Summary of the Master Thesis submitted by Sina Bernhard:

Microgeographic Location Prediction: A Comparison of Software Firms in the U.S. and Germany

1 Introduction

This thesis delves into the distribution and concentration of the software industry in the U.S. and Germany, analyzing the impact of location factors on software company placement. The significance of location for business success, dating back to Johann Heinrich von Thünen's work in 1826, is evident in how economic activities concentrate in specific areas. The research highlights the recognized economic importance of software companies for regional development, emphasizes spatial proximity to other companies, spatial networks, research institutions and universities as a key indicator of spatially concentrated innovation activities.

Choosing densely populated regions for software company locations is deemed advantageous due to higher population diversity, promoting innovation and creativity. The research underscores the differences between the high-tech industry and traditional sectors, predicting a diminished importance of physical space in the information age.

The study advocates for a shift in research focus from larger to smaller spatial units to reveal previously unseen relationships between location factors and companies. The study also explores the impact of agglomeration advantages and contributes to understanding essential location factors for the software industry, specifically comparing the U.S. and Germany in a case study, within the broader context of location and cluster theories in economic geography. In doing so, the thesis addresses the following research questions:

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?

This study identifies geographic clusters and analyzes microgeographic factors via logistic regression on varied data, including agglomeration, infrastructure, socio-economic, topography and amenities. The primary goal is to determine the most influential factors in software company locations. Finally, the study validates these location prediction models and compares them with Kinne and Resch's (2018) findings.

2 Data & Methods

The data used comes from four different sources:

- Institutional Data: Around 15 million street-level geocoded firm observations for the U.S. from Infogroup (2016) and approximately 1.4 million for Germany from the Orbis database (Bureau van Dijk, 2022). Socio-economic data for Germany was obtained from infas360 (202), Nexiga (2022) and Real Estate Pilot (2022).
- Official administrative agencies: Population density data from the European Commission (Pesaresi et al., 2019). Socio-economic information for the United States from the U.S. Census Bureau's American Community Survey and the Department of Housing and Urban Development, based on the 2020 census data (United States Census Bureau, 2022). American life expectancy data gathered from the University of Wisconsin Population Health Institute's County Health Rankings 2018 (County Health Rankings & Roadmaps, 2022), while for Germany, it is obtained from the Federal Institute for Research on Building, Urban Affairs and Spatial Development (Bundesinstitut für Bau-, Stand- und Raumforschung, 2020).
- ArcGIS Living Atlas: The Global Multi-resolution Terrain Elevation Data (GMTED) with a 250-meter cell resolution grid created from the 2010 Global Multi-resolution Terrain Elevation Data (United States Geological Survey, 2022). "Global Fixed Broadband" data from Ookla for broadband Internet network coverage (Ookla, 2022).
- Open Street Map (OSM): OSM data was used for infrastructure variables in the absence of official agency geodata, including entertainment, culture, recreation, universities, airports, inter-state / highway and public transport stops (bus, metro, tram).

To correctly reflect the share of software companies on all companies, the U.S. classifications by the six-digit NAICS-code was chosen. The NAICS-codes used are: 511210, 518210, 541511, 541512, 541513, 541519, which are based on the primary activity of companies. In Germany, the NACE-codes used (6201, 62011, 62019, 62020, 62030, 62090) correspond to the American NAICS-codes.

Figure 1 gives an overview of the methods used:

