Special thanks to my supervisor Mark Bagley and my opponents for their support in the formulation of this thesis.

"I have made this longer than usual because I have not had time to make it shorter."
— Blaise Pascal in "Lettres Provinciales", 1657.

Proximity to Dissimilar Firms and Universities: Investigating the Impact on Start-up Survival Rates in Sweden and the United States

Birgersson, Johan

School of Business Economics and Law, University of Gothenburg, Gothenburg, Sweden

E-mail: gusbirjoa@student.gu.se
Supervisor: Mark Bagley

29 April 2023

This study explores the impact of proximity to dissimilar firms and universities on start-up survival rates in Sweden and the United States. By investigating the relationship between firm density and start-up survival in two countries with distinct institutional contexts, this research aims to contribute to the existing literature. Using survival analysis and data from CB Insights, SCB, and the U.S. Department of Education, it is found that a higher number of firms within a ZIP code or city significantly decreases the risk of start-up failure or discontinuation in both countries. However, the effect of proximity to universities on start-up survival rates varies depending on the type of graduates present, revealing contrasting outcomes to the Knowledge Spillover Theory of Entrepreneurship. Our findings carry important implications for policymakers, entrepreneurs, and researchers working to foster thriving start-up ecosystems, while also highlighting the need for further research to better understand the underlying factors influencing start-up survival rates in various contexts.

Introduction

The purpose of this thesis is to investigate the impact of proximity to dissimilar firms and universities on start-up survival rates in two distinct countries, Sweden and the United States. This study aims to expand on previous literature by examining the proportional effect of proximity to dissimilar firms and universities in two countries with differing approaches to the institution of the university.

Prior research has identified the clustering of firms and proximity to universities as two important factors that influence start-up survival rates. Clustering has been found to positively impact firm survival rates by providing access to resources, knowledge spillovers, and increasing productivity and innovation. Proximity to universities has also been found to positively influence firm survival rates by increasing entrepreneurial activity through knowledge spillovers. However, the impact of these factors on start-up survival rates varies depending on the industry, location, and firm size.

In the United States, there is a higher density of universities compared to Sweden, with both private and public options, but education in the US can be more expensive. As a result, there may be a higher concentration of highly educated individuals in the United States, but they may also face greater financial pressure to join established companies instead of taking the risk of starting their own business.

In contrast, higher education is publicly funded and free for all students in Sweden, which could lead to a higher proportion of graduates with entrepreneurial aspirations. However, Swedish universities may have fewer financial resources to invest in research and development or to support entrepreneurial ventures, which could limit the availability of resources and knowledge spillovers for start-ups.

To investigate the impact of proximity to dissimilar firms and universities on start-up survival rates, this study poses two research questions. Firstly, how does this impact differ between Sweden and the United States? Secondly, what is the proportional effect of proximity to dissimilar firms and universities on start-up survival rates in both countries, and how does this vary based on the presence of STEM and Business graduates in the area?

To address the research questions, this study builds upon existing research examining the relationship between firm clustering and start-up survival rates. Three concepts related to firm clustering are discussed: MAR externalities, Jacobs externalities, and Klepperian mechanisms. Additionally, the Knowledge Spillover Theory of Entrepreneurship (KSTE) is explored, which posits that entrepreneurship is fueled by external knowledge accessible through spillovers from other firms, industries, and research institutions. Survival analysis techniques, including the Kaplan-Meier estimator and Cox proportional hazards model, are employed to analyze the relationship between predictor variables and the hazard rate of firm survival.

The data used in the study were retrieved from CB Insights and university statistics from SCB and the U.S. Department of Education. The dependent variable is the firm's status, and the independent variables include the number of firms in the same ZIP code and the number of graduates in STEM and Business education[2].

This research reveals that an increase in the number of firms within a ZIP code and city had a significant impact on start-up survival rates in both countries, with variations in the effects of proximity to business and STEM students. The findings contribute to a more nuanced understanding of the Knowledge Spillover Theory of Entrepreneurship and propose a novel theory based on the notion of diversity benefits.

In conclusion, this study contributes to a better understanding of the impact of proximity to dissimilar firms and universities on start-up survival rates in Sweden and the United States. The findings suggest that the clustering of firms positively impacts start-up survival rates and that the effect of proximity to universities on start-up survival rates is dependent on the type of students present. These results have important implications for policymakers and entrepreneurs seeking to improve start-up survival rates. Future research should be conducted to confirm these findings and explore the implications further.

Theoretical and empirical background

The *success of a startup* can be defined in various ways, with firm growth, innovation, and survival being commonly used measures. This study focuses on the survival of start-ups as the key indicator of success. Previous research has identified two factors that affect firm survival: proximity to firms and universities Agarwal and Audretsch (2001); Mata and Portugal (1994); Cefis and Marsili (2005). However, the use of firm survival as a metric for firm success is not new. In fact, Alchian (1950) argues that a firm's ability to survive and adapt to changing market conditions is a key determinant of its success. He suggests that firms that are able to learn from past experiences and adjust their strategies and operations accordingly are better equipped to deal with uncertainty and competition.

Therefore, in this theoretical framework section, we examine the role of clusters and proximity to universities in influencing start-up survival, building on Alchian's (1950) argument for using firm survival as a key metric of success. We start by reviewing the relevant literature on the concept of clustering and finally its impact on firm survival, which provides a framework for understanding how geographic concentration of similar and dissimilar firms can enhance the survival prospects of start-ups. This is followed by a review of the

[2] Examine and evaluate the data and analysis presented in this study at https://masterthesissurvivaljb.onrender.com/. Note that the website is slow due to free hosting.

impact of proximity to universities on firm survival, exploring how access to knowledge, research and development, and other resources may help start-ups improve their chances of survival over the long term. Overall, our theoretical framework is grounded in the idea that a firm's ability to survive and adapt to changing market conditions is a key determinant of its success, and that clustering and proximity to universities may play important roles in facilitating this survival and success.

Clustering

The term *cluster* originates from the research field of agglomeration economies, which refer to the benefits that firms and industries can gain from locating close to each other. The field has been extensively studied by economists over the past decades. The concept was first introduced by Alfred Marshall in the late 19th century, but it wasn't until the 1960s that economists began to develop formal models of the concept (Glaeser, 2010). In 1967, Charles Tiebout published a paper on "Local Governments and Spatial Mismatch" that explored the benefits of clustering similar firms in a geographic area. This paper helped to establish the idea of agglomeration economies as a distinct area of research and served as a starting point for further theoretical and empirical work in the field.

In the 1970s, researchers began to study the specific mechanisms that drive agglomeration economies. They focused on the role of knowledge spillovers, which occur when firms share information and ideas with one another. As such, the 1970s was a decade of "rediscovery" of the importance of agglomeration economies (Rosenthal and Strange, 2004). Economists such as Alfred Marshall, Walter Christaller, and August Lösch had previously discussed the concept in the early 20th century, but it was largely neglected in the post-World War II period (Glaeser, 2010). In 1979, Michael Porter published his book "The Competitive Advantage of Nations," which argued that clusters of related firms in particular regions could achieve significant productivity gains Porter (1990). Porter's work provided a framework for understanding the specific factors that contribute to the success of industrial clusters, including access to specialized suppliers, a skilled labor force, and the presence of supporting institutions.

The 1980s saw a growing interest in the empirical study of agglomeration economies. Researchers began to use data on real-world clusters of firms to test the predictions of theoretical models. One notable contribution was the work of Paul Krugman (1980), who developed a model of trade based on economies of scale and imperfect competition. Krugman's model helped to explain why certain industries tend to cluster in particular regions and why trade patterns are often characterized by the exchange of similar goods between countries. In short, the 1980s is brought and developed the empirical studies of agglomeration economies. This development then acts as a stepping stone for additonal contributions. Researchers such as Edward Glaeser and David Cutler explored how the concentration of skilled workers in certain areas could lead to higher wages and greater innovation (Glaeser and Mare, 1994; Cutler et al., 2001). This research built on earlier work by Paul Krugman, who in the 1980s developed a model of spatial agglomeration that emphasized the importance of local labor markets. Glaeser and Mare (1994) examined the relationship between the concentration of skilled workers and wages in urban areas. They found that cities with a higher proportion of college-educated workers had higher wages. They argued that this was evidence of agglomeration economies at work, as the concentration of skilled workers in these areas allowed for greater specialization and knowledge spillovers. (Cutler et al., 2001) looked at the impact of local labor markets on health care innovation. They found that areas with higher concentrations of health care workers tended to have more innovative medical treatments, as the close proximity of these workers allowed for more collaboration and knowledge sharing. This research provided further evidence of the importance of local labor markets in driving agglomeration economies.

In summary, the study of agglomeration economies has come a long way since its early beginnings with Alfred Marshall in the late 19th century. This study continues by discussing two different strains from agglomeration economies *Marshall-Arrow-Romer Externalities* (MAR Externalities) which explores the cross-firm effect of co-location of similar firms while *Jacobs Externalities* discuss the intra-firm effect of co-location of dissimilar firms in diverse environments, which is the focus of this study.

The effects of these theories can be viewed through the lens of clusters we see today in a variety of industries, including manufacturing, technology, and creative industries. For example, Silicon Valley is a well-known tech cluster where similar firms benefit from a large pool of skilled workers and a culture of innovation. Similarly, New York is a creative cluster where firms benefit from access to talent, resources, and networks and where dissimilar firms benefit from an environment where individuals and firms can exchange ideas, knowledge, and resources in ways that lead to new forms of economic activity and innovation (Jacobs, 1961).

Subsequently, the study explores *Klepperian Mechanisms*, which argues that clusters of firms in specific industries or regions can facilitate these informal mechanisms of knowledge transfer, leading to increased innovation, productivity, and competitiveness. Klepper's research highlights the importance of understanding the social and economic context in which knowledge is created and transmitted in order to develop effective policies for promoting economic growth and innovation.

MAR Externalities

Marshall-Arrow-Romer (MAR) externalities refer to a type of external economy that arises when the productivity of one firm increases due to the presence of other firms in the same industry or location (Arrow, 1962; Romer, 1990). In other words, these externalities are positive spillover effects that occur between firms, as opposed to within firms.

MAR externalities are named after the economists Alfred Marshall, Kenneth Arrow, and Paul Romer, who all contributed to the study of external economies of scale. Alfred Marshall (1890) was the first to introduce the concept of external economies in his book "Principles of Economics". Marshall's work was built upon the ideas of several earlier researchers. He drew upon Charles Babbage's (1832) concept of economies of scale and the benefits of specialization, Émile Durkheim's (1893) emphasis on social institutions and norms in shaping economic outcomes, Alfred Weber's (1909) theory of industrial location and the concept of "location rent", and Johann von Thünen's (1826) ideas about the spatial distribution of economic activities and the factors that influence industrial location. Arrow extended Marshall's work to include technological progress and the knowledge spillovers that can arise from it (Arrow, 1962). Romer further developed the concept of knowledge spillovers, referred to as "new growth theory". Their ideas built on Schumpeter's (1934) theory by emphasizing the role of knowledge creation and human capital investment in driving technological progress and innovation (Arrow, 1962; Romer, 1990).

MAR externalities can occur in several ways. For example, the concentration of firms in a particular industry or location can lead to a greater availability of specialized labor, infrastructure, or specialized suppliers. This can lead to lower costs, higher productivity, and greater innovation, benefiting all firms in the location. Additionally, the presence of highly productive firms can serve as a demonstration effect for other firms, encouraging them to adopt more efficient practices.

In the context of clustering, MAR externalities can be a powerful force driving the agglomeration of firms in certain locations or industries. This can lead to the development of industry clusters, which can provide a range of benefits to firms, including access to a skilled labor force, specialized suppliers, and knowledge spillovers (**?**Arrow, 1962; Romer, 1990).

However, it is important to note that the positive externalities associated with similar firms in a cluster can have the property of diminishing returns (Krugman, 1991; Duranton and Puga, 2004). As more firms join a cluster, the initial benefits of agglomeration might diminish due to factors such as strained resource availability, increased competition, and higher costs for inputs (Rosenthal and Strange, 2004). Congestion may also arise, leading to increased costs associated with transportation and infrastructure. Furthermore, as more firms join, the incremental benefits of new knowledge spillovers may decrease, with the most valuable knowledge already shared among existing firms (Beaudry and Schiffauerova, 2009), and the risk of imitation and competition may rise.

Despite the diminishing returns associated with clustering, the overall benefits of agglomeration can still outweigh the costs, particularly if strong network effects or other factors make the cluster attractive for firms (Duranton and Puga, 2004). The degree of diminishing returns may vary across industries, locations, and time periods, depending on specific case characteristics (Rosenthal and Strange, 2004; 2003; Beaudry and Schiffauerova, 2009).

Jacobs externalities

Jacobs externalities, named after the urban theorist Jane Jacobs, refer to a different type of external economy that arises from the diversity and complexity of cities (Jacobs, 1961). Jacobs argued that cities are dynamic systems that thrive on diversity, and that the interactions between different people, firms, and ideas in urban areas can lead to new forms of economic activity and innovation. According to Jacobs, the key to successful urban economies is a dense, diverse, and mixed-use urban environment that facilitates frequent face-to-face interactions between people with different backgrounds and skills. In such an environment, individuals and firms can exchange ideas, knowledge, and resources in ways that lead to new forms of economic activity and innovation (Jacobs, 1961). This can lead to a virtuous cycle of economic growth, where the diversity and complexity of the city lead to the emergence of new economic opportunities, which in turn attract more people and firms to the city, further increasing its diversity and complexity. Jacobs externalities differ from MAR externalities in several ways. While MAR externalities arise from the concentration of firms in a particular industry or location, Jacobs externalities arise from the diversity and complexity of the urban environment itself. MAR externalities are typically associated with knowledge spillovers and productivity gains within a particular industry or location, while Jacobs externalities are associated with the emergence of new forms of economic activity and innovation across different industries and locations (Jacobs, 1961). Finally, MAR externalities are often thought to be self-limiting, as increasing concentration can lead to diminishing returns, while Jacobs externalities are thought to be self-reinforcing, as increasing diversity and complexity can lead to new forms of economic activity and innovation.

