REGION
The Journal of ERSA
Powered by WU

Volume 11, Number 2, 2024, 63–97
DOI: 10.18335/region.v11i2.530

journal homepage: region.ersa.org
ISSN: 2409-5370

# Varying size and shape of spatial units: Analysing the MAUP through agglomeration economies in the case of Germany

**Rozeta Simonovska**[1], **Egle Tafenau**[1]

[1] University of Goettingen, Goettingen, Germany

Received: 13 February 2024/Accepted: 16 September 2024

**Abstract.** When an analysis over a specific geographic area is performed, the way that area is divided into regions can affect the outcome of the analysis. Results obtained based on different geographic units can be conflicting. This issue is known as the Modifiable Areal Unit Problem (MAUP). The objective of this paper is to understand the extent to which the regional setting influences the results of an analysis with spatially aggregated variables, with a focus on agglomeration effects, in the case of Germany. Relying on a sample of manufacturing firms over 7 years we estimate a fixed effects model to explain the firm-specific total factor productivity in dependence of region-based agglomeration variables. We simulate 1000 regional settings of Germany on three scales and overtake thereby some characteristics of the administrative units, which are used as the baseline. We infer that the spatial scale and shape matter in the case of Germany.

## 1 Introduction

Most regional studies rely on the administrative regions as the standard geographical units. It is well known that the results might vary considerably if an analysis is conducted at different regional aggregation levels. Moreover, if geographic units are constructed on other ground than administrative borders, the results can be conflicting to the ones gathered while using the administrative areas. The issue of statistical results being influenced by the choice of the geographical setting is known as the Modifiable Areal Unit Problem (MAUP).

The MAUP has been analysed in different contexts over the years starting from analysis of correlation coefficients (Gehlke, Biehl 1934, Yule, Kendall 1950, Openshaw, Taylor 1979, Arbia 1989) to different multivariate analyses including agglomeration economies (Briant et al. 2010, Andersson et al. 2016, Békés, Harasztosi 2018). The results on the MAUP have been varied depending on the methods used or the country, especially when analysing agglomeration economies, since they are aggregate variables based on regions.

Industrial agglomeration refers to firms locating in specific geographic area from which they can benefit. In this paper, we focus on two types of agglomeration economies: localisation and urbanisation. The localisation externalities, also known as Marshall-Arrow-Romer (MAR) externalities (Marshall 1890, Arrow 1962, Romer 1986), arise if firms locate in close proximity to other firms from the same industrial sector. This enhances the transfer of industry-specific knowledge and extends the pool of labour with skills that are relevant for this industry. The urbanisation or Jacobs externalities (Jacobs

1969) refer to the advantages from locating in an area with a lot of firms from other industrial sectors. Urbanised locations are often diverse and are believed to contribute to the innovation potential of firms because of a high variety of knowledge and ideas.

Germany has been underexplored, both with respect to the MAUP as well as to agglomeration effects. While Ehrl (2013) is a prominent exception with respect to the agglomeration analysis on different administrative levels based on plant level data, further information is needed on how serious the MAUP is in the case of Germany. We aim to fill this gap.

Germany has a unique federal system, where political decisions that affect economic activities can be taken at all administrative levels from municipalities, over districts (Kreise, in the European Nomenclature of territorial units for statistics NUTS3 regions), governmental regions (Regierungsbezirke, NUTS2) and federal states (Bundesländer, NUTS1) to the central government. With information about the sensitivity of the results of an agglomeration analysis with respect to the underlying regional units the policy for the economic development of regions can be designed more precisely with respect to their spatial impact.

Unlike some other European countries, Germany has very strict privacy laws, which often leads to aggregating data at a higher level. This means that in many cases, analysing data at a smaller regional level is not possible. Therefore, understanding the effect of aggregated data is very important.

The rest of the paper is organised as follows. In the next section an overview of the relevant literature is given. In Section 3, the zoning systems are presented. After that, Section 4 describes the set up of the model, the construction of the variables, the estimation procedure of the model and the data used for the analysis. Empirical results are presented and discussed in Section 5. The final section concludes.

## 2 Literature review

### *2.1 The MAUP*

The analysis of the MAUP goes back to Gehlke, Biehl (1934). They were the first ones to take a close look at the problem of varying size of correlation coefficients in answer to a change in the scale of the underlying regions. Also in the subsequent years the authors that examined the MAUP focused on the correlation coefficients, for example Yule, Kendall (1950) and Openshaw, Taylor (1979). The latter expanded the study of the MAUP from the *scale problem*, 'the variation in results that may be obtained when the same areal data are combined into sets of increasing larger areal units of analysis', to the interconnecting *aggregation problem*, 'any variations in results due to alternative units of analysis where $n$, the number of units, is constant' (Openshaw, Taylor 1979, p. 108).

In the last few decades the MAUP has also received substantial attention in multivariate analysis. For example, Briant et al. (2010) analysed agglomeration economies, spatial concentration and trade determinants in France. They relied on three zoning systems: the administrative, a grid and a partly random system. Moreover, each of the systems was looked at on three different scales. Briant et al. (2010) conclude that in the case of France the underlying regional system is not as relevant for the estimation results as the model specification.

For other countries this result is not confirmed in the context of an agglomeration analysis. For instance, Andersson et al. (2016) observe differences in Sweden while analysing square grid data at different scale. Furthermore, Békés, Harasztosi (2018), tackling both the scale and the aggregation problem, conclude that in the case of Hungary the composition of regions is as important as the model specification.

Therefore, the literature thus far has delivered contradictory results on the MAUP in the context of agglomeration effects. Békés, Harasztosi (2018) argue that the diverging results may result from the underlying regional structure of the analysed countries: while France has a relatively homogeneous regional structure, this is not the case for Hungary.

### 2.2 Agglomeration economies

The impact of agglomeration has been of interest to researchers since the end of the 19th century (Marshall 1890), resulting in hundreds of studies. This research has by summarized in many different meta-analyses (Rosenthal, Strange 2004, Beaudry, Schiffauerova 2009, Melo et al. 2009, De Groot et al. 2016).

Most studies report a positive effect of agglomeration economies. For example, Henderson (2003) and Martin et al. (2011) find mostly significant positive effects of localisation for the manufacturing firms in the US and France, respectively. However, Melo et al. (2009), which analyse estimates of 34 studies where the estimation of agglomeration economies is done by using a production function or a wage model, find that there is a positive reporting publication bias. De Groot et al. (2016) find that analysis outcomes have changed over the years and that more recent studies are more likely to report negative results for diversity.

The literature suggests that conclusions on the effects of agglomeration depend on the country, regional level, time period, industries as well as the methodology used in the analysis.

## 3 Zoning systems

The baseline zoning system in the analysis of this paper corresponds to the administrative regions. Specifically, we use the NUTS (Nomenclature of territorial units for statistics) regions according to the NUTS 2016 classification at 3 scales. NUTS regions are areas created by Eurostat[1] in collaboration with each European country for statistical purposes. We use three scales: the NUTS1 regions (Figure 1a) correspond to the 16 federal states of Germany (Bundesländer), the NUTS2 regions (Figure 1b) to the 38 governmental regions (Regierungsbezirke) and the NUTS3 regions (Figure 1c) to the 401 districts (Kreise).

In order to test for the MAUP, we simulate additional 1,000 regional settings for each of the three administrative types of regions. We consider the investigation of the different regional shapes as relevant since most of the German NUTS regions are not homogeneous, in regard to both population and area, especially NUTS1 regions whose borders have historic origin. For constructing the simulated regions, we use the German municipalities (LAU2), 11,271 in total.[2] We build on previous research from Openshaw, Taylor (1979) and Openshaw (1977) for the USA and Briant et al. (2010) for France. However, in addition to keeping the number of regions in accordance with the number of administrative units, we set some constraints to achieve artificial regional systems that account for the characteristics of the heterogeneous regional system of Germany. Moreover, a nested structure is constructed.

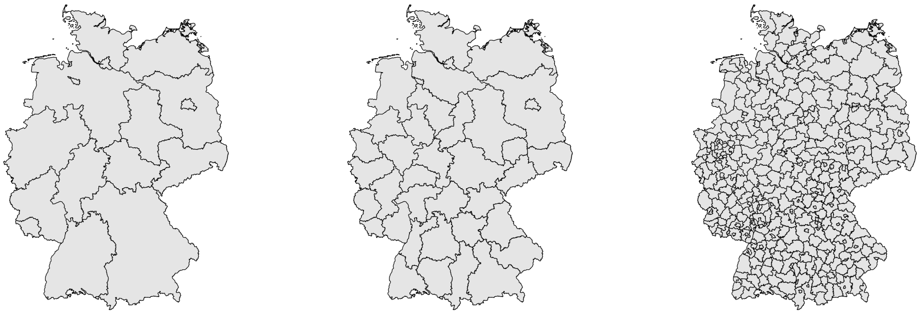
The procedure for creating the artificial regions starts by setting a random seed. After that the municipalities are aggregated to obtain SMALL regions, comparable to NUTS3 regions. Next, the SMALL regions are aggregated into MEDIUM regions, comparable to NUTS2 regions. Finally, the MEDIUM regions are aggregated into LARGE regions, comparable to NUTS1 regions. This way, we obtain a nested structure of each scale, in the same way that the administrative regions are nested in each scale. In addition, during the aggregation process we use some restrictions like population size in order to produce regional structures which resemble the real world administrative regions. The German NUTS regions with the lowest population size, as well as the limits provided by Eurostat for each NUTS level are used as benchmarks. A detailed description of the procedure is given in the Appendix A.

Additionally, we also simulate 1,000 regions at each scale where no nesting structure or any population restrictions are implemented. The only condition is to have regions with approximately similar size.