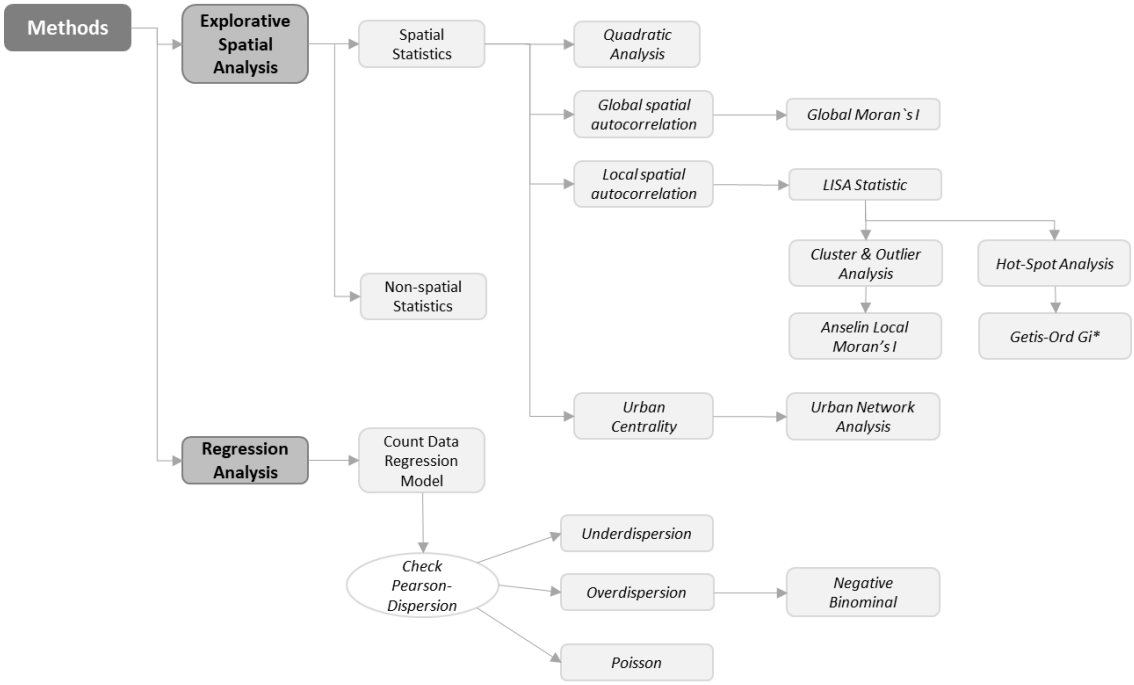

Figure 1: Overview of used methods

The study focuses on Exploratory Spatial Data Analysis (ESDA), emphasizing its role in exploring geospatial data using diverse visualization and descriptive methods (Abelairas-Etxebarria & Astorkiza, 2020). Spatial analyses with aggregated data face the Modifiable Area Unit Problem (MAUP), leading to inconsistent results and inaccurate causal relationships (Arauzo-Carod & Manjón-Antolín, 2012). The research highlights disparities in administrative levels in the U.S., affected by considerable differences in area sizes but relatively minor variations in population density. It attributes these differences to the U.S.'s low population density, particularly in contrast to Germany and the European Union, concentrating heavily in coastal regions.

Spatial analysis efficacy depends on consistent spatial patterns across the study area. Quadrat analysis transforms the two-dimensional distribution of business locations into a one-dimensional distribution (Illian et al., 2008). Spatial weight based on queen contiguity is recommended for cross-sectional analysis.

The Global Moran's I Analysis is used to measure global spatial autocorrelation, assessing the clustering of software companies' business locations (Kelejian & Prucha, 2001; Anselin, 1988). Local statistics, Local Moran statistics and Getis-Ord Gi* Analysis, complement Global Moran's I by identifying anomalous values or concentrations within specific areas (Anselin, 1995; Getis & Ord, 1992). These approaches help detect clusters, outliers, hot spots and spatial distribution patterns of software companies.

The study also employs Urban Centrality Analysis, applying network theory measures to evaluate the importance of traffic nodes in the street network, aiding in assessing urban accessibility and centrality and contributing to future location predictions. Urban network analysis reveals the hierarchical structure of cities and its relevance to location theory principles.

The study employs count data regression analysis to explore microgeographic factors' impact on software company placement and their correlation strength with location. It uses Poisson and Negative Binomial models for count data analysis (Hilbe, 2011).

3 Results

Exploratory Spatial Data Analysis

Descriptive statistics reveal significant variation in dispersion (DI: ratio of variance to mean distribution) across aggregation levels. The variance consistently exceeds the mean, indicating highly clustered and overdispersed patterns in software company locations. Comparing mean and median highlights a substantial proportion of zero values, exceeding expectations for a Poisson distribution, with data influenced by extreme outliers.

United States							
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
Germany							
Scale	**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 1: Descriptive Data - Software Companies

Spatial autocorrelation analysis, using Moran's I, indicates in Germany that autocorrelation increases with decreasing aggregation levels, thus highly dependent on grid size. In the U.S., software and other company autocorrelation is comparatively less pronounced at low aggregation levels and significantly decreases only at a 50 km² grid. Moran's I analysis is consistently highly significant on all scales (p-value ≤ 0.001), with a positive Z-value. The probability and standard deviations (>2.58 and <-2.58 respectively) support rejecting the null hypothesis of random spatial distribution, indicating less than a 1% chance that the observed clustered pattern is random.

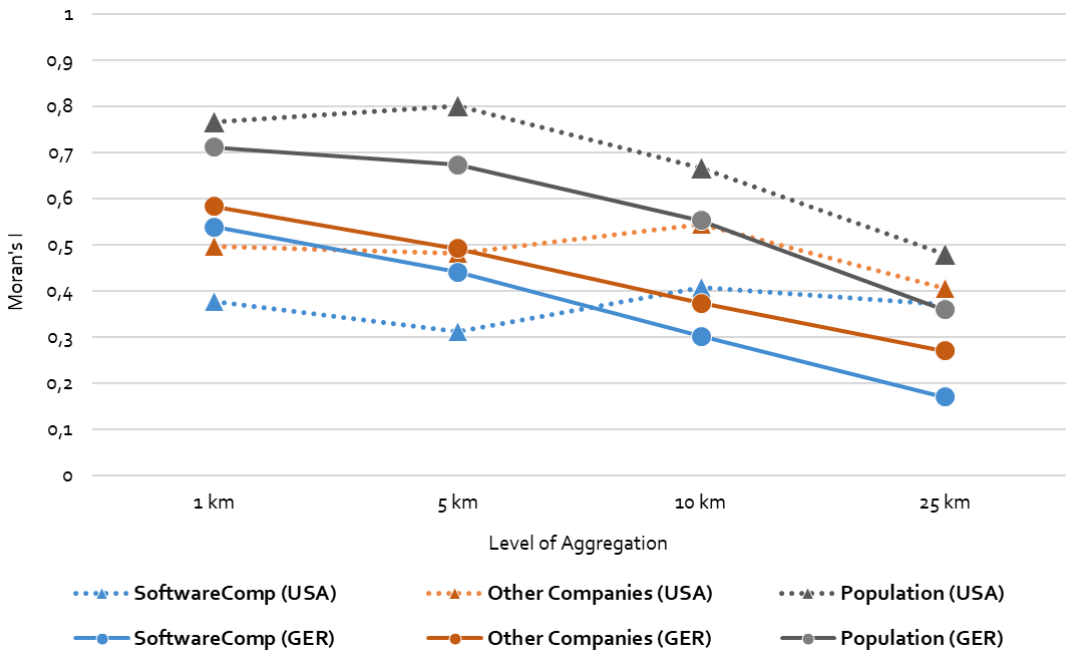

Figure 2: Moran's I of software companies, other companies and population

To comprehend the clustering of software companies in both countries, local spatial autocorrelation analyses were conducted. The Cluster and Outlier Analysis identified numerous statistically significant clusters (p≤0.05), marked by a high concentration of software company locations.

