades of meaning. More properly, we believe that Antràs and Chor (2013) and Alfaro et al. (2017) model a supply chain as a technological sequence made of production stages, from the start of a business line to the delivery to final demand, where each *downstream* output depends on a set of

[4]For a detailed review since the seminal work by Grossman and Hart (1986), see Aghion and Holden (2011). See also Antràs and Yeaple (2014) for a review on trade and firm-level organization of multinational enterprises.

upstream (direct or indirect) inputs. In this framework, all producers shall rely on a surplus from the sale of the final output, and an economic and contractual dependence is established along the supply chain, for how that surplus is optimally generated by and allocated among producers. Eventually, the main prediction by the authors is that final-good producers integrate production stages that are relatively more *downstream* (*upstream*) when final demand is sufficiently elastic (inelastic).

In this respect, we believe that the latter strand of research has a potential to be extended to more complex production structures, while leaving the technological sequence, i.e. the chain, as a corner solution of real-world supply networks. In fact, more recently Antràs and de Gortari (2017) succeed in extending a supposedly linear technological sequence with the introduction of a notion of geographic centrality. In a multi-country setting, the authors predict an optimal location for a production stage to be dependent also on the geographical proximity to other stages. In that framework, *downstream* stages are preferably located in more *central* destinations, where centrality is to be interpreted in terms of geographic proximity.

When it comes to firm-level empirics, both Del Prete and Rungi (2017) and Alfaro et al. (2017) positively test the predictions by Antràs and Chor (2013), assuming as from theory that integration starts from the end of the supply chain. If final demand is sufficiently elastic (inelastic), producers of final goods integrate production stages that are more proximate to (far from) the consumers. However, in the case of *midstream* parents, when integration starts from the middle of the supply chains, Del Prete and Rungi (2017) find that the same theoretical predictions are no more statistically significant. In either case, integrated production tasks tend to be rather proximate to the parent on the *downstreamness* segment, possibly due to some local unexplored technological complementarities.

In our empirical application of the *Input Rank*, we build on the latter framework and test whether the *make or buy* decision can be driven by the technological centrality of an input in the specific production process. In fact, we find that a higher I*nput Rank* is associated with higher odds that the input is vertically integrated, even more, when the elasticity of

demand faced by the parent is more elastic. In this context, our findings seem to show that the network positioning of an input is at least as important as its distance from final consumption, the latter proxied by the *downstreamness*/*upstreamness* segments.

2.3 The *Input Rank*

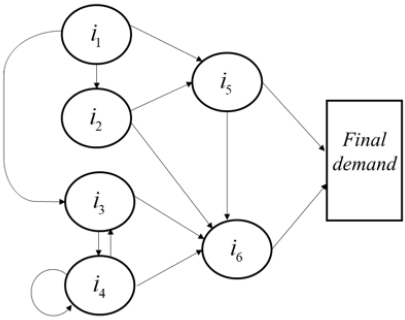
2.3.1 A Producer in a Supply Network

Our aim is to catch the central position of each input, i.e. its technological relevance for the completion of a network-like production process. We start by representing the problem of a producer who plans the delivery of her output based on the requirements of both direct and indirect inputs. In other words, a producer is aware that the completion of her production process requires the contribution of her direct transactions, for which she has immediate knowledge, and of indirect transactions among suppliers of suppliers, whose quantity and quality are not immediately known.

To clarify better the nature of the producer problem, let us consider a stylized production network sketched in Figure 2.3, where we represent an economy made of six producers and a final consumer. A producer i_6 transacts with four suppliers in the set $\{i_2, i_3, i_4, i_5\}$ to buy intermediate goods and services. Besides direct transactions, what happens more *upstream* is not directly disclosed to her. For example, there is a set of suppliers $\{i_2, i_3, i_5\}$ who directly rely on purchases from i_1. On the other hand, suppliers in $\{i_3, i_4\}$ reciprocally exchange part of their output to be used as intermediate inputs in their production process, whereas typically the supplier i_4 employs a share of her output to be reinvested in her production process as an intermediate input. Going further up in the production network, the supplier i_1 is also an indirect supplier of supplier i_4, through the production process made by i_3, who eventually is both a direct and an indirect supplier of our target producer i_6. In a nutshell, the stylized supply network of Figure 2.3 includes the main elements that make the production process recursive in nature.

Against this background, the producer problem reduces to a ranking of

Figure 2.3: A stylized production network made of six producers



Source: Own elaboration.

all direct and indirect suppliers considering their technological centrality in the supply network. What happens if one of them does not deliver? Who increases at most the risk of a disruption on the supply network?

Obviously, beyond the stylized network of Figure 2.3, real-life production processes can see the engagement of an incredibly high number of direct and indirect suppliers, active in many industries and in many countries. For this reason, the representative producer can find it difficult to collect information on *upstream* transactions, especially when global production is more and more fragmented. Realistically, we can imagine that the exploration process started by the producer can be proxied by a Markov discrete chain on paths of suppliers, but with heterogeneous abilities to outreach on the entire web of suppliers.

2.3.2 Random Walks on a Production Network

Let us start by considering an economy in the form of an oriented graph:

$$G(N, E, V, D) \tag{2.1}$$

made of a set of producers,$\forall\ i, j \in N$, connected by a set of direct input-output linkages, $e_{ij} \in E$. Each producer generates an amount of output, $v_i \in V$, that is distributed through linkages whose weights correspond

to (normalized) input requirements, $d_{ij} \in D$, falling in a range $[0, 1]$. Each input requirement, d_{ij}, represents the amount of the i^{th} direct input necessary to produce a unit of the j^{th} output.

From a producer's perspective, her supply network is a sub-graph of the entire economy that can be navigated through *upstream* paths in the form:

$$P_{ir}^{-} = (r, i_l, ..., i_m) \tag{2.2}$$

where the r^{th} node represents headquarters, i_l and i_m are any direct and indirect supplier, respectively. The negative sign on P_{ir}^{-} indicates the orientation of the path[5], from the *root* producer *upstream*, where at each step there is the placement of a demand order for an input. The simpler path is, of course, one that runs from the *root* producer until a direct supplier. More sophisticated paths can run through production cycles, as in the stylized network of Figure 2.3.

In this context, let us define a transitive closure, P_r^{-}, as a collection of all technological paths, such that $P_{ir}^{-} \in P_r^{-}$. In other words, an oriented transitive closure, P_r^{-}, represents a mapping that tells us whether and how any supplier can be reached by a *root* producer, given the entire technological network, G, that defines an economy. Please note how we can find more than one demand path running between a *root* buyer and any supplier[6], possibly given by the recursive nature of the production processes. Usefully, the introduction of a generic *root* buyer allows us changing the reference point in the supply network for how many producers we can find in the entire economy, hence spotting local properties that are

[5]For the sake of generality, we could also define a *downstream* relationship running from suppliers to buyers based on input deliveries, i.e. a supply side on the production network with positively oriented paths, P_{ir}^{+}, which tell us whether a supplier can reach a *root* buyer in a positive transitive closure P_r^{+}. For a reference, see Gilles (2010).

[6]For example, in the simple network described by Figure 2.3, a (positive or negative) transitive closure oriented on i_6 includes a total of 18 paths of variable length: four paths of length one, seven paths of length two, four paths of length 3 and three paths of length 4. Among others, the indirect supplier i_1 is connected to buyer i_6 through seven paths of variable length: $(i_1, i_5, i_6), (i_1, i_2, i_6), (i_1, i_3, i_6), (i_1, i_2, i_5, i_6), (i_1, i_3, i_4, i_6), (i_1, i_3, i_4, i_3, i_6)$, $(i_1, i_3, i_4, i_4, i_6)$. Please note how the presence of reciprocal supplies (i.e., cycles) and in-house production requires the multiple accounting of some suppliers on the same path.

specific for each target producer and her production process.

In essence, any (supply or demand) path that is left outside a (positive or negative) transitive closure oriented to the r^{th} buyer does not participate in the production process of the latter. Fixing an eye on a different *root* buyer would imply the separation of a different sub-network from the entire network economy.

Given the previous framework, we can finally solve the problem of attributing a ranking to direct and indirect suppliers from the perspective of the *root* buyer. The idea is rather simple:

- an input is more technologically relevant to the r^{th} producer if it is also relatively more requested to produce other (direct or indirect) inputs;

- an input is more technologically relevant to the r^{th} producer if it is also relatively more requested to produce other highly-requested (direct or indirect) inputs.

At this point, we can define a measure of the *Input Rank* as a stochastic process[7] started by the *root* producer, who needs travelling *random walks* to collect information on the characteristics of all direct and indirect inputs required to complete her production.

We assume that the *root* producer travels randomly, going from one supplier to another, e.g., calling them by phone and asking on characteristics of deliveries. At any step in the web of transactions, she has a probability α to proceed in the exploration and a probability $(1 - \alpha)$ to fall back to headquarters. The parameter α proxies an information wedge between the buyer and the supplier that prevents full disclosure of the attributes of transactions.

At any time-step t, the *root* producer collects information on the direct requirement, d_{ij}, of each transaction, and she updates her information

[7] To construct the *Input Rank*, we get our inspiration from the *PageRank* centrality introduced by Brin and Page (1998) to organize web pages based on their connectivity with the rest of the web. Nowadays, the *PageRank* centrality is considered a useful tool from network theory and is used across different domains: bibliometrics, biology, physics, etc. (Gleich, 2015). For a previous economic application, see Battiston et al. (2012), who adopt a notion of *Debt Rank* for assessing financial systemic risk.

following a Markov process as follows:

$$\boldsymbol{\pi}_{rt} = \alpha \boldsymbol{D} \boldsymbol{\pi}_{r,t-1} + (1-\alpha)\boldsymbol{h}_r \tag{2.3}$$

where $\boldsymbol{\pi}_{rt}$ and $\boldsymbol{\pi}_{r,t-1}$ are column vectors including rankings of any i^{th} input at time-steps t and $t-1$, respectively. The transition matrix $\boldsymbol{D}$ collects all (column-normalized) direct requirement coefficients, d_{ij}, of transactions in the economy. The vector $\boldsymbol{h}_r$ has all its elements equal to 0 except for the r^{th} element, which is 1 for the selected *root* producer.[8] In our case, the vector $\boldsymbol{h}_r$ uniquely identifies a (negative) transitive closure, P_r^-, hence avoiding that the buyer falls outside her technology when back from her *random walks*. More in general, the vector $\boldsymbol{h}_r$ excludes that the *root* producer lingers in other areas of the production network while assuring her a safe journey back to headquarters.

The probability α falls in a range $(0, 1)$. A value proximate to 0 implies that the producer encounters a higher difficulty to travel *upstream* in the production network. A value proximate to 1 implies that the producer encounters almost no resistance in the exploration of her network. We shall exclude that the parameter α is equal to either 0 or 1 because in either case no exploration is needed or started at all. In the next paragraph, we will discuss an attempt to better qualify this parameter from an economic point of view, and we will discuss how sensitive results are at changing thresholds of this parameter. More in general, a constant parameter α can be seen as an information wedge between a producer and her web of suppliers, such that inputs that are more closely related to the target output are also more easily reached by the *root* producer starting from the headquarters.

What we know from the Perron-Frobenius theorem is that a stationary distribution, $\boldsymbol{\pi}_r^*$, of the Markov process in 2.3 can be found, it is unique, and the sum of its single elements is such that $\sum_i \pi_{ir}^* = 1$, because the transition square matrix $\boldsymbol{D}$ is positive and column-stochastic, with all positive single elements, $d_{ij} \geq 0$. In the stationary status, the single element $\pi_{ir}^* \in [0, 1]$ indicates the final rank of any i^{th} supplier in the supply

[8] See Appendix A for a comparison between our *Input Rank* on supply networks and the original *PageRank* on social networks and web engines.

network of the *root* producer.

To understand the workings of the underlying Markov process, let us consider the updated distance, $\boldsymbol{\pi}_{rk} - \boldsymbol{\pi}_r^*$, which is the distance the *root* producer is from the true *Input Rank* solution after the k^{th} exploration. It satisfies the recursion:

$$\boldsymbol{\pi}_r^* - \boldsymbol{\pi}_{rk} = \alpha \boldsymbol{D}(\boldsymbol{\pi}_r^* - \boldsymbol{\pi}_{r,k-1}) \tag{2.4}$$

Since matrix $\boldsymbol{D}$ has a spectral radius by construction smaller than 1, $\boldsymbol{\pi}_{rk}$ tends to $\boldsymbol{\pi}_r^*$ when the number of time-steps becomes large enough, $k \to \infty$.

Let us assume that we start approximating $\boldsymbol{\pi}_r^*$ with an initial value $\boldsymbol{\pi}_{r0} = \boldsymbol{h}_r$. That is, let us assume that a producer's exploration starts from scratch, with no information at all on any direct or indirect transaction, when leaving headquarters represented by the unitary vector $\boldsymbol{h}_r$, while going through the recursion of eq. 2.4 up to a sufficiently large number of time-steps t.

For smaller networks, the Jacobi or the Power iterative methods for the solution of a linear system of equations are sufficient (Gentle, 1998). For bigger networks, the convergence of eq. 2.4 could be difficult to obtain and some adaptive methods have been suggested (e.g., Kamvar et al. (2003)), according to which single nodes whose centrality has converged are not considered in following iterations, in this way introducing a degree of approximation.[9]

In the simpler network represented by a relatively small input-output table, as the one we will use from Section 2.4, the following solution of the linear production system can be derived:

$$\boldsymbol{\pi}_r^* = (1-\alpha)(\boldsymbol{I} - \alpha \boldsymbol{D})^{-1} \boldsymbol{h}_r \tag{2.5}$$

[9]For detailed references on the mathematical properties of *PageRank* tools, see Langville and Meyer (2011), and Gleich (2015).

2.3.3 Introducing Input-Specific Frictions

So far, the parameter α has been considered as an arbitrary constant that represents a general information wedge that the *root* producer encounters any time she must gather information on the characteristics of any transaction. From another perspective, this parameter discounts the ability of a producer to reach on bigger and complex webs of suppliers, when the time is scarce to navigate through all faraway transactions. The inclusion of such a parameter is certainly useful along increasingly fragmented global supply networks when producers shall navigate through many industries and countries. Conversely, the generic parameter $(1 - \alpha)$ brings back the *root* producer to headquarters, possibly to start navigation on another production path contained in the transitive closure, $P_{ir}^{-} \in P_{r}^{-}$.

In this section, we introduce the possibility that this parameter is a proxy for an input-specific friction in a variant of the *Input Rank*. In other words, we introduce a full vector $\boldsymbol{\alpha}$, whose single elements, α_i, is a (normalized) indicator of input contractibility (Rauch, 1999; Nunn, 2007; Nunn and Trefler, 2008), which assesses how much that input is referenced priced and/or exchanged on thick markets. In other words, a higher input contractibility implies that the *root* producer can more easily gather the information on that transaction, for example because there is some standard contract that has been signed between the parties. If a single transaction is relatively more contractible, the *root* producer proceeds more easily *upstream* to explore the supply network beyond that transaction.

In this case, the Markov process can be expressed as:

$$\boldsymbol{\pi}_{rt} = \boldsymbol{\alpha} \boldsymbol{D} \boldsymbol{\pi}_{r,t-1} + (\mathbf{1} - \boldsymbol{\alpha}) \boldsymbol{h}_{r} \tag{2.6}$$

where $\mathbf{1}$ is a vector with all elements equal to one, and $\boldsymbol{\alpha}$ is the normalized contractibility vector. Main mathematical properties are kept, as from Section 2.3.2, including convergence in bigger networks after a reasonably big number of time steps, following iterative methods. In the following analyses, we will employ both this variant and the one in Section 2.3.2 for an empirical validation of the *Input Rank* on a smaller network generated by input-output tables.

2.4 An Application to U.S. Input-Output Tables

In principle, the *Input Rank* can be computed for any producer plunged into a supply network made of firm-to-firm transactions. In this contribution, we pick more aggregate transactions sourced from the U.S. 2002 input-output tables, compiled by the Bureau of Economic Analysis (BEA), which we consider as a good training case for several reasons.

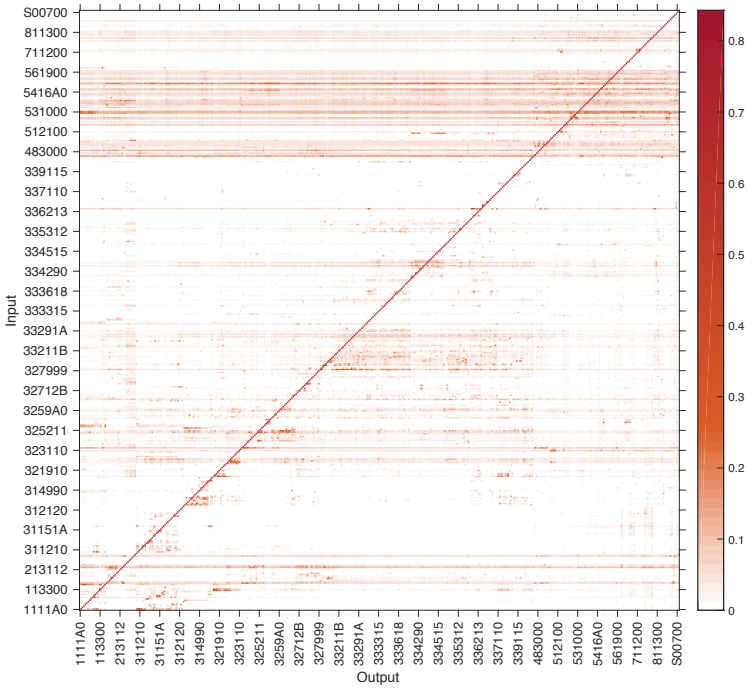
First, U.S. BEA tables represent a reasonably detailed picture of a production networks established among 425 industries, in absence of actual firm-to-firm transactions. Second, the same U.S. tables have been extensively used in previous works to study both production networks (Carvalho, 2014) and vertical integration choices (Acemoglu et al., 2009; Alfaro et al., 2016). Third, the same data have been recently used to compute *upstreamness* and *downstreamness* metrics[10], as a proxy for the technological distance on supposedly linear production sequences.

In fact, we already visualised in Figure 2.1 a solid and complex production network generated by the U.S. input-output tables, which collects 51,768 linkages. After a closer look at it, we can also register a strong heterogeneity of sourcing strategies developed among its 425 industries. In Appendix B Figures B.2.1 and B.2.2, we report both the in-degree and out-degree distributions by industry, i.e. the number of inputs received, and the deliveries made by each node of the U.S. production network. As expected, the industry with the highest number of input industries (296) is the Retail Trade (code 4A0000), because retailers professionally sell physical goods to consumers. Interestingly, the industry with the highest number of purchasing industries (425) is the Wholesale Trade (code 420000), because wholesalers professionally distribute intermediate physical inputs to all industries. However, the 'global' centrality measured by out-degrees in Figure 2.1 is of scarce interest for our scope. More

[10]The 2002 U.S. Input-Output tables have been used for the computation of absolute *downstreamness* metrics in Antràs and Chor (2013), as an exercise based on previous *upstreamness* metrics proposed in Antràs et al. (2012). Alfaro et al. (2017) more recently proposed an alternative output-specific relative *upstreamness* computed on an older 1992 version of the U.S. Input-Output tables.

properly, the *Input Rank* introduced in Section 2.3 shall return a measure of 'local' eigenvector centrality, which better catches the technological relevance of an input with respect to any *root* producer.

Figure 2.4: A visualization of the *Input Rank* computed on U.S. BEA 2002 I-O tables



Source: Own elaboration.
Note: *Input Rank* vectors are computed for each *root* output among 425 industries classified at the 6-digit in the U.S. BEA 2002 tables after using the power iteration method. Inputs on the y-axes and outputs on the x-axes by alphabetical order. A darker cell implies that input is more technologically relevant for that output.

In Figure 2.4, we visualise the results of the computation after following a power iteration method to derive the *Input Rank* as a vector of industry-pair values with a MATLAB code that we cross-validate with a Python code, assuming a stochastic process described as in eq. 2.3. Each 6-digit

industry is an input when on the y-axis, and it is an output when on the x-axis. A darker row implies that that industry is more technologically relevant across most industries. Interestingly, in the upper part of the figure, we find that services industries have a relatively more important role than manufacturing industries used as inputs across either manufacturing or services industries. Within manufacturing outputs, a crucial role is played by inputs coming from more aggregate Primary Metal Manufacturing (code 331) and Fabricated Metal Product Manufacturing (code 332). As expected, Mining industries (code 21) are technologically relevant for manufacturing producers.

Table 2.1: Top 20 inputs (all industries) by *Input Rank* (alpha = 0.5), from U.S. BEA 2002 I-O tables

IO code	Input name	mean	p50	sd	min	max
550000	Management of companies and enterprises	0.0323	0.0306	0.0143	0.0068	0.0936
420000	Wholesale trade	0.0277	0.0279	0.0124	0.0030	0.0949
531000	Real estate	0.0235	0.0170	0.0174	0.0066	0.1215
541800	Advertising and related services	0.0145	0.0125	0.0078	0.0042	0.0606
221100	Electric power generation, transmission, and distribution	0.0116	0.0093	0.0079	0.0023	0.0749
52A000	Monetary authorities and depository credit intermediation	0.0115	0.0092	0.0072	0.0041	0.0589
517000	Telecommunications	0.0093	0.0073	0.0062	0.0032	0.0666
484000	Truck transportation	0.0090	0.0079	0.0065	0.0011	0.0785
331110	Iron and steel mills and ferroalloy manufacturing	0.0088	0.0022	0.0156	0.0003	0.1192
523000	Securities, commodity contracts, investments, and related activities	0.0084	0.0064	0.0142	0.0026	0.2471
324110	Petroleum refineries	0.0083	0.0045	0.0141	0.0017	0.1307
561300	Employment services	0.0078	0.0053	0.0062	0.0028	0.0382
211000	Oil and gas extraction	0.0072	0.0040	0.0144	0.0012	0.1975
541100	Legal services	0.0071	0.0064	0.0029	0.0030	0.0246
533000	Lessors of nonfinancial intangible assets	0.0070	0.0059	0.0052	0.0017	0.0770
541610	Management, scientific, and technical consulting services	0.0065	0.0049	0.0044	0.0018	0.0451
722000	Food services and drinking places	0.0061	0.0048	0.0040	0.0018	0.0250
230301	Nonresidential maintenance and repair	0.0054	0.0043	0.0056	0.0018	0.0790
522A00	Nondepository credit intermediation and related activities	0.0054	0.0041	0.0065	0.0022	0.1042

In Tables 2.1 and 2.2, we report some moments of *Input Rank* distribution: first for all the top 20 inputs, then for the top 20 manufacturing inputs, excluding services. These are useful to look at some details of the input usage. Here, as well, services industries are on average ranked

Table 2.2: Top 20 inputs (manufacturing only) by *Input Rank* (alpha = 0.5), from U.S. BEA 2002 I-O tables

IO code	Input name	mean	p50	sd	min	max
331110	Iron and steel mills and ferroalloy manufacturing	0.0088	0.0022	0.0156	0.0003	0.1192
324110	Petroleum refineries	0.0083	0.0045	0.0141	0.0017	0.1307
336300	Motor vehicle parts manufacturing	0.0052	0.0024	0.0143	0.0010	0.1686
325211	Plastics material and resin manufacturing	0.0052	0.0015	0.0139	0.0002	0.1584
325190	Other basic organic chemical manufacturing	0.0051	0.0019	0.0108	0.0003	0.0934
334413	Semiconductor and related device manufacturing	0.0041	0.0030	0.0061	0.0004	0.0792
322210	Paperboard container manufacturing	0.0039	0.0022	0.0051	0.0003	0.0418
32619A	Other plastics product manufacturing	0.0039	0.0020	0.0044	0.0005	0.0299
334418	Printed circuit assembly (electronic assembly) manufacturing	0.0035	0.0024	0.0047	0.0003	0.0400
321100	Sawmills and wood preservation	0.0030	0.0006	0.0109	0.0002	0.1318
323110	Printing	0.0030	0.0016	0.0057	0.0007	0.0704
322120	Paper mills	0.0028	0.0010	0.0086	0.0002	0.0863
326110	Plastics packaging materials and unlaminated film and sheet manufacturing	0.0027	0.0010	0.0045	0.0001	0.0380
332710	Machine shops	0.0026	0.0019	0.0025	0.0002	0.0143
3259A0	All other chemical product and preparation manufacturing	0.0023	0.0015	0.0026	0.0003	0.0207
322130	Paperboard mills	0.0021	0.0012	0.0051	0.0002	0.0627
33131A	Alumina refining and primary aluminum production	0.0020	0.0003	0.0084	0.0001	0.1146
332800	Coating, engraving, heat treating and allied activities	0.0019	0.0018	0.0015	0.0001	0.0081
325220	Artificial and synthetic fibers and filaments manufacturing	0.0019	0.0001	0.0105	0.0000	0.1271

higher than manufacturing industries. The first highly ranked input is the Management of Companies and Enterprises (code 550000)[11], which unquestionably points to a general professional nature of the management of U.S. companies. Some post-production services also rank relatively high, as expected, as in the case of Wholesale Trade (code 420000) and Advertising (code 541800). Immediately after, we spot Electric Power Generation (code 221100) and bank credit, as included in Monetary Authorities and Depository Credit Intermediation (code 52A000). In Appendix B Table B.2.1, we look at the rank of R&D input services (code 541700) and discover that the latter are more relevant to the General Federal Defense Government Services (code S00500) than to other life sciences industries (In-vitro Diagnostic Substance Manufacturing, code 325413; Biological Product Manufacturing, code 325413; Pharmaceutical Preparation Man-

[11] As from the original definition (BLS, 2018): This sector comprises: i) companies that hold financial activities (securities or other equity interests) in other companies for the purpose of a corporate control to influence management decisions; ii) companies that professionally administer, oversee, and manage other companies through strategic or organizational planning and decision making.

ufacturing, code 325412; Medicinal and Botanical Manufacturing, code 325411).

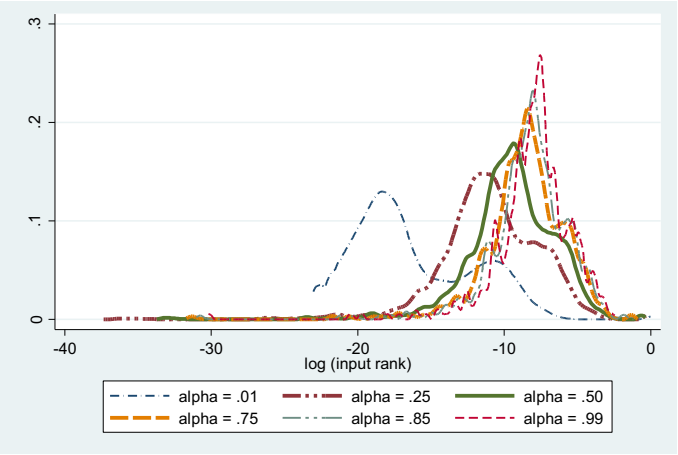
The first manufacturing input encountered among the top 20 is the Iron and Steel Mills and Ferroalloy Manufacturing (code 331110), which comes only after Truck Transportation (code 484000). In general, we observe a high variation of the *Input Rank* across *root* industries, as shown by relatively high values of the standard deviations.[12] In this case, it is more useful to look from the perspective of selected *root* industries (Electronic Computer Manufacturing, code 334111; Automobile Manufacturing, code 336111), in Appendix B Tables B.2.1 and B.2.2 we find that the *Input Rank* is indeed specific to the technological nature of the production processes.

2.4.1 The Sensitivity to the Parameter α

In general, the parameter α can be interpreted as the probability to proceed further in the exploration of a supply network, therefore its complement to one is the probability to stop exploration and fall back to headquarters at each time-step. In Section 2.3.3, we discussed how we can make use of this parameter to introduce an input-specific friction: the contractibility of an input. The latter would catch the thickness of the input market in the fashion of Rauch (1999); Nunn (2007); Nunn and Trefler (2013). In this case, we could assume that a *root* producer can more easily gather information on the implicit characteristics of the transaction thanks, for example, to the existence of a reference price or the signature of a standard supply contract for that input. The alternative is to make reference to other notable constants previously used in the use of some *PageRank* centralities. For example, Brin and Page (1998) originally suggest a damping factor $\alpha = 0.85$. As neutral as possible, $\alpha = 0.5$ implies an equal probability of proceeding through exploration or stopping at each time-step. Indeed, the latter is the reference value we use for main descriptive statistics and for

[12] Please note how in absence of actual firm-to-firm transactions, different aggregations of the input-output industries may alter the *Input Rank*. We expect that a higher aggregation of an input industry entails an overestimation of its *Input Rank* for any other *root* industry. In this case, we shall rely on official statistics offices that separate industries based on their technological relevance in modern economies, as is the scope of periodic updates of input-output tables.

Figure 2.5: Sensitivity of *Input Rank* to changing values of the parameter alpha



Source: Own elaboration.

baseline regressions in this text. However, we will make sure in the next paragraphs that our econometric results are robust to changing thresholds of the parameter α.

In Figure 2.5, we plot the changing shapes of the (log) distributions of the *Input Rank*, as computed on industry pairs from the U.S. 2002 input-output tables, at different constant values of the parameter α. In Table 2.3, we report pairwise correlations among these distributions, including also the case when α_i is the (normalised) input specific contractibility. Finally, in Appendix B Figure B.2.3, we compare the latter case with a parameter $\alpha = 0.5$. In general, we find that shapes can be very similar, although shifting to the left at lower parameter values, because they discount relatively less faraway input industries in the supply network. In the end, the rankings of inputs are highly correlated at varying values of the parameter, at least before approaching the unit value.

From Table 2.3, it is evident that when $\alpha = 0.99$ the distribution becomes very different. In fact, the stochastic process described in eq. 2.4 degenerates when $\alpha \to \infty$. In Appendix B Figure B.2.3, we find that

the distribution after the introduction of an input-specific friction, as in the case of input contractibility, resembles the case of $\alpha = 0.5$. When we perform our econometric exercises in the next sections, we will consider both $\alpha = 0.5$ and input-specific α_i as baseline metrics, while checking for the robustness of our findings at different threshold values of the parameter α. As we will see, the magnitudes can change considerably but statistical significance will not. Therefore, one can choose which value to assign to the parameter α as dependent on the nature of the frictions to proxy in the supply network.

Table 2.3: Pairwise correlations. *Input Rank* computed at several values of the alpha

Values of alpha	**=.01**	**=.25**	**=.50**	**=.75**	**=.85**	**=.99**	**= input contractibility**
=.01	1.00						
=.25	0.99	1.00					
=.50	0.98	0.99	1.00				
=.75	0.92	0.94	0.97	1.00			
=.85	0.80	0.83	0.88	0.97	1.00		
=.99	0.08	0.11	0.18	0.39	0.61	1.00	
= input contractibility	0.78	0.79	0.80	0.78	0.72	0.18	1.00

2.5 The Role of the *Input Rank* in Choices of Vertical Integration

The decision to *make or buy* an input is an example of a situation when a producer needs gathering information on the technological relevance of direct and indirect inputs. In this section, we test whether the *Input Rank* can play a role as a determinant for the decision to integrate a production stage within the firm boundary (i.e. vertical integration), as an alternative to signing supply contracts with independent firms (i.e. outsourcing). For our purpose, we will make use of a dataset of U.S. parent companies that have integrated at least one production stage over time. Our empirical strategy explicitly takes on the theoretical framework by Antràs and Chor (2013), while augmenting the estimates by Del Prete and Rungi (2017) with

the inclusion of the *Input Rank*, therefore assuming that a *root* producer has an albeit reduced ability to collect information on her supply network.