Figure 2 shows the standard deviations of the population in the regions obtained from the two types of simulation approaches, as well as a box-plot of the population of

[1] Statistical Office of the European Union.

[2] Local administrative units (LAU) are a subdivision of NUTS3. The number as well as the borders of municipalities vary over the years. We rely on municipality borders from 31.12.2016. The population of German municipalities is given in Table A1.

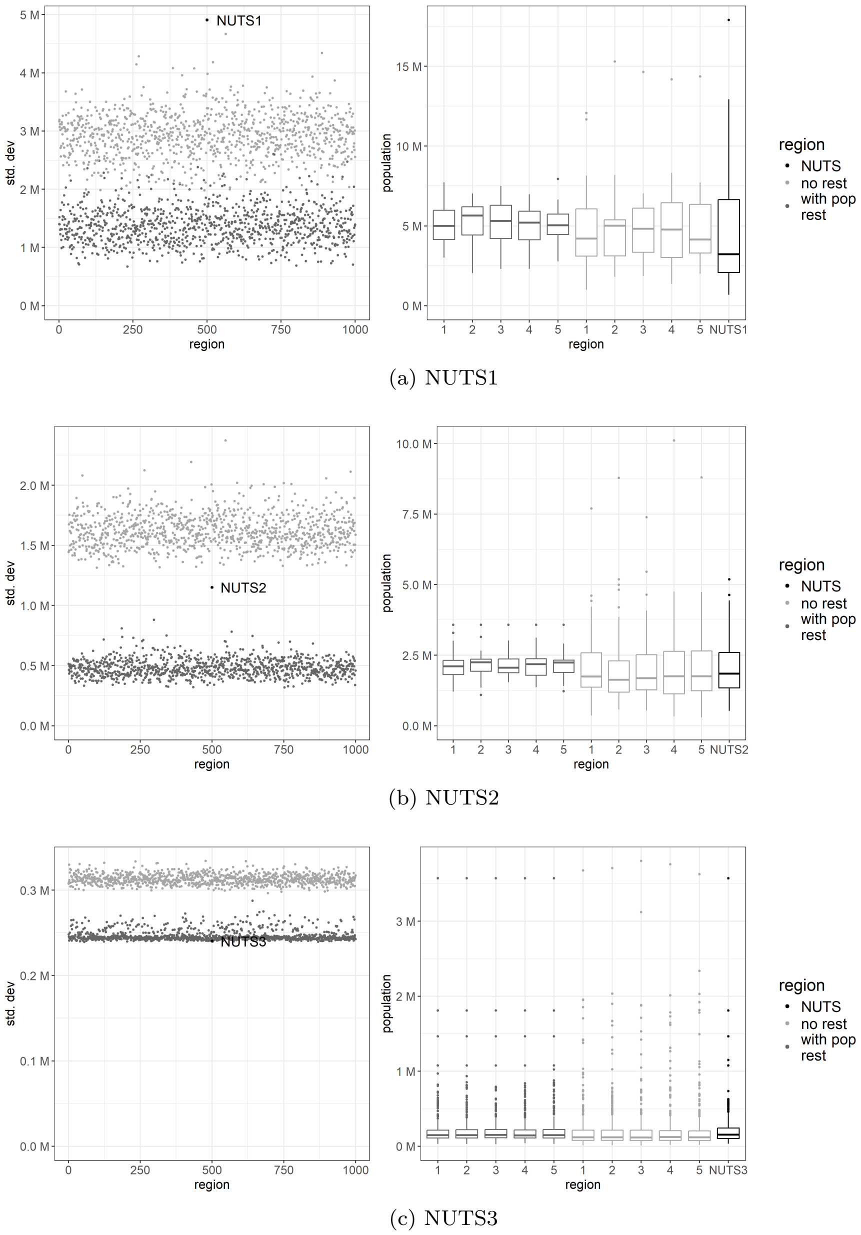


(a) NUTS1

(b) NUTS2

(c) NUTS3

Figure 1: NUTS regions in Germany



(a) NUTS1

(b) NUTS2

(c) NUTS3

Figure 2: Standard deviation and population distribution of simulated German regions

the first 5 regional settings from the two simulation approaches and the corresponding administrative region at each of the 3 scales. We see that introducing a constrain for the regional population size in the simulation algorithm leads to a distribution of the population in the simulated regions which resembles the one of the districts (Kreise), whereas the simulated regions without a population restriction have higher standard deviation. However, at NUTS3 level, even for the algorithm with population restrictions, it was not possible to reduce the standard deviation for the simulated regions below the one of the administrative regions. The reason for this is that there are some municipalities with a very large population size, such as Berlin and Hamburg on one hand, and on another, in order to achieve faster convergence of the algorithm, the minimal population of the simulated regions had to be reduced in a few instances. When it comes to population size in the higher administrative levels, even though the variability in population is reduced when the population restriction is used, because of the non-balanced structure of administrative regions, both types of simulated regions (with or without a population restriction condition) have different variability in population compared to the actual administrative regions.

We further simulated 1,000 artificial regional settings for each of the two simulation approaches for a few other European countries based on data from Eurostat, from 2020. We find that the position of the standard deviation of the administrative NUTS regions compared to the simulated regions with or without population restrictions varies between countries[3]. For the Netherlands and France, for example, in most cases we find that the regions simulated without population restrictions are more similar to the administrative regions. In Hungary, on the other hand, the simulated regions with population restrictions have more similar characteristics to the corresponding administrative regions. Therefore, a simulated regional setting such as the grid-based regional setting is less likely to produce a different results compared to the administrative regional in the case of the Netherlands and France. We expect, that based on our simulations, there should be no effect of the MAUP, more specifically the aggregation or shape effect, on the Netherlands and France, while we expect to find distortions related to the MAUP in Hungary, as reported by Békés, Harasztosi (2018). Therefore, we expect that the MAUP analysis on Germany would be more similar to the one of Hungary, compared to MAUP analysis on France.

## 4 Data and model

### *4.1 Data*

The main source of data for this analysis is the AMADEUS database. This database, managed by the Bureau van Dijk (BvD), contains information on over 21 million companies across Europe with more than 1.4 million of those having their headquarters in Germany. Information about the number of employees, tangible fixed assets, cost of materials and value added for firms in the manufacturing sector is downloaded from this database. Also the location information (city, ZIP code, NUTS1, NUTS2, NUTS3), the number of branches, NACE Rev. 2 industry code, category of company variable, yearly turnover and the consolidation code are retrieved from AMADEUS. We use the time period 2009-2015. A detailed overview of the data selection and sample construction is given in the Appendix B.

All monetary variables used in this paper, given in thousands of Euros, are deflated with an industry-specific deflator at 2-digit industry level. The deflator for the firm-level value added and the cost of materials is a producer price index. The total fixed assets are deflated with an asset price deflator. Both deflators are obtained from Eurostat.

The focus of this paper is on analysing firms in the manufacturing sector (NACE Rev.2 2-digit codes from 10-33). Data availability was the main reason for the choice of this sector, as well as the possibility for better comparability with previous research. However, because of a small number of firms we excluded the sectors manufacture of tobacco products (NACE 2-digit code 12), manufacture of coke and refined petroleum

[3] See Figures A2 - A6 in the Appendix for additional box-plots for each country.

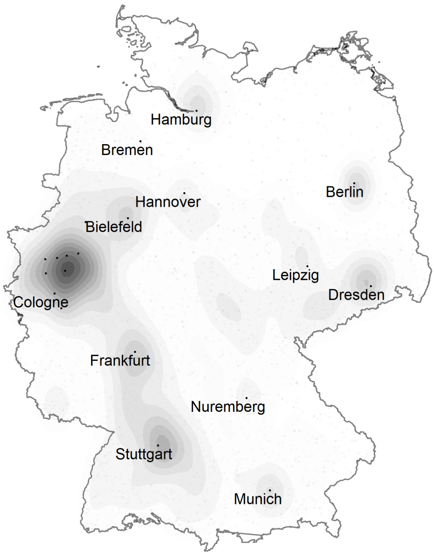


Figure 3: Heat map of the 7317 firms

products (NACE 2-digit code 19) and manufacture of leather (NACE 2-digit code 15). This leads to a sample of 7,317 firms. The spatial distribution of the firms is shown in Figure 3[4]. We notice that most of manufacturing is located in the most populated cities and we can also observe a large cluster in western Germany.

For calculating the agglomeration variables a larger sample of 54,529 firms is used, forming an unbalanced panel of 328,881 observations (see Figure B1 in Appendix B). An additional source of data for the independent variables is the Federal Statistical Office of Germany, from where we gather information on the number of employees in each 2-digit sector in Germany for the period 2009-2015.

For the construction of the artificial regions we use vector data files from The Federal Agency for Cartography and Geodesy (BKG). The control variables rely on the data from the German Employment Agency and the Federal Statistical Office of Germany.

### *4.2 Model*

In order to estimate the strength of the agglomeration externalities, we set up a firm-level model for total factor productivity (TFP). In this model, the TFP of a firm depends on two agglomeration variables. The first variable, following the MAR theory, measures the extent of *localisation* of the firm's industrial sector in the home region of the firm. The second variable, *urbanisation*, is used for estimating the Jacobs externalities: it measures how large or diverse the region is bar the sector of the firm under observation.

Formally, the model can be expressed as

$$TFP_{it} = \alpha_1 loc_{it} + \beta_1 urb_{it} + e_{it} \tag{1}$$

where $TFP_{it}$ is the log of total factor productivity of the firm $i$ in the year $t$, $loc_{it}$ and $urb_{it}$ denote the localisation and urbanisation variables, respectively. $e_{it}$ denotes an error term.

