Master Thesis

submitted within the UNIGIS MSc programme

Interfaculty Department of Geoinformatics - Z_GIS

University of Salzburg

"Analysis and prediction of microgeographic location of software firms"

A comparison of the U.S. and Germany

submitted by

Sina Bernhard, M.A.

U104371

A thesis submitted in partial fulfilment of the requirements of
the degree of
Master of Science – MSc

Advisor:

Assoc. Prof. Dr. Bernd Resch

Berlin, 28.03.2023

Abstract

This study examines the spatial distribution and identifies clusters of software companies in the United States and Germany. Using a dataset of over 15 million U.S. and 1.4 million German street-level geocoded company observations and 24 location factors, a predictive model was created to predict the local presence of software companies within each square kilometer in both countries. The study reveals that the location factors of agglomeration, some socioeconomic factors, and terrain have a positive influence on the location of software companies in both countries, while the effects of other location factors, such as amenities and infrastructure, vary. The purpose of this study is to use exploratory spatial data analysis (ESDA) to strengthen the understanding of economic processes and contribute to the empirical literature. This interdisciplinary study between geoinformatics and economics will provide insights into the location theory and cluster theory of the software industry in the United States and Germany.

Keywords: Location Theory, Cluster Theory, Software Industry, United States, Germany, Location Factors, Predictive Modelling

Table of Content

List of Figures

List of Tables

List of Abbreviations

BosWash	Boston-Washington city band
Chipitts	Chicago-Pittsburgh city band
CV	Coefficient of Variation
DI	Dispersion
EDA	Exploratory Data Analysis
ESDA	Exploratory Spatial Data Analysis
ESRI	Environmental Systems Research Institute
FRA	Federal Railroad Administration
GIS	Geoinformation System
IRR	Incidence-rate Rations
NB	Negative Binomial
MAUP	Modifiable Areal Unit Problem
MLE	Maximum Likelihood Estimator
OSM	OpenStreetMap
p	Parameterizations for the NB regression (alpha parameter)
SanSan	San Diego-San Francisco city band
SE	Standard Error
SQL	Structured Query Language
U.S.	United States of America
VGI	Volunteered Geographic Information

1 Introduction

1.1 Motivation

The main motivation of this thesis is to investigate the distribution and concentration of the software industry in the U.S. and the effects of specific location factors on the location of software companies. With the help of a modeling of spatial relationships between location factors and the number of software companies located in the U.S., predictions for software clusters located in Germany will be made and compared and discussed with other scientific findings.

Man has been economically active for thousands of years. Then as now, the decision of the optimal firm location is of particular importance. The factors influencing the choice of a location are decisive for the success or failure of any business.

The systematic study of sites and their relationship to economic activity has a relatively short history. The pioneering work of Johann Heinrich von Thünen in 1826 (O'Kelly & Bryan, 1996) and later of Edward A. Ross (1896) and Alfred Weber (Friedrich, 1929) brought the topic into the public discourse and provided the starting point for further research (Porter, 2000; Marshall, 2014). Analogous to the economic interests of their time, most studies of location theories focused on agricultural and manufacturing industries.

In the process of high industrialization of the Western world and the transformation to modern industrialized countries, the choice of the optimal location gained further importance and the attention of other actors. Unbalanced development opportunities between different regions have created the need for active participation of governments and local authorities in regional development (Fischer & Nijkamp, 2021). A holistic understanding of location factors and their importance on corporate location decisions and business performance can have important implications for all stakeholders.

Since the 1990s, economic activity has been significantly influenced by two megatrends that are in fact mutually dependent: Globalization and digitalization. The interplay of these two developments has led to significant changes in economic life, resulting in the software industry and triggering the transition from the industrial to the digital age (Petersen & Thode, 2015). In this market, some specific distinctions apply. As a result of the non-physical nature of software products themselves, competition between software vendors is global and, unlike in other industries, home advantage plays a negligible role in the vendors' national markets (Buxmann *et al.*, 2015). Thus, it can be assumed that the relevance of location factors for software companies differs from agricultural or industrial companies.

Economic activities and industries are not evenly distributed, but tend to concentrate in certain areas (Dunlap & Santos, 2021). The clustering of like-minded market participants reveals the role of location in competitive advantage, according to Porter (2000). Although space has lost some of its economic importance in the process of globalization and digitalization - an increasingly complex, knowledge-based and dynamic environment - and decision factors have changed over time, spatial concentrations of competing as well as cooperating firms and institutions seem to continue to play an important role (Porter, 2000), such that today production is aggregated into a small number of global clusters that trade their products with the entire world (Swann, 2008).

The economic importance of software companies for regional development is now recognized beyond doubt. Lighthouses such as the Silicon Valley or the Rhine-Main-Neckar IT-Cluster are regarded as the "magic formula" for regional development (Duranton & Overman, 2002). The United States of America (U.S.) was and is the world leader in software development. The PwC ranking "Global Software 100" points very clearly how strong the U.S. dominates the industry. More than three quarters of the one hundred companies listed are headquartered in the U.S, with Germany well behind in second place with only 5 companies (Columbus, 2016). For adequate interventions and sensible regional policies, there is a need for a scientific basis to help decide what causes software companies to settle in a certain location. For this, it is important to understand how general and how strong is the tendency for clustering in the software industry and which micro-geographic factors play a crucial role. It is worth asking whether the examples mentioned are the exception rather than the rule and on which spatial scale this clustering takes place.

1.2 Research Question

In the context of this work, the following research questions should be answered:

i. Are there significant differences in location patterns of software companies between the U.S. and Germany?

ii. Are the locations in both countries explained by the same location factors?

The spatial differences in the characteristics of the location factors lead to a spatial differentiation of the location qualities. The clarification of these questions is intended to improve the understanding of which factors favor the agglomeration of software companies in the U.S., how this is represented in terms of distribution and size at the micro-geographic level, and what significant differences exist compared to Germany.

To answer the questions, the first task is to examine the distribution and density of software companies in the United States. For this purpose, global statistics (Global Moran's I) are used to search for the presence of spatial autocorrelation in the entire data set and to identify spatial patterns. This purely exploratory study provides an initial overview for subsequent analyses.

Provided that a clustering of software companies is identified in the first step, appropriate local statistics (Hot-Spot Analysis, Local Morans'I) are applied. The number and size of software industry clusters are examined, as well as whether or not they tend to be small-scaled and community-based.

Once geographic clusters have been identified, and thus proven that the software industry is clustered, the next step is to investigate which factors explain the location decisions of software companies in the U.S., comparing the results with Germany. Of primary interest are location factors at the micro geographic level, depending on the availability of data. Since company locations are discontinuous data and spatial asymmetries are of interest, a logistic regression analysis of count data is used to determine which location factors at the micro-geographic level are significant for the location of software companies and how strong the relationship of each determinant is on company location. A large number of different location factors (agglomeration, infrastructure, socio-economic factors, topography to amenities) has been included in the regression analysis. Through a model comparison, we aim to identify those location factors that are particularly strongly related to the business locations of software companies.

Finally, the validated location prediction model is applied (location pattern modeling). The results of the developed prediction model are subsequently compared and discussed with those of Kinne and Resch (2018) - "Analyzing and Predicting Micro-Location Patterns of Software Firms" in Germany - and the research questions are attempted to be answered.

1.3 Structure of the paper

After the introduction and the outlining of my personal motivation and research question, this paper is structured as follows.

Chapter 2 provides an overview of the state of the art of research on which this thesis is based. The following Chapter describes the data set and variables used, before Chapter 4 outlines the methods of spatial data analysis. Chapter 5 then presents the results of the research and subsequently, Chapter 6 discusses the results in relation to the research questions and draws important conclusions from this research. The last chapter closes the thesis with a summary of the findings and provides an outlook for further research approaches.

2 Literature review

The software industry is considered extremely important for innovation and competitiveness in many regions worldwide. Therefore, the location decisions of companies are of great importance to them and policy makers are interested in creating a favorable business environment by developing the appropriate location factors. Pioneers of location theory such as Ross (1896), Weber (Friedrich, 1929) and Marshall (2013) were forerunners in the analysis of location patterns and location advantages and their findings from that time still have their unrestricted validity today: a company's decision for or against a location is significantly influenced by the existing location factors. Following this pioneering research, many scientists have paid attention to the process of location decision-making and resulting spatial distributions of economic activity (Maskell & Kebir, 2005; Murray, 2009; Capello, 2014; Farhauer & Kröll, 2014a; Taylor & Francis, 2021).

Location studies, which have been conducted for a long time, have identified a wide range of location factors in order to understand how location decisions of companies in space can be explained (Capello, 2014). For knowledge-intensive industries, such as the software industry, the proximity of other companies and spatial networks are advantageous. Companies compete with each other while learning from each other through formal, as well as informal, communication and collaboration. Spatial proximity facilitates information sharing and creates regional knowledge infrastructures rooted in inter-firm networks, interpersonal relationships, and local learning processes (Asheim & Coenen, 2008; Rammer *et al.*, 2016; Saxenian, 2018). This also includes accessibility to nearby research institutes and universities (Anselin *et al.*, 1997; Faria *et al.*, 2020; Rammer *et al.*, 2020) and consequently to a skilled young workforce (Anselin *et al.*, 1997). One of the location factors to be rated particularly high is access to a sufficiently available and qualified workforce (Combes & Duranton, 2006; Cader *et al.*, 2013). The location of software companies in densely populated regions is a logical choice (Egeln *et al.*, 2004) and also has the advantage that the population diversity, e.g. due to people with a migration background, is higher in these areas, which again promotes innovation and creativity (Lee *et al.*, 2004). In contrast, cultural and entertainment facilities in the proximity seem to have rather no significant influence on business locations (Kinne & Resch, 2018; Rammer *et al.*, 2020; Smętkowski *et al.*, 2021).

The high-tech and software industry is relatively young and has not been the focus of scientific attention for very long. The first studies were conducted in the 1980s. These studies highlighted the very rapid growth of the high-tech industry in the states of Massachusetts and California. Already at that time a discovery was made: The high-tech industry is not subject to the constraints that traditionally explain the location of companies (Dorfman, 1983). The software industry in particular is fundamentally different in terms of its economic characteristics from industrial goods or classic services (Buxmann *et al.*, 2015). Based on the rapid technological development, experts predicted in the mid-1990s' the growing insignificance of physical space in the information age and predicted a world in which social and economic interactions take place in virtual space (Zook, 2005b). Swann's (2008) concept of the "four ages of space" demonstrates the transformation of the economic significance of location due to (modern) technologies. While we can now say that some predictions of that time have come true, "the rhetoric of spacelessness" (Zook, 2005b) has not entirely been realized. Regions such as San Francisco and Boston were key locations for high-tech industry (Zook, 2005a) in the U.S. then as now, and continue to be the subject of numerous empirical studies (Rosegrant & Lampe, 1992; Ashish Arora *et al.*, 2001; Briant *et al.*, 2010; Saxenian, 2018).

Analyses that examine how spatial factors can influence companies' location choices face a crucial question: the selection of the spatial unit to be used. Different spatial units and the scale of analysis used can lead to bias in the results (Briant *et al.*, 2010; Arauzo-Carod & Manjón-Antolín, 2012). This phenomenon is usually referred to as the modifiable area units problem (MAUP), being defined by location, scale, and shape dimension (Openshaw, 1983). Relationships between location factors and companies that are relevant at macro-level may not be so at micro-level. Rapid technological progress and the increasing availability of spatial data favored new research approaches (Elwood *et al.*, 2012) and also initiated the process of shifting research from spatially large units to smaller units (Friedman *et al.*, 1992). Depending on the interpretation, a micro-location analysis includes the immediate environment such as neighborhoods, districts, and nearby areas before moving to a disaggregated analysis of individuals, households, and businesses (Hooton, 2016).

Great importance is attached to spatial proximity and is reflected in related theoretical concepts: industrial districts, milieux innovateurs, new industrial, localized production (Moulaert & Sekia, 2003). The degree of spatial spillover between university research and high-tech innovation was explored by Anselin *et al.* (1997), who found that the spillover effect of university research on innovation extended over a 50-mile radius. Other studies conclude that knowledge spillovers tend to occur between geographically closer actors. The spatial proximity to other companies, research centers and universities is an indication that innovation activities have a tendency to cluster spatially (Maté-Sánchez-Val *et al.*, 2018). The distance where it is advantageous for a company to be located near research centers and universities is quantified far lower by Maté-Sánchez-Val *et al.* (2018) at less than 2.39 Km in the Madrid study area. Rammer *et al.* (2020) even concluded that the micro-geographic coverage of knowledge sources in urban environments decreases sharply within a few hundred meters. The question of knowledge exchange between industries versus within an industry was explored by Altunbaş *et al.* (2013), who concluded - based on data on the growth of 109 British cities between 1951 and 1991 - that the main knowledge spillovers occur across industries rather than within industries. This conclusion has important implications for agglomeration effects in urban areas. Given the high pressure to be innovative in high-tech industries, knowledge spillovers as an innovation driver are crucial (Arauzo-Carod, 2021).

The literature outlines that agglomeration advantages, such as those offered by metropolitan areas, are important for attracting new high-tech firms as these firms benefit from gains in efficiency (Méndez-Ortega & Arauzo-Carod, 2019; Arauzo-Carod, 2021). The number of high-tech companies located in central areas of major cities is increasing, according to Florida and Mellander (2016). Méndez-Ortega and Arauzo-Carod (2019) discovered that agglomeration patterns vary and are related to different local policies, urban structures and other factors.

A study by Cader *et al.* (2013) in the U.S. state of Kansas shows that a location like Silicon Valley is not necessary to attract high-tech companies. Nevertheless, even in the sparsely populated Midwest, metropolitan areas seem to have an advantage over neighboring and rural regions, although this does not necessarily mean that the latter have no potential. Following Dunlap and Santos (2021) study on multinational high-tech companies in the U.S., domestic high-tech companies are more likely to take advantage of the agglomeration benefits of highly developed regions, while foreign companies seem to be more attracted to less developed regions.

Much of the literature on location theories of high-tech industry refers to the U.S. and specific reference regions such as the Silicon Valley. It is only in the recent past that high-tech clusters in Europe have received increasing attention, both from academics and policy makers. The enormous growth rates of the industry, which contributes to a myriad of new companies and hires a large number of qualified employees, makes the increasing interest in this sector understandable. Meanwhile, there are also numerous high-tech clusters in Europe: "Europe's Silicon Valley" of enterprise software is located in Germany's southwest (Peters, 2021).

Location theory has long been a central component of economic geography, providing explanations and predictions of location and the determining factors that influence it. This study will contribute to the empirical location literature by attempting to enhance the understanding of which factors are crucial for the software industry's locations and by making a comparison between the U.S. and Germany.

3 Data

Suitable data for micro-geographic location analyses have only recently become increasingly available. In particular, the engagement of volunteer enthusiasts who collect, organize and subsequently disseminate spatial information to the public - crowdsourcing of geodata - has contributed positively to this development. Meanwhile, official agencies are also increasingly making geodata available for free use. In this master thesis, geodata from four main sources were used:

- Institute / Enterprise data
- Administrative data
- ArcGIS Living Atlas data (ESRI)
- OpenStreetMap data

3.1 Institute Data

A unique dataset of around 1.4 million street-level geocoded firm observations from the Orbis database[1] (Bureau van Dijk, 2022) for Germany and another 15 million from the big data, analytics and marketing service provider Infogroup (Infogroup, 2016) for the United States were obtained. The data provided was cross-checked with official data from federal agencies to compare and verify the number of tax-paying companies in general and software companies in particular in each of the countries.

3.1.1 Definition Software Industry

The decision of what to include in the definition of the software industry is difficult, as there is no single or coherent definition yet. Although "software" is a widely used term, it cannot be easily reduced to a specific sector, business model, or type of company. Software cannot be identified as part of a specific sector but rather as a transversal technology. The fields in which it is applied do not have precise boundaries and have become an essential part of conducting business all over the world. As Müller (2003) correctly points out in his empirical study of software companies, there are difficulties

[1] Data set contains only companies with web address. Information on the method can be obtained from the study by Kinne & Axenbeck (2020).

in differentiating since the telecommunications and electronics sectors, among others, also carry out software development themselves on a large scale.

According to Tyrväinen and Mazhelis (2009), an independent software industry as we know today developed from the vertical spin-off from the computer industry towards the end of the last century. In the literature a distinction is made between software providers in the narrower and broader sense (Buxmann *et al.*, 2015).

The focus of this paper is on companies in the primary software industry, whose business model is based on the following content: programming activity, software development, data processing, website development, IT consulting, web portal, web server services, web hosting. Explicitly excluded are companies whose business model is based on the distribution or publishing of software, IT training or mass reproduction of software. Relevant software companies are derived by using the NACE code for Germany (the statistical classification of economic activities in the European Community) and NAICS code for the U.S. (the North American industrial classification system).

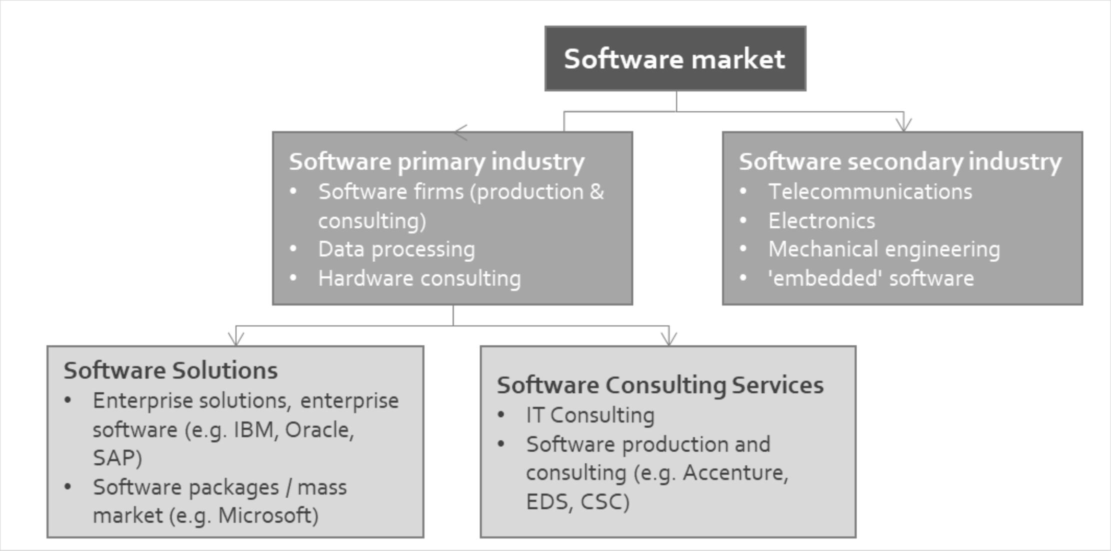

Software market
Software primary industry
• Software firms (production & consulting)
• Data processing
• Hardware consulting
Software secondary industry
• Telecommunications
• Electronics
• Mechanical engineering
• 'embedded' software
Software Solutions
• Enterprise solutions, enterprise software (e.g. IBM, Oracle, SAP)
• Software packages / mass market (e.g. Microsoft)
Software Consulting Services
• IT Consulting
• Software production and consulting (e.g. Accenture, EDS, CSC)

Figure 1: Software market classification (Source: (Müller, 2003)

United States

The 2019 enterprise dataset provided by Infogroup's Historical Business Data (Infogroup, 2016) covers the total stock of businesses in the U.S.. The Infogroup's data set was cross-checked with the "Number of Firms and Establishments" from the United States Census Bureau (2021). The dataset contains numerous characteristics such as the branch of industry through NAICS-codes and postal addresses. To correctly reflect the share of software companies on all companies, the U.S. classifications by the six-digit NAICS-code was chosen. Table 1 below provides the NAICS-codes used in this study, which are based on the primary activity of companies determined by the six-digit code that represents their largest revenue segment in the most recent completed fiscal year.

NAICS-Code	Industry Classification
511210	Software Publishers
518210	Data Processing, Hosting, and Related Services
541511	Custom Computer Programming Services
541512	Computer Systems Design Services
541513	Computer Facilities Management Services
541519	Other Computer Related Services

Table 1: Relevant business sectors (NAICS 2017) in the U.S.

The enterprise dataset consists of over 15 Mio. businesses in total. Exactly 109,267 software companies were identified, representing 0.72 % of all companies in the dataset.

The states Hawaii and Alaska, as well as U.S. overseas territories like Puerto Rico, were removed from the dataset for further analysis, reducing it by 836 software companies, or approximately 1%.[2]

Germany

For Germany, we use a comprehensive business dataset of around 1.4 million geo-coded firm observations at street level, from which 51,734 software companies were extracted. The following table lists the NACE-codes used corresponding to the American NAICS-codes.

NACE-Code	Industry Classification
6201	Programming activities
62011	Development and programming of internet presentations
62019	Other software development
62020	Consulting services in the field of information technology

[2] Known as Contiguous United States or Lower 48.

62030	Data processing equipment operation for third parties
62090	Other information technology service activities
63110	Data processing, hosting and related activities
63120	Web portals

Table 2: Relevant business sectors (NACE 2008) in Germany

Data on socio-demographic topics at a low spatial aggregation level was hardly available from German public authorities and was thankfully provided by German enterprises. Data on proportion of academics and students in the German population at block group or zip code level was provided by infas360 GmbH (infas360, 2022) and salary by Nexiga GmbH (Nexiga, 2022). Data on average age, unemployment rate and foreigner rate was provided by Real Estate Pilot AG (Real Estate Pilot, 2022).

3.2 Administrative Geodata

For the U.S. as for Germany, data released by various official authorities were used. Data on population density comes from the European Commission, a spatial raster dataset of 250 meter that is mapping human settlements worldwide (Pesaresi *et al.*, 2019).

Socio-demographic data for the United States was obtained mainly from the U.S. Census Bureau's American Community Survey and Department of Housing and Urban Development, based on 2020 census data (United States Census Bureau, 2022). Data on American life expectancy was obtained from County Health Rankings 2018 of the University of Wisconsin Population Health Institute (County Health Rankings & Roadmaps, 2022). Data on life expectancy for Germany was only available at county level, made available by the Federal Institute for Research on Building, Urban Affairs and Spatial Development (Bundesinstitut für Bau-, Stand- und Raumforschung, 2020).

3.3 ArcGIS Living Atlas Data

The ArcGIS Living Atlas of the World, hosted by the U.S. company ESRI Inc., is the leading collection of spatial information from around the world (ESRI, 2020). Similar to the publicly provided geographic information collected by volunteers, countless (geospatial) data from public authorities, such as census data, are provided by ESRI on this platform. According to ESRI, a curator reviews the data before publication to ensure quality requirements (ESRI, 2021). ArcGIS Living Atlas of the World is a robust source of data that can be used to analyze and map data from categories such as people, infrastructure, environment, and beyond.

The WMS World Slope GMTED was utilized in this study to compute the terrain for both countries, with a 250-meter cell resolution grid created from the 2010 Global Multi-resolution Terrain Elevation Data (United States Geological Survey, 2022).

The freely available datasets "Global Fixed Broadband" from Ookla were used for obtaining data on broadband Internet network coverage (Ookla, 2022).

3.4 OpenStreetMap Data

OpenStreetMap (OSM) is a crowdsourcing project of geographic data, supported by many contributors and made available to the public. Members use mobile devices to gather geospatial information and contribute it to crowd-sourced datasets that are shared online (Elwood *et al.*, 2012).

In the absence of available geodata from official agencies on cultural, entertainment and recreational facilities as well as infrastructure and universities, OSM data was obtained for the variables on infrastructure: entertainment, culture, recreation, universities, airports, interstate / highway and public transport stops (bus, metro, tram). The corresponding filters for filtering the OSM data can be obtained from the appendix.

Concerns regarding the quality of the data and the potential for specialized applications are continuously discussed in the literature (Flanagin & Metzger, 2008; Elwood *et al.*, 2012; Fonte *et al.*, 2015). A number of studies have examined Volunteered Geographic Information (VGI) data and questioned the completeness and spatial, temporal, and semantic accuracies of the data. There are different ways to investigate the quality of VGI projects. The most common methods - extrinsic quality assessment - involve comparing attributes (e.g. position, location) of the data with other state or commercial geospatial data sets (Haklay, 2010; Wang *et al.*, 2013). This measure was performed as part of the work with data on transportation infrastructure (airports, interstate / highway, public transport stops) and educational institutions (universities). No significant qualitative differences were found.

In places where no reference data sets exist, the methods of intrinsic quality assessment are applied. This method refers exclusively to the characteristics of the data itself. These studies usually involve locally applied spatial analyses, such as cities or municipalities, with a focus on individual object classes (Sehra *et al.*, 2017).

The quality of VGI data depends primarily on the number of peers who review and edit the content, also known as Linus's Law. However, „despite concerns over the quality and trustworthiness of VGI, preliminary assessment seems to indicate that VGI could serve as a potential data source to address research questions across geography" (Elwood *et al.*, 2012).

4 Methods of spatial data analysis

Spatial data analysis is the process of analyzing spatial relationships using a geographic information system (Haining, 2009) in combination with programming languages like python and SQL. This section is about spatial data analysis methods used in this paper.