2.5.1 A Sample of U.S. Parent Companies

Our firm-level data are sourced from the Orbis database, compiled by the Bureau van Dijk, which gathers financial and ownership information for companies on a global scale. For our scope, we collect information on 20,489 U.S. parent companies controlling 154,836 subsidiaries in 210 countries at the end of the year 2015[13]. The selection of U.S. parent companies is coherent with the subsequent use of regressors that are based on U.S. input-output tables and trade data. In Table 2.4, we provide some descriptive statistics of the geographic coverage of the subsidiaries.

Both subsidiaries and parent companies can be active in any industry: manufacturing (28.86%), services (69%), primary (0.29%), and extractive (1.85%). About 81% of subsidiaries integrated by U.S. parents are domestic. Not surprisingly, U.S. parent companies are involved mainly in global supply networks that spread across other OECD economies, where 96% of their subsidiaries are located. The member states of the European Union host the largest number of foreign subsidiaries. Among them, Germany, the United Kingdom, and the Netherlands attract a significant share of U.S. foreign affiliates active in services industries. Not surprisingly, member States of NAFTA, Canada and Mexico, mainly host manufacturing of final and intermediate goods. However, a non-negligible share of subsidiaries is present in Asia, Africa and the Middle East.

To validate our sample, we compare with official 'Data on Activities of Multinational Enterprises' (BEA, 2018) and OECD Statistics on Measuring Globalization (OECD, 2018). In 2015, BEA (2018) reports 6,880 billion dollars of total sales by foreign affiliates and 12,628 billion dollars of total sales by parent companies. The U.S. multinational enterprises present in

[13]To build our sample of parents and subsidiaries, we follow international standards for complex ownership structures (OECD, 2005; UNCTAD, 2009, 2016), according to which the unit of observation is the control link between a parent company and each of its subsidiary after a concentration of voting rights is detected ($> 50\%$). See Rungi et al. (2017). Similar data structures were used in Alviarez et al. (2013), Cravino and Levchenko (2017), Del Prete and Rungi (2017).

our sample account for 94% and 92% of the BEA (2018) values, respectively. The number of foreign affiliates in our sample corresponds to 88.6% on the total of U.S. foreign subsidiaries reported in OECD (2018), although the latter source only reports the latest value valid for the year 2014.

For the scope of our analysis, we map industry affiliations of both parent companies and their subsidiaries, from the NAICS rev. 2012 classification into the 2002 U.S. BEA input-output tables. The match by industry affiliations allows us combining firm-level data with sector-level metrics, like the *Input Rank* we computed in Section 2.4 and the *upstreamness* segments sourced from Alfaro et al. (2017). In absence of actual data on firm-to-firm transactions, such a mapping onto input-output tables[14] allows us proxying the technological relevance of a (direct or indirect) input with reference to a *root* output, in the case of the *Input Rank*, and the relative technological distance of an input from the target output, in the case of the *upstreamness* segment. Finally, we complement our data with industry-level estimates of demand elasticity from Broda and Weinstein (2006), and with a measure of input contractibility retrieved from Antràs and Chor (2013) based on the methodology by Nunn (2007).

2.5.2 Baseline Results

We test a conditional logit model with parent-level fixed effects.[15] The fixed effects conditional logit is a natural empirical strategy for the multinomial case of *ex-ante* alternatives. That is, we can test the determinants of vertical integration choices controlling for the characteristics of the production stages that were both vertically integrated and not integrated by the parent company.

Let $i = 1, 2, ..., N$ denote all the inputs, as from the input-output tables, and let $r = 1, 2, ..., R$ denote the *root* parent companies. The dependent

[14] For similar mappings of firm-level data into input-output tables by industry affiliations, see Alfaro and Charlton (2009), Acemoglu et al. (2010), Alfaro et al. (2016), Del Prete and Rungi (2017).

[15] See McFadden (1974) and Chamberlain (1980) for more details. Present notation is borrowed from Hamerle and Ronning (1995) and Hosmer et al. (2013). See also Head et al. (1995) and Del Prete and Rungi (2017) for previous applications in international economics, the latter with reference to firm-level vertical integration choices.

Table 2.4: Sample geographic coverage by country of subsidiaries

Country of subsidiaries	**Final goods**		**Intermediates**		**Services**		**All industries**	
	N.	**%**	**N.**	**%**	**N.**	**%**	**N.**	**%**
United States	20,571	16.3	24,590	19.5	80,729	64.1	125,890	100.0
European Union	1,934	11.5	2,084	12.3	12,872	76.2	16,890	100.0
of which:								
Germany	273	13.2	306	14.8	1.494	72.1	2,073	100.0
France	171	11.0	213	13.7	1,167	75.2	1,551	100.0
United Kingdom	563	11.4	624	12.7	3,734	75.9	4,921	100.0
Italy	136	19.4	139	19.8	427	60.8	702	100.0
Netherlands	158	6.8	171	7.3	2,005	85.9	2,334	100.0
Canada	980	30.4	923	28.6	1,325	41.1	3,228	100.0
Russia	18	11.7	30	19.5	106	68.8	154	100.0
Asia	251	15.0	312	18.7	1.109	66.3	1,672	100.0
of which:								
Japan	87	11.5	76	10.1	592	78.4	755	100.0
China	92	12.1	66	8.7	605	79.3	763	100.0
India	122	15.7	149	19.1	508	65.2	779	100.0
Africa	67	14.2	93	19.7	313	66.2	473	100.0
Middle East	82	18.2	80	17.8	288	64.0	450	100.0
Latin America	221	12.1	395	21.6	1,210	66.3	1,826	100.0
of which:								
Argentina	24	8.1	70	23.6	203	68.4	297	100.0
Brazil	137	14.6	219	23.3	583	62.1	939	100.0
Mexico	98	23.3	154	36.6	169	40.1	421	100.0
Australia	123	14.2	157	18.1	586	67.7	866	100.0
Rest of the world	489	16.5	585	19.7	1,892	63.8	2,966	100.0
Total	**24,834**	**16.0**	**29,403**	**19.0**	**100,599**	**65.0**	**154,836**	**100.0**

Note: intermediate and final manufacturing categories based on industry affiliates and following the BEC rev. 4 classification provided by the UN Statistics Division.

variable, y_{ir}, takes on a value of 1 when at least one subsidiary has been integrated that produces the i^{th} input, and 0 otherwise. Therefore, for each r^{th} parent company, we have a vector $\boldsymbol{y}_r = (y_{1r}, ..., y_{Nr})$ made of 0s and 1s when each input has been integrated or not, respectively.

We want to consider the probability that a generic parent chooses a value of $\boldsymbol{y}_r$ conditional on $\sum_{i=1}^{N} y_{ir}$:

$$Pr\left(\boldsymbol{y}_r | \sum_{i=1}^{N} y_{ir}\right) = \frac{exp(\sum_{i=1}^{N} y_{ir}\boldsymbol{x}_{ir}\boldsymbol{\beta})}{\sum_{s_i \in S_i} exp(\sum_{i=1}^{N} y_{ir}\boldsymbol{x}_{ir}\boldsymbol{\beta})} \tag{2.7}$$

where the element s_{ir} of the vector $\boldsymbol{s}_i$ is equal to 1 when the i^{th} input is integrated, and 0 otherwise.

For each input-parent pair, we can identify a vector of covariates, $\boldsymbol{x}_{ir}$, which includes: the *Input Rank* relative to the i^{th} input with respect to the output of the r^{th} parent company; the interaction term of the *Input Rank* with the binary variable *Complements* relative to the i^{th} input; the *Input upstreamness* sourced from Alfaro et al. (2017) measuring the technological distance of the i^{th} input to the r^{th} target output; the interaction term of the *Input upstreamness* with *Complements*; the input-specific *Contractibility* derived as in Nunn (2007); the *Direct requirement* coefficient available from the U.S. input-output tables that ranges in $[0, 1]$. In this context, the variable *Complements* is equal to 1 when the elasticity of substitution of the parent industry is above the median $\rho_r > \rho_{med}$, and 0 otherwise ($\rho_r < \rho_{med}$). Errors are clustered by parent companies and variables are standardized. Results from nested specifications are reported in Tables 2.5 and 2.6. In Table 2.5, we report findings when we assume there is an equal probability that a producer proceeds exploring at each time-step the supply network or she falls back to the headquarters. In Table 2.6, we consider that the probability to proceed with the exploration is dependent on the input-specific contractibility. In the latter case, we do not include the input contractibility as a separate variable in the specifications.

The coefficient of immediate interest to us is the one on the *Input Rank*, which indicates whether the odds of vertical integration change for a more central input in the supply network. It is positive and significant throughout our estimations. In the first columns, we consider all parent companies whether active in a manufacturing or a service industry. We find that one additional standard deviation of the *Input Rank* is correlated with 1.35 higher odds of vertical integration. Please note that in further columns, when we introduce subsequent controls, the sample reduces to manufacturing parents only, because the elasticity of substitution by Broda and Weinstein (2006) is estimated on U.S. imports of manufacturing only. Our specification is complete in the fourth column of Table 2.5 and in the third column of Table 2.6, where a standard deviation increase of the *Input Rank* correlates with 1.16 and 1.08 higher odds of vertical integration, respectively.

Our findings are robust after the inclusion of the *Input upstreamness*,

Table 2.5: Baseline regressions I, parent-level fixed effects conditional logit

Dependent variable: **Input is integrated = 1**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Input Rank (alpha = 0.5)	0.318***	0.196***	0.197***	0.149***	0.189***	0.085***
	(0.001)	(0.003)	(0.003)	(0.004)	(0.005)	(0.006)
Input Rank * Complements				0.090***	0.014*	0.209***
				(0.005)	(0.008)	(0.008)
Input upstreamness		-0.595***	-0.583***	-0.782***	-0.443***	-1.166***
		(0.023)	(0.025)	(0.031)	(0.035)	(0.068)
Input upstreamness * Complements		-0.044*	-0.037	0.336***	-0.177***	0.955***
		(0.025)	(0.028)	(0.039)	(0.048)	(0.078)
Contractibility			-0.385***	-0.390***	-0.645***	-0.249***
			(0.018)	(0.017)	(0.032)	(0.025)
Direct requirement	0.063***	0.049***	0.025***	0.015***	0.010*	0.026***
	(0.004)	(0.003)	(0.003)	(0.005)	(0.006)	(0.004)
N. observations	8,564,068	1,437,785	1,151,908	1,151,908	595,218	542,872
N. parent companies	20,294	4,203	4,084	4,084	2,110	1,925
Pseudo R-squared	0.409	0.215	0.250	0.257	0.203	0.342
Log pseudolikelihood	-96,831.2	-29,841.3	-22,779.5	-22,560.8	-12,739.4	-9,281.4
Clustered errors by parent	Yes	Yes	Yes	Yes	Yes	Yes
Activity of parent companies	All	Manu-facturing	Manu-facturing	Manu-facturing	Final goods	Intermediate goods

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

Table 2.6: Baseline regressions II, parent-level fixed effects conditional logit

Dependent variable: **Input is integrated = 1**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Input Rank (alpha = contractibility)	0.307***	0.140***	0.074***	0.129***	0.016**
	(0.001)	(0.003)	(0.004)	(0.006)	(0.007)
Input Rank * Complements			0.118***	0.034***	0.181***
			(0.008)	(0.010)	(0.011)
Input upstreamness		-0.769***	-0.860***	-0.685***	-1.078***
		(0.026)	(0.028)	(0.029)	(0.063)
Input upstreamness * Complements		-0.074**	0.080**	-0.144***	0.399***
		(0.033)	(0.033)	(0.034)	(0.040)
Direct requirement	0.122***	0.076***	0.073***	0.084***	0.065***
	(0.004)	(0.002)	(0.003)	(0.005)	(0.003)
N. observations	8,564,068	1,437,785	1,437,785	745,554	675,473
N. parent companies	20,294	4,203	4,203	2,179	1,975
Pseudo R-squared	0.369	0.166	0.172	0.133	0.219
Log pseudolikelihood	-103,307.9	-31,720.9	-31,485.0	-17,830.4	-13,396.9
Clustered errors by parent	Yes	Yes	Yes	Yes	Yes
Activity of parent companies	All	Manu-facturing	Manu-facturing	Final goods	Intermediate goods

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

which should proxy the relative technological distance from an input and its target output. In this case, more distant inputs are less likely integrated by the parent company. The central tenet of the theoretical framework by Antràs and Chor (2013) and Alfaro et al. (2017) is tested by the sign of the interaction term between the *Input upstreamness* and *Complements*. According to these authors, when final demand is sufficiently elastic (inelastic), parents integrate production stages that are more proximate to (far from) final demand. This seems to be the case for producers of final goods (penultimate columns in Tables 2.5 and 2.6, although a sign reversal is observed in the case of *midstream* parents (last columns), i.e. when considering integration choices by producers of intermediate goods, in line with what tested by Del Prete and Rungi (2017).

In an effort to extend the role of the elasticity of substitution to the case of supply networks, we include a similar interaction term of the variable *Complements* with the *Input Rank*. In this case, when final demand is sufficiently elastic, we find that the odds that a central input is integrated within the boundary of the firm are proportionally higher.

Please note how, as expected, the *Direct requirement* and the input-specific *Contractibility* have a positive and negative coefficient, respectively. In the first case, a higher value of the transaction (if any) is trivially correlated with higher odds of vertical integration. In the second case, a more contractible input is less likely to be integrated because the agreement between a producer and an independent supplier can be more easily enforced by law, thus the incentives for vertical integration are lesser.

2.5.3 Robustness Checks

Our main findings are robust to several checks of robustness. First, in Table 2.7, we check whether sample compositions can have an impact on the sign and significance of coefficients.

In the first column, we exclude cases of inputs coming from the same 2-digit industry of the parent companies. In the second column, we exclude services inputs because some of them could uniquely lead to previous results, as they are more central than manufacturing (see Figure 2.1) in most production processes. In the third column, we modify our indicator of

Table 2.7: Robustness on sample composition, parent-level fixed effects conditional logit

Dependent variable: Input is integrated = 1	**No horizontal**	**Only manuf inputs**	**Input vs output elast**	**Top 100 inputs**
Input Rank (alpha = 0.5)	0.749***	0.152***	0.171***	0.134***
	(0.151)	(0.004)	(0.004)	(0.005)
Input Rank * Complements	0.306***	0.091***	0.064***	0.118***
	(0.076)	(0.006)	(0.005)	(0.006)
Input upstreamness	-0.892***	-0.768***	-0.649***	-0.611***
	(0.035)	(0.030)	(0.026)	(0.051)
Input upstreamness * Complements	0.286***	0.313***	0.137***	0.748***
	(0.046)	(0.040)	(0.038)	(0.062)
Contractibility	-0.505***	-0.289	-0.413***	-0.453***
	(0.025)	(0.016)	(0.017)	(0.023)
Direct requirement	-0.028*	-0.008	0.028***	0.040***
	(0.015)	(0.005)	(0.003)	(0.003)
N. observations	741.066	905.64	1,151,908	156.705
N. parent companies	2.637	3.903	4.084	2.847
Pseudo R-squared	0.080	0.281	0.254	0.398
Log pseudolikelihood	-19,399.7	-18,622.2	-22,653.3	-7,949.5
Clustered errors by parent	Yes	Yes	Yes	Yes
Activity of parent companies	Manu-facturing	Manu-facturing	Manu-facturing	Manu-facturing

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

Complements, explicitly considering the difference between the elasticities of the output and of the input ($\rho_r - \rho_i$), which more specifically provides a reference point to understand how much elastic the demand of the *root* producer is. In the fourth column, we reduce our sample to the top 100 (direct) inputs of the parent output, as from I-O tables, to check whether the role of the *Input Rank* is exclusively driven by direct vs indirect inputs. In all these cases, when an input is more technologically relevant in the supply network, the odds are higher that the parent companies will *make* rather than *buy* the input from an independent supplier.

In Appendix Tables B.2.4, B.2.5, B.2.6 and B.2.7, we further control for: i) sample compositions when the *Input Rank* is built by considering the input-specific contractibility at each time-step of the network exploration (see Section 2.3.3); ii) sample compositions when we consider only midstream manufacturing parents; iii) changing values of the constant parameter α; iv) empirical specifications different from the fixed-effects conditional logit. All main findings are similar in sign and significance with baseline estimates, with the exception of a lack of statistical significance of the coefficient of the *Input Rank* in Table B.2.5, when we exclude possibly horizontal strategies in the 2-digit industries of the parent company.

2.6 Conclusions

In this contribution, we introduced the *Input Rank* as an eigenvector centrality measure applicable to recursive production networks made of transactions among many buyers and suppliers. It measures the technological relevance of suppliers in the entire supply network of a *root* producer. Adapted from previous metrics of information consumption in social networks and web engines, the *Input Rank* proxies a stochastic process that is started by a *root* producer, who needs gathering information on the technology of her entire supply network made of both direct and indirect suppliers. After *random walks* throughout her supply network (e.g., random phone calls), she comes at each time-step with a numerical value and an updated ranking of how important is that (direct or indirect) input for the completion of her production process. Given increasingly complex webs

of suppliers and an increasing fragmentation of the production processes, it is possible that a generic *root* producer is scarcely able to navigate faraway areas of her supply network, therefore a dumping parameter is introduced at each time-step to discount the difficulty in gathering information on a specific transaction, e.g., due to its contractibility, and consequently the probability that the *root* producer falls back home and starts a new journey of exploration.

The *Input Rank* can be computed either on firm-to-firm transactions or on input-output tables, keeping its simple mathematical properties. For the sake of comparison with previous positioning metrics (e.g., *downstreamness* and *upstreamness* segments) of GVCs, we compute it on U.S. 2002 BEA input-output tables, and after that, we test firm-level choices of vertical integration by U.S. parent companies. We find that a higher *Input Rank* correlates with higher odds that input is vertically integrated, even more so when the demand faced by the parent company is more elastic. We argue that vertical integration allows reducing the possibility that otherwise independent suppliers renege on commitments and disrupt the supply network, generating more damage for the completion of the production processes. Even more so when the margins on which the *root* producer can rely are smaller. Our findings are robust to several checks on sample compositions, parameter choices and empirical models.

More in general, we argue that the *Input Rank* better catches the recursive and complex nature of real-world supply networks, which have been so far represented as supposedly linear technological sequences in studies for the international organisation of production. Certainly, both empirics and theory need better considering the technological loops, kinks and corners, which can magnify or dampen a shock in a supply network, finally shaping the organisational response of the company.

Appendices to Chapter 2

A Appendix: From the *PageRank* to the *Input Rank*

The intuition of the *Input Rank* is adapted from the 'personalised' version of the *PageRank* centrality, first used in social networks and search engines (Brin and Page, 1998) to present to users the most pertinent content. Some variants of the *PageRank* have been used in many domains (bibliometrics, biology, physics, engineering of infrastructures, financial exposure, etc.) as an alternative to the Katz (1953) centrality (Gleich, 2015). The underlying assumption is that more important nodes (in our case, inputs) are likely to receive more links from other nodes (in our case, inputs of inputs), and that proximity to central nodes implies, in turn, a relatively higher centrality.

For our scope, the main limitation of the original formulation of the *PageRank* is its 'global' outreach on a supposedly unique network, whereas we are interested in a 'local' outreach of a specific *root* buyer in her oriented supply network. Therefore, we needed a 'personalisation' of the *PageRank*, in the spirit of Haveliwala (2002) and White and Smyth (2003), where different rankings are possible for different *root* nodes, given an initial prior knowledge of the stochastic process.

Starting from the original formulation of the *PageRank*, adopting the notation proposed by (Gleich, 2015), the eigenvalue problem can be represented by the following identity:

$$[(1-\alpha)\boldsymbol{P} + \alpha \boldsymbol{v}\boldsymbol{e}^T]\boldsymbol{x} = \boldsymbol{x} \tag{A.2.1}$$

If one seek an eigenvector $\boldsymbol{x}$ with $\boldsymbol{x} \geq 0$ and $\boldsymbol{e^T x} = 1$, then the *PageRank* formulation in A.2.1 is equivalent to the following linear system:

$$(\boldsymbol{I} - \alpha \boldsymbol{P})\boldsymbol{x} = (1-\alpha)\boldsymbol{v} \tag{A.2.2}$$

For our scope, we substitute each term with a corresponding in our *Input Rank*, to take into account the peculiar economic process at stake:

- In the *PageRank*, a transition matrix $\boldsymbol{P}$ contains the probabilities that an internet user clicks on one page following a web link present

on the one she is visiting, column-normalized by the total number of received links, i.e. its in-degree. In the *Input Rank*, we substitute the matrix $\boldsymbol{P}$ with an input-output matrix $\boldsymbol{D}$, whose single elements are column-normalized buyer-supplier transactions, d_{ij}. Even though in the present work we propose estimates of the *Input Rank* using input-output tables (US BEA, 2002), our framework is open to applications to firm-to-firm transactions. In the latter case, an element of the matrix $\boldsymbol{D}$ would be a firm-to-firm actual shipment, always normalized by column (i.e., expressed by receiving firm as a percentage of all input shipments).

- A vector $\boldsymbol{v}$ is a critical tool that allows for the 'personalisation' of the *PageRank*. In absence of 'personalisation', this vector contains just a uniform distribution of probability across all web pages. Therefore, personalised versions of the *PageRank* make it non-uniform, so that a particular region of the internet is highlighted with a higher probability. At the same time, the vector $\boldsymbol{e}$ is a vector of 1s that algebraically extends the same (uniform or non-uniform) distribution in $\boldsymbol{v}$ to all web users. In our *Input Rank*, we substitute $\boldsymbol{v}$ with a *root*-specific unitary vector, $\boldsymbol{h}_r$ made of all 0s except for the r^{th} element that is equal to 1. Together with the term α, the unitary vector $\boldsymbol{h}_r$ avoids that the r^{th} buyer navigates outside her network of suppliers while pointing at headquarters.

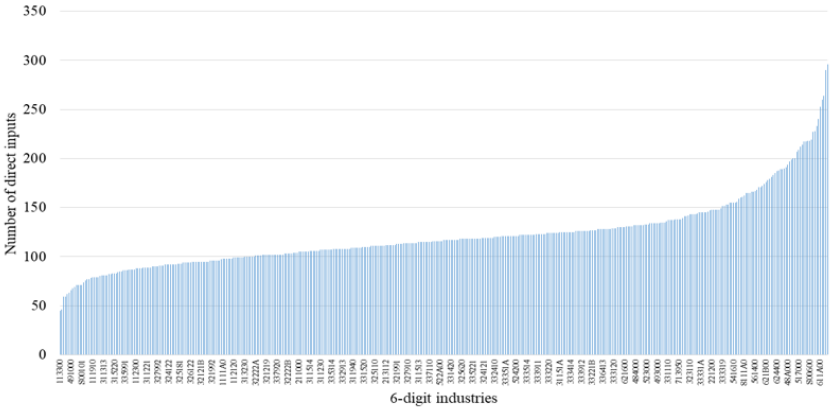
- The term $\alpha \in (0, 1)$ is a teleportation parameter in the *PageRank*, otherwise called a damping factor. It indicates the probability that a 'web surfer' interrupts a random navigation following page-to-page links and falls elsewhere, on any other web page not directly linked to the one she is visiting. By converse, $(1 - \alpha)$ is the probability that the user goes on randomly following her web path made of cross-link citations. In our *Input Rank*, α must be read in connection to the peculiarity of $\boldsymbol{h}_r$. In our case, α is the probability that the *root* producer stops travelling in her production network and goes back to the headquarters (i.e., the 1 in the unitary vector $\boldsymbol{h}_r$). Conversely, $(1 - \alpha)$ is the probability that the producer goes on exploring her

web of suppliers. Later on, in Section 2.3, we further personalize introducing input-specific α_i, which considers the contractual friction that prevents a producer to collect, for example, the input degree of contractibility (Rauch, 1999; Nunn, 2007; Nunn and Trefler, 2013).

- Finally, $\boldsymbol{x}$ is the solution to the eigenvalue problem in A.2.1, which indicates the relevance of the web content in the case of the *PageRank*. In our *Input Rank*, the *root*-specific solution $\boldsymbol{\pi}_r^*$ in eq. 2.5 represents the technological relevance resultant from the exploration of the input-output linkages by the *root* producer.

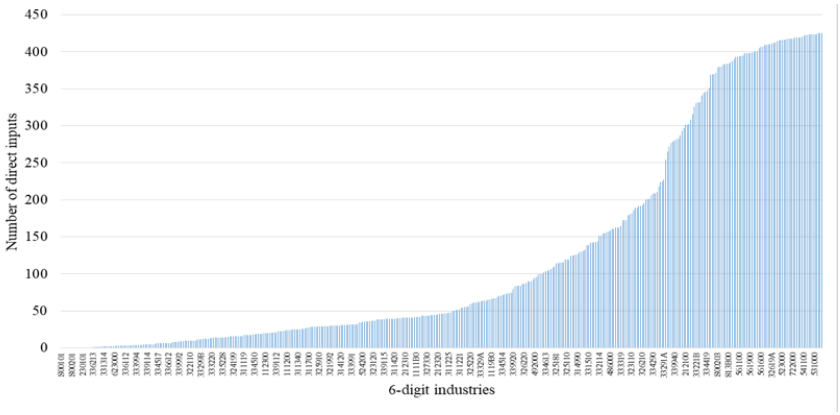
B Appendix: Figures and Tables

Figure B.2.1: In-degree distribution of input-output network from U.S. BEA 2002 I-O tables



Source: Own elaboration.
Note: Number of input industries by output ordered on the x-axis. Average: 122. Minimum at the Logging industry (code 113300) is 45. Maximum at the Retail Trade (code 4A0000) is 296.

Figure B.2.2: Out-degree distribution of input-output network from U.S. BEA 2002 I-O tables



Source: Own elaboration.
Note: Number of buying industries by output ordered on the x-axis. Average: 122. Minimum at the Museums, Historical Sites, Zoos, and Parks (code 712000) is 0. Maximum at the Wholesale Trade (code 420000) is 425.

Figure B.2.3: Distributions of the (logs of) *Input Rank* when alpha is 0.5 and alpha is input-specific contractibility

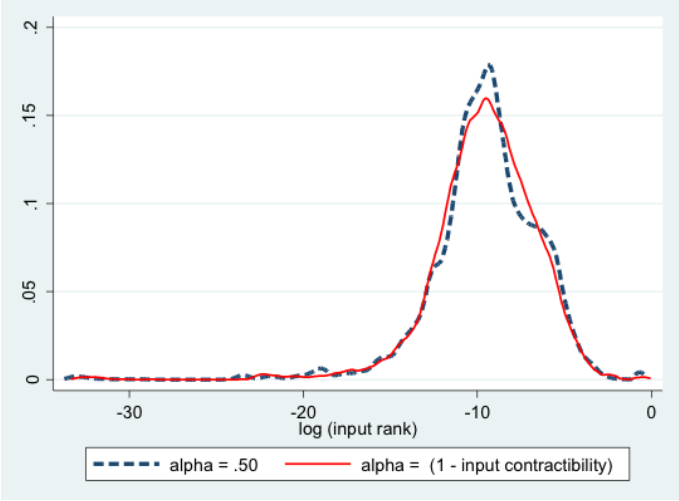


Table B.2.1: Top 10 highest *Input Rank* values of the R&D services (code 541700) by output

IO code	**Output name**	**R&D Input rank (alpha = 0.5)**
S00500	General Federal defense government services	0.0384
325413	In-vitro diagnostic substance manufacturing	0.0317
325414	Biological product (except diagnostic) manufacturing	0.0293
325412	Pharmaceutical preparation manufacturing	0.0247
325411	Medicinal and botanical manufacturing	0.0226
325320	Pesticide and other agricultural chemical manufacturing	0.0211
3259A0	All other chemical product and preparation manufacturing	0.0211
325620	Toilet preparation manufacturing	0.0193
325910	Printing ink manufacturing	0.0192
325610	Soap and cleaning compound manufacturing	0.0190

Table B.2.2: Top 10 direct or indirect inputs by *Input Rank* for the Automotive Manufacturing (code 336111)

IO code	**Input name**	**R&D Input rank (alpha = 0.5)**
336300	Motor vehicle parts manufacturing	0.1686
420000	Wholesale trade	0.0353
550000	Management of companies and enterprises	0.0302
331110	Iron and steel mills and ferroalloy manufacturing	0.0101
531000	Real estate	0.0087
541800	Advertising and related services	0.0078
334413	Semiconductor and related device manufacturing	0.0072
484000	Truck transportation	0.0071
32619A	Other plastics product manufacturing	0.0057
221100	Electric power generation, transmission, and distribution	0.0054

Table B.2.3: Top 10 direct or indirect inputs by *Input Rank* for the Electronic Computer Manufacturing (code 334111)

IO code	**Input name**	**Input rank (alpha=0.5)**
334112	Computer storage device manufacturing	0.0568
420000	Wholesale trade	0.0553
550000	Management of companies and enterprises	0.0467
334418	Printed circuit assembly (electronic assembly) manufacturing	0.0400
334413	Semiconductor and related device manufacturing	0.0374
511200	Software publishers	0.0305
33411A	Computer terminals and other computer peripheral equipment manufacturing	0.0190
541800	Advertising and related services	0.0132
531000	Real estate	0.0121
541700	Scientific research and development services	0.0112

Table B.2.4: Robustness to sample composition when alpha is input contractibility, parent-level fixed effects conditional logit

Dependent variable: **Input is integrated = 1**	**No horizontal**	**Only manuf inputs**	**Input vs output elast**	**Top 100 inputs**
Input Rank (alpha = contractibility)	0.614***	0.067***	0.114***	0.080***
	(0.050)	(0.003)	(0.004)	(0.005)
Input Rank * Complements	0.808***	0.126***	0.109***	0.136***
	(0.060)	(0.009)	(0.007)	(0.014)
Input upstreamness	-0.842***	-0.889***	-0.882***	-0.695***
	(0.032)	(0.028)	(0.026)	(0.046)
Input upstreamness * Complements	0.220***	0.025	0.091**	0.352***
	(0.039)	(0.037)	(0.039)	(0.054)
Direct requirement	0.017***	0.039***	0.064***	0.080***
	(0.005)	(0.003)	(0.003)	(0.003)
N. observations	953.458	925.155	1,151,908	252.897
N. parent companies	2.796	3.903	4.084	3.02
Pseudo R-squared	0.094	0.202	0.210	0.201
Log pseudolikelihood	-25,884.9	-20,793.8	-23,986.9	-14,925.6
Clustered errors by parent	Yes	Yes	Yes	Yes
Activity of parent companies	Manufacturing	Manufacturing	Manufacturing	Manufacturing

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

Table B.2.5: Robustness to sample composition considering *midstream* parents only, parent-level fixed effects conditional logit

Dependent variable: **Input is integrated = 1**	**No horizontal**	**Only manuf inputs**	**Input vs output elast**	**Top 100 inputs**
Input Rank (alpha = 0.5)	0.097	0.080***	0.173***	0.049***
	(0.177)	(0.006)	(0.005)	(0.007)
Input Rank * Complements	0.931***	0.223***	0.090***	0.234***
	(0.080)	(0.008)	(0.008)	(0.009)
Input upstreamness	-1.254***	-1.197***	-0.794***	-1.618***
	(0.073)	(0.065)	(0.047)	(0.116)
Input upstreamness * Complements	0.793***	1.112***	0.385***	1.904***
	(0.084)	(0.078)	(0.062)	(0.132)
Contractibility	-0.191***	-0.163***	-0.271***	-0.276***
	(0.030)	(0.024)	(0.025)	(0.029)
Direct requirement	-0.016	0.012***	0.033***	0.011**
	(0.017)	(0.004)	(0.003)	(0.005)
N. observations	316.429	437.805	542.872	87.847
N. parent companies	1.126	1.887	1.925	1.591
Pseudo R-squared	0.096	0.363	0.313	0.460
Log pseudolikelihood	-8,141.9	-8,063.6	-9,688.9	-3,848.8
Clustered errors by parent	Yes	Yes	Yes	Yes
Activity of parent companies	Intermediate goods	Intermediate goods	Intermediate goods	Intermediate goods

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

Table B.2.6: Robustness to changing values of the parameter alpha, parent-level fixed effects conditional logit

Dependent variable: **Input is integrated = 1**	**alpha = 0.01**	**alpha = 0.25**	**alpha = 0.50**	**alpha = 0.75**	**alpha = 0.85**	**alpha = 0.99**
Input Rank	0.139***	0.142***	0.149***	0.167***	0.198***	0.209***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)	(0.013)
Input Rank * Complements	0.089***	0.089***	0.090***	0.097***	0.115***	0.128***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.015)
Input upstreamness	-0.779***	-0.781***	-0.782***	-0.780***	-0.777***	-0.898***
	(0.030)	(0.031)	(0.031)	(0.031)	(0.032)	(0.030)
Input upstreamness * Complements	0.332***	0.334***	0.336***	0.342***	0.350***	-0.073*
	(0.038)	(0.039)	(0.039)	(0.039)	(0.017)	(0.039)
Contractibility	-0.388***	-0.388***	-0.390***	-0.393***	-0.395***	-0.410***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)
Direct requirement	0.036***	0.026***	0.015***	0.001	-0.007	0.070***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.002)
N. observations	1,151,908	1,151,908	1,151,908	1,151,908	1,151,908	1,151,908
N. parent companies	4.084	4.084	4.084	4.084	4.084	4.084
Pseudo R-squared	0.257	0.257	0.257	0.257	0.257	0.147
Log pseudolikelihood	-22,569.0	-22,564.7	-22,560.8	-22,553.1	-22,548.5	-25,905.4
Clustered errors by parent	Yes	Yes	Yes	Yes	Yes	Yes
Activity of parent companies	Manu-facturing	Manu-facturing	Manu-facturing	Manu-facturing	Manu-facturing	Manu-facturing

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

Table B.2.7: Robustness to changing empirical strategy

Dependent variable: **Input is integrated = 1**	**Linear probability model**	**Logit**	**Probit**
Input Rank (alpha = 0.5)	0.011***	0.151***	0.074***
	(0.001)	(0.005)	(0.002)
Input Rank * Complements	0.024***	0.242***	0.117***
	(0.001)	(0.004)	(0.002)
Input upstreamness	-0.003***	-0.766***	-0.273***
	(0.001)	(0.025)	(0.009)
Input upstreamness * Complements	-0.001***	-0.415***	-0.160***
	(0.001)	(0.025)	(0.009)
Contractibility	-0.001***	-0.354***	-0.141***
	(0.001)	(0.017)	(0.007)
Direct requirement	-0.001***	0.018***	0.011***
	(0.001)	(0.004)	(0.001)
Constant	0.005***	-5.894***	-2.789***
	(0.001)	(0.033)	(0.012)
N. observations	1,257,668	1,257,668	1,257,668
N. parent companies	4.717	4.717	4.717
R-squared/Pseudo	0.124	0.209	0.215
Log pseudolikelihood		-29,320.3	-29,098.1
Clustered errors by parent	Yes	Yes	Yes
Activity of parent companies	Manufacturing	Manufacturing	Manufacturing

Note: Errors clustered by parent in parentheses. ***, **, * stand for p-value<0.01, p-value<0.05 and p-value<0.10, respectively.