For both explanatory variables we use two alternative definitions to check for the robustness of the results, because previous studies, for example Beaudry, Schiffauerova (2009), have shown that the results of an agglomeration analysis depend on the measurement method of the two aspects of agglomeration. We test these two models, one

[4] The 20 most populated cities are shown. For better legibility labels are not displayed for some cities in western Germany (including: Düsseldorf, Dortmund, Essen, Duisburg, Bochum, Wuppertal, Bonn and Münster).

with absolute and one with relative measures, with and without control variables, against simulated regions with and without population restrictions.

We rely on a two-step procedure in which firm level TFP is estimated in the first step based on a production function and the model (1) in the second step.

### 4.2.1 Estimation of TFP

In order to estimate TFP, it is assumed that output of the firm $i$ in the year $t$ follows the Cobb-Douglas production function

$$Y_{it} = A_{it} K_{it}^{\mu} L_{it}^{\nu} \tag{2}$$

where $Y$ denotes the output, $A$ total factor productivity, $K$ capital and $L$ labour. Symbols $\mu$ and $\nu$ represent the capital and labour elasticities, respectively. The production factors and output are understood as physical quantities in a production function. For the practical implementation, however, we will rely in case of the output and capital on their monetary values that are deflated by appropriate price indices. Specifically, the output of a firm is quantified as the value added of the firm, the capital as the value of the total fixed assets. The labour variable is quantified as the number of employees. Unlike the other variables in the equation (2), $A$ is unobservable to the researcher.

Applying natural logarithms to (2) yields the log-linear equation:

$$y_{it} = \beta_0 + \mu k_{it} + \nu l_{it} + \epsilon_{it} \tag{3}$$

where the lower case letters denote the corresponding logarithmic variables from equation (2). The logarithm of the firm-specific TFP is understood as a composition of the general level of productivity in the society ($\beta_0$) and the firm-level deviations from that ($\epsilon_{it}$). Accordingly, $\log(A_{it}) = \beta_0 + \epsilon_{it}$.

Since factor input quantities $k_{it}$ and $l_{it}$ tend to be correlated with the error term, the endogeneity problem appears in the model (3). This implies that the Ordinary Least Squares (OLS) estimator is biased. To account for this issue, different approaches have been used in the literature. Those include instrumental variable (IV), General Method of Moments (GMM), fixed effects (FE) and semi-parametric approaches.

In the following we focus on semi-parametric approaches as according to Van Beveren (2012) the alternative methods tend to be biased or have a poor performance. Among the most widely used are the two-step semi-parametric approaches of Olley, Pakes (1996), Levinsohn, Petrin (2003) and Ackerberg et al. (2015). Subsequently, Wooldridge (2009) combined those with the GMM methodology.

Contrary to the Olly-Pakes (OP), Levinsohn-Petrin (LP) and the Ackerberg-Caves-Frazer (ACF) methods, TFP is estimated in one step in the Wooldridge (2009) approach. As explained in Van Beveren (2012), this enables standard calculation of robust standard errors instead of bootstrapping. Moreover, the resulting estimators are more efficient than in the two-step approaches of OP and LP as the latter cannot account efficiently for heteroscedasticity and serial correlation in the error terms. Additionally, the Wooldridge estimator accounts for the estimation problem in the first stage, which was noted by Ackerberg et al. (2015). Because of those advantages, we use the Wooldridge approach in this paper.

After estimating firm-level TFP, the results are used in the model (1) for the dependent variable in order to estimate the strength of the agglomeration effects.

### 4.2.2 Agglomeration variables

The localisation and urbanisation variables $loc_{it}$ and $urb_{it}$ for the firm $i$ in period $t$ are dependent on the firm's region of location and its sector. The localisation variable characterises the economic volume of the firm's sector in its region of location, the urbanisation variable captures the volume of economic activity in a given region in general. In both cases, the indicators can rely on the absolute or relative magnitude.

The localisation variable varies with time and firm over sectors and regions, $loc_{it} = loc_{it}^{sr}$ ($i \in A_t^{sr}$, where $A_t^{sr}$ denotes the set of firms located in region $r$ from industrial sector $s$ in year $t$, see Martin et al. 2011). The urbanisation variable is constant for all firms for a given time period, sector and region: for each $i \in A_t^{sr}$, $urb_{it} = urb_t^{sr}$.

As the absolute magnitude of a sector in a region, we calculate $loc_{it}^{rs}$ as the number of employees employed by firms that belong to the same sector $s$ and operate in the same region $r$ as the firm $i$, excluding the employees from the firm $i$. Formally,

$$loc_{it}^{sr} = \ln(empl_t^{sr} - empl_{it}^{sr} + 1), \tag{4}$$

where $empl_t^{sr} = \sum_{j \in A_t^{sr}} empl_{jt}^{sr}$ is the number of all employees in the region $r$ in the industrial sector $s$ in the year $t$.

The respective urbanisation measure is calculated as the number of employees in year $t$ in all other sectors different from the sector $s$ to which the firm $i$ belongs to, i.e.

$$urb_{it}^{sr} = \ln(empl_t^{r} - empl_t^{sr} + 1) \quad i \in A_t^{sr}, \tag{5}$$

where $empl_t^r$ is the total number of employees in the region $r$ across all sectors and firms.

In the following, *Model I* will denote the model (1) with the localisation and urbanisation variables as defined in equations (4) and (5), respectively.

As mentioned above, also relative measures can be used for describing the extent of localisation and urbanisation. Accordingly, we analyse the MAR spillovers alternatively with the help of the location quotient, denoted by $locLQ_{it}^{sr}$. It is defined as the share of the own industry employment in a region relative to its national share, i.e.

$$locLQ_{it}^{sr} = \ln\left(\frac{\dfrac{empl_t^{sr} - empl_{it}^{sr} + 1}{empl_t^r + 1}}{\dfrac{empl_t^s}{empl_t}}\right). \tag{6}$$

Thus, the location quotient shows if the industry $s$ is overrepresented in the region $r$ compared to the industry's national share.

As for the Jacobs' externalities we use the diversity index $div_{it}^{sr}$ as a relative measure. It is defined as the inverse of the quotient with the Hirschman-Herfindahl index of industry concentration in a region in the numerator and the Hirschman-Herfindahl index of industry concentration at national level in the denominator:

$$div_{it}^{sr} = \ln\left(\frac{\sum\limits_{s' \neq s}\left(\dfrac{empl_t^{s'r}}{empl_t^r - empl_t^{sr}}\right)^2}{\sum\limits_{s' \neq s}\left(\dfrac{empl_t^{s'}}{empl_t - empl_t^{s}}\right)^2}\right)^{-1} \qquad i \in A_t^{sr}. \tag{7}$$

Accordingly, in the following analysis our *Model II* is expressed as

$$TFP_{it} = \alpha_2 locLQ_{it} + \beta_2 div_{it} + e_{it}. \tag{8}$$

### 4.2.3 Control variables

We expect that some firm-level characteristics that are related to a firm's TFP are omitted from the models. In order to avoid the omitted variable bias, we include firm level fixed effects $\phi_i$ and time fixed effects $\mu_t$ in the Models I and II.

Since further factors can have an effect on the productivity of a firm, we include two control variables in the Models I and II. Both variables vary with time and region.

First, we control for the employment size in neighbouring regions by including a variable for the market potential:

$$mp_{it}^r = \ln\left(\sum_{r'} \frac{allEmpl_{r't}}{D_{r,r'}}\right) \tag{9}$$

where $mp^r_{it} = mp^r_t$ for each $i \in A^r_t$ is the market potential for the region $r$ at time $t$, $allEmpl_{r't}$ is the number of employed people subject to social insurance in the neighbouring regions $r'$ at time $t$ and $D_{r,r'}$ is the Euclidean distance between the centroids of the regions $r$ and $r'$. A region is considered to be a neighbouring region of the region $r$ if they share a border.

The second control variable, transported goods $tg^r_{it}$, measures the accessibility of a region and is constant for all firms for a given time period and region: for each $i \in A^r_t$, $tg^r_{it} = tg^r_t$. It is calculated as the logarithm of the share of the sum of all transported freight in 1,000 t in a region based on airports, sea and river ports, highways and rail, over the total area of the region[5], i.e.

$$tg^r_{it} = \ln\left(\frac{air_{rt} + water_{rt} + road_{rt} + rail_{rt}}{area_r}\right) \tag{10}$$

Therefore, after including these two variables we get two new models:

$$TFP_{it} = \alpha_3 loc_{it} + \beta_3 urb_{it} + \gamma_3 mp_{it} + \delta_3 tg_{it} + \phi_i + \mu_t + e_{it} \tag{11}$$

and

$$TFP_{it} = \alpha_4 locLQ_{it} + \beta_4 div_{it} + \gamma_4 mp_{it} + \delta_4 tg_{it} + \phi_i + \mu_t + e_{it} \tag{12}$$

which we will refer to as the *Model III* and the *Model IV*, respectively.

## 5 Results

The estimation results of the Models I and II (equations (1) and (8), respectively, including firm and time fixed effects) for the three scales of administrative regions are presented in Table 1. We find no statistically significant effect of the agglomeration variables on TFP.[6]

The simulations of the 1,000 sets of artificially created regions based on population restrictions confirm this result: in most of the cases the parameter estimates of the localisation and urbanisation/diversity variables are insignificant, see Figures 4 and 5. However, in the settings with LARGE and MEDIUM regions the parameter estimate of the location quotient is statistically significant for more than 50 or 40 % of the regional settings, respectively, at the 10 % significance level in the Model II (Figure 5). In all those cases, the parameter estimate is negative. In addition, the mean (as well as the median) of the parameter estimates for the location coefficient shifts towards zero when the regions are scaled down. The results for the simulated regions with no population restriction are similar[7].