Klepperian mechanisms

Klepperian mechanisms of knowledge transfer and inheritance refer to the ways in which knowledge and skills are transferred between firms, industries, and generations of innovators Klepper (1996). These mechanisms are named after the economist and innovation scholar, Phillip Klepper, who argued that knowledge and skills are not simply transmitted through formal education and training, but also through a variety of informal mechanisms, including social networks, labor mobility, and knowledge spillovers.

One key mechanism of knowledge transfer identified by Klepper is labor mobility. Firms that employ skilled workers are more likely to generate knowledge spillovers, as these workers may take their knowledge and skills with them when they move to new firms or industries Klepper (1996). This can lead to the diffusion of knowledge across industries and locations, and can contribute to the emergence of new industries and technologies. Another mechanism identified by Klepper is the inheritance of knowledge and skills across generations of innovators. Klepper argued that the development of new industries and technologies often involves a process of recombinant innovation, in which existing knowledge and technologies are combined in new and innovative ways. This process is facilitated by the accumulation of knowledge and skills over time, as new generations of innovators build upon the work of their predecessors Klepper (1996).

The concepts of MAR and Jacobs externalities tie into the Klepperian mechanisms in several ways. MAR externalities can facilitate the transmission of knowledge and skills through knowledge spillovers within a particular industry or location Arrow (1962). Similarly, Jacobs externalities can facilitate the transmission of knowledge and skills through the diversity and complexity of the urban environment, which can lead to new forms of economic activity and innovation Jacobs (1961). Both MAR and Jacobs externalities can contribute to labor mobility, as firms and workers may be attracted to locations with high levels of knowledge spillovers and innovation Arrow (1962); Jacobs (1961). Finally, the inheritance of knowledge and skills is facilitated by the accumulation of knowledge and skills over time, which can be facilitated by MAR and Jacobs externalities that lead to the development of industry clusters and diverse urban environments (Marshall, 1890).

Regional and Institutional Context

This study compares two distinct institutional contexts. Baumol (1990) discusses the role of entrepreneurship in economic growth and highlights the importance of the institutional context in which start-ups operate. Factors such as regulatory environment, access to capital, and the availability of support structures can influence the chances of start-up success.

Previous research

The impact of clustering on the survival of firms has been a topic of interest for researchers and policymakers for many years. Clustering can positively impact firm survival by providing access to resources and knowledge spillovers. According to Porter (1990), geographic clustering creates a competitive advantage by promoting the exchange of knowledge and ideas among firms. This can lead to increased productivity and innovation, which can improve the survival chances of firms.

Multiple sources find impact between clustering and productivity. Frenken et al. (2015) found that while there is strong evidence that clusters promote entry, while there is little evidence that clusters enhance firm growth and firm survival. Meanwhile, Bagley (2019), studies the Swedish ICT sector and finds that knowledge transfer through networks as well as proximity to other firms both significantly impact survival rates of firms. De Vaan et al. (2013) suggest in their study of the global video game industry that in project-based industries, the negative localization externalities associated with competition increase proportionally with cluster size, while positive localization externalities increase more than proportionally with cluster size.

In a study of Manhattan Advertising agencies Arzaghi and Henderson (2008) investigate the networking benefits among agencies that are in close proximity. The study finds a strong effect on productivity of having more nearby advertising agency neighbors, but the benefits of having more nearby neighbors rapidly decrease, even in densely populated areas such as southern Manhattan. Clustering can also lead to increased competition and the creation of barriers to entry. This can negatively impact firm survival, especially for smaller firms. Glaeser et al. (1992) found that in industries with high levels of clustering, smaller firms were more likely to exit the market.

The impact of clustering on firm survival can vary depending on the industry and location. For example, industries that are heavily regulated or require significant capital investment may benefit more from clustering than those that are more flexible. Additionally, the location of the cluster can impact firm survival. For example, clusters located in regions with high levels of economic development and infrastructure may have a greater positive impact on firm survival. Beaudry and Swann (2009) explore the heterogeneity of cluster effects across industries. The study suggests that cluster effects are strongest in manufacturing, manufacturing-related, and infrastructure industries, but weaker in services. Rigby and Brown (2015) investigate whether the benefits of agglomeration vary for different types of business establishments, using plant-level data from the Canadian manufacturing sector. The study finds that larger firms and those in industries with higher R&D intensity benefit more from agglomeration. Interestingly, Maine et al. (2010) examine the impact of being part of a cluster on the growth of new technology-based firms in the US. The study argues that firms benefit from access to specialized resources through clustering. The findings suggest that for a majority of firms, proximity to a cluster has a negative effect on growth, however, in the case of biotech firms, the study indicates the opposite relationship. The study also finds that proximity to a cluster within a diverse metropolitan area is positively related to growth only for firms that heavily rely on broad supply chain effects, such as information and communication technology firms. Larsson (2009) explores the productivity benefits of density externalities in cities, particularly in central business districts (CBDs). The study finds that density externalities drive productivity benefits, especially for non-routine activities.

University Knowledge Spillover

The Knowledge Spillover Theory of Entrepreneurship (KSTE) proposes that entrepreneurship is driven by the external knowledge that is available to individuals and firms through spillovers from other firms, industries, and research institutions Acs et al. (2009), which can lead to the creation of new products, processes, and technologies. KSTE can be viewed through the lens of Klepperian mechanisms of knowledge transfer and inheritance Acs et al. (2009). This aligns with Klepper's idea that knowledge and skills are not only transmitted through formal education and training but also through a variety of informal mechanisms, including social networks and knowledge spillovers Arrow (1962); Jacobs (1961).

According to KSTE, knowledge spillovers can occur through different channels such as labor mobility, informal communication, and collaboration, which can be facilitated by MAR and Jacobs externalities Acs et al. (2009). In particular, labor mobility can play a crucial role in the transmission of knowledge between firms and industries Acs et al. (2009), as highly skilled workers can carry valuable knowledge and skills with them when they move to new firms or locations. This aligns with Klepper's argument that labor mobility as one of the key mechanisms of knowledge transfer (Arrow, 1962).

Furthermore, the KSTE highlights the role of technological knowledge in driving economic growth and development Acs et al. (2009), which is also consistent with Klepper's idea that the accumulation of knowledge and skills over time can contribute to the emergence of new industries and technologies Arrow (1962). The KSTE emphasizes the importance of creating an environment that fosters the creation and diffusion of new knowledge, which can be facilitated by MAR and Jacobs externalities that create opportunities for knowledge spillovers and promote the exchange of ideas and information Acs et al. (2009).

Overall, the KSTE can be viewed as an extension of Klepper's ideas on knowledge transfer and inheritance, highlighting the importance of external knowledge in driving entrepreneurship and economic growth Acs et al. (2009). By understanding the mechanisms through which knowledge is transferred and inherited, policymakers and entrepreneurs can create environments that facilitate the creation and diffusion of new knowledge and technologies Acs et al. (2009); Romer (1990).

Audretsch and Lehmann (2005) could be argued confirming this theory, finding that the number of firms located near a university is positively influenced by the knowledge capacity of the region and the knowledge output of the university. Thus, the knowledge capacity of the region and the knowledge output of universities can be considered important control variables in measuring start-up success.

External sources of knowledge, such as universities, also play a role in the growth of entrepreneurial firms. Cassia et al. (2009) find that universities' knowledge input and output are significant factors in the growth of firms in the UK. This suggests that the interaction between universities and entrepreneurial firms is an important factor in measuring start-up success.

Furthermore, Perkmann et al. (2013) supports the notion that access to specialized knowledge, resources and networks can enhance the innovative capabilities, competitiveness and thereby start-up survival.

Survival Analysis

The Kaplan-Meier estimator, also known as the product-limit estimator, is a non-parametric[3] method for estimating the survival or failure distribution function (Kaplan and Meier, 2009). It is commonly used in medical research, engineering, and other fields where survival analysis is important.

In univariate analysis, the Kaplan-Meier estimator is used to analyze the time until an event of interest occurs, such as death, disease recurrence, or machine failure. The data used in this analysis may include both censored observations, where the event has not yet occurred, and uncensored observations, where the event has occurred.

The Kaplan-Meier estimator calculates the probability of survival, or the probability of an event not occurring, at each time point. This is done by taking into account the number of individuals or units still at risk, or under observation, at each time point. The estimator then multiplies the probability of survival at each time point to obtain an estimate of the survival function.

The Kaplan-Meier estimator is particularly useful in situations where there are a large number of censored observations, since it takes into account the fact that some observations are not fully observed. As is the case for this study. It is also useful for comparing survival curves between different groups, such as treatment groups or patient subgroups, and for identifying factors that may influence survival Kaplan and Meier (2009); Altman (1990).

One limitation of the Kaplan-Meier estimator is that it assumes that the probability of failure or survival is constant over time, which may not be true in all cases. In addition, the estimator does not provide information about the shape of the survival curve or the distribution of survival times. To address these limitations, other methods such as parametric survival models or competing risk models may be used Collett (2015).

The Cox proportional hazards model is a widely used statistical method for analyzing time-to-event data. It was first introduced by David Cox in 1972, and has since become a standard tool in survival analysis.

The Cox proportional hazards model is a semi-parametric model that allows for the examination of the relationship between one or more predictor variables and the hazard rate. The hazard rate is the instantaneous rate at which an event of interest (such as death, failure, or recurrence) occurs, given that the individual has survived up to that point in time. The model assumes that the effect of the predictor variables is constant over time, but allows for a flexible baseline hazard function that can vary with time.

The Cox proportional hazards model has several advantages over other survival analysis methods Kleinbaum and Klein (2012). It is particularly useful when the hazard rate is non-constant over time, as it allows for a flexible baseline hazard function that can vary with time. Additionally, the Cox

[3] A non-parametric model in this context estimates the probability of survival over time without assuming a specific distribution for the underlying hazard or survival function. In this case, the Kaplan-Meier estimator considers the "exited" variable (e.g., event of death) as the only input variable and does not require any other covariates or predictors.

model does not require knowledge of the distribution of survival times, making it more robust to violations of distributional assumptions.

In other statistical models, such as linear regression models, the intercept term represents the value of the response variable when all predictor variables are equal to zero. It is often interpreted as the baseline level of the response variable that is present even in the absence of any predictors.

In the context of survival analysis, an intercept term would represent the baseline hazard rate, which is the hazard rate when all covariates are equal to zero. However, the Cox proportional hazards model does not estimate an intercept term, as the baseline hazard rate is estimated non-parametrically. Instead, the model estimates the effect of each covariate on the hazard rate relative to the baseline hazard function.

One implication of this is that the Cox proportional hazards model does not provide direct estimates of the absolute risk or survival probability for a given set of covariate values. Instead, the model estimates the relative risk or hazard ratio associated with each covariate.

The Cox proportional hazards model has been used in a wide range of applications, including medical research, engineering, and social sciences Therneau and Grambsch (2000). It has been used to analyze survival data in cancer research, to model the failure of mechanical components, and to examine the time to recurrence of psychiatric disorders as well as firm survival.

Several extensions and modifications of the Cox proportional hazards model have been proposed over the years Therneau and Grambsch (2000). These include models for time-varying covariates, competing risks, and frailty models. These extensions have further expanded the usefulness of the Cox model in various fields.

In summary, the Kaplan-Meier estimator and the Cox proportional hazards model are two commonly used methods in survival analysis. The Kaplan-Meier estimator is a non-parametric method used to estimate the survival or failure distribution function from censored data, while the Cox proportional hazards model is a semi-parametric model used to examine the relationship between one or more predictor variables and the hazard rate.

Together, the Kaplan-Meier estimator and the Cox proportional hazards model can provide a robust analysis of survival data. The Kaplan-Meier estimator can be used to estimate the survival or failure distribution function and to compare survival curves between different groups. The Cox proportional hazards model can be used to examine the relationship between predictor variables and the hazard rate, while taking into account the effects of other covariates.

The combination of these two methods allows for a more comprehensive analysis of survival data, providing information about both the probability of an event occurring over time and the factors that may influence that probability. Employing both methods enables a more comprehensive understanding of the survival patterns present in the data, subsequently leading to broader analysis and enhanced outcomes.

Quantitative Methods

The following section is included as a motivation for the use of quantitative methods in the analysis of this study, highlighting the advantages of quantitative methods in comparison to qualitative methods. Quantitative methods of analysis offer numerous advantages for this study, including objectivity, reliability, generalizability, hypothesis testing, and precision in quantifying relationships between variables. By drawing on the insights from Bell, Bryman, and Harley's (2015) and other foundational textbooks, this study finds that the use of quantitative methods is well-motivated and appropriate for this study. Adopting a quantitative approach will enhance the credibility of the findings and provide valuable insights that can inform policy decision-making and strategy development.

Objectivity and Reliability

While qualitative methods emphasize the exploration of subjective experiences and the interpretation of meanings, quantitative research methods are known for their objectivity and reliability (Bryman and Bell, 2015; Creswell, 2014). By using standardized measures and numerical data, quantitative methods minimize subjective biases and allow researchers to replicate the findings of the study. This, in turn, enhances the credibility of the research and enables a better understanding of the phenomena under investigation (Flick, 2015).