CHAPTER 3

Heterogeneous Firms Meet EU Cohesion Policy

3.1 Introduction

In parallel with the European integration, a Cohesion Policy has been developed to offset the imbalances that could benefit some regions in the core of the continent at the expense of regions at its periphery.[1] In the running financial period 2014-2020, regional policy spending amounts to almost a third of the EU budget (EUR 351.8 billion out of a total EUR 1,082 billion) and is the second largest expenditure item after the Common Agriculture Policy (CAP).

The Cohesion Policy aims to reduce regional economic disparities resulting from geographic remoteness, as different levels of prosperity and opportunity may exist both between and within the member states. In broader terms, the overall goal is 'economic, social and territorial cohe-

[1] For details on the core-periphery model and its consequences, see the seminal work by Krugman (1991).

sion', which translates into boosting competitiveness and economic growth, providing people with better services, job opportunities and better quality of life, and connecting regions. EU regional policy is implemented through a range of European Structural and Investment Funds (ESI) in a shared management system, carried out by each Member State in partnership with the European Commission. First, the Commission negotiates and approves the National Strategic Reference Framework (NSRF), setting out the main priorities for spending provided by the EU, and the Operational Programme (OP), establishing specific regions' priorities, objectives, and concrete actions to manage individual projects. Then, managing authorities in each country and/or region select, monitor, and evaluate individual projects submitted by firms, institutions or other entities. The geographical coverage and allocation of transfers are usually based on the level of GDP per capita in PPP compared to the EU average.

In this contribution, we focus on one of the main financial tools of the EU regional policy, i.e. the European Regional Development Fund (ERDF). The timing is crucial, as the European EC (2009) recently launched a pilot action to assess how to improve the local governance of the ERDF, in preparation for the new budget beyond 2020.

In particular, we restrict our attention to specific sources supporting firms in their innovation strategies and competitiveness.[2] In this case, resources are allocated to regional operational programmes that have specified thematic priorities. For instance, the ERDF for *Business Support* has been established to help firms or groups of firms, in particular, SMEs, with services and investments in innovation and sustainable production. Complementary to the latter, the ERDF for *Research, Technology and Development* (RTD) stimulates research and innovation activities through investments in research centres, promoting technology transfers and cooperation between businesses and the scientific environment. Overall, we argue that the ERDF has a dual role: first, it aims at improving the environment of regions, and second, it is a direct financial income for the

[2]One of the priorities of the EU Cohesion Policy and a key component of the renewed Lisbon Strategy Europe 2020 concern the support to firms (European Commission, 2010). Grants to firms across the Member States are mainly used to support private investment to improve private capital stock (European Commission, 2017).

recipient firms, potentially used as a source for investment.

The efficacy of regional policies is usually evaluated at aggregated levels, by country or by region. There is no consensus regarding the outcome of the *Structural Fund Programme*, and research still focuses on aggregate statistics. In fact, increasing availability of firm-level financial data allows us to estimate the benefits on their immediate recipients, i.e. the firms, which usually show heterogeneous competitiveness with power-law distributions.[3] In this case, a policy that is originally designed for the 'representative firm' in a region or in a country may eventually reveal heterogeneous and unintended responses.

In our study, we test the impact of the ERDF tools on the performance of 273,500 EU manufacturing firms, after estimating their total factor productivities (TFPs) according to the most recent semi-parametric econometric technique proposed by Ackerberg et al. (2015). Our purpose is to assess the short-term impact of the ERDF in the period 2007-2015, considering both the *Business Support* (BS) funding, and the *Research, Technology and Development* (RTD) funding.

In the Single Market, increasing economic integration is thought to have a positive impact on productivity due to stronger competitive pressure coming from the elimination of national borders. Firms compete on an EU-wide basis, and so we estimate the TFP of firms that are active across all EU 28 members. The EU-wide approach allows for a comparison of each firm with its peers *within* and *across* the national borders of the integrated market. To account for a possible sample selection bias, due to an unbalanced panel of data that tends to exclude smaller firms, we make our results robust to a Heckman (1979) correction. Further, to account for possible endogeneity between firm-level TFP and the regional allocation of ERDF funding, we make our results robust with instrumental variable panel techniques, on top of a possible sample selection bias, as in Semykina and Wooldridge (2010). In this case, we instrument the regional provisions of ERDF funding in a given region with the average ERDF funding in

[3]Firms' size and productivity are often assumed to be distributed power law, e.g., following a Pareto or a Lognormal. Among others, see the review by Gabaix (2009). For a discussion on the consequences for aggregate economic growth, see for example Luttmer (2010).

neighbouring regions, as the latter is not directly correlated with firms' TFP growth in the given region, but proximate peripheral regions suffer from the same geographical disadvantage and share the same needs of ERDF funds.

We do find a positive and statistically significant association between the *Research, Technology and Development* (RTD) vehicle by the ERDF[4] and firms' productivity growth in our period of analysis. In fact, the firms that seem to benefit more from the RTD measures are the ones in the first quartile of the productivity distribution, i.e., the least efficient in a region. By contrast, the *Business Support*[5] vehicle by the ERDF seems to be not associated with any significant impact on the productivity growth at the firm level, after accounting for possible endogeneity.

We argue that the aim of the RTD vehicle is on average reached thanks to a definition of specific targets, as direct investments in R&D activities can be directly associated to improving firms' competitiveness, boosting innovation of products and production processes. On the other hand, the more general *Business Support* vehicle may not fundamentally change the distribution of inefficiencies in the short run, for example because non-competitive firms can spend the funding to get by and avoid a process of market selection.

This Chapter is organised as follows. The next section collects the state of the literature on the evaluation of *Structural Funds Programme* and the total factor productivity estimation. Then, Section 3.3 provides a thorough description of firm-level data and the TFP estimation, which

[4]The RTD's priority themes for the period 2007-2013, laid down in Commission Regulation (EC) No 1828/2006, were: 01. R&TD activities in research centres; 02. R&TD infrastructure and centres of competence in a specific technology; 03. Technology transfer and improvement of cooperation networks between small and medium-sized businesses (SMEs), between these and other businesses and universities, post-secondary education establishments of all kinds, regional authorities, research centres and scientific and technological poles; 04. Assistance to R&TD, particularly in SMEs; 07. Investment in firms directly linked to research and innovation; 09. Other measures to stimulate research and innovation and entrepreneurship in SMEs.

[5]The *Business Support*'s priority themes for the period 2007-2013, laid down in Commission Regulation (EC) No 1828/2006, were: 05. Advanced support services for firms and groups of firms; 06. Assistance to SMEs for the promotion of environmentally-friendly products and production processes; 08. Other investment in firms; 63. Design and dissemination of innovative and more productive ways of organising work.

then we match with regional policy data. Section 3.4 presents the empirical strategy, and discusses the benchmark results and the robustness checks. The last Section 3.5 offers the summary of key findings and concluding remarks.

3.2 Related Literature

A large body of literature evaluates the regional and national effects of the EU regional policies, motivated by the relevant size of the budget and the supranational role of the European Commission in developing the policy agenda. Among others, Boldrin and Canova (2001) found little evidence that regional policies of the EU-15 were effective in terms of promoting economic growth and fast convergence in per capita income during the period until 1997. They concluded that transfers towards poorer regions had mostly a redistribution purpose. On the same line, Dall'Erba and Le Gallo (2008) found no significant effects of structural funds on the convergence of 145 European regions over the period 1989-1999. Ederveen et al. (2003) revealed that poorer regions caught up with richer regions; however, the extent to which this was due to the Cohesion Policy is ambiguous. They stated that cohesion support has a positive impact in lagging member states if their economies are open. Conversely, Cappelen et al. (2003) found a significant and positive impact of EU regional support on the growth of the European regions after the major reform of structural funds in 1988. Nevertheless, their results show that the effect of the funds was stronger in regions with a favourable industrial structure and with an emphasis put on R&D. Similarly, Leonardi (2006), using sigma- and beta-convergence argues that as a result of the Cohesion Policy from 1989 on, the gap between core and peripheral areas in the EU shrank. Assessing the impact of the Cohesion Policy is challenging because it addresses different economic and social objectives. Thus, aggregate analysis can be misleading when confounding the impacts of diverse policy fields.

A relatively meagre body of literature assesses the impact of this funding system on firms' outcomes. Bernini and Pellegrini (2011) considered subsidies to Southern Italian regions over the period 1996-2004 and found

a positive effect on output, employment and fixed assets in subsidised firms, but slower growth in the TFP than in non-beneficiaries firms. Additionally, Hartsenko and Sauga (2012) positively assess the effectiveness of different types of grants on Estonian firms' net sales. While these studies are restricted to specific European regions, De Zwann and Merlevede (2013) proposed a EU-wide investigation combining regional data with firm-level data for the period 2000-2006. Preliminary results showed no evidence of an average treatment effect on employment and productivity. In this framework, we aim at filling the gap with an EU-wide study of the impact of Cohesion Policy on the distribution of the firm's outcomes.

In the regional economics literature, there are several empirical analysis studying the impact of public subsidies on total factor productivity at the firm level. Bergström (2000) examined the effects on TFP of public capital subsidies to manufacturing firms in Sweden between 1987 and 1993. The study showed that subsidisation could impact on growth in the first year the support is granted, but after that TFP growth deteriorates. Harris and Trainor (2005) used detailed micro panel data distinguishing firms that receive assistance and those that did not for the manufacturing industry in North Ireland. They found that public subsidies to firms throughout the 1983-1998 period had a positive and significant impact on the level of production, and capital transfers are more likely to affect TFP than other forms of financial support positively.

Further, we also relate to the literature on the firm-level total factor productivity (TFP) dispersion as a measure of heterogeneity. Syverson (2011) surveyed recent empirical works and the common finding is that productivity differences among firms within an industry are large and robust to alternative estimation methods. Using data from the 1977 Census of Manufactures (CM), Syverson (2004b) found that establishments at the 90^{th} percentile of the within 4-digit-SIC productivity distribution are nearly twice as productive as those at the 10^{th} percentile. Also, Syverson (2004a,b) showed that the productivity variation across industries and geographic areas is persistent and it can be related to indicators of product substitutability, market structure, and competition. Hsieh and Klenow (2009) showed that under certain assumptions about technology and de-

mand, dispersion in revenue productivity reflects market distortions. In addition, using micro-data on manufacturing, they quantify the potential extent of dispersion as an indicator of misallocation in China and India versus the United States. Recently, Foster et al. (2016) explored the current interpretations of firm-level dispersion in revenue-based productivity measures. Their empirical evidence suggests, under iso-elastic demand, that dispersion may indicate either distortions or variation in demand shocks and/or technical efficiency.

The kinds of literature that we briefly reviewed above are lacking an appropriate assessment of regional policies of the EU on the performance of firms across the whole Single Market. Therefore, the contribution of this paper is twofold. First, we contribute to the existing literature by estimating the firm-level TFP in a EU-wide approach encompassing 28 countries using recent methodological techniques. Second, identifying the location of firms within NUTS-2 regions, we assess the impact of European Regional Development Funds (ERDF) tools on the growth of firms' TFP during the period 2007-2015. The results of the study may provide insightful information and implications for policy-makers at the EU-level and in general across the advanced economies.

3.3 Data

3.3.1 Firm-level Data

In the recent decades, the interconnected and complex global economy has called for an in-depth analysis of micro-agents which from the bottom shape the macro dynamics (Mayer and Ottaviano, 2008). Firm-level data have become a valuable tool for structural analysis and empirical evidence on several issues: assessing and comparing the productivity at different levels of aggregation, investigating innovation and entrepreneurship, understanding the effects of globalisation, moreover, linking the financial and employment decisions of firms to aggregate economic outcomes, among others.

Firm-level data are usually sourced from national and/or local public

agencies (e.g., business registers, production survey, tax returns), but their public access is often restricted mostly because of the risk of disclosing confidential information. As an alternative, commercial databases gather firm-level information about firms located worldwide, when confidentiality is handled. For the purpose of our analysis, the reliability of firm-level balance sheet data is related to the coverage and the quality of information for each firm. We use the Orbis database by Bureau Van Dijk (BvDEP), which contains financial and ownership information on millions of mostly private companies around the world, organised in a standard format after integration and harmonisation.[6]

Because of its broad coverage regarding statistical units[7] and time, and international comparability, it is possible to investigate firm's behaviour by industry, size, country and region over time. Coverage of small firms and balance sheet variables changes from country to country according to the filing requirements by business registers in each country (Kalemli-Ozcan et al., 2015), contributing to measurement errors, classification biases, selection biases, etc..[8] Hence, there is a trade-off between coverage (i.e. the number of firms, variables and countries) and the accuracy of the conducted analysis. For instance, the contemporaneous presence of some balance sheet information, namely turnover, material cost, fixed assets and employees, necessary to compute the total factor productivity reduces the available initial sample significantly.[9]

[6]Orbis financial data have become a common source for productivity estimates. See for example, Gopinath et al. (2017) and Gal (2013).

[7]Some shortcomings may arise when defining the unit of analysis: firms that operate in more than one country have at least one unit counted in each country; only in some countries the business register keeps track of organisational changes (i.e. mergers and acquisitions) within and between firms, and also the definition of legal units may vary across countries.

[8]It is well-known that limited liability companies, although they are required to register their formation, may not report complete balance sheet information in compliance with the national law which differs across countries. Moreover, official business surveys have administrative thresholds (e.g. VAT), below which some businesses are excluded. Therefore, concerns about possible sample selection by country and/or by size must be carefully addressed. Besides, a wider coverage of countries implies the lack of certain variables, i.e. value-added and intermediate inputs which are necessary to measure TFP. Another weakness is the availability of employment information, which is not a mandatory item in balance sheets but rather reported in a memorandum.

[9]Many researchers have experienced a large number of unique firm identifiers, but

As a starting point, we briefly present the coverage and quality of the sample we ended up to compute total factor productivity estimations. Table 3.1 provides information about European firms operating in manufacturing (sectors 10-33 in second revision *Nomenclature générale des activités économiques dans les Communautés européenes*, NACE), for which complete data are available over the period 2007-2015. We show the coverage on the population of firms provided by the Structural Business Statistics of Eurostat for each country and NACE rev.2 2-digit sector, as total economy percentages for the year 2013.[10]

A noteworthy feature of the data is a high coverage of turnover for some countries, although the percentage of operating firms in the year 2013 is much lower. For instance, for Bulgaria, since the percentage of turnover and labour is very good, but the number of firms is lower than the one reported by Eurostat, we can suggest that Orbis has information on large firms. For Austria and some other countries, the coverage of labour is lower than the one of turnover. One reason could be that firms in Orbis are less labour-intensive than the ones in Eurostat. First evidence suggests that for most countries there is an over-representation of medium and large firms in our sample, compared to the majority of small and medium-sized enterprises (SMEs) within the European Union.[11] The latter is confirmed in Table 3.2 where the size distribution regarding the number of firms for each country is reported. For countries like Austria, Belgium, Denmark, Netherlands and the United Kingdom, the highest number of firms in Orbis is of medium size. However, the sample selection on the size is less severe, in terms of representativeness of the sample, when looking

many missing values for financial values, as reported in Kalemli-Ozcan et al. (2015). It appears that there is a reporting lag of about two years on average, so for instance information about a firm in 2010 may fully appear in 2012.

[10]Structural business statistics (SBS) data is collected using statistical surveys, business registers or from various administrative sources across the European Union (EU). Starting in 1995, the SBS provides information on many key variables, such as turnover, value-added, employment, the number of business units, etc., broken down by industries and size for each country. Notice that changes in the specific purposes (e.g., tax collection, government policies, etc.) of the administrative sources may affect the coverage, definition, thresholds, etc., of the data. The sample coverage for the other years in the study is available upon request.

[11]The over-representation for the largest firms in the Orbis database is well understood. See Ribeiro et al. (2010) and Kalemli-Ozcan et al. (2015) for more details.

at the percentage of firms within each sector. Table 3.3 shows that there is a good sector coverage for most countries. Some exceptions are less labour-intensive sectors, such as NACE 13-15 and NACE 16-18, which are overall under-represented.[12]

In following analyses, we will make our results robust to a possible sample selection bias arising from a censoring by firm size.

Table 3.1: Sample coverage, year 2013.

	Turnover			N. of employees			N. of firms		
Country	Sample	Eurostat	%	Sample	Eurostat	%	Sample	Eurostat	%
Austria	77,724	176,744	43.98	215,039	617,441	34.83	760	25,129	3.02
Belgium	224,957	267,274	84.17	297,804	514,258	57.91	2,732	33,468	8.16
Bulgaria	21,742	22,566	96.35	503,980	524,041	96.17	15,607	30,091	51.87
Croatia	13,793	16,044	85.97	181,939	260,534	69.83	7,724	20,673	37.36
Cyprus	693	2,585	26.80	4,312	25,583	16.85	95	4,947	1.92
Czech Rep.	137,125	139,840	98.06	989,866	1,160,215	85.32	14,702	167,688	8.77
Denmark	35,178	93,000	37.83	84,615	351,178	24.09	573	15,062	3.80
Estonia	583	11,142	5.24	80,228	104,565	76.73	3,871	6,381	60.66
Finland	58,559	113,213	51.72	182,904	330,472	55.35	6,369	21,581	29.51
France	471,945	870,241	54.23	1,278,008	2,993,901	42.69	27,953	226,369	12.35
Germany	689,905	1,975,826	34.92	1,830,210	7,220,296	25.35	13,134	202,823	6.48
Greece	23,131	56,478	40.96	134,025	289,187	46.35	4,417	57,736	7.65
Hungary	84,421	93,802	90.00	447,036	664,724	67.25	3,003	47,475	6.33
Italy	661,432	872,479	75.81	2,384,401	3,733,694	63.86	105,399	407,344	25.87
Latvia	5,945	7,517	79.09	98,116	120,761	81.25	5,525	9,535	57.94
Lithuania	7,301	13,508	54.05	90,917	195,701	46.46	1,286	16,120	7.98
Luxembourg	2,480	4,161	59.58	9,137	19,008	48.07	99	733	13.51
Malta	158	383	41.25	2,351	7,786	30.20	52	1,279	4.07
Netherlands	40,741	308,574	13.20	54,109	681,617	7.94	519	60,506	0.86
Poland	51,226	270,727	18.92	350,377	2,347,504	14.93	3,029	174,414	1.74
Portugal	73,470	79,429	92.50	536,092	637,427	84.10	26,797	66,423	40.34
Romania	61,854	65,677	94.18	1,053,333	1,166,313	90.31	30,125	46,761	64.42
Slovenia	18,745	23,848	78.60	124,064	176,175	70.42	6,285	18,148	34.63
Slovakia	58,357	61,104	95.50	357,157	437,796	81.58	6,863	63,185	10.86
Spain	322,576	447,415	72.10	1,115,379	1,736,651	64.23	56,018	168,935	33.16
Sweden	152,804	197,809	77.25	407,280	631,140	64.53	12,633	53,681	23.53
UK	399,185	576,651	69.22	1,278,672	2,368,775	53.98	9,524	127,943	7.44

Notes: This table reports variables aggregated at country-level when information by industry is available in the Orbis and the Eurostat datasets. Turnover is expressed in millions of euros. Firms with consolidated accounts are excluded when considering coverage on turnover and number of employees. Data on Ireland are not available in Eurostat database.

[12]Notice that the primary activity code attributed to each firm may differ in Orbis and Eurostat. While in the latter the criterion of attributing the activity is based on the initial classification of the firm at the time of its set up, Orbis classification is based on the current production of the firm, therefore a more accurate information.

Table 3.2: Size distribution number of firms in the sample and in Eurostat (%), year 2013

	TFP sample					Eurostat				
Country	**0-9**	**10-19**	**20-49**	**50-249**	**250+**	**0-9**	**10-19**	**20-49**	**50-249**	**250+**
Austria	3.3	1.4	6.8	48.1	40.3	72.1	11.6	8.7	5.7	1.9
Belgium	10.7	11.2	30.8	36.0	11.3	81.7	7.6	6.1	3.6	1.0
Bulgaria	50.9	18.4	17.4	11.4	1.9	74.8	9.8	8.6	5.8	1.0
Croatia	66.0	14.4	10.2	7.5	1.8	83.7	7.6	4.8	3.1	0.8
Cyprus	17.5	12.4	39.2	25.8	5.2	90.0	5.9	2.7	1.2	0.1
Czech Rep.	41.0	17.0	18.4	18.1	5.5	92.7	2.7	2.3	1.8	0.5
Denmark	7.5	4.7	10.8	43.1	33.9	71.2	12.4	9.0	6.2	1.2
Estonia	65.9	13.0	11.8	8.1	1.1	74.7	9.3	8.7	6.4	0.9
Finland	55.0	15.9	14.4	11.0	3.7	83.8	6.5	5.3	3.5	0.9
France	58.6	14.4	13.5	10.1	3.4	86.8	5.6	4.4	2.6	0.7
Germany	13.3	13.5	20.3	34.8	18.0	61.4	20.4	8.1	8.1	2.1
Greece	35.0	25.4	23.7	12.9	3.0	95.1	2.0	1.6	1.0	0.2
Hungary	17.8	8.6	19.9	40.2	13.5	84.5	6.5	4.8	3.4	0.8
Ireland	17.8	7.9	17.5	36.7	20.0	n/a	n/a	n/a	n/a	n/a
Italy	52.2	23.8	15.4	7.3	1.4	83.0	9.9	4.8	2.1	0.3
Latvia	66.4	13.3	11.3	8.0	1.0	79.9	8.0	6.6	4.9	0.6
Lithuania	12.4	16.5	28.2	34.4	8.5	82.1	7.0	5.9	4.3	0.8
Luxembourg	3.0	3.0	11.0	43.0	40.0	62.6	12.2	12.6	9.7	3.0
Malta	14.8	6.6	19.7	42.6	16.4	91.1	6.2	0.0	2.7	0.0
Netherlands	7.1	3.5	11.4	48.7	29.3	85.8	5.8	4.5	3.3	0.6
Poland	28.4	2.3	27.2	29.1	12.9	86.9	4.6	4.2	3.5	0.9
Portugal	61.3	18.0	13.0	6.8	0.9	82.6	8.2	5.8	3.0	0.4
Romania	58.8	15.4	13.9	9.5	2.3	70.6	11.4	9.7	6.7	1.6
Slovenia	74.7	10.5	7.1	6.1	1.5	88.5	5.0	3.2	2.7	0.6
Slovakia	52.2	16.8	14.2	12.9	3.8	94.0	2.2	1.8	1.5	0.4
Spain	63.7	16.8	12.4	5.8	1.2	83.6	8.1	5.5	2.4	0.4
Sweden	57.3	18.1	13.3	8.7	2.7	87.8	5.3	3.9	2.4	0.6
UK	4.9	4.7	15.4	57.7	17.3	76.4	10.3	7.4	4.9	1.1

Notes: This table presents the distribution of firms according to their size in the TFP sample and as reported in Eurostat database. It shows the percentages computed on the total number of firms in each country. Each row sums up to 100% for the two samples. Data on size distribution of firms about Ireland are not available in Eurostat database.

Table 3.3: Sector distribution number of firms in the sample and in Eurostat (%), year 2013

	TFP sample								Eurostat							
Country	**10-12**	**13-15**	**16-18**	**19-20,22-23**	**21**	**24-25**	**29-30**	**26-28,31-33**	**10-12**	**13-15**	**16-18**	**19-20,22-23**	**21**	**24-25**	**29-30**	**26-28,31-33**
Austria	12.2	3.5	11.2	17.2	2.6	17.9	5.1	30.2	15.2	6.0	15.0	9.1	0.3	15.7	1.6	37.1
Belgium	18.2	6.0	10.0	21.5	1.8	16.7	2.7	23.1	20.9	5.7	17.3	8.5	0.3	21.6	1.7	24.0
Bulgaria	18.0	14.9	12.8	14.5	0.3	14.0	0.8	24.8	19.4	18.2	11.7	12.1	0.2	12.6	0.6	25.1
Croatia	16.9	7.2	16.1	14.4	0.3	17.3	2.7	25.2	15.7	8.7	17.5	13.5	0.2	17.0	2.1	25.4
Cyprus	30.9	2.1	12.4	21.6	4.1	9.3	2.1	17.5	17.2	6.2	23.9	9.0	0.2	23.6	1.6	18.4
Czech Rep.	8.9	4.7	11.5	13.6	0.4	23.9	3.5	33.4	5.0	8.5	22.2	7.0	0.1	26.0	1.0	30.2
Denmark	20.1	3.0	8.0	13.4	3.0	11.2	2.6	38.7	10.5	4.9	9.9	8.4	0.6	19.9	1.8	44.0
Estonia	9.1	11.8	21.4	7.9	0.1	19.0	2.0	28.6	7.7	12.1	23.3	8.5	0.2	18.4	2.0	27.9
Finland	8.3	3.8	14.1	11.0	0.3	26.8	3.2	32.5	8.1	9.2	15.4	7.6	0.1	22.6	2.9	34.1
France	28.7	4.7	10.4	9.9	0.6	16.1	2.5	27.0	26.8	8.4	14.6	7.2	0.2	9.9	1.3	31.8
Germany	7.4	3.2	9.0	15.2	1.4	23.2	3.9	36.8	14.6	3.7	12.4	10.1	0.3	22.3	1.8	34.7
Greece	28.6	9.4	9.4	20.4	1.7	12.3	1.4	16.8	26.7	15.5	10.8	8.9	0.2	14.7	0.7	22.5
Hungary	19.3	4.6	8.2	17.1	1.2	19.3	5.5	24.7	14.0	9.1	15.0	9.4	0.2	17.2	1.5	33.8
Ireland	9.0	3.8	8.5	14.5	11.0	9.0	2.7	41.4	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Italy	9.3	12.5	8.6	13.0	0.4	22.7	2.6	30.9	14.2	15.0	12.6	8.8	0.1	17.0	1.2	31.2
Latvia	11.3	10.5	28.3	10.8	0.4	12.5	1.3	24.7	10.2	16.9	25.9	9.1	0.3	10.7	1.2	25.7
Lithuania	15.5	9.9	17.3	14.5	0.2	12.5	2.0	28.0	9.2	20.8	23.0	11.1	0.1	9.2	0.6	26.0
Luxembourg	12.0	3.0	8.0	24.0	1.0	22.0	3.0	27.0	19.4	4.2	14.5	9.5	0.1	24.1	1.5	26.6
Malta	23.0	4.9	6.6	18.0	8.2	4.9	8.2	26.2	19.5	1.8	10.2	11.0	0.0	17.9	0.6	39.1
Netherlands	16.6	3.7	7.1	22.2	4.4	9.6	4.2	32.2	9.3	7.9	11.7	7.2	0.3	18.6	3.7	41.3
Poland	16.1	5.7	10.9	19.2	1.1	18.2	3.6	25.1	7.3	10.6	15.6	11.0	0.2	17.2	1.3	36.8
Portugal	16.7	21.3	12.5	10.7	0.3	18.2	1.5	18.7	16.0	22.5	13.1	8.8	0.2	18.1	1.4	19.9
Romania	19.4	14.7	16.4	12.9	0.3	13.6	2.0	20.7	18.3	15.4	17.1	12.9	0.3	12.8	1.8	21.3
Slovakia	10.5	6.6	14.6	14.0	0.3	24.3	2.7	27.1	4.4	8.7	20.4	6.9	0.0	38.6	0.5	20.5
Slovenia	7.5	5.6	17.1	12.6	0.2	26.3	1.9	28.9	10.7	6.9	18.2	10.0	0.1	23.6	1.5	29.1
Spain	17.4	8.5	13.9	13.6	0.4	20.4	2.2	23.7	16.1	11.3	15.5	9.6	0.2	21.5	1.4	24.4
Sweden	9.5	2.8	14.4	9.8	0.6	27.5	4.2	31.3	7.2	8.6	16.7	8.3	0.3	21.1	3.7	34.3
UK	11.6	4.2	8.2	14.7	2.2	15.1	4.6	39.4	6.4	6.1	17.1	9.4	0.4	20.2	3.6	36.8

Notes: This table presents the distribution of firms according to the main sector they belong to, in the TFP sample and as reported in Eurostat database. It shows the percentages computed on the total number of firms in each country for classes of sectors at 2-digit NACE rev. 2 (see Table ?? in the Tables and Graphs Appendix ?? for the corresponding sector name). Each row sums up to 100% for the two samples. Data on sector distribution of firms about Ireland are not available in Eurostat database.