4.1 Exploratory spatial data analysis

Exploratory Spatial Data Analysis (ESDA) is a general term that refers to the analysis of geospatial data in an exploratory manner using a variety of methods to describe and visualize these spatial effects (Abelairas-Etxebarria & Astorkiza, 2020). ESDA is somewhat the opposite of non-spatial analyses (EDA), and puts space and the relative observation of positions at the center of the investigation. In addition to spatial analyses, non-spatial analyses are also used in this work.

4.1.1 Analysis of patterns - Global Spatial Autocorrelation

Spatial autocorrelation distinguishes two different perspectives: Global and local autocorrelation analysis. Global spatial autocorrelation looks for a general pattern between proximity and similarity of spatial data across the overall map sample. This analysis looks at the overall trend of the study area and allows conclusions to be drawn about the degree of clustering of business locations of software companies in the U.S. in a single value.

In contrast to exogenous causes, i.e. location factors such as infrastructure, population density, etc., the focus here is on whether the spatial pattern of the distribution of software companies can also explain the characteristic values of the variables. These are therefore endogenously caused processes. The analysis focuses on investigating whether and to what extent the value of a variable at a particular location in the study area is conditioned by the values of neighboring values (Anselin, 1988). This study examines the location of software companies in relation to other surrounding companies within the same industry.

Spatial analyses using spatially aggregated data, such as administrative data, are problematic due to the modifiable area unit problem (MAUP) and can lead to inaccurate causal relationships and inconsistent results (Kitchin & Thrift, 2009; Arauzo-Carod &

Manjón-Antolín, 2012). As expected, the administrative levels (counties, tracts, block groups) of the U.S. exhibit serious differences in area size (high coefficient of variation: CV=5.38 for Tracts Area), but for population the magnitude of variation appears to be small (low coefficient of variation: CV=0.54 for Tracts Population). The enormous area differences can be explained by the low population density of about 36 people per km^2 (in comparison: Germany - 235, EU (27) - 109) and at the same time a high population concentration in the coastal regions. The boundaries of administrative units in the U.S, as well as in Germany, are based on population. This problem is to be countered as follows: Research has demonstrated that global analyses are most effective when the spatial pattern is consistent across the study area (Briant *et al.*, 2010). The aggregation of point data - such as company locations - is also a prerequisite for performing spatial calculations such as Moran's I and is relevant for the weighting matrix. Therefore, quadrat analysis, a descriptive metric for measuring point distributions, is applied. Instead of calculating the Euclidean distance between two points and its nearest neighbor, quadrat analysis covers the study area with uniform grid cells and summarizes the point frequencies per grid cell via a spatial link. As a result, a two-dimensional distribution of business locations is transformed to a one-dimensional distribution using the grid cells (Illian *et al.*, 2008).

To calculate the optimal length of the side of a square I, it requires the extent of the study area A (United States), divided by the number of features n (locations of software companies) multiplied by two.

$$I = \sqrt{2\frac{A}{n}}$$

Ideally, the squares of the grid cells should be large enough to not have too many squares with zero values, but small enough to have some spatial differentiation. This thesis shows that this cannot always be guaranteed.

For the cross-sectional analysis of spatial relations, the contiguity-based spatial weight of queen contiguity is recommended. This method is particularly suitable for regular grids such as square polygons. Moreover, as pointed out in chapter 2 Literature review, spillover effects decrease sharply with spatial distance. Therefore, the first order contiguity is used, i.e. only polygons that have a common boundary or common vertex.

4.1.1.1 Global Moran’s Index

According to (Kelejian & Prucha, 2001), the Global Moran's I is the most widely used spatial statistic for analyzing big data. The Moran's I index measures spatial similarity by calculating the deviation of two values from their common mean x̄, and it can be defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{i,j} \ (x_i \ - \ \bar{x})\left(x_j \ - \ \bar{x}\right)}{\sum_i \ (x_i \ - \ x)^2}$$

The analysis provides information on whether the pattern of company locations is clustered, dispersed or random and forms the basis for the subsequent analyses. However, the Moran Index does not indicate whether clustering is high or low. The index ranges from 0.0 (absolute polycentricity) to 1.0 (absolute monocentricity) and merely reflects the fact that similar values are located in similar places.

4.1.2 Analysis of clusters – Local Spatial Autocorrelation

Local forms of spatial autocorrelation - the so-called "Local Indicators of Spatial Association" (LISAs) - have been developed to address small spatial units and examine local interrelationships. The local statistics evaluate each location in the context of its immediate neighborhood and compare the local situation to the global one, with an evaluation for significant deviations from the global trend. Clusters or hot spots are identified that either determine the overall cluster pattern or reflect spatial heterogeneity that deviates from the global pattern (Anselin, 1995).

In the context of economic questions such as company locations, positive spatial autocorrelations will mostly be present. In terms of content, this interregional influence can be traced back to spillover effects, which are supported in the endogenous growth theory and the theory of innovative networks. Using LISA statistics, localized concentrations of software clusters and hotspots in the study area can be easily identified and visualized using cartographic representations. For this reason, the study applies the two most common measurement techniques with the aim of discovering spatial clusters and detecting anomalies. These are the cluster and outlier analysis hot spot analysis. The main objective of both techniques is the detection of territorial patterns of software company’s, location and to identify local pockets of dependence that may not be visible using global statistics.

4.1.2.1 Cluster and Outlier Analysis (Anselin Local Moran's I)

The Local Moran statistic was proposed by Anselin (1995) and is used to identify groupings or anomalous values according to the criterion of proximity. First, it accounts for local instabilities in the overall spatial composition and assesses the impact of distinct localities on the global statistic. This allows us to identify contiguous regions that exhibit stronger clustering or opposite spatial autocorrelation (outliers) compared to the global Moran coefficient. For each observation in the dataset, there is an index of the extent of significant spatial clustering of similar values around that observation.

The formula for the local Moran coefficient is:

$$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}\right) \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

The global Moran coefficient is obtained as the average of the local measures.

4.1.2.2 Hot-Spot-Analysis (Getis-Ord Gi*)

The Getis-Ord Gi* Analysis will serve as a complement to indicate the places where software companies appear with different values to that of the surrounding.

The Getis-Ord Gi was introduced by Getis and Ord (1992) shortly before the local Moran's I statistic and focuses directly on the identification of groupings within the area - hot spots and cold spots - rather than classifying them into clusters and outliers. The analysis reflects whether local clusters are concentrations of high values (hot spots) or concentrations of low values (cold spots) and is a valuable complement to the cluster and outlier analysis. Getis-Ord Gi* is used in this work to detect local regions with positive spatial auto-correlation.

The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{x} \sum_{j=1}^{n} w_{i,j}}{s \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j} - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}}$$

The use of both geostatistical techniques will give rise to different results as the conception of their formulation varies. However, due to the size of the U.S. a detailed analysis of the individual clusters identified cannot be covered in this work as the focus is more on the whole country. The same applied to Germany.

4.2 Analysis of relations - Regression Analyses

Logistic regression analysis is used to investigate which location factors are significant at the micro-geographic level for the settling of software companies and how strong the correlation of the different determinants is with the company location.

4.2.1.1 Location Factors

Although there are many location factors influencing the company's choice of location, about two dozen location factors were selected to consider. From a modeler's perspective, it is important to determine which factors are critical to a company's location decision. The selection of location factors in this thesis is primarily based on the study by Kinne and Resch (2018) and was supplemented with regard to the U.S..

Table 3 lists the location factors which are included in the analysis, a majority of which were available as vector data. The variables are grouped into four categories: agglomeration, infrastructure, socio-economic and amenities location factors. In addition, terrain was considered to be of relevance too.

Location Factor	Description (Germany)	Description (US)
	Agglomeration location factors	**Agglomeration location factors**
Company density	Number of local companies (in 10).	Number of local companies (in 10).
Company density squared	Squared number of local firms (in 10).	Squared number of local firms (in 10).
Software companies share	Proportion of software companies in the local business population (in %)	Proportion of software companies in the local business population (in %)
Population density	Population per cell (in 100).	Population per cell (in 100)
Population density squared	Squared number of inhabitants per cell (in 100).	Squared number of inhabitants per cell (in 100)
Street centrality	Street (network) density calculation (1) High value = High density	Street density calculation (1) High value = High density
Universities	Distance to the nearest university (in km).	Distance to the nearest university (in km).
Research institute	Distance to the nearest research institute (in km).	Distance to the nearest research institute (in km).
	Infrastructure location factors	**Infrastructure location factors**
Network coverage broadband Internet	Average latency (upload / download speed) (in Mbps). High value = high internet speed	Average latency (upload / download speed) (in Mbps). High value = high internet speed
Interstate / Highway	Distance to nearest high way / interstate (in km).	Distance to nearest high way / interstate (in km).
Airport	Distance to nearest main civil airport (in km).	Distance to nearest main civil airport (in km).
Public transport	Weighted count of public transport stops.	Weighted count of public transport stops.
	Socio-economic location factors	**Socio-economic location factors**
Salary	Monthly household income (median) (in 100 EUR).	Monthly household income (median) (in 100 EUR).
Educated workforce	Proportion of employees with a university degree (in %)	Proportion of employees with a university degree (in %)
Student rate	Proportion of students in the local population in %.	Proportion of students in the local population in %.

Business tax	Municipal business rate (in 100) fixed by the municipality. High values = high taxes	State cooperate tax rates fixed by the states (in %). High values = high taxes
Life expectancy	Average life expectancy of the population (in years).	Average life expectancy of the population (in years).
Average age	Average age (median) of the population.	Average age (median) of the population.
Unemployment rate	Proportion of unemployed in the working-age population (in %).	Proportion of unemployed in the working-age population (in %).
Migration background	Proportion of people of non-German nationality in the total population (in %).	Proportion of non-U.S.-citizens in the total population (in %).
	Amenities location factors	**Amenities location factors**
Recreation	Number of recreational, community and sports facilities.	Number of recreational, community and sports facilities.
Culture	Number of cultural sites and facilities.	Number of cultural sites and facilities.
Entertainment	Number of dining, nightlife and general entertainment facilities.	Number of dining, nightlife and general entertainment facilities.
	Other	**Other**
Terrain	Average slope or gradient (in degree). High values = hillside location	Average slope or gradient (in degree). High values = hillside location

Table 3: Description of location factors

4.2.1.2 Regression Analysis

A widely used method for modeling the relationship between location factors and the number of local businesses per territorial unit are count data regression models. Count data models were developed in the 1990's and are a subtype of discrete response regression models. Count data is distributed as "non-negative integers, inherently heteroskedastic, right-skewed, and have a variance that increases with the mean" (Hilbe, 2011). If the variance exceeds the mean, the model is said to be overdispersed. The literature abounds with alternative models for count data, however, the Poisson and two forms of the negative binomial models predominate the present applications (Hilbe, 2011).

Table 4 shows that dispersion (DI: ratio of variance to mean distribution) varies strongly with the aggregation level. At all aggregation levels, the variance significantly exceeds the mean, suggesting that the pattern of software company locations seems to be highly clustered and overdispersed. The overdispersion of the data increases even further as the level of aggregation decreases. In addition, the comparison of the mean and median indicates that there is a high proportion of zero values, significantly more than one would expect from a Poisson distribution, as well as extreme outliers influencing the data. The descriptive information demonstrates that selecting the level of aggregation has a significant impact on the statistical features of the analyzed spatial pattern and determines the suitable statistical distribution.

Scale	Obs.	null	Max.	$\bar{x}$	$\tilde{x}$	σ	DI
1 km	8,082,191	99%	336	0.013	0	0.38	10.78
5 km	326,846	94%	1,437	0.33	0	4.70	66.94
10 km	82,553	87%	1,572	1.31	0	12.60	121.19
25 km	13,519	64%	2,820	8.00	0	53.50	357.78
50 km	3,501	38%	4,033	31.00	2	158.30	808.35

Table 4: Descriptive Data - Software Companies U.S.

The decision guidance below is meant here to illustrate the reasonableness for or against a particular regression model in this thesis.

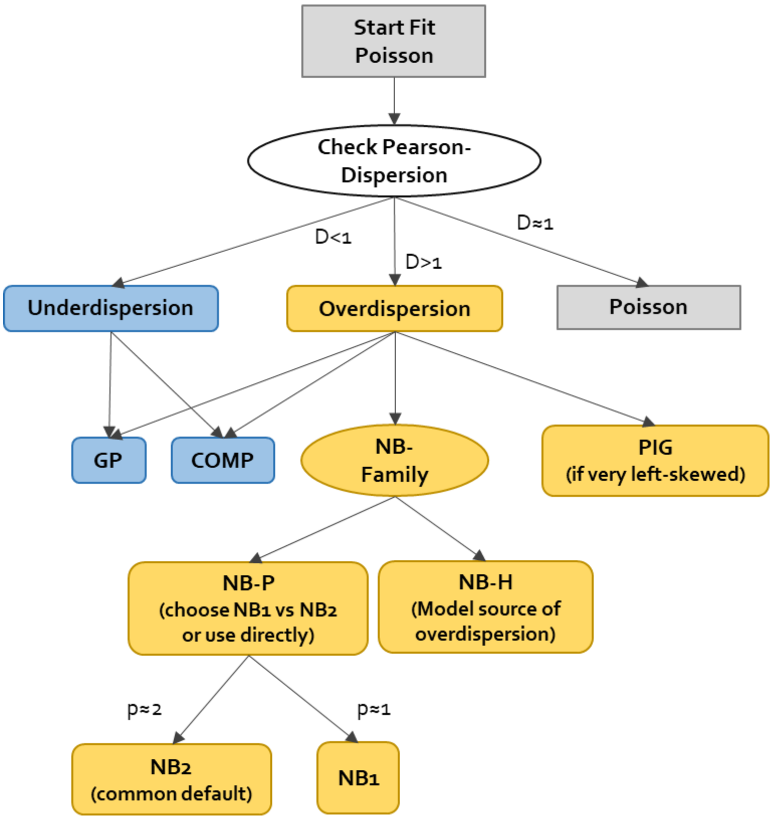


Figure 2: Count data regression - decision tree (Source: Freie Universität Berlin (2023))

The basic assumption is

$$var = \bar{x} + \alpha x^{-p}$$

Where p (alpha parameter) refers to the parameterizations for NB regression.

- p = 0: Poisson
- p = 1: NB1
- p = 2: NB2

Figure 3 shows that the mean $\bar{x}$ of software companies clearly depends on the size of the grid. While the average value is below two companies for a 1 km² grid, the number increases exponentially to over 4,330 for a 50 km² grid.

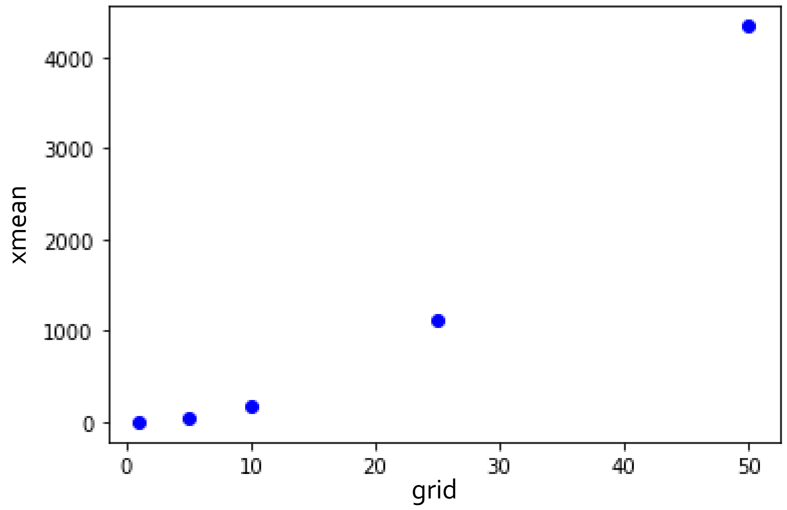


Figure 3: Linear regression: grid vs. xmean U.S.

The graph shows that there appears to be a polynomial relationship. First, a quadratic approach is attempted and the dependent variable is modeled to fit to the polynomial in x.

$$y = ax^2 + bx + c$$

The coefficient of determination of the quadratic regression is R^2 = 99% ($p \leq 0.001$), which confirms that the model assumption is justified and leads to the following consideration. The mean value is proportional to the area.

$$\bar{x} \sim l^2 \sim A$$

For a negative binomial distribution follows:

$$var \sim A^p$$

The following then applies to the standard deviation:

$$\sigma = \sqrt{var} \sim (A^p)^{\frac{1}{2}} = A^{p/2}$$

Thus, the dispersion index is:

$$\frac{var}{n} \sim \frac{A^p}{A} = A^{p-1} = l^{2P-2}$$

And the coefficient of variation:

$$\frac{\sigma}{n} \sim \frac{A^{p/2}}{A} = A^{p/2-1} = l^{P-2}$$

In other words, the coefficient of variation is expected to decrease when $p < 2$. At the same time, the dispersion index for $p > 1$ will increase with increasing cell size. Thus, the observed behavior can be explained by a p-value ranging between 1 and 2. The optimal result with R^2 at around 98 % is observed at a value of P = 1.8, and speaks for a quadratic negative binomial distribution.

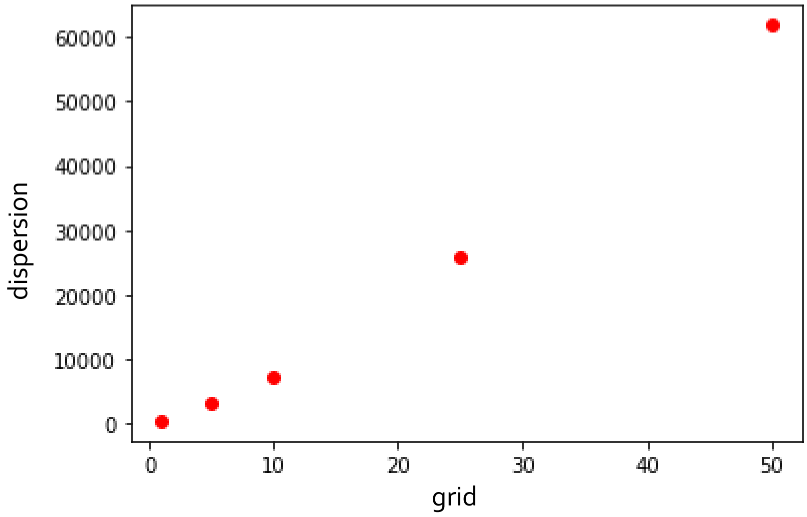


Figure 4: Regression dispersion vs. grid (R^2 = 0.983, $p \leq 0.001$)

The behavior of the coefficient of variation also fits this model quite well.

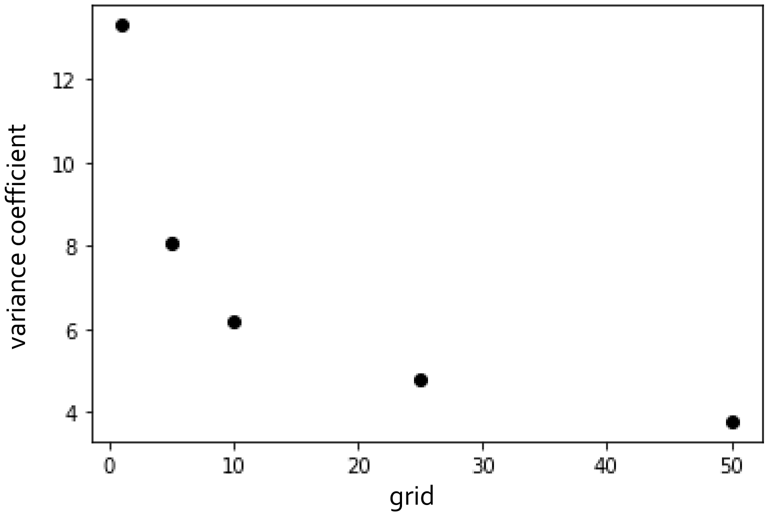


Figure 5: Regression variance coefficient vs. grid (R^2 = 0.983, $p \leq 0.001$)

The negative binomial model adjusts for Poisson overdispersion. Negative binomial distribution is one of the most widely used discrete probability models, but unlike the Poisson model it does not condition on the mean being equal to the variance. Also, the negative binomial model accounts for data with small numbers of events (such as locations of software companies as opposed to grocery stores) and thus has a significantly higher proportion of zero values than would be expected by coincidence (Perumean-Chaney *et al.*, 2013).

There are two different formulations of the negative binomial distribution, NB1 and NB2, whose numbering is due to the exponent in the second term of their variances. The NB1, the linear negative binomial distribution where p = 1 and the more common NB2, a quadratic negative binomial distribution where p = 2 (Hilbe, 2014). According to Hilbe (2014), for large samples overdispersion can exist even from 1.05. The NB2 is the standard form of negative binomial regression used to estimate data with overdispersion, and is the form of model that most statisticians understand by negative binomial. The NB2 is usually the first model to be reverted to when we find that a Poisson model is over dispersive. The line of argument makes it obvious that a Poisson model does not fit well to the given data, regardless of the level of aggregation, and makes the NB model quite plausible.

Figure 6 graphically shows the observations and the expected Poisson distribution at a grid size of 10 Km.

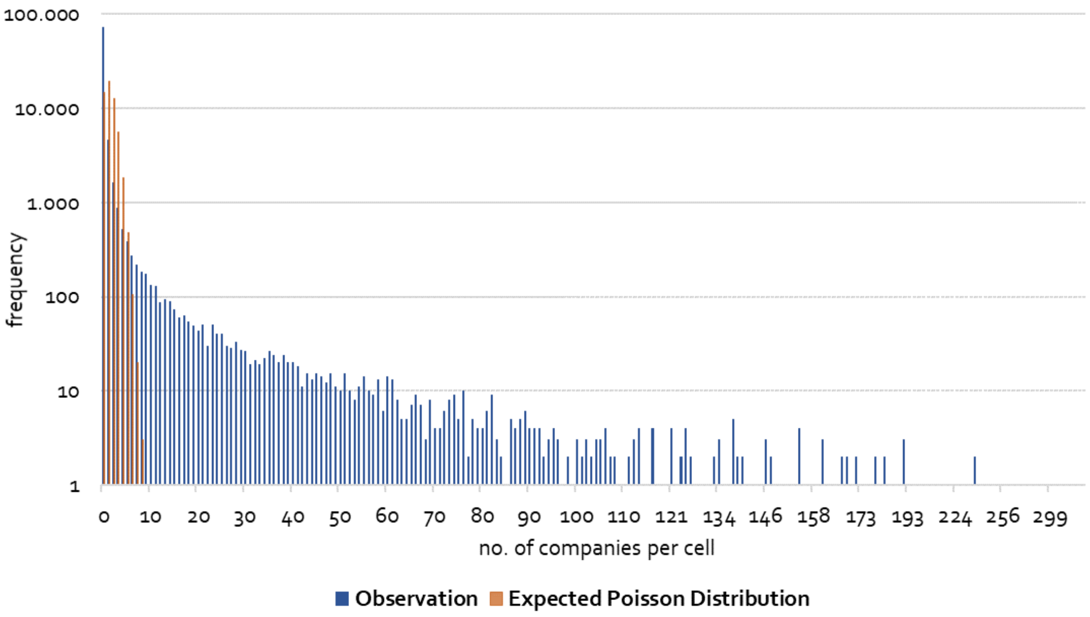


Figure 6: Observation vs. Poisson Model Expectation (10 km grid)

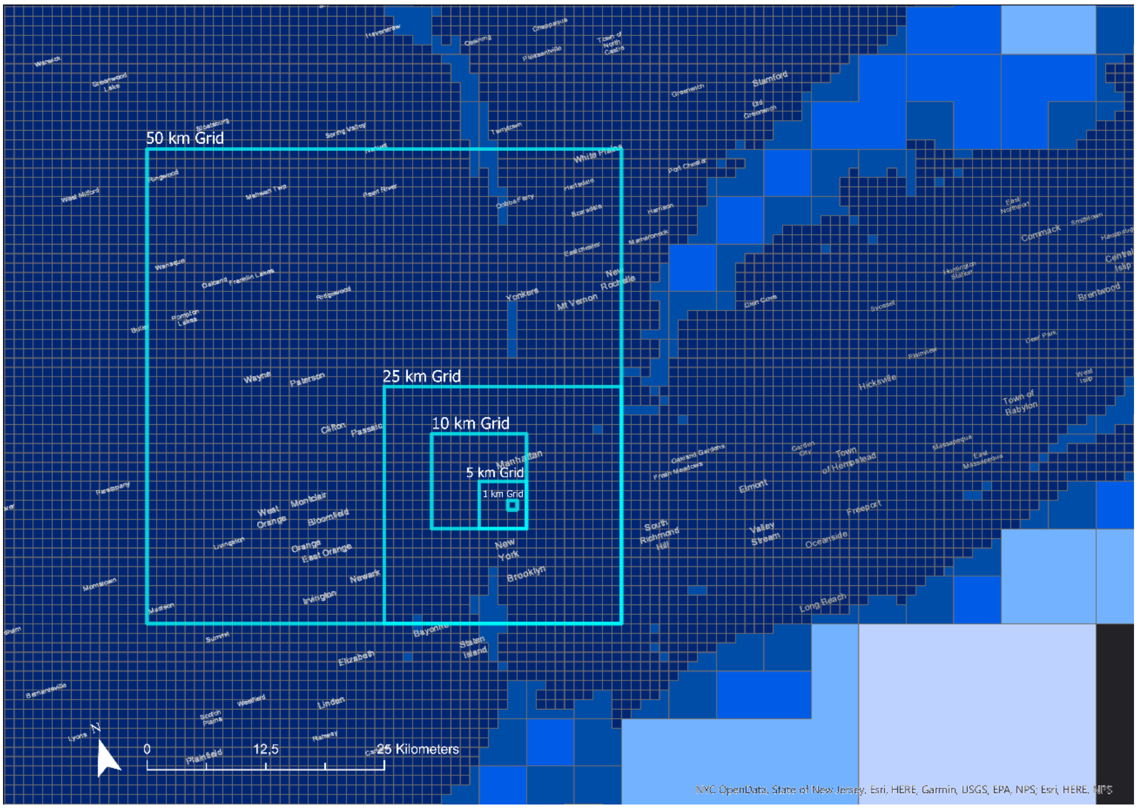
5 Results

This chapter presents the results of the analyses without further interpretation. This will be done in the chapter 6.