In Germany, major high-high (HH) clusters of software companies include Berlin, Munich, and Hamburg, with additional smaller clusters around Stuttgart, Frankfurt, and the Ruhr area. These clusters comprise 63% of German software companies. High-low (HL) outliers are scattered throughout the country, particularly near urban areas, while low-high (LH) outliers primarily exist on the outskirts of metropolitan areas. In the U.S., high-high (HH) clusters are prevalent in coastal regions, along the Great Lakes, and near large cities, constituting 77.7% of software companies. The map displays isolated high-low outliers (HL) mainly on the outskirts of large cities, occasionally in rural areas. Low-high (LH) outliers are concentrated in the peripheries of HH clusters and HL outliers, similar to Germany.

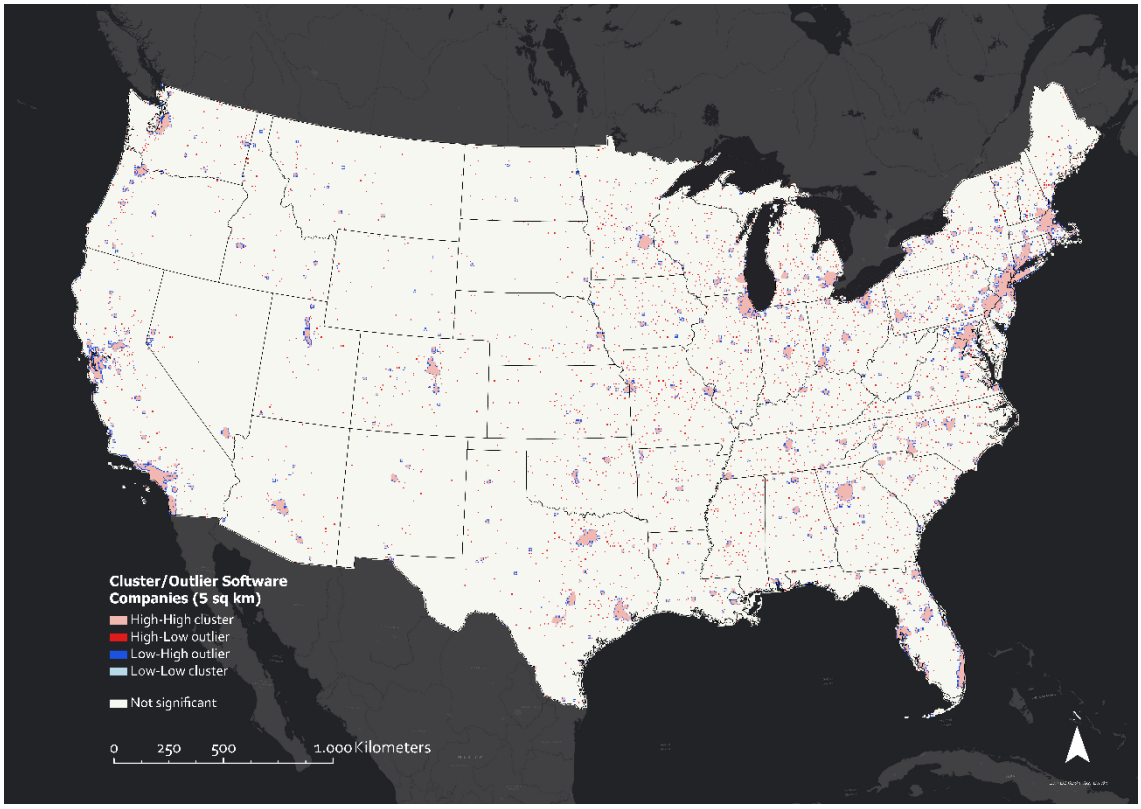

Figure 3: Cluster and Outlier - Software Companies in the U.S. (5 km² grid)