Generalizability

Qualitative methods are often used to provide in-depth and context-specific insights, but they may have limited generalizability. In contrast, quantitative methods allow researchers to generalize their findings to a larger population (Bryman and Bell, 2015; Trochim and Donnelly, 2008). By using random sampling techniques and appropriate sample sizes, quantitative research can offer valuable insights into patterns and trends that can be applied to broader contexts.

Hypothesis Testing

While qualitative research is valuable for generating new ideas and exploring complex phenomena, quantitative research excels in hypothesis testing (Bryman and Bell, 2015; Hair,et al., 2010). Researchers can formulate hypotheses based on existing theories or empirical evidence and use statistical tools to test their validity. This approach enables researchers to make causal inferences and draw conclusions about the relationships between variables, which can ultimately contribute to the development of new theories and models in business research (Saunders et al., 2016).

Precision and Quantification

Qualitative methods provide rich, detailed data but may lack the precision and quantification needed for some research questions. In contrast, quantitative methods enable researchers to quantify the strength and direction of relationships between variables (Bryman and Bell, 2015; Field, 2017). By using various statistical techniques, researchers can measure the effect of one variable on another, making it possible to determine the magnitude of these effects. This precision and quantification are invaluable in business research, where decision-makers rely on accurate and actionable insights (Bryman and Bell, 2015).

Limitations of Quantitative methods

The advantages of quantitative research, as previously discussed, have been widely recognized and supported by numerous scholars and experts in the field of business research. By employing quantitative methods, this study aims to provide objective, generalizable, and precise findings that can significantly impact policy decision-making and strategy development.

However, it is important to acknowledge the limitations of quantitative methods as well. For instance, quantitative methods might not capture the depth and nuance of human experiences, emotions, and perceptions to the same extent as qualitative methods (Creswell, 2014). Moreover, quantitative research can sometimes oversimplify complex phenomena, reducing them to numerical data and potentially overlooking important contextual factors (Flick, 2015).

To address these limitations, this study acknowledges the specific utility of quantitative research. The goal of this study is not to explain how or why firm survival is influenced by proximity to other firms or universities, but rather to determine the direction and extent to which firm survival is correlated with the co-location of other firms or universities. Following this, this study evaluates the research questions in order to explain the underlying mechanisms and causal relationships, which, in future research, can be formulated into new research questions. These questions could then be investigated through the application of either quantitative or qualitative research methods.

Lastly, this study ensures that data collection and data analysis techniques are appropriately designed and executed to capture the nuances of the phenomena under investigation. The study also invites readers to inspect the data and perform the analysis using a publicly available analysis tool[4]. This transparent approach is used to facilitate a deeper understanding of the research context, and to enable readers to appreciate the value and rigor of the quantitative methods employed in this study.

[4] https://masterthesissurvivaljb.onrender.com/

Data, Variables, and Method

The data utilized in this study are a combination of both proprietary and publicly available information. To obtain information on firm survival, the study relied on CB Insights. Additionally, the study gathered public information on university locations and the number of graduates from two sources. The study aggregates data from SCB, and U.S. Department of Education, National Center for Education Statistics to obtain an exhaustive dataset for universities in Sweden and the United States (SCB, 2022; College Scorecard, 2022) In the sections that follow, we will first discuss the data considerations, then move on to describing the process of variable creation, outlining the assumptions made and considerations taken. Finally, the methodology used to process the data for analysis and the steps taken to consider assumptions for the analysis is explained in the methods section.

Data

The data on firm survival was retrieved from CB Insights, a data provider with a focus on the technology market. As such, the data is not a representative sample of the average business in any of the countries analyzed. CB Insights is primarily a platform for market intelligence, and companies monitored within the service have or are in some way of potential interest to investors, leading to a bias of inclusion. CB insight is geared toward firm-level analysis of a given set of firms. As such, it's important to note that this is not aggregate macro-level data. Depending on the methodology that CB Insights employ in their selection, the results of this study are subjected to this selection bias. Therefore, this study assumes that the lifecycle characteristics of the firms in the dataset differ from those of the average general firm.

The sampling performed when collecting the data was random, and the first 10,000 firms within each country selection were retrieved. The subset of companies per country and the distribution among countries is a result of this random sampling. A notable difference to similar studies is that this study had no selection with regards to separation between industries or any other firm categories. The sampling approach employed in this study aims to achieve a diverse sample of companies, aligning with the central limit theorem, in order to investigate the impact of diversity within clusters—an essential aspect of Jacobs Externalities (1961). While this method ensures some degree of diversity within the clusters, it does not isolate the diversity effects. It is natural to expect that firms with similar characteristics or industries tend to locate in the same region, as suggested by the theory of MAR externalities. Nevertheless, the methodology utilized in this study ensures that the sample reflects the actual clustering scenario, incorporating both diversity and absence thereof.

It can be concluded that the aforementioned bias of inclusion is prevalent in the data, foremost, the distribution of survivors vs. exited companies is low by any standard for Sweden, while the percentage of inactive US firms is higher. To some extent, this may also be a case of bias of inclusion. CB insights is a US firm, while it provides it's service globally, it's presumed that the US is the greatest market for CB insights (6sense, 2022). Currently, there is no public information available regarding any other selection or inclusion criteria that CB insights uses.

Additionally, it is also feasible to assume that companies that were already inactive at the time of database creation have not been included. For the purposes of the CB Insights platform, there would be marginal utility in adding information on companies that are no longer active or potential venues of business. However, a caveat to this is their use of machine learning and web scraping to build their dataset (CB Insights, 2023), which would naturally gather information on companies active prior to the inception of CB Insights.

During the data collection process for Swedish universities, it was observed that the classification of graduates' degree programs is not consistent. For example, the SCB data revealed that the University of Gothenburg, including its Business School, registered only five graduates with a Master of Science in Business or Economics during the 2021-22 academic year (SCB, 2022). This discrepancy may compromise the data's reliability and any subsequent analysis of Swedish graduates. To ensure a comparative analysis between the countries, it is essential to have consistent data for both regions. Unfortunately, no alternative sources that satisfied these requirements were found.

Variables

Dependent variable

The variable "exited" used in this study is created based on a company's status at the time of data retrieval. Companies can be classified into one of six statuses, including Alive or Active, Acquired, IPO/Went public, Merged, or Dead or Inactive. For the purpose of this study, a firm's status of "acquired" or "merged" is treated as equivalent, furhtermore, a firm is considered a survivor if its status is not "Dead or Inactive", as seen in Equation 1.

$$IF(Firm_{Status} \neq Dead\ /\ Inactive) \implies exited = 1 \quad (1)$$

Since the primary focus of this study is firm survival, a key data point necessary for the analysis is the lifetime of companies. This data allows us to determine the probability of death or inactivity. Unfortunately, the data were incomplete in this regard. Initially, a sample of 49,000 companies from North America and Scandinavia was obtained, but only 30,000 companies had complete records of duration, and status. Additionally, only firms in Sweden and the United States showed promising distributions among the different statuses of companies[5].

Other necessary datapoints are the founding date and exit date of the firms, where the exit date is the date of either acquisition or death. In cases where the firm is still active, the exit date is censored in the statistical models as it cannot be

[5] Inspect the distribution of firms in dropped countries at https://masterthesissurvivaljb.onrender.com

determined when it will exit. However, for dead or inactive firms, there are instances of missing data on the exit date, in such cases an exit date is approximated using available data on the date of last funding as a proxy[6]. With these pieces of information, i.e., firm status, founding date, and exit date, survival analysis is possible.

Independent variables

For each country, two models are developed, one for each cluster size of interest: city-sized clustering and ZIP-code sized clustering. The number of companies within a given city or ZIP code represents the density of that cluster through the variables "Num Firms In ZIP" and "Num Firms In City". Additionally, two variables for graduates in each respective level in both countries are created. The variable for Science, Technology, Engineering, and Mathematics (STEM) is represented as a percentage of graduates from the following fields of study: communications technology, computer, engineering, engineering technology, biological, mathematics, and science technology. Meanwhile, the business graduates are those pursuing a degree in Business and Administration, Management or in the case of Sweden, any other degree which accredits a Degree in Business or Economics. However, the available data on the direction of study at the universities in each region vary in their detail. For Sweden, each university has number of graduates associated to each category. For universities in the United States, Collegescorecard lists the percentages of graduated students within a given program. As such, in the case of the United States these percentages are multiplied with the latest available number of graduates from the given university.

In short, the variable "Num Firms In City" is simply the aggregate of firms within the same city as a given firm, with the same logic for ZIP size. Similarly, the variable "City Business Students" contains the aggregate number of graduates within business from all universities in the same city as the firm. Conversley, "ZIP Business Students" contains the aggregate number of graduates within business from all universities within the same ZIP code as the firm.

After creating these variables, the natural logarithm is applied to normalize the variables[7].

ZIP codes

Postal codes or ZIP codes, as this study refer to them as, are a system of codes used by postal services around the world to sort and deliver mail to their intended destinations (Postnord, 2017; Terrell, 2013)[8].

This study employs five-digit ZIP codes in both the United States and Sweden to maximize the level of differentiation between the ZIP code and city-level data. US ZIP codes cover an average of 233 square kilometers, while Swedish ZIP codes cover an average of 18 square kilometers. Actual areas of a given ZIP code vary based on population density and geography. Naturally, ZIP codes in Sweden tend to be smaller due to the smaller geographical size of Sweden as compared to the United States.

Method

Prior to preprocessing, there is data on 24,150 firms across the US (85.5 percent) and Sweden (14.5 percent). The percentage of dead or inactive companies have the following distribution: Sweden (4.4 percent dead), and the United States (24 percent dead), see table 1.

[6] To strengthen this proxy, the average duration from last funding until the event of death for all other firms was added to the lifetime of the given firm, 912 days.

[7] The table located at https://masterthesissurvivaljb.onrender.com illustrates how the data appears in the models, except for one variable: the number of firms. This variable undergoes transformation just before fitting the Cox Proportional Hazard Model. This is done in order to maintain interpretability when visualizing the data on clusters within the analysis tool.

[8] Both Sweden and the United States use 5-digit ZIP codes to represent specific geographic areas. The first digit indicates a region, while the second and third digits narrow it down. The final two digits represent a post office or delivery zone. The codes are hierarchical and vary in size and length at each level.

Table 1 Data distribution prior to preprocessing by country

Sweden	Observations
Alive / Active	2446
Acquired	1073
IPO / Went public	333
Dead / Inactive	182
Merged	39
Assets Purchased	24

United States	Observations
Alive / Active	7526
Dead / Inactive	4806
Acquired	3027
IPO / Went public	2793
Merged	1436
Assets Purchased	465

In order to generate the variables described above, necessary sanity checks are conducted, as shown in Appendix Section A. The data is then merged on university and firm by ZIP and city, respectively, to create a complete data set. Observations where ZIP code or city data is missing, are dropped and the merge is verified (Appendix Section A1). However, it's observed that many observations lack data on the newly added universities. This is because, as depicted in Section A2, not all firms are located in the same ZIP code or city as a university, which is what this study aims to investigate the contribution of. After creating the variables for graduates in the respective fields on each level, the new information is merged with the other data. If there is missing data on graduates for a given city or ZIP code, the firm data is kept.

At this stage, utility data to construct the variables are removed. As such, the analysis contains data on the variables in Table 7 and 8.

To prepare the data for Kaplan Meier univariate analysis, rows where the founding year was missing are removed. Subsequently, the duration of each firm is determined as the time between the founding year and the date of exit. Importantly, the data is right-censored on the exit dates of firms with a lifetime longer than 10 years, as this study is specifically interested in the effects pertaining to start-ups[9]. This ensures that the analysis focuses on the early stages of firm survival, and avoids overestimating duration of the firms in the sample.

Prior to modeling, all variables are checked for multicollinearity using the variance inflation factor (VIF) method, see Tables 2-5. Checking for multicollinearity before performing a Cox proportional hazard model is essential because it helps avoid unreliable parameter estimates, inflated standard errors, and reduced model performance. Identifying and addressing multicollinearity issues ensures a more accurate and reliable model, reflecting the true relationships between predictor variables and the outcome. The VIF method examines multicollinearity by quantifying the extent to which the variance of a regression coefficient is inflated due to the correlation among predictor variables.

	feature	**VIF**
0	exited	1.052552
1	Lifetime	1.148432
2	Num Firms In ZIP	1.092788
3	ZIP Business Students	1.198976
4	ZIP STEM Students	1.199426

Table 2 Variance inflation factors for US City

	feature	**VIF**
0	exited	1.067332
1	Lifetime	2.752348
2	Num Firms In CITY	2.306356
3	City Business Students	1.756872
4	City STEM Students	2.072251

Table 3 Variance inflation factors for SWE City

Covariates with a VIF value below the threshold of 10 are considered to have an acceptable level of multicollinearity and were included in the regression model. This threshold is widely accepted in the literature; however, some researchers may opt for a more conservative threshold, such as 5, to minimize the potential impact of multicollinearity on the model's performance (O'Brien, 2007).

[9] Right-censoring is a technique commonly used in survival analysis when the event of interest (ie. firm exit) has not been observed for all subjects by the end of the study period. By right-censoring the exit dates, the study acknowledges the lack of complete information on the duration of firms in the sample, and this uncertainty is accounted for in the analysis.

	feature	VIF
0	exited	1.232378
1	Lifetime	1.884205
2	Num Firms In ZIP	1.601856
3	ZIP Business Students	3.101301
4	ZIP STEM Students	3.171600

Table 4 Variance inflation factors for US ZIP

	feature	VIF
0	exited	1.256085
1	Lifetime	1.927217
2	Num Firms In CITY	1.790545
3	City Business Students	4.839674
4	City STEM Students	5.090520

Table 5 Variance inflation factors for SWE ZIP

The VIF test indicates that all covariates fall below the accepted VIF threshold of 10, confirming their inclusion in the model. However, it is important to mention that the final model exhibits higher levels of correlation in the variables "City Business Students" and "City STEM Students," both with a VIF value of approximately 5 (Table 5).

