Universidade Federal de Minas Gerais – UFMG
Faculdade de Ciências Econômicas
Centro de Desenvolvimento e Planejamento Regional

Pedro Henrique Castro Soares

An investigation on the impact of the relatedness of local productive structures on the survival and growth of new firms

Belo Horizonte

2020

Pedro Henrique Castro Soares

An investigation on the impact of the relatedness of local productive structures on the survival and growth of new firms

Dissertação apresentada ao curso de Mestrado em Economia do Centro de Desenvolvimento e Planejamento Regional da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, como requisito parcial a obtenção do Título de Mestre em Economia.

Orientador: Prof. Dr. João Prates Romero
Coorientador: Dr. Elton Eduardo Freitas

Belo Horizonte
2020

Ficha Catalográfica

S676i
2020

Soares, Pedro Henrique Castro.
An investigation on the impact of the relatedness of local productive structures on the survival and growth of new firms [manuscrito] / Pedro Henrique Castro Soares. – 2020.
82 f.: il.

Orientador: João Prates Romero.
Coorientador: Elton Eduardo Freitas.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Centro de Desenvolvimento e Planejamento Regional.
Inclui bibliografia (f. 79-82).

1. Organização industrial (Teoria econômica) – Teses. 2. Empresas – Finanças – Teses. 3. Desenvolvimento econômico – Teses. I. Romero, João Prates. II. Freitas, Elton Eduardo, 1986-. III. Universidade Federal de Minas Gerais. Centro de Desenvolvimento e Planejamento Regional. IV. Título.

CDD: 330

Elaborado por Leonardo Vasconcelos Renault CRB-6/2211
Biblioteca da FACE/UFMG – LVR/212/2021

FOLHA DE APROVAÇÃO

PEDRO HENRIQUE CASTRO SOARES

TÍTULO DO TRABALHO:

"AN INVESTIGATION ON THE IMPACT OF THE RELATEDNESS OF LOCAL PRODUCTIVE STRUCTURES ON THE SURVIVAL AND GROWTH OF NEW FIRMS"

Dissertação apresentada ao Programa de Pós-Graduação em Economia, da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, para obtenção do título de Mestre em Economia, área de concentração em Economia.

Aprovado EM **24 DE JULHO DE 2020**

BANCA EXAMINADORA

Dr. João Prates Romero
(Orientador) (CEDEPLAR/FACE/UFMG)
(participação por videoconferência)

Dr. Gilberto de Assis Libânio
Coordenador do Curso de Pós-Graduação em Economia

Universidade Federal de Minas Gerais | Faculdade de Ciências Econômicas |
Centro de Desenvolvimento e Planejamento Regional |

Cedeplar

FOLHA DE APROVAÇÃO

PEDRO HENRIQUE CASTRO SOARES

TÍTULO DO TRABALHO:

"AN INVESTIGATION ON THE IMPACT OF THE RELATEDNESS OF LOCAL PRODUCTIVE STRUCTURES ON THE SURVIVAL AND GROWTH OF NEW FIRMS"

Dissertação apresentada ao Programa de Pós-Graduação em Economia, da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, para obtenção do título de Mestre em Economia, área de concentração em Economia.

Aprovado EM **24 DE JULHO DE 2020**

BANCA EXAMINADORA

Dr. Elton Eduardo Freitas
(Coorientador) (FIEMG)
(participação por videoconferência)

Dr. Gilberto de Assis Libânio
Coordenador do Curso de Pós-Graduação em Economia

FACE | Av. Antônio Carlos 6627 | CEP 31270-901 | Belo Horizonte | MG | Brasil
Tel ++ 55 31 3409 7100 | Fax ++ 55 31 3409 7203 | www.cedeplar.ufmg.br

Universidade Federal de Minas Gerais | Faculdade de Ciências Econômicas |
Centro de Desenvolvimento e Planejamento Regional |

Cedeplar

FOLHA DE APROVAÇÃO

PEDRO HENRIQUE CASTRO SOARES

TÍTULO DO TRABALHO:

"AN INVESTIGATION ON THE IMPACT OF THE RELATEDNESS OF LOCAL PRODUCTIVE STRUCTURES ON THE SURVIVAL AND GROWTH OF NEW FIRMS"

Dissertação apresentada ao Programa de Pós-Graduação em Economia, da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, para obtenção do título de Mestre em Economia, área de concentração em Economia.

APROVADO EM **24 DE JULHO DE 2020**

BANCA EXAMINADORA

Dr. Pedro Vasconcelos Maia do Amaral
(CEDEPLAR/FACE/UFMG)
(participação por videoconferência)

Dr. Gilberto de Assis Libânio
Coordenador do Curso de Pós-Graduação
em Economia

Universidade Federal de Minas Gerais | Faculdade de Ciências Econômicas |
Centro de Desenvolvimento e Planejamento Regional |

Cedeplar

FOLHA DE APROVAÇÃO

PEDRO HENRIQUE CASTRO SOARES

TÍTULO DO TRABALHO:

"AN INVESTIGATION ON THE IMPACT OF THE RELATEDNESS OF LOCAL PRODUCTIVE STRUCTURES ON THE SURVIVAL AND GROWTH OF NEW FIRMS"

Dissertação apresentada ao Programa de Pós-Graduação em Economia, da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, para obtenção do título de Mestre em Economia, área de concentração em Economia.

Aprovado **EM 24 DE JULHO DE 2020**

BANCA EXAMINADORA

Dr. Paulo Sérgio de Oliveira Simões Gala
(FGV/SP)
(participação por videoconferência)

Dr. Gilberto de Assis Libânio
Coordenador do Curso de Pós-Graduação em Economia

FACE | Av. Antônio Carlos 6627 | CEP 31270-901 | Belo Horizonte | MG | Brasil
Tel ++ 55 31 3409 7100 | Fax ++ 55 31 3409 7203 | www.cedeplar.ufmg.br

À Crô.

Agradecimentos

O mestrado foi para mim uma experiência de muito auto conhecimento, resiliência e principalmente quebrar e vencer barreiras. É claro que nesse processo tive ajuda de muitas pessoas queridas e aqui vou deixar meus sinceros agradecimentos a elas.

Começo por aqueles que desde o início me mostraram a importância da educação: meus pais. Obrigado por me apoiarem e não medirem esforços para que eu possa estar aqui hoje.

Obrigado aos amigos que o mestrado e a UFMG me trouxeram. Vocês contribuíram para que eu me sentisse em casa em Belo Horizonte e deixaram os dias mais leves para seguir em frente. É impossível citar todos, mas estes não poderiam deixar de ser mencionados: Alex, Vitinho, Goku, Jonas, Pedrinho, Guilherme, Ana Tereza e Diógenes. Meu obrigado especial ao Alex que me apoiou em vários momentos da realização desse trabalho e me ajudou a compreender melhor o mundo acadêmico.

Agradeço à minha namorada Júlia que durante esse tempo também virou minha esposa. Obrigado por ter paciência durante dois anos nas pontes aéreas São Paulo - Belo Horizonte e por todo apoio emocional que você me proporcionou de uma forma muito carinhosa. Por comemorar comigo em cada passo por menor que fosse, mas que deu muita força para continuar.

Quero também deixar agradecimentos especiais ao João e ao Elton por toda disponibilidade, paciência e seriedade que apoiaram esta pesquisa. Meu muito obrigado por acreditarem neste trabalho.

Este trabalho não seria possível sem o suporte das instituições de ensino e de apoio à pesquisa. Agradeço à Universidade Federal de Minas Gerais e ao Centro de Desenvolvimento e Planejamento Regional por me darem a oportunidade de realizar o curso de mestrado. Agradeço também ao Conselho Nacional de Desenvolvimento Científico e Tecnológico por financiar esta pesquisa.

Finalmente, agradeço aos membros da banca examinadora, o Prof. Dr. Pedro Amaral e o Prof. Dr. Paulo Gala, por doarem um pouco do seu tempo e se disponibilizarem a avaliar este trabalho.

“*complex system: a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution.*”

Melanie Mitchell, Complexity: A Guided Tour

Resumo

Há um longo debate sobre quais tipos de externalidades de aglomeração melhor promovem o desenvolvimento econômico. Historicamente, ele se concentra principalmente na discussão de especialização e de diversificação de regiões. Avanços recentes em complexidade econômica e o estudo de atividades relacionadas lançam uma nova luz sobre esse debate. Esses avanços trouxeram novos métodos para medir a similaridade entre as atividades, que têm gerado resultados positivos. Neste estudo, aplicamos essa abordagem para entender uma das facetas do desenvolvimento econômico, como alterar a estrutura produtiva de uma região. Em particular, avaliamos o impacto da similaridade (*relatedness*) que o setor de novas empresas tem com a estrutura industrial existente na região. Para atingir nossa meta, usamos dados no mercado de trabalho formal brasileiro para estudar novas empresas entrando entre 2010 e 2013. Dois métodos são empregados para medir os impactos na sobrevivência, a regressão logística, que estima a probabilidade de uma nova empresa sobreviver por três anos, e o modelo de riscos proporcionais de Cox, que estima o risco contínuo de morte de novas empresas. Além disso, o impacto do relacionamento no crescimento da empresa é avaliado através da regressão linear com efeitos fixos para indústria, região e ano. Nossos resultados apoiam a hipótese de que o *relatedness* tem um impacto positivo na sobrevivência e no crescimento dos participantes. O que, por sua vez, fornece mais evidências para o conceito de *smart specialization*.

Palavras-chave: Similaridade entre indústrias; Complexidade econômica; Sobrevivência de empresas; Diversificação regional

Abstract

There is a long lasting debate on which type of agglomeration externalities best promote economic development. Historically, it is mainly centered on specialization and diversification of regions. Recent developments in economic complexity and the study of related activities shine a new light onto this debate. These advancements brought novel methods for measuring similarity between activities and generated positive results. In this study, we apply this approach to understand one of the facets of economic development, how to change the productive structure of a region. Particularly, we evaluate the impact of the similarity (relatedness) the industry of new firms have to the existing regional industrial structure. To achieve our goal, we use data on the Brazilian formal labor market to study new firms entering between 2010 and 2013. Two methods are employed to measure the impacts on survival, logistic regression, which estimated the probability of a new firm surviving three years, and the Cox Proportional Hazards model, which estimates the continuous risk of death of new firms. Furthermore, the impact of relatedness on firm growth is assessed via linear regression with industry, region and year fixed effects. Our results support the hypothesis that relatedness does have positive impact on the survival and growth of entrants. Which, in turn, gives further evidence for the concept of smart specialization.

Keywords: Relatedness; Industry relatedness; Economic complexity; Firm survival; Regional diversification

List of Figures

List of Tables

Contents

Introduction

It is commonplace to classify countries in three categories, regarding their level of economic development: developed countries, developing countries, and least developed countries. Though this is, in fact, a rather simple approach, it makes for a good base model to grasp the essentials of economic development. Particularly, how does a nation progress from a lower to a higher stage in the development scale? This question motivates a whole field of Economic Theory.

Development economics began based on classical mercantilist economists, such as Smith, Schumpeter, Keynes, and Marx. Until mid-1970s, the field progressed mainly through theoretical models created from studying the historical development of countries. In the 1950s, Robert Solow brought a mathematical approach to economic growth, by developing a simple model, which made the analysis of dynamic growth compatible with statical general equilibrium.

The focus on general equilibrium led mainstream economics for decades. Though mathematically elegant, the theory makes unrealistic assumptions, which raised significant criticism over the years. The views on general equilibrium are changing. More recently, Complexity Theory views on the Economy affirm that its natural state if that of non-equilibrium. Arthur (2014) refutes general equilibrium by saying that:

> Under equilibrium by definition there is no scope for improvement or further adjustment, no scope for exploration, no scope for creation,no scope for transitory phenomena, so anything in the economy that takes adjustment—adaptation, innovation, structural change, history itself—must be bypassed or dropped from theory. The result may be a beautiful structure, but it is one that lacks authenticity, liveness, and creation. [. . .] (ARTHUR, 2014, p. 4)

On our approach to economic development, we focus on the discussion of diversification and specialization, mainly led by Jane Jacobs and Alfred Marshall. Marshall argues that specialization is the path to developing the local economy. Concurrently, Jacobs states that diversification fosters the creation of new products, which in turn increase exports, thus developing the economy. Nevertheless, both authors advocate for new technology as the main motor for economic growth.

Arthur (2009) describes technology in three ways: "*a means to execute human purpose*", it is an "*assemblage of practices and components*", and is it a "*collection of devices and engineering practices available to a culture*". The last two definitions make it clear that technology is something built from combinations of available knowledge. Thus, the question is, how to widen the range of options for new technologies to emerge.

Here, we introduce our research agenda, bearing in mind that we need combinations of existing technologies to further technological advancement, and thus economic growth. We argue that neither specialization nor diversification alone are the best strategy for economic development. On one hand, specialization alone does not provide a sufficient variety of existing technologies for new kinds of technology to emerge, at least not as much as there could be. On the other hand, the more diversified technologies are, the harder it is to combine them, again preventing a multitude of combinations. Rather, we apply the findings of Hidalgo et al. (2007), that countries diversify to related products, to test the hypothesis that diversification towards related industries is the best alternative. Hence, diversification and specialization are not mutually exclusive ideas, but rather we explore their combination.

Regions change their productive structure in three main ways: new firms start producing goods that were not previously present in the region, existing firms change their productive structure, or old firms die. In this work, we pursue the first option. Thus, we analyze what role agglomeration externalities play in new firm survival and growth. Instead of looking at countries, we look into Brazilian regions. We investigate whether new firms are more likely to survive if they are similar to the local industry. However, we also want to look at the diversification and specialization of that location. So, is the firm's similarity to the local industry a relevant factor for its growth and survival?

Recent developments in Economic Complexity offer new methods to measure similarity (*relatedness*). These have been successfully applied in different fields, as we discuss in chapter 1. Moreover, with this tool, it is possible to quantify the similarity between industry and locality. Thus we can measure its impact on the survival and growth of entrant firms in that locality.

Does relatedness have an impact on the survival of new firms? That is, is a firm more likely to survive when its activity is related to others present in the region? In the same manner, how does relatedness impact the growth of entrant firms? These are important questions for developing countries.

Jara-Figueroa et al. (2018) investigate the relationships between the related knowledge workers bring into pioneer firms and the survival rate of said firms. A pioneer firm is a new firm whose industry was not present before in the region. The authors characterize two types of knowledge, industry specific and occupation specific. For example, a salesperson at a shoe store has knowledge about the shoe industry and knowledge about the occupation of seller. If that person were to start working at a car dealership, she would bring occupation specific knowledge, however not industry specific knowledge, since automotive and footwear are not related industries. They find that for pioneer firms, industry specific knowledge is more relevant for their survival and growth.

In a similar way, we estimate the effect of relatedness on a firm's survival and growth.

Nonetheless, we focus on all new firms, rather than only pioneer ones. Furthermore, the way we measure relatedness is different from Jara-Figueroa et al. (2018). While they used the flow of workers between industries and occupations to determine relatedness, we take the approach of Boschma, Balland and Kogler (2015), where the relatedness between a pair of industries is measured from the frequency of co-occurrences in the same regions. Thus, this index tells us, for any given activity and region, how similar that activity is to others present in the region. We expect that the greater this relatedness, the higher the chances of survival of entrant firms and the faster they grow.

We start chapter 1 with a review of the main authors on location theories which sets the theoretical grounds for economies of agglomeration, or the positive externalities caused when economic activities concentrate in a region, the subsequent topic. Examples are larger consumer market, more skilled workers in the region, more suppliers. According to Jacobs (1969), firms also benefit from a diversified region though knowledge spillover. In a later study, Porter (1990) showed that regions also benefit from competition among firms.

This contributes to the relatedness literature, the main concept behind this thesis. Relatedness is one of the main arguments to support the theory of diversification as a driver of economic development and growth. We also present Marshall's theory of specialization as a driver of innovation along with diversification. This discussion is of utmost importance to the debate. Thus, still in chapter 1, we present the main developments and applications of the principle of relatedness (HIDALGO et al., 2018). We finish this chapter with a historical review of works on firm survival and growth.

On chapter 2, we present the data and methods used. First we describe the econometric models used to estimate those effects. We test two models of survival and one of growth. The survival models are logistic regression and Cox proportional hazards regression. On the first, we estimate the probability of a new company surviving to the third year given its relatedness to a region. On the second, we estimate how likely is a company to close given its relatedness. We estimate the impact on growth via linear regression. Afterwards, we start describing the data, we use the Annual Report of Social Information (RAIS), the most complete data source of the Brazilian workforce employed in formal labor. Then we explain the control variables and why they are important to the study. The main variable is the relatedness indicator, it will give us a measure of how related an industry is to the industries in a region. We want to see how this relatedness affects the probability of survival (growth rate) of new firms. We expect it to have a positive impact, meaning that the more related the firm's industry is, the greater the chance of survival (growth rate).

On chapters 3 and 4, we present our empirical results. First, on chapter 3, we do a statistical description of our data, followed by the results of our econometric analysis.

Our findings support the positive effect of relatedness on the survival and growth of newcomers. On chapter 4, we repeat the exercise of the previous chapter with a focus on high and low complexity industries and regions. That is, our goal is to understand whether relatedness plays different roles, conditional on how sophisticated the region (industry) is. Finally, on the last chapter, we conclude the study pointing our main findings and future research topics.

1 Literature Review

1.1 Introduction

In this chapter, we review the main exponents of the literature pertaining to our research topics. First, we present the founding theories relating the economy to space. These foundational concepts are essential to situate our research topic. Thus, we expect to give the reader the necessary ground knowledge.

Second, we present the central authors behind economies of agglomeration. In particular, Alfred Marshall and Jane Jacobs. Furthermore, we expose the fundamental theoretical debate on regional diversification, particularly specialization and diversification. Our empirical investigation seeks to contribute to this discussion by applying the relatedness methodology.

Afterwards, we discuss the recent developments on economic complexity and the principle of relatedness. Moreover, we present relevant publications of its various applications. Finally, we give a brief revision on the literature of survival and growth of new firms.

1.2 Space and location theories

Our research subject, survival and growth of new firms, is a branch of a larger field in economics: the study of the industrial structure and its relationship with regional development. Different regions accommodate distinct industries which makeup the region's industrial structure. This arrangement shapes the region's social, economic, environmental and political development. Therefore, the study of factors influencing the survival of new firms on a regional level is not only of a microeconomic nature, but also requires a deeper understanding of how regional economic structures evolve over time. This is a dynamic complex process with feedback loops, in which the entry of a new firm transforms the structure, which in turn affects the entry of subsequent firms.

Economic geography is intrinsic to this subject. The idea that distinct regions house different industries is based on the perception that industries are not uniformly distributed in space, or not ubiquitous. They are however present in localized spaces. This leads to our background question of why are industries established in their locations and what role does space, and its social and economic structure, play in this process.

To address the spatial issue, it is first necessary to briefly discuss the theories of location, which address the matter in a predominantly microeconomic form. Von-Thünen (1966) is one of the first authors to address the importance of space in the economy and to discuss what defines a location for agricultural activities. To tackle this problem, the author developed a model, in which the soil fertility is homogeneous, there is only one consumer urban center and transportation costs vary, increasing according to the distance to the center. The price that the producer will charge for his product will increase the greater the distance to the population center, due to the increase in transportation costs. Since the market price is given by the highest price, which in turn is defined by those who are furthest from the center, producers closest to the urban center have an advantage, as they obtain a price differential. Thus, the decision to locate productive activities takes into account the proximity to the population center, as it allows to obtain extra income due to the location.

The approach of Von-Thünen (1966) is focused on agricultural production and his model culminates in a hierarchy of the spatial distribution of activities in the form of concentric rings, whose common central point is the population center. Activities with higher transport costs per unit produced and higher transport costs by distance traveled would have to be located in rings closer to the center. However, given the dispute for the central space, it is the activities with the highest productivity per cultivated area that are able to occupy the most central spaces. Hence causing a de-agglomeration effect, in which less productive activities tend to be expelled to the periphery. For our perspective, it is interesting to note that such conditions must weigh in the location chosen by an incoming firm in terms of its survival and growth[1].

With a similar interest, on how the location of productive activities is determined, Weber (1929) proposes an approach in which the cost of transportation is also decisive. His model is based on the perception that inputs are non ubiquitous, that is, they are not evenly distributed in space, the same way consumer centers are. Firms are faced with prices given at their local market, so that their location will be chosen to favor the point that minimizes the cost of transportation between the location of the raw materials and the factory, as well as between the factory and the consumer market. Such transportation cost is given in unit terms, that is, it is considered that each type of raw material and goods produced have different transport costs. In this manner, the decision of the ideal location (minimum transport cost) between the consumer market and the source of raw materials (which are assumed to be located in different places) will be based on the relationship between the cost of transporting raw materials to the factory and the cost of transporting the produced goods to the consumer market.