3.3.2 Total Factor Productivity Estimation: Preliminary Evidence

Exploiting a unique dataset of 542,876 manufacturing firms from EU-28 countries, for the period 2007-2015, we estimate firm-level total factor productivities (TFP) from an industry-level production function.[13] We apply TFPs using the semi-parametric technique proposed by Ackerberg et al. (2015) (ACF hereafter). See Appendix A for further details.

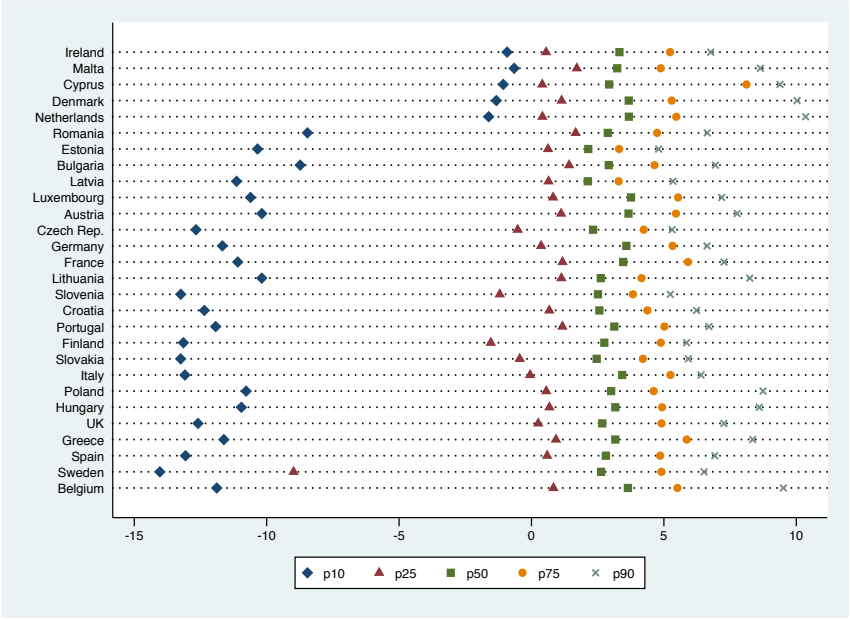
We calculate factor elasticities for each industry on a EU-wide scale to assess the competitiveness of firms horizontally across national borders. In the integrated framework of a Single Market, characterised by increasing economic integration, competitive pressure is usually thought to have a diverse impact on productivity as it is referred in the literature of economic geography (Beaudry and Schiffauerova, 2009). For instance, for the Italian firm "PARMALAT s.p.a." the 2013 productivity computed using elasticities at the country-industry level is quite close to the mean of its industry, while using EU-wide estimations it is among the top productive firms within industry NACE 10. This indicates that the Italian manufacturers of food with very high productivity are in general more competitive than other manufacturers in the Wider Europe. On the contrary, the English firm "LUSH LTD.", operating in industry NACE 20, in 2013 is more productive than the average of its peers using country-industry level elasticities, while using EU-wide industry elasticities it is below the European average. This suggests that while this firm is a very competitive producer of chemicals within England, it is not very highly productive in the Wider Europe beside other competitors in the industry. In this context, empirical evidence suggests that looking at firms across all countries provides more insightful information on the position of a country industry's competitiveness (Altomonte et al., 2010), especially within a Single Market such as the European Union.

Figure 3.1 shows the percentiles of the firm-level log TFP by country. Heterogeneity in firm performance is widely spread across countries in the

[13]Please note that TFP is estimated for all firms in our sample, while the match with following regional data is possible only for the regions that benefitted from ERDF funding.

sample, and it is linked to some extent to the heterogeneity in the size of firms observed within industries (Bartelsman et al., 2013). Nevertheless, there are some empirical regularities. For instance, the median value of productivity is quite similar across countries, and the distance between the 25^{th} and the 10^{th} percentile is in most cases high, meaning that there are few firms in each country with productivity close to zero which is in line with the under-representation of micro firms in our sample. On the other hand, if we look at the 90^{th} percentile only Denmark and Netherlands exhibit high firm-level productivity.

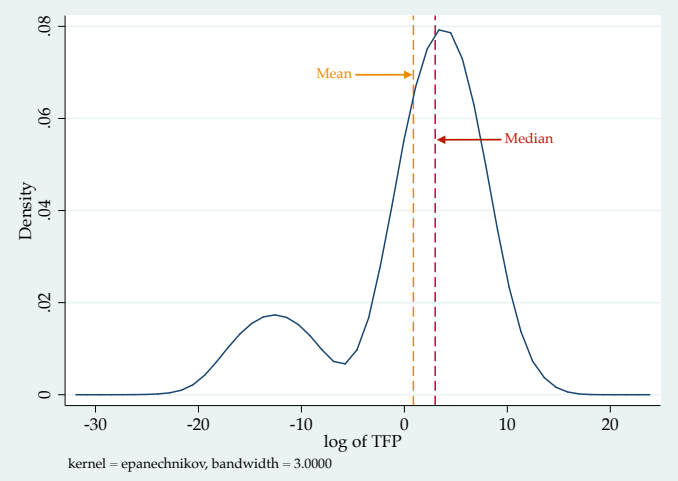
Figure 3.1: Percentiles (log of) TFP by country



Notes: This Figure shows the firm-level total factor productivity percentiles for each country in the sample. Countries are presented in ascending order according to the difference between the 90^{th} and the 10^{th} percentile.
Source: Own elaboration.

Firms are heterogeneous along many dimensions and their distributions have power-law right tails (Gabaix, 2009, 2016; Luttmer, 2010). The

Figure 3.2: Total factor productivity distribution of EU manufacturing firms



Source: Own elaboration.

European case is slightly more sophisticated. In Figure 3.2, we record how the (log of) TFP of EU manufacturing firms shows up as an asymmetric and bimodal distribution. In other words, there are two different sets of firms. On the left side of Figure 3.2, there is a bunch of firms largely inefficient, yet active on the market and far from going bankrupt. These firms operate in the same markets together with more efficient firms, on the right-hand side of Figure 3.2. The origins of such a polarisation in EU-wide firm-level productivity is largely unexplained, and it is not the object of the present study.[14] However, in the context of our exercise, we note how the design of the EU Cohesion Policy does not consider the existence of such heterogeneity, and the average effect of an ERDF financial support may have unexpected and unintended consequences.

[14]See Andrews et al. (2016) for a possible explanation in terms of technological divergence.

3.3.3 Regional Policy Data

We retrieve data on regional policy funding from the report 'Geography of Expenditure - Work Package 13' prepared for the European Commission in 2015 (wiiw and ISMERI EUROPA, 2015). The report studies the cumulative allocations to selected projects and the expenditures of European Regional Development Funds (ERDF) and Cohesion Fund (CF) over the programming period 2007-2013 for all 28 EU countries. Most of the transfers are assigned at the NUTS-2 level.[15] Overall, the *Structural funds Programme* distinguishes transfers by objectives: Convergence (Objective 1), Regional competitiveness and employment (Objective 2), and European territorial cooperation (Objective 3). We restrict our analysis to the Objective 1 which aims at accelerating the economic development of lagging EU regions and it accounts for more than two-thirds of the programme's total budget. The objective covers regions whose GDP per capita in PPP is less than 75% of the EU average. In this work, we focus on the ERDF established in 1975, in particular on two thematic areas which may have a direct impact on firm's productivity: *Business Support*, and *Research, Technology and Development* (RTD).

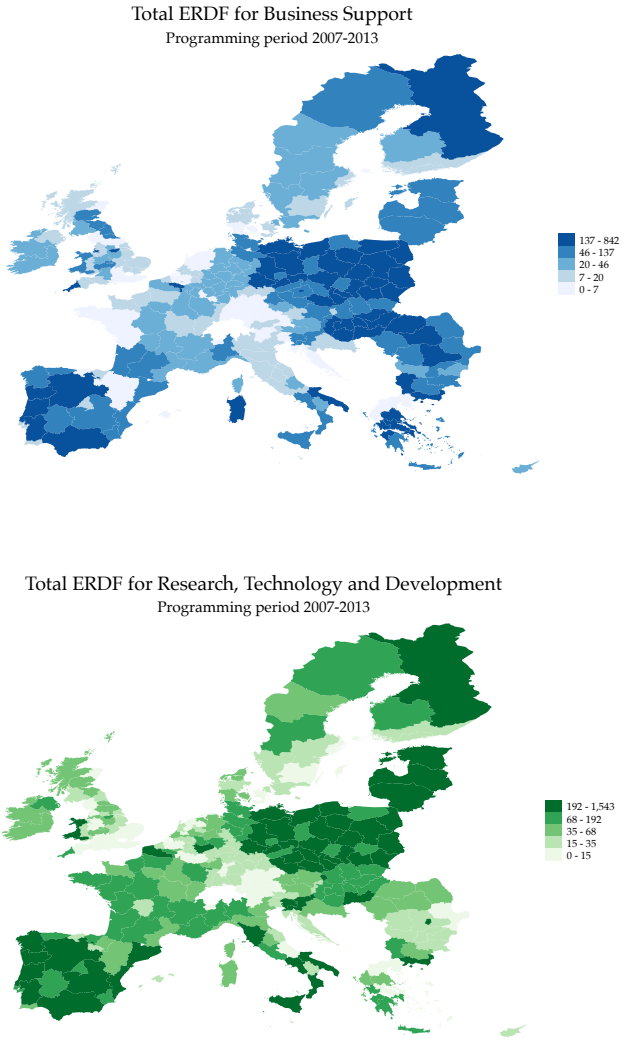
Figure 3.3 provides insights on the distribution of payments across NUTS-2 regions for the two priorities mentioned above. The total value of projects subsidised over the whole programming period 2007-2013 by the ERDF *Business Support* summed up to roughly EUR 21 billion, compared to EUR 35 billion from the ERDF RTD. In a regional approach, the amount of transfers for *Business Support* varied from EUR 53,987 in Schwaben (Germany) to over EUR 842 million in Andalusia (Spain), with an average of about EUR 82 million per region. Financial aid for research, technology and development ranges from EUR 295,576 in South East England to more than EUR 1.5 billion in the Warsaw region (Poland). On average, every region received EUR 132 million for projects involved in innovation and

[15]NUTS is the acronym for *Nomenclature des Unités Territoriales Statistiques* and it is a hierarchical system to for dividing the economic territory of the EU. The highest level of aggregation (NUTS-1) corresponds to major socio-economic regions (e.g., the United Kingdom's regions of England/Scotland/Wales); NUTS-2 refers to basic regions for the application of regional policies (e.g., Italian regions); and NUTS-3 identifies small regions for specific diagnosis (e.g., *Départements* in France).

development activity. In the populous and usually rich regions such as those in England, Belgium, the Netherlands, parts of Germany, Austria and Northern Italy, ERDF payments usually do not exceed 0.1% of the regional GDP over the entire period of financing. The regions with the highest ERDF payments as a share of regional GDP are Észak-Magyarország in Hungary with 0.54% (ERDF *Business Support*) and Alentejo in Portugal with 0.47% (ERDF RTD), respectively.

We match firm-level financial accounts with the regional policy data on ERDF at the NUTS-2 regions, based on the postal address of each firm. It is important to note that while the programme's commitment is in the period 2007-2013, the allocation of funds usually takes 1-2 years longer ending in 2015 for some regions. Moreover, there are some overlaps of the allocated funds from the previous programme in the period 2000-2006. Here we mainly focus on the funds allocated in the program 2007-2013, although our findings are robust to the inclusion of overlaps of the programs.

Figure 3.3: Payments by NUTS-2 region from the European Regional Development Fund



Note: Values in EUR million.
Source: 'Geography of Expenditure – Final Report, Work Package 13', wiiw and ISMERI EUROPA, 2015, own elaboration.

3.4 Empirical Strategy

Our contribution aims at investigating the short-term impact of the ERDF spending on the firm-level TFP growth in the period 2007-2015. First, we make our results robust to the Heckman (1979) correction, accounting for a possible sample selection by firm size, measured by firm sales, as pointed out in the previous section. On top of sample selection, we may want to control for the endogenous relationship between firm-level TFP growth and the regional provision of ERDF funds. In fact, it is quite possible that less (more) ERDF funding is allocated over time when the firms in a region record increasing (decreasing) rates of productivity. The empirical model by Semykina and Wooldridge (2010) suits our case, as it combines instrumental variables in the presence of possible sample selection with panel data.[16] Eventually, we replicate our specification along TFP quartiles, to verify whether a different impact can be found along the heterogeneous and bimodal distribution of inefficiencies, as represented in Figure 3.2.

3.4.1 Baseline Findings

Our outcome equation is:

$$\begin{aligned}\Delta(tfp_{ijrt}) = \alpha_i + \beta_1 BS_{r,t-1} + \beta_2 RTD_{r,t-1} + \beta_3 X_{ir,t-1} \\ + \beta_4 Z_{r,t-1} + \delta_t + \mu_{irt}\end{aligned} \tag{3.1}$$

where the dependent variable $\Delta(tfp_{ijrt})$ is the growth of total factor productivity of firm i operating in industry j located in region r at time t. On the right hand, we highlight our main variables of interest lagged one period: i) $BS_{r,t-1}$ is the ERDF funding for *Business Support* spent in region r and time $t-1$; ii) $RTD_{r,t-1}$ is the ERDF funding for *Research, Technology and Development* in region r at time $t-1$. Under $X_{ir,t-1}$ we collect firm-level controls, and specifically, we include the TFP level and the firm size, both at time $t-1$, since we expect that bigger and already more productive firms tend to grow relatively less in future

[16] For a previous use of the empirical model by Semykina and Wooldridge (2010), see among others Kee et al. (2008) and Essaji (2008).

periods. Then, we refine our specification with $Z_{r,t-1}$, which contains regional level controls to disentangle the impact of regional size, human capital, agglomeration, industrial specialization, industrial diversification and competition externalities on TFP. See Appendix C for more details on all variables and their construction. A set of time fixed effects further controls for overall time trends. We allow the specification to have a firm-level unobserved time-invariant component α_i that is possibly correlated with other regressors, hence testing firm-level fixed effects.

The dependent variable $\Delta(tfp_{ijrt})$ cannot always be observed, and we may have a sample selection problem. As discussed with reference to Table 3.1, our TFP sample underrepresents smaller firms due to the different regulations of entries into national registries. This is a case of unintentional censoring, due to measurement errors. Yet, smaller firms are more likely to benefit from ERDF vehicles, as it is clearly in the original design of the policy. Therefore, we introduce a sample selection process in the form:

$$d_{it} = \theta_0 + \theta_1 size_{it} + s_i + \varepsilon_{it} \tag{3.2}$$

where d_{it} is a binary variable equal to 1 when the TFP growth rate has been computed for company i at time t, and 0 otherwise. We assume that the selection process operates on the firm size, as we know that national registries gathered by our database ask for detailed financial accounts only after some thresholds of activity (see also Kalemli-Ozcan et al. (2015)).[17]

Further, we know that the relationship between TFP growth rates and both the ERDF funding vehicles may be endogenous, because we cannot rule out that decreasing (increasing) firms' competitiveness do have a reversal impact on an increasing (decreasing) allocation of either funds. In fact, it is quite possible that when firms in regions at disadvantage start to reap the benefits from the investments entailed by the ERDF funding, the number of applications and the amount of funds allocated in that region could decrease over time. Therefore, we need a set of instruments, W_{rt}, such that the contemporaneous exogeneity assumption holds, $E(\mu_{irt}|W_{rt}, s_i, \alpha_i) = 0$, conditional on the firm-level unobserved

[17]See Table 3.2 in Section 3.3 for a detail of the sample selection by size at the country level, comparing with census from Eurostat.

selection and heterogeneity components. For this purpose, we instrument the ERDF BS and RTD regional spending in the specification of eq. 3.1 by considering both the BS and RTD spending in the period before in neighbouring regions, i.e. in the regions sharing a common border with the target region r. We can comfortably assume that neighbouring regions have both a BS and RTD funding that is correlated with funding in region r, since they share a similar geographic disadvantage at the periphery of the European Union. On the other hand, we have no reason to believe that the funding in neighbouring regions should have a direct impact on the productivity of firms that are active in the r^{th} region. Please note that by using the technique by Semykina and Wooldridge (2010), we allow firm-level unobserved and time-invariant heterogeneity, α_i and s_i, to be correlated across both the sample selection process and the set of instrumental variables. This is particularly useful in our case, as both the absence of smaller firms in our sample and their exposure to ERDF funding can be explained by some unobserved characteristics. Finally, to overcome a possible aggregation bias resulting from an empirical specification that mixes a firm-level outcome with region-level ERDF spending, we adopt a two-way clustering of the standard errors by firm and by region (Cameron et al., 2011).

In Table 3.4, we report estimates in a sequence. In column 1, we report a simple firms' panel fixed effects specification, and in column 2 we challenge it with a Heckman (1979) correction. The complete empirical model following Semykina and Wooldridge (2010) is augmented with instrumental variables for ERDF funding and reported in column 3 of Table 3.4.

Crucially, we find that the *Business Support* vehicle has a non-significant association with the short-term productivity growth of the firms in our sample, after controlling for endogeneity. We argue that negative coefficients observed in columns 1 and 2 for the BS vehicle are indeed explained by the endogenous trends observed from our data: a generalized decrease in ERDF funding is observed where the TFP levels have increased more, for example in the New Members of the European Union.

On the other hand, a robust impact of the RTD vehicle on TFP growth

Table 3.4: Baseline estimates

Dependent variable: TFP growth	Firm FE	Selection & Firm FE	Endogeneity & Selection & Firm FE
ERDF Business Support	-0.0018***	-0.0017***	0.0020
	(0.0004)	(0.0004)	(0.0018)
ERDF RTD	0.0011***	0.0009**	0.0058***
	(0.0005)	(0.0004)	(0.0014)
Regional GDP	0.2339***	0.3046***	0.3098***
	(0.0173)	(0.0176)	(0.0176)
% Aged 25-64 (Education levels 3-4)	0.0062***	0.0057***	0.0045***
	(0.0005)	(0.0005)	(0.0005)
% Aged 25-64 (Education levels 5-8)	0.0050***	0.0052***	0.0050***
	(0.0006)	(0.0007)	0.0007)
Agglomeration	-0.0192***	-0.0153***	-0.0154***
	(0.0716)	(0.0038)	(0.0040)
Specialization	0.2810**	0.4543***	0.4989***
	(0.1165)	(0.1178)	(0.1183)
Diversification	-0.1882**	-0.2110***	-0.2183***
	(0.0816)	(0.0720)	(0.0722)
Competition	0.2344***	0.2287***	0.2313***
	(0.0707)	(0.0690)	(0.0699)
tfp (t-1)	-0.7822***	-0.7833***	-0.7834***
	(0.0044)	(0.0044)	(0.0044)
Firm Size	0.0441***	-1.7175***	-1.7131***
	(0.0054)	(0.0721)	(0.0723)
N. observations	1,506,240	1,499,673	1,497,580
NUTS 2 regions	273	273	273
Heckman correction	No	Yes	Yes
Instrumental Variables	No	No	Yes
R-squared (within)	0.406	0.410	0.410
Lambda (Heckman)	-	-71.73***	-71.73***
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Two-way clustered errors	Yes	Yes	Yes

Note: Two-way clustered standard errors in parenthesis ***p<0.01, **p<0.05, *p<0.1.

rates is reported across all specifications in Table 3.4. The coefficient is relatively higher after we control for possible endogeneity, in column 3.

3.4.2 Robustness and Sensitivity Checks

We explore the robustness of our main findings along the quartiles of the heterogeneous TFP distribution observed in Figure 3.2. We use the same empirical specification from the baseline eq. 3.1 and eq. 3.2, after considering instrumental variables in a panel dataset with sample selection and endogeneity, as proposed by Semykina and Wooldridge (2010). Results are reported in Table 3.5.

Table 3.5: Robustness of findings across the quantiles of the TFP distribution

Dependent variable: TFP growth	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**4th Quartile**
ERDF Business Support	0.0051	0.0005	-0.0013	-0.0027
	(0.0030)	(0.0007)	(0.0008)	(0.0016)
ERDF RTD	0.0157***	0.0047***	0.0039***	0.0007
	(0.0047)	(0.0010)	(0.0009)	(0.0014)
Regional GDP	1.6786***	-0.1592***	-0.1431***	-0.0878***
	(0.0555)	(0.0114)	(0.0124)	(0.0172)
% Aged 25-64 (Education levels 3-4)	0.0071***	0.0012***	0.0032***	0.0035***
	(0.0019)	(0.0004)	(0.0003)	(0.0005)
% Aged 25-64 (Education levels 5-8)	-0.0011	0.0034***	0.0040***	0.0007
	(0.0023)	(0.0005)	(0.0004)	(0.0006)
Agglomeration	0.0001	-0.0068*	-0.0066*	-0.0085**
	(0.0057)	(0.0035)	(0.0027)	(0.0033)
Specialization	5.0035***	0.9135***	-1.0708***	-0.5882***
	(0.4102)	(0.1580)	(0.1287)	(0.0732)
Diversification	-2.1578***	-0.2701***	0.3381***	0.2414***
	(0.2658)	(0.0681)	(0.0571)	(0.0586)
Competition	1.2201***	-0.8665***	-0.0059	-0.0074
	(0.1024)	(0.0158)	(0.0203)	(0.0116)
tfp (t-1)	-0.8215***	-0.9256***	-0.9126***	-0.8001***
	(0.0045)	(0.0032)	(0.0031)	(0.0052)
Firm Size	6.1880***	3.9034***	1.9665***	2.0675***
	(0.7385)	(0.1014)	(0.1204)	(0.2652)
N. observations	381,964	374,863	379,502	367,998
NUTS 2 regions	273	273	273	113
Heckman correction	Yes	Yes	Yes	Yes
Instrumental Variables	Yes	Yes	Yes	Yes
R-squared (within)	0.4777	0.6642	0.6707	0.5523
Lambda	-168.01***	-94.60***	-44.84***	-40.84***
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Two-way clustered errors	Yes	Yes	Yes	Yes

Note: Two-way clustered standard errors in parenthesis ***p<0.01, **p<0.05, *p<0.1.

The ERDF *Business Support* vehicle has no significant statistical association with TFP growth rates along the TFP distribution. Instead, we record the strongest impact of the ERDF RTD funding for the firms that had the lowest TFP levels in our period of analysis, in the first quartile of the distribution. In fact, we find the strongest impact on the left tail of Figure 3.2, around the first modal value along the asymmetric bimodal distribution. Briefly, a regional increase in the RTD funding is associated with a relatively higher growth rate of the TFP for firms that are less productive, as it was the priority of the policy maker.

Nonetheless, results from Table 3.5 show also that more productive firms may benefit from the EU Cohesion Policy, although the magnitude of the coefficients is three times lower. We argue that firms in the second and third quartiles of the TFP distribution can become more productive: i) directly, because they apply and receive an RTD funding; ii) indirectly, because they benefit from technological spillovers spreading within the region from the original applicants. For example, it is quite possible that after an investment is made by a company under RTD funding, other firms in local markets start to imitate the innovations in products and processes. Interestingly, the positive association of RTD funding with TFP growth rates becomes smaller and smaller in magnitude at increasing levels of TFP, until the correlation becomes statistically non-significant in the fourth quartile, i.e. on the right tail previously observed in Figure 3.2.

Finally, we perform a sensitivity check for our findings taking alternative outcomes: the firm-level value added and the labour productivity. Interestingly, neither ERDF measure is significantly associated with firm-level value added or labour productivity. Indeed, in the first case, we did not expect an impact of the Cohesion Policy on yet another proxy of firm size. In the second case, labour productivity is a less reliable proxy of firms' competitiveness, because it is subject to changes in the combinations of production factors. Among others, we would expect that following investments under RTD imply a relatively higher firm-level capital intensity, hence an endogenous change in the marginal product of labour.

3.5 Conclusions

In this contribution, we tested the impact of the EU Cohesion Policy on firm-level TFP growth, as the latter is a useful indicator of firm-level competitiveness. Specifically, we considered the funding by the ERDF originally separated as two measures, one for *Business Support* and one for *Research, Technology and Development*.

In general, the tendency is to refer to aggregate statistics at the country, region or industry level to evaluate the impact of a regional or industrial policy. The underlying reference is to a representative economic agent (e.g., a firm, or an individual), on which an average impact can be estimated from the aggregate. Thanks to an increasing availability of micro-data at the firm-level, it is possible to check whether the impact is heterogeneous within the aggregated country, region and industry.

In fact, after we estimate total factor productivity (TFP) for firms in the European Union, we already find that inefficiencies are distributed asymmetrically, bimodal and with a long right tail. Therefore, heterogeneous responses may be predicted from the application of the EU Cohesion Policy. We do find that the *Business Support* does not have any statistical association with the evolution of firms' competitiveness in our period of analysis controlling for endogeneity bias. On the other hand, a more targeted support for innovation in products and production processes, under the RTD funding, is significantly associated with an increase in TFP. The impact is relatively higher for the least productive firms, in the first quartile of the TFP distribution, where a support seems to be most needed.

We believe that our findings are important in the early hours of the negotiations for the next budget of the EU Cohesion Policy, which calls for a recalibration of the funding and a revision of the governance. We support the idea that a revision of the EU Cohesion Policy should also consider the heterogeneous distribution of inefficiencies *within* regions, which are often as relevant as differences *across* regions.

Appendices to Chapter 3

A Appendix: Total Factor Productivity (TFP)

A.1 Estimates of Total Factor Productivity

Seminal contributions of Tinbergen (1942) and Solow (1957) tie the aggregate production function to productivity, defined as the proportion of output not explained by some inputs used in production. As an indicator of efficiency, the total factor productivity (TFP) tries to quantify the difference in firms' output obtained using the same level of capital, labour, intermediate goods, etc.. With a single homogeneous output and a single homogeneous input, productivity is a cardinal number, hence units of output per unit of input (Bartelsman and Wolf, 2017). On the other hand, when multiple inputs are used, alternatively, when inputs and outputs are not strictly homogeneous across firms or over time, many econometric issues arise in the estimation of factor elasticities.

We start reviewing the literature on estimations of production functions with some preliminaries, so then we move on several approaches addressing the main methodological issues.

Assuming a production function is in the Cobb-Douglas form:

$$Y_{ijct} = e^{\beta_{j0}} K_{ijct}^{\beta_k} L_{ijct}^{\beta_l} M_{ijct}^{\beta_m} e^{\varepsilon_{ijct}} \tag{A.3.1}$$

where Y, K, L, and M are respectively units of output, capital, labour and materials, observable for a firm i operating in industry j of country c at time t. ε_{ijct} captures unobservables to the researcher which may affect the output (e.g., the weather, the management quality).

Taking the log of equation (A.3.1), the empirical problem reduces to the estimation of the following equation:

$$y_{ijct} = \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} + \beta_m m_{ijct} + \omega_{ijct} + \epsilon_{ijct} \tag{A.3.2}$$

we can think about ε_{ijct} having two components, i.e. ω_{ijct} is predictable to the firm when it makes its inputs decision, but unobserved by the researcher, whereas the firm has no information about ϵ_{ijct}. In the literature ω_{ijct}

is called Hick's neutral productivity shock. ϵ_{ijct} may represent a non-predictable productivity shock or a measurement error in the output variable.

The major identification problem in estimating the production function is that input choices made by the firms depend on ω_{ijct} which affects their marginal products.[18] Therefore, the correlation between k_{ijct}, l_{ijct}, m_{ijct}, and ω_{ijct} generates an endogeneity problem, i.e. a simultaneity bias[19], which makes OLS estimates inconsistent, as for the first time noted in Marschak and Andrews (1944).

Furthermore, if no allowance is made for entry and exit of firms over the sample period, a selection bias emerges (Olley and Pakes, 1996). The bias arises because a firm may decide to entry or exit the market according to its observed level of productivity. This generates a negative correlation between k_{ijct} and ω_{ijct}, conditional on being in the dataset (Ackerberg et al., 2007). Intuitively, firms with higher capital level can survive with lower productivity ω_{ijct}. Consequently, ignoring the entry and exit decision of firms or using a balanced panel result in upward biased TFP estimations.

Another important issue emerging from empirical applications is the omitted price bias. Usually, physical output and inputs are not observed. In the literature, a common solution to eliminate the price effect has been to deflate firm-level sales and input expenditures by an industry producer price index. However, if the choice of inputs is correlated with the unobserved firm-level price variation ($p_{ijct} - p_{jct}$), a bias is generated in the input coefficients.[20] In particular, assuming that inputs are positively

[18] For instance, a firm operating in a perfectly competitive market where the cost of capital, labour, materials is fixed as the output price, chooses the level of inputs that maximises its profit. The first order conditions (FOCs) say that the optimal level of k, l, m, depends on ω_{ijct}, where higher ω_{ijct} implies more use of those inputs. Hence, we expect the inputs coefficients to be positive biased. Usually, there is a greater bias on labour and materials. The intuition is that labour and materials tend to be flexible inputs. Conversely capital is a stock that accumulates over time, therefore less correlated with the productivity shock. Failure to correct for simultaneity bias leads to a downward bias in the TFP estimation for firms using many factors of production.

[19] If the researcher has no knowledge of the time when input decisions are taken, a simultaneity bias may arise because input quantities are determined during the period, when productivity shocks occur, while the researcher observes the data at the end of the period.

[20] Omitted price bias arises in the presence of imperfect competition, when firm-level

correlated with the output and the latter is negatively correlated with prices (as in standard demand-supply system), a negative correlation of inputs with firm-level prices leads to an under-estimation of TFP (De Loecker, 2011). The same applies when there is output price variation across firms.[21]

Traditional solutions to the endogeneity of input choice include fixed effects model (Hoch, 1955, 1962), (Mundlak, 1961, 1963), (Mundlak and Hoch, 1965), instrumental variables (Griliches and Maitresse, 1998) and first order conditions with respect to inputs (Klein, 1953),(Solow, 1957), (Griliches and Ringstad, 1971), and (Hall, 1988).