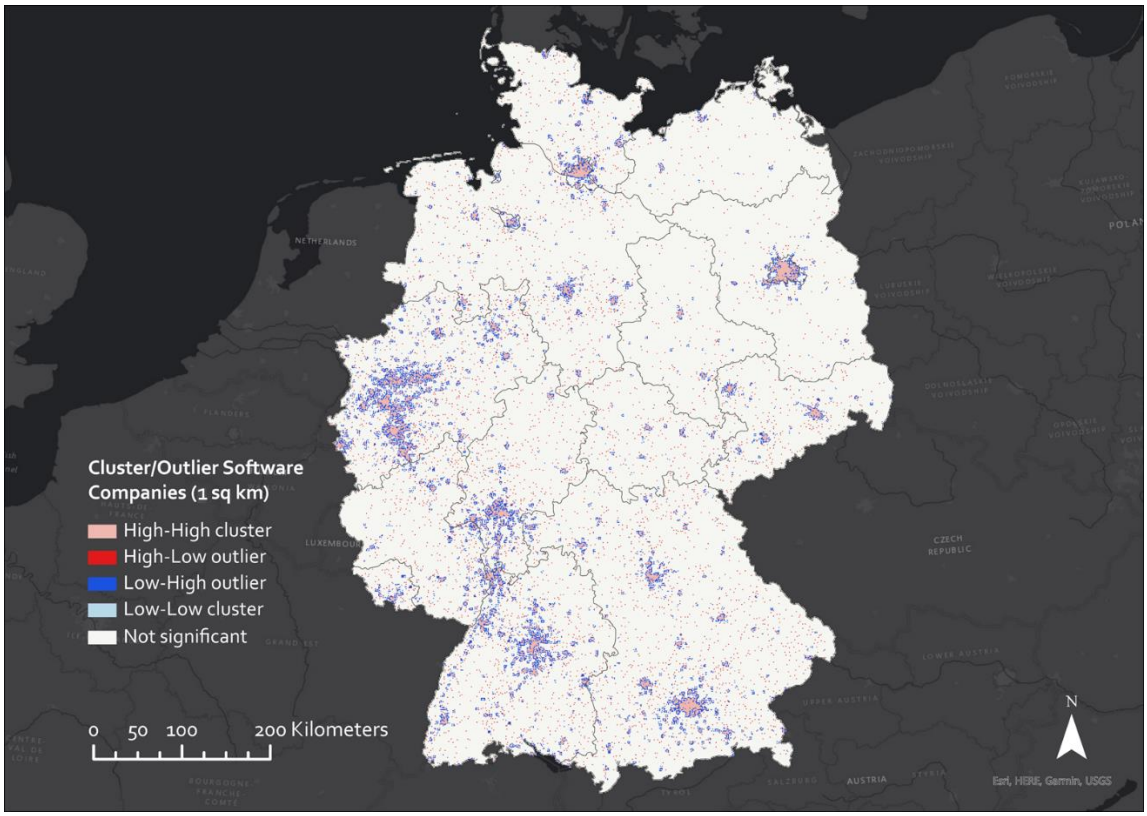

Figure 4: Cluster and Outlier - Software Companies in the Germany (1 km² grid)

The Getis-Ord Gi* analysis was used to verify and complement the Cluster and Outlier Analysis. The high Moran coefficients of the clusters are significant hotspots, with an above-average Z-value of 7.0, indicating a confidence level of at least 90%.

Count Data Regression

First, a a bivariate Spearman's rank correlation analysis, to assess monotonic relationships between variables, was conducted. Grid cells with missing values for location factors (2.3 % of total U.S. area, 1.04% in Germany) were excluded for data quality, without impacting analysis results. Our comprehensive model correlates software company density with 24 location factors per 1 km² grid. Due to data overdispersion, especially in the U.S., negative binomial regression (Maximum Likelihood Estimator (MLE)) was applied to interpret coefficients. This analysis explores the significance of location factors in influencing software company presence, explaining regional concentrations in the U.S. and Germany. Table 2 displays estimated coefficients as incidence-rate ratios (IRR).

		United States			Germany		
Location Factors	**Description**	**IRR**	**sig.**	**SE**	**IRR**	**sig.**	**SE**
Agglomeration location factors							
Company density	Number of local companies (in 10).	1.188	***	.0006	1.147	***	.0074
Company density²	Squared number of local firms (in 10).	1.000	***	.000	0.999	***	.0000
Software companies share	Proportion of software companies in the local business population (in %).	1.248	***	.0002	1.082	***	.0032
Population density	Population per cell (in 100).	1.031	***	.010	1.168	***	.0034
Population density²	Squared number of inhabitants per cell (in 100).	1.000	***	.000	0.998	***	.0000
Street centrality	Street (network) density calculation (1). High value = High density	1.006	***	.000	1.001	***	.0000
Universities	Distance to the nearest university (in km).	0.98	***	.0009	0.994	***	.0009
Research institutes	Distance to the nearest research institute (in km).	0.928	***	.002	0.994	***	.0009
Infrastructure location factors							
Network coverage broadband Internet	Average latency (upload / download speed) (in Mbps). High value = high internet speed	1.001	***	0.0001	1.008	***	.0001
Interstate / Highway	Distance to nearest highway / interstate (in km).	0.995	***	.0003	0.986	***	.00012
Airport	Distance to nearest main civil airport (in km).	0.993	***	.0004	1.002	***	.0004
Public transport	Weighted count of public transport stops.	0.956	***	.0014	1.005	***	.0013
Socio-economic location factors							
Salary	Monthly household income (median) (in 100 EUR).	1.000	***	.000	1.080	***	.0026
Educated workforce	Proportion of employees with a university degree (in %).	1.022	***	.0005	1.004	***	.0017
Student rate	Proportion of students in the local population in %.	0.993	***	.0008	0.994	***	.0023
Business tax	(U.S.) Coperate tax rates fixed by the states | communities (in %). (GE)Municipal business rate (in 100) fixed by the municipality. High values = high rates.	1.013	***	.0017	1.024		.0181
Life expectancy	Average life expectancy of the population (in years).	1.040	***	.0024	1.146	***	.0111
Average age	Average age (median) of the population.	0.987	***	.0010	1.012	***	.0038
Unemployment rate	Proportion of unemployed in the working-age population (in %).	1.007	***	.0017	0.982	***	.0049
Migration background	Proportion of people of non-U.S. | non-German nationality in the total population (in %).	1.003	***	.0003	1.005	***	.0014
Amenities location factors							
Recreation	Number of recreational. community and sports facilities.	1.012	***	.0004	1.012	***	.0011
Culture	Number of cultural sites and facilities.	1.004		.671	0.982		.0151
Entertainment	Number of dining. nightlife and general entertainment facilities.	0.972	***	.0012	0.995	*	.0027
Other							
Terrain	Average slope or gradient (in degree). High values = hillside location	0.701	***	.0123	0.994	**	.0028