As for the urbanisation/diversity variable, in both models the corresponding parameter estimate only in a few cases is found to be statistically significant. Also, here the statistically significant estimates are mostly negative.

We conclude that localisation and urbanisation/diversity are not significant factors in determining a firm's productivity in Germany[8]. However, belonging to an industry that is overrepresented in its wider region of location compared to the industry's national share might have a negative effect to a firm's productivity.

If a model is estimated only for one regional setting, statistically significant parameter estimates might be obtained. This can easily lead to a strong conclusion by an analyst.

[5] The data and the description for the transported goods variable is given in Appendix B.3.2.

[6] We also estimated alternative specifications of the models. Especially, if the time fixed effects are excluded, the parameter estimates for the urbanisation/diversity variables are statistically significant. However, if the model is augmented with control variables like market potential or the quantity of goods that are transported on the infrastructure of the region, the parameter estimates of the urbanisation/diversity variable turn insignificant. Accordingly we conclude that the model without time fixed effects is misspecified.

[7] Appendix C, Figures C3 and C4

[8] The significance of the agglomeration variable can be affected by the model specifications. As previously mentioned, if no time fixed effects are included, the probability of finding significant estimates increases. Furthermore, other model specifications, such as using robust standard errors also affects the amount of significant values, that is, the significance decreases when robust standard errors are used.

Table 1: Estimation results for the Models I and II

| | NUTS1 | NUTS2 | NUTS3 |
|---|---|---|---|
| | *Model I* | | |
| localisation | -0.0250 | 0.0264 | -0.0102 |
| | (0.0796) | (0.0481) | (0.0174) |
| urbanisation | -0.0879 | -0.1221 | -0.0043 |
| | (0.3318) | (0.1099) | (0.0376) |
| firm FE | yes | yes | yes |
| year FE | yes | yes | yes |
| observations | 25,676 | 25,676 | 25,676 |
| R-squared | 0.00006 | 0.00037 | 0.00003 |
| | *Model II* | | |
| localisationLQ | -0.0545 | -0.0135 | -0.0115 |
| | (0.0361) | (0.0259) | (0.0129) |
| diversity | 0.0157 | 0.0209 | -0.0182 |
| | (0.0354) | (0.0312) | (0.0137) |
| firm FE | yes | yes | yes |
| year FE | yes | yes | yes |
| observations | 25,676 | 25,676 | 25,676 |
| R-squared | 0.00062 | 0.00012 | 0.00026 |

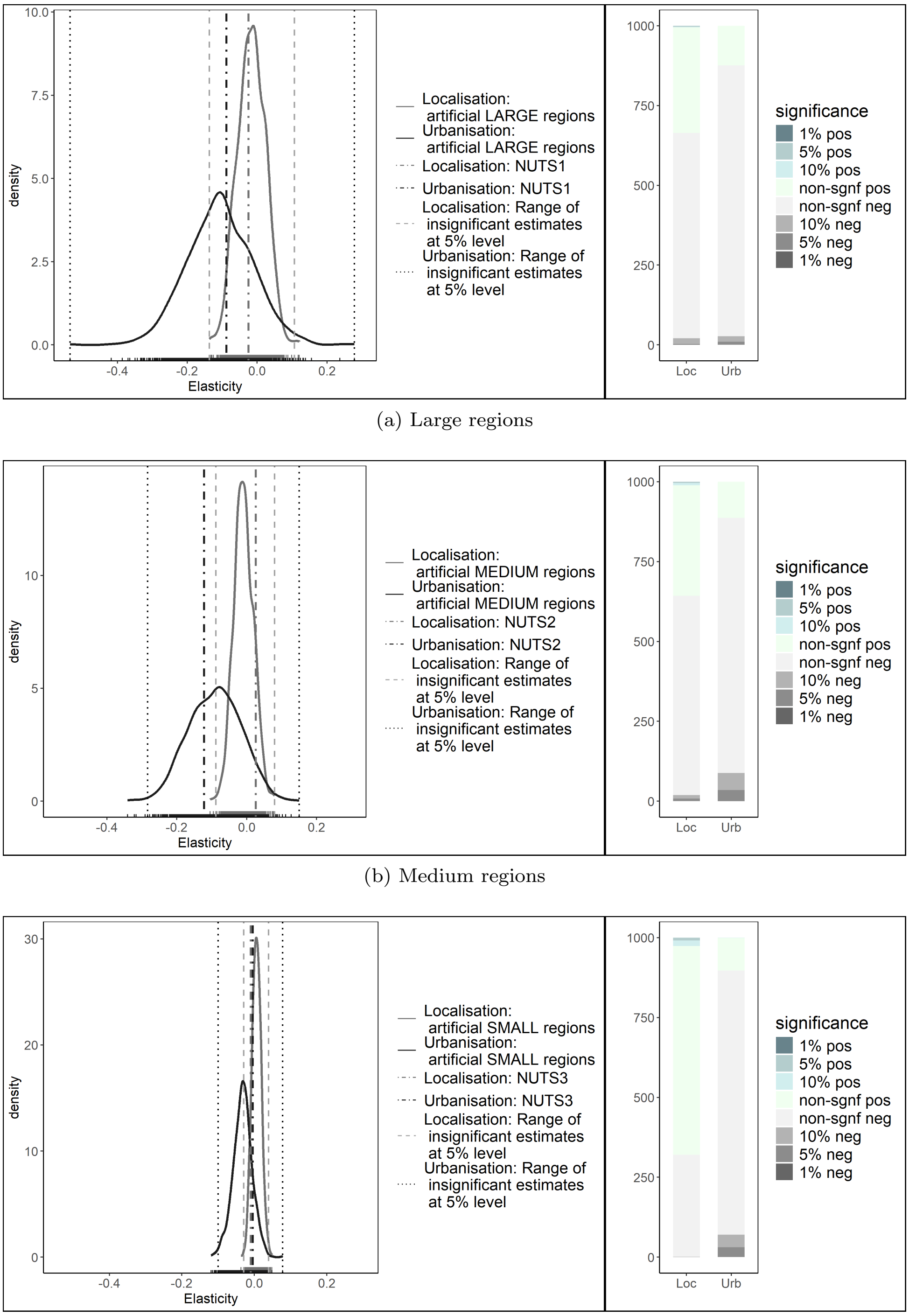
*Note*: Agglomeration variables are based on employment (equations (4)-(7)). Independent variables are standardised to have zero mean and standard deviation 1. Dependent variable is $\ln(TFP)$ estimated with the Wooldridge method. The time period is 2009-2015, the number of firms 7,317. Cluster robust standard errors at region level are given in parentheses.

Table 2: Parameter estimates in Models III and IV

| | NUTS1 | NUTS2 | NUTS3 |
|---|---|---|---|
| | *Model III* | | |
| localisation | -0.0302 | 0.0246 | -0.0095 |
| | (0.0770) | (0.0477) | (0.0174) |
| urbanisation | 0.0301 | -0.0978 | 0.0075 |
| | (0.2885) | (0.1085) | (0.0373) |
| market potential | 0.6274** | 0.4583* | 0.3437*** |
| | (0.2222) | (0.2350) | (0.0994) |
| transported goods | 0.2460* | 0.1173 | 0.2987* |
| | (0.1551) | (0.1667) | (0.1544) |
| firm FE | yes | yes | yes |
| time FE | yes | yes | yes |
| observations | 25,676 | 25,676 | 25,676 |
| R-squared | 0.00131 | 0.00132 | 0.00167 |
| | *Model IV* | | |
| localisationLQ | -0.0600 | -0.0159 | -0.0122 |
| | (0.0346) | (0.0253) | (0.0130) |
| diversity | 0.0026 | 0.0198 | -0.0186 |
| | (0.0324) | ( 0.0318) | (0.0135) |
| market potential | 0.6342** | 0.4881** | 0.3450*** |
| | ( 0.2538) | (0.2345) | (0.0994) |
| transported goods | 0.2591 | 0.1202 | 0.2972* |
| | (0.1538) | (0.1642) | (0.1534) |
| firm FE | yes | yes | yes |
| firm FE | yes | yes | yes |
| observations | 25,676 | 25,676 | 25,676 |
| R-squared | 0.00195 | 0.00121 | 0.00192 |

*Note*: Agglomeration variables are based on employment (equations (4)-(7)). Explanatory variables are standardised to have zero mean and standard deviation 1. Dependent variable is $\ln(TFP)$ estimated with the Wooldridge method. The time period is 2009-2015, the number of firms 7,317. Cluster robust standard errors at region level are given in parentheses.

*, ** and *** denote significance at 10%, 5% and 1% level, respectively.