5.1 The software industry in the U.S.

As mentioned in the previous chapter, a grid was applied covering the entire study area for ease of processing the data. With a total area of the Contiguous United States of about 8,080,464 km² and 108,431 software companies, the square analysis resulted in a side length for the grid of 12.21 km. Various distance and spatial correlation measures calculated at global and local scales are used in the work. For a comparison of pattern and structure of the company locations on different aggregation levels, uniform grids of 1, 5, 10, 25 as well as 50 km side length were generated for the U.S..

$$12.21\ km^2 = \sqrt{2 * \frac{8{,}080{,}464\ km^2}{108{,}431}}$$



NYC OpenData, State of New Jersey, Esri, HERE, Garmin, USGS, EPA, NPS, Esri, HERE, NPS

Figure 7: 1, 5, 10, 25, 50 km grids (New York City area)

Due to the highly varying population density in the U.S., a 5-km grid is used for the exploratory analyses and maps. This provides a differentiated picture of the business locations of software companies and thus provides a better understanding at the micro-geographic level.

5.1.1 Analysis of patterns

Figure 8 presents an overview map that displays the pattern and structure of the gridded distribution of the software industry in the United States. The map indicates that the populated coastal regions and areas close to the Great Lakes have a significant concentration of software companies, indicating that the pattern is highly correlated with the population distribution. In contrast, the sparsely populated regions of the Rocky Mountains and the Great Plains have a relatively low number of software companies, with some exceptions such as the areas around Denver, Phoenix, Dallas, and Atlanta.

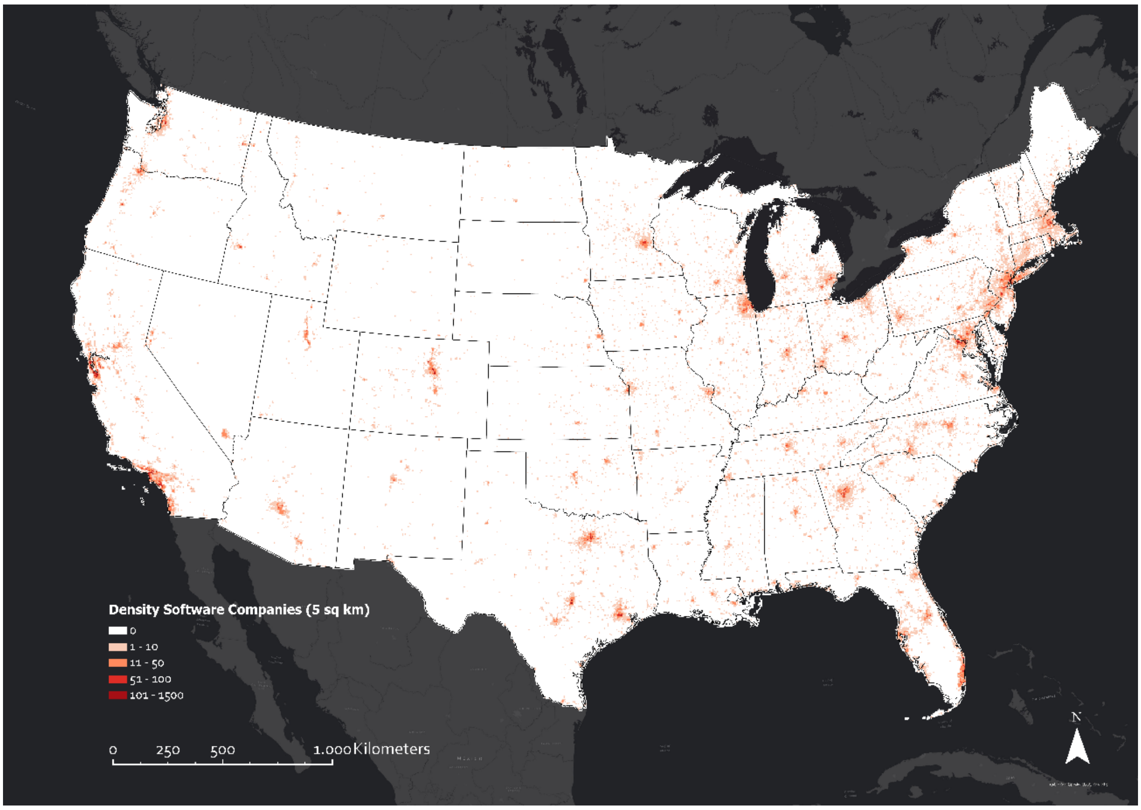

Density Software Companies (5 sq km)
0
1 - 10
11 - 50
51 - 100
101 - 1500
0 250 500 1.000Kilometers
N

Figure 8: Overview map - Pattern of software industry in the U.S. (5 km² grid)

Figure 9 shows an exemplary focus map of one of the most famous American cities, New York City, to get an impression of the spatial data precision. The pattern shows that software companies are largely located in the financially strong city center, with some exception around the center. Outside the center, the number of software companies per km² tends to be in the low single digits.

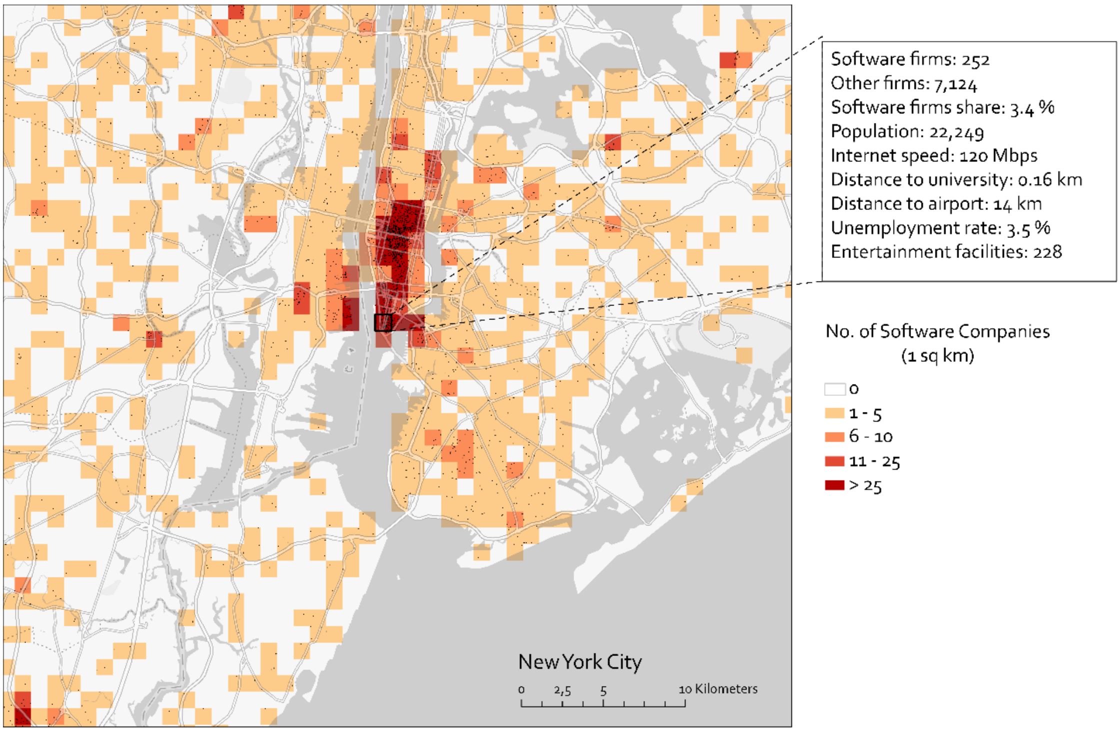


Figure 9: New York City on 1 km scale with selection of location factors for exemplary grid cell

The measure of spatial autocorrelation of the variable's software companies, other companies and population at different levels of aggregation is shown in Figure 10. The analysis shows that the highest autocorrelation of software companies with a Moran's I of 0.407 exists at a 10-km grid. A somewhat stronger autocorrelation is found for other companies at the 1 km to 10 km aggregation levels. This drops to the level of software companies at higher aggregation levels. On top of that, the population shows a very high spatial autocorrelation, with a Moran's I above 0.8 at a 5 km grid. As the level of aggregation increases, the autocorrelation decreases for all three variables and is similarly low for 50 km grid. The Moran's I analysis is highly significant at all scales (p-value ≤ 0.001) and has throughout a positive Z value. Further details of the analysis can be found in Table 13 in the appendix The probability and (>2.58 or <–2.58) standard deviations imply that the null hypothesis on the randomness of the spatial distribution can be rejected with a less than 1% likelihood that this clustered pattern could be the result of random chance. The spatial distribution of software company locations in the dataset shows a significant larger spatial clustering than expected.

The maximum autocorrelation is where the cells are maximally dependent on each other. For population, this is at 5 km. As the aggregation level increases, the spatial autocorrelation decreases. This is because larger grid cells have more averaged values, resulting in a decrease in deviation from the mean value. The population density depends very strongly on the grid size, whereas this is less the case for both software companies and other companies' density.

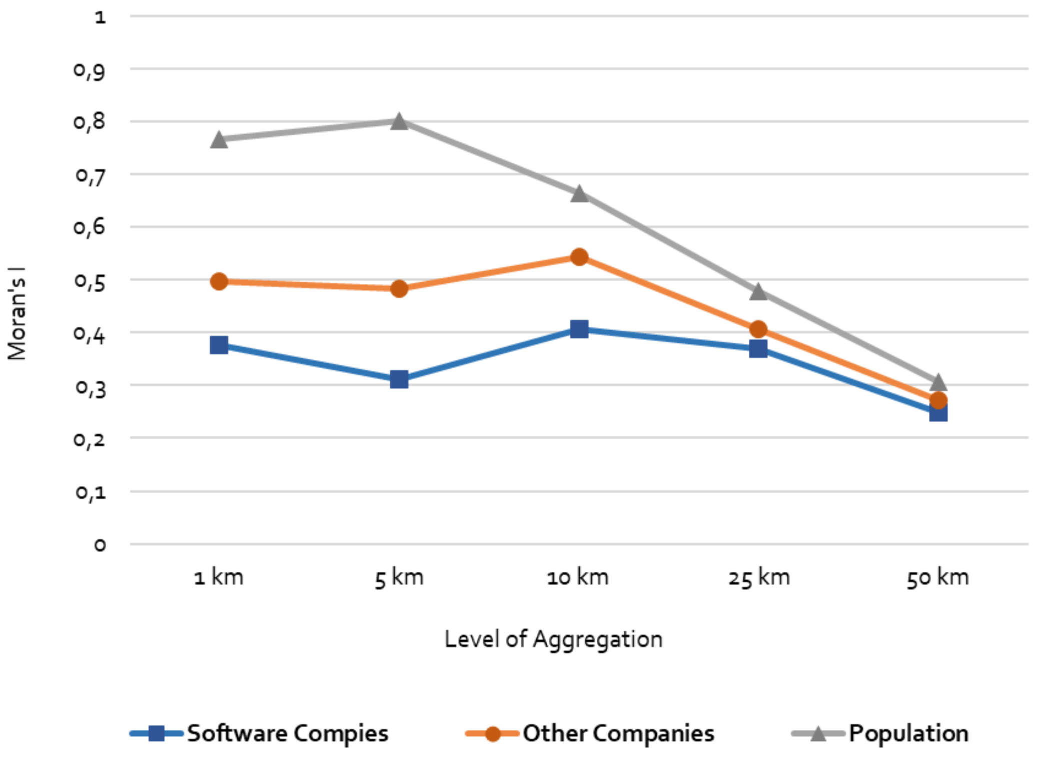


Figure 10: Moran's I of software companies, other companies and population – U.S.

5.1.2 Analysis of clusters

Having proved in the previous section that it is not a random distribution of software companies but a clustering, this section addresses the local spatial autocorrelation analyses, applying spatial grouping techniques, to determine whether the software companies cluster around certain areas. In order to carry out the cluster and outlier analysis on the software companies in the U.S. the queen contiguity is used again and a row standardization applied.

At first glance, the cluster map in Figure 11 appears to be shaded in the same locations as the choropleth map before. The cluster outlier analysis identifies a large number of statistically significant clusters ($p \leq 0.05$) with high numbers of company locations of the software industry in the U.S.. The presence of high-high clusters (HH) is most noticeable in coastal regions and on the shores of the Great Lakes. Some HH-clusters were also identified in the hinterland of the U.S.. There are 89,067 out of 108,431 software companies in this category (82 %), with an average of eleven companies per 5 km grid. No low-low cluster (LL) were identified.

If negative values of the local Moran coefficient occur in the presence of positive global spatial autocorrelation, spatial non-stationarities are present in the form of outliers ("pockets of non-stationarity"). Many single high-low outliers (HL) are spread across the country, somewhat stronger in the eastern half of the U.S.. Around 4 % of software companies belong to the HL-outlier category, with 1.3 software companies per 5 km grid cell. Meanwhile, outliers in the low-high (LH) category are largely found in the outskirts of metropolitan areas, a sort of border between urban areas with a high number of software companies and rural areas with few or none.

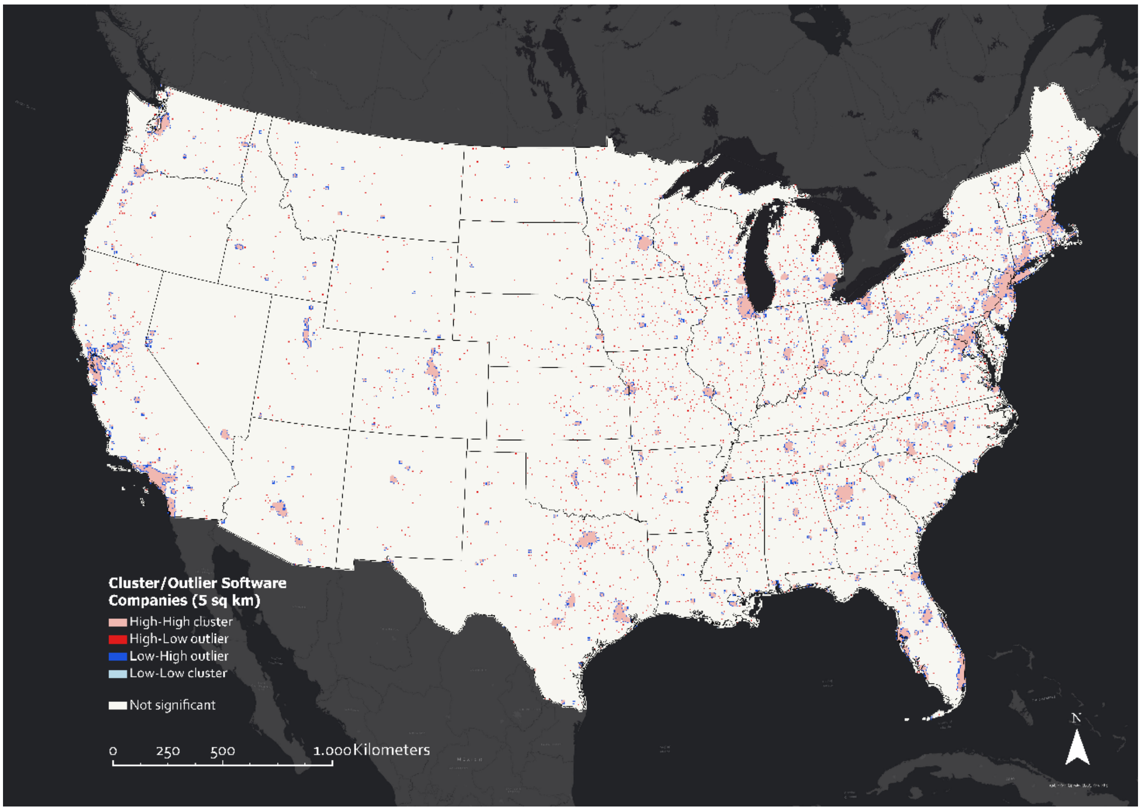


Figure 11: Cluster and Outlier - Software Companies in the U.S. (5 km grid)

A closer look on the city bands San Diego-San Francisco (SanSan) as well as Boston-Washington (BosWash) reveals vast HH-clusters, parallel along the coast, bordered by LH-outliers. HH-clusters are all located in urban areas. HL-outliers are occasionally found in more rural areas, but mostly with a certain proximity to urban areas, which can be observed particularly well in Figure 12.

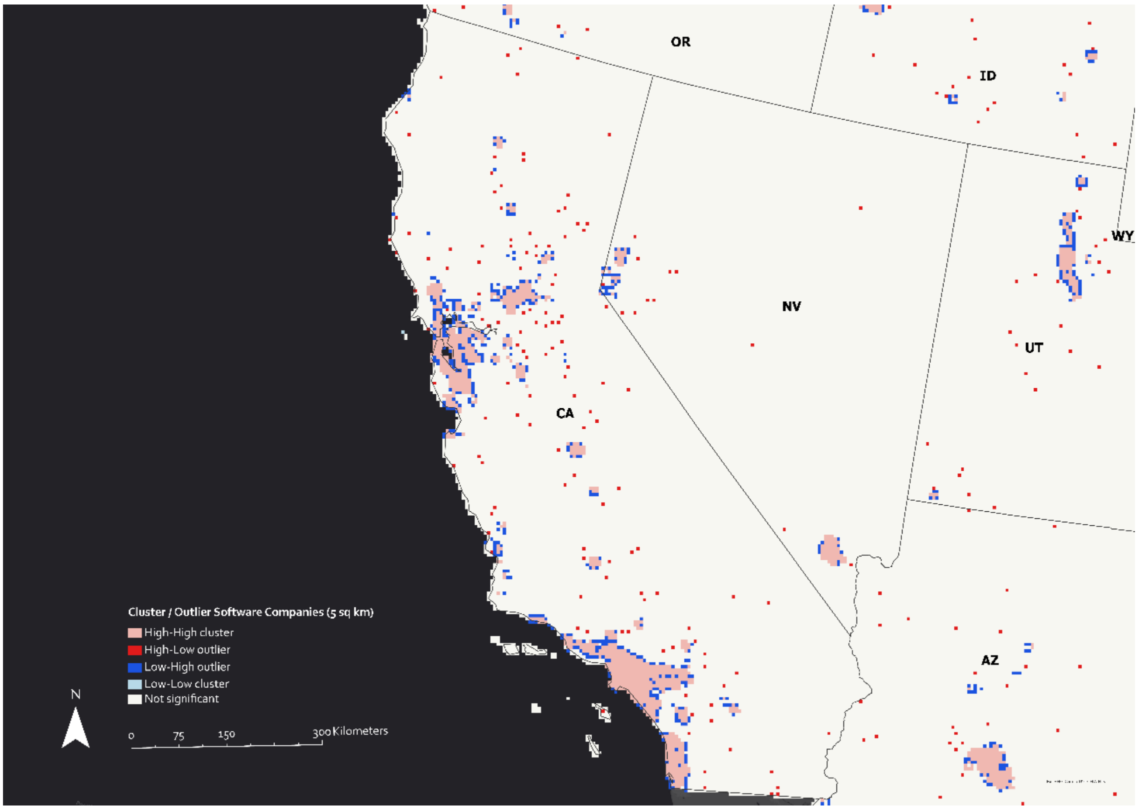


Figure 12: Cluster and Outlier - Software Companies SanSan (5 km grid)

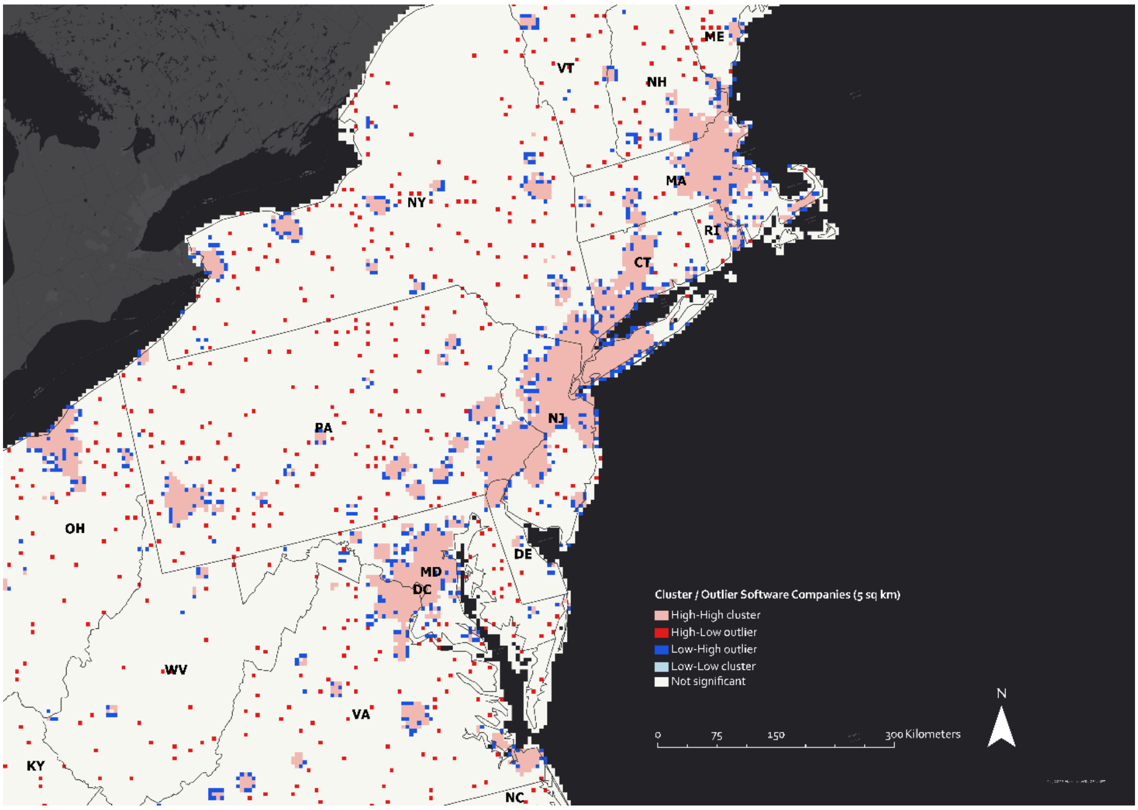


Figure 13: Cluster and Outlier - Software Companies BosWash (5 km grid)

Whether the local clusters are hot spots or cold spots cannot be determined from the local Moran coefficient (cluster / outlier analysis), since in both cases there is regionally limited spatial autocorrelation. The Ge-tis-Ord Gi* analysis was used to verify and complement the cluster and outlier analysis.