If the cost of transporting raw materials to the factory is higher than the cost of transporting the produced goods to the consumer market, the factory should be located closer to

[1] Although the firm's location in relation to centralities is an important factor, this variable was not explored in the models developed in the following chapters, because it requires data on location, which were not available, and a geolocation effort that was outside the scope of this work.

the source of inputs. The opposite is also true. Therefore, it is present the notion that the relationship between the weight of the raw material and the final products is decisive, because the heavier something is, the more difficult it is to transport. Thus, it is important to consider that the decision of location will vary according to the type of productive activity, as there is a change in weight when the raw material is transformed into a final product that varies according to the type of production. This idea is summarized in the so-called index of raw materials, which expresses the relationship between their weight and the weight of the final product[2]. In addition to this general framework, the author adds other issues such as the location and price of labor, which have a secondary role.

The simplified descriptions of the models by Weber and Von Thünen bring elements to reflect on the relationship between firms and space. However, among the classic works of the theories of location, the perspectives of Lösch and Christaller, which deal with the issue from a regional point of view, are more useful to this research. Allowing for a more fertile intersection with the idea of relatedness that will be discussed later.

Lösch (1974) develops a model that, to some extent, is close to the authors discussed earlier following a hypothetical-deductive method. However, there are relevant differences from Weber. The first difference regards the assumptions of the models, as Lösch assumes that raw materials and inputs are ubiquitous, population and consumers are distributed equally in space, there is free entry and exit for firms and transportation costs are homogeneous and are passed directly to the consumer. The second is that the author seeks to present a perspective of general equilibrium and not just the choice of location for individual firms. The third, and most important, concerns the assumption that firms have gains of scale according to the growth of their production due to the expansion of demand (LEMOS, 1988).

Based on this, the author develops the idea of a market area. Starting from the assumption that with the growth of production, the firm presents gains of scale, which allows it to obtain more competitive prices and to receive demand from more distant consumers. However, such gains have decreasing returns. On the other hand, there is a higher transportation cost for consumers who are further away from the company. Thus, while gains in scale act to promote long distance sales, transportation costs act in an opposite direction, limiting the firm's reach. In other words, Lösch endogenizes transport costs in the demand curve faced by the firm. Hence, there is an area where the firm is able to sell its products, which is limited by the distance at which gains in scale no longer surpasses the cost of transportation. Lösch model's intent to be a general equilibrium derived from the idea that another competing firm tends to settle in a position symmetrical to that of the first firm, so that their market areas are adjacent.

[2] It is beyond the scope of the present study to work with these models in formal terms or in more detail. The objective is simply to present the author's general idea, which contributes to problematize the decision to locate new firms and, consequently, their survival and growth.

From Lemos (1988), we observe that the idea of a market area can be used, not only for the individual firm, but also for a set of firms in the same sector, that is, an industry. When considering an agglomeration of different industries, in a diversified environment, we have the overlap of several market areas. Through Lösch's perspective, we conclude that the market zone of a firm or industry competing with the incoming firm influences its decision to locate and consequently the demand it faces, which in turn affects its survival and growth. In other words, the survival and growth of a new firm may be affected by the type of industry existing in the region in which it is entering, that is, by its economic structure.

We would like to stress that the classic authors of location theory present models within a microeconomic perspective, often based on unrealistic hypotheses. Nonetheless, from this brief review, we note that their ideas support an initial understanding on how the space and types of goods, existing in a region, are related to the economic performance of new firms, which is a central issue in this work.

1.3 Economies of Agglomeration

In order to address the interactions between the economic structure of a region and the entry and growth of new firms, the literature on agglomeration economies is more appropriate. It addresses the issue of how the spatial proximity between firms influences the economic performance of a region, the firms in the region, and particularly on the diversification process of regions. Furthermore, diversification is addressed in this research through the economic complexity approach. Consequently, we present the main theoretical ideas from selected works on in the economies of agglomeration literature. By doing so, we aspire to elucidate the field and how it is related to our empirical investigation.

According to Freitas (2019, own translation), agglomeration externalities "*are advantages or disadvantages that local productive entities derive from the concentration of economic agents and activities*"[3] and "*to formally constitute externalities, costs or benefits must be experienced by a productive entity, but caused by another entity*"[4]. In other words, from the discussion in the previous section, in which space and the productive structure influence the entry of new firms, we appreciate that understanding the idea of agglomeration externalities is fundamental to be able to discuss entrant firm's survival and growth. Therefore, we will focus the discussion on two authors, who bring the main ideas of agglomeration externalities: Alfred Marshall and Jane Jacobs.

[3] "são vantagens ou desvantagens que as unidades produtivas locais tiram da concentração de agentes econômicos e de atividades" (FREITAS, 2019).

[4] "para constituir formalmente externalidades, custos ou benefícios devem ser experimentados por uma unidade produtiva, mas causados por outra unidade" (FREITAS, 2019).

Alfred Marshall initiates the discussion on economies of agglomeration. In his book, Principles of Economics, Marshall brings a historical discussion on the determinants of firm location. The author also introduces the idea that firm agglomerations can lead to economic gains, what later became known as *Marshallian* externalities. Marshall states that the causes for the wealth of a nation are rooted in international trade. Nevertheless, one must also consider the gains obtained by industrial districts, i.e., observe *"groups of skilled workers who are gathered within the narrow boundaries of a manufacturing town or a thickly peopled industrial district"* (MARSHALL, 2013, p. 225).

From a production point of view, Marshall appreciates the spatial proximity of industries that require the same skills from workers. He states that this environment facilitates the transmission of knowledge and the discussion of ideas to improve productive processes, machinery, and the emergence of inventions in that industry. In addition, it is relevant to our analysis that auxiliary branches, regarding an industry, are likely to develop in the region, yielding a space with complementary industries.

The aforementioned complementarity allows for a greater specialization by the firm. For instance, by having closely located suppliers, a company can afford more expensive, and more specialized machinery, which would be too costly otherwise. Furthermore, the proximity to companies of the same industry means there are other players interested in older models of the company's machines, to whom it could sell to, thus reducing maintenance costs. Equivalently, specialized regions attract highly skilled workers, which ensures employers a steady supply of labor. In other words, the author supports the formation of specialized industrial districts, with greatly complementary industries, as they bring economic gains and promote regional economic growth (MARSHALL, 2013, p. 225-226).

Likewise, from a demand point of view, Marshall states that closely located companies of the same industry compels consumers, who look for those products, to go to the place of agglomeration. That is, the proximity of industries increases the global demand from the region where they are situated by attracting buyers. The author refer to these gains in agglomeration as "external economies" (MARSHALL, 2013, p. 221).

Marshall's angle on the subject is relevant for this study, because it allows us to construct a hypothesis that firms, with a higher similarity (*relatedness*) to the region they enter, are susceptible to higher probability of survival and larger growth rate. Be it by greater demand or by gains associated with sharing resources with similar industries.

Frequently, Marshall is associated mainly with the advantages of specialization in agglomerations. Little, however, is mentioned about his approach to diversification. Indeed, it is not a major part of his work. Nevertheless, it is worth noting that the author uncovers at least two gains from diversification in industrial agglomerations. First, regions with a wide variety of jobs (which require different skills from the worker) grow faster, after all, the diversity of skills within a family increases the chances more members are employed at any given time, thus raising family income. Second, in a diversified city, even if there is a sharp decline in demand for a certain industry, the other existing industries will still remain in good condition to dissipate the impacts of the shock over the local economy. This would not happen in the case of total specialization (MARSHALL, 2013, p. 225-227). Yet, Marshallian externalities are currently used when referring to benefits and costs relative to agglomeration of firms in the same economic activity.

Glaeser et al. (1992) summarize how Marshallian externalities, or Marshall-Arrow-Romer (MAR) externalities, are used today to refer to the benefits of the spatial concentration of firms in the same industry. The authors argue that such externalities can be synthesized by the idea that "*the concentration of an industry in a city helps knowledge spillovers between firms and, therefore, the growth of that industry and of that city*". According to that study, from the perspective of MAR externalities, the existence of local monopolies would be better than local competition, as they restrict the flow of ideas to that territory. This allows externalities to be more easily internalized by innovative agents, favoring growth [5].

Even though Marshall writes about the benefits of regional diversification, the main reference to discuss how the introduction of new economic activities in a region begets economic development is Jane Jacobs. In *The Economy of Cities* (JACOBS, 1969), shows how diversification is the process of adding new kinds of work to old work. Namely, there is a tendency for firms to produce goods that are similar to those previously produced, as they already have the expertise for such. The author summarizes this idea by stating that "*New goods and services[...] do not come out from thin air. New work arises upon existing work; it requires 'parent' work.*" (JACOBS, 1969, p. 52).

Her approach agrees with Marshall's in stressing the importance of the division of labor in a region (both inspired by the work of Adam Smith). However, her reasoning is that said division of labor fragments production and creates the conditions for the emergence of new kinds of work. Which in turn are a response to the demand for resources, given the new production, and the demand from final consumers, allowing for productive diversification. The author states that division of labor, on its own, is not enough to yield diversification. Nevertheless, it organizes production in such a way that it is more prone to the emergence of new activities (JACOBS, 1969, p. 76).

Through this avenue, we introduce the discussion relevant to this study. In that, it contradicts the idea that long term economic development derives from the division of

[5] Still, according to Glaeser et al. (1992), this view on greater efficiency of monopolies on incorporating technological spillovers is contrasted by the approach of Porter (1990), in which competition for innovation in a specialized geographic region is fundamental for innovation. Here, however, we will not go into that debate.

labor within the same activities. Simply dividing production, without adding new kinds of work, increases efficiency and makes way for diversification. Still, it does not have the capability of generating new work, the actual vehicle of long term development:

> "It is ironic that division of labor gets no credit for its genuinely bountiful effect. It prepares the way, it provides the special footholds, for adding new goods and services into economic life. Ants, no matter how efficiently they divide their tasks, do nothing so terrifying and wonderful. Seen as a source of new work, division of labor becomes something infinitely more useful than Adam Smith suggested when he limited its function to the efficient rationalization of work." (JACOBS, 1969, p. 76).

In an attempt to measure the development of a city, the best way would be to check how many new types of work arise, in a given time span, relative to the work already existing in that city (JACOBS, 1969, p. 84)[6]. In other words, it is the city's ability to creatively reinvent its productive activity that, in the long run, leads to its continuing growth and development. Such process of creating new work and diversification derives precisely from the urban agglomerations (JACOBS, 1969, p. 76). It allows direct exchanges of knowledge between different sectors, which leads to the development of new products.

Many of the new types of work arise directly from urban agglomerations and the "impracticality of big cities" that they imply. The interaction of the different types of knowledge that are clustered in diverse cities is reverted to creative solutions to problems originated from agglomeration itself. In other words, agglomeration generates innovations that favor development.

On the other hand, the growth of cities is given by a process of diversification as it allows the expansion of their exports. This increase, in addition to directly having a multiplicative effect on employment, allows for import expansion of goods not produced in the city. These, in turn, are important as they facilitate the generation of new local production based on the resources that are imported. Once there is a market that demands such resources, there is the possibility of creating industries focused on the production of such products within the city itself. An import substitution process. As a result of this process, according to Jacobs (1969, p. 142): i) economic activity expands rapidly; ii) agricultural market grows rapidly due to changes in the composition of the city's imports; and iii) jobs in the city expand rapidly.

Exports are, therefore, essential for the beginning of a process of sustained economic growth in the region. Moreover, where there is greater diversification, there tends to be greater exports, because diversified places are more likely to start a new type of work, since it is easy to acquire the most varied types of inputs from local suppliers, such as parts or services, necessary to new work and to assemble new goods. Furthermore,

6 Her approach deals with cities, however we can also think about it in regional terms.

diversification within a given space allows for easy readaptation of the existing production to supply the need for a new industry, by the absorption of similar knowledge, as well as the assimilation of types of knowledge that is not exactly similar but that when combined end up generating a new type of good, that is, an innovation. Jacobs (1969) makes it clear that the spatial proximity of different productive activities, agglomeration, favors the productive diversification that fosters development. The key concept is to create "new work" using "old work", hence a greater variety of "old work" means greater possibilities for "new work". The idea of agglomeration externalities, to Jacobs, is intertwined with the idea of diversification.

Among the empirical works discussing agglomeration economies and agglomeration externalities, some seek to identify whether it is Marshallian externalities or Jacobs' perspective that has greater adherence to reality. Others, however, seek only to evaluate the existence of agglomeration externalities, as is the case of Sveikauskas (1975), which tests whether large cities tend to show productivity gains. The author first highlights static factors that favor productivity, given by the greater division of labor and its reversion in the gains in scale related to the size of firms, industries and the amount of economic activity. This is a direct reference to Marshallian externalities. However, he also highlights that there are productivity gains arising out of dynamic factors, which occur due to the greater flow of ideas from the concentration of different types of businesses in a city that, in his understanding, are the most important as they favor technological advancement. These viewpoints bring up the perspective of Jacobs' externalities.

To discuss this issue, the author uses cross section data referring to 14 industrial sectors in metropolitan areas of the United States for 1967, aiming to measure the impact of population growth on productivity[7], controlling for the years of study of the population and for different regions. His hypothesis that population growth is related to higher productivity is confirmed by the result that by doubling the size of a city's population, productivity increases by 5.98%. His work, however, does not seek to assess whether there is a predominance of effects between Marshallian and Jacobs externalities.

Segal (1976) evaluates the returns of the factors of production of large metropolitan regions of the United States for the year 1967. The authors start from the observation that the product per worker in large cities is greater than in small cities and present three possible explanations for this. First, there is an increase in the capital to labor ratio in large cities (different production functions by city size), due to higher wages relative to the profit rates, as wages must compensate for the difficulties imposed by population agglomeration (such as higher rental and commute costs), or due to the higher qualification of labor in urban centers. Second, the existence of returns to scale. Third, large and small cities have production functions with the same capital to labor

[7] He uses added value per worker as a measure of productivity.

ratio, but with different multiplier constants, being higher in the first case. Although this type of modelling is not used in the present work, it is worth noting that its results show that metropolitan areas with two million inhabitants or more have an approximately 8% higher factor productivity than smaller areas. However, they emphasize that this is not due to increasing returns to scale, but rather an agglomeration effect that displaces the production function.

The works of Sveikauskas (1975) and Segal (1976) use population to discuss agglomeration economies. However, there is no variable aimed at directly evaluating the existence of agglomeration externalities or checking for greater returns when the city's production is specialized or diversified. The work of Moomaw (1983) argues that, in studies that use population as a measure of agglomeration, there is an understanding it functions as a substitute for agglomeration economies. These, in turn, actually have their economic dynamics determined by other variables such as urbanization economies and location economies, or, more generally, agglomeration externalities as discussed so far.

An effort to verify the types of agglomeration externalities is more evident in the work of Henderson (1986), which divides the economies of scale obtained into location economies and urbanization economies. The first are similar to what we present here as Marshallian externalities, that is, gains from the existence of a large volume of employment within the same industry whose firms have spatial proximity. The latter concerns the economies of scale obtained by the operation of firms in large agglomerations that provide interaction between different economic sectors. In this manner, the level of specialization or the specific composition of that economy does not play a large role, but the existence of a large labor market and high level of employment and economic activity that imply gains of scale external to firms. Note that these urbanization externalities are not the same as what we call Jacobs' externalities, as it does not necessarily state that economic gains come from diversification. Despite being similar to Jacobs (1969) approach, the main variable is the size of the city and not its diversification.

However, Henderson (1986) makes the fundamental contribution that it is necessary to evaluate the economies of agglomeration separately for different industries, with different production functions per city. Hence, this opens the possibility to understand the kinds of agglomeration externalities and gains of scale present in each industry. This idea is important because the growth of cities is likely to simultaneously represent an attraction for those industries that obtain greater location economies to the detriment of those that obtain less gains. If location economies are prevalent, this tends to shift the central research question on agglomeration economies:

> In that case, the relevant question is not whether resources are more productive in large versus small cities, but what types of industries are best off (and likely to be found) in what sizes of cities. (HENDERSON, 1986, p.49)

In his econometric investigation, he seeks to qualify the type of agglomeration externalities. The author uses data on Brazilian urban areas from the 1970 Demographic and Industrial censuses and from the Census of Manufacturers of metropolitan regions in the United States. The results for Brazil show that location externalities have the greatest effects for most industries, so a 10% growth in the production of a firm from that industry in that city is related to an approximately 1% growth in the production of a firm from that industry in that city. For the United States, the results also show strong effects of location economies (HENDERSON, 1986).

The empirical discussion on which types of agglomeration externality prevail come to attention with the work of Glaeser et al. (1992). The authors seek to empirically assess agglomeration externalities with data of 170 major American cities in a cross sectional model. Their work focusses on large firms, and is based on the assumption that externalities are a source of permanent income growth through technological spillovers. That is, they follow the entire life cycle of the company and not just the emergence of new activities. This way, their analysis tests in which cities firms grow faster, taking into account local levels of specialization and competition.

Their main results indicate that a certain industry grows slower in cities where this industry already has a large share compared to cities with smaller share. In other words, Marshallian externalities (specialization externalities) do not seem to have a real positive impact on industry growth in the data. Moreover, their results indicate that an industry tends to grow more when the city is specialized in other industries, that is, the data suggests that Jacobian externalities (diversification externalities) favor growth. However, Glaeser et al. (1992) highlight that the perspective analyzed by them seeks to simultaneously explain the growth of cities and the formation of cities, which gives the analyzed externalities a dynamic characteristic[8].

The study by Henderson, Kuncoro and Turner (1995) also seeks to explore dynamic agglomeration externalities. That is, related mainly to knowledge flows between spatially close firms that consolidate over time[9]. Dialoguing with the previous study, they demonstrate that Jacobs' externalities occur only for some specific sectors, while for others there is a predominance of MAR externalities, i.e., related to greater productive specialization. To this end, the authors used employment data from eight industrial

[8] There are works focussing on the perspective of static externalities, in which they do not specifically concentrate in the relationship between technological spillovers and growth, but in the relationship between location and specialization (called location or urbanization externalities), reaching different results. Some examples cited by Glaeser et al. (1992) are the works of Henderson (1986), Wheat (1986), and Krugman (1991).

[9] It is interesting to not that, in their study, we find a simple and illuminating definition that summarizes the idea of dynamic externalities that was worked out in the present literature review "*Dynamic externalities come from local accumulations of knowledge enhanced by long-term relationships and histories of interactions, creating a stock of 'local trade secrets' that benefit local firms*" (HENDERSON; KUNCORO; TURNER, 1995, p. 1083).

sectors in 224 metropolitan areas in the United States between 1970 and 1987. More specifically, their results point out that Jacobian externalities tend to be more important for new high-tech industries, that benefit from environments with greater diversification. Cities with greater diversification tend to attract new industries. However, it is also found that MAR externalities are important over time for some of these industries. So that, although diversification is an important factor for nascent industries, their continuity in a city tends to be favored by greater specialization. On the other hand, mature industries tend to benefit only from locations with more specialized economies in their fields of activity, that is, specialization externalities. The results, therefore, are contrary to those of Glaeser et al. (1992), in which diversification externalities always dominate.

The work of Combes (2000) analyzes how sector specialization, diversification, competition, factory size, and employment density affect employment growth in French regions, using data from 1984 to 1993. An element of his work is to use a wider range of industrial sectors and to also encompass the service sector, which is important as services have different characteristics from the industrial sector, for often being non-tradable, and with their consumption restricted to the place of production. Furthermore, many services use inputs from distinct industrial sectors, which suggests that services should benefit more from diversification externalities. Another difference is that his work uses data from all over France, while previous studies were focused on metropolitan regions, which accounts for a possible selection bias in the choice of regions from previous works.

The results indicate that specialization has a negative impact on employment growth for most sectors, especially for services, with a positive impact in only 13 of the 94 sectors evaluated. The author points out that such negative results tend to be strongly influenced by the sectoral cycles that occurred in France in the period, in which many of the sectors studied were still emerging and are, therefore, growing on a national scale. Considering the life cycle of products, there is a tendency for them to first appear in a concentrated manner and then to disperse, resulting in negative impacts from specialization on job growth, since employment growth in each region is being considered relative to its participation in the French economy as a whole. The other explanation would be the opposite, as there is a group of sectors that, during this period, have reduced employment at the national level. In this scenario, economies that are less specialized would have greater flexibility to replace the production of decadent sectors with rising sectors. This result is consistent with the study by Glaeser et al. (1992) (COMBES, 2000).

When it comes to diversification, that is, Jacobs externalities, Combes (2000) finds positive effects for most service sectors and negative effects for most industrial sectors. The magnitude of its impact being, on average, small compared to the impacts of specialization. His general assessment is that the economies of externality related to diversification are significant only in some sectors, agreeing broadly with Henderson, Kuncoro and Turner (1995).

Jacobs' approach on agglomeration externalities is a fundamental theoretical basis for works in economic complexity that will be discussed shortly in the next section, in which the complexification of the economy corresponds essentially to the expansion of its diversification. This is evident when taking into account the *smart specialization policies* discussed by Balland et al. (2019). In other words, from Jacobs' perspective on agglomerations, there is a great tendency for new kinds of firms that appear in a region to maintain a degree of similarity with existing firms, which, from the perspective of complexity, implies that there is a tendency of emergence of firms with greater *relatedness* with the industries of the region.