*Table 2: Location factors and estimated coefficients (IRR) with robust standard errors (SE) - */**/*** indicate significance at 10/5/1 per cent, respectively.*

Interpretation of Regression Coefficients

Following Kinne & Resch (2018), the squared location factors of company and population density was incorporated in the analysis, based on the idea that urbanization can lead to congestion effects. Both the U.S. and Germany show an inverted U-shaped influence of density on the location economy, indicating agglomeration shadows. In both countries, urban network centrality slightly boosts the number of local software firms. A significant proportion of software companies among local businesses significantly increases the presence of additional local software firms, suggesting that clusters stimulate the formation of more software firms in the same location.

In knowledge-intensive sectors like software, proximity to knowledge hubs is crucial. A slight decrease in local software companies occurs as the distance to research institutes and universities increases, observed in both countries. The trend applies to highways, but not to airports and public transport in both countries. High-speed Internet availability (mbps) has minimal influence.

Proximity to universities has a positive impact, but a high local student population has a negative effect. University graduates in the local workforce increase software firms. Despite higher incomes for university degree holders, there's no (strong) effect on new software firms in the U.S. and Germany. Surprisingly, higher state corporate income taxes show a positive effect on software firms (Germany not significant). A significant migrant population increases local software companies. High life expectancy is linked to more software firms, except in the U.S. where an older population negatively impacts them. In the U.S., a rise in the unemployment rate correlates with an increase in software companies, unlike in Germany, which seems implausible.

Recreational facilities are positive for the number of software company in both countries, whereas entertainment has a negative impact. Cultural facilities show a positive effect in the U.S., but a negative in Germany, though not significantly in either. These amenities may impact cities or counties positively but may not enhance immediate neighborhood appeal. Locations with slopes strongly deter software company settlement.

Model comparison

The NB MLE model exhibits the best fit based on GoF indicators (log likelihood, BIC, AIC) for both countries, especially in Germany. However, AIC and BIC are not absolute measures of quality. In the U.S., Pseudo R^2 values show good explanatory power across all models. In Germany, the Poisson model has the best fit, while the quadratic NB has the least. RMSE values differ significantly, with the Poisson model providing the best prediction quality. All models, with the full set of predictors, significantly outperform the null model according to highly significant Omnibus Tests ($p < 0.001$).

United States				
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

Germany				
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 3: Goodness of fit (GoF) (Method: Fisher, Scale parameter measure: Deviance, Chi-squared statistic: Wald, Estimator: Robust)

The graphs depict observed and predicted values for each model. Zero values are appropriately handled by all models in both countries. Models tend to underestimate predictions for cells with few software companies. In the U.S., the Poisson Model consistently underestimates predictions, while in Germany, it generally aligns well with observations, despite higher underestimation at low count cells. NB models initially underestimate low count cells but drastically overestimate high count cells. It seems the Poisson model is a better predictor at the 1 km^2 scale, supported by lower RMSE and higher Pseudo-R^2, despite less sensitivity in log-likelihood, AIC, and BIC to overestimation and underestimation.

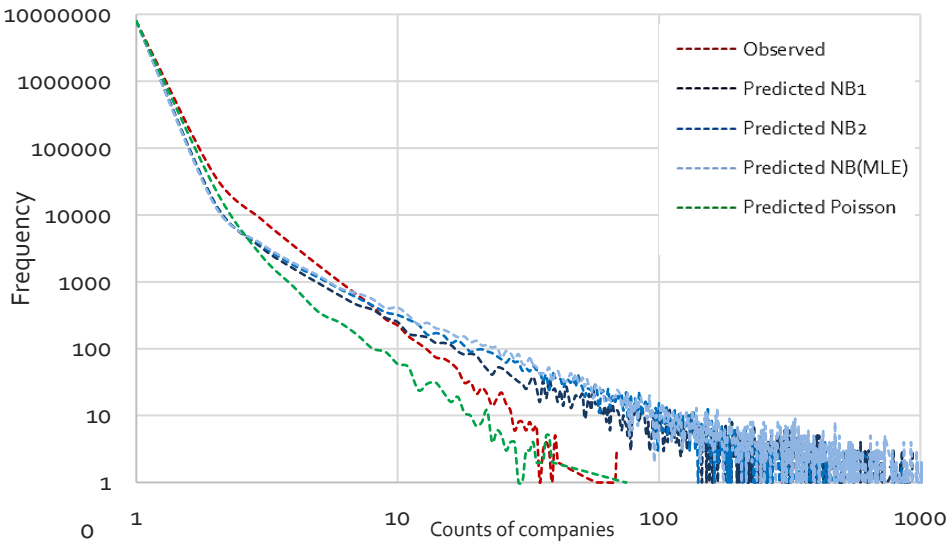

Figure 5: Frequencies of observed and predicted software firm counts - U.S.