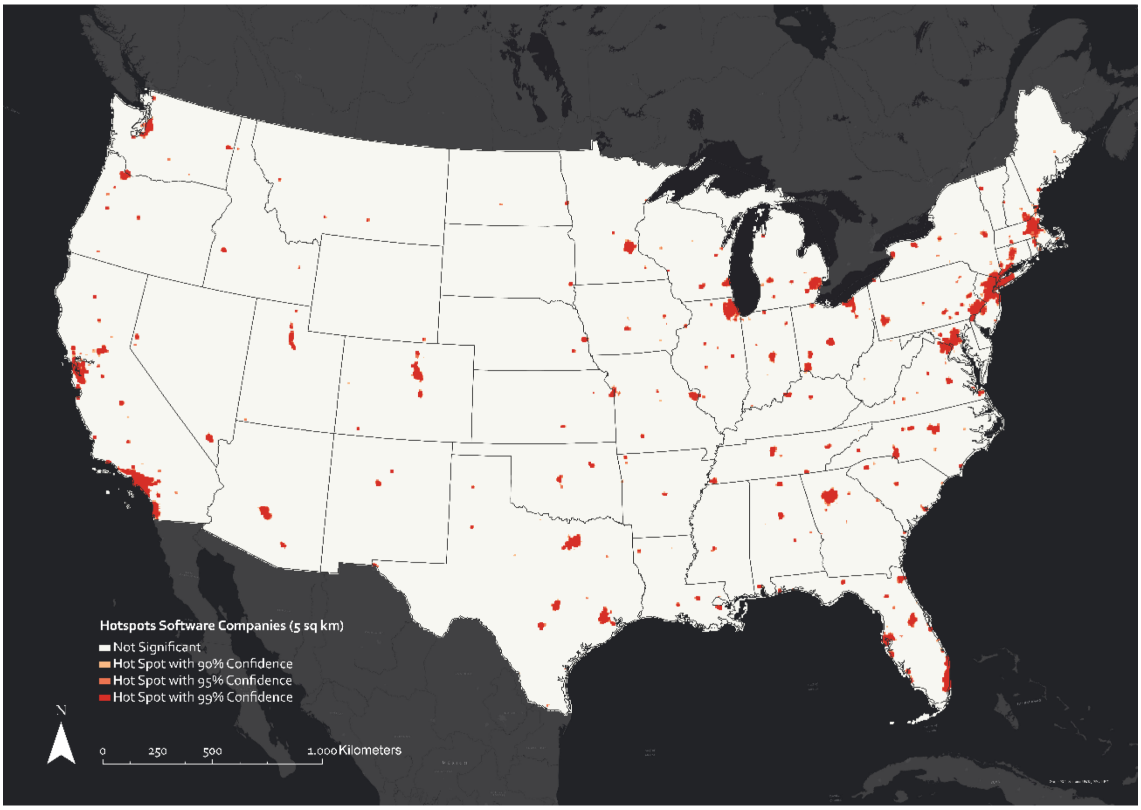

Hotspots Software Companies (5 sq km)
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
N
0
250
500
1.000 Kilometers

Figure 14: Hotspot-Analysis - Software Companies in the U.S. (5 km grid)

Figure 14 shows that there are significant overlaps with the regions identified in the cluster and outlier analysis. The high Moran coefficients of the clusters are significant hotspots with an above-average Z-value of 7.0, at a confidence level of at least 90 %. There are 84,726 software companies (78 %) in the hotspot's category, similar to the HH-clusters of the cluster and outlier analysis, with an average of about 12 companies per 5 km² grid. Considering only hotspots with a confidence level of 95 %, the Z-value increases to 7.6 and the average number of software companies to 13. The high Z-value represents a dense clustering with high values of software companies. It seems there is a high concentration of software companies particularly in the metropolitan center areas, as demonstrated by the city of Boston in Figure 15. However, it is also worthy to note that the areas that are not significant cover a vast area in the United States.

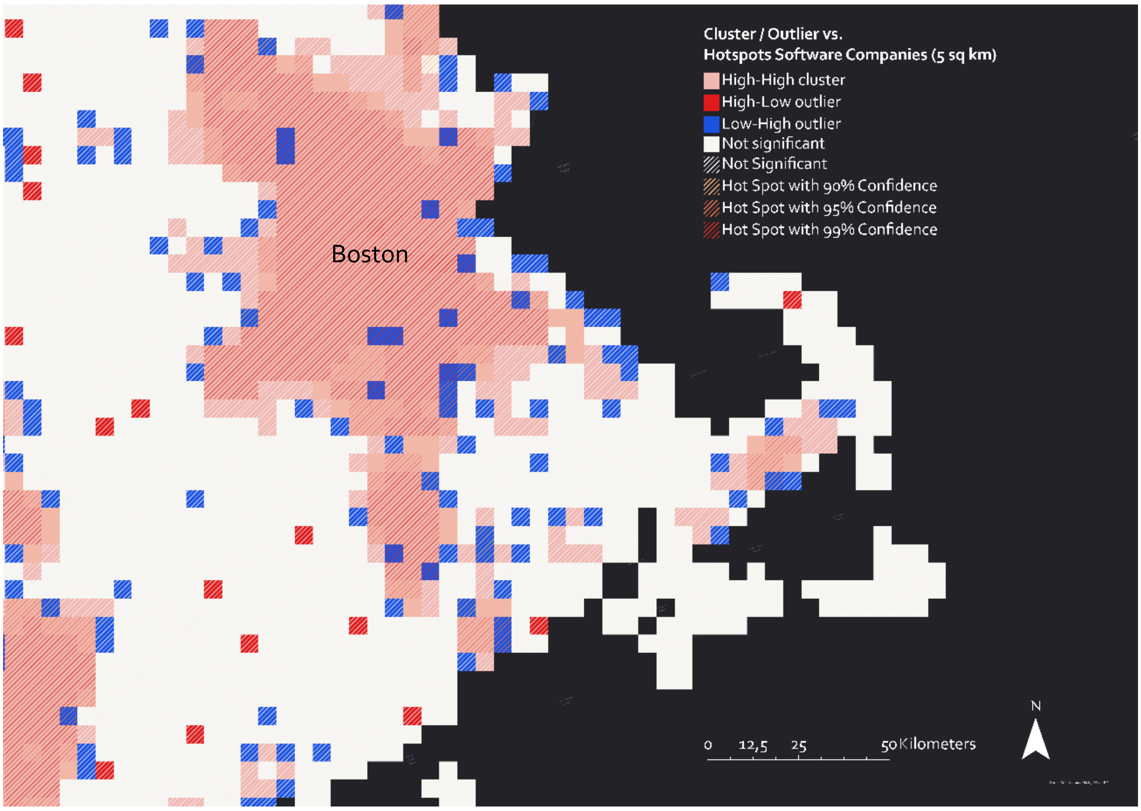


Figure 15: Cluster vs. Hotspot Analysis - Boston City (5 km grid)

Both geostatistical techniques were used exploratory to deepen one's understanding of the dataset at hand, but do not answer the research questions. It would certainly be interesting to prove homogeneous as well as heterogeneous clusters of software companies, which are not only analyzed on their spatial proximity but also on other criteria such as founding year, company size, etc. But this would go beyond the scope of the thesis.

5.1.3 Analysis of relations[3]

Prior to conducting the negative binomial regression, a bivariate non-parametric Spearman's rank correlation analysis was carried out to assess the strength of the monotonic relationship between the variables. The correlation matrix in Figure 16 shows which two characteristics have a reciprocal relationship as well as the direction of the

[3] Grid cells with missing values were removed from the dataset before being analyzed. This represents about 2.3% of the total area. It does not affect the result of the analysis, but was performed in the interest of data quality.

relationship. The five strongest positive vs. negative correlation can be taken from the following table (Table 5).

Of great interest are the extremely strong correlations between Population density and Population square (r = 0.99), Company density and Company square (r = 0.99) and Software Company and Software share (r = 0.99). There is also a strong positive correlation between universities and research institutes in the U.S. (r = 0.72). Strong negative correlations were found for Research Institutes and Street Centrality (r = -0.50) as well as Universities and Street Centrality. Furthermore, there is a rather strong negative correlation between the variables Population and Universities (Population density | Population square = - 0.42).

Positive correlation	**Negative correlation**
Pop_sq – Pop_density: 0.99	Research_inst – Street_cenrality: -0.53
Software_share – Software_comp: 0.99	Universities – Street_centraliy: -0.53
Company_sq – Company_density: 0.99	Universities – Pop_sq: -0.42
Universities – Research_inst: 0.72	Universities – Pop_density: -0.42
Research_inst – Airport: 0.59	Network – Research_inst: -0.41

Table 5: Strongest positive and negative correlation - U.S.

Besides the high positive correlation of software companies with the share of software companies in all companies (r = 0.99), there are further positive correlations with Company density (r = 0.25) (square (sq), r = 0.25), Public transport (r = 0.30), Network coverage (r = 0.18) and the amenities location factors; Entertainment (r = 0.34), Recreation (r = 0.29) and Culture (r = 0.19). Weak negative correlations exist with Airport (r = -0.11), Interstate (r = -0.10), Universities (r = -0.12) and Research institutes (r = -0.13). No correlations with software companies were found with Terrain, Business tax and all socio-demographic factors: Unemployment rate, Life expectancy, Average age, Student rate, Educated workforce, Migration rate and Salary.

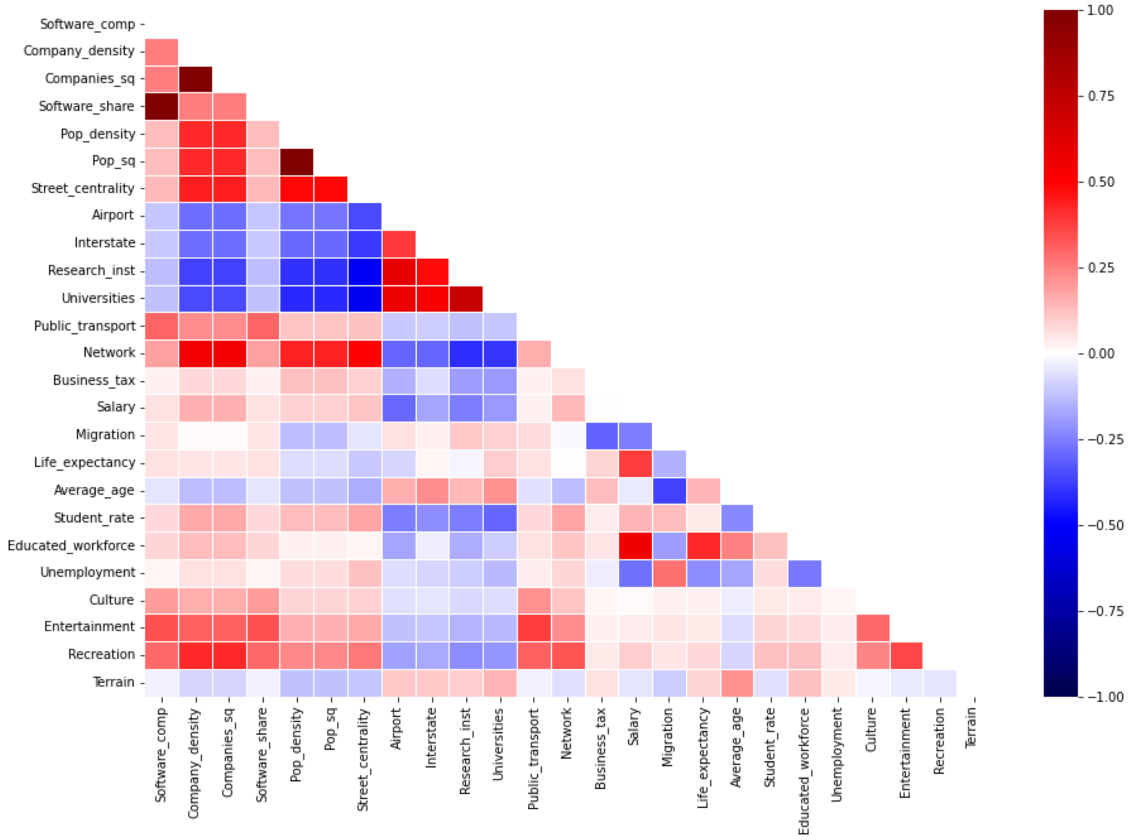


Figure 16: Correlation matrix - Location Factors U.S. (p ≤ 0.05)

A descriptive analysis of all variables (N, min, max, $\overline{x}$, σ) can be found in the appendix.

A comprehensive model was applied that correlates the number of software companies per 1 km² grid with the 24 different location factors. Due to the strong overdispersion of the data, a negative binomial regression (Maximum Likelihood Estimator (MLE)) was applied for the interpretation of the coefficients. The regression analysis looks at the significance of the various location factors on the location of software companies and helps - in an explorative approach - to understand why certain regions of the U.S. have a high number of software companies, while large regions have no software companies at all. Table 6 shows the estimated coefficients of the location factors. The regression coefficients are expressed as incidence-rate rations (IRR).

Providing that the result is significant, a value greater than 1 indicates that with increasing scores on the predictor, the incidence rate changes by a factor of the IRR. A value less than one indicates that with increasing scoring on the predictor, the incidence rate decreases by a factor of the IRR. Looking at the table below, a one-unit increase in (non-software) company density (equivalent to 10 companies) would expect an increase of local software firms by a factor of 1.188 (+18,8 % in a 1 km² grid), while holding all other variables in the model constant. A one-unit increase in distance to the next airport (1 km) is associated with a 0.993 (-0.7 % in a 1 km^2 grid) smaller number of local software firms. The Wald Chi-Square Test was applied testing whether the model containing the full set of predictors fits significantly better than the null model.

Location Factors	Description	IRR	SE
Agglomeration location factors			
Company density	Number of local companies (in 10).	1.188***	.0006
Company density^2	Squared number of local firms (in 10).	1.000***	.000
Software companies share	Proportion of software companies in the local business population (in %)	1.248***	.0002
Population density	Population per cell (in 100).	1.031***	.010
Population density^2	Squared number of inhabitants per cell (in 100).	1.000***	.000
Street centrality	Street (network) density calculation (1). High value = High density	1.006***	.000
Universities	Distance to the nearest university (in km).	.980***	.0009
Research institutes	Distance to the nearest research institute (in km).	.928***	.002
Infrastructure location factors			
Network coverage broadband Internet	Average latency (upload / download speed) (in Mbps). High value = high internet speed	1.001***	.0001
Interstate / Highway	Distance to nearest high way / interstate (in km).	.995***	.0003
Airport	Distance to nearest main civil airport (in km).	.993***	.0004
Public transport	Weighted count of public transport stops.	.956***	.0014
Socio-economic location factors			
Salary	Monthly household income (median) (in 100 EUR).	1.000***	.000
Educated workforce	Proportion of employees with a university degree (in %).	1.022***	.0005
Student rate	Proportion of students in the local population in %.	.993***	.0008
Business tax	Corporate income tax rates fixed by the states (in %). High values = high rates	1.013***	.0017
Life expectancy	Average life expectancy of the population (in years).	1.040***	.0024
Average age	Average age (median) of the population.	.987***	.0010
Unemployment rate	Proportion of unemployed in the working-age population (in %).	1.007***	.0017
Migration background	Proportion of people of non-German nationality in the total population (in %).	1.003***	.0003
Amenities location factors			

Recreation	Number of recreational, community and sports facilities.	1.012***	.0004
Culture	Number of cultural sites and facilities.	1.004	.671
Entertainment	Number of dining, nightlife and general entertainment facilities.	.972***	.0012
Other			
Terrain	Average slope or gradient (in degree). High values = hillside location	.701***	.0123

*Table 6: Location factors and estimated coefficients (IRR) with robust standard errors (SE) – U.S. - */**/*** indicate significance at 10/5/1 per cent, respectively.*

5.1.3.1 Interpretation of Regression Coefficients

Following the work of Kinne & Resch (2018) the square was included in the analysis for the location factors company density and population density. The underlying approach is that urbanization leads to congestion effects. A concentration of economic activity, such as an increasing number of companies or population, is initially positive for the attractiveness of a location, but after exceeding a certain threshold it turns negative due to factors such as increasing pollution or living costs. Thus, density has an inverted u-shaped influence on the location economy (Rocha, 2008; Liviano & Arauzo-Carod, 2011). The settlement of both companies and population in the United States significantly and positively impact the location of local software companies. Although the coefficients of their squared values do not explicitly indicate a negative correlation, they do suggest the presence of agglomeration shadows. However, the explanatory power of the squared urbanization variables as evidence of overload effects is controversial according to the World Bank (2009). The street density shows a significant weak positive effect on the location of local software companies. It is estimated that a high proportion of software companies among all local companies significantly increases the number of additional local software companies. Clusters of software companies encourage company formation of further software companies in the same place. For knowledge-intensive sectors such as the software industry, proximity to places of knowledge is equally essential. There is a significant reduction in the number of local software companies associated with an increase in the distance to research institutes and universities. The same effect is also observed for highways and airports. Whereas, good accessibility of public transportation is not linked to an increase of local software companies. Availability of high-speed Internet (mbps) has little effect on the entry of new local software companies. Locations with good availability and accessibility of infrastructure as well as knowledge cluster seem to be preferred for software companies in the U.S..

While proximity to universities has a significantly positive impact on the number of local software companies, a high percentage of students in the local population appears to have a converse effect. Well-educated workers with academic degrees in the local labor force raises the probability for the settlement of software companies. Although workers with university degrees presumably earn higher incomes than non-academics (see correlation matrix), there seems to be no effect on the entry of new software firms. Higher state corporate income taxes show a significantly positive effect and seem rather implausible at first sight. A high life expectancy of the local population is associated with a high number of software companies. Conversely, an increase in the average age appears to have a negative impact on the number of local software companies. Moreover, a higher proportion of individuals with a migration background in the local population likely results in an increase in the number of additional local software companies. As for the relationship between the unemployment rate in the local population and the number of local software companies, the current evidence suggests that an increase in the former leads to an increase in the latter, although this appears to be an implausible relationship at this time.

When looking at the coefficients of the amenities, an increase in recreational facilities raises the probability of software companies by 1.2 %, whereas entertainment tends to have a negative effect on the number of local software companies. Cultural facilities also show a positive effect, although this is not significant. It is important to note that while these amenities may have a positive impact at lower levels of aggregation, such as the city or county level, they do not necessarily enhance the attractiveness of the immediate neighborhood. The presence of locations with a slope has a significantly strong negative impact on the settlement of software companies.

5.1.3.2 Model comparison

Assessing model fit can be difficult. There are a number of indices in the literature that help to evaluate how well the model represents the data. The various measures of Goodness of Fit (GoF) as well as measures of exploratory power were considered as suitable for this study. It helps to evaluate the fit and accuracy of the various models applied and allows to draw conclusions.

Table 7 presents the results of the GoF as well as prediction accuracy of the different models. The Poisson model corresponds to P = 0 (alpha parameter), while the linear negative binomial is P = 1 (NB1) and the quadratic negative binomial regression is P = 2 (NB2). The negative binomial (MLE) where the coefficient of the regression model is estimated is P = 3.4.

Pseudo-R^2 can be understood in analogy to the R^2 of linear regressions and indicates which model better represents reality and can only be compared to the Pseudo-R^2 values of another model. A higher Pseudo-R^2 value corresponds to a better fit of the model to the given data; Pseudo-R^2 values of 0.2 to 0.4 can already be taken as indicators of a very good explanatory power. The most common formula for calculating this indicator was used; McFadden's pseudo-R2 (IBM, 2022). The Root Mean Square Error (RMSE) is also a measure for model evaluation and judges the quality of predictions. The RMSE explicitly shows how much our predictions deviate on average from the actual values of the data set. A smaller value indicates better model (Coxe *et al.*, 2009). The goodness of fit of a statistical model describes how well an estimated model can explain a set of observations. The log-likelihood is used to evaluate the model quality, the probability that given data are consistent with a particular given model. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are commonly used measures to compare different models, with lower values indicating better model selection (Coxe *et al.*, 2009). The overall consideration of the statistical measures allows a profound decision to be made as to which model is most appropriate.

Looking at the GoF indicators (log likelihood, BIC, AIC), the NB MEL model clearly shows the lowest values and thus the best model fit according to these three indices. The AIC should not be taken as an absolute measure of quality. Even the model that is shown to be the best by the AIC may have a very poor fit to the data. The same is valid for the BIC. The Poisson model, on the other hand, has the highest values and thus the most inferior fit, according to the AIC and BIC. The Pseudo-R^2 returns relatively similar values among all models and demonstrates good explanatory power. The values for RMSE differ greatly between the models. The Poisson model has by far the lowest root mean square error and thus represents the best prediction quality, whereas, all NB models show very strong deviations in the predicted values.

The Omnibus Tests of all models were highly significant ($p < 0.001$). The results indicate that all models containing the full set of predictors are a significant improvement over a null model.

Measure	**Poisson**	**NB1**	**NB2**	**NB(MLE)**
Pseudo-R^2	0.584	0.637	0.633	0.621
RMSE	0.451	4.61E+14	1.88E+20	6.15E+15

Log Likelihood	-272,032	-207,939	-200,207	-197,958
AIC	544,114	415,927	400,464	395,968
BIC	544,461	416,275	400,464	396,329

Table 7: Goodness of fit (GoF) – U.S.[4]

The observed values and predicted values for each model are illustrated in the graph below (Figure 17). The extreme number of zero values is neither overestimated nor underestimated by the models. For all models, underestimation of predictions occurs for cells with a low number of software companies. In the case of the Poisson Model, there is an underestimation of the prediction almost throughout. For the NB models, the initial underestimation of low count cells turns into a very drastic overestimation of the high count cells, reflecting the extreme RMSE values. Based on the descriptive statistics in Table 7, the Poisson model appears to be the better predictive model at the 1 km^2 scale.

[4] Method: Fisher, Scale parameter measure: Deviance, Chi-squared statistic: Wald, Estimator: Robust

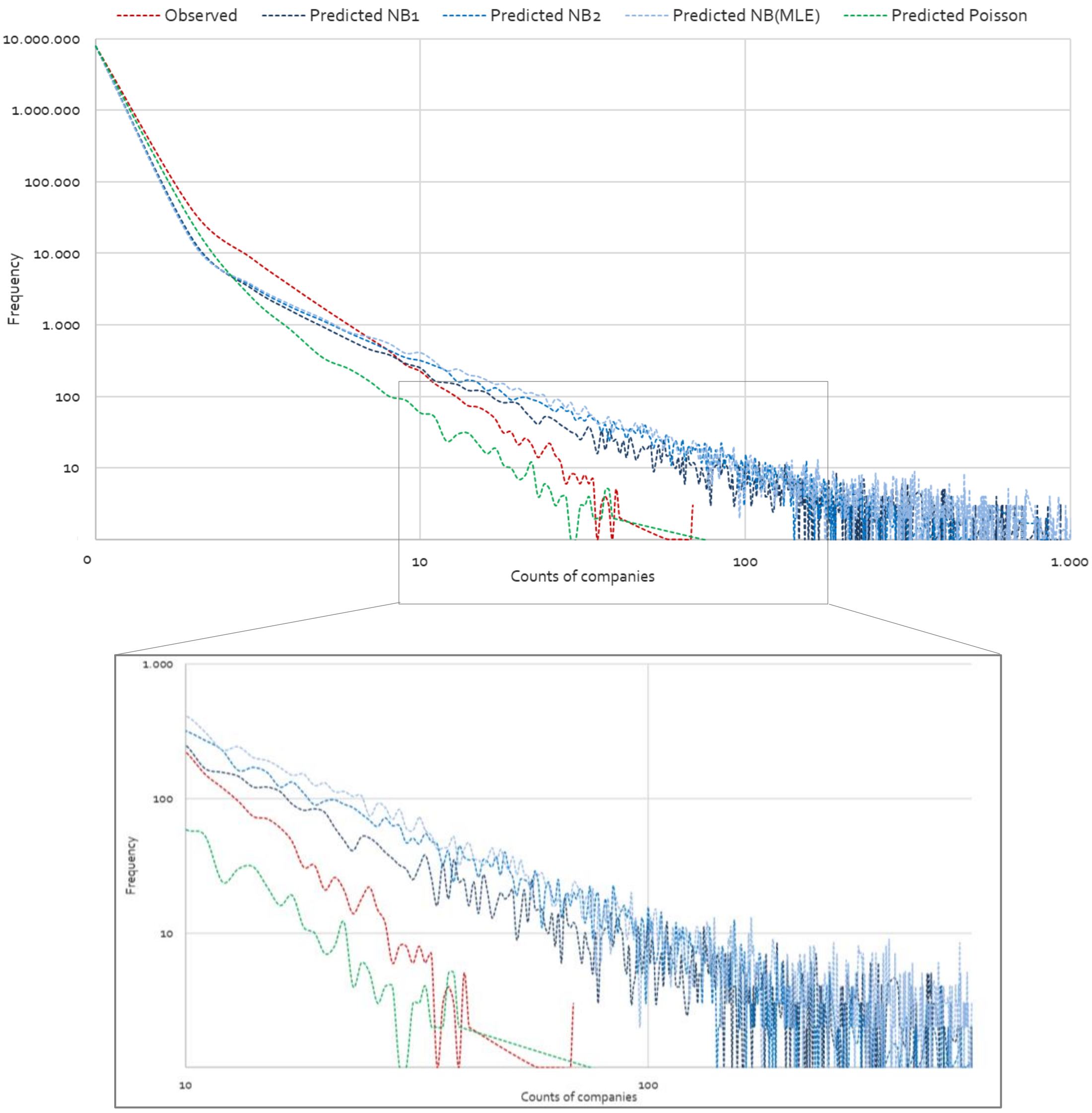


Figure 17: Frequencies of observed and predicted software firm counts - U.S.

Figure 18 shows the regression residuals - the prediction errors; deviation of the observed values from their expected values - aggregated on a 5 km grid. The grid cells with a higher number of software firms than predicted by the model are shown in warm colors. The model underestimated those cells. Grid cells with cold colors indicate overestimation of software counts by the model.

The residuals of the models with the best model fit - Poisson and NB MEL - were illustrated in maps. The upper map shows the vast overestimation of the NB MEL model. Software company counts are largely overestimated, with some underestimates primarily in metropolitan centers such as Minneapolis or Atlanta, as well as in smaller cities such as Pittsburgh, Buffalo or Austin.

The Poisson Model, on the other hand, shows primarily underestimations. The number of software companies in the city centers is uniformly underestimated, while overestimations tend to be found on the outskirts. This pattern can be found in most cities. For both models, it can be stated that the prediction errors - overestimation or underestimation - occur almost only in urban areas in the U.S..