Certainly, our exposition on agglomeration economies does not contemplate the whole debate. Nevertheless, it is clear how the debate ends up being progressively outlined around two views in which certain moments seem to be antagonistic, which are the benefits of specialization and the gains from the diversification of economic agglomerations. However, as the empirical literature emphasizes, the division is not clear and the effects of diversification and specialization are defined according to the sectors analyzed and other variables that are taken into analysis.

Thus, it seems more appropriate to say that agglomeration economies are favored by a combination of specialization and diversification rather than predominantly one or another. For this reason, it is proposed in this work to advance this agenda on agglomerations using the concept of relatedness. It provides instruments to elaborate the idea of smart specialization, which in short is the diversification of economies towards specific sectors that present greater affinity with the existing productive structure in that region, overcoming the idea of generalized diversification or specialization and looking at the real possibilities of the region. The following section introduces this perspective.

1.4 Relatedness

As we have discussed thus far, it is evident that modifications to the productive structure are not a novel research subject. Neither is the idea of related economic activities. Hirschman (1958) brings a perspective on related industries through input output links between them. The author describes two main linkage effects *backward linkage effects* and *forward linkage effects*. These refer to the fact that an economic activity, present in a region, has certain demands and produces goods. Therefore, the industry induces the need in the region to supply it with its inputs (*backward linkage*). Furthermore, if this activity does not supply the final consumer, it induces other economic activities that use these supplies to come to the region (*forward linkage*).

However, we take the approach of capabilities, that is, we are interested in measuring similarities between industries not by the productive chain in which it is inserted, rather by what the region needs to have in order to foster an economic activity. We are interested in which knowledge and inputs are common to industries, for example, institutions, economic system, worker skills. Recent years have brought significant advances in how to quantify and empirically evaluate the economic diversification and specialization processes, mainly at regional and urban levels. Particularly, due to the development of the economic complexity approach, and its practical method to examine such processes.

In recent years, research on economic complexity brought about a new methodology with prominent instruments to study the industrial structure of a region. Among them, is *the principle of relatedness*, which proposes ways to measure similarity between products, industries, or economic activities based on the existing production in a locality.

According to Hidalgo et al. (2018, p. 452), "*two activities, such as products, industries, or research areas are related when they require similar knowledge or inputs*". The authors based themselves on the idea that different goods are similar as they need similar inputs, be it resources, knowledge or technology employed in their production. According to Balland et al. (2019), relatedness is conceptualized as:

> [...] knowledge has an architecture that is based upon similarities and differences in the way that different types of knowledge can be used. When knowledge subsets are close substitutes for one another, or when they demand similar sets of cognitive capabilities and skills for their use, we think of them as being related or proximate to one another [...] (BALLAND et al., 2019, p. 2)

By measuring the similarity among various sectors of economic activity present in a certain region, it is feasible to obtain an indicator of the homogeneity (or heterogeneity) of the current industrial structure. This enables us to advance in the investigation of the modifications suffered by the local economy through time and associate these changes with distinct variables. Which leads to the identification of correlations and causalities that assist the comprehension of economic phenomena. Ultimately, these measurements facilitate the inference of the similarity between economic activity not yet present in a region and those previously existent.

The principle of relatedness has had quite a few applications in recent work, being calculated in different ways. Thus, it is enlightening to examine these variations in order to understand the importance and the potential this principle has.

The notable works of Hidalgo et al. (2007), Hidalgo and Hausmann (2009) e Hausmann et al. (2014) apply the concept of relatedness to countries exports. The authors verify the similarity between two products based on the frequency they are simultaneously exported by countries. Put in another words, if countries that export product A usually export product B, then it is assumed that the two products have a high coincidence of inputs and technologies required for their manufacturing. Drawing on this similarity between products and the products countries already export, Petralia, Balland and Morrison (2017) show that nations have a higher probability of diversifying to related products.

Neffke and Henning (2013) use worker flows between industries to identify similarities in the employment of human capital (skill relatedness) between them. They also discuss diversification pathways firms take, regarding the types of human capital qualification they use. The authors measure relatedness from estimated employee flows through different activities by a regression model that controls for size of industry, average salaries and industry growth rates. For that, they use a database containing the information of 4 million Swedish workers and 415 economic sectors. The study identifies that firms are more likely to diversify their production by entering activities with higher skill relatedness with those it already produces. Hence, via relatedness, the authors demonstrated that the characteristics of human capital are a relevant factor to predict the diversification paths of the firms.

Another example is the work by Balland et al. (2019), which, among other topics, uses patent data from the European Patent Office (EPO), from 1977 to 2011, to calculate relatedness between technological classes of the patents developed by different regions in Europe in an effort to discuss the practiced Smart Specialization Policies. They measure relatedness between two technologies by the frequency they appear simultaneously in patents. As do Boschma, Balland and Kogler (2015), and Hidalgo et al. (2007), the authors use this measure of similarity to identify "the core productions" of each region. This allows them to conclude that it is easier for a region to diversify its technological matrix by developing new technologies related to the already existing ones.

Also focused on a regional perspective, the work of Freitas (2019) introduces a measure of similarity between two sectors of the economy based on three distinct aspects. The first is equivalent to Hidalgo et al. (2007), called co-localization, i.e., the more frequently two sectors are concurrently present in regions, the more related they are. The second, called co-occupation, relates sector according to the amount of common occupations between them. The third, co-corporation, link sectors by their recurrence in a single firm (often, firms are categorized in more than one sector). His final relatedness indicator is a composition of the previous three. The author uses it to understand how the diversification process of Brazilian microregions is associated with their productive trajectories. The results imply that novel economic sectors in a region are largely related to those already present.

Expanding on the notion of relatedness, we emphasize that it can be applied to purposes other than exclusively technological and production diversification, as reiterate Hidalgo et al. (2018). As an example, the work by Guevara et al. (2016) employs relatedness to understand the relationship between different academic research fields. This relatedness indicator is built from the likelihood an author publishes in two different areas, with its construction based on data from Google Scholar. The results reveal that relatedness is a better predictor of the research areas individuals, and even institutions, will eventually publish in, when compared to models based on citations.

Countless other publications the make use of the concept of relatedness for various purposes. Many of them concentrating on the topic of diversification in countries or regions[10]. Nonetheless, we did not find publications regarding the relationship between regional industrial structure and the survival and growth of entrant firms. In other words, little is known about how the productive trajectory of regions affects the survival of new firms from the perspective of their industries. Do firms, with higher similarity with what is produced in the region they enter, have a greater chance of surviving? Do they experience larger growth rates?

The only comparable work we could find is by Jara-Figueroa et al. (2018). The authors investigate the impact workers' previous experience have on the survival and growth of pioneer firms in Brazil. This is done by testing two types of related knowledge, associated with human capital. First, the industry-specific knowledge, which assesses worker's previous experience in other industries. Second, the occupation-specific knowledge, which relates an individual's background to his occupation in the firm. They find industry specific knowledge to be the most relevant for the survival of new firms. Though, the authors do not investigate relatedness relative to the existing industrial structure in the microregions. That is specifically what we intend to do in our research.

When addressing the issue of survival and growth of new firms from the perspective of Jacobs or Marshall agglomeration economies that we discussed earlier, the underlying question would be whether an entrant firm ina a region's market will perform better if that region is specialized or if that region is diversified. From a relatedness perspective, the question is directed to another path: will a new firm have better economic performance if it is more similar to the economic structure already existing in the region or if its entry is disruptive in relation to the structure? In other words, do firms that show greater similarity to what is produced in the region they are entering tend to survive longer and show greater growth? We hope that the efforts of this research can contribute to the literature exposed in this chapter.

[10] Other examples of publications utilizing relatedness are Boschma and Iammarino (2009), Neffke, Henning and Boschma (2011), Hidalgo and Castañer (2015), Boschma, Balland and Kogler (2015), Romero et al. (2015), Castaldi, Frenken and Los (2015), Neffke et al. (2018), Britto et al. (2019).

1.5 Survival and Growth of New Firms

Among the earliest research on survival and growth of new firms is the study done by Phillips and Kirchhoff (1989). The authors explore the creation of firms and their growth in the first few years of existence. Through a cohort analysis, they conclude that firms that grow in those initial years have a higher chance of survival, regardless of the amount of growth experienced. Though the authors do not offer a causal explanation, they conclude that growth and survival are intrinsically related. With the same methodology, Dunne, Roberts and Samuelson (1989) study the growth and exit of manufacturing plants in the United States from 1967 to 1977. Among their findings, is the insight that death and growth of plants decrease the larger the initial size.

Adopting a new approach, Audretsch (1991) studies the survival of new firms over a ten year period, from 1976 to 1986, in the United States. He uses logistic regression to investigate what factors cause a newcomer to close or succeed in different time spans. Particularly, the work concludes that a higher market concentration has a positive impact on firm survival on a four year span, while it is not significant for a ten year span. He also finds that innovation in small firms has a positive impact on survival.

Audretsch and Mahmood (1995) build on the previous work by adding firm characteristics absent in the 1991 study, while adopting the proportional hazards model (COX, 1972). They are able to substantiate that higher wages and larger initial establishment size increase chances of survival. This is also supported by Mata and Portugal (1994), who performed a similar study for Portuguese manufacturing firms, which opened in 1983. These authors did not observe a significant impact of industry concentration on survival. However, they detected that larger average firm size in an industry has a negative effect on the survival of newcomers.

In a more recent study, with data for small companies in Brazil, Mizumoto et al. (2010) examine the impact of human capital, social capital and managerial practices in new firms' chances of survival. That is done in two ways, first by modeling the probability of new firms closing with the use of a logistic model, and second by predicting the risk of closure of a company using the Cox proportional hazards model. They conclude that not one single factor of those three is responsible for the failure of new firms, but rather a combination of them.

An investigation of the impact different types of agglomeration externalities have on the survival of new firms is the work of Renski (2011). The author assesses the impact of specialization, urbanization, and diversification externalities on the survival of entrant firms in the years 1994 and 1995 in the United States. With the Cox model, he finds that higher diversity lowers the risk of death in 5 out of the 8 industries in the study, while urbanization externalities were beneficial for 2 industries. Moreover, specialization was also found to positively affect the majority of the industries.

Howell et al. (2018) also uses the approach of agglomeration economies to perform a survival analysis for new companies in China. The authors test effects of five types of externalities, the two we discussed earlier, specialization (Marshall), controlled for by the location quotient, and diversity (Jacobs). Moreover, they also explore related variety, unrelated variety, and city size externalities. Their notion of related variety externalities is conceptually closest to our main variable, relatedness. Though, in that study, similarity between industries is measured by shared inputs from input-output tables. Among the five kinds of externalities tested, they find related variety has the strongest positive impact on survival. Nevertheless, we would like to stress that despite these findings being in favor of our hypothesis, input-output tables do not capture the complete range of similarities between local capabilities needed to support two different related industries.

As did Howell et al. (2018), Ebert, Brenner and Brixy (2019) also evaluate the impact of agglomeration economies on the survival of new firms. The authors measure these effects with six different indexes on the survival chances of 6776 German start-ups born between 2007 and 2011. Two for localization externalities, which measure the concentration of industries in a region and the specialization the region in an industry[11]. Two measures to capture diversity: the Krugman index and the Herfindahl index. One index to capture unrelated variety and another to capture related variety[12]. They find that localization (specialization) externalities are least important, or even detrimental, to innovative firms, while positively affecting the survival of firms that do not bring innovation. Furthermore, neither the diversification and variety indexes were not found relevant for the survival of those firms.

While the previously mentioned works have focused on studying survival and growth of entrant firms, none has taken the approach of the principle of relatedness into consideration. In this regard, Jara-Figueroa et al. (2018) investigate the impact workers' previous experience have on the survival and growth of pioneer firms in Brazil. This is done by testing two types of related knowledge, associated with human capital, to determine which has a greater impact on the success of new pioneer firms, which they define as those who are new to a region in that its industry is not yet present. The first type of knowledge is the industry-specific knowledge, that relates the worker's previous experience to the industry of the new firm. The second kind is the occupation-specific knowledge, which relates an individual's background to his occupation in the pioneer firm.

The framework of relatedness used in that study is different than the chosen for this work. Instead of measuring the relatedness from the coexistence of industries in a region, they

[11] Measured by the Local Quotient. More details on this index in section 2.3.3.
[12] Both indexes are variations of the Shannon diversity index, described in section 2.3.2.

use revealed relatedness labor flows between industries (NEFFKE; HENNING, 2008; NEFFKE; HENNING; BOSCHMA, 2011; NEFFKE; HENNING, 2013). Rigorously, Jara-Figueroa et al. (2018) define industry relatedness as the residual of a linear regression in which the dependent variable is labor flow and the exogenous variables are the size of an industry and its growth rate. Namely, two industries are related when labor flows between them more than it does on average among others, controlled by size and growth.

The authors find that, for pioneer entrant firms, industry related previous knowledge has an impact on survival and growth, while occupation related knowledge does not. In comparison, they find industry related knowledge to be a more relevant predictor of survival and growth rates for pioneer firms than for non pioneer new firms. Furthermore, occupation specific knowledge was found to have a significant impact on survival and growth of non pioneer firms.

1.6 Conclusion

This chapter set the theoretical background necessary to support our study. As we discussed, the principle of relatedness describes an important methodology to measure similarity. In this research, we contribute to the literature by applying this principle to estimate the effect the similarity between industry and regional industrial structure has on the survival and growth of new firms in that region. We expect that the approach of related diversification reveals itself a valuable tool to analyse and comprehend the determinants of regional development.

Moreover, production is crucial for the development of regions, hence the question that remains is how to improve this productive capacity. As presented in this chapter, the debate on which of the two, specialization or diversification externalities, is more beneficial to regional development has inconclusive support. Enters the concept of smart specialization as an alternative to measure the combination of both. We propose to investigate whether this third option better adheres to the reality of new firms in Brazilian microregions. To achieve that, we employ the principles of relatedness to measure the similarity between industry and locality.

The following chapter presents the methodology for our empirical exercise. We present the econometric methods to assess the impact of relatedness on the survival and growth of firms entering a region. With them, we estimate the probability of surviving to the third year, the probability of staying alive and the growth rate. Furthermore, we explain the controls used. Besides the main variables indicated by our revision of the survival literature, we also use a diversity index to control for Jacobs externalities, the Local Quotient as a proxy for agglomeration externalities. Finally, we describe how the economic complexity indexes and relatedness are calculated.

In the subsequent chapters, we perform two empirical exercises. In the first, we estimate the effect of relatedness over all new firms that opened in Brazil between 2010 and 2013. Later, we repeat the exercise, but separating the industries and regions into two groups each, high and low economic complexity.

2 Methodology

2.1 Introduction

This chapter presents the methods and data used. First, we present the econometric models we use on our study. To study the survival of new firms, we test two different models. The first is the logistic regression model, which estimates the probability of a firma surviving to the third year. The second is the Cox Proportional Hazards regression, which estimates the instantaneous risk of death a new firm has. In order to model the growth of firms, we fit a regression model on the growth rate of new firms at the third year, compared to the first. All models are controlled for fixed effects in microregion, industry and year.

Afterwards, we present the Annual Report of Social Information (RAIS) dataset, the most complete data source of formal labor in Brazil. Then, we explain the control variables and how they are constructed. Particularly, the Economic Complexity indexes, the Local Quotient, the Shannon Diversity index, and the Relatedness indicator. We close this chapter by mentioning the python package we developed to calculate most of these indexes.

2.2 Econometric Models

2.2.1 Specification

In this study, we define a econometric specification largely inspired by Jara-Figueroa et al. (2018). For a new firm (f), created in a year (t), our goal is to assess the effect of relatedness between the industry (i) of that firm and the region (m) it is situated on its survival and growth.

The impact on survival is estimated with two different methods. The first is a logistic regression, in which the dependent variable is a dummy indicating whether the newcomer is still operating at the third year $(t + 3)$ since it opened $S^{t+3}_{f,m,i}$, as in equation 2.1.

$$\begin{aligned} S^{t+3}_{f,m,i} = \; & \beta_0 + \beta_1 R^t_{m,i} + \beta_2 LQ^t_{m,i} + \beta_3 MCI^t_m + \beta_4 ICI^t_i + \beta_5 S^t_m \\ & + \beta_6 \ln\left(GDPpc^t_m\right) + \beta_7 \ln\left(\delta^t_m\right) + \beta_8 \% Grad^t_f + \beta_9 \ln\left(n^t_{0\ f}\right) \\ & + \beta_{10} \ln\left(\overline{Sal}^t_f\right) + \mu_m + \upsilon_i + \tau_t + \varepsilon^t_{f,m,i} \end{aligned} \tag{2.1}$$

The second is a Cox Proportional Hazards regression on firms that started operating in 2010 (COX, 1972). Rather than having to use a dummy, this technique allows us to treat duration as a continuous variable. With it, we estimate the impact of our covariates in the chance the firm is still alive in the following time step. We discuss more about this model in section 2.2.3.

The impact on wage growth is estimated with a linear regression, as in equation 2.2. The dependent variable is the number of employees of the entrant firm three years after it opened divided by the number of employees on its first year $G^{t+3}_{f,m,i}$. On this regression, we only use firms that survive to the third year.

$$\begin{aligned} G^{t+3}_{f,m,i} = \; & \beta_0 + \beta_1 R^t_{m,i} + \beta_2 LQ^t_{m,i} + \beta_3 MCI^t_m + \beta_4 ICI^t_i + \beta_5 S^t_m \\ & + \beta_6 \ln\left(GDPpc^t_m\right) + \beta_7 \ln\left(\delta^t_m\right) + \beta_8 \% Grad^t_f + \beta_9 \ln\left(n^t_{0\,f}\right) \\ & + \beta_{10} \ln\left(\overline{Sal}^t_f\right) + \mu_m + \upsilon_i + \tau_t + \varepsilon^t_{f,m,i} \end{aligned} \quad (2.2)$$

In both equations 2.1 and 2.2, $R^t_{m,i}$ is the relatedness of the firms industry i to the other industries present in the region m at its opening year t. $LQ^t_{m,i}$ is the Local Quotient, MCI^t_m and ICI^t_i are the complexity indexes, S^t_m is the Shannon diversity index, $GDPpc^t_m$ is the GDP per capita, δ^t_m is the population density, $\%Grad^t_f$ is the percentage of employees with a university degree, $n^t_{0\,f}$ is the firm's initial size, and $\overline{Sal}^t_f$ is the average salary of all employees in the month of December.

As do Jara-Figueroa et al. (2018) and Boschma, Balland and Kogler (2015), on both regressions, we use three way fixed effects to control for unobserved effects that do not change over region, industry, and time,represented respectively by μ, υ, and τ.

The initial size of new firms has been shown to be relevant for survival (DUNNE; ROBERTS; SAMUELSON, 1989; MATA; PORTUGAL, 1994; AUDRETSCH; MAHMOOD, 1995). Another important firm level control is the average salary, as demonstrated by Jara-Figueroa et al. (2018). Furthermore, Mata and Portugal (2002) show that human capital is an important predictor of the survival of new firms. Hence, to control for the firm's initial stock of human capital, we include the percentage of employees with a university degree.

Because our regional delimitation is defined a priori as administrative regions, we can not distinguish whether high employment is due to an extensive region or high urban agglomeration. Hence, to control for this, we include the population density. Moreover, we add the region's GDP per capita to control for capital stock.

Audretsch (1991) has observed the importance of concentration in the survival of new firms. Here, we make a different specification by, instead, adding diversity to the model in the form of the Shannon diversity index. This variable has the advantage of also controlling for Jacobs type externalities. Furthermore, we include the Local Quotient as a regional specialization index to account for localization externalities.

Finally, in addition to measure the impact of relatedness on the survival of entrant firms, we also test two complexity indexes. We name them the microregion complexity index, which measures the region's intrinsic capabilities, and the industry complexity index, which measures the industry's intrinsic capabilities. Both indexes are inspired by Hidalgo and Hausmann (2009). We further discuss each variable in section 2.3.

2.2.2 Logistic Regression

A positive advantage of the logistic regression model is it allows us to analyze the effect of the covariates on the probability of survival of new firms. The dependent variable in equation 2.1 is a *dummy* that assumes value of 1 if the firm is operating at its third year and 0 if it has closed. OLS is not suited to model a binary random variable because it extrapolates and predicts values outside of the range between 0 and 1. Hence, we follow the literature and choose logistic regression to accomplish this task.

Given a set of *n* exogenous variables $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, the probability of the firm surviving is shown in equation 2.3.

$$P\left(y=1|\mathbf{X}\right) = G\left(\beta_0 + \mathbf{X}\beta\right) \tag{2.3}$$

where G is a function strictly between 0 and 1 for all real values, and β is a vector of coefficients (WOOLDRIDGE, 2010). For the logistic regression model, G is the logistic function in equation 2.4.

$$G\left(z\right) = \frac{e^z}{1+e^z} \tag{2.4}$$

This is a strictly increasing monotonic function, which tends to 0 as z tends to negative infinity and tends to 1 as z tends to positive infinity. The logistic regression is estimated through the Maximum Likelihood method. For each example in the data set, the function returns the probability it will survive to the third year.

The interpretation of the regression results are not as straight forward as in a linear model. However, their sign and significance is enough for our study to identify a causal relationship between relatedness and firm survival.