The models in this study, employ the Huber sandwich estimator, also known as the Wei-Lin estimate, to account for possible heteroscedasticity and obtain robust standard errors (Huber, 1967). The Huber sandwich estimator is widely recognized for its effectiveness in providing reliable inferences in the presence of potential model misspecification or violation of assumptions (White, 1980; Lin and Wei, 1989). By using the Huber sandwich estimator, this study is able to provide more robust findings to support the analysis, enhancing the credibility and generalizability of the results. The results from the models with this estimator are presented in Table 9. The same models are also fitted without the Huber sandwich estimator (See Appendix B).

What this study seeks to understand

Consider the observations in Table 6. For instance, the first firm listed, Bioassay Systems, survived for 10 years in the United States with 46 firms in its ZIP code, the firm has 7.74 logarithmic units of business students which translates to 2298 business students, and 6.13 logarithmic units of STEM students, which translates to 459 students in its ZIP code. On the other hand, Media Systems Technology also survived for 7 years in the United States but had only 1 firm in its ZIP code and no business or STEM students.

Name	Bioassay Systems	Media Systems Technology
exited	1	1
Lifetime	10	7
Num Firms In ZIP	46	1
ZIP Business Students	7.741	0
ZIP STEM Students	6.313	0

Table 6 Illustrative Example

Media Systems Technology, located in Irvine, California, is situated in a large city sized cluster in the United States, with a low number of represented firms in its ZIP code. Additionally, the company does not appear to benefit from proximity to universities, as there are no business or STEM students in the ZIP code where it is located. This lack of proximity to universities could potentially make it difficult for the company to access a pool of talented graduates and research facilities. This study seeks to understand if this is the case.

Bioassay Systems, on the other hand, is located in Woburn, Massachusetts, which is also a large city-sized cluster in the United States, but with a relatively higher number of firms in the ZIP code where it is located. Additionally, Bioassay Systems appears to also have ZIP-level distance to universities, as there are both business and STEM students in the ZIP code where it is located. This proximity to universities could potentially provide Bioassay Systems with access to a pool of talented graduates and research facilities, which may help them to innovate and remain competitive in their industry. The analysis will help us answer if this is the case.

Overall, these differences in proximity to universities and clustering patterns could potentially impact the survival and success of Media Systems Technology and Bioassay Systems in different ways. This is what this study seeks to understand.

Results

Descriptive statistics

Having prepared the data for analysis, a substantial amount of data has been removed. In the interest of being able to generalize the results of the analysis, we want to inspect the distribution of data in the analysis, and uncover potential biases introduced by the data cleaning process. We see, as compared to the initial data, that observations have been dropped seemingly even among the different statuses for each respective country, suggesting that our data processing have not disproportionately affected any particular group[10]. The vast majority of firms remain active in both countries. Furthermore, there is a comparable distribution of firm statuses between the two nations, with the notable exception that the proportion of dead or inactive firms is significantly higher in the United States.

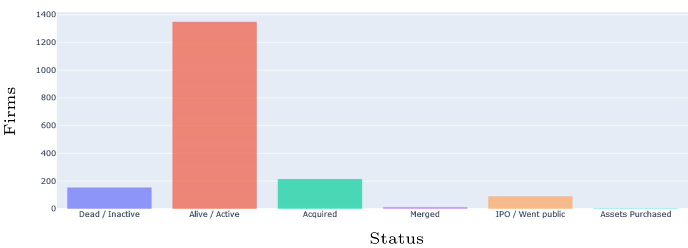


Figure 1 Distribution of status swedish firms

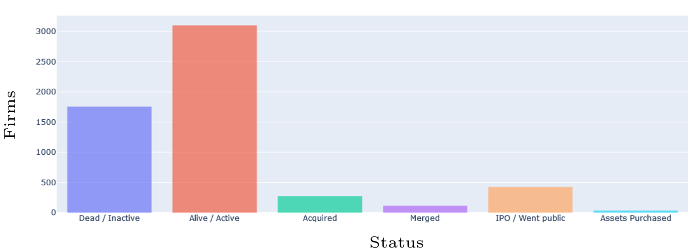


Figure 2 Distribution of status US firms

[10] Visit https://masterthesissurvivaljb.onrender.com to inspect the data, and follow the analysis, utilize the presets in the toolbar to inspect each model.

United States	mean	std	min	25%	75%	max
exited	0.308	0.461	0	0	1	1
Lifetime	6.305	2.353	0	4	8	10
country United States	1.000	0.000	1	1	1	1
Num Firms In ZIP	62.682	74.931	1	7	93	314
ZIP Business Students	0.904	2.886	0	0	0	14.010
ZIP STEM Students	1.046	2.927	0	0	0	12.804
Num Firms In CITY	495.991	633.367	1	40	431	1682
City Business Students	4.374	5.550	0	0	10.926	13.285
City STEM Students	4.733	6.066	0	0	10.429	15.263

Table 7 Descriptive statistics of 5696 observations in United States

Table 7 presents descriptive statistics for 5696 observations in the United States. The variable "exited" has a mean of 0.308. This indicates that 30.8% of the start-up firms in the sample have failed or discontinued operations. The variable "Lifetime" has a mean of 6.305 years and a standard deviation of 2.353 years. This indicates that, on average, start-up firms in the sample have been in operation for 6 years and four months before failure or discontinuation. The variable "country United States" is a binary variable with a mean of 1.000 and a standard deviation of 0.000. This indicates that all observations in the sample are from the United States. The variable "Num Firms In ZIP" has a mean of 62.682 and a standard deviation of 74.931. This indicates that, on average, there are 63 start-up firms in the ZIP codes where the observations are located but that there is a great variation around the mean. The variables "ZIP Business Students" and "ZIP STEM Students" represent the number of business and STEM students in the ZIP codes where the observations are located, respectively. "ZIP Business Students" has a mean of 0.904 and a standard deviation of 2.886, while "ZIP STEM Students" has a mean of 1.046 and a standard deviation of 2.927. This indicates that, on average, there are very few business and STEM students in the ZIP codes where the observations are located. Note that both the city and zip level business and stem data are transformed using logarithmic transformation at this stage so these figures translate to one to three students on average, The variable "Num Firms In CITY" has a mean of 495.991 and a standard deviation of 633.367, this data is not transformed. As such, it indicates that, on average, there are almost 496 start-up firms in the cities where the observations are located. The variables "City Business Students" and "City STEM Students" represent the number of business and STEM students in the cities where the observations are located, respectively. City Business Students has a mean of 4.374 and a standard deviation of 5.550, while "City STEM Students" has a mean of 4.733 and a standard deviation of 6.066. Again, these figures are subjected to logarithmic transformation. This indicates that, on average, there are more business and STEM students in the cities where the observations are located compared to the ZIP codes, naturally. Overall, the descriptive statistics suggest that the sample consists of start-up firms that have been in operation for an average of 6.305 years, and that are located in cities with a relatively large number of start-up firms and business and STEM students.

Sweden	mean	std	min	25%	50%	75%	max
exited	0.084	0.278	0	0	0	0	1
Lifetime	6.005	2.419	1	4	6	8	10
country Sweden	1.000	0.000	1	1	1	1	1
Num Firms In ZIP	38.751	97.553	1	1	5	17	351
ZIP Business Students	0.008	0.201	0	0	0	0	6.425
ZIP STEM Students	0.016	0.350	0	0	0	0	8.413
Num Firms In CITY	994.662	830.919	1	111	1763	1763	1763
City Business Students	0.692	2.132	0	0	0	0	9.472
City STEM Students	1.825	3.909	0	0	0	0	11.665

Table 8 Descriptive statistics of 1818 observations in Sweden

Presented in Table 4, are the descriptive statistics for 1818 observations of start-ups in Sweden. The "exited" variable indicates whether the start-up has failed or discontinued (1) or not (0). The mean and standard deviation are 0.084 and 0.278, respectively. This means that the majority of the start-ups in Sweden (92%) are still operating. The "Lifetime" variable represents the length of time in years from the start-up's founding until the end of the observation period. The mean and standard deviation are 6.005 and 2.419, respectively. The minimum and maximum values are 1 and 10 years, respectively. The "country Sweden" variable is, as with the United States, a binary variable indicating whether the start-up is located in Sweden (1) or not (0). As expected, the mean is 1, indicating that all observations are in Sweden. The "Num Firms In ZIP" have a mean and standard deviation of 38.751 and 97.553, respectively, with a minimum value of 1 and a maximum value of 351. For Sweden "ZIP Business Students" and "ZIP STEM Students" have mean values of 0.008 and 0.016, respectively, with very low standard deviations of 0.201 and 0.350, respectively. This suggests that the majority of ZIP codes have very few business and STEM students, this low number of graduates partly pertains to the issue mention in the methods section. The "Num Firms In CITY" have a mean and standard deviation are 994.662 and 830.919, respectively, with a minimum value of 1 and a maximum value of 1763. The "City Business Students" and "City STEM Students" variables represent the number of business and STEM students located in the same city as the start-up, respectively. The mean values are 0.692 and 1.825, respectively, with standard deviations of 2.132 and 3.909, respectively. These variables have higher mean and standard deviation values than their ZIP code counterparts, indicating that cities have more students than ZIP codes on average in Sweden aswell.

Variable Distributions

In the following stage, the distribution of the variables in the study are inspected. First, by examining the distribution of cluster densities in both Sweden and the United States, for both the city and ZIP level. Lastly, by investigating the distribution of graduates across both countries.

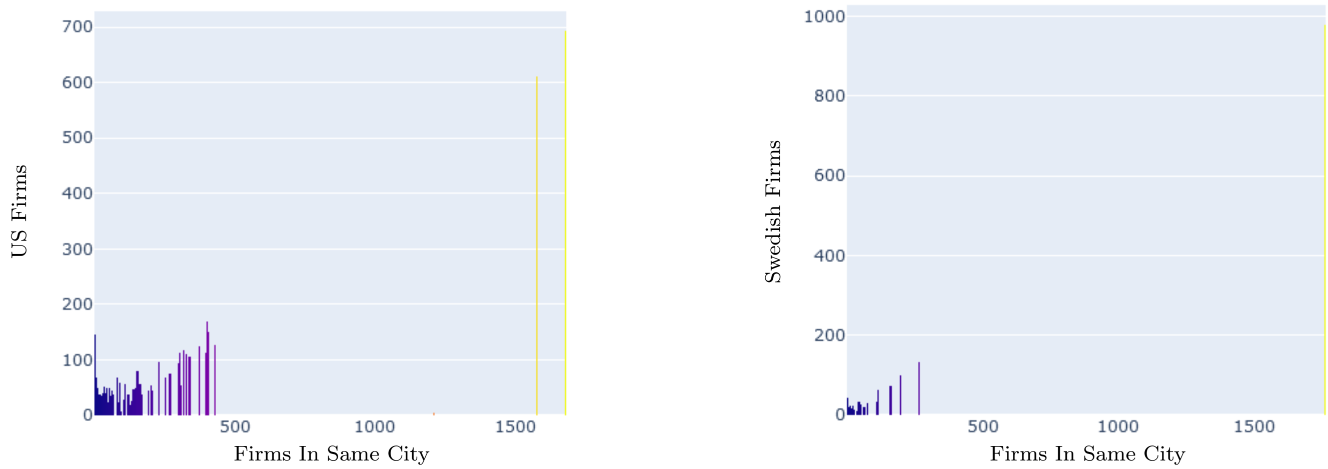


Figure 3 Dist. of firms in city-sized clusters.

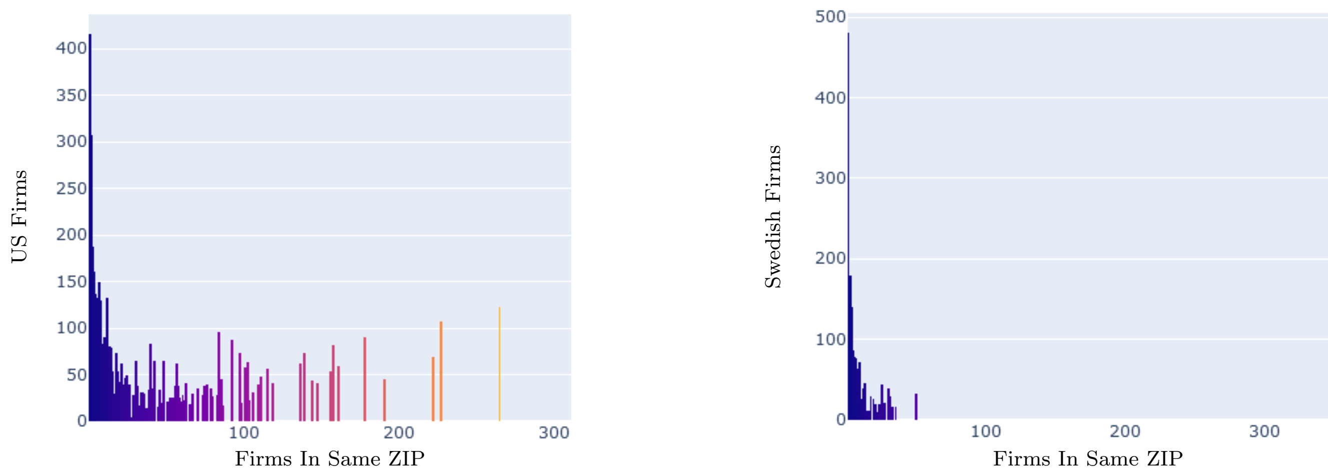


Figure 4 Dist. of firms in ZIP-sized clusters.