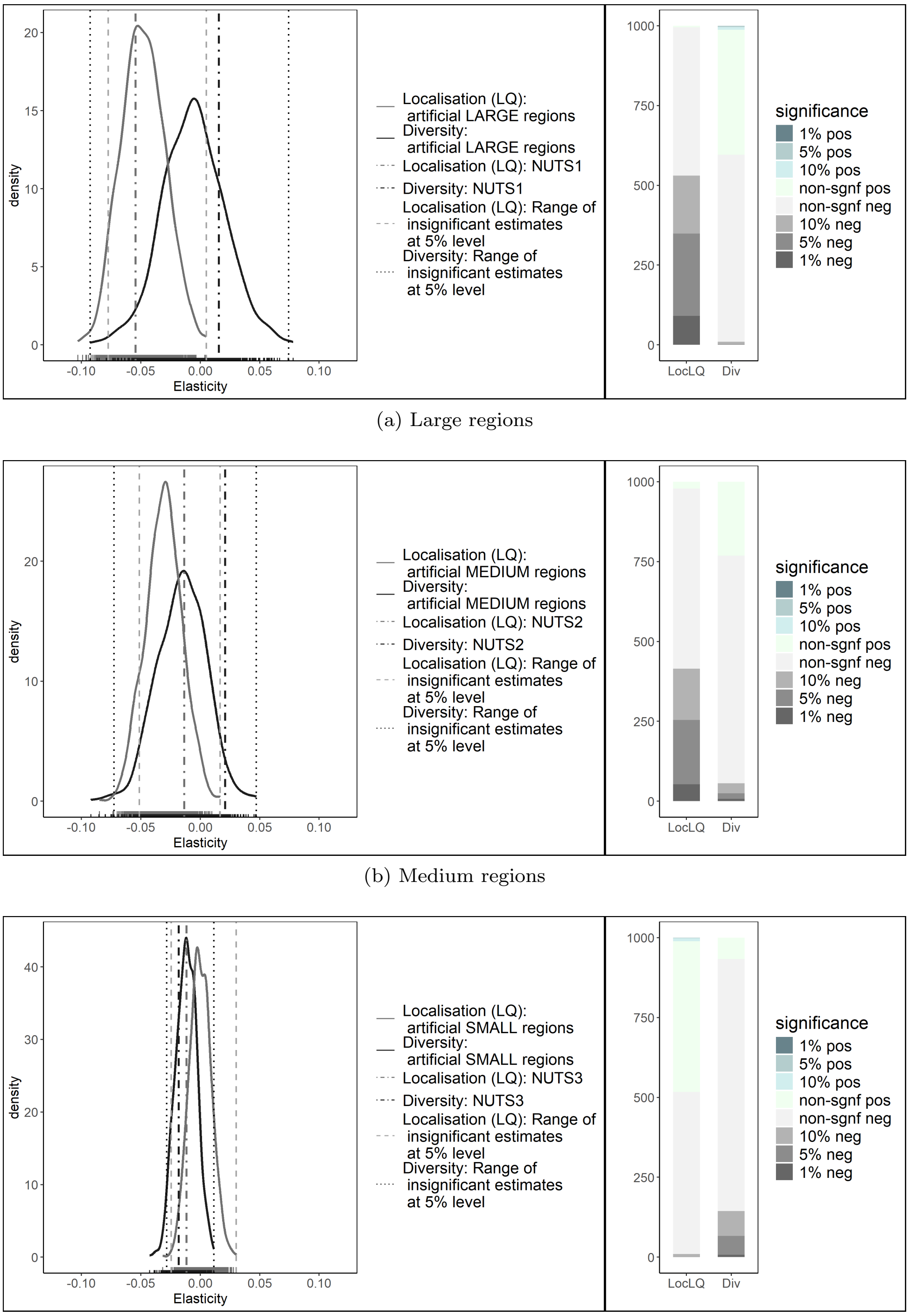
(a) Large regions

(b) Medium regions

(c) Small regions

Figure 4: Parameter estimates of the localisation and urbanisation variables in the Model I and the distribution of their significance

*Note*: The densities of the parameter estimates rely on 1,000 settings of artificial regions. The range of insignificant estimates at 5 % level covers all parameter estimates that are not statistically significant at 5 % significance level.



(a) Large regions

(b) Medium regions

(c) Small regions

Figure 5: Parameter estimates of the localisation and urbanisation variables in the Model II and the distribution of their significance

*Note*: The densities of the parameter estimates rely on 1,000 settings of artificial regions. The range of insignificant estimates at 5 % level covers all parameter estimates that are not statistically significant at 5 % significance level.

A careful analysis could reveal, however, that such a result arose only because of randomness: if 1,000 different regional settings are analysed, it is probable that statistically significant parameter estimates are found for one or a few of them even if the true parameter value is zero. One of those random cases could correspond to the administrative regions. Thus, the simulations with artificial regional settings help to assess the validity or strength of the conclusions obtained from a statistical model.

Examples for such a scenario are the extended Models III and IV. As revealed in Table 2, the parameter estimate of the market potential variable is for all three scales positive and statistically significant at 10% level, in the case of the NUTS3 regions even at 1% level. However, even though over 99% of the estimates are statistically significant at SMALL level, for MEDIUM regions in model III, in only around 55% of the regional settings the parameter estimate is statistically significant at 5% level (see Tables C3-C4 and Figures C1-C2 in Appendix C). For the transported goods variable, the mean of the 1,000 estimates from the simulated regional settings is far from the estimated coefficient from the administrative regions when looking at a higher aggregation level. However, at a smaller scale the mean of the 1,000 estimates and the transported goods estimate for the NUT3 regions are closer. Also both the administrative and the simulated regions suggest a higher probability of finding a positive significant effect of the transported goods variable at the smallest scale. Furthermore, a very high estimation uncertainty is revealed by the simulations. Therefore, simulating artificial regional settings helps to assess the validity of the results of a model that relies on regional data.

## 6 Conclusion

The goal of this paper was to understand the relevance of the underlying regional setting when analysing the effect of spatially aggregated variables such as localisation and urbanisation in the case of Germany. To achieve the goal, we looked at 1,000 artificially created zoning systems at three different scales and with two types of simulation methods and for each of the settings as well as for the administrative regional units we estimated regression models with Total Factor Productivity as the dependent variable and varying measures of localisation and urbanisation as explanatory variables.

As expected, the statistical significance of the localisation and urbanisation effects varies with the geographical settings. Based on the administrative regions, no significant results for the base models are found. However, the analysis of the artificial regional settings provides evidence of possibly adverse effects to TFP of a firm if the firm locates in a region with an over-proportional share of the firm's branch (as compared to the national average). This result holds only if sufficiently large regions are examined. The simulated results for the extended models confirm the results of the agglomeration estimates from the base models.

Though there is evidence for the MAUP if agglomeration effects are analysed in the context of Germany, we also find that the model specification is important – possibly even more important than the MAUP. For example, the choice of certain specifications in the model, such as removing time fixed effect can lead to significant results. Furthermore, the way of measuring the agglomeration variables localisation and urbanisation should be carefully considered.

In addition, we show that the estimation uncertainty rises with the aggregation level of the regions. This results is expected if some variables are defined at the level of the regions as aggregation leads to a loss of information. Therefore, small regional units should be preferred for an analysis of the effects of localisation and urbanisation. Moreover, the results for one regional level cannot necessarily be transferred to other regional levels. Accordingly, the policy implications of a regional analysis are reliable only if the goals of the corresponding policy measures are to be achieved at the regional aggregation level that was used in the underlying analysis.

## Acknowledgement

This research was supported by the German Research Association (DFG) Research Training Group 1644 “Scaling Problems in Statistics”, grant no. 152112243.

## References

Ackerberg DA, Caves K, Frazer G (2015) Identification properties of recent production function estimators. *Econometrica* 83: 2411–2451. CrossRef

Andersson M, Klaesson J, Larsson JP (2016) How local are spatial density externalities? Neighbourhood effects in agglomeration economies. *Regional Studies* 50: 1082–1095. CrossRef

Arbia G (1989) *Spatial data configuration in statistical analysis of regional economic and related problems.* Kluwer, Dordrecht

Arrow KJ (1962) The economic implications of learning by doing. *Review of Economic Studies* 29: 155–173. CrossRef

Beaudry C, Schiffauerova A (2009) Who’s right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy* 38: 318–337. CrossRef

Békés G, Harasztosi P (2018) Grid and shake: spatial aggregation and the robustness of regionally estimated elasticities. *The Annals of Regional Science* 60: 143–170. CrossRef

Briant A, Combes PP, Lafourcade M (2010) Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics* 67: 287–302. CrossRef

De Groot HL, Poot J, Smit MJ (2016) Which agglomeration externalities matter most and why? *Journal of Economic Surveys* 30: 756–782. CrossRef

Ehrl P (2013) Agglomeration economies with consistent productivity estimates. *Regional Science and Urban Economics* 43: 751–763. CrossRef

Gehlke CE, Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association* 29: 169–170. CrossRef

Henderson JV (2003) Marshall’s scale economies. *Journal of Urban Economics* 53: 1–28. CrossRef

Jacobs J (1969) *The Economies of Cities.* Random House, New York

Levinsohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies* 70: 317–341. CrossRef

Marshall A (1890) *Principles of Economics.* Mac-Millan, London

Martin P, Mayer T, Mayneris F (2011) Spatial concentration and plant-level productivity in France. *Journal of Urban Economics* 69: 182–195. CrossRef

Melo PC, Graham DJ, Noland RB (2009) A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics* 39: 332–342. CrossRef

Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1297. CrossRef

Openshaw S (1977) Algorithm 3: a procedure to generate pseudo-random aggregations of n zones into m zones, where m is less than n. *Environment and Planning A* 9: 1423–1428. CrossRef

Openshaw S, Taylor P (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed), *Statistical Applications in the Spatial Sciences*. Pion, London, 127–144

Romer PM (1986) Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037. CrossRef

Rosenthal SS, Strange WC (2004) Evidence on the nature and sources of agglomeration economies. In: *Handbook of regional and urban economics*, Volume 4. Elsevier, 2119–2171

Van Beveren I (2012) Total factor productivity estimation: A practical review. *Journal of Economic Surveys* 26: 98–128. CrossRef

Wooldridge JM (2009) On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters* 104: 112–114. CrossRef

Yule GU, Kendall MG (1950) *An introduction to the theory of statistics*. Charles Griffin, London

## Appendices

### A Creating the artificial regions

In order to create the artificial regions, we use German municipalities (LAU2), as the smallest administrative unit in Germany. However, over the 7 year period considered in our analysis, there have been a number of changes in the borders of many municipalities in Germany. Therefore, we use the municipality and NUTS stand from 2016. Furthermore, for the variables in which we use municipality level data[9], we use the municipality stand in 2016 and for the whole period of analysis, that is the municipality data is transformed to correspond to the stand in 2016.