2.2.3 Cox Proportional Hazards

One challenge of survival analysis is the fact that before every subject has died, naive estimates of lifetime will be biased. This is called right censored data. Nevertheless, we would like to estimate the lifetime and the effects of covariates on life duration before we see the end of life of every subject. We work around this with the logistic regression by estimating the effects over a three year period. However, another solution is to use survival analysis methods, such as the Cox Proportional Hazards model (COX, 1972).

The Cox Proportional Hazards model studies the length of time until failure. This failure time is modeled as a continuous random variable T by the survival function in equation 2.5.

$$F_T(t) = P(T \geq t) \tag{2.5}$$

The regression estimates the instantaneous risk of failure, or the derivative of the survival function, conditional on exogenous variables. This failure rate is defined in equation 2.6.

$$\lambda\left(t\right) = \frac{dF_T(t)}{dt} = \lim_{\Delta t \to 0^+} \frac{P\left(t \leq T < t + \Delta t | t \leq T\right)}{\Delta t} \tag{2.6}$$

Thus, Cox (1972) defines the Proportional Hazards model as equation 2.7.

$$\lambda\left(t\right) = \lambda_0\left(t\right)\exp\left(\mathbf{X}\beta\right) \tag{2.7}$$

where λ_0 is the baseline hazard function. Dividing both sides by the baseline hazard and taking the logarithm, we get equation 2.8, which is what is estimated.

$$\ln\left({}^{\lambda(t)}/_{\lambda_0(t)}\right) = \mathbf{X}\beta \tag{2.8}$$

The actual estimated model is shown in equation 2.9. In order to analyze over the longest possible period, we only include new firms born in the first year of this study, 2010. Moreover, given that we were unable to find an implementation of this model that supports fixed effects, we choose to include dummy variables. Due to the computational intractability arisen from adding an excessive amount of dummy variables, we were not able to include dummies for each microregion and industry. Hence, we opt to include dummies for state (μ_{state}) and industry on a two digit aggregation (υ_{i2d}), rather that the five digits used on the other techniques. All other exogenous variables are the same as in equations 2.1 and 2.2.

$$\begin{aligned}\ln\left({}^{\lambda(t)}/_{\lambda_0(t)}\right)_{f,m,i} = {} & \beta_1 R_{m,i} + \beta_2 LQ_{m,i} + \beta_3 MCI_m + \beta_4 ICI_i + \beta_5 S_m \\ & + \beta_6 \ln\left(GDPpc_m\right) + \beta_7 \ln\left(\delta_m\right) + \beta_8 \%Grad_f + \beta_9 \ln\left(n_{0\,f}\right) \\ & + \beta_{10} \ln\left(\overline{Sal}_f\right) + \mu_{state} + \upsilon_{i2d} + \varepsilon_{f,m,i}\end{aligned} \tag{2.9}$$

Because we do not know the baseline hazard and it is allowed to vary over time, we can only interpret the coefficients as the logarithm of odds ratio. That is, if a covariate is a dummy, its estimated coefficient is the difference of the log-hazard function between the two groups. For a continuous covariate, the interpretation is over a one unit change in its value.

For example, if, for a treatment group dummy, we get a coefficient of -0.05, that means that the difference between treatment versus the control groups' log-hazard is -0.05, and the treatment group is 95.12% as likely as the control group to die in the next time step ($e^{-0.05} = 0.9512$). Hence, we conclude the treatment reduces the chance of death by about 5%.

2.3 Data

We use the Annual Report of Social Information (RAIS) data set, the main source of information about formal labor in Brazil. It consists of information from registered firms regarding their employees. The entirety of the Brazilian formal workforce is reported in this data set. In RAIS, industries are identified by the National Classification of Economic Activity (CNAE) on a 5 digit aggregation level.

These are the steps for preparing the data. The available data comprises the period from 2006 to 2016. For every firm, only employees holding a job position on December 31st are considered. The data set is summarized so that each line represents one company per microregion. The constructed variables for the firm are: average wage per worker, total wage paid by the company, total number of workers, percentage of workers with a university degree. Furthermore, for each region and each industry, we construct the percentage of workers with a university degree, the average salary and the average firm size.

Brazil is divided in a hierarchical geographical structure. Municipalities are grouped in microregions, which in turn are grouped in mesoregions and later states. Finally, the states are grouped in 5 macroregions: North, Northeast, Center-West, Southeast and South (IBGE, 1990). To account for variability in the city level, here we use the microregion aggregation level as the regional unit for our analysis.

The data regarding the information on population, area and GDP was obtained from the Automatic Recovery IBGE System, Sidra (2019), organized by the Brazilian Institute of Geography and Statistics (IBGE). The information is presented in the municipality aggregation level and later we group it to the microregion level. All monetary variables, GDP and salary, are converted to 2010 present values using the Extended National Consumer Price Index (IPCA) compiled by the National System of Consumer Price Indexes (IBGE/SNIPC) also obtained from Sidra (2019).

2.3.1 New firms, Survival and Growth

In order to identify a new firm in a region, we require that it does not have any employees in RAIS for at least 4 consecutive years. Hence, the time span for this study starts in 2010 and ends in 2016. A firm is considered closed if it does not report any employee to the annual RAIS data set.

For the growth model, the dependent variable is constructed as the division of the number of employees three years after firm creation by the initial number of employees. Thus, only four years are available to build this variable, 2010 through 2013.

For the survival models, the dependent variables are constructed in two different ways. First, to create the binary dependent variable of the logistic regression model, we assign 1 to firms that still have employees three years after its creation and 0 otherwise. Again, we can only use firms born in 2010 through 2013. As for the Cox proportional hazards model, it requires the firm's maximum observed age in years and a binary indicator of whether the event happened. In this case, the event is the closure of the new firm. To perform a longer experiment, we only use newcomers from 2010. Hence, this indicator is 0 for all firms that are in the RAIS in 2016 and 1 if they disappear from the data set before 2016.

2.3.2 Diversity

We discussed how Jacobs externalities play a role in the economic growth. According to Henderson, Kuncoro and Turner (1995), diversified regions have a greater a capacity for this type of externality. Also, Glaeser et al. (1992) state that diversity is a measure of the variety of industries present in a region. Hence, in order to also control the impact this has on firm survival, we use the Shannon entropy (SHANNON, 1948), shown in equation 2.10:

$$
\begin{gathered}
H_m = -\sum_i \left(p_i \cdot \ln p_i\right), \\
\text{where } p_i = \frac{n_{m,i}}{\sum_i n_{m,i}}
\end{gathered}
\tag{2.10}
$$

in which i represents an industry, m represents a microregion and $n_{m,i}$ is the number of employees in industry i and microregion m.

Following Freitas (2019), we make an adaptation to H_m. What we actually define as Shannon diversity index is shown in equation 2.11:

$$
S_m = e^{H_m} \tag{2.11}
$$

This index balances the amount of industries and how they are concentrated in a region. More diverse regions will have a larger Shannon index than otherwise, so long as the workforce is divided evenly among industries. Furthermore, the more industries a region has, the larger the entropy tends to be. However if an industry has a large share of the region's workforce, and that region still has numerous other industries, the entropy is potentially lower than another region with less industries present.

2.3.3 Local Quotient

The Local Quotient (LQ) compares a region's share of an activity with the share of that activity in a greater region. It can be used with any base significant for the problem or region studied (ISARD, 1967). The method indicates the degree of specialization a region has in the activity. The higher the LQ, the higher is the specialization.

In measuring how specialized a region is in an industry, we control for the size of that industry in larger cities. In other words, a micro region will be specialized in a CNAE class not simply because of the size of the industry in the city, but by the relative presence of that industry compared to others present in the region. Thus, the LQ controls for the size of the micro region (GLAESER et al., 1992).

> In computing the location quotient, a quantity named in various ways, an investigator can use any base he considers significant for the problem and region under study (ISARD, 1967, p. 124).

In this study, we are interested in the level of specialization a region has on a given CNAE class, hence the quantity of interest is the total number of employees of that class in a micro region. Following the adaptation by Freitas (2019), equation 2.12 shows how we calculate the Local Quotient:

$$LQ_{m,i} = \frac{n_{m,i} \big/ \sum_i n_{m,i}}{\sum_m n_{m,i} \big/ \sum_m \sum_i n_{m,i}} \tag{2.12}$$

in which i represents a CNAE class, m represents a microregion and $n_{m,i}$ is the number of employees of CNAE i in micro region m.

2.3.4 Complexity Indexes

One might expect the aviation industry to be more sophisticated than retail. One possible way to verify that, would be to count the number of places that produce airplanes and compare with the number of places where there are stores. The former number is much smaller than the latter. However fast to compute this approach may be, it is easy to see its limitations, for example, vanilla crops are only found in few places. The same argument follows when thinking of how capable a microregion is to accommodate different industries.

Hidalgo and Hausmann (2009) proposed a method to measure the complexity of products and countries. The Product Complexity Index and the Economic Complexity Index are calculated extracting the intrinsic information regarding the amount of capabilities needed to manufacture a product and how capable a country is to vary its production.

We follow their guidelines to construct the Industry Complexity Index (ICI) and the Microregion Complexity Index (MCI). First, we must define when an industry is present in a microregion. This is done by constructing a matrix (M) with 1 where the Local Quotient is greater than one and 0 otherwise as in equation 2.13.

$$M_{m,i} = \begin{cases} 1 & LQ_{m,i} \geq 1 \\ 0 & otherwise \end{cases} \tag{2.13}$$

Each row of the matrix represents a microregion and each column a CNAE class. It follows that for every row, the sum of all columns amounts to the quantity of different industries present in that microregion. Likewise, for every column, the sum of all rows equals the number of places that CNAE class is present. This yields equations 2.14 and 2.15.

$$U_{i,0} = \sum_{m} M_{m,i} \tag{2.14}$$

$$D_{m,0} = \sum_{i} M_{m,i} \tag{2.15}$$

$U_{i,0}$ is the ubiquity of industry i, i.e., in how many microregions is that industry relevant. $D_{m,0}$ is the diversity of the microregion.

An industry will be more sophisticated if it is present in few places and if those places are capable of housing multiple industries, hence a more capable microregion. Equivalently, a region is more capable when it is home to multiple industries and those industries are rare. These corrections make it possible to attain the intrinsic information. They are represented in the recursions in equations 2.16 and 2.17

$$U_{i,N} = \frac{1}{U_{i,0}} \sum_{m} M_{m,i} \cdot D_{m,N-1} \tag{2.16}$$

$$D_{m,N} = \frac{1}{D_{m,0}} \sum_{i} M_{m,i} \cdot U_{i,N-1} \tag{2.17}$$

After inserting 2.17 in 2.16, we get:

$$U_{i,N} = \frac{1}{U_{i,0}} \sum_m M_{m,i} \cdot \frac{1}{D_{m,0}} \sum_{i'} M_{m,i'} \cdot U_{i',N-2} \tag{2.18}$$

$$U_{i,N} = \sum_{i'} \left(\sum_m \frac{M_{m,i} M_{m,i'}}{U_{i,0} D_{m,0}} \right) \cdot U_{i',N-2} \tag{2.19}$$

Notice that equation 2.20,

$$\sum_m \frac{M_{m,i} M_{m,i'}}{U_{i,0} D_{m,0}} \tag{2.20}$$

is a square matrix and that, in equation 2.19, $U_{i,N}$ is the eigenvalue associated with the largest eigenvector of said matrix. However, this gives us no information since all values of the eigenvector are equal, because equation 2.19 is satisfied when $U_{i,N} = U_{i,N-2}$ and the associated eigenvalue is 1. Thus, we select the eigenvalue corresponding to the second largest eigenvector, which yields the intrinsic industry's intrinsic information about its complexity. The actual index is taken after this eigenvector is standardized. The same argument follows in order to obtain the Microregion Complexity Index.

2.3.5 Relatedness

Following Boschma, Balland and Kogler (2015), we define relatedness as the density index proposed by Hidalgo et al. (2007), which is intended to measure the proximity of a new product to the products a country is currently able to manufacture. In our case, the industry of the entrant firm is not necessarily new to the region. We work around this issue by not computing the economic activity of the entrant firm in our index. Thus, when calculating the relatedness of an industry to a region, the industry will not be considered as present in the region. This might potentially underestimate the influence of relatedness in the survival and growth of new firms. Nevertheless, if this underestimated impact of relatedness is shown to be relevant, it will be significant to our research question.

Before we define the formula for relatedness, we must specify the similarity between two industries. For a pair of industries i and j, the similarity between them is measured as the number of regions both are present divided by the maximum ubiquity of the two. As shown in equation 2.21.

$$\phi_{i,j} = \frac{\sum_m M_{m,i} M_{m,j}}{max\left(U_{i,0}, U_{j,0}\right)} \tag{2.21}$$

in which M is the matrix obtained from the Local Quotient, m is a microregion, $U_{i,0}$ is the ubiquity of industry i, and i and j are distinct economic activities.

Now, we can define our measure of relatedness as in equation 2.22. The relatedness of an industry to a region is the sum of similarities of that industry to all other present in the region divided by the sum of similarities of that industry to all other existing industries. In other words, every economic activity has a certain number of related activities and a level of similarity between the two, let's call it the neighborhood of that industry. The percentage of that neighborhood present in a region is the relatedness of that industry to the region. This index is always between 0 and 1. Hence, a value of 0.25 means that 25% of the neighborhood is present in the region. A value of 1 means that all related economic activities are present in the region.

$$R_{m,i} = \frac{\sum\limits_{j \neq i} M_{m,j} \phi_{i,j}}{\sum\limits_{j \neq i} \phi_{i,j}} \tag{2.22}$$

2.3.6 Econci Python Package

During the course of this work, we developed python code to calculate the main indexes. The code was encapsulated in the python package `econci` [1]. This package has a main focus on economic complexity as presented in Hausmann et al. (2014). Hence, it does not feature all of the variables we use.

Nevertheless, `econci` calculates the Microregion Complexity index, the Industry Complexity index, the Relatedness and the Local Quotient. These indexes are referred to in Hausmann et al. (2014) as the Economic Complexity index, the Product Complexity index, the Density of a product in a country, and the Revealed Comparative Advantage (BALASSA; NOLAND, 1989) respectively.

2.4 Conclusion

The purpose of this chapter was to give the reader an understanding of the methods employed in this work, as well as justifying their use. To evaluate the impact relatedness has on the probability of survival of new firms, we use the logistic regression, in which the dependent variable is a dummy indicating whether the firm survived to the third year. With the Cox model, we take it one step further and estimate the effect of relatedness on the continuous risk of death. In other words, at any moment in a company's life span, we estimate the impact the similarity of its industry to the region has on the chance of closure. Finally, once we establish the effect on survival, remains the question as to

[1] For documentation, refer to https://github.com/phcsoares/econci

whether relatedness has an impact on the growth of this new firm. To assess that, we employ linear regression with the growth rate at third year as the dependent variable and control industry, microregion and year fixed effects.

In the following chapter, we apply these methods in two empirical exercises. In the first, we estimate the effect of relatedness over all new firms that opened in Brazil between 2010 and 2013. Later, we repeat the exercise, but separating the industries and regions into two groups each, high and low economic complexity.

3 Empirical Investigation: Firm Survival and Growth

3.1 Introduction

In this chapter, we interpret and discuss our results. First, we analyze the data and present the geographical distribution of economic indicators over the Brazilian territory. Here, we give the reader an overview of the data and how wealth, economic complexity, and the number of new firms are distributed across Brazil throughout the years.

Afterwards, we present the results for both survival methods. With the logistic regression, we estimate the impact of relatedness on the probability of survival of entrant firms. With the Cox Proportional Hazards regression, we estimate the impact of relatedness on the chance of this new firm to remain alive. Subsequently, we bring the results for the growth model, that is, we estimate the impact of relatedness on the growth rate of newcomers. Finally, we close the chapter with a discussion of our main results.

3.2 Descriptive statistics

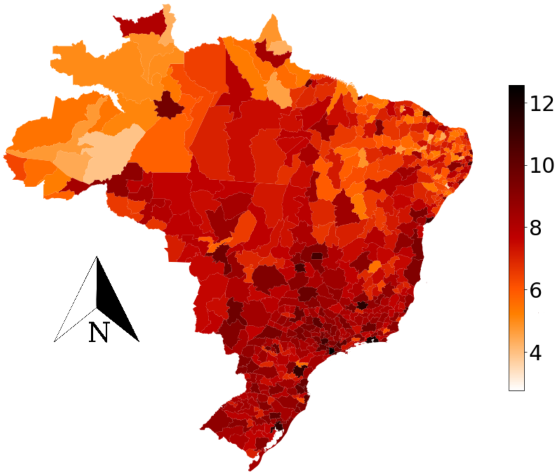
To identify new firms, we require that they do not appear in the data base for at least 4 years. Hence, our study starts in 2010. Furthermore, because we are interested in the survival rate at the third year, the last year we look at entrant firms is 2013. Thus, our study comprises new firms that opened between 2010 and 2013. Table 1 shows the number of entrant firms per year grouped by macroregion. As we observe, the number of new firms stays roughly the same throughout the years. In addition, regions South and Southeast have the largest number of newcomers per year.

Table 1 – Number of new firms per region and year

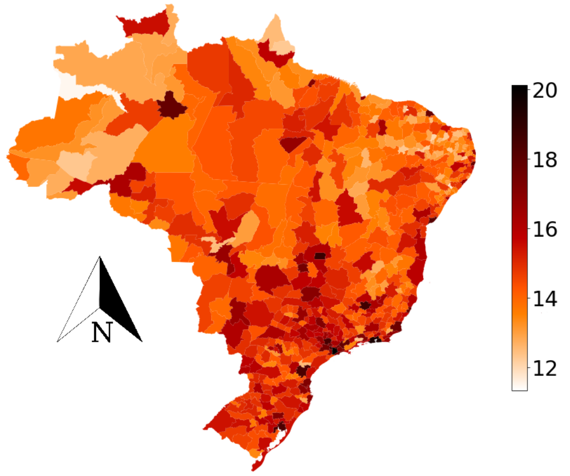
	Entry Year			
Region	2010	2011	2012	2013
Center-West	31,837	32,937	32,109	33,459
North	14,330	15,961	15,418	16,290
Northeast	53,880	57,225	55,953	57,211
South	68,070	67,279	64,086	64,977
Southeast	154,061	157,158	148,472	150,333
Brazil	322,178	330,560	316,038	322,270

Source: Elaborated by the author based on RAIS (2010 - 2013)

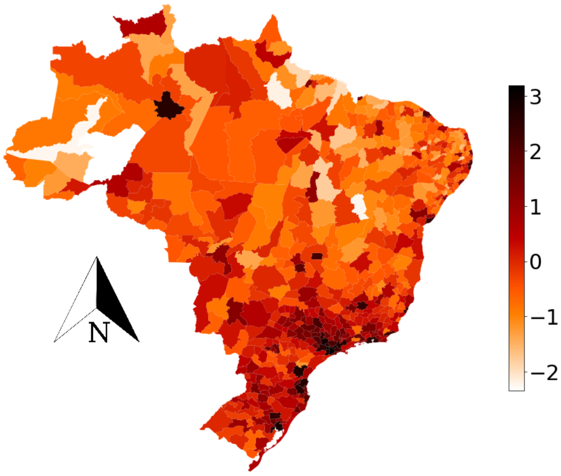
Figure 1 – Microregion indexes



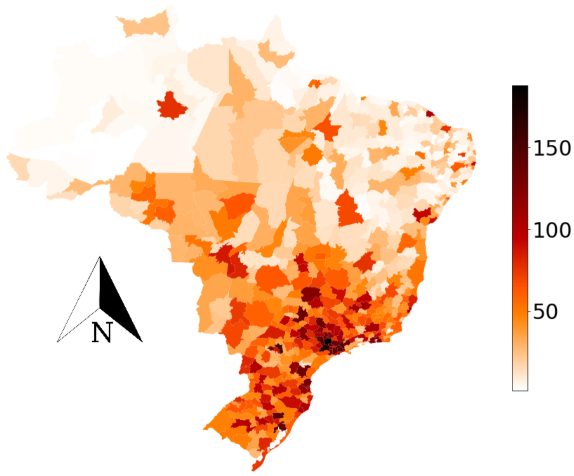
(a) Logarithm of the number of firms



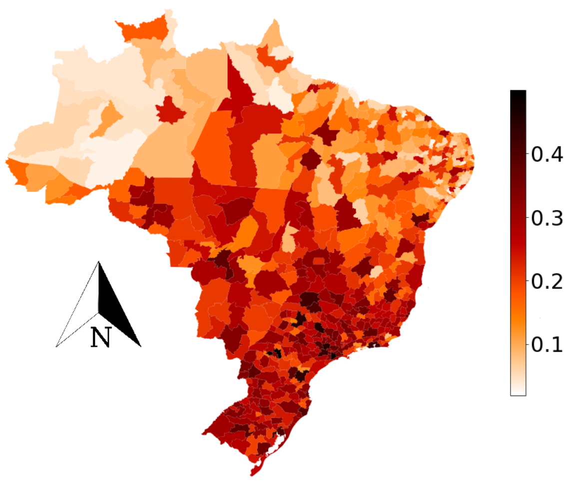
(b) Logarithm of GDP



(c) Microregion complexity index



(d) Diversity index



(e) Average relatedness of new firms

Source: Elaborated by the author

In figure 1, we see a spatial representation of the number of firms in the microregion, the GDP of the region, the microregion complexity index, the diversity index and the average relatedness of new firms for all microregions. All values are for the year 2010, the first of this study. It is worth noting that, on all images, there is a concentration of higher values in the South and Southeast regions of Brazil and of lower values in the North and Northeast regions.