Examining the distribution of city-sized clusters of firms, Figure 3, the different densities and the frequency of the given density can be inspected. The high level distribution shows that the majority of firms are part of a less dense cluster. Inspecting the distribution at a level of higher granularity it is noted that the majority of firms are located in a city with 80 or less companies. Additionally, it is found that, in the sample, 42 Swedish firms and 145 US firms are alone in their city. Here, it is important to note that this study is not looking at every firm, the study is inspecting a subset of firms. Firms that are presented on CB insight. It's reasonable to anticipate these firms as firms located in rural cities. Cities that in reality have more firms in proximity, however, less or no firms alike to the type of firm this study investigates.

Similar to the distribution of firms in cities, it is observed that a majority of firms in both Sweden and the US are not located in high-density zipcodes see Figure 4. However, there are some differences between the two countries. For instance, in the sample, 480 Swedish firms and 415 US firms are the only ones with their specific zipcode. This could be due in part to the difference in zipcode size between the two countries, as well as the fact that the US has a 41 percent higher population density (SCB, 2021; World Bank, 2021; USCB, 2023; 2021). These factors suggest that more people in Sweden are living in the same zipcodes compared to the US, which may impact the distribution of firms in these areas. As with cities, there are specific regions with high density, 314 companies in the dataset are part of a remarkably dense zipcode. Looking through the data, it is confirmed that these companies are all within New-York city.

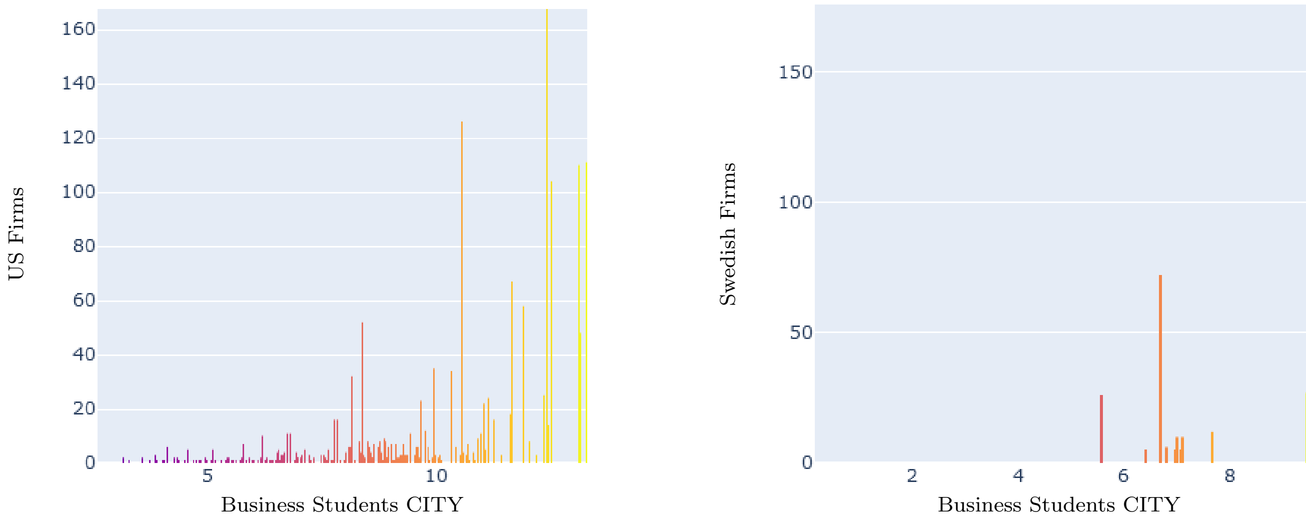


Figure 5 Dist. of business graduates in city-sized clusters.

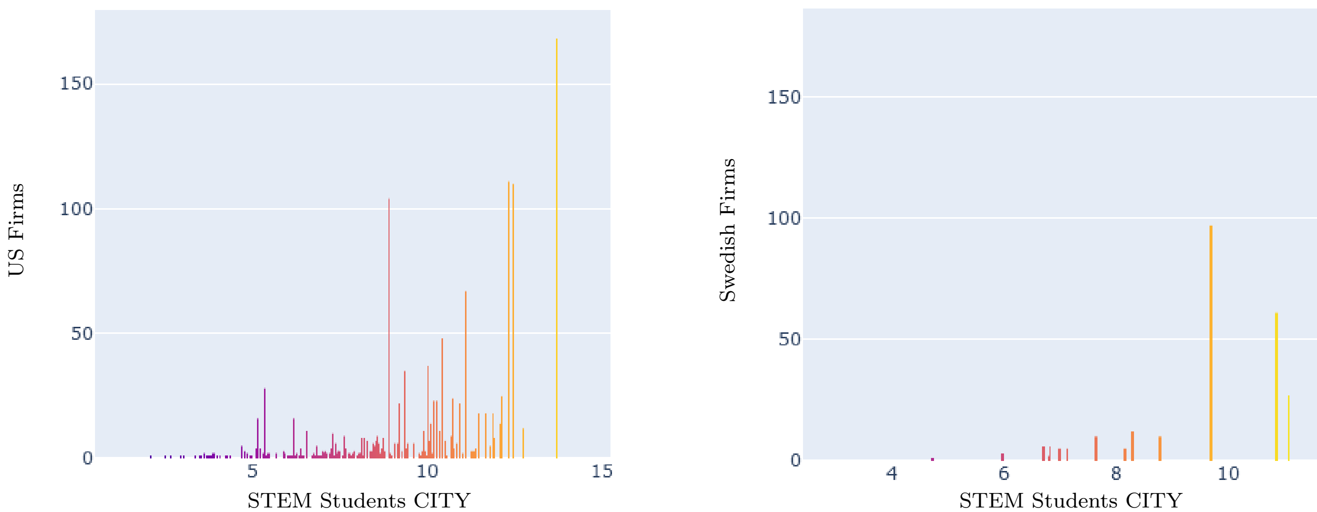


Figure 6 Dist. of STEM graduates in city-sized clusters.

Looking at the level of business graduates on the city level, see Figure 5, a higher concentration of students in the US as compared to Sweden is noted. This due to two reasons: Firstly, the United States has a greater number of universities as compared to Sweden, which naturally translates into a higher number of business graduates within the country. Secondly, there is the issue with the SCB data that has been previously discussed in the section that pertains to the data in Figure 5 and 6. As such, it is important to take this information into consideration when analyzing and interpreting the data presented in Figure 5 and 6.

Similar to the situation with business students, it is found that there is a comparable scarcity of data on STEM graduates in Sweden at the city level, see Figure 6. Moreover, the study observes a similar pattern of a higher density of students in the United States as compared to Sweden. It is important to note that the distribution of business and STEM students at the city level in both countries is relatively comparable. [11]

[11] Inspect the data on https://masterthesissurvivaljb.onrender.com/

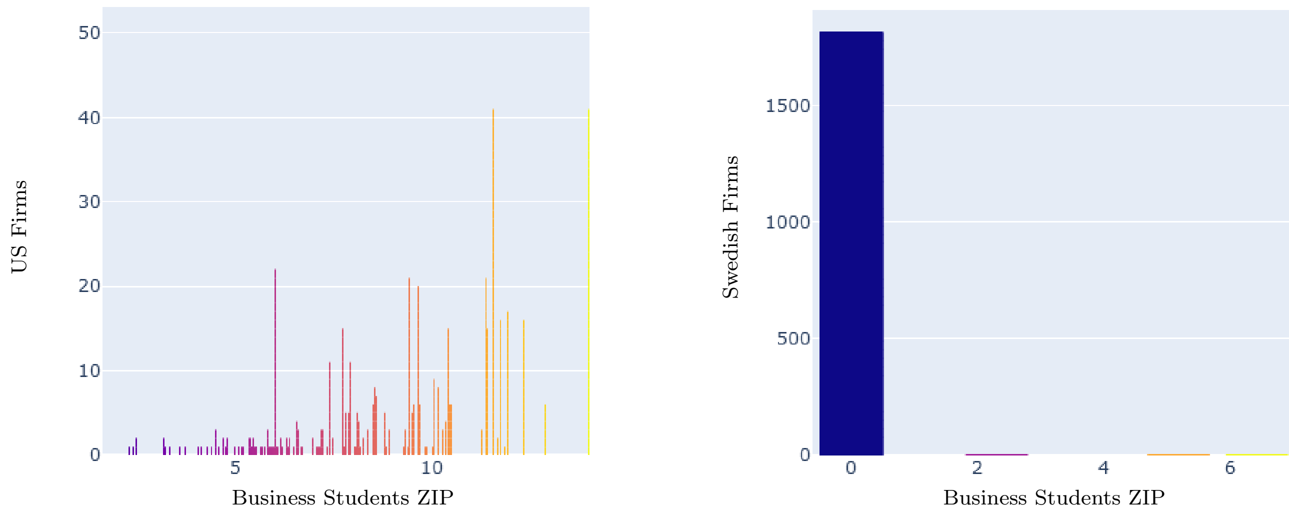


Figure 7 Dist. of business graduates in ZIP-sized clusters.

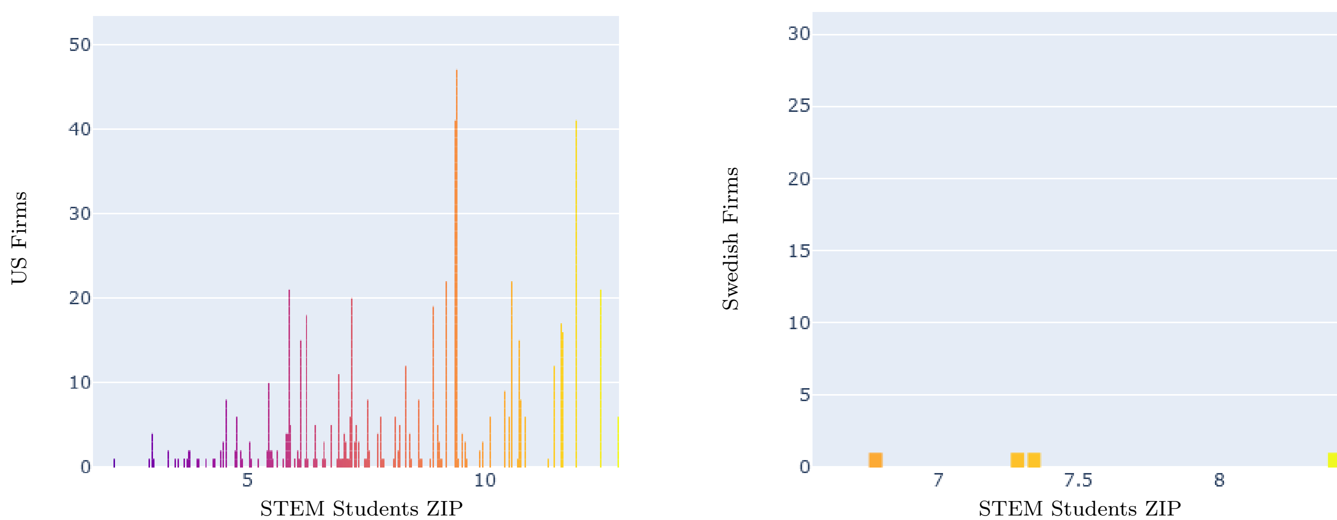


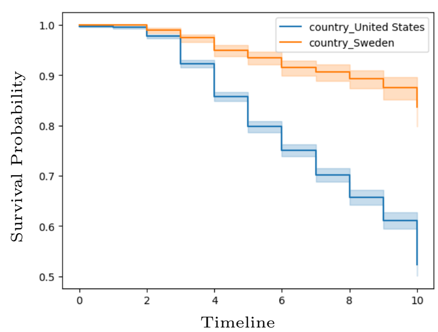
Figure 8 Dist. of STEM graduates in ZIP-sized clusters.

ZIP clusters are particularly affected by the issue of the data on graduates in Sweden, as expected. The quality of data on graduates is crucial for this study as it is essential in determining the impact of proximity to universities on start-up survival rates. However, limitations in the Swedish data may lead to uncertainty in the analysis, particularly regarding the effect of proximity to universities on start-up survival rates in Sweden.

Figure 7 and 8 reveal a comparable distribution of students in the business and STEM fields within ZIP sized clusters in the United States. Specifically, there exist similar groupings of STEM and business clusters with densities ranging between approximately 130 and 500 graduates in the respective field. Furthermore, there are notable peaks in both cases, indicating a higher number of firms that are part of clusters with densities above 500.

Non-parametric analysis

According to the results obtained from the Kaplan-Meier survival analysis, it can be confirmed that the survival rate of the firms in the United States is lower than that of Swedish firms, see Figure 9. However, it is noteworthy that both countries exhibit a comparatively lower risk of death during the initial three years. Beyond this period, US firms' survival probability drops significantly. For more detailed information regarding the survival rates of individual countries, see Table 5 and 6 in the appendix. Figure 9 displays start-up survival over time, with 7,514 firms initially observed. In the first year, 54 firms failed and 45 firms were no longer being tracked. By the end of year two, 344 firms had failed and 224 firms were no longer being tracked, leaving 7,442 firms still being observed. This trend continues each year until year 10, where only 737 firms remained. However, this doesn't account for factors like proximity to universities that could impact survival rates.



event at	removed	censored	entrance	at risk
0	18	0	7514	7514
1	54	45	0	7496
2	344	224	0	7442
3	710	384	0	7098
4	893	507	0	6388
5	984	668	0	5495
6	982	753	0	4511
7	962	779	0	3529
8	997	860	0	2567
9	833	741	0	1570
10	737	647	0	737

Figure 9 Survival plot and table Swedish versus US firms

Semi-parametric analysis

The results in Table 5 show the hazard ratios of clusters, proximate business graduates, and STEM graduates on zip and city level in start-up firms in Sweden and the United States, see appendix Table 7 for standard models results. The hazard ratios indicate the effect of each covariate on the risk of the event (i.e., start-up firm failure or discontinuation). A hazard ratio greater than 1 indicates an increased risk of the event associated with the covariate, while a hazard ratio less than 1 indicates a decreased risk.