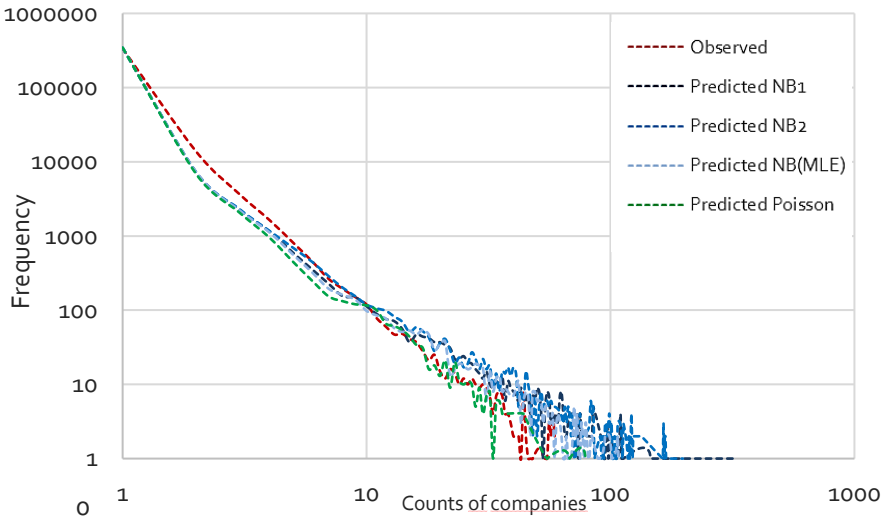

Figure 6: Frequencies of observed and predicted software firm counts - Germany

The maps show the prediction errors, deviation of the observed values from their expected values, is partly enormous. Grid cells with a higher number of software firms than predicted by the model are shown in red. The model underestimated those cells. Grid cells with colored in blue 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 the U.S., both the NB MEL and Poisson models exhibit prediction errors mainly in urban areas, with large-scale forecast errors, especially for the NB MEL model. The San Francisco Bay Area and New York City are highlighted, showing significant overestimation and underestimation. The systematic prediction error may be attributed to factors like spatial autocorrelation, omitted explanatory variables, and the unique urban structure.

In Germany, the models consistently underestimate areas with low count of software companies and overestimate those with a high count. The latter does not hold true for the Poisson model. Prediction errors occur in economically strong metropolitan areas, whereas rural regions show minor deviation from the observed values. The NB MEL model tends to overestimate in metropolitan areas but underestimates in less populated regions, while the Poisson model exhibits inconsistent patterns in various cities.

Overall, both countries share common findings of prediction errors in urban areas, but the specific patterns and causes differ, reflecting the unique characteristics of their urban structures and technological landscapes.