Following the analysis, table 2 brings descriptive statistics for the relatedness of newcomers by macroregion and year, the first two columns, respectively. The table contains the mean, standard deviation, minimum and maximum values, and the quartile boundaries. Both regions Southeast and South have higher relatedness values. While North and Northeast estimates are similar and concentrated on lower values. We also notice that the estimates for Brazil are closer to those of Southeast and South. That is expected since the two regions are the primary economical forces of Brazil and where most new firms are created.

Table 2 – Relatedness descriptive statistics

Region	Entry year	Mean	Std. dev.	Min.	25%	50%	75%	Max.
Center-West	2010	0.2734	0.0753	0.0148	0.2143	0.2815	0.3270	0.4633
	2011	0.2726	0.0806	0.0227	0.2058	0.2807	0.3417	0.4502
	2012	0.2786	0.0789	0.0165	0.2153	0.2864	0.3519	0.4209
	2013	0.2761	0.0868	0.0278	0.1967	0.2890	0.3548	0.4755
North	2010	0.2158	0.0843	0.0088	0.1494	0.2146	0.2741	0.4263
	2011	0.2143	0.0834	0.0110	0.1503	0.2167	0.2710	0.5704
	2012	0.2216	0.0837	0.0038	0.1588	0.2138	0.2849	0.4377
	2013	0.2217	0.0812	0.0120	0.1623	0.2288	0.2756	0.4316
Northeast	2010	0.2368	0.0903	0.0072	0.1707	0.2504	0.3099	0.4796
	2011	0.2339	0.0859	0.0105	0.1748	0.2423	0.3016	0.4467
	2012	0.2393	0.0829	0.0117	0.1817	0.2488	0.3044	0.4273
	2013	0.2426	0.0894	0.0049	0.1763	0.2533	0.3155	0.4856
South	2010	0.3412	0.0907	0.0385	0.2880	0.3440	0.4025	0.5648
	2011	0.3450	0.0930	0.0409	0.2894	0.3466	0.4080	0.5632
	2012	0.3367	0.0871	0.0334	0.2834	0.3417	0.3921	0.5646
	2013	0.3366	0.0888	0.0525	0.2863	0.3391	0.3924	0.6029
Southeast	2010	0.3622	0.1052	0.0252	0.2987	0.3569	0.4205	0.8255
	2011	0.3573	0.1043	0.0207	0.2949	0.3520	0.4148	0.7860
	2012	0.3567	0.1025	0.0274	0.2970	0.3510	0.4142	0.8365
	2013	0.3611	0.1048	0.0278	0.2988	0.3581	0.4206	0.7698
Brazil	2010	0.3215	0.1099	0.0072	0.2511	0.3235	0.3889	0.8255
	2011	0.3181	0.1096	0.0105	0.2459	0.3200	0.3858	0.7860
	2012	0.3174	0.1056	0.0038	0.2480	0.3203	0.3804	0.8365
	2013	0.3192	0.1088	0.0049	0.2485	0.3246	0.3841	0.7698

Source: Elaborated by the author based on RAIS (2010 - 2013)

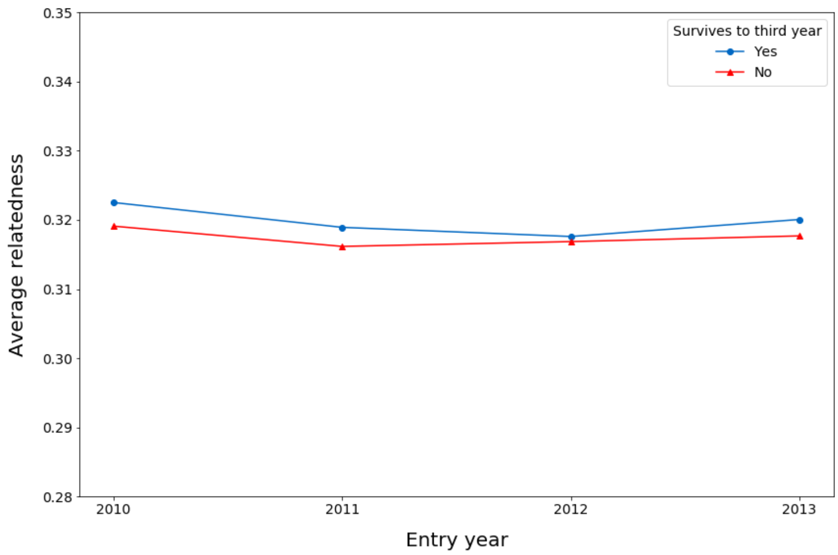
Now we compare the average of relatedness, industry complexity index, microregion complexity index, and the Shannon diversity index for firms that survive to the third year and those that do not. Table 3 shows these averages separated by entry year. We notice that relatedness is the only indicator that has a larger average value for surviving firms on all periods, this relationship is also represented in figure 2. For all other variables, the surviving firms have higher values for 2010 and 2011, and lower values for the other two years. This suggests the causal effects of MCI, ICI and Shannon index on survival, if they exist, might be hard to capture.

Table 3 – Average of indexes conditional on firm survival

		Entry Year			
Index	Survives 3 years	2010	2011	2012	2013
Relatedness	No	0.3191	0.3162	0.3169	0.3177
	Yes	0.3225	0.3189	0.3176	0.3201
MCI	No	1.3816	1.3775	1.3782	1.3535
	Yes	1.4087	1.3975	1.3685	1.3468
ICI	No	-0.6248	-0.5920	-0.5990	-0.5692
	Yes	-0.6067	-0.5884	-0.6047	-0.5876
Shannon index	No	88.3586	89.4985	92.4820	91.8594
	Yes	89.3098	90.0193	92.1266	91.5022

Source: Elaborated by the author based on RAIS (2010 - 2013)

Figure 2 – Average relatedness of new firms by entry year



Source: Elaborated by the author

The dependent variables investigated in this study, survival and growth rate of entrant firms, are summarized in table 4, which contains the mean, standard deviation, minimum and maximum values for each variable. We observe that close to 68% of entrant firms survive for 3 years, nonetheless with a relatively high dispersion. As for the growth rate of employees, on average, the companies have a small decline on their size, though with a large variation of growth rates among them. It is worth mentioning that, on the table, we only compute growth rates for companies that are still in business on the third year.

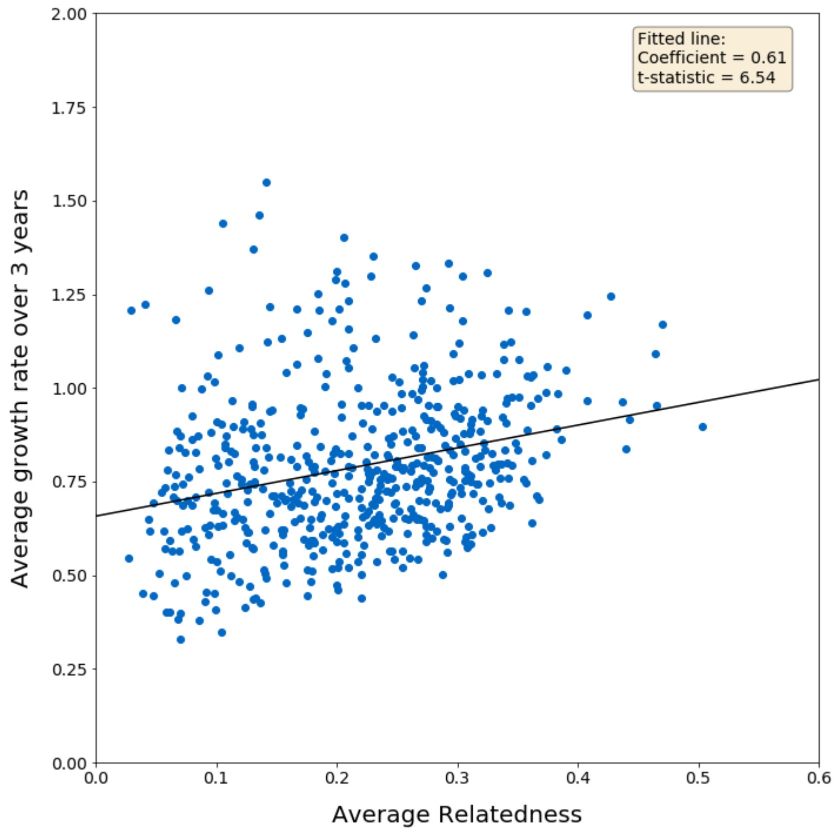
Table 4 – Endogenous variables descriptive statistics.

Variable	Mean	Std. dev.	Min.	Max.
Survival at third year	0.6813	0.4659	0.0000	1.0000
Growth rate at third year	0.9757	8.4987	-0.9994	3,558.00

Source: Elaborated by the author based on RAIS (2010 - 2013)

To better illustrate the relationship between relatedness and the growth rate of new firms, we refer to figure 3. In the image, each point is one microregion, and the coordinates are the average relatedness for new firms per microregion versus the average growth rate of new firms over their initial three years. The fitted line has a positive slope, which indicates a positive correlation between the two axis.

Figure 3 – Scatter plot of growth rates versus regional average relatedness of new firms



Source: Elaborated by the author

Table 5 contains the means of the exogenous variables by year. More specifically, each column represents the means relative to all new firms in that year. The local quotient and the economic complexity index of the regions new firms are born follow an overall negative trend, indicating there are more companies opening in less complex regions. Moreover, the Shannon diversity index is higher in the last couple of years than in the first two. GDP per capita and percentage of graduates follow a positive trend. Finally, population density increases overall while average firm size and salary decreases.

Table 5 – Exogenous variables means by year.

	Entry Year			
Variable	2010	2011	2012	2013
LQ	2.5545	2.5680	2.4721	2.3758
MCI	1.4005	1.3914	1.3716	1.3492
ICI	-0.6122	-0.5895	-0.6029	-0.5813
Shannon index	89.0232	89.8600	92.2412	91.6256
GDP per capita (thousands)	23.1740	24.1159	24.8327	24.9090
Pop. density (pop./km^2)	937.1291	928.9015	947.7102	963.7643
% of graduates	0.0646	0.0650	0.0696	0.0728
Firm size	6.3412	6.3135	6.3347	6.2159
Firm avg. salary	873.11	848.05	849.63	821.48

Source: Elaborated by the author based on RAIS (2010 - 2013) and SIDRA (2010 - 2013)

3.3 Survival Analysis

The impact of relatedness on the survival of new firms is assessed in two ways. The first by a logistic regression on the dependent variable indicating whether a new firm is alive three years after its creation, and the other by the Cox proportional hazards model, which models the risk factors contributing to the company's insolvency.

3.3.1 Logistic Regression

With the Logistic Regression method, we model the survival of the entrant firm three years after its opening (equation 2.1). On all models, the standard errors are clustered by microregion. Furthermore, the table shows the models with fixed effects on the year, microregion and industry. The results are presented in the following order, model 1 considers only relatedness as exogenous, model 2 shows the results for all variables except for relatedness, and model 3 presents the results for relatedness and the control variables together.

Table 6 presents the results of the model on the firm level, that is the percentage of graduates, firm size and average salary are calculated for the firm itself rather than averaged over the region or the industry. We first analyze the estimates of the control variables.

The local quotient (LQ), which measures the specialization of the region in that industry, is not statistically significant in any of the models. That is also the case for the Shannon diversity index.

The microregion complexity index (MCI) has a negative sign and is statistically significant at 1% in model 2 and at 5% in model 3. The magnitude of the effect of MCI in survival rate decreases in the presence of relatedness. Hence there is correlation between the two. The negative sign indicates that regions with more capabilities constitute a harder environment for newcomers.

The industry complexity index (ICI) is statistically significant at 5% and has a positive sign. A positive effect on survival implies that it is easier for firms to survive in more complex industries. The addition of relatedness to the models does not significantly change the magnitude of ICI's impact.

The Shannon index is not statistically significant. While Audretsch (1991) found that market concentration has a positive impact on firm survival, here we do not find evidence of diversity's effect.

We now analyze the regional controls. First, logarithm of GDP per capita is statistically significant at 10% on all models. The positive sign indicates that richer regions gives newcomers a higher chance of survival. For logarithm of population density, none of the coefficients are statistically significant.

The variables on the firm level represent the state of the firm on its first year. All of them are statistically significant at 1% on both models 2 and 3. The percentage of graduates in the new firm, a proxy for human capital, is significant and positive across all models. Hence firms with more educated employees have a larger chance of surviving to the third year. Firm size, here the logarithm of firm size, is likewise positive and significant, thus larger firms reach the third year of existence in greater numbers. The impact of initial firm size is in accordance with the findings of Audretsch and Mahmood (1995) and Mata and Portugal (1994). The logarithm of the firm's average salary is also positive, indicating that higher paying firms have a better chance of survival.

Finally, we analyze the impact that relatedness has on the survival rate at the third year. First, we see that its effect is positive and statistically significant at 1% on models 1 and 3. The proximity of the industry of a new firm has to the local economic structure increases the chance of surviving to the third year. Jara-Figueroa et al. (2018) also evaluate the impact of relatedness, though they measure it by the flow of workers between industries. Nonetheless, they identify a similar impact of relatedness on survival of new firms. Our result are as expected and further supports the argument of related diversification.

Table 6 – Logistic regression results

Dependent Variable:	Survival rate at third year		
Model:	(1)	(2)	(3)
Relatedness	0.3320***		0.2263***
	(0.0620)		(0.0609)
LQ		0.0002	0.0001
		(0.0003)	(0.0003)
MCI		-0.0834***	-0.0766**
		(0.0319)	(0.0317)
ICI		0.0659**	0.0657**
		(0.0264)	(0.0271)
ln(Shannon Index)		0.0658	0.0393
		(0.0749)	(0.0754)
ln(GPD per capita)		0.0937*	0.0951*
		(0.0547)	(0.0554)
ln(Pop. density)		0.2328	0.2194
		(0.3972)	(0.4060)
% of graduates		0.0713***	0.0699***
		(0.0127)	(0.0131)
ln(Firm size)		0.2667***	0.2667***
		(0.0084)	(0.0083)
ln(Firm avg. salary)		0.0923***	0.0909***
		(0.0093)	(0.0092)
Fixed-Effects			
Year	Yes	Yes	Yes
Microregion	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes
Fit statistics			
Observations	1,291,007	1,291,007	1,291,007
Adj-pseudo R^2	0.02042	0.03067	0.0307
BIC	1,614,792.17	1,598,426.17	1,598,415.06

Standard-errors clustered by microregion in parenthesis.
*Signif Codes: ***: 0.01, **: 0.05, *: 0.1*

3.3.2 Cox Proportional Hazards

For the Cox proportional hazards model, we consider only entrant firms in 2010. This method estimates the impacts of the covariates on the chance of death in the following time period, in this case, the chance of closure (equation 2.9). A negative coefficient means a smaller risk of death, thus a greater chance of survival. The value of relatedness and percentage of graduates are multiplied by 100 to make it possible to interpret a 1% change in their values. Furthermore, given that we were unable to find an implementation of this model that supports fixed effects, we choose to include dummy variables. Due to the computational intractability arisen from adding an excessive amount of dummy variables, we were not able to include dummies for each microregion and industry. Hence, we opt to include dummies for state and industry on a two digit aggregation, rather that the five digits used on the other techniques.

Table 7 presents the results for the Cox regression on the age of new firms in years. On all models, the standard errors are clustered by state. Model 1 considers only Relatedness as exogenous variable, model 2 considers all other variables, and model 3 models the firm's risk of closure using all variables. We first analyze the estimates for the control variables.

The Local Quotient (LQ) is not statistically significant on any of the models. This is the same result as in the logistic regression.

The microregion complexity index (MCI) is statistically significant in models 2 and 3. The positive sign means a greater chance of death compared to the baseline hazard. By raising Euler's number to the coefficient we conclude that regions with MCI one standard deviation above average yield a 3% greater chance of death for newcomers. The industry complexity index (ICI), is also relevant in both models 2 and 3. From this, we infer that new firms have a greater chance of death, hence, lower chance of survival the more sophisticated their industries. New firms in an industry with one standard deviation above average ICI have about 5% greater risk of closure.

The Shannon index is only significant in the presence of relatedness. This suggests a negative interaction between them. From model 3, we infer that a one standard deviation more diverse region increases the chance of closure by about 3.4%.

We now analyze the estimates for the microregion GDP per capita and population density. Only GDP per capita is statistically significant at 10%, in model 2, and impacts survival chances negatively. However in the presence of relatedness it is no longer significant. The Cox regression did not detect any impact of population density on the survival of new firms.

All estimates of firm specific variables are statistically significant at 1% in both models 2 and 3. Moreover, percentage of graduates, firm size and average salary all have a negative correlation with the firm's risk of closure. From model 3, a 1% increase in the percentage of graduates in a new firm lowers the risk of death in the following period by about 0.1%. As for firm size and average salary, one unit increase of their logarithms yield, respectively, 12.3% and 5.4% lower risk of closure.

Finally, we discuss the estimates for relatedness, which is statistically significant only in model 3. From model 3, we observe that an increase in relatedness of 1% decreases the risk of death by 0.3%. Moreover, it is also worth noting the interaction between relatedness and diversity. While a more diverse environment is prejudicial for new firms survival, related diversity is actually beneficial.

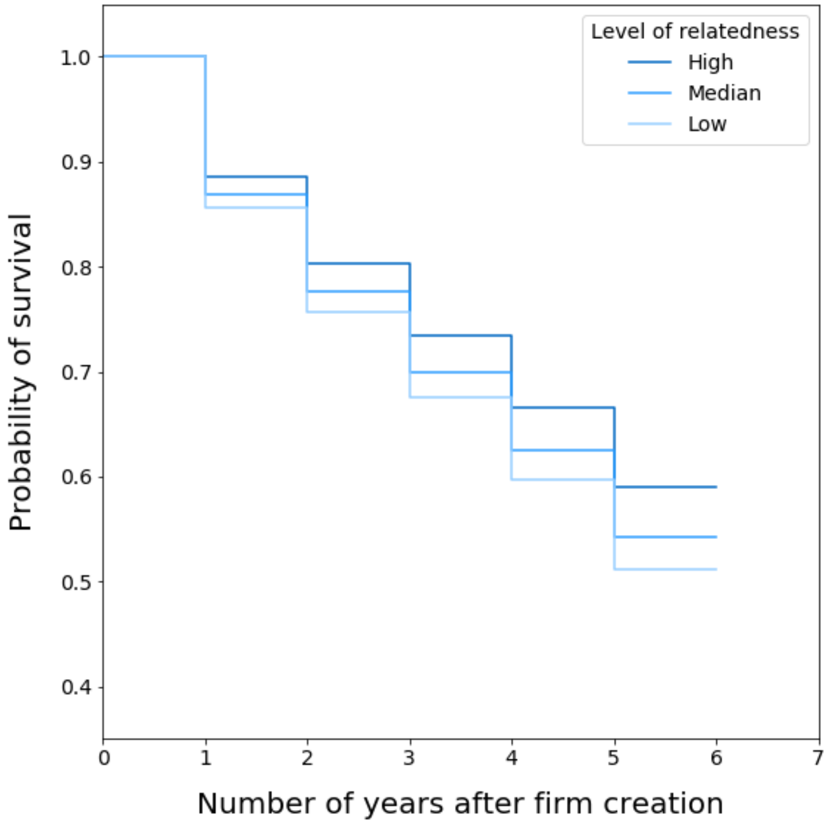
Table 7 – Cox Proportional Hazards results

Dependent Variable:		Firm age in years	
Model:	(1)	(2)	(3)
Relatedness	-0.0004		-0.0029***
	(0.0006)		(0.0006)
LQ		-0.0001	-0.0001
		(0.0003)	(0.0003)
MCI		0.0325***	0.0295***
		(0.0109)	(0.0113)
ICI		0.0513***	0.0491***
		(0.0090)	(0.0073)
ln(Shannon Index)		0.0030	0.0336**
		(0.0133)	(0.0155)
ln(GPD per capita)		0.0313*	0.0223
		(0.0178)	(0.0179)
ln(Pop. density)		-0.0020	0.0042
		(0.0052)	(0.0045)
% of graduates		-0.0009***	-0.0009***
		(0.0002)	(0.0002)
ln(Firm size)		-0.1306***	-0.1307***
		(0.0050)	(0.0049)
ln(Firm avg. salary)		-0.0592***	-0.0560***
		(0.0109)	(0.0108)
Dummies			
State	Yes	Yes	Yes
Industry (2 dig.)	Yes	Yes	Yes
Fit statistics			
Observations	322178	322178	322178
Events	156641	156641	156641
R^2	0.0258	0.0344	0.0346
AIC	3872293.0415	3869458.3160	3869401.8283

Standard-errors clustered by state in parenthesis.
*Signif Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 4 depicts the estimated survival function from table 7 model 3. This function predicts the probability of survival of an individual as time passes. Moreover, we present the average values for three levels of relatedness: high stands for the highest value of the relatedness distribution, low the bottom lowest and median represents the median of the distribution. We use the average value of all other variables to generate the predictions. As expected, firms with higher relatedness have a greater chance of survival, hence their survival functions drop slower. In all, the top of the distribution benefits more from the higher relatedness values than the loss the bottom experiences. That is, a higher value has a larger impact than a lower value.