All models in Table 5 were run using the Huber sandwich estimator, also known as the Wei-Lin estimate for the standard error. This estimator is robust to heteroscedasticity and allows for more accurate inference in the presence of outliers. By using this estimator, this study is able to obtain more reliable estimates of the hazard ratios and their associated standard errors, view the standard model in the appendix Table 7.

In the case of Zip sized clusters in the US, the number of firms in the ZIP has a hazard ratio of 0.759, statistically significant at $p < 0.01$, which means that an increase in the number of firms in the ZIP decreases the risk of start-up firm failure or discontinuation. This finding is opposed to the results of (Frenken et al., 2015) and contradictory to (Maine et al., 2010) and in line with findings from Bagley (2019) albeit Bagley studied the Internet and Telecommunications sector in Sweden. Additionally, in the non-robust model, an increase in the number of business students in the ZIP has a hazard ratio of 0.97 $p < 0.05$, indicating that an increase in the number of business students in the ZIP decreases the risk of start-up firm failure or discontinuation which is inline with (Cassia et al., 2009). Moreover, the hazard ratio for STEM students in the ZIP is not statistically significant (hazard ratio = 1.02, $p > 0.05$), which suggests that the number of STEM students in the ZIP is not associated with the risk of start-up firm failure or discontinuation.

For zip-sized clusters in Sweden, the number of firms in the ZIP has a hazard ratio of 0.570 $p < 0.01$, showing that an increase in the number of firms in the ZIP decreases the risk of start-up firm failure or discontinuation in Sweden. Interestingly, zip level proximity to business students in the ZIP is not statistically significant (hazard ratios = 0.00 $p > 0.98$) in the standard model, but significant in the robust model. Furthermore, in contrast to the United States the Swedish firms have a lower survival rate when close to a university with STEM students, as shown by the hazard ratio of 1.36 at statistical significance $p < 0.05$. Note that the data on Swedish graduates could be considered unreliable due to the reasons mentioned in the method section. Hopefully, better data will become available to confirm these findings.

In the case of City sized clusters in the US, an increase in the number of firms in the CITY has a hazard ratio of 0.90, statistically significant at $p < 0.01$, indicating that an increase in the number of firms in the CITY decreases the risk of start-up firm failure or discontinuation. Moreover, an increase in the number of business students in the CITY has a hazard ratio of 1.02** ($p < 0.05$), indicating that an increase in the number of business students on the city-level increases the risk of start-up firm failure or discontinuation. Moreover, the hazard ratio for STEM students in the CITY is not statistically significant (hazard ratio = 1.00, $p > 0.05$).

For City sized clusters in Sweden, an increase in the number of firms in the CITY has a hazard ratio of 1.02 with $p > 0.05$, indicating that the number of firms in the CITY is not associated with the risk of start-up firm failure or discontinuation. Additionally, the hazard ratios for business and STEM students in the CITY are not statistically significant (hazard ratios = 1.03 and 1.00, respectively; $p > 0.05$). Again likely a result of the unrepresentative data.

The goodness-of-fit and overall significance of the models can also be assessed by considering the Pseudo R2 and F-test results.

For the United States-Zip sized clusters model, the Pseudo R2 is 0.06, meaning that the model explains 6 percent of the variation in the outcome variable. The F-test is 56.3, with an associated Fp-value of 0.00, indicating that at least one of the covariates has a statistically significant effect on the outcome variable.

For the Sweden-Zip sized clusters model, the Pseudo R2 is 0.04, indicating that the model explains 4 percent of the variation in the outcome variable. The F-test is 19.8, with an associated Fp-value of 0.00, indicating that at least one of the covariates has a statistically significant effect on the outcome variable.

These results demonstrate that the models for both the United States and Sweden at the zip-sized cluster level provide some explanatory power, with at least one of the covariates having a significant effect on the outcome variable.

For the United States-City sized clusters model, the Pseudo R2 is 0.07, meaning that the model explains 7 percent of the variation in the outcome variable. The F-test is 78.4, with an associated Fp-value of 0.00, indicating that at least one of the covariates has a statistically significant effect on the outcome variable.

However, for the Sweden-City sized clusters model, the Pseudo R2 is -0.04, meaning that the model does not explain much of the variation in the outcome variable. The F-test is 0.65, with an associated Fp-value of 0.88, indicating that none of the covariates have a statistically significant effect on the outcome variable.

These results suggest that the model for the United States at the city-sized cluster level has some explanatory power, while the model for Sweden at the city-sized cluster level does not perform well in explaining the variation in the outcome variable. This could be due to the unrepresentative data mentioned earlier in the analysis.

It is important to note that the results of these analyses should be interpreted in the context of the logarithmic transformation. In particular, a hazard ratio of 1.0 indicates that a one-unit increase in the logarithm of the covariate is not associated with any change in the risk of start-up firm failure or discontinuation. Moreover, hazard ratios less than 1.0 indicate decreasing risk, while hazard ratios greater than 1.0 indicate increased risk on a logarithmic scale.

Table 9 Robust Hazard ratios of clusters, proximate business graduates, and stem graduates on zip and city level in start-up firms in Sweden and the United States. Huber sandwich estimator robust standard errors in parenthesis.

Covariates	**United States** **Zip sized clusters (N = 5696)**	**Sweden** **Zip sized clusters (N = 1818)**
Num Firms In ZIP (log)	0.759***	0.570***
	(0.038)	(0.15)
ZIP Business Students (log)	0.973**	0.00***
	(0.01)	(0.58)
ZIP STEM Students (log)	1.021	1.36***
	(0.01)	(0.11)
Pseudo-R^2	0.06	0.04
F-test (P-value)	56.3 (0.00***)	19.8 (0.00***)
	City sized clusters (N = 5696)	**City sized clusters (N = 1818)**
Num Firms In CITY (log)	0.900***	1.02
	(0.01)	(0.04)
CITY Business Students (log)	1.02**	1.03
	(0.01)	(0.04)
CITY STEM Students (log)	1.00	1.00
	(0.01)	(0.03)
Pseudo-R^2	0.07	-0.04
F-test (P-value)	78.4 (0.00***)	0.65 (0.88)

A hazard ratio greater than 1 indicates an increased risk of the event associated with the covariate, while a hazard ratio less than 1 indicates a decreased risk. Analyses were performed on transformed data using logarithmic transformations to correct for non-linearity. See the appendix for further details on the regressions. ****P*-value < 0.01. ***P*-value < 0.05.

The analyses reported in Table 5 were performed on logarithmically transformed data to correct for non-linearity. Specifically, the covariates representing the number of firms in the ZIP and CITY, as well as the number of business and STEM students in both the ZIP and CITY were logarithmically transformed prior to conducting the regression analyses. Therefore, the hazard ratios reported in Table 5 represent the effect of a one-unit increase in the logarithm of each covariate on the risk of start-up firm failure or discontinuation.

Going back to the previous example with Bioassay Systems and Media Systems Technology in table 6. Bioassay Systems, survived for 10 years in the United States with 46 firms in its ZIP code, the firm has zip level proximity to 2298 business students, and 459 STEM students in its ZIP code. Media Systems Technology survived for 7 years in the United States but had only 1 firm in its ZIP code and no business or STEM students.

Based on these results it can be concluded that the zip level proximity to a university impacted Bioassay survival, additionally, the findings support that Bioassay Systems survival chances were positively impacted by being part of a large zip cluster.

Overall, the results find that the number of firms in the ZIP and CITY has a significant effect on the risk of start-up firm failure or discontinuation in both Sweden and the United States. An example of which could be the case in Table 6. Moreover, the results find indication that an increase in the number of business students in the city decreases the risk and evidence that an increase in the ZIP level increases the risk of start-up firm failure or discontinuation, respectively, in the US. The direct interpretation is that zip level proximity to business students has a protective effect in the United States, while city-level proximity has a detrimental effect on survival rate. While the data for Sweden does show supporting evidence, these results are not considered reliable.

Discussion

Proximity to dissimilar firms

Our research aimed to investigate the impact of proximity to dissimilar firms and universities on start-up survival rates in Sweden and the United States. Our findings indicate that an increase in the number of firms in the ZIP and CITY decreases the risk of start-up firm failure or discontinuation in the United States. Our findings were only significant on the ZIP level for Sweden. These findings align with the theory of agglomeration economies and clustering. Furthermore, these result supports the theory that clustering dissimilar firms enhances the survival prospects of start-ups through knowledge spillovers, access to diverse knowledge sources, and labor, as suggested by Jacobs Externalities 1961. Moreover, in the United States, it is found that the effect is stronger on smaller areas of clusters, that is, more firms in closer proximity. As such, this study finds evidence supporting Jacobs Externalities' self-reinforcing property of diverse clusters.

Proximity to universites

The presence of graduates in the ZIP and CITY was found to have different effects on start-up survival in the United States and Sweden. The risk of start-up firm failure or discontinuation increased on the ZIP level for STEM graduates in Sweden, while in the United States, the results were found insignificant. However, in the United States, an increase in business students on the city level is found to be associated with an increased risk of start-up firm failure or discontinuation. This finding contrasts with the notion of Knowledge Spillover Theory of Entrepreneurship, which emphasizes the role of technological knowledge in driving economic growth and development Acs et al. (2009). Moreover, it contradicts the prior assumption that Lehman and Audretths confirm the theory of KTSE. By our measure of firm success, we find that there is a negative effect associated with the co-location of Business graduates and the survival of start-up firms. Furhtermore, our findings support Cassia et al. (2009) findings that universities play a significant role in the health of start-up firms, although we find evidence that this effect can also be negative. This study proposes a possible explanation for this result in the spirit of Jacobs Externalities. As this study uses a novel approach in two regards; Firstly, the type of graduates is specified as either STEM or Business. Secondly, the spillover is analyzed on a five-digit proximity level. This study proposes, that the reason for the contradicting results are found in the same benefits associated with dissimilarity in firm co-location. Following the notion that dissimilar firms benefit from each other's diverse knowledge spillover, it is proposed that this effect is present with university knowledge spillover as well. Prior research, only specifies that firms co-locate with universities. From this, the benefits are derived implicitly. However, this study finds that when the knowledge source is constrained to only that of Business students there is a negative effect associated with the knowledge spillover. This suggests that start-ups located near universities with diverse skill sets may benefit more from the knowledge spillovers and potential collaborations that proximity provides. Further research into the nature of this effect at greater distances to the university needs to be conducted to confirm this notion. As the work of Rosenthal and Strange (2003) suggests that the impact of proximity on firm survival may vary depending on the geographic scale at which it is analyzed. An additional explanation for this discrepancy could be that STEM and business graduates are more likely to contribute to the growth of established firms rather than new start-ups, or that the impact of STEM graduates is more nuanced and not captured by the analysis.

The institutional context

The analysis also highlights the role of the institutional context in shaping start-up survival rates. Differences in regulatory environments, access to capital, and support structures between Sweden and the United States can influence the relationship between proximity to dissimilar firms, universities, and start-up survival rates (Baumol, 1990). This underscores the need to consider institutional factors when examining the impact of proximity on start-up survival rates across different countries. Additionally, the results show that the dynamics between start-ups and dissimilar firms within the same geographic area can significantly influence start-up survival rates. This emphasizes the importance of understanding the specific dynamics between start-ups and dissimilar firms within a given geographic area in order to develop policies aimed at fostering thriving start-up ecosystems.

Limitations

The models for both countries at the zip-sized cluster level provided explanatory power, whereas the model for Sweden at the city-sized cluster level did not perform well. This may be due to data limitations, as the Swedish data may not be as reliable as the US data, or it could be indicative of different dynamics at different geographic scales previously mentioned. Further research could explore these differences and their implications for start-up survival in various geographic contexts.

Future research

Future research should investigate the implications of specific differences in the institutional context, considering factors such as cultural, economic, and institutional factors present in different geographic contexts.

Future research should also aim to incorporate additional control variables such as firm size and industry to provide a more comprehensive understanding of the relationship between clustering, knowledge spillovers, and start-up survival. This would allow us to better compare these findings with those of previous studies, such as (De Vaan et al., 2013), (Glaeser et al., 1992), (Beaudry and Swann, 2009), (Rigby and Brown, 2015), (Maine et al., 2010), (Larsson, 2009). By doing so, research may be able to identify any nuances and variations in the impact of proximity to dissimilar firms and universities on start-up survival across different industries and firm sizes. Moreover, it would be valuable to further explore the role of universities and graduates in the start-up ecosystem. The findings indicate a negative relationship between the presence of STEM and business graduates and start-up survival in certain contexts, which nuances the Knowledge Spillover Theory of Entrepreneurship. Additional research could help investigate the mechanisms through which these negative effects occur and identify potential strategies for leveraging the knowledge and skills of graduates to support start-up growth and success. By addressing these gaps in our understanding and building on this study's findings, future research has the potential to offer additional insights into the role of proximity, clustering, and diverse knowledge spillovers in promoting start-up survival and fostering innovation and economic growth.

Conclusion

This research study analyzed the impact of proximity to dissimilar firms and universities on start-up survival rates in Sweden and the United States. The results revealed that an increase in the number of rims within a ZIP code and city significantly decreased the risk of start-up failure or discontinuation in the United States, with similar findings on the ZIP level for Sweden. This supports the concept of agglomeration economies and clustering, as well as Jacobs Externalities' self-reinforcing property of diverse clusters.