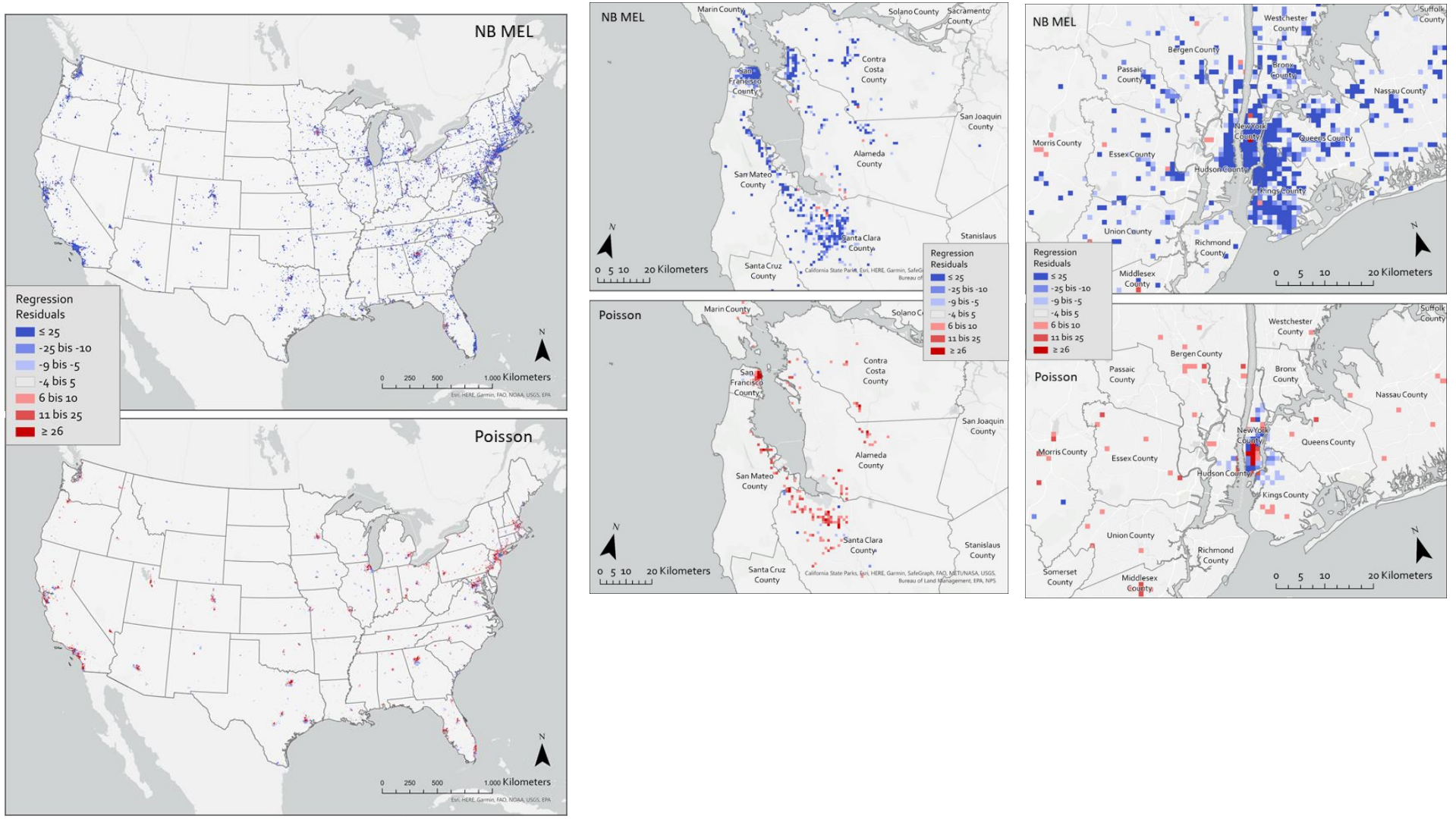

Figure 7: Regression residuals aggregated at 5 km grid - U.S (left). *Regression residual original at 1 km grid of the San Francisco Area (middle) and NYC (right)*

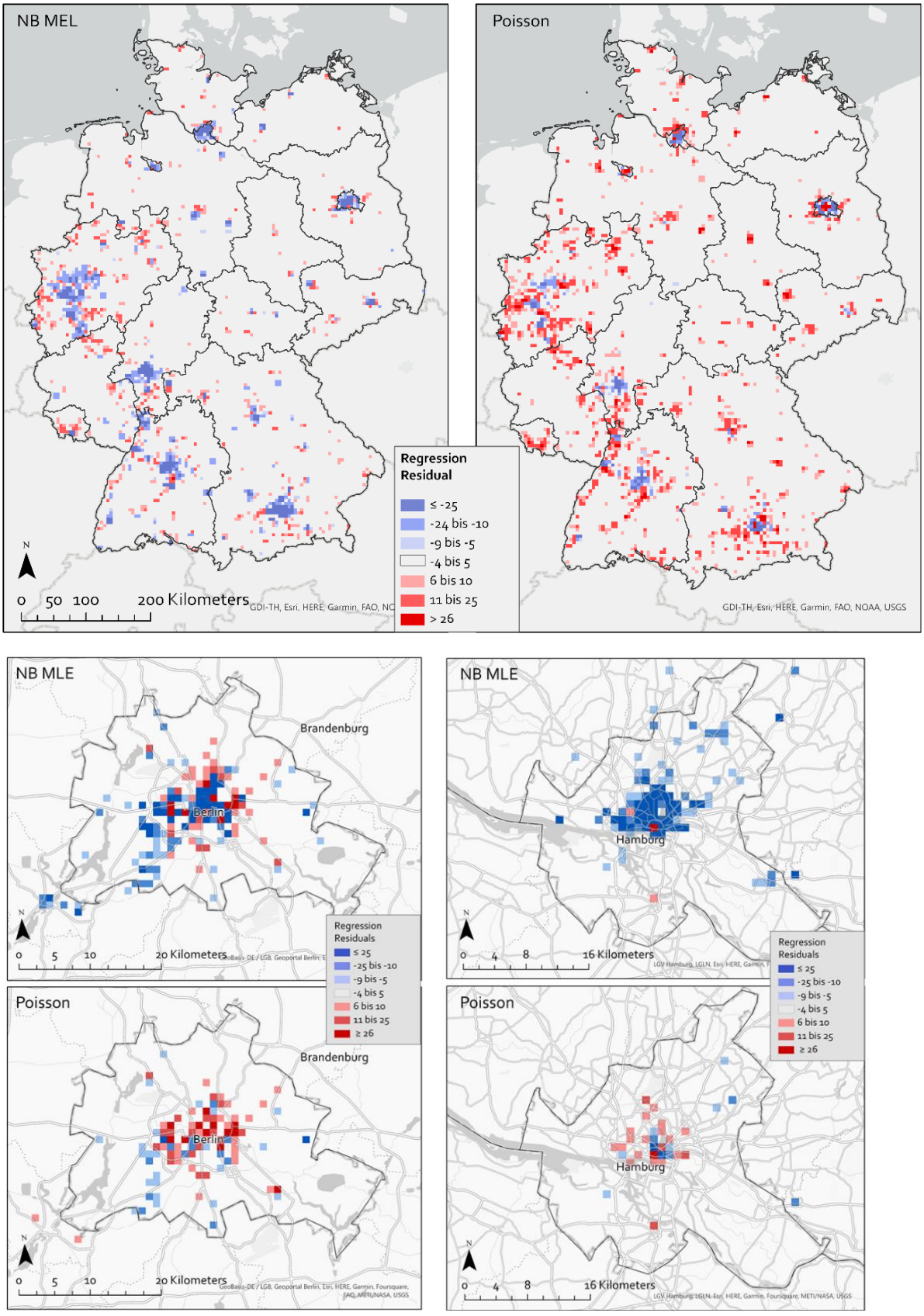

Figure 8: Regression residuals aggregated at 5 km grid in Germany (upper maps)

Regression residual original at 1 km grid of cities in Germany (Berlin, Hamburg) (lower maps)

4 Discussion

Agglomeration location factors

Company and population density reveal microgeographic differences in both countries. While a positive correlation between agglomeration and company location decisions is acknowledged, an "agglomeration shadow" effect is identified, indicating that the positive impact of density eventually turns negative after reaching a threshold. Squaring business and population density factors in Germany confirms an inverted U-shape effect.