Figure 4 – Predicted Survival Function



Source: Elaborated by the author

3.4 Growth

We now move on to analyze the impact of relatedness on the growth of new firms, measured by the growth of numbers of employees three years after the firm opened. The method used is linear regression with fixed effects. On all models, the standard errors are clustered by microregion. Similarly to table 6, the table shows the models with fixed effects on the year, microregion and industry. The results are presented in the following order, model 1 considers only relatedness as exogenous, model 2 shows the results for all variables except for relatedness, and model 3 presents the results for relatedness and the control variables together.

Table 8 presents the results of the growth model. We first analyze the estimates of the control variables. The first thing we notice is the change in statistical significance of the Local Quotient and the three indexes when compared to the survival model. For the logistic regression, both LQ and Shannon index were not significant, while MCI and ICI were. The opposite occurs in the growth model.

The local quotient (LQ), which measures the specialization of the region in that industry, is positive and significant at 1%. It indicates that, given an industry, the more specialized the region, the faster new firms grow.

The growth model does not identify the relevance of the microregion complexity index (MCI) and the industry complexity index (ICI) on the growth rate of new firms.

The Shannon index is statistically significant in both models 2 and 3. Moreover, the negative sign indicates that firms have smaller growth rates in more diverse environments. It is worth noting its interaction with relatedness. When relatedness is included in the model, the coefficient for diversity is the one that chances magnitude the most, effectively becoming more negative. This suggests an interaction between the two.

Table 8 – Growth rate linear regression results

Dependent Variable:	Growth rate at third year		
Model:	(1)	(2)	(3)
Relatedness	0.9210***		0.4893**
	(0.1795)		(0.2289)
LQ		0.0071***	0.0070***
		(0.0020)	(0.0020)
MCI		0.0562	0.0710
		(0.1332)	(0.1307)
ICI		0.0040	0.0024
		(0.1402)	(0.1399)
ln(Shannon Index)		-0.3689**	-0.4265***
		(0.1559)	(0.1544)
ln(GPD per capita)		-0.1592	-0.1569
		(0.1714)	(0.1709)
ln(Pop. density)		-1.4115	-1.4430
		(1.0063)	(1.0283)
% of graduates		0.9160***	0.9129***
		(0.1088)	(0.1089)
ln(Firm size)		-0.8234***	-0.8233***
		(0.0252)	(0.0252)
ln(Firm avg. salary)		1.2669***	1.2635***
		(0.0746)	(0.0752)
Fixed-Effects			
Year	Yes	Yes	Yes
Microregion	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes
Fit statistics			
Observations	879,684	879,684	879,684
R^2	0.0018	0.00327	0.00327
BIC	6,281,141.58	6,272,176.96	6,272,195.73

Standard-errors clustered by microregion in parenthesis.
*Signif Codes: ***: 0.01, **: 0.05, *: 0.1*

While neither one of the regional controls are statistically significant for firm growth at third year, all of the firm level variables are. The percentage of graduates in the new firm is significant and positive across all models. Hence firms with more educated employees grow more in their three initial years than others. A firm with 1% more graduates than average is expected to grow 0.9% more after three years. The firm size, here the logarithm of firm size, is negative and significant, thus larger firms grow slower up until their third year, which should not come as a surprise. The firm's average salary has a strong positive influence on growth, thus higher paying new companies also grow faster.

Finally, we analyze the impact relatedness has on growth rate at third year. First, we see that its effect is positive and statistically significant. The proximity of the industry of a new firm has to the local economic structure increases the rate at which firms grow to the third year. A firm opening in a region where its industry is 1% more related than the average is expected to grow 0.5% more reaching the third year. This result is as expected and further supports the argument of related diversification.

3.5 Conclusion

Among our controls variables, the main determinants of survival were the three controls at firm level. Namely, the firm's initial size, its average salary, which controls for how attractive a position in the firm is, and the percentage of graduates, which controls for the stock of human capital the firm initially has. They all have positive impact on survival. As for growth, only the initial size has a negative impact. That is expected, since it is easier to double the size of a 5 person company than a 500 person one.

The economic complexity indexes, ICI and MCI, are both significant for survival and continual existence of the firm. However, their effects are different depending on the situation. First, for the probability of survival to the third year, more complex regions are a harder environment for new firms, while more complex industries are beneficial. Second, for the continuous risk of death over a six year period, both indexes are detrimental to the success of the firm. Furthermore, neither economic complexity index was found to have significant impacts on the growth rate of entrant firms.

While the local quotient and the diversity index were not particularly significant for survival, they were found to be significant for new firm growth. Additionally, they have opposite effects on the growth rate. First, the local quotient, which controls for regional specialization in the new firm's industry, has a positive effect. Second, the diversity index, which controls for Jacobs externalities, has a negative effect on growth. In other words, it is harder for new firms to grow in more diverse and competitive environments.

As for the similarity of the newcomer's industry and the existing regional economic activity, it was statistically significant on all models. More importantly, it has a positive effect on both survival and growth. This corroborates to our hypothesis that related diversification is a more advantageous path for regional economic development.

Differently from the works of Howell et al. (2018) and Ebert, Brenner and Brixy (2019), specialization (Local Quotient) and diversity (Shannon index) were not found significant for the survival of new firms. There are two possible reasons for it. First, those authors did not control for neither regional nor industry complexity. The composition of these indexes could partly account for the effects of said agglomeration externalities. Second, as we explore in chapter 4, the Local Quotient appears to impact firms differently according to the level of technology in the region. Likewise, the Shannon index presents the same behavior when we divide the entrant firms by level od industry complexity.

In all, our results support the hypothesis of smart specialization (BALLAND et al., 2019). Furthermore, relatedness is a valid alternative to measure agglomeration externalities, besides specialization and diversification. Both in the context of giving new firms a greater chance of surviving and larger growth rates.

In the following chapter, we conduct a similar empirical exercise. Though, we are interested on the effect of relatedness by the level of economic complexity of industry and region. Particularly, we investigate the impact relatedness has on new firms in a higher and a lower complexity microregion (industry).

4 Empirical Investigation: Firm Survival and Growth by Level of Economic Complexity of Industries and Microregions

4.1 Introduction

In this chapter we repeat the exercise done in chapter 3. However, we are now focused on the effect of relatedness in different regions of the distributions of the Microregion Complexity Index (MCI) and the Industry Complexity Index (ICI). Our main objective is to verify whether relatedness has different impacts on survival and growth of new firms for higher versus lower sophistication of industries and microregions.

For this exercise, we divide the industries and microregions into the top 50% and the bottom 50% of their distributions by year. That is, for every year, we group the microregions (and industry) in two groups, the upper half of the distribution of MCI and the lower half. Later, we perform our analysis on both groups and compare the results between them. Hence it is possible for a region (or industry) to be in present in the one group in the first year and in another group the following year. Nevertheless, the majority of microregions (and industries) remains in the same group throughout the four years of our study. Furthermore, each firm is a single observation in our research. Thus, for all variables, we only consider the values for the year the firm was created.

4.2 Descriptive statistics

Before we move on to the regression results, table 9 displays the number of entrant firms by year and by group. For MCI, there are on average 7 times more new firms in the top half of the distribution than on the bottom half, which is in accordance with the concept of this complexity index. Regions with higher values of MCI are more sophisticated, have more capabilities and are more diversified. All of this contributes to there being more opportunities for new companies to appear.

Table 9 – Number of new firms per Complexity Index and entry year

	MCI		ICI	
Entry Year	Bottom 50%	Top 50%	Bottom 50%	Top 50%
2010	39,205	282,973	242,454	79,724
2011	41,270	289,290	245,892	84,668
2012	37,719	278,319	232,797	83,241
2013	42,189	280,081	235,368	86,902

Source: Elaborated by the author based on RAIS (2010 - 2013)

An analogous behavior is observed for the two groups of the Industry Complexity Index distribution. However, we see the opposite for this index, there are 3 times more firms opening on less sophisticated industries. This suggests that higher complexity industries are harder for entrepreneurs to venture into. Also, if an industry is more sophisticated, it is expected that it has a larger entry barrier.

Table 10 – Survival and growth rates per Microregion Complexity Index group

	Survival rate		Growth rate	
Entry Year	Bottom 50%	Top 50%	Bottom 50%	Top 50%
2010	68.99%	69.99%	83.97%	112.13%
2011	69.05%	69.46%	79.04%	104.9%
2012	68.73%	67.64%	72.28%	99.95%
2013	66.56%	65.29%	61.7%	85.28%

Source: Elaborated by the author based on RAIS (2010 - 2013)

Table 11 – Survival and growth rates per Industry Complexity Index group

	Survival rate		Growth rate	
Entry Year	Bottom 50%	Top 50%	Bottom 50%	Top 50%
2010	69.66%	70.51%	97.45%	142.66%
2011	69.18%	70.09%	92.19%	128.93%
2012	67.67%	68.06%	89.37%	116.73%
2013	65.37%	65.7%	74.48%	102.8%

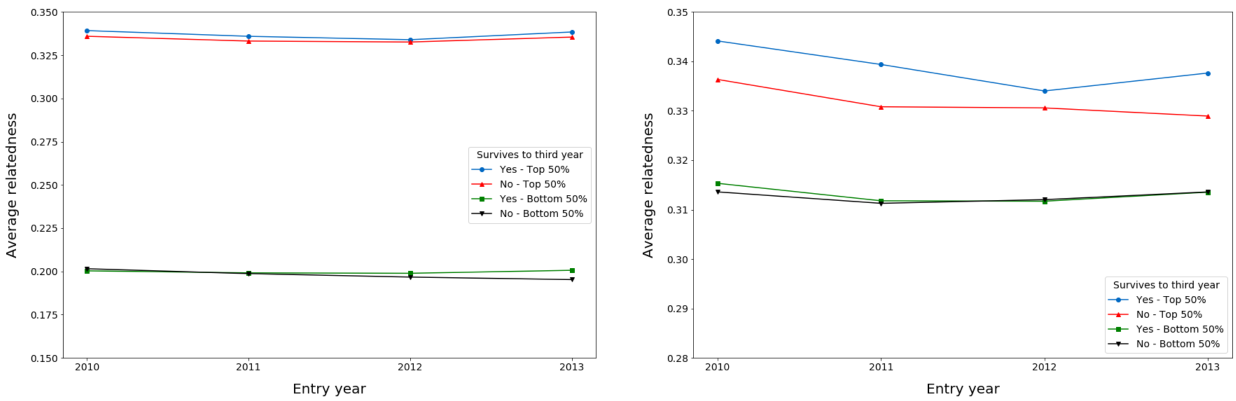
Source: Elaborated by the author based on RAIS (2010 - 2013)

Tables 10 and 11 bring the survival and growth rates of the new firms by year for MCI and ICI groups respectively. On both tables, the survival rate of the top and bottom halves of the distributions are very similar. As for the growth rates, there is a clear difference. For both, ICI and MCI, firms in the top half of the distribution experience larger growth. Moreover, all values present a negative trend across the years.

From a first look at figure 5, we see that relatedness impacts survival differently, depending on the group. Figure 5a shows the average relatedness of new firms by level of

Microregion Complexity for the years of the study, as well as for firms that survived to the third year and those that did not. Although, the top 50% of the distribution having considerably higher average relatedness than the bottom 50%, there is only a small difference between those that survived versus those that did not. The first has higher values.

Figure 5 – Average relatedness of new firms by entry year and complexity level



(a) Microregion average relatedness (b) Industry average relatedness

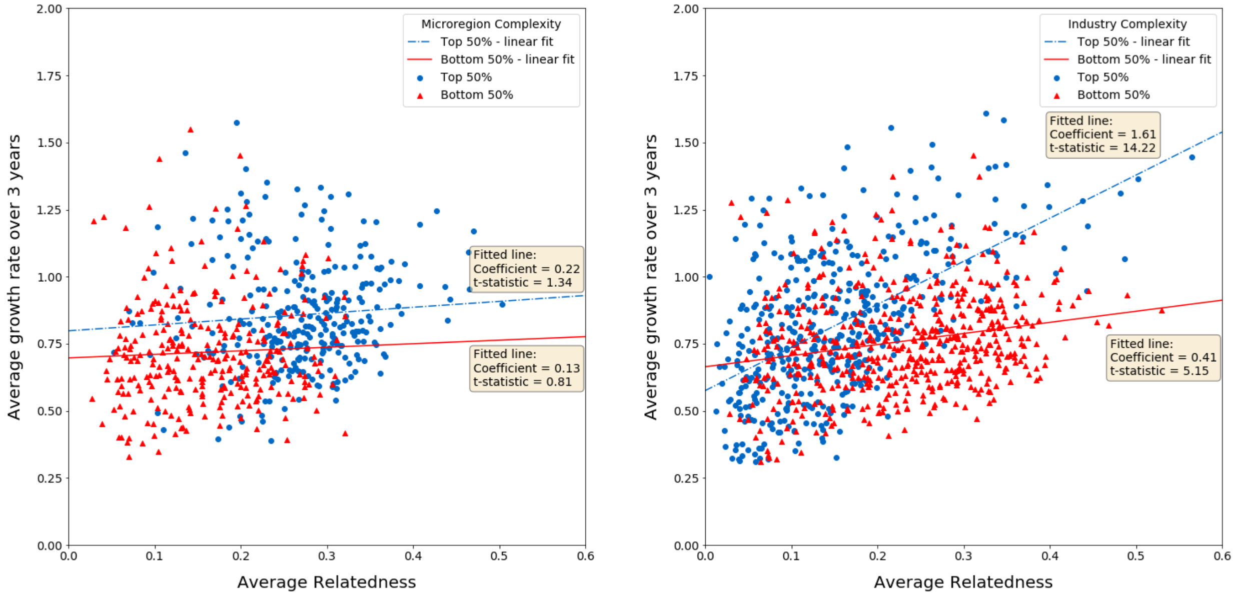
Source: Elaborated by the author

On figure 5b, we have a similar representation, though, this time, the new firms are separated by level of Industry Complexity. It is worth noting that, differently from the previous split, there is a larger distance between the averages of relatedness of survivors compared to non survivors for the top 50%. Furthermore, the distance to the bottom 50% is smaller. For both images, the average relatedness of firms that survive three years and those who do not is very similar in the lower half of the distribution of the complexity indexes.

A visual analysis of the relationship between relatedness and the growth rate of new firms by level of industry and microregion complexity is depicted in figure 6. On the left, the microregions are separated by level of MCI. We observe a slight different in the slopes of the fitted lines, with the line for the top 50% of the distribution having a higher slope. Nevertheless, neither inclination is statistically significant, hence we can not affirm there appears to be a correlation between average relatedness and new firm growth. On the right side, each point is the average of growth rates of new firms in an industry. We observe a much larger difference in the inclination of the fitted line than we had for microregions. Moreover, both inclinations are statistically significant. This leads us to infer a possible correlation between relatedness and firm growth. Finally, the similarity between firm and region appears to have a larger impact in growth for new firms in more complex industries.

In the following sections, we analyse the regression results for both index groups. First, we present the results for the Microregion Complexity Index and later for the Industry complexity index.

Figure 6 – Growth rates versus relatedness of new firms by complexity level



(a) Microregion average growth rate

(b) Industry average growth rate

Source: Elaborated by the author

4.3 Regression results by MCI level

In this section, we analyse the regression results by Microregion Complexity Index group. Here, we compare the results of models estimated on the group of firms that opened in microregions with MCI value in the lower half of the distribution of MCI by entry year. On all tables, models (1), (2) and (3) represent the results estimated on the bottom 50%, while models (4), (5) and (6) are the estimates for the top 50% of the distribution. As before, models (1) and (3) consider only relatedness as exogenous, models (2) and (5) show the results for all control variables, and models (3) and (6) have the results for all variables. Furthermore, all models are controlled for fixed effects. First, we analyse the results of the regressions modelling survival, logistic regression and Cox Proportional Hazards regression. Later, we analyse the results of the linear model for the growth rate.

We use the logistic regression method to model the survival of the entrant firm three years after its opening (equation 2.1). On all models, the standard errors are clustered by microregion. Furthermore, the table shows the models with fixed effects on the year, microregion and industry. Table 12 presents the results for this method. We first analyze the estimates of the control variables.

The local quotient (LQ), which measures the specialization of the region in that industry, is statistically significant in all models it is considered. This did not happen in the previous chapter. Furthermore, the results show that specialization has a negative effect on survival of new firms in less complex microregions, while it has a positive effect in more complex regions. This dichotomy may have prevented us from detecting any causal effect before. These results suggest that sophisticated regions are a better environment for new firms, in therms of survival, than least sophisticated ones.

The microregion complexity index (MCI) has a negative sign and is statistically significant at 10% in the bottom 50% of MCI distribution and at 5% in the top 50%. The magnitude of the effect of MCI in survival rate does not change for the bottom half, while it decreases in the presence of relatedness for the top half. Hence there is correlation between the two in the second case. The negative sign indicates that regions with more capabilities constitute a harder environment for newcomers. This was also detected on previous results.

The industry complexity index (ICI) is statistically significant at 1% and has a positive sign for the lower half of the MCI distribution and is not statistically significant in the top half. A positive effect on survival implies that it is easier for firms to survive in more complex industries. The addition of relatedness to the models does not significantly change the magnitude of ICI's impact. The fact that we only observe an effect on the bottom 50% suggests that new firms of more sophisticated industries have a higher chance of survival in least complex regions, compared to least sophisticated industries in the same environment.

Table 12 show that the Shannon diversity index, and the regional controls (GDP per capita, and population density) are not statistically significant in any of the models.

The variables on the firm level represent the state of the firm on its first year. While only firm size is statistically significant for the bottom 50%, all three are significant for the top 50%. The percentage of graduates in the new firm, a proxy for human capital, is significant and positive on models (5) and (6). Hence, firms with more educated employees have a larger chance of surviving to the third year. Firm size, here the logarithm of firm size, is likewise positive and significant on all models, thus larger firms reach the third year of existence in greater numbers. The logarithm of the firm's average salary is also positive and significant for the top half of the distribution, indicating that higher paying firms have a better chance of survival. We note that percentage of graduates (human capital) and salary (an attraction for qualified workers) are significant in more complex regions. These results might indicate that new firms need a higher stock of human capital in complex regions, which might be a demand to compete in those markets.

Table 12 – Logistic regression results - divided in MCI median

Dependent Variable:	Survival rate at third year					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	-0.1574		-0.3308	0.2684***		0.1515**
	(0.2881)		(0.3345)	(0.0634)		(0.0645)
LQ		-0.0008**	-0.0007*		0.0011***	0.0010***
		(0.0004)	(0.0004)		(0.0004)	(0.0004)
MCI		-0.1218*	-0.1218*		-0.0813**	-0.0755**
		(0.0638)	(0.0637)		(0.0367)	(0.0362)
ICI		0.3545***	0.3415***		0.0414	0.0410
		(0.0882)	(0.0903)		(0.0264)	(0.0268)
ln(Shannon Index)		0.0685	0.0943		0.0217	0.0014
		(0.0973)	(0.1057)		(0.0932)	(0.0928)
ln(GPD per capita)		0.1223	0.1174		0.0544	0.0551
		(0.0958)	(0.0966)		(0.0643)	(0.0649)
ln(Pop. density)		0.7943	0.7964		0.2530	0.2460
		(0.5406)	(0.5396)		(0.4574)	(0.4642)
% of graduates		0.0424	0.0416		0.0726***	0.0717***
		(0.0362)	(0.0363)		(0.0138)	(0.0142)
ln(Firm size)		0.3185***	0.3186***		0.2618***	0.2618***
		(0.0093)	(0.0093)		(0.0086)	(0.0086)
ln(Firm avg. salary)		0.0255	0.0262		0.1011***	0.1001***
		(0.0226)	(0.0226)		(0.0101)	(0.0100)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Microregion	Yes	Yes	Yes	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	159,984	159,984	159,984	1,130,622	1,130,622	1,130,622
Adj-pseudo R^2	0.0255	0.03578	0.03579	0.02198	0.03233	0.03235
BIC	214,788.89	212,924.26	212,946.89	1,411,253.37	1,396,811.92	1,396,823.86

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Finally, we analyze the impact that relatedness has on the survival rate at the third year. First, we see that its effect is positive and statistically significant on models 4 and 6, the top 50%, but not significant for the bottom 50%. The proximity of the industry of a new firm has to the local economic structure increases the chance of surviving to the third year. We stress the importance of this result. Given, relatedness is significant in more complex regions, it becomes more important as regions develop.