Fixed effects approach overcomes simultaneity and selection bias, assuming that productivity shock is time-invariant, i.e. $\omega_{ijct} = \omega_{ijc}$, and imposing strict exogeneity on ϵ_{ijct}. Then, the unobservable is differenced out by mean differencing

$$\begin{aligned} y_{ijct} - \bar{y}_{ijc} = \beta_k(k_{ijct} - \bar{k}_{ijc}) + \beta_l(l_{ijct} - \bar{l}_{ijc}) \\ + \beta_m(m_{ijct} - \bar{m}_{ijc}) + (\epsilon_{ijct} - \bar{\epsilon}_{ijc}) \end{aligned} \tag{A.3.3}$$

alternatively, first differencing

$$\begin{aligned} y_{ijct} - y_{ijc,t-1} = \beta_k(k_{ijct} - k_{ijc,t-1}) + \beta_l(l_{ijct} - l_{ijc,t-1}) \\ + \beta_m(m_{ijct} - m_{ijc,t-1}) + (\epsilon_{ijct} - \epsilon_{ijc,t-1}) \end{aligned} \tag{A.3.4}$$

Although we can estimate these equations with OLS, the fixed effect estimator has not performed well in practice (Griliches and Maitresse,

prices deviate from industry-price deflators, as suggested by Foster et al. (2008) in the empirical results of the Colombian manufacturing sector. Because information on actual firm prices is rare, several studies solve for firm-level prices after introducing a demand for the output, as in Klette and Griliches (1996) and De Loecker (2011). Similarly, some studies such as Eslava et al. (2004) and Ornaghi (2006) address the omitted input price bias.

[21] Input and output price variation may interact in such a way that the output price bias offsets the inputs price bias if the following assumptions hold true (Loecker and Goldberg, 2014):

i. Monopolistic competition within the industry.

ii. Firms produce horizontally differentiated products and face the same constant elasticity of substitution demand system.

iii. Constant return to scale.

iv. Input price variation (across firms and time) is input neutral.

1998). A common finding is the low estimate of the capital coefficient and returns to scale. Probably this is due to the strong fixed-effect assumption or to problems with data (Ackerberg et al., 2015).

An alternative method is to find instruments that are correlated with the endogenous inputs, but not with ω_{ijct} (and ϵ_{ijct}), and do not directly determine the output. Theoretically, there exist such instruments, for instance, input and output prices, but their validity is restricted to strong assumptions. Specifically, these prices affect firms' optimal choices of inputs, without directly determining the output. However, one key issue is the form of competition in input and output markets. If the input market is imperfectly competitive (i.e. a firm faces an upward sloping supply curve), higher ω_{ijct} implies that the firm uses a higher level of inputs, raising their price. If the output market is imperfectly competitive (i.e. a firm faces a downward sloping demand curve), then higher ω_{ijct} implies that the firm produces more, decreasing the price. In these circumstances, the correlation between the productivity shock and prices invalidates the exogeneity of these instruments. Moreover, imposing perfect competition in both markets is misleading because of little variation in capital and labour cost across firms. Many concerns also arise from the labour cost variation which may represent unobserved labour quality. Blundell and Bond (2000) offer an extended GMM estimator using as instruments the lagged first differences of inputs and imposing an error component model, which allows for firm-specific dynamics in productivity.

Another approach uses the information provided by the first order conditions of optimising firms (i.e. cost minimization, or profit maximisation). For instance, in the Cobb-Douglas framework, the output elasticities w.r.t. an input must be equal to its share of revenue. Thus, they are the factor coefficients of the production function. A caveat of this static first order conditions is that it does not take into account that input choices may be subject to dynamics, e.g., adjustment costs. If this is the case, we should impose additional auxiliary assumptions on the firms' environment.

More recent approaches have developed semi-parametric estimators to overcome some of the methodological issues described above. The basic idea underlying Olley and Pakes (1996) (OP hereafter) to eliminate the

endogeneity problem is to find an equation that makes the production shock, i.e. ω_{ijct}, "observable". They develop a dynamic model of firm behaviour to solve for the simultaneity and sample selection bias. The assumptions underlying the semi-parametric estimator are the following:

i. *First-order Markov process.* The productivity shock ω_{ijct} evolves according to a stochastic process, as follows:

$$p(\omega_{ijct}|I_{ijc,t-1}) = p(\omega_{ijct}|\omega_{ijc,t-1}) \tag{A.3.5}$$

where I_{ijct} is the firm i's information set of current and past productivity shocks.

ii. *Timing and dynamics of input choices.* Labor input l_{ijct} is non-dynamic in the sense that its choice upon observing ω_{ijct} at time t only affects current profits. Conversely, capital k_{ijct} is accumulated according to a dynamic investment process, i.e. $k_{ijct} = \delta k_{ijc,t-1} + i_{ijc,t-1}$, where $i_{ijc,t-1}$ is the investment decision of the previous period.

iii. *Scalar unobservable.* The dynamic investment function has the form:

$$i_{ijct} = f_t(k_{ijct}, \omega_{ijct}) \tag{A.3.6}$$

iv. *Strict monotonicity.* $f_t(k_{ijct}, \omega_{ijct})$ is strictly increasing in ω_{ijct}.

Following from the latter assumption, one can invert the investment function

$$\omega_{ijct} = f_t^{-1}(k_{ijct}, i_{ijct}) \tag{A.3.7}$$

to obtain an expression for the productivity shock ω_{ijct} in terms of variables observed by the econometrician. In the literature, i_{ijct} is called control or proxy variable.

The OP algorithm consists of two steps which follow the substitution of expression (A.3.7) in the log-linearised production function (A.3.2). Then, the estimating function is the following:

$$\begin{aligned} y_{ijct} &= \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} + \beta_m m_{ijct} + f_t^{-1}(k_{ijct}, i_{ijct}) + \epsilon_{ijct} \\ &= \beta_l l_{ijct} + \Phi_t(k_{ijct}, i_{ijct}) + \epsilon_{ijct} \end{aligned} \tag{A.3.8}$$

where we can use a polynomial approximation to $\Phi_t(k_{ijct}, i_{ijct})$ in k_{ijct} and i_{ijct}. In this way, the unobservable causing the endogeneity problem is eliminated, and the result of this stage is a consistent estimate of the labour and materials coefficients. We can obtain these coefficients running OLS of y_{ijct} on l_{ijct}, or using GMM on the following moment condition:

$$E[\epsilon_{ijct}|I_{ijct}] = E[y_{ijct} - \beta_l l_{ijct} + \Phi_t(k_{ijct}, i_{ijct})|I_{ijct}] = 0 \tag{A.3.9}$$

To estimate the coefficient on capital, a second stage is required. Here, it is necessary to exploit information on firm dynamics. Following the assumption that productivity shock is a Markov process, we can decompose ω_{ijct} into its conditional expectation at time $t-1$, and an innovation component ξ_{ijct} uncorrelated with productivity and capital

$$\begin{aligned} \omega_{ijct} &= E[\omega_{ijct}|I_{ijc,t-1}] + \xi_{ijct} = E[\omega_{ijct}|\omega_{ijc,t-1}] + \xi_{ijct} \\ &= g(\omega_{ijc,t-1}) + \xi_{ijct} \end{aligned} \tag{A.3.10}$$

Substituting (A.3.10) in the production function (A.3.2), we get:

$$\begin{aligned} y_{ijct} &= \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} + \beta_m m_{ijct} + g(\omega_{ijc,t-1}) + \xi_{ijct} + \epsilon_{ijct} \\ &= \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} + g(\Phi_{ijc,t-1}(k_{ijc,t-1}, i_{ijc,t-1}) - \beta_{j0} \\ &\quad - \beta_k k_{ijc,t-1}) + \xi_{ijct} + \epsilon_{ijct} \end{aligned} \tag{A.3.11}$$

Given that $E[\xi_{ijct}|I_{ijc,t-1}] = 0$ and $E[\epsilon_{ijct}|I_{ijct}] = 0$, we can substitute the first stage estimates on labour and materials and use the following moment condition to identify the capital coefficients:

$$\begin{aligned} E[(\xi_{ijct} + \epsilon_{ijct})|I_{ijc,t-1}] = E[&y_{ijct} - \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} \\ &- g(\Phi_{ijc,t-1}(k_{ijc,t-1}, i_{ijc,t-1}) + \\ &- \beta_{j0} - \beta_k k_{ijc,t-1})|I_{ijc,t-1}] = 0 \end{aligned} \tag{A.3.12}$$

Wooldridge (2009) propose estimating both stages simultaneously in a single GMM step, specifying lags of capital and inputs as instrumental variables. This improves the two-step estimation in terms of efficiency and robust standard errors are easy to obtain.[22] A drawback is that the

[22]Because the OP estimation involves two stages, deriving analytic standard errors is non trivial, so clustered (by firm) bootstrap is often used. Moreover, two steps estimators ignore the possible error correlation across equations and heteroskedaticity.

non-linear search over a large set of parameters is very demanding. OP also faces the sample selection problem, considering that unproductive firms may exit the market. Thus, It introduces survival probabilities, i.e. $P_{ijc,t-1}$, based on the following exit rule:

$$\chi_{it} \begin{cases} 1 & \text{if } \omega_{it} \geq \underline{\omega}_{it} \\ 0 & \text{otherwise} \end{cases}$$

The probability is estimated by fitting a probit model[23] of χ_{ijct} on $i_{ijc,t-1}$ and $k_{ijc,t-1}$, as well as their squares and cross products. Then, the predicted probability $\hat{P}_{ijc,t-1}$ enters the second stage of the algorithm.

Levinsohn and Petrin (2003) (LP hereafter) advocate that the assumption of *strict monotonicity* of i_{ijct} in ω_{ijct} is a strong assumption, in the sense that it is not always true that firms with the same k_{ijct} and i_{ijct} have the same productivity level. In practice, investment activity is lumpy, and i_{ijct} is often 0. Although, OP can be extended to allow weak monotonicity, this implies discarding zero investment observations from the analysis. As an alternative, LP propose to use a non-dynamic variable, i.e. intermediate input, as a proxy for the unobservable.[24] Similarly to the investment choice in OP, the firm's optimal choice of m_{ijct} is a function of k_{ijct} and ω_{ijct}, i.e.

$$m_{ijct} = f_t(k_{ijct}, \omega_{ijct}), \tag{A.3.13}$$

assuming *strict monotonicity* it can be inverted, as follows

$$\omega_{ijct} = f_t^{-1}(k_{ijct}, m_{ijct}), \tag{A.3.14}$$

and substituted in the production function as in (A.3.8). First and second stages follows from OP, though an additional moment condition is required

[23] The dependent variable is a survival dummy, i.e. equal to 1 if the firm does not exit at period t. The left-hand side includes the same polynomial terms of the first stage in capital and investment, plus the constant. Refer to Ackerberg et al. (2007) for an exhaustive description.

[24] An important advantage of using intermediate inputs for the unobserved productivity is that it does not rule-out any firm-specific unobservables affecting investment choices (e.g., the price of investment or capital adjustment costs). Nevertheless, both OP and LP do not allow for serially correlated, unobserved firm-specific shocks to intermediate inputs and labour prices.

in the second step, i.e. $E[(\xi_{ijct} + \epsilon_{ijct})|m_{ijc,t-1}] = 0$.[25]

Ackerberg et al. (2015) extend the semi-parametric estimator of OP to overcome some issues related to the identification of the labour coefficient in the first stage. The critique arises from the fact that l_{ijct} may be collinear to the non-parametric function in k_{ijct} and m_{ijct}.[26] A solution is not to identify the labour coefficient in the first stage, but do it in the second stage, assuming a conditional intermediate inputs demand function.

$$m_{ijct} = f_t(k_{ijct}, \omega_{ijct}, l_{ijct}) \tag{A.3.18}$$

Satisfying the *Strict monotonicity* assumption, equation (A.3.18) can be inverted and substituted into the production function.[27]

De Loecker (2011) improve on OP algorithm correcting for the omitted output price bias in the presence of imperfect competition when estimating revenue productivity. The underlying idea is to explicitly define a demand system and solve for firm-level prices.[28] Then, by adding the industry output as an additional regressor in the output deflated production function[29], it is possible to proxy for unobserved firm-level prices, as shown in

[25]LP do not correct for the selection bias, because the empirical results provided by OP are small when an unbalanced panel is used.

[26]The first LP stage can be written as:

$$y_{ijct} = \beta_l l_{ijct} + np(k_{ijcT}, m_{ijct}) + \epsilon_{ijct} \tag{A.3.15}$$

where $np(k_{ijc}, m_{ijct})$ is a non-parametric function. Then, a natural model for firms to choose l_{ijct} can be

$$l_{ijct} = h_t(k_{ijct}, \omega_{ijct}). \tag{A.3.16}$$

Replacing ω_{ijct} with (A.3.14), we have

$$l_{ijct} = h_t(k_{iJCt}, f_t^{-1}(k_{ijct}, m_{ijct})) = \tilde{h}_t(k_{ijct}, m_{ijct}) \tag{A.3.17}$$

[27]Note that even if the first does not identify any parameter, it is still crucial to separate ω_{ijct} from ϵ_{ijct}.

[28]It is based on the fact that the industry price level is constructed as the weighted average of all firm-level prices, where weights usually refer to turnover information.

[29]The output deflated production function is the following:

$$\begin{aligned} \tilde{r}_{ijct} &= p_{ijct} + y_{ijct} - \bar{p}_{ijct} \\ &= \beta_{j0} + \beta_k k_{ijct} + \beta_l l_{ijct} + \beta_m m_{ijct} + (p_{ijct} - \bar{p}_{ijct}) + \omega_{ijct} + \epsilon_{ijct} \end{aligned} \tag{A.3.19}$$

where $\bar{p}_{ijct}$ is the industry-level price deflator, i.e. $\bar{p}_{ijct} = p_{jct}$ and intermediate inputs may be correlated with the unobserved price difference, i.e. $E(x_{ijct}(p_{ijct} - \bar{p}_{ijct})) \neq 0$, with $x_{ijct} = (k_{ijct}, l_{ijct}, m_{ijct})$.

Van Beveren (2012) using a conditional Dixit-Stiglitz demand system.

Eventually, stochastic frontier techniques have become popular in the estimation of firm-level productivity. The idea is to measure a firm's performance as a deviation from a frontier isoquant, as for the first time addressed in Farrell (1957). Aigner et al. (1977) and Meeusen and van Den Broeck (1977) decompose the residual into two components: the unobserved productivity and the measurement error. The first component is assumed to follow a one-side distribution (e.g., half-normal distribution), while the latter is symmetrically distributed around the frontier (e.g. normal distribution).[30]

In the contribution presented in this Chapter 3, we exploit the recent semiparametric technique by Ackerberg et al. (2015), which solves the simultaneity bias elaborating on the functional form and on the data generating processes, originally present in seminal works by Olley and Pakes (1996), and Levinsohn and Petrin (2003).

A.2 Deflation and International Comparability

To analyse and compare the TFP estimations over time, across countries and industries, we deflate nominal values of variables necessary for the computation and hence report all values in a common real currency-year.

First, we convert variables in Orbis from euros to local currencies at the end of the year for countries outside the Eurozone in each period.[31] Then, following Smarzynska Javorcik (2004), we separately deflate output, intermediate inputs, and capital using 2-digit industry producer price index (PPI) for each country-year, with the base year 2010. In particular, the output which refers to turnover is deflated using the total PPI of the focal industry from Eurostat and World Bank (WDI). Capital, that is defined as the value of fixed assets at the beginning of the year, is deflated by the simple average of the deflators for the following 2-digit NACE rev. 2 industries: 26, *Manufacture of computer, electronic and optical products*;

[30] See Van Biesebroeck (2007) for a comparison of several TFP estimation techniques including stochastic frontiers.

[31] The values of the variables in Orbis are expressed in euro currency at market exchange rate, at the end of the corresponding year.

27, *Manufacture of electrical equipment*; 28, *Manufacture of machinery and equipment n.e.c.*; 29, *Manufacture of motor vehicles, trailers and semi-trailers*; 30, *Manufacture of other transport equipment.*

Intermediate inputs, corresponding to material costs, are deflated by the intermediate inputs' deflator that is calculated as the weighted average of PPI of the supplying industries, with technical coefficients (expenditure shares of input industries) as weights retrieved from the WIOD input-output tables (2016 Release).[32] In fact, these weights are representing the proportion of inputs sourced from a given sector.

We deal with missing PPI in several ways, after checking the availability in National Statistic Office tables: if the data are missing for the whole industry, we use the more aggregated PPI on manufacturing; if the data is missing for one or more years, we interpolate using the closest years; if it is not possible to construct the deflator for material cost, we use the aggregate PPI on intermediate goods; also, some adjustments for specific industries' aggregations are done to correspond to WIOD industry groups (e.g., we use NACE 31-32 in WIOD for the aggregation 31-33 in PPI).

Finally, we convert back the deflated variables in the domestic currency to euro, using the relevant exchange rate of the (base) year 2010. This procedure ensures that the change in prices does not distort the level and the growth of TFP.

[32]The extended World Input-Output Database (WIOD) covers 28 EU countries and 15 other major countries in the world for the period 2000-2014. All the details are offered in Timmer et al. (2016). For the year 2015, which is part of our study, we use the weights of 2014.

B Appendix: Tables

Table B.3.1: Industry classification for manufacturing

Code	2-digit NACE rev.2
10	Manufacture of food products
11	Manufacture of beverages
12	Manufacture of tobacco products
13	Manufacture of textiles
14	Manufacture of wearing apparel
15	Manufacture of leather and related products
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
17	Manufacture of paper and paper products
18	Printing and reproduction of recorded media
19	Manufacture of coke and refined petroleum products
20	Manufacture of chemicals and chemical products
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations
22	Manufacture of rubber and plastic products
23	Manufacture of other non-metallic mineral products
24	Manufacture of basic metals
25	Manufacture of fabricated metal products, except machinery and equipment
26	Manufacture of computer, electronic and optical products
27	Manufacture of electrical equipment
28	Manufacture of machinery and equipment n.e.c.
29	Manufacture of motor vehicles, trailers and semi-trailers
30	Manufacture of other transport equipment
31	Manufacture of furniture
32	Other manufacturing
33	Repair and installation of machinery and equipment

C Appendix: Construction of variables

ERDF Business Support: Data are sourced from the report 'Geography of Expenditure – Work Package 13' prepared for the European Commission in 2015 wiiw and ISMERI EUROPA (2015). Projects in this category should help firms or groups of firms, in particular, SMEs, with services and investments in innovation and sustainable production. We use payments by NUTS-2 region for the programming period 2007-2013.
ERDF Research Technology Development: Data sourced from the report 'Geography of Expenditure - Work Package 13' prepared for the European Commission in 2015 wiiw and ISMERI EUROPA (2015). Projects in this category should stimulate research and innovation activities through investments in research centres, promoting technology transfers and cooperation between businesses and the scientific environment. We use payments by NUTS-2 region for the programming period 2007-2013.

Firm-level controls
Firm size: Computed from Orbis data. It is the (log of) firm-level turnover.
TFP $(t-1)$: Total Factor productivity estimate from Orbis data, after implementation of the semiparametric technique by Ackerberg et al. (2015).
Region-level controls
Human capital is controlled after including two variables at the NUTS-2 region level: i) the percentage of people aged 25-64 with upper secondary and postsecondary education; ii) and the percentage of people aged 25-64 with tertiary education, following International standard classification of education (ISCED) 2011. Data are sourced from Eurostat.
Agglomeration (or geographic concentration) is proxied using data from the Structural Business Statistics (SBS) in Eurostat. It represents the concentration at the regional level of the industry j in which the firm i is active, and it is computed taking the ratio of the regional industrial employment to the total regional area, following Henderson et al. (1995).

$$Agl = \frac{L_{jrt}}{Area_{rt}} = \frac{\sum_i L_{ijrt}}{Area_{rt}} \tag{C.3.1}$$

where L_{jrt} is the total employment for industry j in the region r at time t, and $Area_{rt}$ is the area of the region r at time t in square kilometres.

Industrial specialisation is proxied using Structural Business Statistics (SBS) data from Eurostat. It captures the concentration of production in the industry j within region r at time t. It is measured using the ratio of regional industrial employment to the total regional employment, following Henderson et al. (1995).

$$S = \frac{L_{jrt}}{\sum_j L_{jrt}} = \frac{\sum_i L_{ijrt}}{\sum_j \sum_i L_{ijrt}} \tag{C.3.2}$$

Industrial diversification is proxied using the Structural Business Statistics (SBS) data from Eurostat. It shows the within-regional concentration of industries (j') other than the one under investigation (j), following a Hirschman-Herfindahl Index, i.e., the sum of squares of the shares of other industries' employment ($L_{j'rt}$) in the region relative to the total employment in the region, except for the industry of the firm.

$$D = \sum_{j' \neq j} S^2_{j'rt} = \sum_{j' \neq j} \left(\frac{L_{j'rt}}{\sum_{j' \neq j} L_{j'rt}} \right)^2 \tag{C.3.3}$$

Competition is proxied using the Structural Business Statistics (SBS) data from Eurostat. It captures the competition of firms within the local industry according to the number of local units resulting in the inverse average employment of the local firms.

$$C = n_{jrt}/L_{jr} \tag{C.3.4}$$

where n_{jrt} is the number of local units (firms) in the region r at time t, operating in industry j.

Chapter 4

Spatial Clustering and Survival: Evidence from Firms in Umbria and Puglia

4.1 Introduction

Despite the increasing breakdown of production by tasks at a global level and the offshore of some manufacturing activities, especially toward developing countries, many local clusters of firms, and thereby industries, manage to survive in the new global scenario.[1] Competencies and special-

[1]The general reduction of transportation costs has led many analysts to claim the "death of distance", however as Storper (2013) suggests "distance is not dead; it is not even sick". Still, communication and coordination are not easy along geographical distances and cultural borders. There are many examples on this line. For instance, after an initial wave of offshoring service centres and data processing in the early 2000s to countries such as India, many firms had decided to re-onshore part of their customer service, because of reduction in quality standards. The Boing's company experienced four years delay in the production of the 787 Dreamliner aircraft, after the outsourcing of production to about 50 suppliers around the world. Nevertheless, some firms, such as Walmart and IKEA, source from global supply chains while others, such as Zara,

isation in core activities of firms based in local districts co-evolve with the needs and features of their global buyers and production partners (De Marchi et al., 2017). In this respect, the organization of industrial clusters is mainly based on subcontracting relationships within vertically fragmented supply chains.

Traditional trade theory emphasizes the role of endowments, factor prices, technologies and policy regimes in explaining the different patterns of production across different locations. For instance, this can explain why some manufacturing activities have moved from high-wage developed countries to low-wage developing countries. However, this theory fails to explain why *a priori* similar regions can develop very different production structures and economic activity is unevenly distributed across space (Ottaviano and Puga, 1998).

Another approach is provided by theories of economic geography that look at the spatial agglomeration of economic activities, i.e. the physical proximity of firms, workers and consumers. The first agglomeration theory dates back to Marshall (1890) and explains the tendency of firms from the same industry to concentrate in industrial clusters (e.g., Silicon Valley). Marshall (1890) identifies three agglomeration's mechanisms: labour pooling, knowledge spillover, and input-output linkages, which are associated with costs or benefits to firms. Clustering of firms and proximity effects cannot be disregarded.

A broad literature on industrial clusters (or similarly industrial districts) supporting and explaining the advantages related to local production systems developed with the seminal contribution of Beccatini (1986), Piore and Sabel (1984), Krugman (1991) and Porter (1986). Although alternative theoretical and empirical studies emerged, policy-makers inspired by the idea of industrial clustering Porter (1990) have used cluster-economy development as a policy tool to promote local development.

Since the seminal contribution of Glaeser et al. (1992) many scholars have attempted to understand the impact of agglomeration externalities[2]

has strong local clusters for some production stages, including design.

[2]In the contemporary literature, agglomeration externalities are often divided into three types of economies of scale and scope: urbanisation externalities, localisation, and diversification. Urbanization externalities reflect the effects of city size, irrespective

(Duranton and Puga, 2004; Rosenthal and Strange, 2004; Beaudry and Schiffauerova, 2009; Combes et al., 2012)), with a focus on the possible impact that agglomeration has on firm's productivity, innovation, real wages, employment and location choice. By contrast, few studies have investigated the role of agglomeration economies in business demography, i.e. firm entry and exit (Cainelli et al., 2014; Ferragina and Mazzotta, 2015; Basile et al., 2017).

In particular, the literature on firm survival has paid much attention to firm and industry characteristics [3], ignoring the geography and therefore costs and benefits of firms being located in proximity. For instance, firms producing similar products within the localised cluster may face higher competition, and thus only the most productive firms are more likely to survive; or firms located in highly specialised areas, such as industrial districts, may be more vulnerable to idiosyncratic shocks (Basile et al., 2017). Also, the related local variety (Jacobs, 1969), arising from the co-existence of diverse industries sharing common or complementary knowledge, may protect firms from idiosyncratic demand shocks and therefore increase the survival probability. Finally, urban agglomeration captured by the overall size of the economy may increase the likelihood of survival when firms face higher local demand and supply of public services, and congestion costs (e.g., land prices, crime rates) are negligible (Ciccone and Hall, 1996). More details about state of the art in this research field are offered in the next Section 4.2.

In this contribution, we aim at enriching the empirical literature on firm

of sectoral composition. Big cities often provide higher access to local market and better infrastructure, but they are also plagued by congestion, resulting in pollution and high factor costs, high real estate rents and other disamenities. Hence, urbanisation externalities can just as well represent economies as diseconomies to local firms. Localization externalities are intra-industry spillover benefits that firms derive from the local presence of other firms belonging to the same industry (Marshall, 1890). Diversification externalities capture inter-industry spillover benefits that arise from the presence of a large variety of industries in the local area (Jacobs, 1969).

[3] Although it is not very widespread in externality studies, survival analysis has been used intensively in the fields of industrial dynamics and business studies. Most of these studies have looked at survival of firms or plants with respect to their size and age (Disney et al., 2003), pre-entry experience (Thompson, 2005), the structure of the market (Cantner et al., 2006; Buenstorf, 2007), the maturity of the industry (Agarwal and Gort, 2002), or combinations of these dimensions (Klepper, 2002).

survival testing the impact of agglomeration externalities in continuous space arising from the spatial concentration of activity in the same industry (localisation), while controlling for other firm- and industry-level characteristics. Particular attention is given to asymmetric industry effects. More generally, we separate the analysis of manufacturing and service sectors. In fact, over the last decades, many manufacturing firms concentrated in clusters have spread out because of an international market orientation towards foreign demand. On the contrary, firms providing services tend to be more oriented to local demand.

Our focus is on Italian firms, in particular, those located in the regions of Umbria and Puglia operating in several industries.[4] The Italian case is worth studying because it has a long tradition in local clusters, or similarly industrial districts[5], of small and medium-enterprises (SMEs) interacting in the same industry. Since the theoretical and empirical contribution of Becattini (1979), many efforts have been done to detect industrial districts along the peninsula. ISTAT currently identifies 141 industrial districts by Local Market Areas (LMAs) and economic specialisation by using the data on economic units from the 9^{th} Industry and Services Census. Although there has been a decline in the number of districts[6], Local systems represent about one-fourth of the Italian economic system, and districts still characterise the Italian manufacturing industry. They represent 64.1% of mainly manufacturing LMAs which employ 65.8% of

[4]We took two Italian regions, i.e. Umbria and Puglia, as case studies, although the methodology implemented in this work can be applied to any other areas or countries.

[5]The National Statical Office, ISTAT, designs local labour market areas (LMAs, 'local labour systems' - 'SSL' in Italy) as functional geographic areas beyond the administrative boundaries where the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. Then, it classifies LMAs as clusters if they satisfy three requirements: i) a higher percentage of employees in manufacturing than in agriculture; ii) specialisation in one particular industry, mainly within manufacturing; iii) a high concentration of workers in firms with less than 250 employees, all compared to the national average.

[6]The structural changes of the Italian economy driven by process of tertiarisation can explain the reduction in the number of districts. Besides, over the years there have been changes in the ISTAT classification criteria. Eventually, the rise of international competition in low labour cost countries, the stagnation of domestic and EU markets, the increasing technological complexity and the organisation of production along global value chains (GVCs) are reshaping the industrial districts Rabellotti and Giuliani (2017).

workers in manufacturing, as reported by ISTAT in 2011.

First, we introduce a data-driven approach to identify industrial clusters. Specifically, after a careful process of geocoding to retrieve the geographical coordinates (i.e. latitude and longitude) from the postal address of each firm in our sample, we implemented a density-based spatial clustering algorithm, namely the *DBSCAN*. Using this method we were able to identify industrial clusters by NACE rev.2 2-digit in each region, beyond existing administrative boundaries and to overcome the limitations of the qualitative assessment of clusters, often based on specific sectors case studies.

Second, we used the spatial information to group firms within and outside an industrial cluster. In this way, using the Kaplan and Meier (1958) estimator, we explored the survival rate of firms for the two groups. Interestingly, we found that: i) for manufacturing firms, it is more likely to survive outside an industrial cluster after about one year from establishment; ii) for firms in services, it is more likely to survive in geographically concentrated areas, i.e. inside a cluster, after three years and a half of activity.

Third, we tested the impact of several variables on the survival probability of a cohort of firms established between 1st January 2008 and 31st December 2016 using the semiparametric Cox (1972) model with a cluster-specific baseline hazard. Focusing on agglomeration externalities, we found that: i) an increase in the concentration of firms operating in different types of activities, captured by the *unrelated variety* (Frenken et al., 2007), increases the survival rate of firms in the service sectors, while we did not find any significant effect for the manufacturing; *within* two-digit industry externalities, captured by *related variety* (Frenken et al., 2007) has a negative impact on firms' longevity in services, while the effect on manufacturing firms is not significant; iii) the degree of urbanization, measured by the *population density* has a significant non-linear effect on firm's survival. Finally, we also found a significant association of financial variables with survival probabilities which are often neglected in the related literature.

The rest of the chapter is organised as follows. The next Section 4.2

positions our contribution with respect to literature. Section 4.3 presents the framework of spatial clustering using micro-geographic data and the *DBSCAN* algorithm implementation in our research. Section 4.4 describes the empirical strategy and results. Concluding remarks are offered in Section 4.5..

4.2 Related Literature

A significant number of empirical studies analyses the factors that impact the survival probability of firms in the market. At the firm level, there is a common agreement that the age and the size significantly affect the firm survival as predicted by the model provided by Jovanovic (1982). In particular, firms in the first few years of their life are expected to survive less than the older cohorts also operating in the same market Dunne et al. (1989); Mata and Portugal (1994); Segarra and Callejón (2002). According to Jovanovic (1982)'s model, as new firms get some experience by carrying out their activities, they adapt their output to the 'true' efficiency level and reduce their risk of failure. Therefore, the survival probability increases over the life cycle of a firm, although not linearly. Besides, several studies argued that small firms, which are more likely to operate below the minimum efficient scale and have less access to capital markets, are less likely to survive Evans (1987); Hall (1987); Agarwal and Audretsch (2001). Also, the ownership structure of a firm has been empirically investigated, and findings are ambiguous. A group of studies claimed that foreign-owned firms are less likely to survive than domestic firms (among others, Colombo and Delmastro (2000) for Italy; Gorg and Strobl (2003) for Ireland; Ferragina and Mazzotta (2015) for Italy), while others found no differences in the chance of survival between domestic and foreign firms (among others, Mata and Portugal (1994) for Portugal; Blanchard et al. (2016) for Belgium).