Due to the aggregation, the picture that emerges is that large-scale forecast errors exist, especially for the NB MEL model. The detailed maps in the original 1 km^2 grid shall provide further insights (Figure 19). The San Francisco Bay Area, as part of the Sansan city band and home of the Silicon Valley - the most important high-tech hub in the world - is shown. In addition, New York City was chosen, as part of the BosWash city band and the city with the highest population and business density in the U.S..

When looking at the residual pattern, it suggests that the prediction of both models systematically fails in some areas. The detailed maps suggest that a positive spatial autocorrelation is present and may be due to one or more omitted explanatory variables, with a high collinearity. Likewise, software companies themselves could be a significant location factor. Perhaps software companies gather where other software companies are and this develops a self-reinforcing interaction process of firm location accumulation. This is supported by the strong bivariate correlation coefficient between these two variables (see Figure 16). For the Poisson model, Moran's I is 0.087 and for the NB MEL model it is 0.151, which is not very large for both models, but still above zero ($p \leq 0.001$).

Another possible explanation for the systematic prediction error could be the urban structure prevalent in North America. The Central Business District consists almost exclusively of business and office quarters directly in the center. The area is usually not very large, but due to skyscrapers there is a very high density of businesses. Residential space, on the other hand, is not to be found in the Central Business Districts (Hahn, 2014). The high concentration in the city centers and the related traffic issues have led to the establishment of more industrial, commercial and high-tech parks along the main connecting roads in recent years, the so-called edge cities. The long history of Europe and thus of Germany has produced different urban structures. Both models show incorrect predictions of New York City's Central Business District, Lower Manhattan, as illustrated in the detailed map on the right side of Figure 19. This is also observed in other cities, as shown in Figure 18.

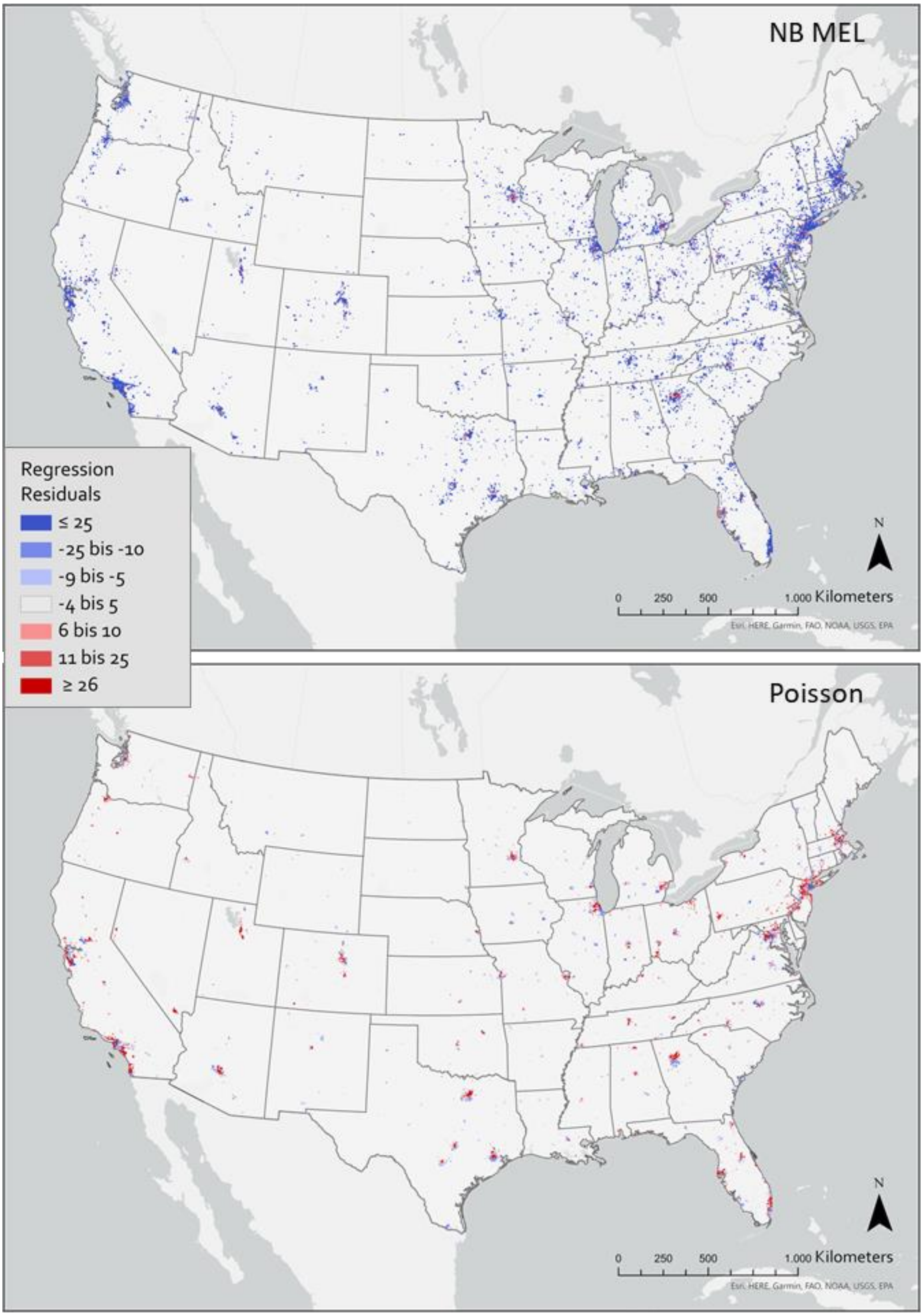

NB MEL
Regression Residuals
≤ 25
-25 bis -10
-9 bis -5
-4 bis 5
6 bis 10
11 bis 25
≥ 26
N
0 250 500 1.000 Kilometers
Esri, HERE, Garmin, FAO, NOAA, USGS, EPA
Poisson
N
0 250 500 1.000 Kilometers
Esri, HERE, Garmin, FAO, NOAA, USGS, EPA

Figure 18: Regression residuals aggregated at 5 km grid in U.S.

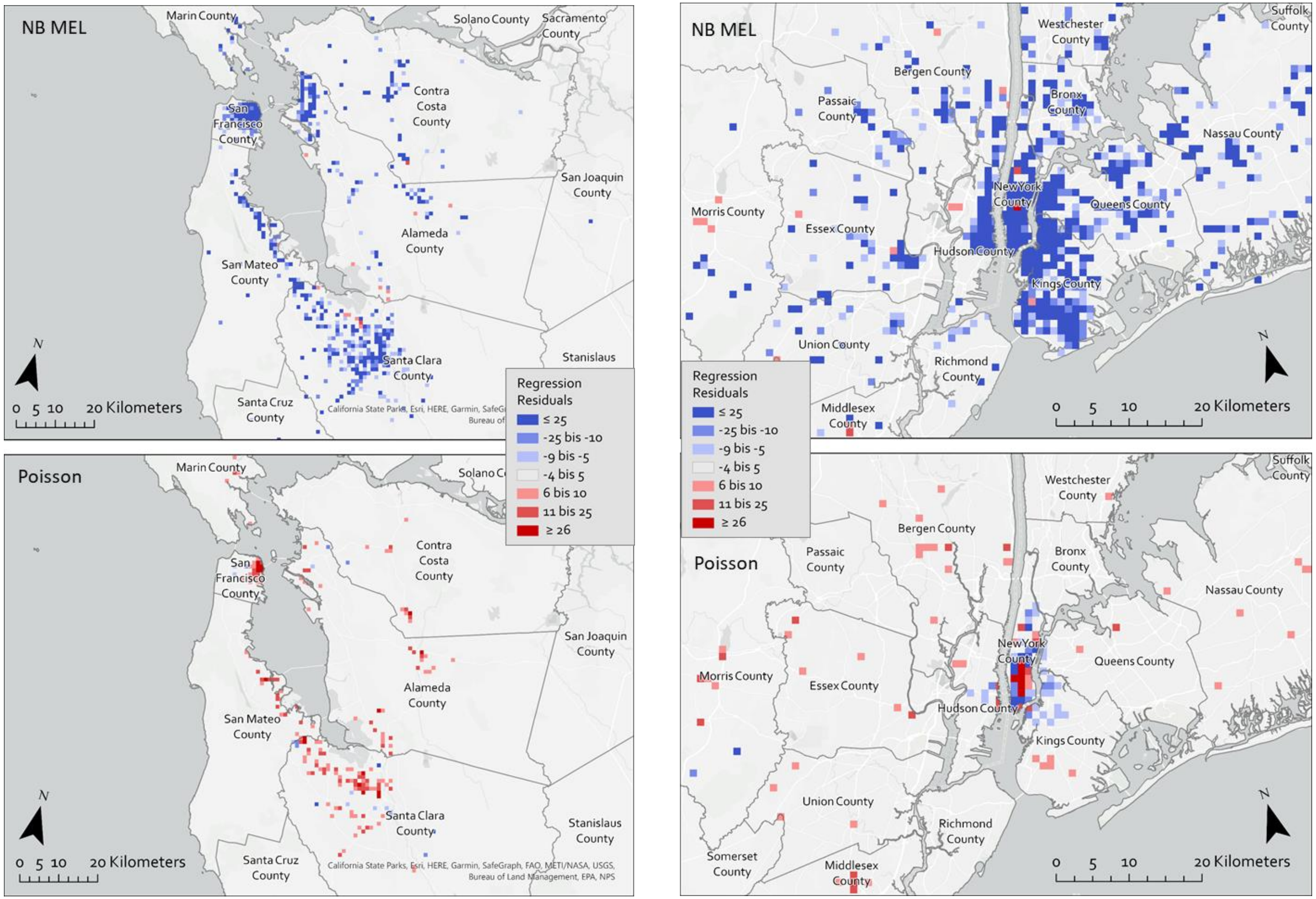


Figure 19: Regression residual original at 1 km grid of cities in the U.S. (San Francisco, New York City)

5.2 The software industry in Germany

For a better understanding of the software industry in Germany and the differences between the U.S. and Germany, a brief exploratory analysis is conducted before the regression model is applied and conclusions are drawn.

The square analysis for Germany with regard to software companies resulted in a grid of 3.75 km². Since Germany is several times smaller and has a higher population density than the U.S., uniform grids of 1, 3, 5, 10 and 25 km² were generated. Table 8 contains the descriptive statistics for the different scales. The comparison of the mean and median shows a high proportion of zero values. However, this proportion is significantly lower in Germany compared to the U.S. (Table 4). The variance exceeds the mean at all aggregation levels (DI: ratio of variance to mean of distribution), albeit significantly lower than in the U.S.. Nevertheless, an overdispersion of the data is also visible in Germany; the locations of software companies in Germany allow the conclusion that there is a clustering.

grid	obs.	null	max.	$\bar{x}$	$\tilde{x}$	σ	DI
1 km	361,482	94%	216	0.14	0	1.30	12.07
3 km	40,889	73%	577	1.24	0	8.50	58.27
5 km	14,930	53%	1,209	3.40	0	19.90	116.47
10 km	3,863	23%	1,797	13.10	3	60.20	276.64
25 km	672	7%	3,103	75.50	22	227.80	687.32

Table 8: Descriptive Data - Software Companies Germany

Figure 20 visualizes the exponential slope of the regression line. In Germany, too, the mean $\bar{x}$ of the software companies depends on the raster size. A polynomial relation is also apparent here.

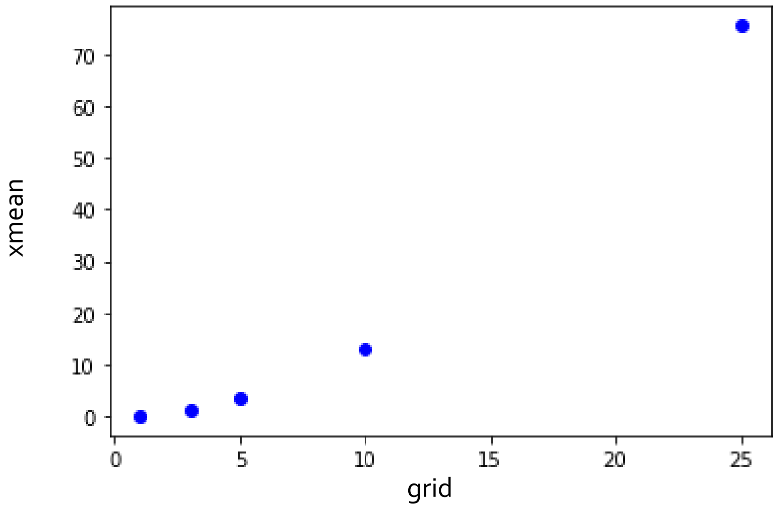


Figure 20: Linear regression: grid vs. xmean Germany

5.2.1 Analysis of patterns

Figure 21 shows the pattern and structure of the software industry in Germany in an overview map. Germany does not have large expansive areas without software company locations like in the U.S. (see Figure 8). Very high concentrations of software companies can be identified in the cities of Berlin, Hamburg and Munich in particular. Other large concentrations, although less dense, can be found in the metropolitan regions of Stuttgart, Frankfurt-Rhine-Main and Rhine-Ruhr. A north-east-south-west gradient can be observed in the settlement of software companies in Germany.

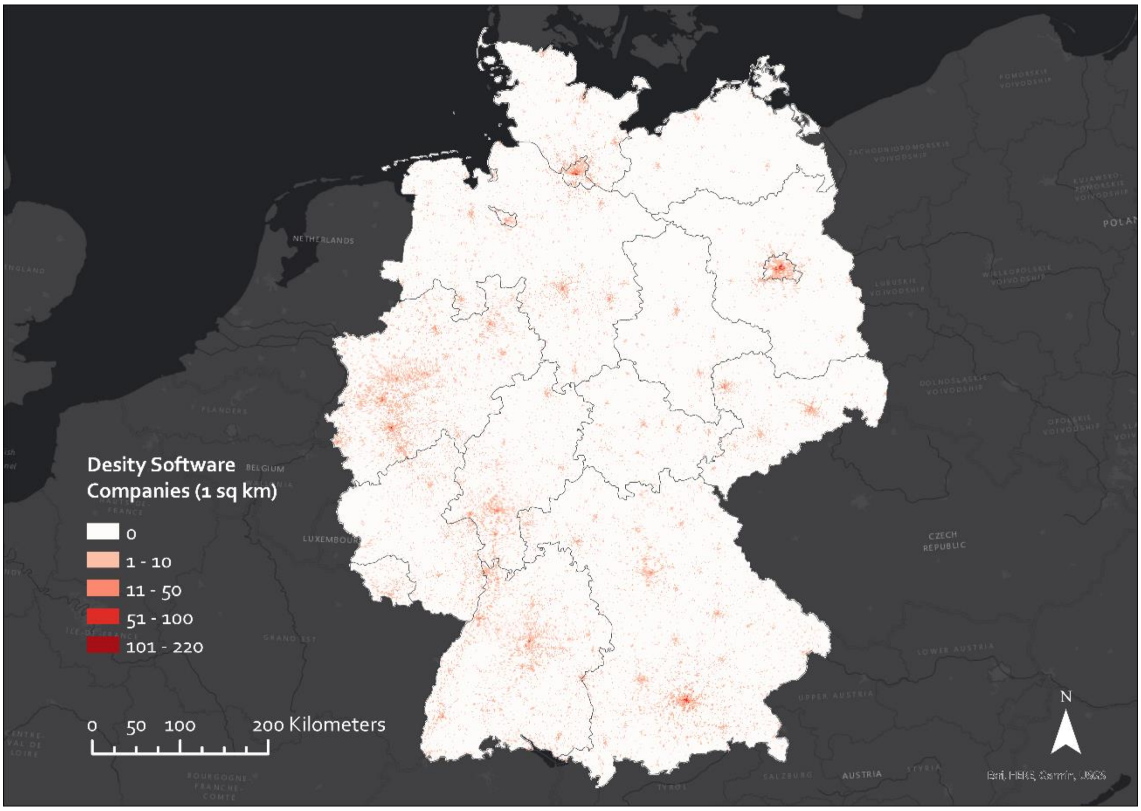


Figure 21: Overview map - Pattern of software industry in Germany (1 km grid)

High concentrations of software companies can be found largely in inner-city areas. The number of software companies gradually decreases as the distance from the center of the location increases. A few small conglomerations of software companies can also be found on the outskirts, as shown in the example of Berlin in Figure 22.

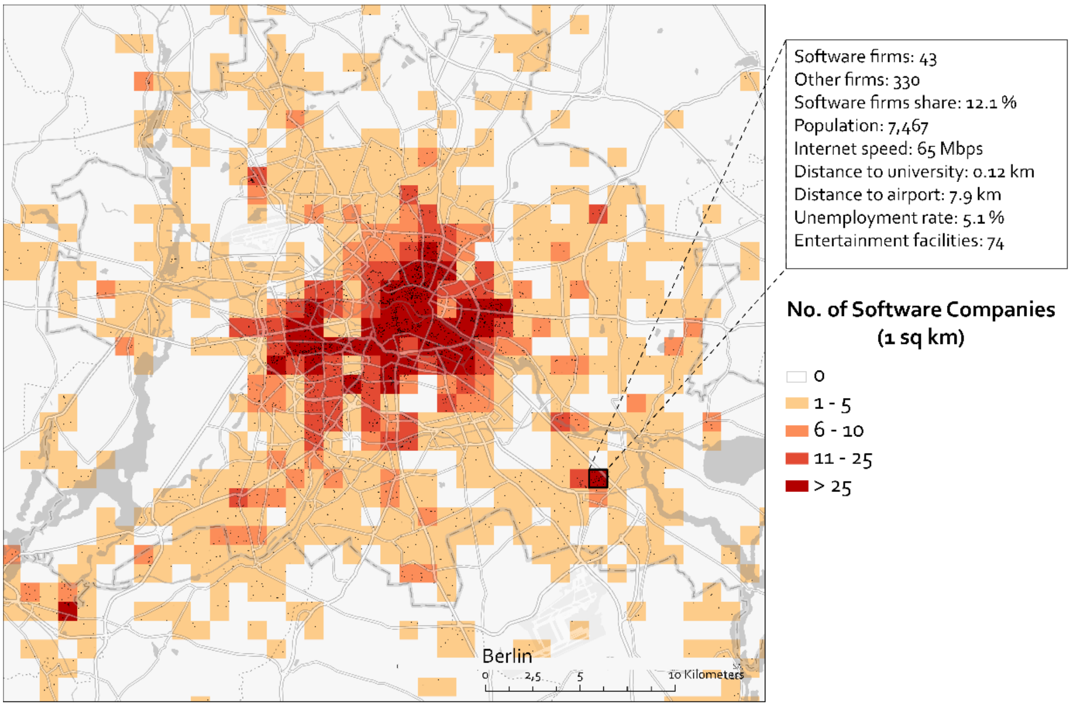

Software firms: 43
Other firms: 330
Software firms share: 12.1 %
Population: 7,467
Internet speed: 65 Mbps
Distance to university: 0.12 km
Distance to airport: 7.9 km
Unemployment rate: 5.1 %
Entertainment facilities: 74

Figure 22: Berlin on 1 km scale with selection of location factors for exemplary grid cell

The following graph shows the global Moran's Index for the variables software companies, other companies and population at different aggregation levels. The highest autocorrelation of software companies with a Moran's I of 0.537452 exists at a 1 km grid and drops comparatively steeply to 0.169253 at a 25 km grid. The situation is similar for the other companies, but the clustering is almost always somewhat higher than for the software companies. As in the U.S., the population is most clustered with a Moran's I of 0.710964 at the 1 km level of aggregation. Nevertheless, the population concentration is lower than in the United States.

The global Moran's I analysis shows highly significant (p-value ≤ 0.001) values at all aggregation levels for all three variables, as well as a positive Z-value. Details can be found in the appendix. Thus, random distribution can be discarded and the null hypothesis can be rejected. The spatial distribution of the software company locations shows a stronger spatial clustering in Germany than expected.

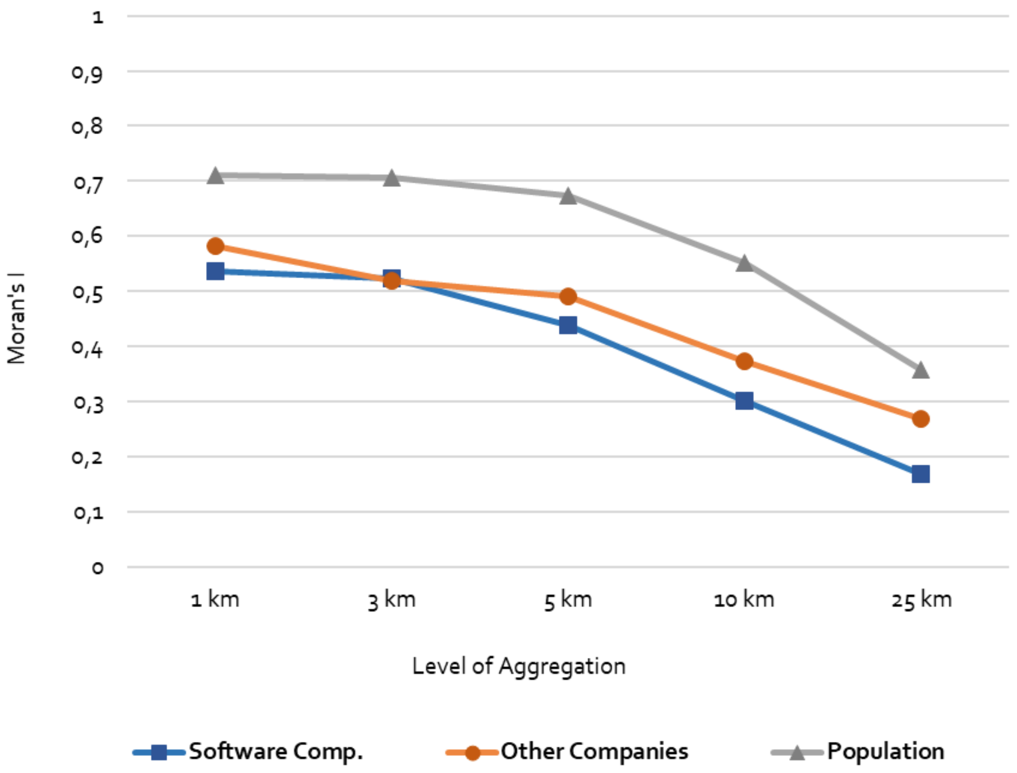


Figure 23: Moran's I of software companies, other companies and population – Germany

5.2.2 Analysis of clusters

Following the global Moran's I analysis, local spatial autocorrelation analyses are performed in this section. Berlin, Munich and Hamburg are identified as the largest high-high-clusters (HH). Smaller HH-clusters, but in higher numbers and with high spatial proximity, cover the area around Stuttgart, Frankfurt, and the Ruhr region. Other individual medium-sized clusters range from Kiel in the north across Germany to Augsburg in the south. Around 63 % of German software companies belong to this category, with an average of four companies per 1 km². Low-low clusters (LL) were not identified.

Looking at the outliers, low-high-outliers (HL) are found scattered across the country. Nevertheless, a spatial proximity towards HH-clusters can be identified. This category has on average 1.17 software companies per km² and contains about 12 % of all software companies. The low-high-outliers (LH) can be found mostly on the outskirts of metropolitan areas and represent the transfer from urban areas with high concentrations to rural areas with low concentrations of software companies.

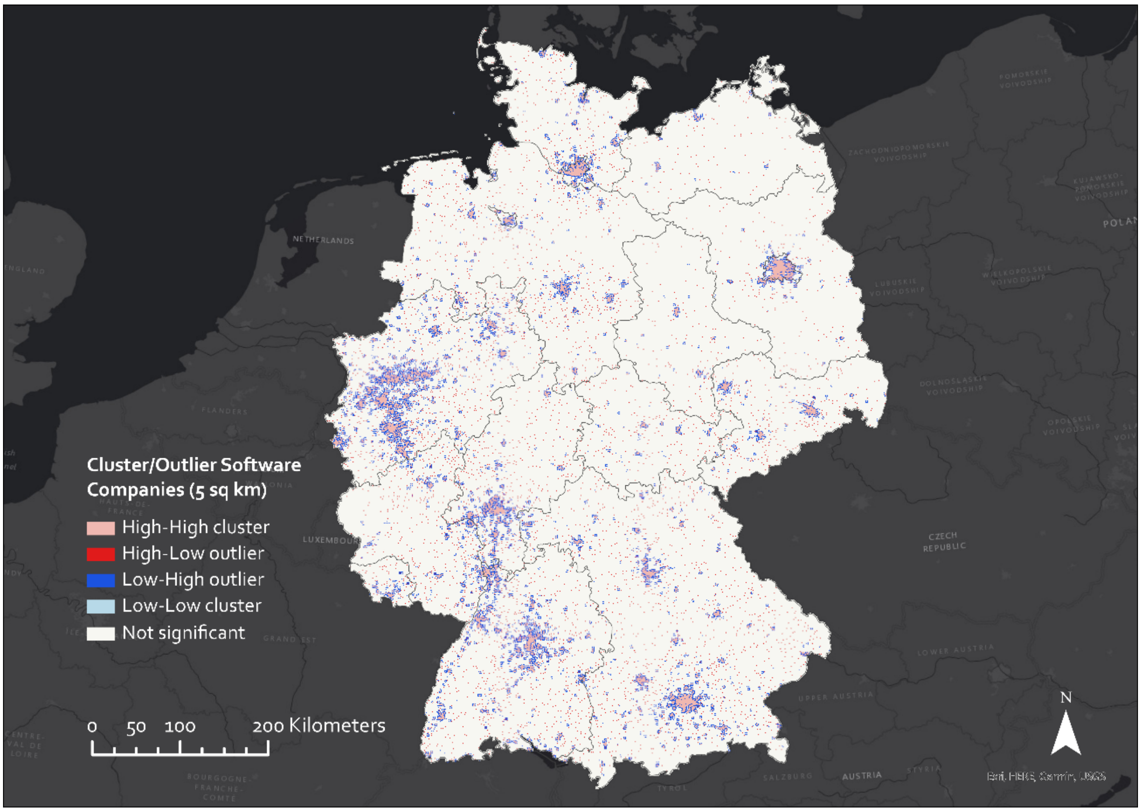


Figure 24: Cluster and Outlier - Software Companies in Germany. (1 km grid)

To complete the exploratory analysis, the Getis Ord Gi* is applied. Figure 25 shows a clear overlap of the HH-clusters from the previous analysis with the identified hotspots.