For the Cox proportional hazards model, we consider only entrant firms in 2010. This method estimates the impacts of the covariates on the chance of death in the following time period, in this case, the chance of closure (equation 2.9). A negative coefficient means a smaller risk of death, thus a greater chance of survival. The value of relatedness and percentage of graduates are multiplied by 100 to make it possible to interpret a 1% change in their values. Furthermore, given that we were unable to find an implementation of this model that supports fixed effects, we choose to include dummy variables. Due to the computational intractability arisen from adding an excessive amount of dummy variables, we were not able to include dummies for each microregion and industry. Hence, we opt to include dummies for state and industry on a two digit aggregation, rather that the five digits used on the other techniques.

Table 13 presents the results for the Cox regression on the age of new firms in years. Model (1) and (4) consider only relatedness as exogenous variable, models (2) and (5) consider all other variables, and models (3) and (6) estimate the firm's risk of closure using all variables. We first analyze the estimates for the control variables.

Contrary to the logistic regression results in table 12, the Local Quotient (LQ) is not statistically significant on any of the models. This might be due to the difference in time spans between the models, the logistic model evaluates over a fixed three year span, while the Cox model estimates on a continuous 6 years period. Hence, Specialization might be significant in the shorter span, but not in the long run.

The microregion complexity index (MCI) is only statistically significant in the top half of the distribution. The positive sign means a greater chance of death compared to the baseline hazard. By raising Euler's number to the coefficient we conclude that regions with MCI one standard deviation above average yield a 3.7% greater chance of death for newcomers. We conclude that the complexity of regions becomes more relevant to the survival as they develop. And in less sophisticated region, regional complexity does not play a large role on firm survival.

The industry complexity index (ICI), is significant and positive in all models. From this, we infer that new firms have a greater chance of death, hence, lower chance of survival the more sophisticated their industries. From model (6), new firms in an industry with one standard deviation above average ICI have about 4.8% greater risk of closure.

The Shannon index is only significant in the absence of relatedness in the top 50%, with a negative sign. This suggests an interaction between them, if relatedness is not accounted for, its effect is captured by the diversification index.

We now analyze the estimates for the microregion GDP per capita and population density. Only GDP per capita is statistically significant at 10%, in model (5), and impacts survival chances negatively. However in the presence of relatedness it is no longer significant. The Cox regression did not detect any impact of population density on the survival of new firms.

Table 13 – Cox regression results - divided in MCI median

Dependent Variable:	Firm age in years					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	-0.0013		-0.0042	-0.0010***		-0.0024***
	(0.0012)		(0.0022)	(0.0003)		(0.0004)
LQ		0.0007	0.0006		-0.0006	-0.0005
		(0.0005)	(0.0005)		(0.0003)	(0.0003)
MCI		0.0599	0.0694		0.0453***	0.0382***
		(0.0377)	(0.0381)		(0.0093)	(0.0093)
ICI		0.0855***	0.0599**		0.0473***	0.0470***
		(0.0181)	(0.0226)		(0.0060)	(0.0060)
ln(Shannon Index)		0.0103	0.0480		-0.0329**	0.0048
		(0.0226)	(0.0302)		(0.0124)	(0.0142)
ln(GPD per capita)		0.0318	0.0186		0.0246*	0.0192
		(0.0240)	(0.0250)		(0.0110)	(0.0110)
ln(Pop. density)		0.0062	0.0051		-0.0039	0.0020
		(0.0127)	(0.0127)		(0.0041)	(0.0043)
% of graduates		-0.0015**	-0.0015**		-0.0009***	-0.0009***
		(0.0005)	(0.0005)		(0.0002)	(0.0002)
ln(Firm size)		-0.1378***	-0.1379***		-0.1310***	-0.1310***
		(0.0092)	(0.0092)		(0.0030)	(0.0030)
ln(Firm avg. salary)		-0.0398	-0.0379		-0.0608***	-0.0586***
		(0.0255)	(0.0255)		(0.0078)	(0.0079)
Dummies						
State	Yes	Yes	Yes	Yes	Yes	Yes
Industry (2 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	39205	39205	39205	282973	282973	282973
Events	18217	18217	18217	138424	138424	138424
R^2	0.0275	0.0350	0.0350	0.0265	0.0352	0.0353
AIC	374269.6419	373983.0951	373981.5237	3385282.2501	3382757.4854	3382728.0014

Standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All estimates of firm specific variables are statistically significant at 1% for the top half of the distribution, while only salary is not significant for the bottom half. Moreover, percentage of graduates, firm size and average salary all have a negative correlation with the firm's risk of closure. From model (6), a 1% increase in the percentage of graduates in a new firm lowers the risk of death in the following period by about 0.1%. As for firm size and average salary, one unit increase of their logarithms yield, respectively, 12.3% and 5.7% lower risk of closure.

The estimates for relatedness in table 13, which are statistically significant only in the top 50%, as in the logistic regression results. From model (6), we observe that an increase in relatedness of 1% decreases the risk of death by 0.2%. Moreover, as it did in the logistic regression results, relatedness is significant in complex regions.

We now move on to analyze the impact of relatedness on the growth of new firms, measured by the growth of numbers of employees three years after the firm opened. The method used is linear regression with fixed effects. On all models, the standard errors are clustered by microregion. Table 14 shows the models with fixed effects on the year, microregion and industry. The results are presented in the following order, models (1) and (4) consider only relatedness as exogenous, models (2) and (5) show the results for all variables except for relatedness, and models (3) and (6) presents the results for relatedness and the control variables together.

Table 14 presents the results of the growth model. We first analyze the estimates of the control variables. The first thing we notice is the change in sign of the Local Quotient on estimates for the bottom half of the distribution, when compared to the survival model.

The local quotient (LQ), which measures the specialization of the region in that industry, is positive and significant on all models. It indicates that, given an industry, the more specialized the region, the faster new firms grow. For the lower half of the distribution of MCI, regional specialization is detrimental to survival, yet beneficial to growth of new firms.

The growth model does not identify the relevance of the microregion complexity index (MCI) and the industry complexity index (ICI) on the growth rate of new firms.

The Shannon index is statistically significant in both models (5) and (6). Moreover, the negative sign indicates that firms have smaller growth rates in more diverse environments. It is worth noting its interaction with relatedness. When relatedness is included in the model, the coefficient for diversity is the one that chances magnitude the most, effectively becoming more negative. This suggests an interaction between the two.

While neither one of the regional controls are statistically significant for firm growth at third year, all of the firm level variables are, except for percentage of graduates in the bottom half of the distribution. The percentage of graduates in the new firm is significant and positive for the top 50%. Hence firms with more educated employees grow more in their three initial years than others. A firm with 1% more graduates than average is expected to grow 1% more after three years. The firm size, here the logarithm of firm size, is negative and significant, thus larger firms grow slower up until their third year, which should not come as a surprise. Moreover, we note that is has a larger magnitude in the top 50%, which suggests that more complex regions are a harder environment for large firms to grow. The firm's average salary has a strong positive influence on growth, thus higher paying new companies also grow faster, particularly in the top half of the distribution.

Finally, we analyze the impact relatedness has on growth rate at third year. First, we see that its effect is positive and statistically significant only for the top 50% of the distribution of MCI. The proximity of the industry of a new firm has to the local economic structure increases the rate at which firms grow to the third year. From model (6), a firm opening in a region where its industry is 1% more related than the average is expected to grow 0.5% more reaching the third year. Again, as it did in the logistic regression and the Cox regression, relatedness is particularly important for more complex regions.

Table 14 – Growth rate linear regression results- divided in MCI median

Dependent Variable:	Growth rate at third year					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	0.3372		0.0431	0.8378***		0.4219*
	(0.4269)		(0.4991)	(0.2008)		(0.2457)
LQ		0.0068**	0.0068**		0.0055*	0.0053*
		(0.0031)	(0.0032)		(0.0029)	(0.0029)
MCI		-0.0495	-0.0495		0.0592	0.0755
		(0.1014)	(0.1014)		(0.1812)	(0.1775)
ICI		-0.2024	-0.2007		0.0191	0.0172
		(0.1550)	(0.1576)		(0.1529)	(0.1527)
ln(Shannon Index)		0.0825	0.0792		-0.6167***	-0.6732***
		(0.1171)	(0.1166)		(0.1994)	(0.1959)
ln(GPD per capita)		-0.2096**	-0.2090**		-0.1389	-0.1372
		(0.1055)	(0.1044)		(0.2414)	(0.2408)
ln(Pop. density)		-0.6109	-0.6111		-1.338	-1.359
		(1.197)	(1.197)		(1.259)	(1.28)
% of graduates		0.1126	0.1127		0.9956***	0.9932***
		(0.1213)	(0.1218)		(0.1180)	(0.1182)
ln(Firm size)		-0.6446***	-0.6446***		-0.8480***	-0.8479***
		(0.0274)	(0.0274)		(0.0265)	(0.0265)
ln(Firm avg. salary)		0.8848***	0.8847***		1.301***	1.298***
		(0.1086)	(0.1090)		(0.0811)	(0.0817)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Microregion	Yes	Yes	Yes	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	109,546	109,546	109,546	770,138	770,138	770,138
R^2	0.02723	0.05566	0.05566	0.01658	0.02656	0.02657
BIC	589,836.85	586,773.05	586,796.25	5,582,829.60	5,575,187.43	5,575,209.35

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

4.4 Regression results by ICI level

In this section, we analyse the regression results by Industry Complexity Index group. Here, we compare the results of models estimated on the group of firms that opened in industries with ICI value in the lower half of the distribution of ICI by entry year. On all tables, models (1), (2) and (3) represent the results estimated on the bottom 50%, while models (4), (5) and (6) are the estimates for the top 50% of the distribution. As before, models (1) and (3) consider only relatedness as exogenous, models (2) and (5) show the results for all control variables, and models (3) and (6) have the results for all variables. Furthermore, all models are controlled for fixed effects. First, we analyse the results of the regressions modelling survival, logistic regression and Cox Proportional Hazards regression. Later, we analyse the results of the linear model for the growth rate.

We use the logistic regression method to model the survival of the entrant firm three years after its opening (equation 2.1). On all models, the standard errors are clustered by microregion. Furthermore, the table shows the models with fixed effects on the year, microregion and industry. Table 15 presents the results for this method. We first analyze the estimates of the control variables.

The local quotient (LQ), which measures the specialization of the region in that industry, is not statistically significant in any of the models it is considered. This is the same result as in chapter 3.

The microregion complexity index (MCI) has a negative sign and is statistically significant at 5% in the bottom 50% of MCI distribution and is not significant in the top 50%. The magnitude of the effect of MCI in survival rate decreases in the presence of relatedness for the bottom half. Hence there is correlation between the two in the first case. The negative sign indicates that regions with more capabilities constitute a harder environment for newcomers. This was also detected on previous results.

The industry complexity index (ICI) is not statistically significant for the lower half of the MCI distribution and is statistically significant at 1% and has a positive sign for the top half. A positive effect on survival implies that it is easier for firms to survive in more complex industries. The addition of relatedness to the models does not significantly change the magnitude of ICI's impact. The fact that we only observe an effect on the top 50% further supports that the more sophisticated the industry, the higher the chances of survival of new firms. Particularly, compared to table 6.

Table 15 show that the Shannon diversity index, and GDP per capita are not statistically significant in any of the models. While population density is significant at 10% for the top of the distribution of ICI. That is, more complex industries have a greater chance of survival in more densely populated areas.

The variables on the firm level represent the state of the firm on its first year. All of them are statistically significant for the bottom 50%, and for the top 50%. The percentage of graduates in the new firm, a proxy for human capital, is significant and positive on all models. Hence, firms with more educated employees have a larger chance of surviving to the third year. Firm size, here the logarithm of firm size, is likewise positive and significant on all models, thus larger firms reach the third year of existence in greater numbers. The logarithm of the firm's average salary is also positive and significant for both halves of the distribution of ICI, indicating that higher paying firms have a better chance of survival. We note that percentage of graduates (human capital) have a larger impact on more complex industries than least complex ones. On the other hand, salary presents the opposite behavior, it has a larger impact in least complex industries than on more complex ones.

Finally, we analyze the impact that relatedness has on the survival rate at the third year. First, we see that its effect is positive and statistically significant on the bottom 50%, but not significant for the top 50%, when all controls are included (model (6)). The proximity of the industry of a new firm has to the local economic structure increases the chance of surviving to the third year for least complex industries.

Table 15 – Logistic regression results - divided in ICI median

Dependent Variable:	Survival rate at third year					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	0.4532***		0.3807***	0.5311***		0.2282
	(0.1138)		(0.1114)	(0.1462)		(0.1514)
LQ		-0.0000	-0.0001		0.0017	0.0013
		(0.0003)	(0.0003)		(0.0011)	(0.0011)
MCI		-0.0903**	-0.0761**		-0.0640	-0.0636
		(0.0358)	(0.0350)		(0.0633)	(0.0637)
ICI		-0.0212	-0.0199		0.1010***	0.1012***
		(0.0432)	(0.0443)		(0.0364)	(0.0371)
ln(Shannon Index)		0.0646	0.0175		0.0558	0.0308
		(0.0756)	(0.0776)		(0.1137)	(0.1092)
ln(GPD per capita)		0.0954	0.0973		0.0667	0.0622
		(0.0587)	(0.0597)		(0.1045)	(0.1040)
ln(Pop. density)		0.0210	0.0020		0.9007*	0.8728*
		(0.4567)	(0.4725)		(0.4931)	(0.4893)
% of graduates		0.0560***	0.0540***		0.1097***	0.1094***
		(0.0138)	(0.0143)		(0.0230)	(0.0232)
ln(Firm size)		0.2866***	0.2867***		0.2204***	0.2203***
		(0.0075)	(0.0074)		(0.0099)	(0.0098)
ln(Firm avg. salary)		0.1061***	0.1049***		0.0586***	0.0575***
		(0.0092)	(0.0093)		(0.0151)	(0.0148)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Microregion	Yes	Yes	Yes	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	956,468	956,468	956,468	334,443	334,443	334,443
Adj-pseudo R^2	0.02368	0.03485	0.03488	0.01863	0.02682	0.02683
BIC	1,196,358.58	1,183,184.61	1,183,176.21	431,800.47	428,590.99	428,612.11

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

For the Cox proportional hazards model, we consider only entrant firms in 2010. This method estimates the impacts of the covariates on the chance of death in the following time period, in this case, the chance of closure (equation 2.9). A negative coefficient means a smaller risk of death, thus a greater chance of survival. The value of relatedness and percentage of graduates are multiplied by 100 to make it possible to interpret a 1% change in their values. Furthermore, given that we were unable to find an implementation of this model that supports fixed effects, we choose to include dummy variables. Due to the computational intractability arisen from adding an excessive amount of dummy variables, we were not able to include dummies for each microregion and industry. Hence, we opt to include dummies for state and industry on a two digit aggregation, rather that the five digits used on the other techniques.

Table 16 presents the results for the Cox regression on the age of new firms in years. Model (1) and (4) consider only relatedness as exogenous variable, models (2) and (5) consider all other variables, and models (3) and (6) estimate the firm's risk of closure using all variables. We first analyze the estimates for the control variables.

As in the logistic regression results in table 15, the Local Quotient (LQ) is not statistically significant on any of the models. The microregion complexity index (MCI) is statistically significant on both halves of the distribution. The positive sign means a greater chance of death compared to the baseline hazard. By raising Euler's number to the coefficient we conclude (model (6)) that regions with MCI one standard deviation above average yield a 6.6% greater chance of death for newcomers. We conclude that the complexity of regions becomes more relevant to the survival as they develop. And in less sophisticated region, regional complexity plays a smaller role on firm survival.

The industry complexity index (ICI), is significant and positive in all models. From this, we infer that new firms have a greater chance of death, hence, lower chance of survival the more sophisticated their industries. From model (3), new firms in an industry with one standard deviation above average ICI have about 8.4% greater risk of closure. Furthermore, the bottom half of the distribution of ICI is more susceptible to the negative effect higher ICI has on survival.

The Shannon index is positive and significant in model (3), while negative and significant in models (5) and (6). It is worth nothing this contrast. Our results suggest that diversity is detrimental lower complexity industries, while beneficial to higher complexity ones.

We now analyze the estimates for the microregion GDP per capita and population density. Only GDP per capita is statistically significant, and for the bottom 50%, and impacts survival chances negatively. The Cox regression did not detect any impact of population density on the survival of new firms.

All estimates of firm specific variables are statistically significant for both halves of the distribution of ICI. Moreover, percentage of graduates, firm size and average salary all have a negative correlation with the firm's risk of closure. For model (3), considering firm size and average salary, one unit increase of their logarithms yield, respectively, 13.7% and 6% lower risk of closure. It is worth noting that these firm controls are more relevant for the survival of new firms in the bottom 50% of the distribution o ICI.

Table 16 – Cox regression results - divided in ICI median

Dependent Variable:	Firm age in years					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	0.0005		-0.0031***	-0.0007		-0.0027**
	(0.0004)		(0.0006)	(0.0004)		(0.0010)
LQ		-0.0001	-0.0001		0.0005	0.0007
		(0.0003)	(0.0003)		(0.0010)	(0.0011)
MCI		0.0280**	0.0212*		0.0533**	0.0640***
		(0.0092)	(0.0093)		(0.0175)	(0.0179)
ICI		0.0865***	0.0810***		0.0364*	0.0374*
		(0.0081)	(0.0082)		(0.0179)	(0.0179)
ln(Shannon Index)		0.0136	0.0542***		-0.0798***	-0.0637**
		(0.0093)	(0.0120)		(0.0210)	(0.0218)
ln(GPD per capita)		0.0352**	0.0238*		0.0229	0.0181
		(0.0109)	(0.0111)		(0.0214)	(0.0215)
ln(Pop. density)		0.0014	0.0046		-0.0075	0.0036
		(0.0044)	(0.0044)		(0.0077)	(0.0087)
% of graduates		-0.0008***	-0.0008***		-0.0007**	-0.0007**
		(0.0002)	(0.0002)		(0.0003)	(0.0003)
ln(Firm size)		-0.1472***	-0.1475***		-0.0899***	-0.0898***
		(0.0034)	(0.0034)		(0.0053)	(0.0053)
ln(Firm avg. salary)		-0.0652***	-0.0627***		-0.0303*	-0.0279*
		(0.0089)	(0.0090)		(0.0138)	(0.0138)
Dummies						
State	Yes	Yes	Yes	Yes	Yes	Yes
Industry (2 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	242454	242454	242454	79724	79724	79724
Events	117718	117718	117718	38923	38923	38923
R^2	0.0316	0.0419	0.0420	0.0247	0.0292	0.0293
AIC	2841851.0271	2839282.0172	2839255.2996	853611.5391	853261.5810	853255.4460

Standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The estimates for relatedness in table 16, which are statistically significant only in the presence of all controls. From model (6), we observe that an increase in relatedness of 1% decreases the risk of death by 0.3%. Moreover, in opposition to the logistic regression results, relatedness is significant in the top 50%.

We now move on to analyze the impact of relatedness on the growth of new firms, measured by the growth of numbers of employees three years after the firm opened. The method used is linear regression with fixed effects. On all models, the standard errors are clustered by microregion. Table 17 shows the models with fixed effects on the year, microregion and industry. The results are presented in the following order, models (1) and (4) consider only relatedness as exogenous, models (2) and (5) show the results for all variables except for relatedness, and models (3) and (6) presents the results for relatedness and the control variables together.

Table 17 presents the results of the growth model. We first analyze the estimates of the control variables. The first thing we notice is the change in sign of the Local Quotient on estimates for both halves of the distribution, when compared to the survival model.

The local quotient (LQ) is positive and significant on all models. It indicates that, given an industry, the more specialized the region, the faster new firms grow. Furthermore, specialization plays a larger role on growth for firms in the top 50%.

The growth model does not identify the relevance of the microregion complexity index (MCI) and the industry complexity index (ICI) on the growth rate of new firms.

The Shannon index is statistically significant in both models (5) and (6). Moreover, the negative sign indicates that firms have smaller growth rates in more diverse environments. It is worth noting its interaction with relatedness. When relatedness is included in the model, the coefficient for diversity is the one that chances magnitude the most, effectively becoming more negative. This suggests an interaction between the two.