At the industry level, market competition, usually proxied by the concentration ratio, favours firms' growth, although it is expected to have an impact on firms' survival. On the one hand, firms struggle to survive in an environment with a high concentration in an industry. On the other

hand, the market concentration may lead to higher price-cost margins in the industry which increase the survival probability (Basile et al., 2017). Also, the empirical evidence is quite ambiguous: some studies found that a competitive environment threatens the survival probability of a firm (Audretsch and Mahmood, 1995; Gorg and Strobl, 2003), while others found a negative effect (Mata and Portugal, 1994; Strotmann, 2007). In addition, it appears that firms live longer in growing industries than in declining industries, thanks to better demand conditions (Audretsch and Mahmood, 1995; Mata and Portugal, 1994; Strotmann, 2007; Ferragina and Mazzotta, 2015; Basile et al., 2017).

The literature on firm survival has only recently started to consider the possible effects of geographical proximity, including measures of spatial concentration of a given industry within a specific area. Among others, De Silva and McComb (2012) analysed the agglomeration effects on firm survival in Texas (U.S.) using establishment data. They found evidence that a higher concentration of firms in the same industry within small geographic areas (1 mile) increases the mortality rate, whereas over a larger geographic area the localisation effects reduce establishment mortality. Additionally, Van Oort et al. (2012) using a multilevel analysis assessed the extent to which variance in the survival and growth rates of firms in the advanced producer service sector in the Netherlands can be attributed to between-firm variance, between-area variance or between-sector variance. They showed that localisation has a small, positive effect on new firm survival. Neffke et al. (2012) examined the agglomeration effects on manufacturing plant survival in Sweden during the period 1970-2004. They found that diversification contributes to firm survival in the first 15 years of existence, while localisation externalities have no impact. Interestingly, they observed that the local presence of technologically related industries increases the likelihood of firm survival. Further, Boschma and Wenting (2007) illustrated the spatial evolution of the British automobile industry between 1895 and 1968. They showed that the local presence of related sectors, such as bicycles and coach making, had a significant and negative impact on firm survival.

On this strand of literature, there are some empirical studies which

focus on the behaviour of Italian firms. Italy represents an interesting case study because since the 1980s the entire peninsula develops an industrial growth model based on industrial districts. Remarkable is the spatial agglomeration of small and medium-sized enterprises (SMEs), specialised in different phases of the production process, which had achieved economies of scale comparable to those obtained by large firms (Rabellotti and Giuliani, 2017).

However, empirical evidence on recent developments shows that many firms have reorganised their production to compete in a global context and to participate in global value chains. In this framework, Carree et al. (2011) analysed a range of determinants of the exit rate for 12 different sectors in the Italian provinces for 11 years. They observed that the presence of industrial districts reduces the likelihood of firm exit in several industries, including the manufacturing sectors: Food, Clothing, Commerce, and Transport. Then, Cainelli et al. (2014), using a large panel dataset for the Italian economy at the industry-province level over 13 years, found that localisation significantly affects firm exit in the short-run, especially for low-tech firms. Also, their results showed that the impact of diversification is less clear, although negative and significant for low-tech firms. More recently, Basile et al. (2017) assessed the impact of spatial agglomeration externalities on Italian start-up firms' survival between 2004 and 2010. Overall, their results suggested that agglomeration economies exert a significant effect on firms' survival. In particular, localisation reduces the likelihood of firm's exit but only in services. Then, the diversity of technologically related activities (related variety) positively affect manufacturing industries, while unrelated variety has a significant effect on services. Eventually, urbanisation does not have any significant effect on Italian start-up's survival.

Finally, the effects of financial indicators have been largely neglected in the literature of firm survival. According to the OECD (2009) report, Italian SMEs strongly depend on banks and structural financial constraints, and these financial aspects were particularly evident during the recent crisis.[7] Only a bunch of studies considered financial variables and found

[7]See ECB (2013) for a detailed review of the corporate finance structure within the

a significant association with firms' survival probabilities (Zingales, 1998; Fotopoulod, 2000; Bunn and Redwood, 2003; Vartia, 2004).

4.3 Spatial Clustering using Micro-Geographic Data

Recent contributions show that firms with different characteristics such as size, technology and productivity coexist within industrial clusters and productive systems (Rabellotti and Schmitz, 1999; Wang, 2015). A central issue in this framework is how to identify this tendency of localisation and to measure it properly.

Several discrete-based indexes have been suggested to measure the spatial concentration of economic activities in spatial units. Among others, there are the LQ (location quotient), the locational Gini coefficient, the spatial Hirshman-Herfindahl Index, the Ellison and Glaeser (1997) concentration index, etc. These indexes are widely used to identify the industrial pattern and check whether there is a high concentration that could detect specialisation in an area. They usually rely on Census employment data.

As a drawback, these measures fail to identify the spatial dimension of a cluster. Results depend on the level of agglomeration chosen, e.g., cities, regions, states. In fact, outcomes may vary a lot according to the select area under investigation (Briant et al., 2010; Burger et al., 2010). Therefore, it is not reasonable to assume that economic structures follow boundary lines, especially between regions located in the same countries. This scaling and regional's boundary problem is known as 'Modifiable Areal Unit Problem' as first described in Openshaw (1984). Still, in the literature, there is no agreement on a local threshold to detect the presence of clusters. Another crucial neglected aspect is the form of the agglomeration patterns since there could be a concentration of sectors in one spatial unit or a concentration in several spatial units randomly distributed. We have to take into account that location of economic activities is highly restricted due to planning regulations, land peculiarities (e.g., mountains, lakes,

Euro area in turbulent times, such as the 2008 financial crisis, using firm-level data.

swamps) or political system's restrictions.

Distance-based methods have been used to overcome the limitations mentioned above. They take into account the spatial concentration of a sector relative to the overall location of firms in the area under investigation. These agglomeration indices usually require micro-geographic data (Arbia, 2001), i.e. latitude and longitude. They consider firms as spatial units of analysis and measure the density of economic activity along the link, i.e. geographical distance, between pairs of firms.

Two micro-geographical distance-based approaches have been elaborated in the literature. One method, using distance bands, counts geocoded firms within a specific distance to evaluate localisation. Among others, these works include Duranton and Overman (2005) which use K-density functions, Marcon and Puech (2010) using M-functions and Arbia (2001), which investigate Ripley's K-function. None of these indexes aims at detecting the spatial location of clusters. In fact, they allow for a test of significant dispersion or concentration, but they do not provide information where in space this concentration can be found (Scholl and Brenner, 2016). Another method uses distance-decay functions and builds firm-specific values of spatial concentration as a proxy of agglomeration. It cumulates the inverted distances of one geocoded firm to all other firms in the same industry (Scholl and Brenner, 2016). Various decay functions exist in the literature, from simple linear (Audretsch et al., 2005), exponential (Drucker and Feser, 2012), and log-logistic (De Vries et al., 2009), and whereby results depend on the chosen distance function. For policy-makers, it is of great interest to know the location of firm's industrial clusters and to identify the firms within them. This information, along with the heterogeneity of firms within a sector, can enrich the debate when preparing industrial policies. Hence, new agglomeration indices should be developed.

There is no general agreement on the principles a measure should satisfy, although Duranton and Overman (2005) propose five criteria for measuring agglomeration in a meaningful way. According to them, an agglomeration index should: i) be comparable across industries; ii) control for the overall agglomeration of manufacturing; iii) control for industrial concentration; iv) be unbiased to scale and agglomeration; v) and indicate the significance

of the results. Additionally, Combes et al. (2008) identify some additional properties: vi) be computable in closed form from accessible data; vii) be justified by a suitable model; viii) and be comparable across spatial scales. Eventually, a measurement of agglomeration must be able to separate random from non-random clusters, to provide ways to measure the exceptional concentration of economic activity.

4.3.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Its Implementation

A challenging identification method of industrial clusters is the Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) algorithm (Ester et al., 1996).[8] This data clustering algorithm identifies high concentration of points (e.g., firms) without relying on existing boundaries definitions. The idea is to group density-connected points (e.g., firms), i.e. points that are close in space and with many nearby neighbours while leaving alone-points in low-density regions. Differently from other clustering algorithms: i) it does not require knowledge on the number of clusters in advance; ii) it is a density-based approach; iii) it is designed to discover clusters of arbitrary shape, size, and density; iv) it can unambiguously detect noise points not belonging to any clusters; v) and it is efficient on large databases.

Given a dataset D containing a set of points $p \in D$, *DBSCAN* estimates the density around a point using the concept of ε-neighbourhood.

Definition 1. ***ε-Neighbourhood.*** *The ε-neighbourhood of a point p, $N_\varepsilon(p)$, is the set of points within a specified radius ε around the point p. It is defined by*

$$N_\varepsilon(p) = \{q | d(p,q) \leq \varepsilon\}$$

where d is some distance measure and $\varepsilon \in \mathbb{R}^+$.

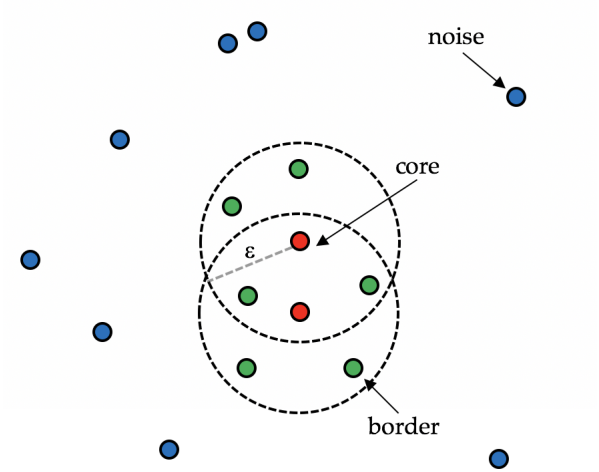
[8] *DBSCAN* is one of the most common clustering algorithms in Machine Learning and also the most cited in the scientific literature. In this respect, in 2014, *DBSCAN* received an award at a leading data mining conference (KDD) because of the remarkable attention given by the Scientific Community.

In the ε-neighbourhood of p a minimum number of points ($MinPts$) is required to identify a cluster. Each point of the dataset D is classified into *core*, *border*, or *noise* point. A visual representation is offered in Figure 4.1.

Definition 2. ***Point Classification.*** *A point $p \in D$ is classified as:*

- *A core point if $|N_\varepsilon(p)| \geq MinPts$, i.e. $N_\varepsilon(p)$ has a high density, where $MinPts \in \mathbb{Z}^+$ is a threshold parameter chosen by the user;*
- *A border point if p is not a core point, but $p \in N_\varepsilon(q)$, i.e. it is in the ε-neighbourhood of a core point $q \in D$;*
- *A noise, otherwise.*

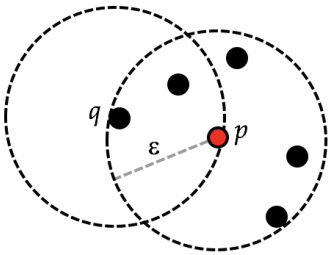
Figure 4.1: Core, border and noise points



Source: Own elaboration.
Note: This figure shows an example of points' classification. Red points are classified as *core* points, because the area in an ε radius contains at least 5 points, assuming *MinPts*=5. Green points are classified as *border* points, and the blue points are *noise* points.

DBSCAN identifies contiguous dense regions from individual points using the notions of *density-reachable* and *density-connected*, as formerly defined in (Ester et al., 1996).

Figure 4.2: Directly density-reachable, *MinPts*=5



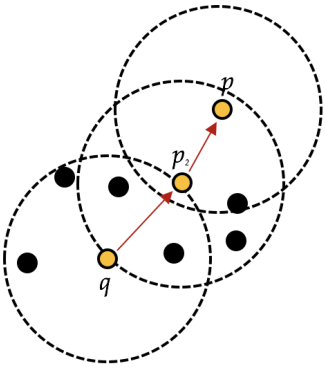
Source: Own elaboration.

Definition 3. ***Directly density-reachable.*** *A point $q \in D$ is directly density-reachable from a point $p \in D$ with respect to ε and $MinPts$ iff*

1. *$q \in N_\varepsilon(p)$, and*
2. *$|N_\varepsilon(p)| \geq MinPts$ (core point condition).*

Directly density-reachability is symmetric for pairs of *core* points and asymmetric if one *core* point and one *border* point are considered.

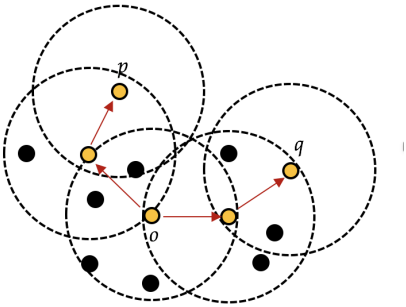
Figure 4.3: Density-reachable, *MinPts*=5



Source: Own elaboration.

Definition 4. ***Density-reachable.*** *A point p is density-reachable from a point q if there is an ordered sequence of points $(p_1, p_2, ..., p_n) \in D$, with $p_1 = q$ and $p_n = p$, such that p_{i+1} is directly density-reachable from p_i, $\forall i \in \{1, 2, ..., n-1\}$.*

Figure 4.4: Density-connected, *MinPts*=5



Source: Own elaboration.

Definition 5. ***Density-connected.*** *A point $p \in D$ is density-connected to a point $q \in D$ if there is a point $o \in D$ such that both p and q are density-reachable from o.*

The notions of *density-connected* and *density-reachable* are employed to identify clusters.

Definition 6. ***Cluster.*** *A cluster C is a non-empty subset of D satisfying the following conditions:*

1. *(Maximality) $\forall p, q$: if $p \in C$ and q is density-reachable from p, then $q \in C$, and*

2. *(Connectivity) $\forall p, q$: p is density-connected to q.*

In other words, a cluster is a set of density-connected points which is maximal with respect to density-reachability.

The algorithm starts with any arbitrary point p[9] and gets its ε-neighbourhood.[10] If p is a *core* point, a new cluster is created and it includes all points in its neighbourhood. Then, if an additional *core* point is present in the neighbourhood, the search is expanded to include also its neighbours. When no more points are classified as *core* points in the enlarged neighbourhood, the cluster is complete and the remaining points are investigated to possibly identify a *core* point to start a new cluster. If p is a *border* point, so that no points are *density-reachable* from p, the algorithm proceeds visiting the next point in the dataset D. Finally, the set of points in D not belonging to any cluster C_i are classified as *noise*.

Two global input parameters, i.e. equal for all clusters, need to be defined a *priori*: the maximum radius (ε) which defines the area of the neighbourhood of a point p, and the minimum number of points ($MinPts$) in it as density threshold. The choice of these input parameters is heuristic and depends on the context in which the algorithm is implemented, although a few 'rules of thumb' may help. For instance, the density threshold can be set at least equal to the dimensions of the dataset plus 1. However, in case of large datasets, noisy data or presence of duplicates points larger values would be more appropriate. For the choice of the ε-parameter, Ester et al. (1996) propose plotting the kNN distances, i.e. the distance to the kth nearest neighbour, for each point in the dataset in descending order. The desired ε-parameter values are those in correspondence of the 'knee' shown in the plot. Therefore, points inside a cluster are those close to other points in the same cluster with a small k-nearest neighbour distance, while remote noise points have larger kNN distance.

[9]The ordering in which the points are investigated does not matter for the assignment of *core* points to a cluster. However, *DBSCAN* is not entirely deterministic since *border* points which are reachable from more than one cluster can be part of several clusters, and the algorithm assigns them to the first cluster identified. Fortunately, the latter situation is uncommon for most datasets, and it should have a little impact on the algorithm's results (Schubert et al., 2017).

[10]Most algorithm's implementations, available in several software packages and programming languages, use by default the Euclidean distance for the neighbourhood computation, but any other distance function can be chosen.

DBSCAN Implementation Using Firms Location

In our framework, the *DBSCAN* algorithm allowed us to compute localisation of firms within actual distances in kilometres and describe the distribution of sample firms by distance and sector.[11]

We used micro-data coming from the commercial Orbis database, compiled by the Bureau Van Dijk, for Italian firms based in Umbria and Puglia, operating mostly in the manufacturing and the service sector.[12] We exploited the maximum level territorial detail by using the postal address of each firm, upon availability. This allowed translating the postal addresses to geographical coordinates, i.e. latitude and longitude, through geocoding[13] which have been then used to provide insightful spatial information.

We ended up with a dataset containing geocode information for 48,934 firms located in Umbria and 184,206 firms located in Puglia. Next, we combined the latter information with the 'status' (e.g., active, bankruptcy, in liquidation, dissolved, etc.) of each firm reported in the financial statement for the period 2008-2016, to assess the presence of the firms in the market in the time span studied. Finally, we implemented the *DBSCAN* algorithm[14] to identify industrial clusters by industry-year for each region using latitude and longitude data on active firms. As global input parameters, we set $MinPts = 10$, i.e. an industrial cluster must contain at least 10 firms, and $\varepsilon = 5km$. We chose these values heuristically in our computation, after inspecting different values suitable to our firm-level analysis.

[11] Similarly, the *DBSCAN* identification method has been used in BEIS (2017) to investigate the geographical agglomeration of UK companies within three sectors, i.e. Digital-Health, Financial-Services and Processing Industry.

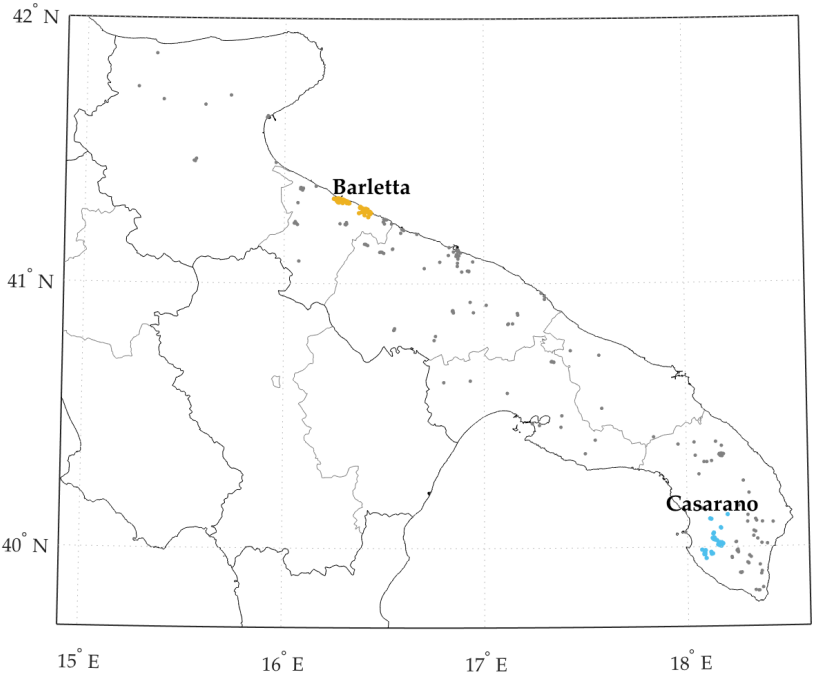
[12] Further details about the sample of firms are provided in Section 4.4.1.

[13] To obtain geographic positioning of our sample of firms, we mainly used the open source geocoding service OpenStreetMap (https://nominatim.openstreetmap.org/search. Last Accessed: 31/03/2018.), and we complemented with the information retrieved from the commercial Google's Geocoding API's (https://developers.google.com/maps/documentation/geocoding/start. Last Accessed: 31/03/2018.) service. For details about geocoding micro-data and usability of spatial information, see (Lennert, 2015).

[14] All the computations using the *DBSCAN* clustering algorithm to our sample have been run on MATLAB software.

In Figure 4.5, we report the localisation of firms active in 2011, in NACE rev. 2 sector 15 (Manufacture of leather and related products), and we show the two industrial clusters as resulting from the implementation of the *DBSCAN*. Interestingly, using this method we detected two footwear districts in Puglia, one around Casarano and the other around Barletta which have been investigated in the literature in the context of globalisation in the recent years (Capestro et al., 2014; Amighini and Rabellotti, 2006).

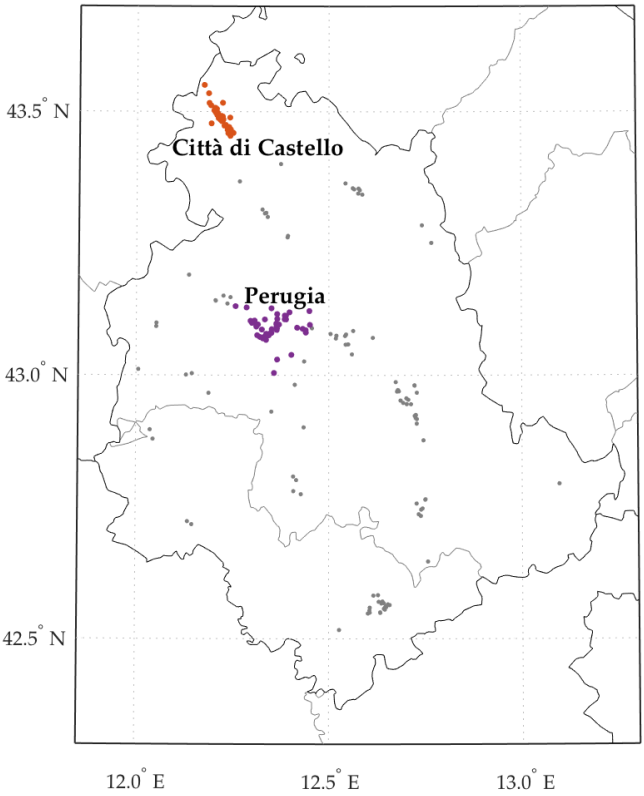
Figure 4.5: Firm localisation in NACE rev. 2 sector 15, year 2011



Source: Own elaboration.
Note: This figure shows the industrial clusters identified implementing the *DBSCAN* algorithm to our sample of firm-level data. Each point represents an active firm in the NUTS-3 Italian region Puglia.

In Figure 4.6, we report the localisation of firms active in 2011, in NACE rev. 2 sector 18 (Printing and reproduction of recorded media) in Umbria. Also, in this case, two industrial clusters have been identified, one in Città di Castello and the other around Perugia. In fact, in these areas, it is well known that there are many firms with a long tradition in the paper processing industry, graphics and printing (ISTAT, 2011).

Figure 4.6: Firm localisation in NACE rev. 2 sector 18, year 2011



Source: Own elaboration.
Note: This figure shows the industrial clusters identified implementing the *DBSCAN* algorithm to our sample firm-level data. Each point represents an active firm in the NUTS-3 Italian region Umbria.

The *DBSCAN* approach for cluster identification is a good alternative to the traditional methods described in Section 4.3. We built industrial clusters from the bottom up from location data of firms. In this way, we introduced flexibility of scale of analysis beyond existing administrative boundary inside each region, and we employed analytical technique from Machine Learning that allowed overcoming the limitations of the qualitative assessment of clusters, often based on case studies. Furthermore, it has been not necessary to know the number of clusters in advance and within each industry, we could find one or more clusters or even none since the *DBSCAN* has a robust approach to outliers. Also, the resulting clusters have an arbitrary shape.

In the following sections, we show how the industrial clustering information can be complemented with firm-level data to analyse the survival probability of firms operating in different sectors, within and outside the cluster environment.

4.4 Survival Analysis of Firms

4.4.1 Data

We used micro-data coming from the commercial Orbis database, compiled by the Bureau Van Dijk, for firms located in Umbria and Puglia, mainly operating in the manufacturing and service sectors. We based our study mostly on private firms' financial information (turnover, number of employees, assets, liabilities, etc.), economic activity (4-digit NACE rev. 2), and dates of entry and exit from the market. Also, we exploited the maximum level territorial detail by using the postal address of each firm, upon availability. Through geocoding, we translated the postal addresses to geographical coordinates, i.e. latitude and longitude, as described in Section 4.3.1.

In Table 4.1, we present the coverage of our sample on the actual population of firms provided by ISTAT for each NUTS-3 (province) in the Italian regions under study in 2015. As known, the highest number of firms is recorded around the biggest urban agglomeration in the regions.

Table 4.1: Sample coverage by NUTS-3 (province) in Umbria and Puglia, year 2015

	Umbria			
	ORBIS sample		**ISTAT**	
	N. firms	**%**	**N. firms**	**%**
Perugia	37,246	74.95%	50,677	76.26%
Terni	12,448	25.05%	15,778	23.74%
Total	49,694	100.00%	66,455	100.00%
	Puglia			
Foggia	31,958	15.52%	35,276	14.16%
Bari	66,848	32.46%	82,908	33.27%
Taranto	25,485	12.38%	29,406	11.80%
Brindisi	19,815	9.62%	22,616	9.08%
Lecce	43,592	21.17%	53,807	21.59%
Barletta-Andria-Trani	18,226	8.85%	25,183	10.11%
Total	**205,924**	**100.00%**	**249,196**	**100.00%**

Note: This table presents the number of active firms in the year 2015 in our sample and as sourced from the ISTAT database, broken by province. The number of firms in the Agriculture industry is not included in the computation.

Table 4.2: Sample coverage by industry aggregates in Umbria and Puglia, year 2015

	Umbria				Puglia			
	ORBIS sample		ISTAT		ORBIS sample		ISTAT	
NACE rev.2 aggregates	**N. firms**	**%**	**N. firms**	**%**	**N. firms**	**%**	**N. firms**	**%**
C: Manufacturing	4,927	11.71%	6,569	9.92%	17,814	10.41%	20,928	8.51%
E: Water supply; sewerage, waste management and remediation activities	114	0.27%	132	0.20%	636	0.37%	691	0.28%
F: Construction	7,311	17.38%	8,007	12.09%	28,872	16.87%	28,625	11.65%
G: Wholesale and retail trade; repair of motor vehicle and motorcycles	12,927	30.73%	16,856	25.46%	64,951	37.96%	81,954	33.34%
H: Transportation and storage	1,178	2.80%	1,722	2.60%	5,966	3.49%	6,584	2.68%
I: Accommodation and food service activities	3,725	8.85%	4,959	7.49%	15,825	9.25%	18,583	7.56%
J: Information and commmunication	1,145	2.72%	1,320	1.99%	3,378	1.97%	3,418	1.39%
K: Financial and insurance activities	1,257	2.99%	1,553	2.35%	3,582	2.09%	4,513	1.84%
L: Real estate activities	2,418	5.75%	3,485	5.26%	4,261	2.49%	5,314	2.16%
M: Professional, scientific and technical activities	1,774	4.22%	10,960	16.55%	5,973	3.49%	39,366	16.02%
N: Administrative and support service activities	1,648	3.92%	1,943	2.93%	5,529	3.23%	5,593	2.28%
P: Education	245	0.58%	433	0.65%	1,197	0.70%	1,396	0.57%
Q: Human health and social work activities	408	0.97%	3,918	5.92%	2,389	1.40%	14,859	6.05%
R: Arts, entertainment and recreation	689	1.64%	864	1.30%	1,115	0.65%	1,542	0.63%
S: Other service activities	2,303	5.47%	3,487	5.27%	9,619	5.62%	12,430	5.06%
Total	**42,069**	**100.00%**	**66,208**	**100.00%**	**171,107**	**100%**	**245,796**	**100.00%**

Note: This table presents the number of active firms in the year 2015 in our sample and as sourced from the ISTAT database, broken by NACE rev. 2 aggregates.

In Table 4.2, we show the distribution of firms across industry aggregates in our sample and in ISTAT, when a comparison was feasible. In line with statistics at the national level, the highest number of firms is involved in the service and trade sectors.

4.4.2 Empirical Strategy

We investigated the survival time of our sample of firms, together with information that describes their individual characteristics, industry and spatial patterns, focusing on agglomeration economies.

We restricted the study to a cohort of firms established between 1st January 2008 and 31st December 2016 and examine their likelihood of

surviving up to 31st December 2016.[15]

First, using the Kaplan and Meier (1958) estimator, we explored the survival rate at different spell lengths by economic activity inside and outside the industrial clusters.[16] We considered as the year of firm's entry in the market, the incorporation date reported in the financial statement which corresponds to the firm's registration as a legal entity. Further, we assumed that a firm exits from the market based on the information contained in the variable 'status' (e.g., bankruptcy, liquidation, dissolved, etc.) and 'status date', and we complemented by checking the financial statement data. Second, we employed the semiparametric Cox (1972) model testing the impact of several covariates on the survival probability of firms.

Several issues were carefully addressed: a right censoring and the discreteness of the dependent variable. As for the latter, although firms exit the market in the continuous time, the database collects information on a yearly basis, i.e. at the end of the fiscal year. Thus, for each firm, we can observe the spell length from its birth year to the end of the year t. This justifies the use of discrete-time hazard models. As for the right censoring of the time variable, information about the firm may be incomplete because the firm does not have an event (i.e. exit from the market) during the time of the study. In fact, for some firms, we may know that the survival is at least equal to time t, whereas for other firms we know their exit time from the market. This happens when the study does not span enough time to observe the event for all the units of analysis in the sample, or firms drop out of the study for unrelated reasons, such as lost to follow-up or withdraw from the analysis.[17]

Unlike standard regression models, survival analysis combines censored

[15] We reduced our initial sample to firms entering the market from 2008 because more financial statement information was available in the Orbis database (Kalemli-Ozcan et al., 2015).

[16] We used a dummy variable which indicates whether a firm in its entry year belongs to an industrial cluster as identified by the *DBSCAN* algorithm and described in Section 4.3.1.

[17] Also, time to an event, often referred to as survival time is always positive and has a skewed distribution. Therefore, standard regression techniques, such as linear regression may be inadequate.

and uncensored observations in estimating model parameters. Usually, the dependent variable is composed of two parts: the time to the event T (i.e., firm's exit), and the event status, which records if the event T has occurred for the k^{th} firm. In this framework, we are interested in the probability that the period of survival is of at least length t. Assuming that intervals are of unit length (e.g., one year), the probability of survival until the end of year a_t, i.e. S_t, is the product of non-experiencing the event T in each of the year up to the t^{th}. Hence, we have:

$$S(t) \equiv S_t = (1-h_1)(1-h_2)...(1-h_{t-1})(1-h_t) = \prod_{j=1}^{t}(1-h_t) \quad (4.1)$$

$$\text{with } 0 \leq S(t) \leq 1, \text{for } t \in [0, \infty)$$

where the discrete time survival function $S(t)$ is written in terms of interval hazard rates (h_t). The latter represents the risk to the event T occurence, hence $h_t = \frac{d_j}{n_j}$, where d_j denotes the number of firms that exit and n_j denotes the number of firms remaining in the cohort. Conversely, the probability of firm's exit within the t^{th} interval $(a_{t-1}, a_t]$, i.e. the discrete-time density function $f(t)$, is the following:

$$\begin{aligned} f(t) &= Pr(a_{t-1} < T \leq a_t) \\ &= S(t-1) - S(t) \\ &= \frac{S(t)}{1-h_t} - S(t) \\ &= \left(\frac{1}{1-h_t} - 1\right) S(t) \\ &= \frac{h_t}{1-h_t} \prod_{k=1}^{t}(1-h_t) \end{aligned} \quad (4.2)$$

$$\text{with } 0 \leq f(t) \leq 1, \text{and } \leq h(t) \leq 1, \text{for } t \in [0, \infty)$$

Hence, the discrete-time density function is the product of the probability of surviving up to the end of interval $t-1$ and the probability of firm's exit the market in the t^{th} interval.