With a confidence level of 90%, the Z value is 7.0 and increases to 8.7 with a confidence level of 95%. The high Z-value reflects the high concentration of software companies, which can be found in particular in metropolitan areas, but also in medium-sized cities.

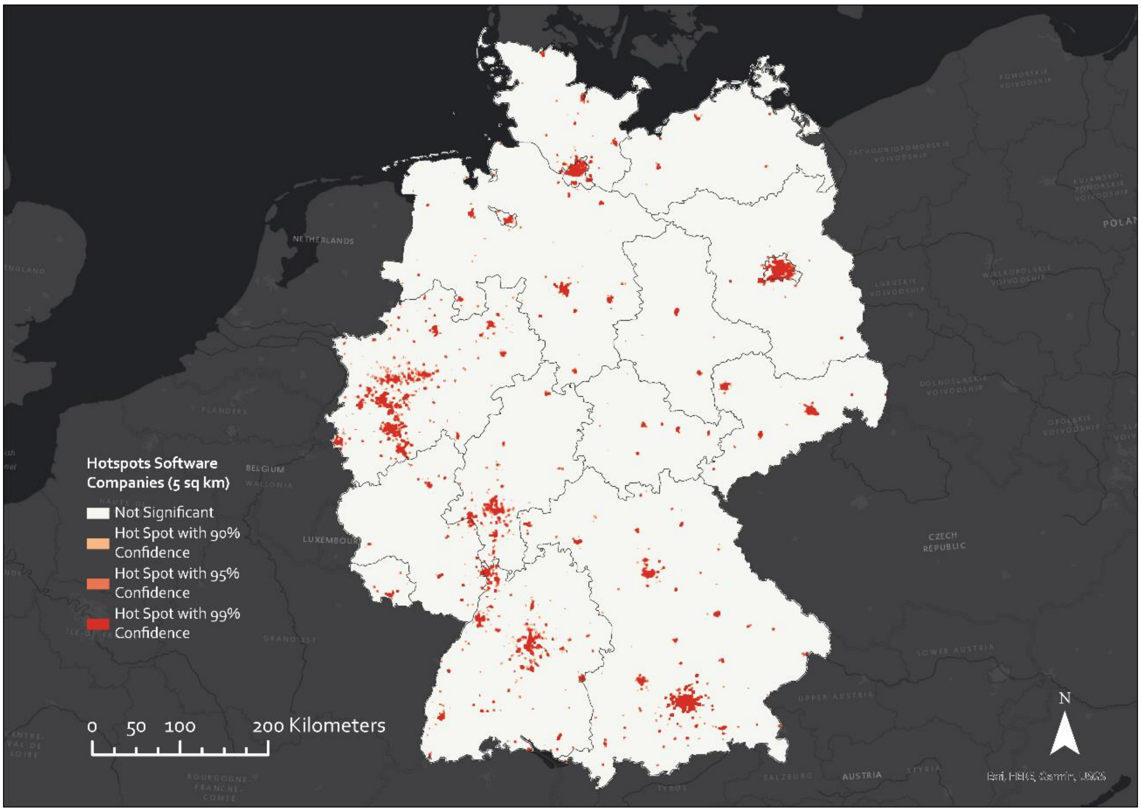

Hotspots Software Companies (5 sq km)
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
0 50 100 200 Kilometers
N
NETHERLANDS
BELGIUM
LUXEMBOURG
CZECH REPUBLIC
AUSTRIA
Esri, HERE, Garmin, USGS

Figure 25: Hotspot-Analysis - Software Companies in Germany (1 km grid)

5.2.3 Analysis of relations[5]

A bivariate nonparametric Spearman's rank correlation analysis was also performed here. The correlation matrix in Table 9 shows which two characteristics have a strong reciprocal relationship as well as the direction of the relationship.

Positive correlation	**Negative correlation**
Company_density – Company_sq: 0.99	Student_rate - Universities: -0.63
Pop_density - Pop_sq: 0.86	Unemployment - Life_expectany: -0.58
Street_centrality - Pop_density: 0.71	Life_expectancy - Average_age: -0.53
Pop_sq - Street_centrality: 0.70	Unemployment - Salary: -0.52
Pop_density - Network_converage: 0.64	Salary - Average_age: -0.51
Companies_sq - Pop_sq: 0.64	

Table 9: Strongest positive and negative correlation - Germany

Strong positive correlations with Software density (Software_companies) exist with Company density (r = 0.57) (square (sq), r = 0.57), Population density (r= 0.37) (square (sq), r= 0.42) and Street centrality (r = 0.35). Further positive correlations exist with the location factors of Public transport stops (r = 0.37), Entertainment (r = 0.40) and Recreation (r = 0.37). All these location factors show a high concentration in primarily urban areas. Also, worth mentioning is the Network Internet coverage (r = 0.3), Migration background (r = 0.21) and Student rate (r = 0.25). Weak negative correlations with software companies exist with the following location factors: Airport, Interstate, Research Institutes and Universities as well as Average age (r ≈ 0.12 - 0.19). No correlations with software companies were found with Terrain, Unemployment rate and Life expectation.

[5] Municipality-free areas as well as cells with missing values were removed from the data set, which corresponds to 1.04 % of the data. Although it does not affect the results of the analysis, it was done to guarantee a high quality of the data.

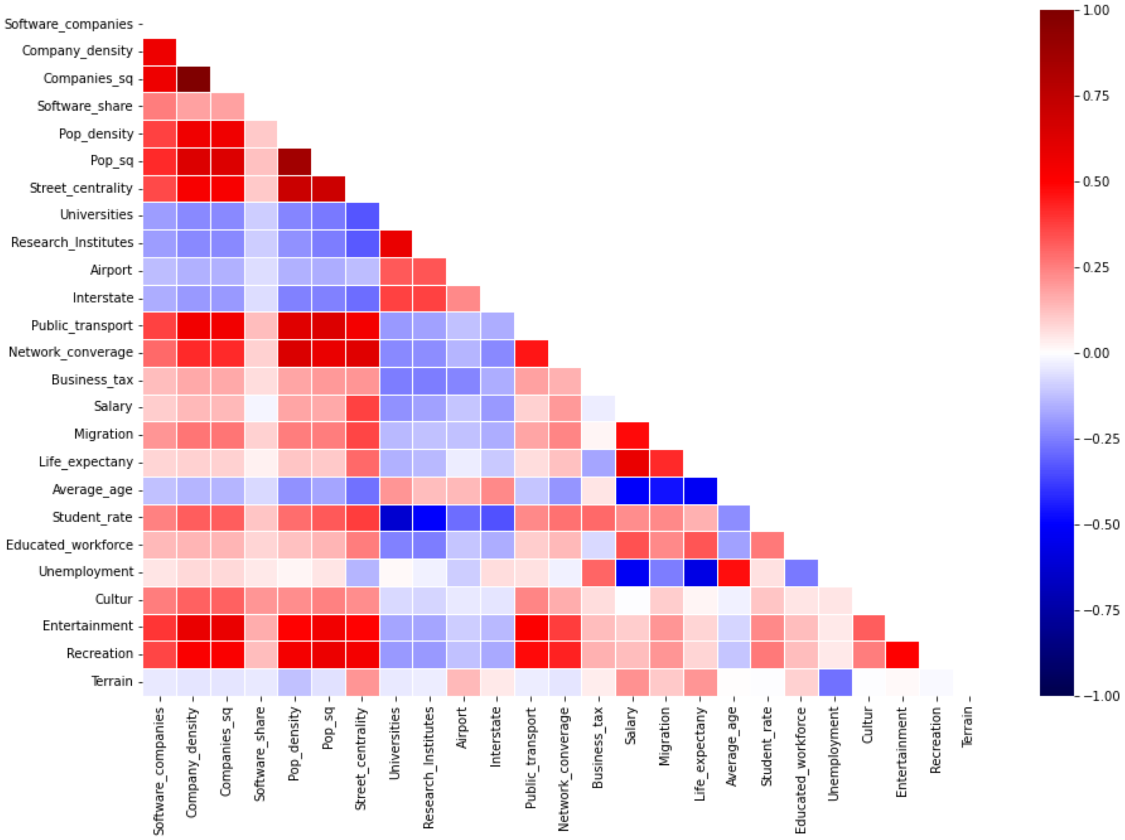


Figure 26: Correlation matrix - Location Factors Germany (p ≤ 0.05)

A descriptive analysis (N, min, max, $\bar{x}$, σ) was conducted for all variables too, the results can be taken from the appendix.

Location Factors	Description	IRR	SE
Agglomeration location factors			
Company density	Number of local companies (in 10).	1.147***	.0074
Company density^2	Squared number of local firms (in 10).	0.999***	.0000
Software companies share	Proportion of software companies in the local business population (in %)	1.082***	.0032
Population density	Population per cell (in 100).	1.168***	.0034
Population density^2	Squared number of inhabitants per cell (in 100).	0.998***	.0000
Street centrality	Street (network) density calculation (1). High value = High density	1.001***	.0000
Universities	Distance to the nearest university (in km).	0.994***	.0009
Research institutes	Distance to the nearest research institute (in km).	0.994***	.0009
Infrastructure location factors			
Network coverage broadband Internet	Average latency (upload / download speed) (in Mbps). High value = high internet speed	1.008***	.0001
Interstate / Highway	Distance to nearest high way / interstate (in km).	0.986***	.00012
Airport	Distance to nearest main civil airport (in km).	1.002***	.0004
Public transport	Weighted count of public transport stops.	1.005***	.0013
Socio-economic location factors			
Salary	Monthly household income (median) (in 100 EUR).	1.080***	.0026
Educated workforce	Proportion of employees with a university degree (in %).	1.004***	.0017
Student rate	Proportion of students in the local population in %.	0.994***	.0023
Business tax	Municipal business rate (in 100) fixed by the municipality. High values = high taxes	1.024	.0181
Life expectancy	Average life expectancy of the population (in years).	1.146***	.0111
Average age	Average age (median) of the population.	1.012***	.0038
Unemployment rate	Proportion of unemployed in the working-age population (in %).	0.982***	.0049
Migration background	Proportion of people of non-German nationality in the total population (in %).	1.005***	.0014
Amenities location factors			

Recreation	Number of recreational, community and sports facilities.	1.012***	.0011
Culture	Number of cultural sites and facilities.	0.982	.0151
Entertainment	Number of dining, nightlife and general entertainment facilities.	0.995*	.0027
Other			
Terrain	Average slope or gradient (in degree). High values = hillside location	0.994**	.0028

*Table 10: Location factors and estimated coefficients (IRR) with robust standard errors (SE) – Germany - */**/*** indicate significance at 10/5/1 per cent, respectively*

5.2.3.1 Interpretation of Regression Coefficients

Company density and population density have a significant positive effect on the number of local software companies in Germany. The squared values of these factors have significant negative coefficients, indicating an inverse u-shaped relationship as previously stated by Kinne and Resch (2018). The impact of street centrality on the number of local software firms is found to be only marginally positive. In contrast, a high proportion of software firms among the local firms is believed to have a substantial positive effect on the number of additional local software firms. This finding suggests that the existence of industry-owned firms stimulates the creation of new firms within the same industry.

The geographical location of the business sites plays a role, as it involves significant differences in terms of infrastructures and accessibility. An increase in the distance to research institutes and universities leads to a significant reduction in the number of software companies. This effect is also observed for highways in this study, but not for airports. The ease of access to public transportation appears to have a beneficial impact on the number of software companies in the local area. Furthermore, the study indicates that the availability of broadband internet connectivity has a significant positive impact on the establishment of new local software firms. Locations with good availability and accessibility to infrastructure and internet connection seem to be preferred.

Despite the significant positive effect of proximity to universities, it is rather contradictory that a high proportion of students in the local population shows a significant negative effect. However, a high proportion of university graduates in the local labor force is found to increase the number of software firms. The availability of skilled labor and the proximity to innovation drivers such as universities and research institutes seem to be decisive factors for the local economic development of software companies.

Average household income has a positive effect on the entry of new software companies. Higher business taxes show a rather positive, but not significant, effect and are not very plausible. It is worth mentioning that unlike the other socio-demographic factors that are detailed at the block group and ZIP code level, business taxes are set at the municipality level, as each municipality determines them individually. A significant increase in the number of software companies is associated with a high life expectancy and a higher average age. An increase in the share of people with migration background in the total local population is estimated to result in more local software companies. The unemployment rate seems to have a opposite effect; an increasing unemployment rate is associated with a decreasing number of local software companies. The various categories of amenities show that recreational facilities have a significant positive effect, while entertainment has rather a negative effect on the number of local software companies. The result of the amenity cultural facilities is not significant. Sites with slopes have a significant negative impact on the number of local software companies.

5.2.3.2 Model comparison

Table 11 presents the results of the Goodness of Fit (GoF) and prediction accuracy of the different models for Germany. The Poisson model corresponds to P = 0, while the linear Negative Binomial is P = 1 (NB1) and the quadratic negative binomial P = 2 (NB2). The Negative Binomial (MLE) where the coefficient of the regression model is estimated is P = 0.72.

Looking at the Pseudo-R² measure, the Poisson model has the best model fit, the quadratic NB the most inferior fit. The RMSE value shows that the Poisson model has the highest prediction accuracy, the deviation from the observed values is much smaller than for the NB models. The Poisson model has a better fit than all NB models, as indicated by the RMSE and pseudo-R² measures. When looking at the AIC and the BIC, the NB MEL model has the lowest values on these indices, but the deviations among the models are very small. The log-likelihood also barely differs between the models.

Measure	**Poisson**	**NB1**	**NB2**	**NB(MLE)**
Pseudo-R^2	0.614	0.525	0.473	0.543
RMSE	0.84	120.81	1,129.91	51.85
Log Likelihood	-77,194	-72,828	-73,952	-72,721
AIC	154,438	145,706	147,954	145,493
BIC	154,707	145,976	148,223	145,774

Table 11: Goodness of fit (GoF) – Germany[6]

[6] Method: Fisher, Scale parameter measure: Person chi-square, Chi-squared statistic: Wald, Estimator: Robust

Omnibus tests for all models were highly significant ($p < 0.001$). The results show that all models containing the full set of predictors represent a significant improvement over a null model.

Figure 27 visualizes the frequencies of observed against the predicted software firm counts for each model. The number of zero values was estimated correctly by all models. Cells with a low count were underestimated by all four models. All negative binomial models overestimate cells with high numbers of software companies. Although the Poisson model shows the highest underestimation at low count cells, it generally matches the observations best. This is also reflected in the RMSE values and Pseudo-R^2, even though not in the log-likelihood, AIC and BIC, which are less sensitive to overestimation and underestimation, respectively.

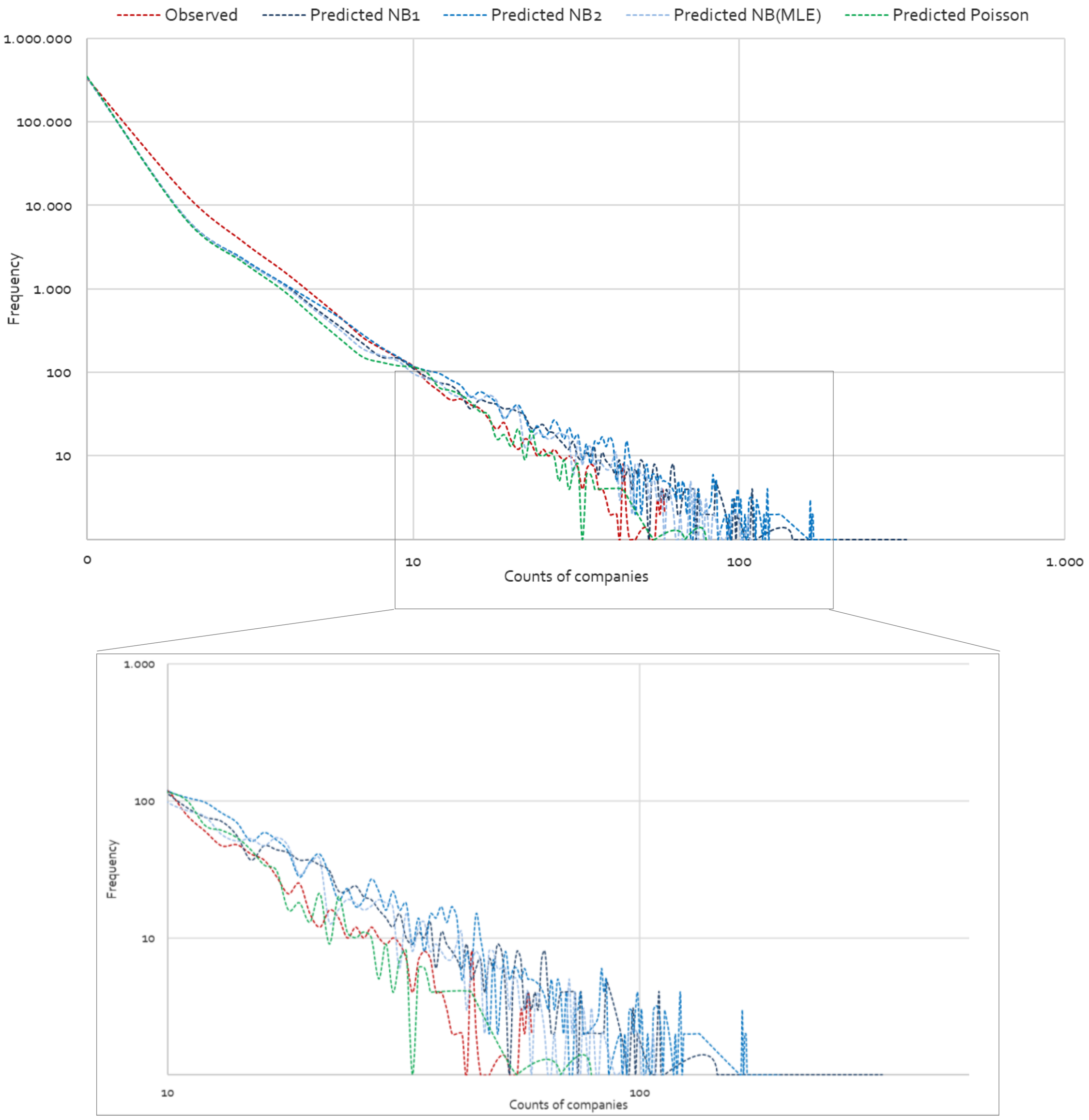


Figure 27: Frequencies of observed and predicted software firm counts - Germany

The models accurately estimate the zero values, but consistently undervalue cells with a low count of software companies and overvalue cells with a high count of software companies. The latter does not hold true for the Poisson model. Different from the study of Kinne and Resch (2018), municipality-free areas as well as the areas of large lakes (such as Chiemsee) and large nature protection areas (such as Bavarian forest), where it is impossible for companies to settle, were removed from the data set beforehand. This could be the reason why the number of zeros was estimated correctly, but does not provide an explanation for the underestimation respectively overestimation. In addition, there are other factors that account for the discrepancies between the results of this study and those of Kinne and Resch (2018). The explanatory variables, especially the socio-demographic location factors, are at a lower level of aggregation in this study (block group or zip code level). Additionally, the data set[7] used on the locations of (software) companies differs and - despite a sufficient data basis in both studies - can lead to deviations.

The two most suitable models - the Poisson and the NB MEL model - are compared in the following maps. The maps show that prediction errors occur in both models, especially in the economically strong metropolitan areas as well as along the agglomerations in the Rhine-Main region.

[7] This study: No. of all companies = 1,320,909 | No. of Software companies = 50,734
Study of Kinne & Resch (2018): No. of all companies = 2,970,000 | No. of Software companies = 70,009

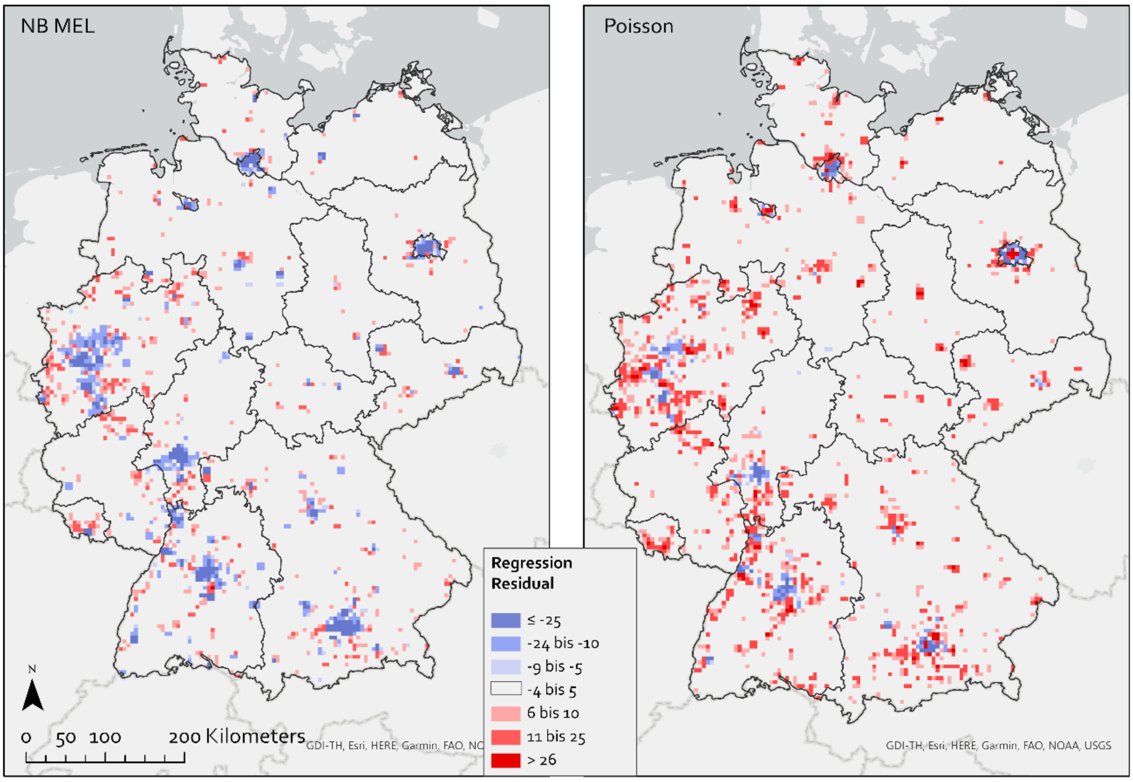


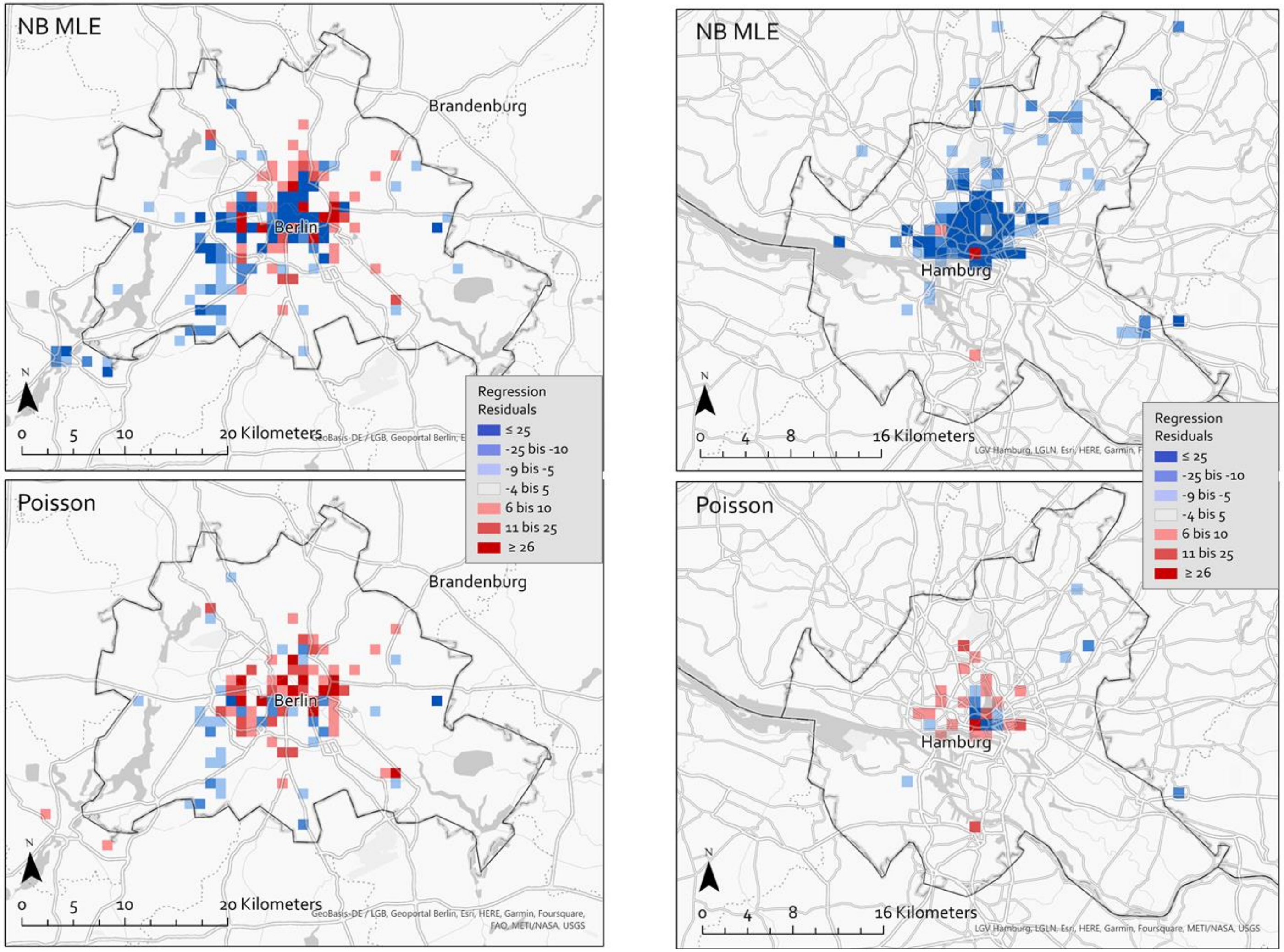
Figure 28: Regression residuals aggregated at 5 km grid in Germany

The further catchment area of Hamburg, Berlin, Munich, but also Stuttgart and Frankfurt, as well as along the Ruhr, which were identified in the ESDA as major hotspots of the software industry, the number of software companies was uniformly underestimated, although much less so in the case of the NB MEL model. The NB MEL model shows significant overestimates in the metropolitan areas, whereas underestimates in less population regions. The Poisson model overestimates as well as underestimates individual parts of city centers and shows a rather inconsistent picture as can be seen in Berlin, Hamburg or Munich. On the other hand, the Poisson model systematically underestimates peripheral areas of major cities as well as smaller cities such as Leipzig, Hanover, Saarbrucken, Kiel, etc.