Clusters of software companies positively influence the establishment of new companies in the same industry, with the U.S. emphasizing this factor as a critical predictor. Unlike Kinne and Resch's Urban Centrality Index, this study conducted a Centrality Analysis of the street network, revealing a link between road network densification, increased software companies, and potentially high mobility.

Additionally, the significance of proximity to universities and research institutes for software companies, facilitating knowledge spillovers. Regions with dense social networks and open labor markets, encourage entrepreneurship. A negative coefficient for universities and research institutes in both countries emphasizes the importance of proximity, as the number of software companies decreases with greater distance from these institutions.

Infrastructure location factors

In the U.S., there's a positive correlation between advantageous infrastructure and local software companies, excluding access to public transport, possibly due to differing population structure and mobility behavior. Germany, on the other hand, shows a positive relationship between infrastructure location factors, broadband network coverage, airport proximity, and public transport accessibility. These findings deviate partly from Kinne and Resch's (2018) results.

Socio-economic location factors

Well-educated employees significantly impact the presence of local knowledge-intensive companies in both countries, aligning with Kinne and Resch's (2018) findings. Positive relationships exist between the number of local software companies and factors such as people with migration backgrounds, average life expectancy, and proximity to universities. Contrary to previous studies, this analysis doubts the negative impact of high taxes on company location.

While higher local unemployment rates deter businesses in Germany, the U.S. exhibits a positive correlation, possibly influenced by the microgeographic level of analysis. Proximity to universities positively affects software companies, but a high proportion of students in the local population negatively impacts, in line with Kinne and Resch (2018).

Amenities location factors and other

This study categorized amenities into recreation, culture and entertainment. A significant positive correlation with Recreation and a negative correlation with Entertainment were found in both countries. Culture did not have a significant effect, making it challenging to pinpoint the influential amen-

ity. The analysis also considered terrain, finding a significant negative relationship between the average slope and software company location decisions in both countries, consistent with Kinne and Resch's (2018) findings.

Comparison and discussion of model adequacy

This study chose Negative Binomial and Poisson Regression models for analyzing software company data due to their suitability for over-dispersed data. Despite the NB-MEL model's better performance based on AIC, BIC and log-likelihood, significant differences in RMSE favor the Poisson model at the microgeographic level. Unlike Kinne and Resch's study, this research didn't encounter prediction errors in zero values, possibly due to excluding unsuitable areas in advance. The study acknowledges the potential benefits of zero-inflated models and suggests the Hurdle Model for handling excess zeros.

Concerns about socio-economic data heterogeneity are deemed less significant, given the detailed aggregation level. Multicollinearity is recognized as a potential issue, and the correlation matrix is suggested for diagnosis. Endogeneity problems, such as simultaneity and omitted variable biases, are acknowledged, particularly in assessing the causal relationship between local amenities and software companies.

The study emphasizes the challenge of operationalizing location factors at the microgeographic level, noting the sensitivity of scales and potential errors introduced by data aggregation. The impossibility of creating a single set of homogeneous spatial units is highlighted, suggesting that some location factors may be more meaningful at larger levels of aggregation. Arauzo-Carod and Manjón-Antolín's approach to address the Modifiable Areal Unit Problem (MAUP) is mentioned, using spatially lagged variables.

5 Conclusion

This study analyzes software company distribution in Germany and the U.S. using geocoded firm data at the street level. Employing spatial exploratory data analysis (ESDA), the research identifies 24 predictor variables influencing location decisions, inspired by Kinne and Resch's study. The microgeographic approach successfully utilizes low-level aggregated data. Key findings include:

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

Both countries exhibit global spatial autocorrelation, with Germany showing stronger clustering at the microgeographic level, while the U.S. displays strong clusters not only at this level but also beyond, potentially due to larger metropolitan areas. Cluster outlier and hotspot analyses identify significant spatial clusters in metropolitan areas of both countries, decreasing significantly outside these areas.

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

Agglomeration factors, population density, street centrality, proximity to universities, research institutes, access to broadband internet, and proximity to highways consistently influence software company numbers in both countries. Socioeconomic factors, including an educated workforce, life expectancy, and a larger proportion of the population with a migrant background are crucial for software company development in both countries. Differences emerge in location factors related to public transport stops and airport proximity. In terms of amenities, the recreation location factor significantly positive impacts local software company counts in both countries, while the factor entertainment has a negative effect. Terrain, specifically steeper gradients, is associated with a significant reduction in the number of software companies in both Germany and the U.S.

The microgeographic prediction model, utilizing 24 location factors, performs well in both countries. It effectively addresses excess zero values but tends to underestimate in low count cells and overestimate in high count cells (excluding the Poisson Model).