If we assume that every firm follows the same survival function $S(t)$, i.e. we assume that there are non-individual differences $S(t)$ as presented

in Equation 4.1, can easily be estimated by using the non-parametric Kaplan-Meier product-limit estimator (Kaplan and Meier, 1958).

On the contrary, to investigate the effect of several variables upon the time firm's exit takes place, the Cox proportional hazard model Cox (1972) is employed. The method does not assume any form of the survival function, but it is semi-parametric because the effects of the explanatory variables are parametrized to multiplicatively shift the baseline hazard function. Moreover, the results derive from a partial likelihood method of estimation. The hazard rate for the k^{th} firm in the data is:

$$h_k = h_{0c}(t)\exp(\boldsymbol{\beta}' X_i) \tag{4.3}$$

where $h_{0c}(t)$ is the cluster-specific baseline hazard function[18], which corresponds to the overall risk for firms when $X_i = 0$, and it represents the non-parametric part of the model. Moreover, $\exp(\boldsymbol{\beta}' X_i)$ is the relative risk, i.e. a proportionate increase or reduction in risk associated with the set of characteristics X_i. A value of β_i greater (lower) than 0, indicates that as the value of the i^{th} covariate increases, the event hazard increases (decreases), or equivalently the length of survival decreases (increases). The logarithm expression of Equation 4.3 gives an additive linear model for the log of the hazard:

$$\log h_k(t|X_i) = \log h_{0c}(t) + \beta' X_i \tag{4.4}$$

As in all additive model, the effect of the explanatory variables X_i is assumed to be the same at all times t.

We tested the effect of firm-specific variables[19]: *size*, measured by the natural logarithm of turnover, and the square is also introduced to allow

[18] In our study, we assumed the baseline hazard to differ by firms in an industrial cluster and outside it. Nevertheless, the coefficients are the same regardless of the group.

[19] We did not control for firm age, although it is a common practice in the empirical literature of firm survival since we considered a cohort of firms established between 2008 and 2016. Moreover, we were not able to test the effect of productivity measures (e.g., labour productivity and TFP), also relevant to this type of analysis, because we did not have enough information. It is well known that limited liability companies may not report complete financial statement's information in compliance with the national law.

for non-linearities; *number of patents* and *number of trademarks*, which are the stock of granted patents and registered trademarks owned by a firm and they are suitable proxies of firms' innovativeness; financial indicators, i.e. *solvency ratio* and *current ratio*, which measure the firms' ability to pay short-term and long-term obligations. Industry-specific conditions at the 2-digit NACE rev. 2 is captured by the Herfindahl-Hirschman Index (HHI), taken in logarithm form, as an approximation of market competition. As main variables of interest, capturing the geography-firm nexus, we included agglomeration variables computed by NUTS-3 (province) using census data for 2012 provided by ISTAT: diversification externality has been decomposed into *related variety* capturing *within* industry relatedness and *unrelated variety* measuring *between* industry externalities; finally, urbanization externalities are proxied by the *population density*. See Appendix B for more details on all variables and their construction.

4.4.3 Econometric Results

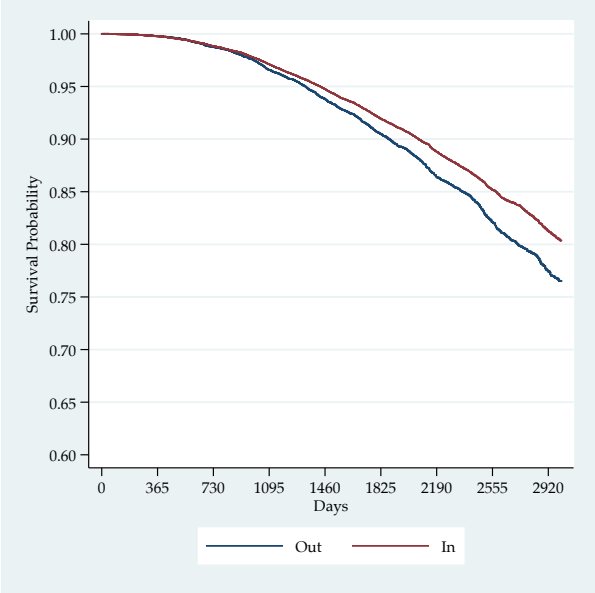
Survival Estimate Kaplan-Meier

In Figure 4.7 we compare the survival probability, estimated by the non-parametric Kaplan and Meier (1958) estimator, for a cohort of firms in an industrial cluster versus firms not belonging to any cluster, at each point in time (days) for the nine years of study.

Crucially, we can see that there is a difference between the two survival probabilities which is also statistically significant as confirmed in Table 4.3, by the tests (Log Rank, Breslow, Tarone-Ware) for equality between the two functions. Considering all firms in our sample, we found that after about nine years, 74% of firms remained active. In particular, in an industrial cluster was still active the 69.30% of firms, while outside any industrial agglomeration only the 63.20% was still operating on the market altogether in Umbria e Puglia at the end of the year 2016. Hence, from the third year firms in industrial clusters seem to have been less exposed to exit from the market than firms outside-cluster.

However, as reported in Figure 4.8, exploring separately the survival

Figure 4.7: Kaplan-Meier survival estimate outside versus inside industrial clusters



Source: Own elaboration.

Table 4.3: Test for equality of survival functions

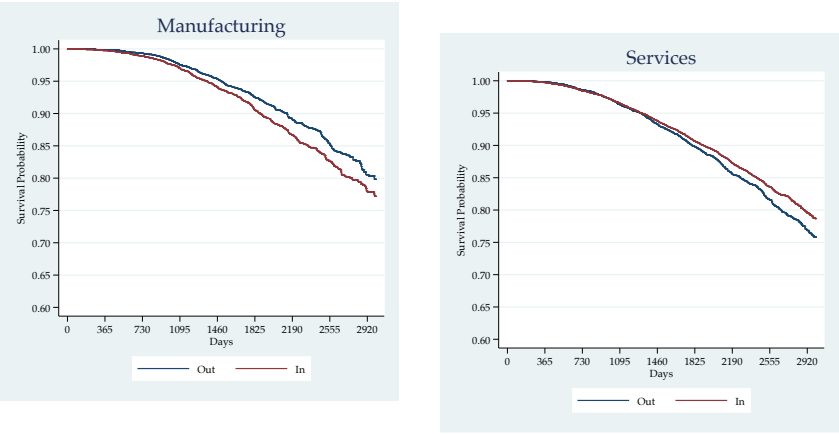
	Log Rank	**Breslow**	**Tarone-Ware**
Outside versus inside industrial clusters	51.1	29.84	40.02
	(0.0000)	(0.0000)	(0.0000)

Note: P-values in parenthesis.

probability for firms operating in the manufacturing and the service sectors, we can provide different conclusions. For firms in the manufacturing, it seems more likely to survive outside a manufacturing cluster than inside it after about one year of activity.

These results are in line with the structural changes occurring at the industrial districts or cluster level in Italy, following the challenges of

Figure 4.8: Kaplan-Meier survival estimate outside versus inside industrial clusters in manufacturing and services



Source: Own elaboration.

globalization and the information and communication technology (ICT) revolution (Iuzzolino and Minucci, 2011). In the new scenario, some manufacturing clusters are experiencing deep crisis, while others are successfully engaging in global value chains which involve global buyers and production partners. Overall, many firms have been unable to survive in the highly competitive global context.[20] On the contrary, we see that service firms, primarily oriented towards local demand, are more likely to survive if they are geographically concentrated. By remaining close to other companies operating in the same sector, service firms can benefit from the pool of skills and skilled workers in the area, which encourages them to stay on the market.

Cox Proportional Hazard Model Results

In Table 4.4, we check which variables could drive differences on firm's hazards, or similarly but with opposite sign, on the survival rate. The

[20]See De Marchi et al. (2017) for a collection of detailed studies investigating local clusters in global value chains.

Table 4.4: Estimation results

	(1)	(2)	(3)	(4)	(5)	(6)
Unrelated variety	-0.088***		-0.079		-0.10***	
	(0.023)		(0.067)		(0.033)	
Related variety		0.098***		0.081		0.11***
		(0.021)		(0.065)		(0.030)
Urbanisation	-36.2***	-34.3***	-26.0*	-23.4	-38.8***	-36.4***
	(5.08)	(5.03)	(15.3)	(15.2)	(7.24)	(7.19)
Urbanisation^2	8.08***	7.65***	5.75*	5.18	8.69***	8.17***
	(1.14)	(1.13)	(3.42)	(3.38)	(1.62)	(1.61)
Firm size	0.29***	0.072**	-0.024	-0.096	0.50***	0.10*
	(0.034)	(0.034)	(0.10)	(0.10)	(0.058)	(0.058)
Firm size^2	-0.0063***	-0.0064***	0.0043	0.0043	-0.0090***	-0.0091***
	(0.0020)	(0.0020)	(0.0059)	(0.0059)	(0.0034)	(0.0034)
Number of trademarks	-0.027	-0.029	-0.70	-0.70	0.17	0.16
	(0.12)	(0.12)	(0.43)	(0.43)	(0.13)	(0.13)
Number of patents	0.047	0.045	0.15***	0.15***	-0.039	-0.042
	(0.061)	(0.062)	(0.029)	(0.029)	(0.11)	(0.11)
Solvency ratio	-0.0087***	-0.0086***	-0.017***	-0.017***	-0.0081***	-0.0081***
	(0.0012)	(0.0012)	(0.0038)	(0.0038)	(0.0015)	(0.0015)
Current ratio	-0.000071**	-0.000072**	-0.0046**	-0.0045**	-0.000032	-0.000032
	(0.000029)	(0.000029)	(0.0022)	(0.0022)	(0.000024)	(0.000024)
HHI	-0.017	-0.018	-0.035	-0.036	-0.12***	-0.12***
	(0.020)	(0.020)	(0.068)	(0.068)	(0.032)	(0.032)
N. observations	31,969	31,969	4,019	4,019	14,592	14,592
Wald test	131.2 (10)	135.6 (10)	50.5 (10)	50.7 (10)	86.3 (10)	90.0 (10)
Log pseudolikelihood	-13862.8	-13860.1	-1186.1	-1186.0	-6038.4	-6036.5
Two-way clustered errors	Yes	Yes	Yes	Yes	Yes	Yes
Activity of firms	All	All	Manufacturing	Manufacturing	Services	Services

Coefficient are reported. Errors clustered by firm and NUTS-3. Estimates are stratified by the dummy Cluster, indicating if a firm belongs or not to an industrial cluster. ***, **, * stand for p-value < 0.01, p-value < 0.05 and p-value < 0.10, respectively.

coefficient and robust standard errors adjusted for clustering at the firm and NUTS-3 level are reported. The first two column shows estimates for the whole sample of firms, while the second and the third columns refer to firms operating in the manufacturing sectors, and the last columns refer to services.

In light of studying the geography-firm nexus, our focus is on spatial agglomeration variables. We observe that *Unrelated variety*, capturing diversification across industries at the province-level, increases the survival rate, in particular for firms in the service sectors; while the variable's coefficient for manufacturing is not significant. The geographic concentration of firms operating in different types of activities has been interpreted in the literature as a portfolio strategy (Frenken et al., 2007) to protect the regional environment against industry-specific shocks. Specifically, the significant effect for firms in service sectors is in line with the local demand orientation of those firms.

Related Variety, which captures the geographic concentration of firms in complementary or related industries, has a negative effect on firms' longevity in services; while the effect on manufacturing firms is not significant. In the literature, the latter kind of variety has been associated to Jacobs-types externalities arising from spillovers (e.g., sharing knowledge, competencies, technology, etc.) between industries which should enhance firm's innovation and production processes. Firms in the service sectors are usually more exposed to local competition since they provide their service locally, therefore diversification within two-digit industries may threaten their survival. Our results are in line with the findings in Basile et al. (2017) and Ferragina and Mazzotta (2014).

As expected, the degree of urbanisation measure by the *population density* shows that when the province population density increases, firms are more likely to survive. Bigger cities usually can provide more incentives and support to business activities, and local demand is higher than in smaller ones. However, we also find a significant non-linear effect which confirms the fact that diseconomies (e.g., congestion, pollution, high factor costs, etc.) may decrease the survival rate in highly inhabited areas.

Focusing on firm-level control variables, we found that the firm size,

measured by the turnover, has a significant and non-linear effect on the survival of firms. It seems that for low increases in the firm size, the hazard rate increases, then from a certain size threshold, it starts decreasing. A possible explanation is the presence of mostly micro and small firms in our sample (as also along the entire Italian peninsula) which are more vulnerable. For instance, only after a certain dimension, they can start operating at an efficient scale and have better access to the capital market.

Contrary to what expected, from our results an increase in the *Number of patents* decreases the survival rate of manufacturing firms. However, the uncertainty associated with innovation especially in high-tech activities *ex ante* can be misleading. In fact, the commercial risk related to a 'new-to-the-world' innovation, is high in the short-run, as discussed in Buddelmeyer et al. (2009), but after a while, some innovations have a clear positive effect on firm survival which is also positively reflected on financial, management and economic capabilities.

The likelihood of firm survival is also affected by its financial situation. Although financial constraints are not readily observable empirically, many variables are commonly computed from the financial statement.[21] From our study, we find that the more firms can meet their long-term financial obligations, as proxied by the *Solvency ratio*, the lower (higher) is the hazard (survival) rate. Also, an increase in *Current ratio*, which translates into reducing the financial uncertainty of a firm in the short-run, increases the likelihood of firm survival.

Finally, at the industry level, a rise in the market concentration in services significantly decreases the hazard rate, as captured by the Herfindahl-Hirschman Index (HHI) coefficient.

4.5 Conclusions

In this contribution, we explored the geography-firm nexus using micro-data. Specifically, we implemented a challenging density-based spatial

[21] See Ferrando et al. (2015) for an encompassing description of several indicators elaborated on the financial statements of the European firms, collected in the financial module of CompNet.

clustering algorithm, the *DBSCAN*, to identify the spatial concentration of firms operating in the same NACE rev. 2 2-digit sector. Starting from the postal address of each active firm, and after a careful process of geocoding which allowed retrieving geographical coordinates, i.e. latitude and longitude, we could identify several industrial clusters in the Italian regions, Umbria and Puglia. Without relying on existing administrative boundaries, we have been able to locate in space industrial clusters and to identify the firms *within* and *outside* them. We argue that nowadays, using a data-driven approach to investigate the patterns of geographical clustering of firms on continuous space is of utmost relevance also thanks to the increasing availability of micro-data. Much richer and detailed spatial information can be used to understand the economic performance of firms and the actual drivers of regional economic development.

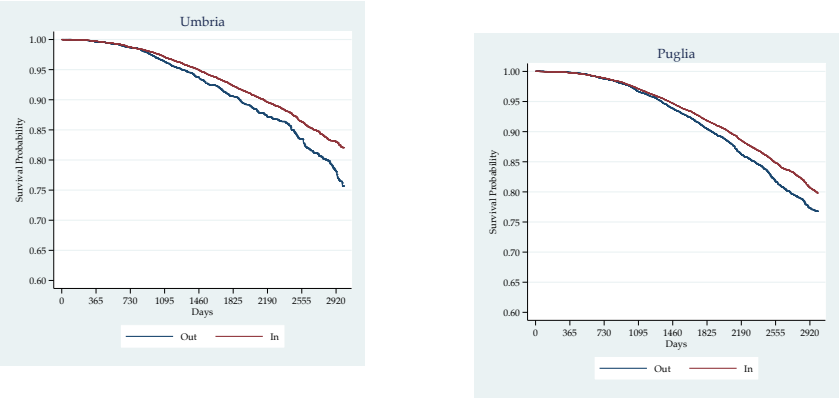
Then, our findings from the survival analysis of firms, show that manufacturing activities are more likely to survive *outside* an industrial cluster, while the contrary is true for firms in services. These results reflect the current trajectory of some manufacturing companies at the cluster level struggling to compete in the global scenario. Besides, we confirm the fading of the 'district effect' (Di Giacinto et al., 2013), especially in Italy, combined with the increasing heterogeneity of firms along many dimensions (e.g., size, performance, productivity, etc.) *within*, *between* and *outside* clusters (De Marchi et al., 2017). On the contrary, services activities seem still to be able to capture value locally, also when in proximity to similar businesses.

More in general, in this study we showed the importance of looking at the geography of firms that have often been neglected in the literature, but which provides additional information when combined with other specific firm and industry characteristics. Finally, we argue that a more completed picture of the local region using firm-level data can help policy-makers when planning industrial policies.

Appendices to Chapter 4

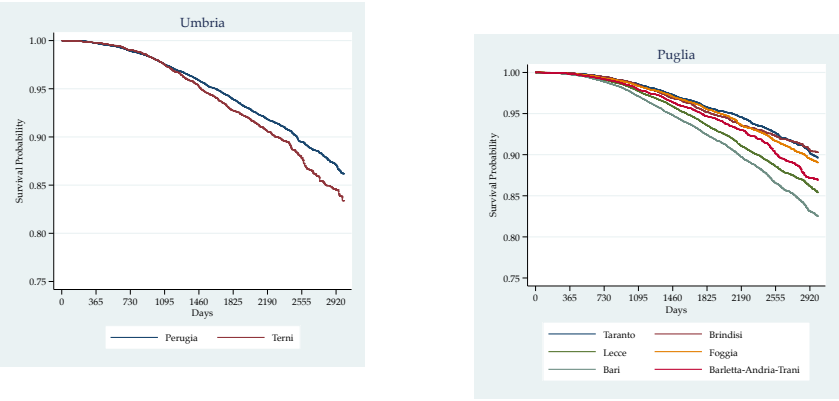
A Appendix: Figures

Figure A.4.1: Kaplan-Meier survival estimate outside versus inside industrial clusters in Umbria and Puglia



Source: Own elaboration.

Figure A.4.2: Kaplan-Meier survival estimate by province in Umbria and Puglia



Source: Own elaboration.

B Appendix: Construction of variables

Firm-specific variables

Firm size: Computed from ORBIS data. It is the (log of) firm-level turnover.

Agglomeration variables

Urbanisation: It is proxied by the *population density*, i.e. the ratio between the number of inhabitants and the area in km^2 of each NUTS-3 (province), as provided by ISTAT. We computed this measure using data for the year 2008, our starting period of investigation. However, there are no significant differences in the following years.

Unrelated variety: Computed using employment data at the NUTS-3 (province) level from ISTAT for the year 2012. This diversification index, proposed in Frenken et al. (2007), measures the extent to which an area is diversified in very different types of activities. It is computed across two-digit industries, based on the assumption that at this level of aggregation they are unrelated to each other. Thus, it aims at capturing *between* industry externalities.

$$\mathrm{UV}_r = \sum_{s=1}^{S} P_{s,r} \log_2 \left(\frac{1}{P_{s,r}} \right) \tag{B.4.1}$$

where $P_{s,r}$ is the employment share of each two-digit sector, $s = 1, .., S$, in a province r.

Related variety: Computed using employment data at the NUTS-3 (province) level from ISTAT for the year 2012. This diversification index, proposed in Frenken et al. (2007), captures *within* industry externalities (degree of relatedness). It is measured as the weighted sum of entropy at disaggregated industry-level within each two-digit industry. The underlying assumption is that industries within this level of aggregation are related and can learn from each other (e.g., knowledge spillover).

$$\mathrm{RV}_r = \sum_{s=1}^{S} P_{s,r} \left(\sum_{i \in S_g} \frac{p_{i,r}}{P_{s,r}} \log_2 \left(\frac{1}{p_{i,r}/P_{s,r}} \right) \right) \tag{B.4.2}$$

where $p_{i,r}$ is the employment share of each five-digit sector i falling exclusively under the two-digit sector S_g.

Industry-specific variables

Herfindahl-Hirschman Index (HHI): Computed from Orbis data. It approximates the level of concentration within the industry.

$$\text{HHI}_s = \sum_{i=1}^{N} m_i^2 \tag{B.4.3}$$

where m is a proxy for the market share measured using firm i's share turnover within an industry s at NACE 2-digit level and N is the number of firms in that industry. The higher the value, the higher the level of industry concentration. Low values of the Herfindahl-Hirschman Index, therefore, characterise a competitive setting.

CHAPTER 5

Conclusions

In this thesis, we shed light on some interesting firm-level features within a global context, contributing to the literature on international trade, industrial organisation and economic geography.

We started analysing network-like production processes and providing new insights on vertical integration choices within the respective theoretical framework. Then, stressing the fact that firms are heterogeneous along many dimensions, we tested the impact of EU regional policy on their productivity. Finally, we looked at the geography-firm nexus exploiting fine-grained spatial information, enriching the empirical literature on firm survival.

Here follows a summary of the main contributions and some considerations on possible future works for each chapter of the thesis.

5.1 Summary

In Chapter 2, we introduced the *Input Rank* as a measure to study the organisation of supply networks at the firm level. We assume that a

Markov process of exploration may be started by a producer throughout her web of direct and indirect suppliers, to assess the technological relevance of each direct and indirect input, when her ability to outreach in the supply network may be limited. Therefore, each producer ends up with an input-output eigenvector centrality, which is higher when a direct or indirect input is relatively more requested to produce other direct or indirect inputs, and when that input is relatively more requested to produce other highly-requested inputs. Finally, we computed the *Input Rank* on U.S input-output tables and tested its empirical validity for choices of vertical integration on a dataset made of 20,489 U.S. parent companies controlling 154,836 affiliates worldwide. Results showed that a higher *Input Rank* is positively associated with a higher probability that the input is vertically integrated, relatively more when the demand faced by the parent company is more elastic. We argue that a producer reduces the risk of disruption in her supply network when a central input is vertically integrated. In this framework, the *Input Rank* is at least complementary to previous sequential metrics (e.g., *upstreamness* or *downstreamness*), because it better catches the recursive nature of real-world supply networks, whereas linear technological sequences may be just corner solutions.

In Chapter 3, we empirically tested the effect of the EU Cohesion Policy using a unique dataset of 273,500 European manufacturing firms, combining regional policy data at NUTS-2 level with firm-level total factor productivities (TFP). In a framework of heterogeneous firms and different absorptive capacity by regions, we show that financing by the European Regional Development Fund (ERDF) aimed at direct investments in R&D is associated with improvement of firms' productivity in a region. The association is stronger in the first quartile of the TFP distribution, i.e. for firms that are least efficient in a region. Conversely, funding designed at overall *Business Support* is not significantly correlated with any improvement in firm-level competitiveness. We finally argue that considering the heterogeneous distribution of firms' inefficiencies within a region is critical for a better design of the Cohesion Policy.

In Chapter 4, we presented a data-driven approach to identify industrial clusters from micro-data. Specifically, we borrow from Machine Learning a density-based spatial clustering algorithm, the *DBSCAN*, and show how its implementation to firms' location allows overcoming some limitations of distance-based methods currently used in the literature to explore the spatial concentration of economic activities. Investigating the survival probability of a cohort of firms from the Italian regions Umbria and Puglia, we found that hazard rate of firms inside an industrial cluster is statistically significant different to the one of firms not belonging to any cluster. Therefore, we argue that although firms compete in a global scenario, it is still of utmost importance to look at the geography-firm nexus to better catch their dynamics.

5.2 Future Work

Here follow some considerations on possible future work for each of the topics covered in the thesis chapters.

In Chapter 2, we highlighted the importance of looking at the network dimensions in the global organisation of production. In line with recent theoretical and empirical works, we argue that the shape of production networks is the outcome of endogenous collective choices by buyers and suppliers, which through input linkages drive both individual and aggregate dynamics. Further research should consider the recursive and complex structure of real-world production networks also from a theoretical perspective to better understand the implications of possible shocks and their propagation, and likewise the organisational response of firms. Then, in the framework of choices of vertical integration, we believe that the proposed *Input Rank* can also be extended to include the geographic dimension, either using firm-to-firm transactions (see, for example, ongoing works in the review by Bernard et al. (2017)) or multi-country input-output tables. The local technological centrality of an input combined with the geographical proximity of production stages can

provide insightful information on the determinants of vertical integration choices, which until now have been considered neutral to country-level factors. Finally, it would be interesting to empirically investigate the evolution of production networks overtime by sector, at the heart of technological progress. In fact, production networks are dynamic by nature both organizationally and geographically. Adjustments are continuously being made in responses to internal and external circumstances.

In Chapter 3, we showed that inefficiencies, as per TFP estimation, are distributed asymmetrically, bimodal and with a long right tail. Therefore, we may expect heterogeneous responses from the implementation of the EU Cohesion Policy, also *within* regions. Future research should overcome the policy data limitations at the firm-level. For the programming period 2014-2020, improvements in the administrative capacity and coordination by public administration in member states and regions is expected. Therefore, data on individual projects, such as expenditure, category, and beneficiaries should be available in a more harmonised way across European regions. Then, this detailed information could be integrated with other firm-level characteristics to evaluate the impact of the Cohesion Policy's tools also on the productivity of individual firms.

The identification method of industrial clusters, i.e. the *DBSCAN* algorithm, detailed in Chapter 4, offers opportunities for empirical exploration of clusters in whatever country or region. Therefore, it would be interesting to extend the research to the entire Italian peninsula. Further, the *DBSCAN* algorithm requires two input parameters: the $MinPts$, i.e. the minimum number of firms to form a cluster and ε the maximum radius that defines the neighbouring area of a firm. In our implementation, we follow the original formulation in Ester et al. (1996), setting global input parameters, i.e. the same for all clusters, after heuristically inspecting several alternatives. However, specific input parameters can be introduced to account for different spatial scales of regions and thus obtain better clustering results.

Bibliography

Daron Acemoglu, Pol Antràs, and Elhanan Helpman. Contracts and Technology Adoption. *The American Economic Review*, 97(3):916–943, 2007. 6, 23

Daron Acemoglu, Simon Johnson, and Todd Mitton. Determinants of Vertical Integration: Financial Development and Contracting Costs. *The Journal of Finance*, 64(3):1251–1290, 2009. 32

Daron Acemoglu, Rachel Griffith, Philippe Aghion, and Fabrizio Zilibotti. Vertical Integration and Technology: Theory and Evidence. *Journal of European Economic Association*, 8(5):989–1033, 2010. 40

Daron Acemoglu, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The Network Origins of Aggregate Fluctuations. *Econometrica*, 80(5): 1977–2016, 2012. 22

Daron Acemoglu, Azarakhsh Malekian, and Asu Ozdaglar. Network Security and Contagion. *Journal of Economic Theory*, (166):536–585, 2016. 22

Daniel Ackerberg, Lanier Benkard, Steven Berry, and Ariel Pakes. Econometric Tools for Analyzing Market Outcomes. *Handbook of Econometrics*, 6:4171–4276, 2007. 86, 91

Daniel Ackerberg, Kevin Caves, and Garth Frazer. Identification Properties of Recent Production Function Estimators. *Econometrica*, 83(6):2411–2451, 2015. 61, 71, 88, 92, 93, 96

Rajshree Agarwal and David B. Audretsch. Does Entry Size Matter? The Impact of the Life Cycle and Technology on Firm Survival. *The Journal of Industrial Economics*, 49(1):21–43, 2001. 104

Rajshree Agarwal and Michael Gort. Firm and Product Life Cycles and Firm Survival. *American Economic Review*, 92(2):184–190, 2002. 101

Philippe Aghion and Richard Holden. Incomplete Contracts and the Theory of the Firm: What have we learned over the past 25 years? *The Journal of Economic Perspectives*, 25(2):181–197, 2011. 23

Dennis Aigner, Knox Lovell, and Peter Schmidt. Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*, 6 (1):21–37, 1977. 93

Laura Alfaro and Andrew Charlton. Intra-industry Foreign Direct Investment. *American Economic Review*, 99(5):2096–2119, 2009. 40

Laura Alfaro, Paola Conconi, Harald Fadinger, and Andrew F. Newman. Do Prices Determine Vertical Integration? *Review of Economic Studies*, 83(3): 855–888, 2016. 32, 40

Laura Alfaro, Pol Antràs, Davin Chor, and Paola Conconi. Internalizing Global Value Chains: A Firm-Level Analysis. *National Bureau of Economic Research, Working Paper No. 21582*, 2017. 6, 16, 18, 21, 23, 24, 32, 40, 42, 44

Carlo Altomonte, Marcella Nicolini, Armando Rungi, and Laura Ogliari. Assessing the Competitive Behaviour of Firms in the Single Market: A Micro-based Approach. Economic papers, European Economy, 2010. 71

Carlo Altomonte, Tommaso Aquilante, and Gianmarco Ottaviano. *The Triggers of Competitiveness: the EFIGE Cross-country Report.* Bruegel Blueprint 17, 2012. 9

Carlo Altomonte, Maria Bas, Matteo Bugamelli, Italo Colantone, Lionel Fontagné, Emanuele Forlani, Thierry Mayer, Filippo di Mauro, Philippe Martin, Giorgio Navaretti, Gianmarco Ottaviano, Maddalena Ronchi, Gianluca Santoni, and Elena Zaurini. Measuring Competitiveness in Europe: Resource Allocation, Granularity and Trade. *Bruegel Blueprint Series Volume XVIV*, 2016. 9

Vanessa Alviarez, Javier Cravino, and Andrei A. Levchenko. The Growth of Multinational Firms in the Great Recession. *Journal of Monetary Economics*, 85:50–64, 2013. 39

Alessia Amighini and Roberta Rabellotti. How do Italian Footwear Industrial Districts Face Globalization? *European Planning Studies*, 14(4):485–502, 2006. 115

Dan Andrews, Chiara Criscuolo, and Peter N. Gal. The Best versus the Rest. The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy. *OECD Productivity Working Papers, No. 5*, 2016. 73

Pol Antràs. *Global production: Firms, Contracts, and Trade Structure*. Princeton University Press, 2015. 3

Pol Antràs and Davin Chor. Organizing the Global Value Chain. *Econometrica*, 81(6):2127–2204, 2013. 4, 6, 7, 16, 17, 19, 21, 23, 24, 32, 38, 40, 44

Pol Antràs and Davin Chor. On the Measurement of Upstreamness and Downstreamness in Global Value Chains. *NBER Working Paper No. 24185*, 2018. 6, 16

Pol Antràs and Alonso de Gortari. On the Geography of Global Value Chains. *NBER Working Paper No. 23456*, 2017. 16, 24

Pol Antràs and Stephen R. Yeaple. *Multinational Firms and the Structure of International Trade*, volume 4. 2014. 6, 23

Pol Antràs, Davin Chor, Thibault Fally, and Russell Hillberry. Measuring the Upstreamness of Production and Trade Flows. *The American Economic Review*, 102(3):412–416, 2012. 6, 16, 32

Giuseppe Arbia. Modelling the Geography of Economic Activities on a Continuous Space. *Papers in Regional Science*, 80(4):411–424, 2001. 11, 108

David B. Audretsch and Talat. Mahmood. New-Firm Survival: New Results Using a Hazard Function. *Review of Economics and Statistics*, 77(1):97–103, 1995. 105

David B. Audretsch, Erik E. Lehmann, and Susanne Warning. University Spillovers and New Firm Location. *Research Policy*, 34(7):1113–1122, 2005. 108

Richard Baldwin. Globalisation: the great unbundling (s). *Economic Council of Finland*, 20(3):5–47, 2006. 2, 16

Richard Baldwin and Javier Lopez-Gonzalez. Supply-chain Trade: A Portrait of Global Patterns and Several Testable Hypotheses. *The World Economy*, 38 (11):1682–1721, 2015. 2