Prior to starting with the simulation process, the neighbours list and the distance between municipality centroids are determined. Since neighbours are used for aggregating, we assigned the two closest municipalities as neighbours to the municipalities which do not have a shared border with any other municipalities (i.e. islands and Büsingen am Hochrhein, a German enclave surrounded by Swiss municipalities). Additionally, islands consisted of multiple municipalities are connected by mainland Germany, that is, a municipality from the island and the closest mainland municipality are considered as neighbours.

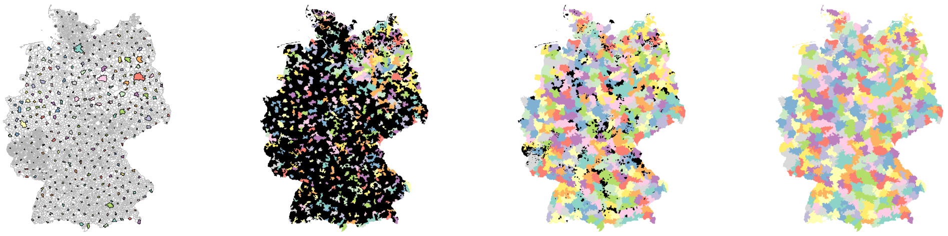
First, for each of the 1,000 settings a different starting seed is set in order to aggregate the 11,271 municipalities into 401 regions. From the total set of municipalities, 401 initial ones are selected. However, in Germany there are a number of municipalities with a large population, for example Berlin and Hamburg. These municipalities cause non-convergence of the algorithm if they are added to a region in a later step. To avoid this, all municipalities with a population above a threshold[10] are selected as part of the 401 initial regions. For these regions no additional municipalities will be added in later steps since they are over the population threshold. Next, one by one, the remaining starting municipalities are added to the initial 401 such that every other that is selected has to be at a certain predetermined distance from the previously selected starting municipalities. After the starting 401 municipalities are selected, if the population of that municipality in 2016 is smaller than the population of the smallest district in Germany (34,270 inhabitants), a neighbouring municipality is added to the initial one and they are aggregated. The step is repeated until the population in the aggregated region is larger than the threshold or the region has run out of neighbours to be added. If there are aggregated regions whose population is smaller than the threshold[11] and they have no available neighbours to be added to that region, then the procedure is restarted and the initial seed is increased by one. Next, the threshold for adding neighbouring municipalities to a region is increased step by step. First it is set to be the total population of Germany divided by the number of regions (401) and afterwards in following steps it is duplicated. In a next step, the threshold is set as the maximum population for a NUTS3 region (800,000 inhabitants). In the final step, any remaining municipalities which are not assigned to an aggregated region, are then added to a neighbouring region. With this procedure the first scale of the artificially created regions is completed. These regions are referred to as SMALL regions and they correspond to the NUTS3 regions (Kreise).

Similarly to the creation of SMALL regions, we use population properties of NUTS2 in Germany and the general intervals from the NUTS classification to create the artificial regions corresponding to NUTS2 regions. Because NUTS regions are nested in each other, the goal is for the artificial regions to be nested as well. Therefore, already created SMALL regions are used for creating the next scale, MEDIUM regions. As in the procedure for small regions where municipalities were used as starting point, in the creation of MEDIUM regions we use SMALL regions as the starting point for aggregation. Firstly, the initial 38 regions are selected such that they are at a minimal predetermined distance. Next, similar to the procedure for SMALL regions, neighbours of the initial regions are added until the threshold is exceeded. At the end, any remaining

[9] Market potential.

[10] 800,000 for creating the first scale, based on the NUTS criteria for NUTS3 regions.

[11] The starting threshold for this step is 80% of the population of the least populated NUTS3 region (27,542) and with every iteration we reduce it with the final being 50% of the population (17,214).



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4

Figure A1: Simulation steps

non-assigned SMALL regions are added to aggregate regions. The final scale, LARGE regions, are created in a similar procedure, aggregating the 38 MEDIUM regions into 16 regions.[12]

The procedure for simulating regions with no population restrictions involves fewer steps. As in the procedure using population restrictions, the initial step is choosing $N$ initial regions at a certain distance, where $N = 401$ for NUTS3, $N = 38$ for NUTS2 and $N = 16$ for NUTS1. No nesting structure is implemented, therefore each scale starts with municipalities. In the next step, if the area of the initial $N$ regions (municipalities) is less than 70% of the average area of the corresponding level, then neighbouring regions are added until the area is over this condition or there are no more neighbouring regions to be added. In the following step, similarly to the previous one, the average area is used as a condition, however it is increased to 90% of the average area. In the final step, the remaining regions are added to the aggregated regions.

Figure A1 shows the process of simulating regions.

1. Set a different starting seed for each simulation. Select a random region. Repeat selecting regions until $N$ is reached, such that the centroid of each region which is selected as next is at a distance of at least $0.7 * \sqrt{(totalarea/n)}$ of previously selected regions, Figure A1a.

2. If the population size of the $N$ selected regions is smaller than the smallest population size of the corresponding NUTS region (i.e. NUTS3 for SMALL regions), then merge neighbouring regions with the starting region. If there are regions which have no remaining neighbouring regions, but still do not fulfil the criteria for having population size larger or equal to the smallest corresponding administrative region, then return to Step 1 and increase the seed by one (after 3 seeds-increase the threshold is reduced to 80% of the population size of the smallest corresponding administrative region, then to 66,7% and to 50%), Figure A1b. This step is only performed when population restrictions are used. When there are no population restrictions, then this step is omitted.

3. Use average population/area to add neighbouring regions to regions from *Step 2*(*Step 1* when no population restriction is used), Figure A1c.

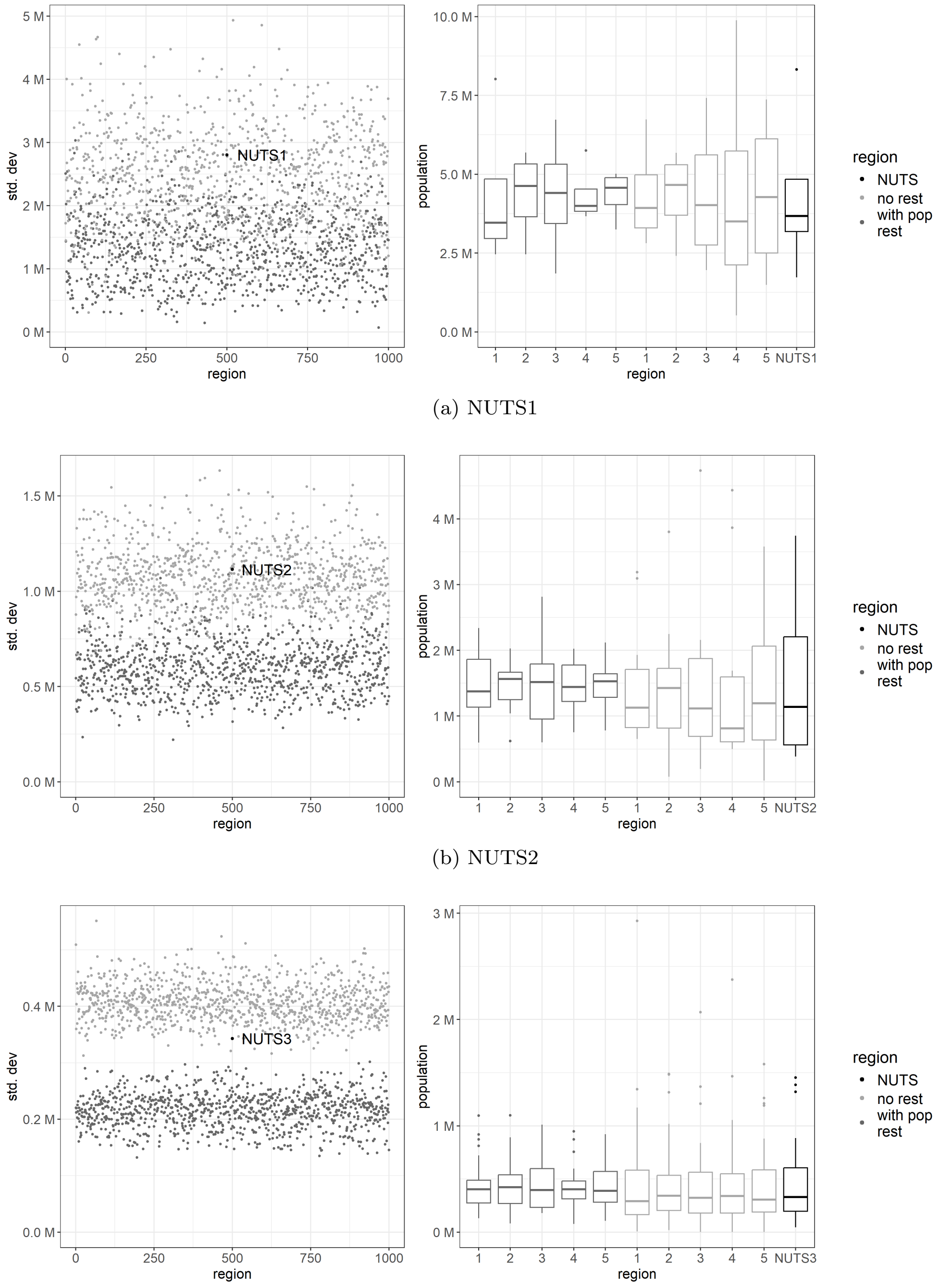
4. Any remaining regions (regions in black in Figure A1c) are assigned to one of the neighbouring regions from the $N$ groups, Figure A1d.