Due to the aggregation, one gets the impression that the forecast errors are large-scale in the populated areas. However, looking at Figure 29, a somewhat different picture is shown. The initial evaluation on a 1 km² grid also shows an error in the forecast values - NB MEL rather overestimation, Poisson rather underestimation - but mainly in the centers of the metropolises. The suburbs of the illustrated cities (Berlin, Hamburg) and the rural regions show minor deviation from the observed values.

One possible explanation for the prediction errors in the metropolitan and densely populated areas could be the population characteristics. Although the data used are at a relatively low level of aggregation (block group, ZIP code), one can still assume a spatial bias that may generate systematic errors in some urban districts. At an aggregation level of 1 km^2, the spatial autocorrelation of the Moran's Index is 0.082 for the Poisson Model and 0.148 for the NB MEL Model. The Morans's I is low for both models, but above zero ($p \leq 0.001$) and might partially explain the prediction error.

Another reason for underestimation and distortion could be attributed to isolated local technology centers, such as Germany's largest technology park, Adlershof, located in the southeast of Berlin, and the Schöneweide Technology and Start-up Center in the northwest.



GeoBasis-DE / LGB, Geoportal Berlin, Esri, HERE, Garmin, Foursquare, FAO, METI/NASA, USGS

LGV Hamburg, LGLN, Esri, HERE, Garmin, Foursquare, METI/NASA, USGS

Figure 29: Regression residual original at 1 km grid of cities in Germany (Berlin, Hamburg)

6 Discussion of the results and methods used

The results of the IRR coefficient are first discussed and interpreted in context with other, previous studies. After, the model fit is discussed and the advantages and disadvantages as well as limitations of the respective models are considered.

6.1 Discussion of regression coefficients

6.1.1 Agglomeration location factors

The interest in location theory and business agglomerations has been present for quite some time on the academic side (Friedrich, 1929; Marshall, 2013; van Oort & Bosma, 2013; Farhauer & Kröll, 2014b). The approach of modeling agglomeration economies as a function of density is a common empirical approach. Agglomeration effects occur due to the geographic proximity and interrelation of economic activities, labor force, knowledge exchange and available infrastructure that makes this possible (Glaeser & Gottlieb, 2009; van Oort & Bosma, 2013; Faria *et al.*, 2020).

This study looked at the density of companies as well as the population. Although there is a correlation between the variables in both countries, there are differences at the microgeographic level, as the exploratory analysis has shown. The positive effects of agglomeration on companies' location decisions have been thoroughly investigated and confirmed in numerous studies (Bondonio & Greenbaum, 2007; Puga, 2010; van Oort & Bosma, 2013; Dunlap & Santos, 2021), which is present when productivity increases with density (Glaeser & Gottlieb, 2009). Recent studies have also shown that the relationship between agglomeration and economic activity is not linear. This is referred to as the "agglomeration shadow" effect, whereby an increase in the concentration of economic activities initially has a positive impact on location attractiveness, but eventually reaches a threshold where it turns negative. Density has an inverted u-shaped influence on the location economy (Rocha, 2008; Liviano & Arauzo-Carod, 2011). The study included an approach that involved squaring the two location factors of business and population density. The coefficient of squared location factors is significantly negative in Germany, confirming the assumed inverted u-shape of the effect and is consistent with the results by Kinne and Resch (2018). In the U.S., the coefficient for both location factors is significant, but neither positive nor negative. However, both coefficients are significantly lower than the coefficients of the non-squared location factors, so that an inversion of the positive effect can be assumed here as well.

Clusters of companies of one industry often encourage the location of new companies of the same industry (Rocha, 2008; Rammer *et al.*, 2016). Therefore, we considered the share of software companies in the population of local companies. The positive and significant coefficient confirms this assumption in both countries. Despite the underlying correlation, this factor appears the most important predictor in the U.S.. Software companies seem to like to settle where they meet their peers.

While the study by Kinne and Resch (2018) looks at the Urban Centrality Index, this study performed a Centrality Analysis of the street network. The network theory uses measures of centrality to determine the relative importance of a (traffic) node within an overall (road) network. The density of the nodes provides information about the density of the local road network. A study with this approach was not found. The analysis shows that a densification of the road network (= increase in local traffic nodes) is linked to an increase in software companies, and may be an indication of high mobility of people (and goods).

A significant number of software companies are born as spin-offs from universities and research institutes or are founded by researchers themselves, which often locate close to their parent organization. The close proximity means that they continue to benefit from knowledge spillovers. Given the high dynamics and innovation pressure in this industry, knowledge spillovers are of crucial importance (Anselin *et al.*, 1997; Maté-Sánchez-Val *et al.*, 2018; Arauzo-Carod, 2021). As in the case of Silicon Valley, entrepreneurship and experimentation are encouraged in the region due to its dense social networks and open labor markets. Companies engage in intense competition while also learning from each other about changing markets and technologies through informal communication and collaboration (Saxenian, 2018). The results also show a significant and negative coefficient for universities and research institutes (as the distance to these facilities increases, the number of software companies decreases) in both countries, confirming the importance of the relationship between these actors.

6.1.2 Infrastructure location factors

The importance of public transportation infrastructure has been examined in various studies and its importance has been confirmed. A well accessible transport infrastructure is likely to have an impact on local economic growth, as it affects the accessibility of an area (Swann, 2008; Liviano & Arauzo-Carod, 2011; Arauzo-Carod, 2021). Thanks to portable computers and wireless Internet, software companies are much more spatially flexible than other industries. Their most important and usually only significant input is a qualified workforce, so good public transportation connections seems to be important. The weighting of local public transport is based on the transport capacities of the means of transportation considered (Peter, 2005). In addition to the weighted number of public transport stops (bus, tram, subway / commuter rail), the distance (linear) to major civil airports and highways was also evaluated. Since long-distance trains play a subordinate role in the U.S., they were not included. The software industry is more dependent on broadband Internet than most other industries (Buxmann *et al.*, 2015). The analysis of the American data confirms the positive correlation of an advantageous infrastructure on the number of local software companies, apart from access to public transport. This could possibly be related to the unbalanced population structure and mobility behavior of the U.S., which is very different from the German one (see Chapter 4.1.1). In Germany, a positive relationship with infrastructure location factors can be identified with network coverage, distance to airport and accessibility public transport, but does not reflect the results of Kinne and Resch (2018).

6.1.3 Socio-economic location factors

The most important resource for knowledge-intensive companies is their workforce. Studies show that the availability of well-educated employees has a positive effect on the number of local companies (Panizza & Santis, 2018; Berger & Fisher, 2023). The analysis results indicate that a well-educated local workforce is a significant location factor in both countries and in line with the findings by Kinne and Resch (2018). Likewise, there is a positive relationship with the location factors people with migration background (Lee *et al.*, 2004) as well as average life expectancy (Kinne & Resch, 2018) and the number of local software companies.

The negative effect of high taxes on the location of companies has been well studied (Friedman *et al.*, 1992; Bondonio & Greenbaum, 2007). This study cannot confirm this for either of the two countries examined, especially since the result for Germany is not significant and also disagrees with the result by Kinne and Resch (2018).

Higher local unemployment rates can act as a deterrent to businesses by indicating less developed and dynamic areas (Egeln *et al.*, 2004). In Germany, the coefficient is significantly negative, as expected, whereas in the U.S. it is positive. Perhaps trying to capture these effects at the microgeographic level makes little sense, since cities tend to be a single labor market, and workers commute between neighborhoods or work increasingly from home offices these days.

While proximity to universities has a positive effect on the number of local software companies, a high proportion of students in the local population has a significant negative effect and is in line with the findings by Kinne and Resch (2018). One explanation could be that in the U.S., about one-third of students live on campus (Urban Institute, 2019), while in Germany, about 10 percent live in student housing (IWD, 2021), and thus may live in areas with a low agglomeration of businesses.

6.1.4 Amenities location factors and other

Young, well-educated and creative people have a strong preference for a rich and social life and are accordingly attracted to urban areas with a wide range of amenities. Software companies depend on this very workforce and may consequently follow them (Glaeser & Gottlieb, 2009; Florida & Mellander, 2016). As a result, people and businesses are increasingly locating in dense, amenity-rich areas (Moeller, 2014). In this study, amenities were divided into three groups: Recreation, Culture and Entertainment. In Germany, the analysis shows a positive relationship with recreation and culture, in the U.S. only recreation. The location factor culture was not significant in either country. It is difficult to determine exactly which amenity has a positive effect and at what scale level this occurs.

A factor that captures the terrain was also included in the analysis. Average slope or gradient (in degrees) was calculated by cell, to identify slopes. In both countries, the coefficient shows that the gradient is a significant and negative location factor, and is consistent with the results by (Kinne & Resch, 2018).

6.2 Comparison and discussion of model adequacy

After an exploratory analysis of the data, two count data models were chosen that are most commonly used in location analysis and for data with overdispersion: Negative Binomial Regression and Poisson Regression (Arauzo-Carod *et al.*, 2010; Arauzo-Carod & Manjón-Antolín, 2013; Cader *et al.*, 2013). Both models are appreciated because they naturally allow for large quantities of zero values (Rocha, 2008). Once a suitable model class had been chosen, the aim was to identify the most accurate model possible to reflect the reality of the given data. Based on exploratory analysis of U.S. software company data, a quadratic negative binomial regression was deemed a valid model. However, after conducting an overall evaluation of the goodness of fit parameters, the Poisson model was found to be the best option at the microgeographic level. Although the NB-MEL model outperformed the Poisson model in terms of AIC, BIC, and log-likelihood in both countries, and the pseudo-R2 performed well for all models, the significant differences in RMSE indicate the superiority of the Poisson model.

Unlike the study by Kinne and Resch (2018), there is no major prediction error in the zero values for any model in this study. It is possible that this is because certain areas, such as national parks, military bases, and large bodies of water, which are not suitable for the location of (software) companies, were excluded from the data in advance to avoid structural zeros. Nevertheless, I support the approach by Kinne and Resch (2018) to use zero-inflated models by including a land use classification model in the analysis. By assigning each grid cell uniquely by land use designation to one of two latent groups - construction land vs. non-construction land - structural zeros can be accurately identified. It is likely that this approach could achieve a better prediction result than the simple Poisson Model or NB Model used in this study. There are also other models, such as the Hurdle Model, which are good at modeling the excess of zeros. (Winkelmann, 2003; Buczkowska & Lapparent, 2014).

The problem of heterogeneity in socio-demographic data, as in Kinne and Resch (2018), is probably significantly lower in this study. With the exception of the location factor life expectancy, which was only available at district level in Germany and at county level in the U.S., the remaining socio-demographic location factors and population density are available at a quite detailed aggregation level. Unobserved heterogeneity cannot be ruled out with absolute certainty because the data are not available per household.[8]. But the geographic detail of the data used largely ensures a consideration of the different socio-demographic profiles of small spatial units such as individual neighborhoods. The low Moran's Index shows a rather weak autocorrelation of the predicted values on a microgeographic level.

One problem could be multicollinearity. Multicollinearity in the context of regression analysis occurs when there is a high correlation of two or more variables. If this is the case, the legitimate question arises whether two highly correlated independent variables do not actually measure the same thing and it would therefore be better to omit one of them (Field, 2018). Therefore, the existence of multicollinearity may lead to biased estimates of the regression coefficients for location factors, which may not accurately reflect their true impact on the number of software companies. The correlation matrix is a suitable diagnostic tool to detect collinearity. Figure 16 and Figure 26 illustrate the bivariate correlation coefficients of the analyzed location factors. According to Field (2018), a correlation value above 0.8 or below -0.8 is an indication of multicollinearity.

Another problem concerns endogeneity. There are two types of endogeneity. First, estimates may be subject to simultaneity bias. This bias is present when the outcome variable is a predictor of x, rather than simply a response of x. For example, it is difficult to distinguish whether local amenities (bars, restaurants, gyms, etc.) attract software companies or, if the causality runs in the other direction and certain amenities follow the location of companies. Second, the probability of unobserved variables in the error term is very high, so-called omitted variable bias (Hill *et al.*, 2021). For example, both local businesses as well as amenities may be influenced by a favorable local rent or a positive image of the neighborhood.

The results of this study suggest that operationalizing location factors at the microgeographic level is partially difficult. Different location factors work at different levels of aggregation, thus showing the sensitivity of scales (Openshaw, 1983). Therefore, aggregation of data is a cause of error and can affect spatial studies. When spatially varying data are aggregated into spatial units (grid cells, ZIP codes, blocks, counties), the

[8] Block level data (the lowest level) is not available in Germany as well as U.S. for data protection reasons.

resulting summary values can be affected and the underlying original patterns can be biased (Cattan, 2002). As Su *et al.* (2011) states, creating a single set of homogeneous spatial units is simply impossible, as each variable might have its own spatial pattern. Thus, the presumption here is that some location factors, such as cultural amenities or unemployment rate, are not meaningful at the microgeographic level, but are significant at larger levels of aggregation, such as city size. Arauzo-Carod and Manjón-Antolín (2012) show a possible approach to address the MAUP problem by using spatially lagged variables.

7 Conclusion

This study investigates the distribution of software companies in the countries Germany and the USA. A comprehensive dataset with geocoded firm data at street level was used for each country and possible location patterns of software firms were investigated in a comprehensive spatial exploratory data analysis (ESDA). In addition, models were presented and compared for the prediction of software companies' locations. For the location prediction, 24 factors were identified as predictor variables, that may have an influence on the location decision of software companies. The selection of factors was based on, but not limited to, the study by Kinne and Resch (2018). For an analysis on a microgeographic level, it was of great importance to use data on a low aggregation level, which was on the whole successful. The predefined research questions can be answered as follows:

i. Are there significant differences in location patterns of software companies between the U.S. and Germany?

In both countries, a global spatial autocorrelation of software companies at different levels of aggregation was found (see Figure 30). The analysis of the global trend provides conclusions about a high degree of spatial autocorrelation of software companies in Germany. The highest autocorrelation exists on a 1 km² grid. This value decreases continuously with increasing aggregation level. The agglomeration rate of software company locations in the U.S. is lower at this scale, but remains similarly high with increasing aggregation levels and only drops significantly at a 50 km². The general spatial autocorrelation differs between countries in that the software company in Germany clusters more strongly at the microgeographic level, whereas a strong cluster effect in the U.S. is observed not only at the microgeographic level (below 5 km²), but beyond. Comparatively, metropolitan areas in the U.S. are much larger than those in Germany and connect to huge megalopolises, so-called city bands (see BosWash, ChiPitts and SanSan), which do not yet exist in Germany.

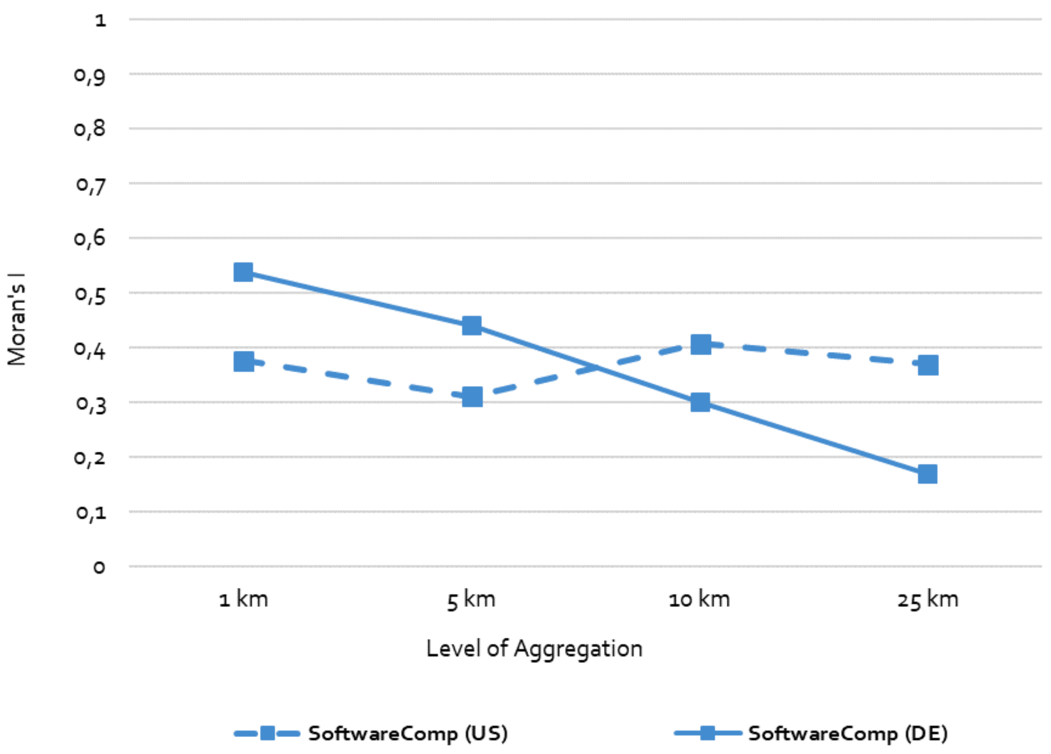


Figure 30: Moran's I of software companies - Germany vs. U.S.

The cluster outlier analysis and hotspot analysis identify statistically significant spatial clusters with high values of software companies in the metropolitan areas of both countries. Outside the metropolitan areas, the number of clusters decreases significantly, and software companies tend to be sporadically present in rural regions.

i. Are the locations in both countries explained by the same location factors?

An analysis on a microgeographic level requires adequate data. These should ideally be available as point data, which is not possible due to justified data protection concerns, or aggregated at a low administrative level (such as block group). This helps to reduce prediction errors. The results of the IRR coefficients are almost exclusively significant in both countries.

The impact of agglomeration location factors on the number of local software companies is almost identical in both countries. The concentration of companies, both within and outside the software industry, and a high population density have a positive effect on the number of local software companies in Germany and the U.S. Additionally, street centrality and proximity to universities and research institutes also have a positive effect in both countries.

There is a significant consensus regarding the socioeconomic location factors. Factors such as the availability of a well-educated workforce, high life expectancy, and a larger proportion of the local population with a migrant background appear to be crucial for the local economic development of software companies in both the U.S. and Germany.

Both countries show a positive influence of access to broadband internet and proximity to highways on the location of software companies. However, the location factors of public transport stops and proximity to the airport differ between the countries.

In terms of amenities, only the recreation location factor has a significant positive impact on the local software company count in both countries. Terrain appears to be a crucial factor in Germany and the U.S., with a steeper gradient associated with a significant reduction in the number of software companies.

The microgeographic prediction model created, based on the 24 location factors, was able to predict the location to a satisfactory degree in both countries. All applied models were able to deal with the excess zero values. However, there was an underestimation of low count cells in both countries and for all models and - except for Poisson Model - a significant overestimation of high count cells. Whether there are models that provide a better result for an entire country at the microgeographic level cannot be answered here. Rather, a comparison at the country level has more of an informative side. But to be able to plan practical measures such as a location choice, one certainly needs small-scale analyses.

8 References

Abelairas-Etxebarria, P. & Astorkiza, I. (2020) From Exploratory Data Analysis to Exploratory Spatial Data Analysis. *Mathematics and Statistics*, 8(2), 82–86. Available from: https://doi.org/10.13189/ms.2020.080202.

Altunbaş, Y., Jones, E. & Thornton, J. (2013) Knowledge spillovers and the growth of British cities. *Applied Economics Letters*, 20(2), 162–166. Available from: https://doi.org/10.1080/13504851.2012.684773.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Springer Netherlands: Dordrecht.

Anselin, L. (1995) Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27(2), 93–115. Available from: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L., Varga, A. & Acs, Z. (1997) Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics*, 42(3), 422–448. Available from: https://doi.org/10.1006/juec.1997.2032.

Arauzo-Carod, J.-M. (2021) Location determinants of high-tech firms: an intra-urban approach. *Industry and Innovation*, 1–24. Available from: https://doi.org/10.1080/13662716.2021.1929868.

Arauzo-Carod, J.-M., Liviano-Solis, D. & Manjón-Antolín, M. (2010) EMPIRICAL STUDIES IN INDUSTRIAL LOCATION: AN ASSESSMENT OF THEIR METHODS AND RESULTS*. *Journal of Regional Science*, 50(3), 685–711. Available from: https://doi.org/10.1111/j.1467-9787.2009.00625.x.

Arauzo-Carod, J.-M. & Manjón-Antolín, M. (2012) (Optimal) spatial aggregation in the determinants of industrial location. *Small Business Economics*, 39(3), 645–658. Available from: https://doi.org/10.1007/s11187-011-9335-6.

Arauzo-Carod, J.-M. & Manjón-Antolín, M. (2013) New Insights in Industrial Location. *Tijdschrift voor economische en sociale geografie*, 104(2), 133–135. Available from: https://doi.org/10.1111/tesg.12016.

Asheim, B.T. & Coenen, L. (2008) The role of regional innovation systems in a globalising economy: Comparing knowledge bases and institutional frameworks of Nordic clusters, in: Vertova, Giovanna (Hrsg.), The Changing Economic Geography of Globalization: Reinventing Space. *Economic Geography*, 84(2), 166–183. Available from: https://doi.org/10.1111/j.1944-8287.2008.tb00408.x.

Ashish Arora, Alfonso Gambardella & Salvatore Torrisi (2001) *In the footsteps of Silicon Valley? Indian and Irish software in the international division of labor*.

Berger, N. & Fisher, P. (2023) *A Well-Educated Workforce Is Key to State Prosperity*. Available from: https://www.epi.org/publication/states-education-productivity-growth-foundations/ [Accessed 22 February 2023].

Bondonio, D. & Greenbaum, R.T. (2007) Do local tax incentives affect economic growth? What mean impacts miss in the analysis of enterprise zone policies. *Regional Science and Urban Economics*, 37(1), 121–136. Available from: https://doi.org/10.1016/j.regsciurbeco.2006.08.002.

Briant, A., Combes, P.-P. & Lafourcade, M. (2010) Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, 67(3), 287–302. Available from: https://doi.org/10.1016/j.jue.2009.09.014.

Buczkowska, S. & Lapparent, M. de (2014) Location choices of newly created establishments: Spatial patterns at the aggregate level. *Regional Science and Urban Economics*, 48, 68–81. Available from: https://doi.org/10.1016/j.regsciurbeco.2014.05.001.

Bundesinstitut für Bau-, Stand- und Raumforschung (2020) *Startseite*. Available from: https://www.bbsr.bund.de/BBSR/DE/startseite/_node.html [Accessed 4 September 2022].