The impact of proximity to business and STEM students differed between the two countries. This study found that an increase in the number of business students has a negative effect on firm survival at the city level in the United States. In Sweden, the risk of start-up failure or discontinuation increased at the ZIP level. These findings contrast the expectations from the Knowledge Spillover Theory of Entrepreneurship and underscore the need for further research to examine the exact cause. Albeit, the findings contribute towards a more nuanced understanding of the Knowledge Spillover theory and postulates a novel theory based on the notion of diversity benefits similar to Jacobs Externalities which explains the results in relation to prior research.

It is important to note that this study has some limitations, including potential issues with data reliability and variations in model performance at different geographic scales, particularly in the case of Sweden. Therefore, further research is needed to gain a better understanding of the factors that influence start-up survival rates in different contexts and to explore the unique cultural, economic, and institutional factors present in various geographic settings.

Despite these limitations, the research offers valuable insights for policymakers, entrepreneurs, and researchers interested in fostering a thriving start-up ecosystem. By understanding the impact of proximity to dissimilar firms and universities on start-up survival rates, policymakers can develop strategies to create a supportive environment for start-ups, while entrepreneurs can use this knowledge to make informed decisions about where to locate their businesses. Additionally, researchers can build on our findings to advance our understanding of the complex dynamics involved in the survival of start-up firms and the role of proximity, clustering, and knowledge spillovers in promoting start-up survival and fostering innovation and economic growth.

References

Acs, Zoltan J., David B. Audretsch, Pontus Braunerhjelm, and Bo Carlsson (2009). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, 32(1), 15-30. https://doi.org/10.1007/s11187-013-9505-9

Agarwal, R., and Audretsch, D. B. 2001 Does Entry Size Matter? The Impact of the Life Cycle and Technology on Firm Survival.. *The Journal of Industrial Economics* **49(1), 21–43**, https://doi.org/10.1111/1467-6451.00136

Alchian, A. A. (1950). Uncertainty, Evolution, and Economic Theory. *Journal of Political Economy*, 58(3), 211–221.

Altman, D. G. 1990 Practical Statistics for Medical Research. *Chapman and Hall/CRC* , https://doi.org/10.1201/9780429258589

Audretsch, D. B., and Lehmann, E. E. 2005 Does the Knowledge Spillover Theory of Entrepreneurship hold for regions?. *Research Policy* **34(8), 1191–1202**, https://doi.org/10.1016/j.respol.2005.03.012

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In R. R. Nelson (Ed.), *The rate and direction of inventive activity: Economic and social factors* (pp. 609-625). Princeton University Press. https://doi.org/10.1515/9781400879762-024

Arzaghi, M., and Henderson, J. V. 2008. Networking off Madison Avenue. *The Review of Economic Studies* **75(4), 1011–1038**, http://www.jstor.org/stable/20185068

Babbage, C. 1832 *On the Economy of Machinery and Manufactures*. London: Charles Knight.

Baumol, W. J. (1990). Entrepreneurship: Productive, unproductive, and destructive. *Journal of Political Economy*, 98(5), 893-921.

Bagley, M.J.O. 2019. Networks, geography and the survival of the firm.. *Journal of Evolutionary Economics* **29(4) 1173–1209**, https://doi.org/10.1007/s00191-019-00616-z

Beaudry, C., and Swann, G. M. P. 2009 Firm growth in industrial clusters of the United Kingdom. *Small Business Economics* **32(4), 409–424**, https://doi.org/10.1007/s11187-007-9083-9

Beaudry, C. and Schiffauerova, A. 2009. Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy* **38(2) 318-337**, https://doi.org/10.1016/j.respol.2008.11.008

Birgersson. J Analysis Tool: Survival Analysis Master Thesis https://masterthesissurvivaljb.onrender.com

Boschma, R. (2015). Do spinoff dynamics or agglomeration externalities drive industry clustering? A reappraisal of Steven Klepper's work.. *Industrial and Corporate Change* **24(4), 859–873**, https://doi.org/10.1093/icc/dtv024

Bryman, A., and Bell, E. 2015. *Business research methods*. Oxford University Press, USA.

Cefis, E., and Marsili, O. 2005 A matter of life and death: Innovation and firm survival. *Industrial and Corporate Change* **14(6), 1167–1192**, https://doi.org/10.1093/icc/dth081

Cassia, L., Colombelli, A., and Paleari, S. 2009 Firms' growth: Does the innovation system matter?. *Structural Change and Economic Dynamics* **20(3), 211–220.**, https://doi.org/10.1016/j.strueco.2009.01.001

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. 2012. The Productivity Advantages of Large Cities: Distinguishing Agglomeration From Firm Selection.. *Econometrica* **80(6), 2543–2594**, https://doi.org/10.3982/ECTA8442

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**(2), 187-220.

Creswell, J. W. 2014. *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications, Inc.

Collett, D. 2015 Modelling Survival Data in Medical Research. *Routledge and CRC Press.* , ISBN: 9781439856789

CB Insights. (n.d.). Data and Insights. Retrieved April 4, 2023, from https://www.cbinsights.com/what-we-offer/data/

College Scorecard. (2022). College Scorecard. Retrieved April 06, 2023, from https://collegescorecard.ed.gov/

Cutler, D. M., Glaeser, E. L., and Norberg, K. E. (2001). Explaining the Rise in Health Care Spending: The Role of Medical Innovation. In *Frontiers in Health Policy Research* (Vol. 4, pp. 1-32). MIT Press.

De Vaan, M., Boschma, R., and Frenken, K. 2013 Clustering and firm performance in project-based industries: The case of the global video game industry. *Journal of Economic Geography* **13(6), 965–991**, https://doi.org/10.1093/jeg/lbs038

Duranton, G. and Puga, D. 2004. Micro-foundations of Urban Agglomeration Economies. In J. V. Henderson and J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics* (Vol. 4, pp. 2063-2117). Elsevier. https://doi.org/10.1016/S1574-0080(04)80005-1

Durkheim, E. 1893 *The Division of Labor in Society*. Paris: Alcan.

Easterby-Smith, M., Thorpe, R., and Jackson, P. R. 2015. *Management and business research*. SAGE Publications Ltd.

Field, A. 2017. *Discovering statistics using IBM SPSS statistics*. SAGE Publications Ltd.

Flick, U. 2015. *Introducing research methodology: A beginner's guide to doing a research project*. SAGE Publications Ltd.

Frenken, K., Cefis, E., and Stam, E. 2015 Industrial Dynamics and Clusters: A Survey. *Regional Studies* **49(1), 10–27**, https://doi.org/10.1080/00343404.2014.904505

Glaeser, E. L., and Mare, D. C. (1994). Cities and Skills. *National Bureau of Economic Research*.

Glaeser, E. L., Kallal, H.D, Scheinkman, J, Shleifer, A 1992 Growth in Cities. *Journal of Political Economy* **100(6), 1126-52** , http://dx.doi.org/10.1086/261856

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

Glaeser, E. L., and Maré, D. C. 2001 Cities and Skills. *Journal of Labor Economics* **19(2), 316–342**, https://doi.org/10.1086/319563

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2010. *Multivariate data analysis*. Prentice Hall.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221–233.

Jacobs, J. (1961). *The death and life of great American cities*. Random House.

Kaplan, E. L., and Meier, P. 1958 Nonparametric Estimation from Incomplete Observations.. *Journal of the American Statistical Association* **53(282), 457–481**, https://doi.org/10.1080/01621459.1958.10501452

Kleinbaum, D. G., and Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). New York: Springer.

Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *The American Economic Review*, 86(3), 562-583. https://www.jstor.org/stable/2118212

Krugman, P. (1980). Scale Economies, Product Differentiation, and the Pattern of Trade. *American Economic Review*, **70(5)**, 950–959.

Krugman, P. 1991. Increasing Returns and Economic Geography. *Journal of Political Economy* **99(3) 483-499**, https://www.jstor.org/stable/2937739

Larsson, J. P. 2017Non-routine activities and the within-city geography of jobs.. *Urban Studies* **54(8), 1808–1833.**, https://doi.org/10.1177/0042098016643266

Lin, D. Y., and Wei, L. J. 1989. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84(408)**, 1074–1078.

Maine, E. M., Shapiro, D. M., and Vining, A. R. 2010 The role of clustering in the growth of new technology-based firms. *Small Business Economics* **34(2), 127–146**, https://doi.org/10.1007/s11187-008-9104-3

Mata, J., and Portugal, P. 1994 Life Duration of New Firms. *The Journal of Industrial Economics* **42(3), 227–245**, https://doi.org/10.2307/2950567

Marshall, A. (1890). *Principles of Economics*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137375261

O'Brien, R. M. 2007 A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity* **41(5), 673-690**, https://doi.org/10.1007/s11135-006-9018-6

Porter, M. E. (1990). The Competitive Advantage of Nations. *Harvard Business Review*, **68(2)**, 73–93.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., ... and Krabel, S. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy*, 42(2), 423-442.

Postnord. (2017). Postnummer Systemet I Sverige. Retrieved from https://www.postnord.se/siteassets/pdf/faktablad/postnummersystemet-i-sverige-171213.pdf

Rigby, D. L., and Brown, W. M. (2015) Who Benefits from Agglomeration?. *Regional Studies* **49(1), 28–43.**, https://doi.org/10.1080/00343404.2012.753141.

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

Rosenthal, S. S. and Strange, W. C. 2004. Evidence on the nature and sources of agglomeration economies. In J. V. Henderson and J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics* (Vol. 4, pp. 2119-2171). Elsevier. https://doi.org/10.1016/S1574-0080(04)80006-3

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71-S102. https://www.jstor.org/stable/2937632

Sorenson, O., and Audia, P. G. 2000 The Social Structure of Entrepreneurial Activity: Geographic Concentration of Footwear Production in the United States. *American Journal of Sociology* **106(2), 424–462**, https://doi.org/10.1086/316962

Statistics Sweden. Population statistics. Accessed on April 13, 2023. https://www.scb.se/en/finding-statistics/statistics-by-subject-area/population/population-composition/population-statistics/.

Statistics Sweden. (2022). Studenter och examina i högskoleutbildning på grundnivå och avancerad nivå [Students and degrees in higher education at the undergraduate and advanced level]. Retrieved April 06, 2023, from https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__UF__UF0205/ExaLasarOversikt/

Stuart, T., and Sorenson, O. 2003. he geography of opportunity: Spatial heterogeneity in founding rates and the performance of biotechnology firms. *Research Policy*, **2(2), 229–253.**, https://doi.org/10.1016/S0048-7333(02)00098-7

Schumpeter, J. A. (1934). *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Harvard University Press.

Saunders, M., Lewis, P., and Thornhill, A. *Research methods for business students*. Pearson Education Limited.

Tiebout, C. M. 1967 Local Governments and Spatial Mismatch. *Journal of Economic Literature* **5(4), 1443-1461**, https://www.jstor.org/stable/2720237

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer.

Trochim, W. M. K., and Donnelly, J. P. 2008. *The research methods knowledge base*. Atomic Dog Publishing, Inc.

Terrell, E. (2013). Research Guides: This Month in Business History: ZIP Code Introduced [Research guide]. Retrieved 27 April 2023, from https://guides.loc.gov/this-month-in-business-history/july/zip-code-introduced

United States Census Bureau. U.S. and World Population Clock. Accessed on April 13, 2023. https://www.census.gov/popclock/.

United States Census Bureau. 2020 Census State Population and Apportionment Results. Accessed on April 13, 2023. https://www.census.gov/library/publications/2021/dec/2020\-census-state-population-apportionment.html.

von Thünen, J. H. 1826 *The Isolated State*. Hamburg: Friedrich Perthes.

Weber, A. 1909 *Theory of the Location of Industries*. Chicago: The University of Chicago Press.

Wennberg, K., and Lindqvist, G.. The effect of clusters on the survival and performance of new firms. *Small Business Economics* **34(3), 221**–**241**, https://doi.org/10.1007/s11187-008-9123-0

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48(4)**, 817–838.

World Bank. Land area (sq. km) - Sweden. Accessed on April 13, 2023. https://data.worldbank.org/indicator/.