[12] The R package for creating simulated regions *RegionSim* is available on https://github.com.

Table A1: Regional structure of Germany

| | *number of regions* | | | |
|---|---|---|---|---|
| | NUTS1 | NUTS2 | NUTS3 | municipality |
| | Baden-Württemberg | 4 | 44 | 1,103 |
| | Bavaria | 7 | 96 | 2,237 |
| | Berlin | 1 | 1 | 1 |
| | Brandenburg | 1 | 18 | 417 |
| | Bremen | 1 | 2 | 2 |
| | Hamburg | 1 | 1 | 1 |
| | Hesse | 3 | 26 | 430 |
| | Mecklenburg Western Pomerania | 1 | 8 | 753 |
| | Lower Saxony | 4 | 45 | 969 |
| | North Rhine-Westphalia | 5 | 53 | 396 |
| | Rhineland-Palatinate | 3 | 36 | 2,305 |
| | Saarland | 1 | 6 | 52 |
| | Saxony | 3 | 13 | 426 |
| | Saxony-Anhalt | 1 | 14 | 218 |
| | Schleswig-Holstein | 1 | 15 | 1,112 |
| | Thuringia | 1 | 23 | 849 |
| Total: | 16 | 38 | 401 | 11,271 |
| | *population* | | | |
| min | 678,753 | 528,728 | 34,428 | 9 |
| | (Bremen) | (Trier) | (Zweibrücken) | (Gröde) |
| max | 17,894,969 | 5,191,702 | 3,574,830 | 3,574,830 |
| | (North Rhine-Westphalia) | (Dusseldorf) | (Berlin) | (Berlin) |
| mean | 5,157,908 | 2,171,751 | 205,801 | 7,462 |

*Note*: Population numbers are for the year 2016. A number of municipalities, called unincorporated areas (in German Gemeindefreies Gebiet) are not populated.

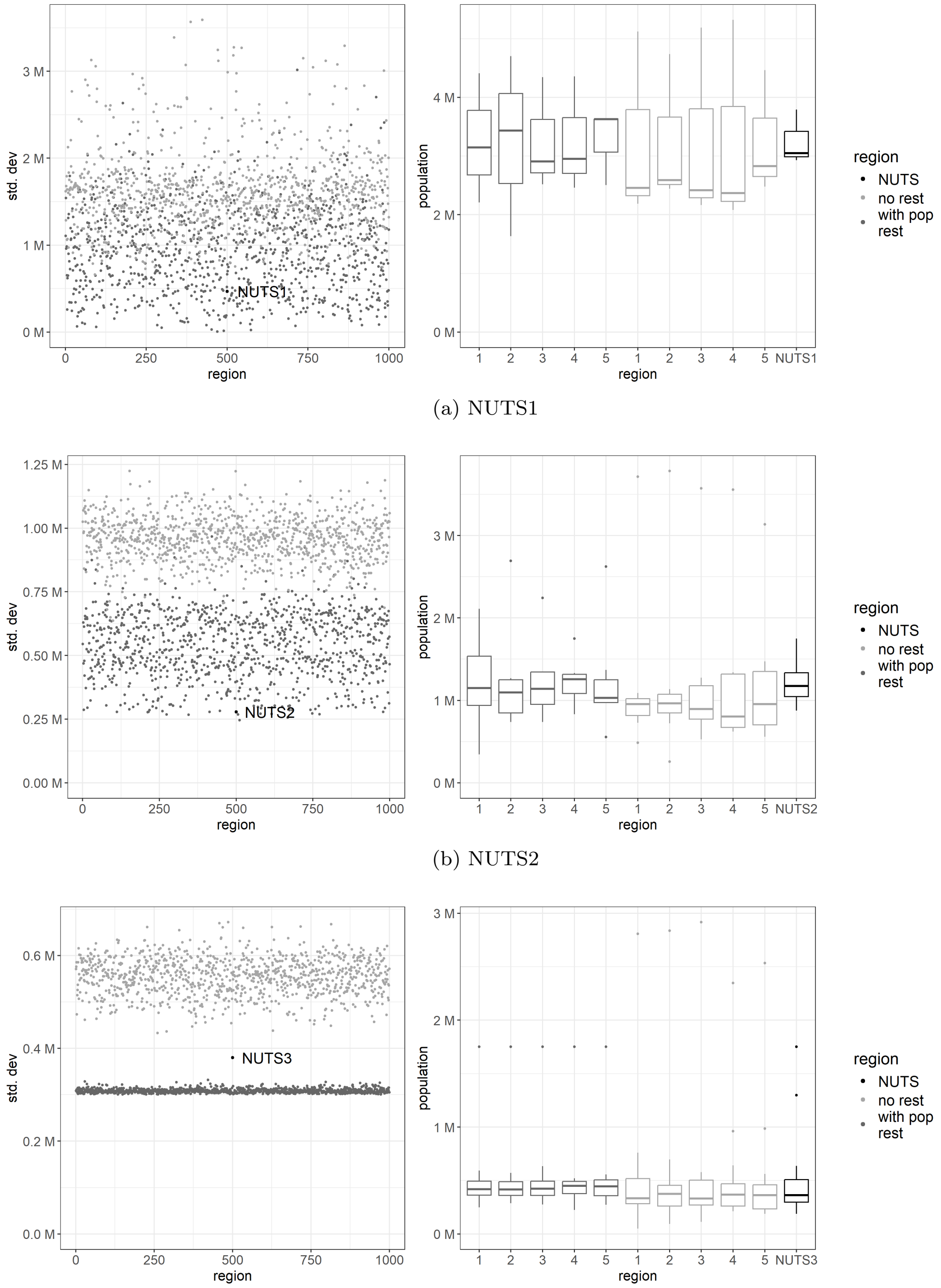


(a) NUTS1

(b) NUTS2

(c) NUTS3

Figure A2: Standard deviation and population distribution of simulated regions for the Netherlands

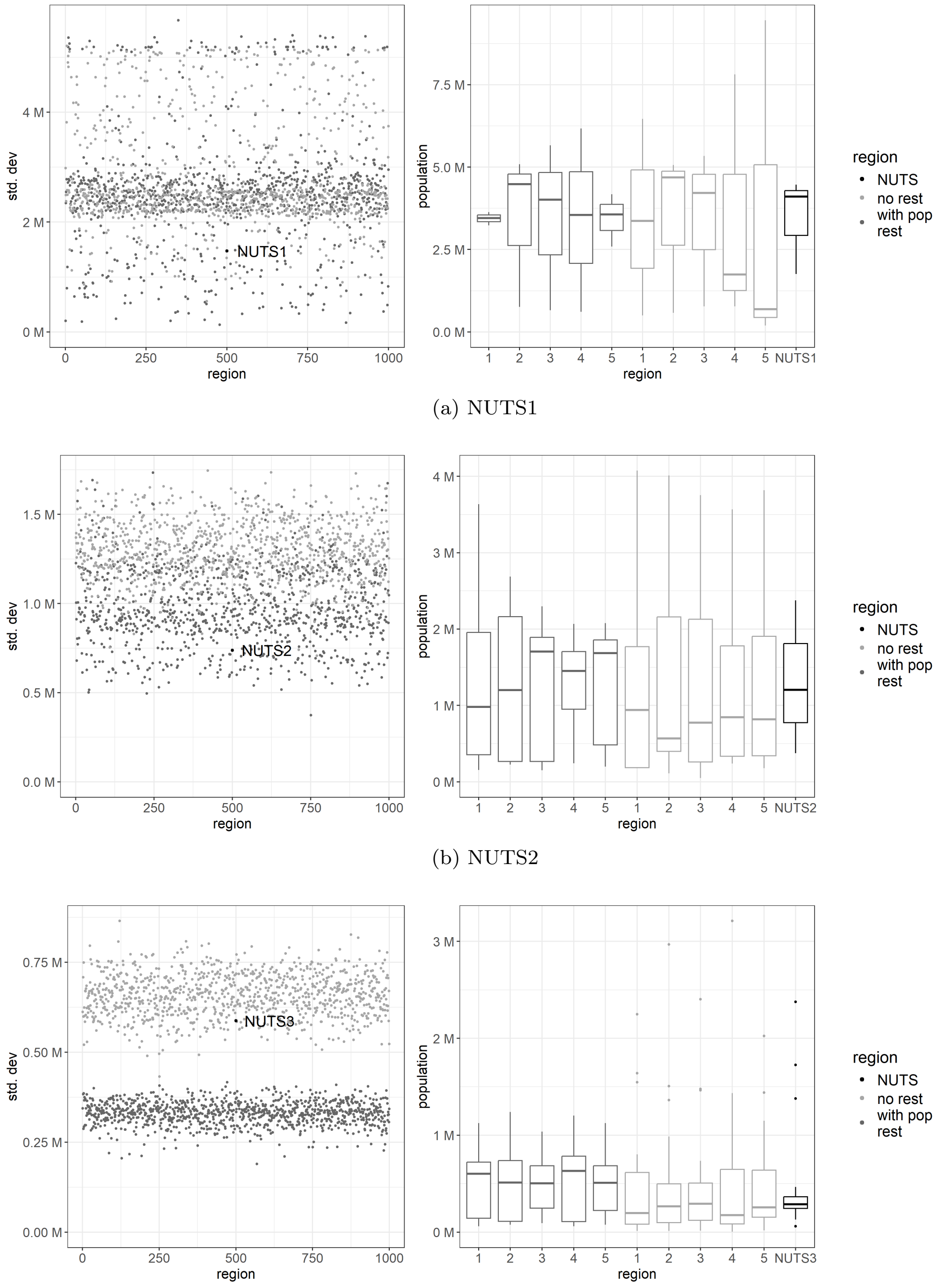


(a) NUTS1

(b) NUTS2

(c) NUTS3

Figure A3: Standard deviation and population distribution of simulated regions for Hungary