Table 17 – Growth rate linear regression results- divided in ICI median

Dependent Variable:	Growth rate at third year					
Group:	Bottom 50%			Top 50%		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Relatedness	0.6929*		0.1499	2.464***		2.256***
	(0.3924)		(0.4557)	(0.5626)		(0.4718)
LQ		0.0063***	0.0063***		0.0169***	0.0142***
		(0.0020)	(0.0021)		(0.0055)	(0.0051)
MCI		0.1628	0.1684		-0.3586	-0.3536
		(0.1463)	(0.1372)		(0.3077)	(0.3049)
ICI		-0.0502	-0.0513		-0.0056	-0.0054
		(0.2036)	(0.2056)		(0.2199)	(0.2155)
ln(Shannon Index)		-0.0919	-0.1105		-1.86**	-2.105***
		(0.0974)	(0.1033)		(0.7630)	(0.7535)
ln(GPD per capita)		-0.0921	-0.0916		-0.5646	-0.6007
		(0.1479)	(0.1472)		(0.5342)	(0.5338)
ln(Pop. density)		-0.9583	-0.9668		-2.49	-2.776
		(0.7719)	(0.7821)		(2.424)	(2.453)
% of graduates		0.7866***	0.7858***		1.158***	1.155***
		(0.1334)	(0.1348)		(0.2458)	(0.2449)
ln(Firm size)		-0.7587***	-0.7586***		-0.9816***	-0.9828***
		(0.0317)	(0.0317)		(0.0442)	(0.0441)
ln(Firm avg. salary)		1.186***	1.185***		1.414***	1.402***
		(0.1142)	(0.1146)		(0.1203)	(0.1196)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Microregion	Yes	Yes	Yes	Yes	Yes	Yes
Industry (5 dig.)	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	650,377	650,377	650,377	229,307	229,307	229,307
R^2	0.01413	0.02547	0.02547	0.01752	0.02691	0.02696
BIC	4,421,655.44	4,414,341.59	4,414,367.99	1,788,514.79	1,786,510.10	1,786,523.82

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

While neither one of the regional controls are statistically significant for firm growth at third year, all of the firm level variables are. The percentage of graduates in the new firm is significant and positive. Hence firms with more educated employees grow more in their three initial years than others. From model (6), a firm with 1% more graduates than average is expected to grow 1.2% more after three years. The firm size is negative and significant, thus larger firms grow slower up until their third year. Moreover, we note that is has a larger magnitude in the top 50%, which suggests that more complex industries are harder activities for larger firms to grow. The firm's average salary has a strong positive influence on growth, thus higher paying new companies also grow faster, particularly in the top half of the distribution.

Finally, we analyze the impact relatedness has on growth rate at third year. First, we see that its effect is positive and statistically significant only for the top 50% of the distribution of ICI. The proximity of the industry of a new firm has to the local economic structure increases the rate at which firms grow to the third year. From model (6), a firm opening in a region where its industry is 1% more related than the average is expected to grow 2.3% more reaching the third year. Again, as it did in the logistic regression and the Cox regression, relatedness is particularly important for more complex regions.

4.5 Conclusion

As we found in chapter 3, the firm's initial size, its average salary, and the percentage of graduates were highly significant, though with differences. They all have positive impact on survival for low and high complexity industries, however only firm initial size has a positive impact for low complexity regions. As for growth, we see a similar behavior, only the initial size has a negative impact. Furthermore, the percentage of graduates is not significant in low complexity regions.

The economic complexity indexes, ICI and MCI, also behave differently depending on the complexity level. For the probability of a new firm surviving to the third year in low complexity regions, ICI has a positive impact on firm survival, while MCI is not significant. Hence, in these regions, the more sophisticated the industry, the better the chances of survival, regardless how much capabilities are present in the region. While in high complexity regions, MCI has a negative impact and ICI is not significant. Here, the opposite happens, with more capable regions being a harder environment for new firms. Furthermore, the industry type does not seem to be a significant determinant of survival for new firms in regions with a larger stock of capabilities.

For the Cox model, ICI is significant for both high and low complexity regions, while MCI only for high complexity, though both impact survival negatively. In this model, we see a similar result for MCI as in the previous. Although, industry complexity impacts firm survival negatively in a longer time span.

For industry complexity levels, the results of the logistic model are symmetric, with ICI benefiting firms in high complexity industries and MCI being detrimental to firms in low complexity industries. From the Cox model, we conclude that, in a longer time span, both ICI and MCI have negative impacts in the survival of new firms. Neither ICI, nor MCI had significant impact in firm growth. This suggests that, although capabilities play a role in firm survival, they are not determinants of firm growth. Thus, different externalities come into play for firm survival than for firm growth rates.

Interestingly, the local quotient, which controls for specialization externalities, has significant impact on the probability of survival at third year by the level of regional complexity. Though with opposite effects. For low complexity regions, LQ has a negative impact on survival, while it has a positive effect for firms in high complexity regions. This dichotomy might explain why it was not significant in the results of chapter 3. Furthermore, in the growth models, LQ was found to have a positive impact.

Diversity, represented by the Shannon index, does not impact growth in neither low complexity industries or regions. It does however maintain its negative effect in the high complexity groups. Moreover, the results from the Cox models by industry complexity suggest that diversity also has opposite effects on new firms. For the long run survival of the new firm, diversity is detrimental for low complexity industries, but beneficial for forms in high complexity activities. From table 16, we identify again a dichotomy in the behavior of a control for agglomeration externalities, this could explain why diversity was not significant for the survival of new firms in chapter 3.

Finally, while relatedness did not show a dual effect as LQ and diversity did, it does affect high and low complexity differently. It has no effect in the survival and growth of entrant firms in low complexity regions, while it maintains its positive effect in high complexity regions. In other words, similarity is relevant for more developed microregions. As for the groups of industries, relatedness is significant and positive for the probability of survival of firms in low complexity industries, and reduces the risk of death of firms in all industries in the lung run. With respect to growth, relatedness was only found significant and beneficial for firms in high complexity activities and microregions.

As we found in chapter 3, perhaps even more so now, the similarity between industry and region is of great importance for regional diversification. Our results show that new firms in lower complexity industries benefit from the similarity to the present economic activity. Hence, it opens up the opportunity to identify better activities to foster in these regions with an interest in generating jobs. Moreover, Brazil, as a developing country, still has a ways to go in therms of creating new capabilities. From our results, we see that the highest complexity regions in Brazil are tougher environments for new firms, but relatedness makes their path easier. Thus, strategies to develop the country, which would undoubtedly consider the development of said regions, can also use relatedness as a tool.

Final Considerations

This study aimed to identify a causal relationship between the similarity (relatedness) of a new firm's industry and the local industrial structure. Particularly, how this affects the firm's probability of survival and growth rate.

We employ three main methods, two for survival estimation and one for growth. For survival, the first method is logistic regression, in which the dependent variable is a dummy indicating whether the new firm is alive three years after its creation, this lets us estimate the probability a company will survive to the third year. The second survival method is Cox Proportional Hazards estimation, which estimates, at any given moment, the risk of closure in the following time period. The final method, for growth rate estimation, is an OLS regression.

In order to identify the relationship between industry and region, we resorted to the relatedness literature. Among the different ways to measure it, we chose the coexistence metric. It evaluates similarity between industries, so that the new firm's economic activity can be associated with those present in the region it opens. This is a different metric than that used by Jara-Figueroa et al. (2018), which did not take into account the industry versus locality relationship. The authors asses relatedness relative to the knowledge workers bring to the firm.

Our results point to the importance of relatedness to the survival and growth of new firms. Particularly, the higher the similarity between the new firm's industry and the regional economic structure, the higher the probability of surviving to the third year and the faster the newcomer grows. It is also worth noting the interaction between relatedness and diversity (Shannon index). Once relatedness is controlled for, diversity becomes more detrimental to survival and growth.

This indicates that a diverse environment is a harsh one for new firms to grow. However when the diversity is towards related industries, it is in fact beneficial. From the point of view of closely related industries, we can interpret high relatedness in a region as a specialization on the group of related industries. Hence the region is diversified and specialized at the same time. Diversification provides Jacobs externalities while specialization adds Marshallian externalities.

Ultimately, our results give more support for smart specialization policies (BALLAND et al., 2019). Furthermore, we contribute to the literature by bringing evidence that relatedness is a valid alternative to measure agglomeration externalities, aside from the traditional specialization and diversification. This is done by showing that relatedness has an impact in the context of increasing the chances new firms survive and increasing their growth rates. Thus the similarity of industry and region can not be overlooked and should be used as a tool to help determine new, more complex, activities regions should focus on. Which in turn furthers economic growth (HIDALGO et al., 2007).

Over all results, the initial size of newcomers, the control for the stock of human capital in the firm and the average salary are the most relevant of all variables to explain the survival of new firms. Unsurprisingly, the initial size of firms only has a negative impact on the growth rate at third year. Moreover, the complexity indexes (MCI and ICI) significantly impact firm survival but not firm growth, as seen in the Cox and logistic models. Moreover, GDP per capita and population density are hardly ever statistically significant.

When the dataset is divided into higher and lower complexity industries and region, there are interesting differences to be noted regarding the causal effect of relatedness. First, for the regional complexity split, relatedness was only found to be statistically significant for the top 50% of the distribution. In other words, firm's opening in more sophisticated regions benefit most from similarity. Moreover, relatedness does not seem to impact the survival and growth of new firms in least complex regions. Consequently, firms in complex regions seem to benefit more from similarity the similarity to the local economic activity than those in least complex region.

Second, for the industry complexity split, relatedness affects both groups, but in different ways. Looking at survival, the logistic model, which estimates the probability of surviving to the third year, implies that firms similarity is only relevant for the the lower half of the distribution of the Industry Complexity Index (ICI). While the Cox model, which estimates the risk of death in the subsequent time period, finds relatedness to be statistically significant for both halves of the distribution of ICI, albeit less significant for the top 50%. Looking at growth, only the top half of the distribution of ICI benefits from similarity to the local industrial structure. In all, firms in complex industries, need not to open in regions with high relatedness to survive, but if they do chose those locations, they grow faster.

To sum up, this dissertation provides several contributions to the understanding of firm survival and growth: i) relatedness, measured by the coexistence of industries, has positive impact on both survival and growth of new firms; ii) capabilities, measured by the economic complexity indexes ICI and MCI, affect firm survival but not growth; iii) specialization has a positive impact on growth, and a dual effect on survival, depending on how much capabilities the region has, positive for regions with many capabilities and negative otherwise; iv) diversity has negative impact on growth, and a dual effect on long term survival, depending on the level of capabilities the industry has, positive for industries with many capabilities and negative otherwise.

Our results have implications for public policies aiming at regional development. Particularly, those focused on expanding the local industry. Relatedness was shown to be a relevant variant to consider when choosing which economic activities to foster, conditional on the region's already existing industrial structure. Furthermore, as Hidalgo et al. (2007) argues, more complex activities increases economic growth rates, and regions tend to develop economic activities related to those already exiting (JACOBS, 1969; ARTHUR, 2009; HIDALGO et al., 2007; BOSCHMA, 2017; ALSHAMSI; PINHEIRO; HIDALGO, 2018).

As any academic endeavour, there are limitations to our work. First, we measure firm growth by the number of employees. However, there are other business indicators that would also be interesting to test, such as revenue, which might be better signal of success, especially for tech companies and start-ups. Second, it would also be worth controlling for the distance between the firm and the urban center, which takes a considerable effort. Finally, while we use coexistence, Neffke and Henning (2008) and Jara-Figueroa et al. (2018) use labor flows between industries, there are other possible forms to assess relatedness. Today, with the advent of social media networks and ever more available data, it opens up the possibility to explore the connections employees have with workers in other industries.

Nonetheless, despite its limitations, this dissertation provides several contributions to the existing literature, and we hope it serves as incentive to foster further research on those topics.

References

ALSHAMSI, A.; PINHEIRO, F. L.; HIDALGO, C. A. Optimal diversification strategies in the networks of related products and of related research areas. *Nature Communications*, Springer US, v. 9, n. 1, 2018.

ARTHUR, W. B. *The nature of technology: what it is and how it evolves*. [S.l.]: Simon and Schuster, 2009.

ARTHUR, W. B. *Complexity and the Economy*. [S.l.]: Oxford University Press, 2014.

AUDRETSCH, D. B. New-Firm Survival and the Technological Regime. *The Review of Economics and Statistics*, JSTOR, v. 73, n. 3, p. 450, aug 1991.

AUDRETSCH, D. B.; MAHMOOD, T. New Firm Survival: New Results Using a Hazard Function. *The Review of Economics and Statistics*, The MIT Press, v. 77, n. 1, p. 97, feb 1995.

BALASSA, B.; NOLAND, M. "Revealed" Comparative Advantage in Japan and the United States. *Journal of International Economic Integration*, Center for Economic Integration, Sejong University, v. 4, p. 8–22, 1989.

BALLAND, P.-A.; BOSCHMA, R.; CRESPO, J.; RIGBY, D. L. Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, Taylor & Francis, v. 53, n. 9, p. 1252–1268, 2019.

BOSCHMA, R. Relatedness as driver of regional diversification: a research agenda. *Regional Studies*, Routledge, v. 51, n. 3, p. 351–364, mar 2017.

BOSCHMA, R.; BALLAND, P.-A.; KOGLER, D. F. Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, Narnia, v. 24, n. 1, p. 223–250, feb 2015.

BOSCHMA, R.; IAMMARINO, S. Related variety, trade linkages, and regional growth in italy. *Economic geography*, Wiley Online Library, v. 85, n. 3, p. 289–311, 2009.

BRITTO, G.; ROMERO, J. P.; FREITAS, E.; COELHO, C. La gran brecha: complejidad económica y trayectorias de desarrollo del brasil y la república de corea. *Revista CEPAL*, 2019.

CASTALDI, C.; FRENKEN, K.; LOS, B. Related variety, unrelated variety and technological breakthroughs: an analysis of us state-level patenting. *Regional studies*, Routledge, v. 49, n. 5, p. 767–781, 2015.

COMBES, P.-P. Economic Structure and Local Growth: France, 1984–1993. *Journal of Urban Economics*, Academic Press, v. 47, n. 3, p. 329–355, may 2000.

COX, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley, v. 34, n. 2, p. 187–202, jan 1972.

DUNNE, T.; ROBERTS, M. J.; SAMUELSON, L. The Growth and Failure of U. S. Manufacturing Plants. *The Quarterly Journal of Economics*, Oxford University Press (OUP), v. 104, n. 4, p. 698, nov 1989.

EBERT, T.; BRENNER, T.; BRIXY, U. New firm survival: the interdependence between regional externalities and innovativeness. *Small Business Economics*, Springer, v. 53, n. 1, p. 287–309, 2019.

FREITAS, E. *Indústrias relacionadas, complexidade econômica e diversificação regional: uma aplicação para microrregiões brasileiras*. 149 p. Tese (Doutorado) — UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2019.

GLAESER, E. L.; KALLAL, H. D.; SCHEINKMAN, J. A.; SHLEIFER, A. Growth in Cities. *Journal of Political Economy*, The University of Chicago Press, v. 100, n. 6, p. 1126–1152, dec 1992.

GUEVARA, M. R.; HARTMANN, D.; ARISTARÁN, M.; MENDOZA, M.; HIDALGO, C. A. The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics*, Springer, v. 109, n. 3, p. 1695–1709, 2016.

HAUSMANN, R.; HIDALGO, C. A.; BUSTOS, S.; COSCIA, M.; CHUNG, S.; JIMENEZ, J.; SIMOES, A. J. C.; YILDIRIM, M. A. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. [S.l.]: MIT Press, 2014.

HENDERSON, J. V. Efficiency of resource usage and city size. *Journal of Urban economics*, Elsevier, v. 19, n. 1, p. 47–70, 1986.

HENDERSON, V.; KUNCORO, A.; TURNER, M. Industrial Development in Cities. *Journal of Political Economy*, v. 103, n. 5, p. 1067–1090, oct 1995.

HIDALGO, C. A.; BALLAND, P.-A.; BOSCHMA, R.; DELGADO, M.; FELDMAN, M.; FRENKEN, K.; GLAESER, E.; HE, C.; KOGLER, D. F.; MORRISON, A. et al. The principle of relatedness. In: SPRINGER. *International conference on complex systems*. [S.l.], 2018. p. 451–457.

HIDALGO, C. A.; CASTAÑER, E. E. The amenity space and the evolution of neighborhoods. *arXiv preprint arXiv:1509.02868*, 2015.

HIDALGO, C. A.; HAUSMANN, R. The Building Blocks of Economic Complexity. n. 186, p. 22, 2009.

HIDALGO, C. A.; KILINGER, B.; BARABÁSI, A.-L.; HAUSMANN, R. The Product Space Conditons the Develpment of Nations. *Science*, v. 317, n. July, p. 482–487, 2007.

HIRSCHMAN, A. O. *The Strategy of Economic Development*. New Haven: Yale University Press, 1958.

HOWELL, A.; HE, C.; YANG, R.; FAN, C. C. Agglomeration,(un)-related variety and new firm survival in china: Do local subsidies matter? *Papers in Regional Science*, Wiley Online Library, v. 97, n. 3, p. 485–500, 2018.

IBGE. *Divisão Regional do Basil em Mesorregiões e Microrregiões Geográficas*. Rio de Janeiro, 1990.

ISARD, W. *Methods of regional analysis*. Cambridge: The M.I.T. Press, 1967.

JACOBS, J. *The economy of cities*. New York: Vintage Books, 1969.

JARA-FIGUEROA, C.; JUN, B.; GLAESER, E. L.; HIDALGO, C. A. The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, v. 115, n. 50, p. 12646–12653, dec 2018.

KRUGMAN, P. Increasing returns and economic geography. *Journal of political economy*, The University of Chicago Press, v. 99, n. 3, p. 483–499, 1991.

LEMOS, M. B. *Espaço e Capital: um estudo sobre a dinâmica centro x periferia. Campinas*. Tese (Doutorado) — Instituto de Economia da Universidade Estadual de Campinas, 1988.

LÖSCH, A. *The economics of location*. New Haven: Yale University, 1974.

MARSHALL, A. Industrial organization, continued. the concentration of specialized industries in particular localities. In: *Principles of economics*. [S.l.]: Springer, 2013. p. 222–231.

MATA, J.; PORTUGAL, P. Life Duration of New Firms. *The Journal of Industrial Economics*, Wiley, v. 42, n. 3, p. 227, sep 1994.

MATA, J.; PORTUGAL, P. The survival of new domestic and foreign-owned firms. *Strategic Management Journal*, John Wiley & Sons, Ltd, v. 23, n. 4, p. 323–343, apr 2002.

MIZUMOTO, F. M.; ARTES, R.; LAZZARINI, S. G.; HASHIMOTO, M.; BEDÊ, M. A. A sobrevivência de empresas nascentes no estado de São Paulo: um estudo sobre capital humano, capital social e práticas gerenciais. *Revista de Administração*, Elsevier, v. 45, n. 4, p. 343–355, oct 2010.

MOOMAW, R. L. Is population scale a worthless surrogate for business agglomeration economies? *Regional Science and Urban Economics*, Elsevier, v. 13, n. 4, p. 525–545, 1983.

NEFFKE, F.; HARTOG, M.; BOSCHMA, R.; HENNING, M. Agents of Structural Change: The Role of Firms and Entrepreneurs in Regional Diversification. *Economic Geography*, Routledge, v. 94, n. 1, p. 23–48, jan 2018.

NEFFKE, F.; HENNING, M. Revealed Relatedness: Mapping Industry Space. *Papers in Evolutionary Economic Geography (PEEG)*, Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography, 2008.

NEFFKE, F.; HENNING, M. Skill relatedness and firm diversification. *Strategic Management Journal*, John Wiley & Sons, Ltd, v. 34, n. 3, p. 297–316, mar 2013.

NEFFKE, F.; HENNING, M.; BOSCHMA, R. How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions. *Economic Geography*, Wiley/Blackwell (10.1111), v. 87, n. 3, p. 237–265, jul 2011.

PETRALIA, S.; BALLAND, P.-A.; MORRISON, A. Climbing the ladder of technological development. *Research Policy*, North-Holland, v. 46, n. 5, p. 956–969, jun 2017.

PHILLIPS, B. D.; KIRCHHOFF, B. A. Formation, growth and survival; Small firm dynamics in the U.S. Economy. *Small Business Economics*, Kluwer Academic Publishers, v. 1, n. 1, p. 65–74, mar 1989.

PORTER, M. E. *The Competitive Advantage of Nations Harvard Business Review*. [S.l.], 1990.

RENSKI, H. External economies of localization, urbanization and industrial diversity and new firm survival. *Papers in Regional Science*, Wiley Online Library, v. 90, n. 3, p. 473–502, 2011.

ROMERO, J. P.; FREITAS, E.; BRITTO, G.; COELHO, C. The Paths of Industrial Competitiveness in Brazil and South Korea. p. 1–35, 2015.

SEGAL, D. Are there returns to scale in city size? *The Review of Economics and Statistics*, JSTOR, p. 339–350, 1976.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, v. 27, n. 3, p. 379–423, jul 1948.

SIDRA. *Sistema IBGE de Recuperação Automática*. 2019.

SVEIKAUSKAS, L. The productivity of cities. *The Quarterly Journal of Economics*, MIT Press, v. 89, n. 3, p. 393–413, 1975.

VON-THÜNEN, J. H. Isolated state: an english edition of der isolierte staat. Pergamon Press, Oxford, 1966.

WEBER, A. *Theory of the Location of Industries*. Chicago, IL: University of Chicago Press, 1929.

WHEAT, L. F. The determinants of 1963–77 regional manufacturing growth: Why the south and west grow. *Journal of Regional Science*, Wiley Online Library, v. 26, n. 4, p. 635–659, 1986.

WOOLDRIDGE, J. M. *Introdução à Econometria: Uma Abordagem Moderna*. 4. ed. São Paulo: Cengage Learning, 2010. 725 p.