6sense. (2022). CB Insights market share. Retrieved April 26, 2023, from https://6sense.com/tech/market-research/cb-insights-market-\share

A Data-processing

A.1 *Merge check ZIP*

We check random values to see if the merge was performed correctly. Not all firms are in the same ZIP as a university, hence not all ZIP_firm have a match. We see that the original ZIP values match for each row. We see how there are now duplicates, this is because we have taken the cartesian product of the data frames and because some ZIP codes have multiple universities. We verify this by selecting all universities with ZIP "77079", and find two.

```
check_merge = zip_df[['Name','ZIP_firm','ZIP_uni']]
print(check_merge[34238:34243])
print(zip_df['school.name'].loc[zip_df['ZIP_uni']== '77079'].unique())
```

```
>index        Name                     ZIP_firm     ZIP_uni
>34238        HighRadius               77079        77079
>34239        Evolution Petroleum      77079        77079
>34240        Evolution Petroleum      77079        77079
>34241        Erskine Energy           77079        77079
>34242        Erskine Energy           77079        77079

>['University of Phoenix-Texas', 'Texas Health School']
```

```
#Looks good so we remove duplicates of each firm
zip_df = zip_df.drop_duplicates('Name')
zip_df
```

A.2 *Merge check city*

Similar to the ZIP codes, we check random city values post-merge. We see that the original city values match. We see duplicates, this is, again, because we have taken the cartesian product of the data frames. We verify this by selecting all universities in Rye, and Alhambra, we expect to find 4 universities in Alhambra, and 0 in Rye.

```
check_merge_city = city_df[['Name','school.name','City_firm','City_uni','STEM_students','Business_students']]
print('Rye:',city_df['school.name'].loc[city_df['City_uni']== 'Rye'].unique())
print('Alhambra',city_df['school.name'].loc[city_df['City_uni']== 'Alhambra'].unique())
print(pd.DataFrame(pd.concat([check_merge_city.loc[check_merge_city['City_uni']=='Rye'])
print(check_merge_city.loc[check_merge_city['City_uni']=='Alhambra']])))
```

```
>index      Name                            school.name                         City_firm   City_uni
>686        Propertyfirst.com               Platt College-Los Angeles           Alhambra    Alhambra
>687        Propertyfirst.com               California Inst. of Adv. Management Alhambra    Alhambra
>688        Propertyfirst.com               Alhambra Medical University         Alhambra    Alhambra
>689        Propertyfirst.com               Alhambra Beauty College             Alhambra    Alhambra
>690        Allied Pacific of California    Platt College-Los Angeles           Alhambra    Alhambra
>691        Allied Pacific of California    California  Inst. of Adv. Management Alhambra   Alhambra
>692        Allied Pacific of California    Alhambra Medical University         Alhambra    Alhambra
>693        Allied Pacific of California    Alhambra Beauty College             Alhambra    Alhambra

>index      Name                            school.name                         City_firm   City_uni
```

```
#Looks good, So we remove duplicates of each firm
city_df = city_df.drop_duplicates('Name')
city_df
```

B Survival tables per country

event at	removed	censored	entrance	at risk
0	18	0	5696	5696
1	36	27	0	5678
2	221	119	0	5642
3	512	212	0	5421
4	673	325	0	4909
5	746	450	0	4236
6	758	550	0	3490
7	728	553	0	2732
8	780	651	0	2004
9	644	559	0	1224
10	580	497	0	580

Table B1 Survival Table for the United States

event at	removed	censored	entrance	at risk
0	0	0	1818	1818
1	18	18	0	1818
2	123	105	0	1800
3	198	172	0	1677
4	220	182	0	1479
5	238	218	0	1259
6	224	203	0	1021
7	234	226	0	797
8	217	209	0	563
9	189	182	0	346
10	157	150	0	157

Table B2 Survival Table for Sweden

C Standard Model

Table C3 Hazard ratios of clusters, proximate business graduates, and stem graduates on zip and city level in start-up firms in Sweden and the United States.

Covariates	**Zip sized clusters US**	**Zip sized clusters Sweden**
Num Firms In ZIP (log)	0.758***	0.570***
	(0.037)	(0.15)
ZIP Business Students (log)	0.97**	0.00
	(0.01)	(462.77)
ZIP STEM Students (log)	1.02	1.36**
	(0.01)	(0.13)
	City sized clusters US	**City sized clusters Sweden**
Num Firms In CITY (log)	0.900***	1.02
	(0.01)	(0.04)
CITY Business Students (log)	1.02**	1.03
	(0.01)	(0.05)
CITY STEM Students (log)	1.00	1.00
	(0.01)	(0.03)

A hazard ratio greater than 1 indicates an increased risk of the event associated with the covariate, while a hazard ratio less than 1 indicates a decreased risk. Analyses were performed on transformed data using logarithmic transformations to correct for non-linearity. See the Tables below for further details on the regressions. $^{***}P$-value < 0.01.
$^{**}P$-value < 0.05.

covariate	coef	exp(coef)	se(coef)	p
Num Firms In ZIP	-0.276	0.759	0.037	0.000
ZIP Business Students	-0.027	0.973	0.014	0.047
ZIP STEM Students	0.021	1.021	0.013	0.108

Table C4 Fit summary US ZIP (Standard)

covariate	coef	exp(coef)	se(coef)	p
Num Firms In CITY	-0.105	0.900	0.012	0.000
City Business Students	0.018	1.018	0.008	0.020
City STEM Students	0.002	1.002	0.007	0.819

Table C5 Fit summary US CITY (Standard)

covariate	coef	exp(coef)	se(coef)	p
Num Firms In ZIP	-0.562	0.570	0.149	0.000
ZIP Business Students	-6.252	0.002	462.774	0.989
ZIP STEM Students	0.308	1.361	0.126	0.014

Table C6 Fit summary SWE ZIP (Standard)

covariate	coef	exp(coef)	se(coef)	p
Num Firms In CITY	0.019	1.019	0.040	0.636
City Business Students	0.031	1.032	0.046	0.493
City STEM Students	-0.005	0.995	0.027	0.846

Table C7 Fit summary SWE CITY(Standard)

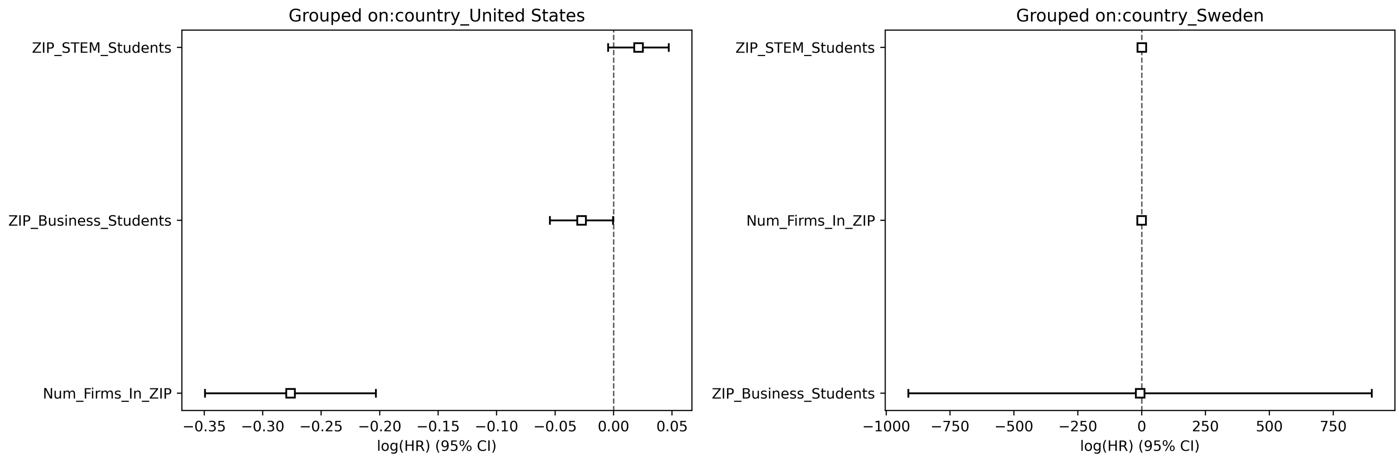


Figure C10 ZIP level proportional hazard.

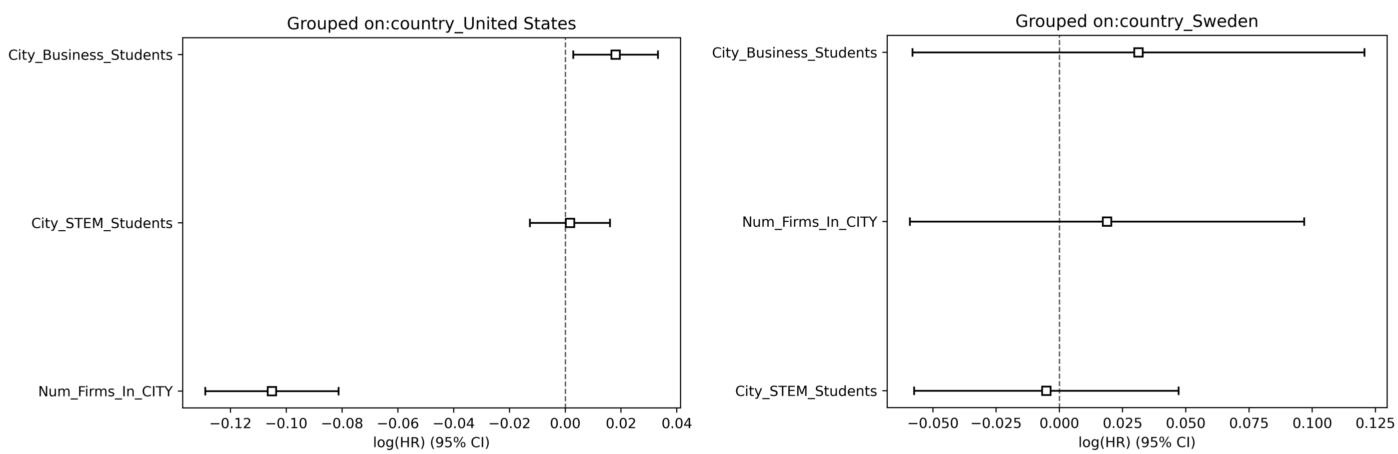


Figure C11 City level proportional hazard.

D Robust Model

covariate	coef	exp(coef)	se(coef)	p
Num Firms In ZIP	-0.276	0.759	0.038	0.000
ZIP Business Students	-0.027	0.973	0.013	0.038
ZIP STEM Students	0.021	1.021	0.013	0.106

Table D8 Fit summary US ZIP (Robust). The model's goodness of fit was assessed using the Pseudo-R2 measure, which was found to be 0.0579. Additionally, an F-test was conducted to test the overall significance of the model, resulting in a test statistic of 56.2946 and an associated p-value of 0.0000.

covariate	coef	exp(coef)	se(coef)	p
Num Firms In CITY	-0.105	0.900	0.012	0.000
CITY Business Students	0.018	1.018	0.008	0.017
CITY STEM Students	0.002	1.002	0.007	0.815

Table D9 Fit summary US CITY (Robust). The model's goodness of fit was assessed using the Pseudo-R2 measure, which was found to be 0.0712. Additionally, an F-test was conducted to test the overall significance of the model, resulting in a test statistic of 78.4499 and an associated p-value of 0.0000.

covariate	coef	exp(coef)	se(coef)	p
Num Firms In ZIP	-0.562	0.570	0.148	0.000
ZIP Business Students	-6.252	0.002	0.575	0.000
ZIP STEM Students	0.308	1.361	0.111	0.006

Table D10 Fit summary SWE ZIP (Robust). The model's goodness of fit was assessed using the Pseudo-R2 measure, which was found to be 0.0382. Additionally, an F-test was conducted to test the overall significance of the model, resulting in a test statistic of 19.8812 and an associated p-value of 0.0002.

covariate	coef	exp(coef)	se(coef)	p
Num Firms In CITY	0.019	1.019	0.037	0.615
CITY Business Students	0.031	1.032	0.041	0.449
CITY STEM Students	-0.005	0.995	0.025	0.834

Table D11 Fit summary SWE CITY (Robust). The model's goodness of fit was assessed using the Pseudo-R2 measure, which was found to be -0.042. Additionally, an F-test was conducted to test the overall significance of the model, resulting in a test statistic of 0.659 and an associated p-value of 0.883.

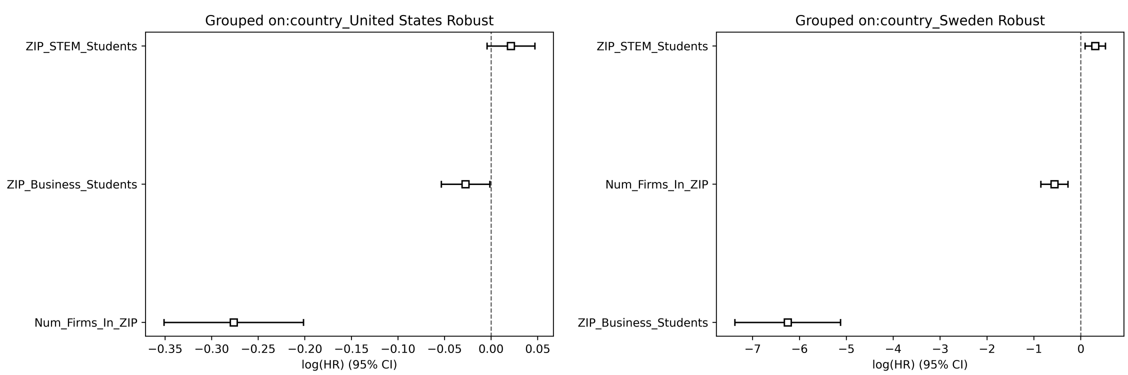


Figure D12 ZIP level proportional hazard.

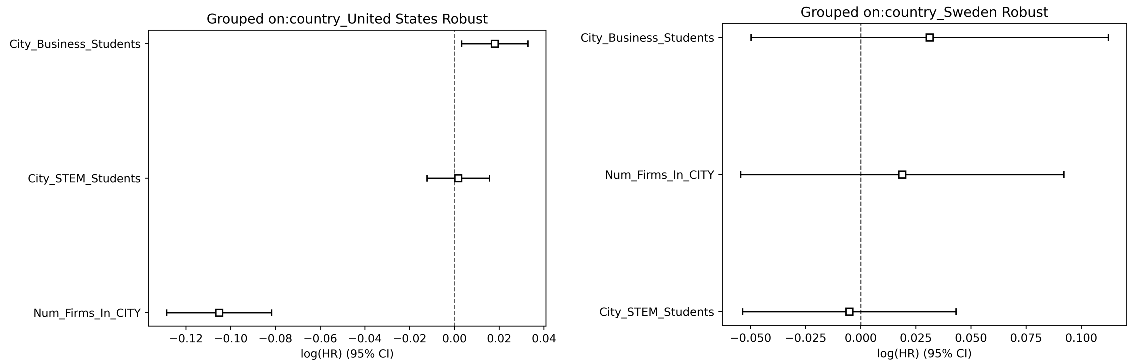


Figure D13 City level proportional hazard.