Bureau van Dijk (2022) *Informationen zu nicht börsennotierten Unternehmen – Orbis*. Available from: https://www.bvdinfo.com/de-de/ [Accessed 15 May 2022].

Buxmann, P., Diefenbach, H. & Hess, T. (2015) *Die Softwareindustrie: Ökonomische Prinzipien, Strategien, Perspektiven*, 3rd edition. Springer Gabler: Berlin.

Cader, H.A., Crespi, J.M. & Leatherman, J.C. (2013) What Factors Affect Information Technology Firm Location Choices in Middle America? An Examination of Regional and Industrial Variation in Kansas. *International Regional Science Review*, 36(2), 207–234. Available from: https://doi.org/10.1177/0160017611415268.

Capello, R. (2014) Classical Contributions: Von Thünen, Weber, Christaller, Lösch. In: Fischer, M.M. & Nijkamp, P. (Eds.) *Handbook of Regional Science.* Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 507–526.

Cattan, N. (2002) REDEFINING TERRITORIES: FUNCTIONAL REGIONS. *OECD Territorial Outlook*. Available from: https://www.istat.it/it/files/2014/12/Cattan-redifining-Territories-Functional-Regions.pdf [Accessed 12 December 2021].

Columbus, L. (2016) *PwC Global 100 Software Leaders, 2016: Subscription & Cloud Apps Revolutionizing Enterprise Software*, personal communication. 30 July. Available from: https://www.forbes.com/sites/louiscolumbus/2016/07/30/pwc-global-100-software-leaders-2016-subscription-cloud-apps-revolutionizing-enterprise-software/?sh=7352aa5f4466 [Accessed 7 February 2021].

Combes, P.-P. & Duranton, G. (2006) Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics*, 36(1), 1–28. Available from: https://doi.org/10.1016/j.regsciurbeco.2005.06.003.

County Health Rankings & Roadmaps (2022) *How Healthy is your County? | County Health Rankings*. Available from: https://www.countyhealthrankings.org/ [Accessed 4 September 2022].

Coxe, S., West, S.G. & Aiken, L.S. (2009) The analysis of count data: a gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2), 121–136. Available from: https://doi.org/10.1080/00223890802634175.

Dorfman, N.S. (1983) Route 128: The development of a regional high technology economy. *Research Policy*, 12(6), 299–316. Available from: https://doi.org/10.1016/0048-7333(83)90009-4.

Dunlap, D.R. & Santos, R.S. (2021) Storming the Beachhead: An Examination of Developed and Emerging Market Multinational Strategic Location Decisions in the U.S. *Journal of Risk and Financial Management*, 14(7), 325. Available from: https://doi.org/10.3390/jrfm14070325.

Duranton, G. & Overman, H.G. (2002) Testing for Localisation Using Micro-Geographic Data. *Centre for Economic Performance*, 2002. Available from: https://doi.org/10.1111/0034-6527.00362.

Egeln, J., Gottschalk, S. & Rammer, C. (2004) Location Decisions of Spin-offs from Public Research Institutions. *Industry & Innovation*, 11(3), 207–223. Available from: https://doi.org/10.1080/1366271042000265384.

Elwood, S., Goodchild, M.F. & Sui, D.Z. (2012) Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, 102(3), 571–590. Available from: https://doi.org/10.1080/00045608.2011.595657.

ESRI (2020) *Was ist ArcGIS Living Atlas of the World?* Portal for ArcGIS. Available from: https://enterprise.arcgis.com/de/portal/latest/use/what-is-living-atlas.htm [Accessed 10 May 2021].

ESRI (2021) *Living Atlas of the World | ArcGIS*. Available from: https://livingatlas.arcgis.com/de/participate/ [Accessed 10 May 2021].

Farhauer, O. & Kröll, A. (2014a) Technologischer Fortschritt und der Lebenszyklus von Städten. In: Farhauer, O. & Kröll, A. (Eds.) *Standorttheorien.* Springer Fachmedien Wiesbaden: Wiesbaden, pp. 289–296.

Farhauer, O. & Kröll, A. (Eds.) (2014b) *Standorttheorien*. Springer Fachmedien Wiesbaden: Wiesbaden.

Faria, J.R., Ferreira, J.J., Johnson, K.H., Mixon, F.G. & Wanke, P.F. (2020) Agglomeration economies and university program creation in the knowledge economy. *Socio-Economic Planning Sciences*, 72, 100800. Available from: https://doi.org/10.1016/j.seps.2020.100800.

Field, A.P. (2018) *Discovering statistics using IBM SPSS statistics*, 5th edition. SAGE Publications: London, Thousand Oaks, California.

Fischer, M.M. & Nijkamp, P. (2021) *Handbook of Regional Science*. Springer Berlin Heidelberg: Berlin, Heidelberg.

Flanagin, A.J. & Metzger, M.J. (2008) The credibility of volunteered geographic information. *GeoJournal*, 72(3-4), 137–148. Available from: https://doi.org/10.1007/s10708-008-9188-y.

Florida, R. & Mellander, C. (2016) Rise of the Startup City: The Changing Geography of the Venture Capital Financed Innovation. *California Management Review*, 59(1), 14–38. Available from: https://doi.org/10.1177/0008125616683952.

Fonte, C.C., Bastin, L., Foody, G., Kellenberger, T., Kerle, N. & Mooney, P. et al. (2015) VGI QUALITY CONTROL. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5, 317–324. Available from: https://doi.org/10.5194/isprsannals-II-3-W5-317-2015.

Freie Universität Berlin (2023) *Wahl der Modellklasse: lineare Regression, Logit Modell etc.* Available from: https://wikis.fu-berlin.de/pages/viewpage.action?pageId=735543500 [Accessed 28 February 2023].

Friedman, J., Gerlowski, D.A. & Silberman, J. (1992) WHAT ATTRACTS FOREIGN MULTINATIONAL CORPORATIONS? EVIDENCE FROM BRANCH PLANT LOCATION IN THE UNITED STATES. *Journal of Regional Science*, 32(4), 403–418. Available from: https://doi.org/10.1111/j.1467-9787.1992.tb00197.x.

Friedrich, C.J. (1929) *Alfred Weber´s theory of the location of industries*. Cambridge University Press: Chicago, III.

Getis, A. & Ord, J.K. (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), 189–206. Available from: https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Glaeser, E.L. & Gottlieb, J.D. (2009) The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States. *Journal of Economic Literature*, 47(4), 983–1028. Available from: https://doi.org/10.1257/jel.47.4.983.

Hahn, B. (2014) *Die US-amerikanische Stadt im Wandel*, 2014th edition. Springer Berlin Heidelberg: Berlin, Heidelberg.

Haining, R.P. (2009) *Spatial data analysis: Theory and practice*, 6th edition. Cambridge Univ. Press: Cambridge.

Haklay, M. (2010) How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, 37(4), 682–703. Available from: https://doi.org/10.1068/b35097.

Hilbe, J.M. (2011) Modeling Count Data. In: Lovric, M. (Ed.) *International Encyclopedia of Statistical Science.* Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 836–839.

Hilbe, J.M. (2014) *Modeling Count Data*. Cambridge University Press: Cambridge.

Hill, A.D., Johnson, S.G., Greco, L.M., O'Boyle, E.H. & Walter, S.L. (2021) Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research. *Journal of Management*, 47(1), 105–143. Available from: https://doi.org/10.1177/0149206320960533.

Hooton, C.A. (2016) *Micro-geographic economic analysis : the theory, techniques, and evidence of micro-level economic policies and their evaluations*. Cambridge, United Kingdom, Apollo - University of Cambridge Repository. Available from: https://doi.org/10.17863/CAM.22432.

IBM (2022) *IBM Documentation*. Available from: https://www.ibm.com/docs/de/spss-statistics/saas?topic=model-pseudo-r-square [Accessed 12 February 2023].

infas360 (2022) *infas360*. Available from: https://www.infas360.de/ [Accessed 10 October 2022].

Infogroup (2016) *Infogroup US Historical Business Data*. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PNOFKI [Accessed 15 March 2022].

IWD (2021) Wohnungsnot der Studenten verschärft sich weiter. *IWD,* 2 October. Available from: https://www.iwd.de/artikel/wohnungsnot-der-studenten-verschaerft-sich-weiter-499592/ [Accessed 24 February 2023].

Kelejian, H.H. & Prucha, I.R. (2001) On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics*, 104(2), 219–257. Available from: https://doi.org/10.1016/S0304-4076(01)00064-1.

Kinne, J. & Axenbeck, J. (2020) Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics*, 125(3), 2011–2041. Available from: https://doi.org/10.1007/s11192-020-03726-9.

Kinne, J. & Resch, B. (2018) Analyzing and Predicting Micro-Location Patterns of Software Firms. *ISPRS International Journal of Geo-Information*, 7(1), 1–21. Available from: https://doi.org/10.3390/ijgi7010001.

Kitchin, R. & Thrift, N.J. (Eds.) (2009) *International encyclopedia of human geography*. Elsevier: Amsterdam, Oxford.

Lee, S.Y., Florida, R. & Acs, Z. (2004) Creativity and Entrepreneurship: A Regional Analysis of New Firm Formation. *Regional Studies*, 38(8), 879–891. Available from: https://doi.org/10.1080/0034340042000280910.

Liviano, D. & Arauzo-Carod, J.-M. (2011) *Industrial Location and Space: New Insights*. Working Papers, Tarragona, Universitat Rovira i Virgili.

Marshall, A. (2013) *Principles of Economics*. Palgrave Macmillan UK: London.

Marshall, A. (2014) *Principles of Economics: Palgrave Classics in Economics*. Palgrave Macmillan UK: London.

Maskell, P. & Kebir, L. (2005) What Qualifies as a Cluster Theory? Available from: https://doi.org/10.4324/9780203640890.

Maté-Sánchez-Val, M., López-Hernandez, F. & Rodriguez Fuentes, C.C. (2018) Geographical factors and business failure: An empirical study from the Madrid metropolitan area. *Economic Modelling*, 74, 275–283. Available from: https://doi.org/10.1016/j.econmod.2018.05.022.

Méndez-Ortega, C. & Arauzo-Carod, J.-M. (2019) *Do software and video game firms share location patterns across cities? Evidence from Barcelona, Lyon and Hamburg*. Reus, Universitat Rovira i Virgili. Available from: https://doi.org/10.1007/s00168-019-00917-y.

Moeller, K. (2014) *Culturally Clustered or in the Cloud? Location of Internet Start-ups in Berlin:*. SERC Discussion Paper 157, London, UK, London School of Economics.

Moulaert, F. & Sekia, F. (2003) Territorial Innovation Models: A Critical Survey. *Regional Studies*, 37(3), 289–302. Available from: https://doi.org/10.1080/0034340032000065442.

Müller, T.A. (2003) *Kunden- und Wettbewerbsorientierung neugegründeter Softwareunternehmen: Eine empirische Untersuchung von Teamgründungen*. Deutscher Universitätsverlag: Wiesbaden.

Murray, A.T. (2009) Location Theory. In: Kitchin, R. & Thrift, N.J. (Eds.) *International encyclopedia of human geography.* Elsevier: Amsterdam, Oxford, pp. 237–243.

Nexiga (2022) *Nexiga - Location Intelligence Lösungen vom Experten*. Available from: https://www.nexiga.com/ [Accessed 4 September 2022].

O'Kelly, M. & Bryan, D. (1996) Agricultural location theory: von Thunen's contribution to economic geography. *Progress in Human Geography*, 20(4), 457–475. Available from: https://doi.org/10.1177/030913259602000402.

Ookla (2022) *Ookla's Open Data Initiative | Ookla® - Providing network intelligence to enable modern connectivity*. Available from: https://www.ookla.com/ookla-for-good/open-data [Accessed 4 September 2022].

Openshaw, S. (1983) *The modifiable areal unit problem*. Geo Books: Norwich.

Panizza, A. de & Santis, S. de (2018) *Education of employers and the workforce, enterprise performance and the adoption of IT tools and innovations: evidence from the population of Italian small firms*, Paris.

Perumean-Chaney, S.E., Morgan, C., McDowall, D. & Aban, I. (2013) Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation*, 83(9), 1671–1683. Available from: https://doi.org/10.1080/00949655.2012.668550.

Pesaresi, M., Florczyk, A., Schiavina, M., Melchiorri, M. & Maffenini, L. (2019) *GHS settlement grid, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015), R2019A*. Available from: https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218.

Peter, R. (2005) *Kapazitäten und Flächenbedarf öffentlicher Verkehrssysteme in schweizerischen Agglomerationen*. Term Paper, Zürich, Schweiz, ETH Zürich.

Peters, S. (2021) *Europe's Silicon Valley of Enterprise Software*. Available from: https://www.clusterplattform.de/CLUSTER/Redaktion/EN/Cluster/software_cluster.html [Accessed 3 November 2021].

Petersen, T. & Thode, E. (2015) *Globalisierung, Digitalisierung und Einkommensungleichheit*, personal communication. 2015 [Accessed 30 April 2021].

Porter, M.E. (2000) Location, Competition, and Economic Development: Local Clusters in a Global Economy, 14(1), 15–34. Available from: https://doi.org/10.1177/089124240001400105.

Puga, D. (2010) THE MAGNITUDE AND CAUSES OF AGGLOMERATION ECONOMIES. *Journal of Regional Science*, 50(1), 203–219. Available from: https://doi.org/10.1111/j.1467-9787.2009.00657.x.

Rammer, C., Kinne, J. & Blind, K. (2020) Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. *Urban Studies*, 57(5), 996–1014. Available from: https://doi.org/10.1177/0042098018820241.

Rammer, C., Kinne, Jan & Blind, K. (2016) Urban Innovative Neighbourhoods: Micro-Location Patterns of Innovative Firms in Berlin, 2016. Available from: https://doi.org/10.2139/ssrn.2882503.

Real Estate Pilot (2022) *Home - Real Estate Pilot*. Available from: https://realestatepilot.com/ [Accessed 4 September 2022].

Rocha, N. (2008) *Firm location determinants: Empirical evidence for France*. HEID Working Paper No. 08/2008, Geneva, Graduate Institute of International and Development Studies.

Rosegrant, S. & Lampe, D.R. (1992) *Route 128: lessons from Boston's high-tech community*. Basic Books: New York, NY.

Ross, E.A. (1896) The Location of Industries. *The Quarterly Journal of Economics*, 10(3), 247–268. Available from: https://doi.org/10.2307/1882585.

Saxenian, A. (2018) Inside-Out: Regional Networks and Industrial Adaptation in Silicon Valley and Route 128. In: Granovetter, M. & Swedberg, R. (Eds.) *The Sociology of Economic Life.* Routledge, pp. 357–374.

Sehra, S., Singh, J. & Rai, H. (2017) Assessing OpenStreetMap Data Using Intrinsic Quality Indicators: An Extension to the QGIS Processing Toolbox. *Future Internet*, 9(2), 15. Available from: https://doi.org/10.3390/fi9020015.

Smętkowski, M., Celińska-Janowicz, D. & Wojnar, K. (2021) Location patterns of advanced producer service firms in Warsaw: A tale of agglomeration in the era of creativity. *Cities*, 108, 102937. Available from: https://doi.org/10.1016/j.cities.2020.102937.

Su, M.D., Lin, M.-C. & Wen, T.H. (2011) Spatial Mapping and Environmental Risk Identification. In: Nriagu, J.O. (Ed.) *Encyclopedia of environmental health.* Elsevier Science: Amsterdam, London, pp. 228–235.

Swann, P.G.M. (2008) Place is what we think with: Or spatial history, intellectual capital and competitive distinction, in: Vertova, Giovanna (Hrsg.), The Changing Economic Geography of Globalization: Reinventing Space. *Economic Geography*, 84(2), 102–117. Available from: https://doi.org/10.1111/j.1944-8287.2008.tb00408.x.

Taylor & Francis (2021) *Regional industrialization | Determinants of industrial location | Gle*. Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9780367197537-13/regional-industrialization-glenn-rayp-stijn-ronsse [Accessed 23 September 2021].

Tyrväinen, P. & Mazhelis, O. (Eds.) (2009) *Vertical Software Industry Evolution*. Physica-Verlag HD: Heidelberg.

United States Census Bureau (2021) *2017 County Business Patterns and Economic Census: The Number of Firms and Establishments, Employment, Annual Payroll, and Receipts by State, Industry, and Enterprise Employment Size: 2017*.

United States Census Bureau (2022) *Census Bureau Data*. Available from: https://data.census.gov/cedsci/ [Accessed 4 September 2022].

United States Geological Survey (2022) *Coastal Changes and Impacts | U.S. Geological Survey*. Available from: https://www.usgs.gov/coastal-changes-and-impacts [Accessed 4 September 2022].

Urban Institute (2019) *Understanding College Affordability*. Available from: https://collegeaffordability.urban.org/prices-and-expenses/room-and-board/#/room_and_board_by_type_of_institution [Accessed 24 February 2023].

van Oort, F.G. & Bosma, N.S. (2013) Agglomeration economies, inventors and entrepreneurs as engines of European regional economic development. *The Annals of Regional Science*, 51(1), 213–244. Available from: https://doi.org/10.1007/s00168-012-0547-8.

Wang, M., Li, Q., Hu, Q. & Zhou, M. (2013) Quality Analysis of open street map data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-2/W1, 155–158. Available from: https://doi.org/10.5194/isprsarchives-XL-2-W1-155-2013.

Winkelmann, R. (2003) *Econometric analysis of count data*. Springer: Berlin, New York.

World Bank (2009) *Reshaping economic geography*, Washington D.C. World development report. Available from: https://doi.org/10.1093/jeg/lbp019.

Zook, M.A. (2005a) Mapping the Internet Industry. In: Zook, M.A. (Ed.) *The Geography of the Internet Industry.* Blackwell Publishing Ltd: Oxford, UK, pp. 24–39.

Zook, M.A. (2005b) Uncovering the Geography of the Internet Industry. In: Zook, M.A. (Ed.) *The Geography of the Internet Industry.* Blackwell Publishing Ltd: Oxford, UK, pp. 1–9.

Appendix

Appendix 1 - Global Moran's I Summary

Grid size	Moran's Index	Expected Index	Variance	z-score	p-value
Software companies					
1 km	0.537452	-0.000003	0.000001	647.036986	0.000000
3 km	0.523033	-0.000024	0.000006	213.064417	0.000000
5 km	0.439516	-0.000067	0.000016	110.426466	0.000000
10 km	0.300828	-0.000259	0.000060	38.756307	0.000000
25 km	0.169253	-0.001490	0.000346	9.185306	0.000000
Other companies					
1 km	0.582203	-0.000003	0.000001	698.992249	0.000000
3 km	0.519657	-0.000024	0.000006	210.135457	0.000000
5 km	0.490850	-0.000067	0.000017	119.900111	0.000000
10 km	0.372681	-0.000259	0.000063	46.820452	0.000000
25 km	0.269300	-0.001490	0.000374	14.003548	0.000000
Population					
1 km	0.710964	-0.000003	0.000001	849.549232	0.000000
3 km	0.706304	-0.000025	0.000006	281.727315	0.000000
5 km	0.672998	-0.000067	0.000017	161.506647	0.000000
10 km	0.552564	-0.000259	0.000067	67.357697	0.000000
25 km	0.358083	-0.001497	0.000394	18.115190	0.000000

Table 12: Global Moran's I Summary – Germany

Grid size	Moran's Index	Expected Index	Variance	z-score	p-value
Software companies					
1 km	0.376568	-0.000000	0.000000	2166.223333	0.000000
5 km	0.310374	-0.000003	0.000001	372.176125	0.000000
10 km	0.407163	-0.000012	0.000003	237.995903	0.000000
25 km	0.369365	-0.000074	0.000018	86.946715	0.000000
50 km	0.248958	-0.000286	0.000071	29.586508	0.000000
Other companies					
1 km	0.495868	-0.000000	0.000000	2823.676634	0.000000
5 km	0.481751	-0.000003	0.000001	559.978293	0.000000
10 km	0.543731	-0.000012	0.000003	313.495482	0.000000
25 km	0.405377	-0.000074	0.000018	95.563618	0.000000
50 km	0.270874	-0.000286	0.000070	32.458000	0.000000
Population					
1 km	0.765396	-0.000000	0.000000	4341.872004	0.000000
5 km	0.801475	-0.000003	0.000001	913.614927	0.000000
10 km	0.664755	-0.000012	0.000003	380.657432	0.000000
25 km	0.477774	-0.000074	0.000018	111.705295	0.000000
50 km	0.305718	-0.000286	0.000071	36.339590	0.000000

Table 13: Global Moran's I Summary – U.S.

Appendix 2 – Descriptive Statistic Location Factors

Location Factor	N	Min	Max	Mean	Std. Deviation
Agglomeration Location Factors					
Company density	357,723	0	246.00	0.33	1,94
Company density squared	357,723	0	60516,00	3,88	172,76
Software companies	357,723	0	216.00	0.14	1.31
Software companies share	357,723	0	50.00	0.04	0.76
Population density	357,723	0	94.42	2.26	6.20
Population density squared	357,723	0	8915.00	43.49	265.28
Street centrality	357,723	0	2384.98	52.00	92.47
Universities	357,723	0.019	91.31	23.34	13.83
Research institute	357,723	0.011	86.65	21.26	13.23
Infrastructure Factors					
Network coverage broadband Internet	357,723	0	626.91	25.25	35.27
Interstate / Highway	357,723	0.000003	69.52	11.57	9.91
Airport	357,723	0.160	156.84	47.47	24.21
Public transport	357,723	0	284.00	1.47	4.88
Socio-economic Factors					
Salary	357,723	9.00	65.00	30.95	3.10
Education level	357,723	0	50.46	12.32	4.03
Student rate	357,723	0	50.00	0.65	2.20
Business tax	357,723	2.00	6.00	3.75	0.45
Life expectancy	357,723	78.28	83.92	81.13	0.92
Average age	357,723	26.30	69.50	45.91	2.55
Unemployment rate	357,723	0	41.30	3.44	1.87
Migration background	357,723	0	100.00	6.21	5.64
Recreational value Factors					
Recreation	357,723	0	152.00	0.88	3.12
Culture	357,723	0	40.00	0.04	0.42
Entertainment	357,723	0	494.00	0.76	5.43
Other					
Terrain	357,723	0	47.00	2.57	3.24

Table 14: Descriptive Statistic - Germany

Location Factor	N	Min	Max	Mean	Std. Deviation
Agglomeration Location Factors					
Company density	7,899,190	0	1,666.00	0.19	2.53
Company density squared	7,899,190	0	2,775,556.00	6.42	1425.51
Software companies	7,899,190	0	336.00	0.014	0.39
Software companies share	7,899,190	0	100.00	0.06	1.75
Population density	7,899,190	0	593.35	0.40	2.93
Population density squared	7,899,190	0	352,068.53	8.74	424.19
Street centrality	7,899,190	0	1,263.21	3.59	14.64
Universities	7,899,190	0.004	287.14	51.16	40.57
Research institute	7,899,190	0.006	257.58	40.70	34.66
Infrastructure Factors					
Network coverage broadband Internet	7,899,190	0	2,966.00	7.90	52.37
Interstate / Highway	7,899,190	0	307.94	54.26	49.90
Airport	7,899,190	0.022	246.98	62.09	36.01
Public transport	7,899,190	0	234.00	0.04	0.85
Socio-economic factors					
Salary	7,899,190	7,741.00	24,6750.00	49,075.51	16,014.70
Education level	7,899,190	0	100.00	19.43	9.93
Student rate	7,899,190	0	100.00	3.61	3.13
Business tax	7,899,190	0	12.00	5.35	3.22
Life expectancy	7,899,190	61.63	104.74	77.98	2.99
Average age	7,899,190	16.50	84.00	43.39	6.63
Unemployment rate	7,899,190	0	100.00	4.40	4.41
Migration background	7,899,190	0	100.00	20.33	21.11
Recreational value Factors					
Recreation	7,899,190	0	375.00	0.12	1.46
Culture	7,899,190	0	38.00	0.00	0.11
Entertainment	7,899,190	0	363.00	0.05	0.93
Other					
Terrain	7,899,190	0	15.51	0.81	1.15

Table 15: Descriptive Statistic – U.S.

Appendix 3 – OSM Filter

gis_osm_transport_a_free_1.shp: Transport

fclass = 'bus_station' Or fclass = 'bus_stop' Or fclass = 'railway_halt' Or fclass = 'railway_station' Or fclass = 'tram_stop'

gis_osm_pois_free_1: Culture

fclass = 'arts_centre' Or fclass = 'cinema' Or fclass = 'museum' Or fclass = 'theatre'

gis_osm_pois_free_1: Recreation

fclass = 'golf_course' Or fclass = 'ice_rink' Or fclass = 'park' Or fclass = 'pitch' Or fclass = 'sports_centre' Or fclass = 'stadium' Or fclass = 'swimming_pool' Or fclass = 'track'

gis_osm_pois_free_1: Entertainment

fclass = 'bar' Or fclass = 'beverages' Or fclass = 'biergarten' Or fclass = 'cafe' Or fclass = 'fast_food' Or fclass = 'nightclub' Or fclass = 'pub' Or fclass = 'restaurant' Or fclass = 'food_court'