Richard Baldwin and Anthony J Venables. Spiders and Snakes: Offshoring and Agglomeration in the Global Economy. *Journal of International Economics*, 90(2):245–254, 2013. 4, 16, 22

Eric J. Bartelsman and Zoltan Wolf. Measuring Productivity Dispersion. *Tinbergen Institute Discussion Paper*, 2017. 85

Eric J Bartelsman, John Haltiwanger, and Stefano Scarpetta. Cross-Country Differences in Productivity: The Role of Allocation and Selection. *American Economic Review*, 103(1):305–34, 2013. 72

Roberto Basile, Rosanna Pittiglio, and Filippo Reganati. Do Agglomeration Externalities Affect Firm Survival? *Regional Studies*, 51(4):548–562, 2017. 11, 101, 105, 106, 127

Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk. *Scientific Reports*, 2(541), 2012. 20, 28

Catherine Beaudry and Andrea Schiffauerova. Who's right, Marshall or Jacobs? The Localization versus Urbanization Debate. *Research policy*, 38(2):318–337, 2009. 11, 71, 101

Giacomo. Becattini. Dal Settore Industriale al Distretto Industriale. Alcune Considerazioni Sull'Unità d'Indagine dell'Economia Industriale. *Rivista di Economia e Politica Industriale*, 1(1):7–21, 1979. 102

Giacomo Beccatini. Small Firms and Industrial Districts: The Experience of Italy. *Economia Internazionale*, 39(2-3-4):98–103, 1986. 100

BEIS. Industrial Clusters in England. *Department for Business, Energy and Industrial Strategy. Research Paper No. 4*, 2017. 114

Fredrik Bergström. Capital Subsidies and the Performance of Firms. *Small Business Economics*, 14(3):183–193, 2000. 64

Andrew B. Bernard and Andreas Moxnes. Networks and Trade. *Annual Review of Economics*, 10(1), 2018. 22

Andrew B. Bernard, Andreas Moxnes, and Yukiko U. Saito. Production Networks, Geography and Firm Performance. *Forthcoming on Journal of Political Economy*, 2017. 137

Cristina Bernini and Guido Pellegrini. How are growth and productivity in private firms affected by public subsidy? Evidence from a regional policy. *Regional Science and Urban Economics*, 41(3):253–265, 2011. 63

Pierre Blanchard, Emmanuel Dhyne, Catherine Fuss, and Claude Mathieu. Not So Easy Come, Still Easy Go? Footloose Multinationals Revisited. *The World Economy*, 39(5):679–707, 2016. 104

Ina Blind, Matz Dahlberg, and Gustav Engström. Introduction to the Special Issue 'On the Use of Geo-Coded Data in Economic Research'. *CESifo Economic Studies*, 64(2):123–126, 2018. 11

BLS. Industries at a Glance. Technical report, Bureau of Labor Statistics, 2018. 35

Richard Blundell and Stephen Bond. GMM Estimation with Persistent Panel Data: an Application to Production Functions. *Econometric Reviews*, 19(3): 321–340, 2000. 88

Rafael Boix, Josè Luis Hervas-Oliver, and Blanca De Miguel-Molina. Micro-geographies of creative industries clusters in Europe: From hot spots to assemblages. *Papers in Regional Science*, 94(4):753–772, 2015. 11

Michele Boldrin and Fabio Canova. Inequality and Convergence in Europe's Regions: Reconsidering European Regional Policies. *Economic policy*, 16(32): 206–253, 2001. 63

Ron A. Boschma and Rik. Wenting. The Spatial Evolution of the British Automobile Industry: Does Location Matter? *Industrial and Corporate Change*, 16(2):213–238, 2007. 105

Anthony Briant, P-P. Combes, and Miren. Lafourcade. Dots to Boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations? *Journal of Urban Economics*, 67(3):287–302, 2010. 107

Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 1998. 5, 13, 20, 23, 28, 36, 49

Christian Broda and David E Weinstein. Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585, 2006. 40, 42

Hielke Buddelmeyer, Paul H. Jensen, and Elizabeth Webster. Innovation and the Determinants of Company Survival. *Oxford Economic Papers*, 62(2):261–285, 2009. 128

Guido Buenstorf. Evolution on the Shoulders of Giants: Entrepreneurship and Firm Survival in the German Laser Industry. *Review of Industrial Organization*, 30(3):179–202, 2007. 101

Philip Bunn and Victoria Redwood. Accounts-Based Modelling of Business Failures and the Implications for Financial Stability. *Bank of England. Working Paper No. 210*, 2003. 107

Martijn J. Burger, Frank G. van Oort, and Bert. van der Knaap. A Treatise on the Scale-dependence of Agglomeration Externalities and the MAUP. *Scienze Regionali*, 9(1):19–40, 2010. 107

Giulio Cainelli, Sandro Montresor, and Giuseppe Vittucci Marzetti. Spatial Agglomeration and Firm Exit: A Spatial Dynamic Analysis for Italian Provinces. *Small Business Economics*, 43(1):213–218, 2014. 101, 106

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2): 238–249, 2011. 79

Uwe Cantner, Kristina Dreßler, and Jens J. Krüger. Firm Survival in the German Automobile Industry. *Empirica*, 33(1):49–60, 2006. 101

Mauro Capestro, Elisabetta Tarantino, Fabrizio Morgangni, Eleonora Tricarico, and Gianluigi Guido. Distretti Calzaturieri in Crisi: Cause del Declino e Strategie di Rinnovamento. *Economia e Società Regionale*, 32(1):187–212, 2014. 115

Aadne Cappelen, Fulvio Castellacci, Jan Fagerberg, and Bart Verspagen. The Impact of EU Regional Support on Growth and Convergence in the European Union. *Journal of Common Market Studies*, 41(4):621–644, 2003. 63

Martin A. Carree, Ingrid Verheul, and Enrico. Santarelli. Sectoral Patterns of Firm Exit in Italian Provinces. *Journal of Evolutionary Economics*, 21(3): 499–517, 2011. 106

Vasco M Carvalho. From Micro to Macro via Production Networks. *The Journal of Economic Perspectives*, 28(4):23–47, 2014. 22, 32

Gary Chamberlain. Analysis of Covariance with Qualitative Data. *Review of Economic Studies*, 47(1):225–238, 1980. 40

Antonio Ciccone and Robert E. Hall. Productivity and the Density of Economic Activity. *American Economic Review*, 86(1):54–70, 1996. 12, 101

Ronald H. Coase. The Nature of the Firm. *Economica*, 4(16):386–405, 1937. 6

Neil M. Coe, Peter Dicken, and Martin Hess. Global Production Networks: Realizing the Potential. *Journal of economic geography*, 8(3):271–295, 2008. 3, 5, 10

Massimo G. Colombo and Marco. Delmastro. A Note on the Relation Between Size, Ownership Status and Plant's Closure: Sunk Costs vs. Strategic Size Liability. *Economics Letters*, 69(3):421–427, 2000. 104

Pierre-Philippe Combes, Thierry Mayer, and Jacques-François Thisse. *Economic Geography: The Integration of Regions and Nations*. Princeton University Press, 2008. 109

Pierre-Philippe Combes, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux. The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection. *Econometrica*, 80(6):2543–2594, 2012. 11, 101

Arnaud Costinot, Jonathan Vogel, and Su Wang. An Elementary Theory of Global Supply Chains. *Review of Economic Studies*, 80(1):109–144, 2012. 6, 16, 23

David R. Cox. Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*, 34(1):187–220, 1972. 103, 120, 122

Javier Cravino and Andrei A. Levchenko. Multinational Firms and International Business Cycle Transmission. *The Quarterly Journal of Economics*, 132(2): 921–962, 2017. 39

Sandy Dall'Erba and Julie Le Gallo. Regional Covergence and the Impact of European Structural Funds over 1989–1999: A Spatial Econometric Analysis. *Papers in Regional Science*, 87(2):219–244, 2008. 63

Jan De Loecker. Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity. *Econometrica*, 79(5): 1407–1451, 2011. 87, 92

Valentina De Marchi, Eleonora Di Maria, and Gary Gereffi. *Local Clusters in Global Value Chains: Linking Actors and Territories through Manufacturing and Innovation.* Routledge, 2017. 100, 125, 129

Dakshina G. De Silva and Robert P. McComb. Geographic Concentration and High Tech Firm Survival. *Regional Science and Urban Economics*, 42(4): 691–701, 2012. 105

Jacob J. De Vries, Peter Nijkamp, and Piet Rietveld. Exponential or Power Distance-decay for Commuting? An Alternative Specification. *Environment and Planning A*, 41(2):461–480, 2009. 108

Matthijs De Zwann and Bruno Merlevede. Regional Policy and Firm Productivity. *ETSG 2013*, 377, 2013. 64

Davide Del Prete and Armando Rungi. Organizing the Global Value Chain: a firm-level test. *Journal of International Economics*, 109:16–30, 2017. 21, 24, 38, 39, 40, 44

Valter Di Giacinto, Matteo Gomellini, Giacinto Micucci, and Marcello Pagnini. Mapping Local Productivity Advantages in Italy: Industrial Districts, Cities or Both? *Journal of Economic Geography*, 14(2):365–394, 2013. 129

Peter Dicken, Philip F. Kelly, Kris Olds, and Henry Wai-Chung Yeung. Chains and Networks, Territories and Scales: Towards a Relational Framework for Analysing the Global Economy. *Global Networks*, 1(2):89–112, 2001. 3

Richard Disney, Jonathan Haskel, and Ylva Heden. Entry, Exit and Establishment Survival in UK Manufacturing. *The Journal of Industrial Economics*, 51(1):91–112, 2003. 101

Joshua Drucker and Edward Feser. Regional Industrial Structure and Agglomeration Economies: An Analysis of Productivity in Three Manufacturing Industries. *Regional Science and Urban Economics*, 42(1–2):1–14, 2012. 108

Timothy Dunne, Mark J. Roberts, and Larry. Samuelson. The Growth and Failure of U.S. Manufacturing Plants. *Quarterly Journal of Economics*, 104 (4):671–698, 1989. 104

Gilles Duranton and Henry G.. Overman. Testing for Localization Using Micro-geographic Data. *The Review of Economic Studies*, 72(4):1077–1106, 2005. 108

Gilles Duranton and Diego Puga. Micro-foundations of Urban Agglomeration Economies. In *Handbook of Regional and Urban Economics*, pages 2063–2117, 2004. 11, 101

EC. Strengthening Good Governance and Administrative Capacity for Cohesion Policy. Technical report, European Commisssion, 2009. 13, 60

ECB. Corporate Finance and Economic Activity in the Euro Area. Structural Issues Report 2013. *European Central Bank, Occasional Paper Series No. 151*, 2013. 106

Sjef Ederveen, Joeri Gorter, Ruud De Mooij, and Richard Nahuis. Funds and Games: the Economics of European Cohesion Policy. Technical report, European Network of Economic Policy Research Institutes, 2003. 63

Glenn Ellison and Edward L. Glaeser. Geographic Concentration in US Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy*, 105(5):889–927, 1997. 107

Marcela Eslava, John Haltiwanger, Adriana Kugler, and Maurice Kugler. The Effects of Structural Reforms on Productivity and Profitability Enhancing Reallocation: Evidence from Colombia. *Journal of Development Economics*, 75(2):333–371, 2004. 87

Azim Essaji. Technical Regulations and Specialization in International Trade. *Journal of International Economics*, 76(2):166–176, 2008. 77

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD*, 96(34):226–231, 1996. 109, 110, 113, 138

European Commission. *Europe 2020. A strategy for smart, sustainable and inclusive growth. Communication from the Commission.* COM (2010) 2020 final, 2010. 60

European Commission. 7th Report on Economic, Social and Territorial Cohesion. Technical report, European Commission, 2017. 60

David. Evans. The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries. *The Journal of Industrial Economics*, 35 (4):567–581, 1987. 104

Thibault Fally. Production staging: measurement and facts. *Boulder, Colorado, University of Colorado Boulder, May*, pages 155–168, 2012. 6, 16

Thibault Fally and Russell Hillberry. A Coasian Model of International Production Chains. *NBER Working Paper No. 21520*, 2015. 23

Michael James Farrell. The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3):253–290, 1957. 93

Anna M. Ferragina and Fernanda Mazzotta. Agglomeration Economies in Italy: Impact on Heterogeneous Firms' Exit in a Multilevel Framework. *Economia e Politica Industriale*, 42(4):395–440, 2015. 101, 104, 105

Anna Maria Ferragina and Fernanda Mazzotta. Local Agglomeration Economies: What Impact on Multinational and National Italian Firms' Survival? *Procedia-Social and Behavioral Sciences*, 110(2014):8–19, 2014. 127

Annalisa Ferrando, Matteo Iudice, Carlo Altomonte, Sven Blank, Marie-Hélène Felt, Philipp Meinen, Neugebauer Katja, and Iulia Siedschlag. Assessing the financial and financing conditions of firms in europe: the financial module in compnet. *European Central Bank, Working Paper No. 1836*, 2015. 128

Lucia Foster, John Haltiwanger, and Chad Syverson. Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review*, 98(1):394–425, 2008. 87

Lucia Foster, Cheryl Grim, John Haltiwanger, and Zoltan Wolf. Firm-Level Dispersion in Productivity: Is the Devil in the Details? *American Economic Review*, 106(5):95–98, 2016. 65

Helen Fotopoulod, Georgios; Louti. Determinants of Hazard Confronting New Entry: Does Financial Structure Matter? *Review of Industrial Organization*, 17(3):285–300, 2000. 107

Linton C. Freeman, Douglas Roeder, and Robert R. Mulholland. Centrality in Social Networks. Experimental results. *Social networks*, 2(2):119–141, 1979. 5

Koen Frenken, Frank Van Oort, and Thijs Verburg. Related variety, Unrelated variety and Regional economic growth. *Regional Studies*, 41(5):685–697, 2007. 103, 127, 132

Thomas MJ Fruchterman and Edward M. Reingold. Graph Drawing by Force-Directed Placement. *Software – Practice and Experience*, 21(11):1129–1164, 1991. 17, 18

Xavier Gabaix. Power Laws in Economics and Finance. *Annual Review of Economics*, 1(1):225–293, 2009. 8, 61, 72

Xavier Gabaix. Power Laws in Economics: An Introduction. *Journal of Economic Perspectives*, 30(1):185–206, 2016. 8, 72

Peter N. Gal. Measuring Total Factor Productivity at the Firm Level using OECD-ORBIS. *OECD Publishing*, 2013. 66

James E. Gentle. *Numerical Linear Algebra for Applications in Statistics*. Springer Publishing, 1998. 30

Robert P. Gilles. The Cooperative Game Theory of Networks and Hierarchies. *Springer Science and Business Media*, 44, 2010. 27

Edward L Glaeser, Hedi D Kallal, Jose A Scheinkman, and Andrei Shleifer. Growth in Cities. *Journal of Political Economy*, 100(6):1126–1152, 1992. 100

David F. Gleich. PageRank Beyond the Web. *Society for Industrial and Applied Mathematics – SIAM Rev.*, 57(3):321–363, 2015. 13, 20, 23, 28, 30, 49

Gita Gopinath, Şebnem Kalemli-Özcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. Capital Allocation and Productivity in South Europe. *The Quarterly Journal of Economics*, 132(4):1915–1967, 2017. 66

Holger Gorg and Eric Strobl. Footloose Multinationals? *The Manchester School*, 71(1):1–19, 2003. 104, 105

Zvi Griliches and Jacques Maitresse. Production Functions: The Search for Identification. *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, 1998. 87

Zvi Griliches and Vidar Ringstad. *Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data*. North-Holland Amsterdam, 1971. 87

Sanford J. Grossman and Oliver D. Hart. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, 94(4):691–719, 1986. 6, 23

Bronwyn Hall. The Relationship Between Firm Size and Firm Growth in the US Manufacturing Sector. *Journal of Industrial Economics*, 35(1):583–606, 1987. 104

Robert E. Hall. The Relation between Price and Marginal Cost in US Industry. *Journal of Political Economy*, 96(5):921–947, 1988. 87

Alfred Hamerle and Gerd Ronning. *Panel Analysis for Qualitative Variables*, pages 401–451. New York: Plenum, 1995. 40

Philipp Harms, Oliver Lorz, and Dieter Urban. Offshoring along the Production Chain. *Canadian Journal of Economics/Revue canadienne d'économique*, 45 (1):93–106, 2012. 6, 23

Richard Harris and Mary Trainor. Capital Subsidies and their Impact on Total Factor Productivity: Firm-level Evidence from Northern Ireland. *Journal of Regional Science*, 45(1):49–74, 2005. 64

Jelena Hartsenko and Ako Sauga. Does Financial Support from the EU Structural Funds has an Impact on the Firms' Performance: Evidence from Estonia. In *Proceedings of 30th International Conference Mathematical Methods in Economics*, pages 260–65, 2012. 64

Taher H. Haveliwala. Topic-Sensitive Pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002. 23, 49

Keith Head, John Ries, and Deborah Swenson. Agglomeration Benefits and Location Choice: Evidence from Japanese Manufacturing Investments in the United States. *Journal of International Economics*, 38(3):223–247, 1995. 40

James J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979. 61, 77, 79

Vernon Henderson, Ari Kuncoro, and Matt Turner. Industrial Development in Cities·. *Journal of Political Economy*, 103(5):1067–1090, 1995. 96, 97

Irving Hoch. Estimation of Production Function Parameters and Testing for Efficiency. *Econometrica*, 23(3):325–326, 1955. 87

Irving Hoch. Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data. *Econometrica*, pages 34–53, 1962. 87

David W. Hosmer, Stanley Lemeshow, and Rodney Sturdivant. *Applied Logistic Regression*. John Wiley and Sons, 2013. 40

Chang Tai Hsieh and Peter J Klenow. Misallocation and Manufacturing TFP in China and India. *The Quarterly Journal of Economics*, 124(4):1403–1448, 2009. 64

Ray Hudson. Conceptualizing Economies and their Geographies: Spaces, Flows and Circuits. *Progress in Human Geography*, 28(4):447–471, 2004. 3

Charles R. Hulten. *Total Factor Productivity: A Short Biography*, pages 1–54. University of Chicago Press, 2001. 8

David Hummels, Jun Ishii, and Kei-Mu Yi. The Nature and Growth of Vertical Specialization in World Trade. *Journal of international Economics*, 54(1): 75–96, 2001. 2

ISTAT. I Distretti Industriali 2011. 9 Censimento dell'Industria e dei Servizi e Censimento delle Istituzioni Non Profit. Technical report, ISTAT, 2011. 116

Giovanni Iuzzolino and Giacinto Minucci. Le Recenti Trasformazioni nei Distretti Industriali Italiani. *Osservatorio Nazionale Distretti Italiani, II Rapporto, Roma.*, 2011. 125

Jane Jacobs. *The Economy of Cities*. Vintage Books, 1969. 12, 101

Boyan Jovanovic. Selection and Evolution of Industry. *Econometrica*, 50(3): 649–670, 1982. 104

Sebnem Kalemli-Ozcan, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. How to Construct Nationally Representative Firm Level data from the ORBIS Global Database. Working paper, NBER, September 2015. 66, 67, 78, 120

Sepandar Kamvar, Taher Haveliwala, Christopher Manning, and Gene Golub. Exploiting the Block Structure of the Web for Computing PageRank. Technical report, Stanford InfoLab, 2003. 30

Edward L. Kaplan and Paul. Meier. Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*, 53(1282): 457–481, 1958. 103, 120, 122, 123

Hajime Katayama, Shihua Lu, and James R Tybout. Firm-level Productivity Studies: Illusions and a Solution. *International Journal of Industrial Organization*, 27(3):403–413, 2009. 8

Leo Katz. A New Status Index Derived from Sociometric Analysis. *Psychometrika*, 18(1):39–43, 1953. 5, 49

Hiau Looi Kee, Alessandro Nicita, and Marcelo Olarreaga. Import Demand Elasticities and Trade Distortions. *Review of Economics and Statistics*, 90(4): 666–682, 2008. 77

Martin Kennedy and Richard Florida. *Locating Global Advantage: Industry Dynamics in International Economy*. Stanford University Press, 2004. 2

Lawrence Robert Klein. *Textbook of Econometrics*. Prentice Hall, 1953. 87

Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999. 5

Steven Klepper. Firm Survival and the Evolution of Oligopoly. *RAND Journal of Economics*, 33(1):37–61, 2002. 101

Tor Jakob Klette and Zvi Griliches. The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous. *Journal of Applied Econometrics*, 11(4):343–361, 1996. 87

Paul Krugman. Increasing Returns and Economic Geography. *Journal of Political Economy*, 99(3):483–499, 1991. 59, 100

Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond. The Science of Search Engine Rankings*. Princeton University Press, 2011. 30

Moritz Lennert. The Use of Exhaustive Micro-Data Firm Databases for Economic Geography: The Issues of Geocoding and Usability in the Case of the Amadeus Database. *SPRS International Journal of Geo-Information*, 4(1):62–86, 2015. 10, 114

Robert Leonardi. Cohesion in the European Union. *Regional Studies*, 40(2): 155–166, 2006. 63

James Levinsohn and Amil Petrin. Estimating Production Functions Using Inputs to Control for Unobservables. *The Review of Economic Studies*, 70(2): 317–341, 2003. 91, 93

Jan De Loecker and Pinelopi Koujianou Goldberg. Firm Performance in a Global Market. *Annual Review of Economics*, 6(1):201–227, 2014. 87

Erzo GJ. Luttmer. Models of Growth and Firm Heterogeneity. *Annual Review of Economics*, 2(1):547–576, 2010. 8, 61, 72

Eric Marcon and Florence. Puech. Measures of the Geographic Concentration of Industries: Improving Distance-based Methods. *Journal of Economic Geography*, 10(5):745–762, 2010. 108

Jacob Marschak and William H. Andrews. Random Simultaneous Equations and the Theory of Production. *Econometrica*, 12(3/4):143–205, 1944. 86

Alfred Marshall. *Local clusters in global value chains: linking actors and territories through manufacturing and innovation*. London: Macmillan, 1890. 100, 101

Jose Mata and Pedro. Portugal. Life Duration of New Firms. *The Journal of Industrial Economics*, 42(3):227–245, 1994. 104, 105

Thierry Mayer and Gianmarco Ottaviano. The Happy Few: The Internationalisation of European Firms. *Intereconomics*, 43(3):135–148, 2008. 1, 9, 65

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics. Academic Press New York*, 1974. 40

Wim Meeusen and Julien van Den Broeck. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review*, 18(2):435–444, 1977. 93

Ronald E. Miller and Umed Temurshoev. Output Upstreamness and Input Downstreamness of Industries/Countries in World Production. *International Regional Science Review*, 40(5):443–475, 2017. 6, 16

Yair Mundlak. Empirical Production Function Free of Management Bias. *American Journal of Agricultural Economics*, 43(1):44–56, 1961. 87

Yair Mundlak. Estimation of Production and Behavioral Functions From a Combination of Time Series and Cross Section Data, 1963. 87

Yair Mundlak and Irving Hoch. Consequences of Alternative Specifications in Estimation of Cobb-Douglas Production Functions. *Econometrica: Journal of the Econometric Society*, pages 814–828, 1965. 87

Frank M.H. Neffke, Martin Henning, and Ron Boschma. The Impact of Aging and Technological Relatedness on Agglomeration Externalities: A Survival Analysis. *Journal of Economic Geography*, 12(2):485–517, 2012. 105

Nathan Nunn. Relationship-Specificity, Incomplete Contracts, and the Pattern of Trade. *The Quarterly Journal of Economics*, 122(2):569–600, 2007. 31, 36, 40, 42, 51

Nathan Nunn and Daniel Trefler. The Boundaries of the Multinational Firm: An Empirical Analysis. *The organization of firms in a global economy*, pages 55–83, 2008. 31

Nathan Nunn and Daniel Trefler. Incomplete Contracts and the Boundaries of the Multinational Firm. *Journal of Economic Behavior and Organization*, 94: 330–334, 2013. 36, 51

Ezra Oberfield. A Theory of Input-Output Architecture. *Econometrica*, 86(2): 559–589, 2018. 22

OECD. Guidelines for Multinational Enterprises. Technical report, OECD, 2005. 39

OECD. The Impact of the Global Crisis on SME and Entrepreneurship Financing and Policy Responses. Technical report, OECD, 2009. 106

OECD. Interconnected Economies. Benefiting from Global Value Chains. Technical report, OECD Publishing, Paris, 2013. 2

G. Steven Olley and Ariel Pakes. The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6):1263–1297, 1996. 86, 88, 93

Stan Openshaw. Ecological Fallacies and the Analysis of Areal Census Data. *Environment and planning A*, 16(1):17–31, 1984. 11, 107

Carmine Ornaghi. Assessing the Effects of Measurement Errors on the Estimation of Production Functions. *Journal of Applied Econometrics*, 21(6):879–891, 2006. 87

Gianmarco Ottaviano and Diego Puga. Agglomeration in the Global Economy: A Survey of the New Economic Geography. *The World Economy*, 21(6): 707–731, 1998. 100

Michael J. Piore and Charles F. Sabel. *The Second Industrial Divide: Possibilities for Prosperity.* New York: Basic Books, 1984. 100

Michael Porter. Clusters and the New Economics of Competition. *Harvard Business Review*, 76(6):77–90, 1986. 100

Michael Porter. *The Competitive Advantage of Nations.* New York: Free Press, 1990. 8, 100

Roberta Rabellotti and Elisa Giuliani. Italian Industrial Districts Today: Between Decline and Openness to Global Value Chains. In *Local Clusters in Global Value Chains*, pages 35–46, 2017. 102, 106

Roberta Rabellotti and Hubert. Schmitz. The Internal Heterogeneity of Industrial Districts in Italy, Brazil and Mexico. *Regional Studies*, 33(2):97–108, 1999. 107

James E. Rauch. Networks versus Markets in International Trade. *Journal of International Economics*, 48(1):7–35, 1999. 31, 36, 51

Samuel Pinto Ribeiro, Stefano Menghinello, and Koen De Backer. The OECD ORBIS Database: Responding to the Need for Firm-level Micro-data in the OECD. *OECD Statistics Working Papers*, 2010. 67

Stuart S. Rosenthal and William C. Strange. Evidence on the Nature and Sources of Agglomeration Economies. In *Handbook of Regional and Urban Economics*, pages 2119–2171, 2004. 11, 101

Jan Ruffner and Andrin Spescha. The Impact of Clustering on Firm Innovation. *CESifo Economic Studies*, 64(2):176–215, 2018. 11

Armando Rungi, Gregory Morrison, and Fabio Pammolli. Global Ownership and Corporate Control Networks. *IMT Lucca EIC WP Series 07/2017*, 2017. 39

Tobias Scholl and Thomas Brenner. Detecting Spatial Clustering Using a Firm-level Cluster Index. *Regional Studies*, 50(6):1054–1068, 2016. 108

Erich Schubert, Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems TODS*, 42(3):1–21, 2017. 113

Agusti Segarra and Maria Callejón. New firms' Survival and Market Turbulence: New Evidence from Spain. *Review of Industrial Organization*, 20(1):1–14, 2002. 104

Anastasia Semykina and Jeffrey M. Wooldridge. Estimating Panel Data Models in the Presence of Endogeneity and Selection. *Journal of Econometrics*, 157 (2):375–380, 2010. 61, 77, 79, 81

Beata Smarzynska Javorcik. Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers Through Backward Linkages. *American Economic Review*, 94(3):605–627, 2004. 93

Robert M. Solow. Technical Change and the Aggregate Production Function. *The review of Economics and Statistics*, pages 312–320, 1957. 85, 87

Michael Storper. *Keys to the city: How economics, institutions, social interaction, and politics shape development.* Princeton University Press, 2013. 99

Harald. Strotmann. Entrepreneurial Survival. *Small Business Economics*, 28(1): 87–104, 2007. 105

Chad Syverson. Market Structure and Productivity: A Concrete Example. *Journal of Political Economy*, 112(6):1181–1222, 2004a. 64

Chad Syverson. Product Substitutability and Productivity Dispersion. *The Review of Economics and Statistics*, 86(2):534–550, 2004b. 64

Chad Syverson. What Determines Productivity? *Journal of Economic Literature*, 49(2):326–365, 2011. 8, 64

Peter Thompson. Selection and Firm Survival: Evidence from the Shipbuilding Industry . *Review of Economics and Statistics*, 87(1):26–36, 2005. 101

Marcel P. Timmer, Bart Los, Robert Stehrer, and Gaaitzen J. de Vries. An Anatomy of the Global Trade Slowdown based on the WIOD 2016 Release. GGDC Research Memorandum GD-162, Groningen Growth and Development Centre, University of Groningen, 2016. 94

Jan Tinbergen. Zur theorie der langfristigen wirtschaftsentwicklung. *Weltwirtschaftliches Archiv*, 55:511–549, 1942. 85

UNCTAD. Training Manual on Statistics for FDI and the Operations of TNCs Vol. 2. Technical report, United Nations, 2009. 39

UNCTAD. World Investment Report 2016. *United Nations*, 2016. 39

Ilke Van Beveren. Total Factor Productivity Estimation: A Practical Review. *Journal of Economic Surveys*, 26(1):98–128, 2012. 8, 93

Johannes Van Biesebroeck. Robustness of productivity estimates. *The Journal of Industrial Economics*, 55(3):529–569, 2007. 93

Frank G. Van Oort, Martijn J. Burger, Joris Knoben, and Otto. Raspe. Multilevel Approaches and the Firm-Agglomeration Ambiguity in Economic Growth Studies. *Journal of Economic Surveys*, 26(3):468–491, 2012. 105

Laura Vartia. Assessing plant entry and exit dynamics and survival–does firms' financial status matter? *European University Institute*, 2004. 107

Cassandra C. Wang. Geography of Knowledge Sourcing, Search Breadth and Depth Patterns, and Innovative Performance: A Firm Heterogeneity Perspective. *Environment and Planning A*, 47(3):744–761, 2015. 107

Zhi Wang, Shang-Jin Wei, Xinding Yu, and Kunfu Zhu. Measures of Participation in Global Value Chains and Global Business Cycles. *NBER Working Paper No. 23222*, 2017. 6, 16

Scott White and Padhraic Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM, 2003. 23, 49

wiiw and ISMERI EUROPA. Geography of Expenditure Final Report Work Package 13 - Ex post evaluation of Cohesion Policy programmes 2007-2013, focusing on the European Regional Development funds (ERDF) and the Cohesion Fund. Technical report, European Commission, 2015. 74, 96

Oliver E. Williamson. The Vertical Integration of Production: Market Failure Considerations. *The American Economic Review*, 61(2):112–123, 1971. 6

Oliver E. Williamson. Markets and Hierarchies: Analysis and Antitrust Implications: A Study in the Economics of Internal Organization. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1975. 6

Oliver E. Williamson. Transaction-Cost Economics: the Governance of Contractual Relations. *The Journal of Law and Economics*, 22(2):233–261, 1979. 6

Jeffrey M. Wooldridge. On Estimating Firm-Level Production Functions using Proxy Variables to Control for Unobservables. *Economics Letters*, 104(3): 112–114, 2009. 90

Luigi Zingales. Survival of the Fittest or the Fattest? Exit and Financing in Trucking Industry. *Journal of Finance*, 53(3):905–938, 1998. 107