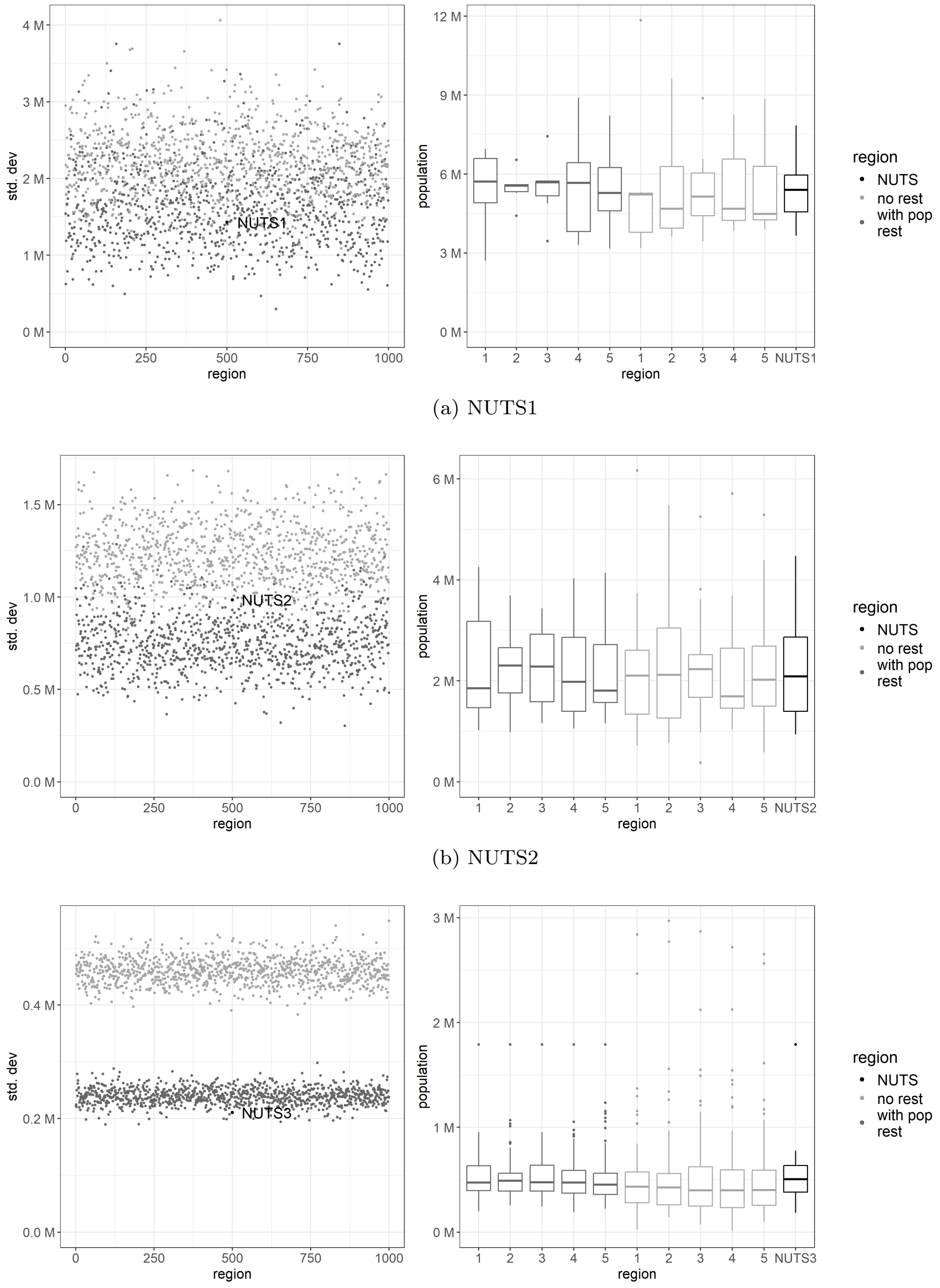


(a) NUTS1

(b) NUTS2

(c) NUTS3

Figure A4: Standard deviation and population distribution of simulated regions for Sweden

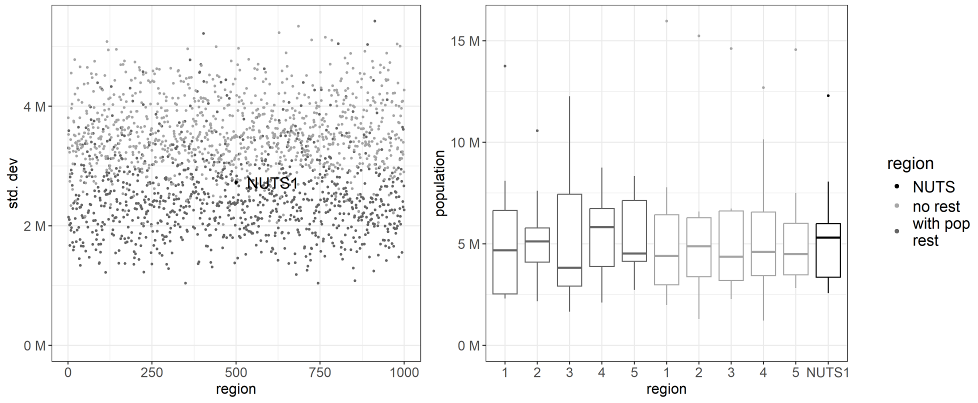


(a) NUTS1

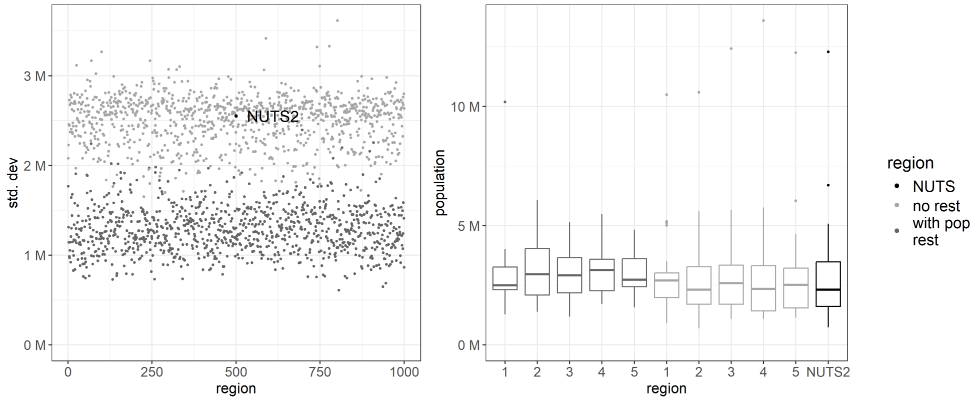
(b) NUTS2

(c) NUTS3

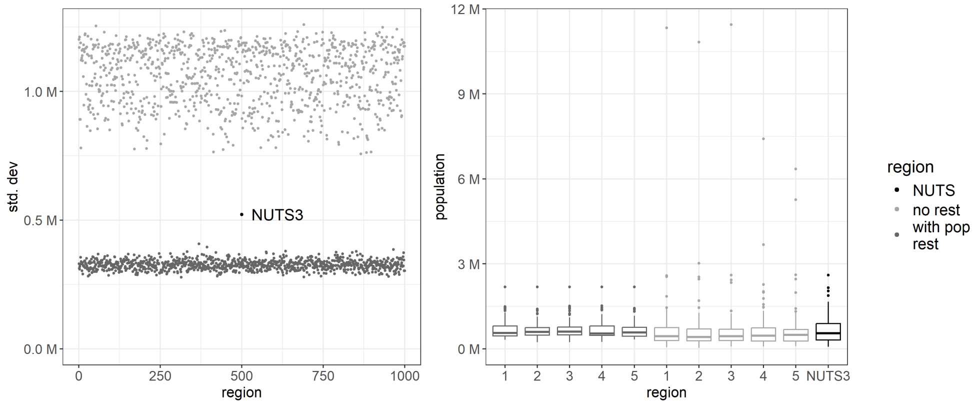
Figure A5: Standard deviation and population distribution of simulated regions for Poland



(a) NUTS1



(b) NUTS2



(c) NUTS3

Figure A6: Standard deviation and population distribution of simulated regions for France (The outermost regions and Corsica were excluded from the analysis)

## B Sample Construction and Data Selection

### *B.1 TFP*

The firm data set, obtained from the BvD AMADEUS database, is constructed based on multiple criteria. First, firms in the manufacturing sector located in Germany are selected. Our starting dataset of over 120,000 has a range between 2004 and 2018. However, the earlier and later years are excluded because in those years the number of firms for which the relevant data are available is small. The final time period of our analysis is 2009-2015.

Specifically, the following characteristics of firms are used to exclude units that might distort our analysis.

- Not enough location information (no information about the city and the NUTS 3 region or no data on Zip code and NUTS 1 region) or conflicting location information (city and NUTS regions do not match).
- Unambiguous mother companies: The statement of a company integrates the statements of its controlled subsidiaries or branches (consolidation codes C1 and C2).
- Firms with more than 10 branches.
- No data for at least one of the variables value added, total fixed assets, number of employees and cost of materials.
- The number of employees is below 10 for any of the available years.
- The number of employees exceeds 3,500 for any of the available years.
- Non-positive values for the financial variables.
- Large changes in the number of employees, material cost, value added.

Next, two and three in-between missing values over the available period are imputed. If the number of missing values between two years is larger than 3, the smaller available edge period is excluded. Finally, after the estimation of TFP, outliers for log(TFP) are also excluded (firms with productivity lower than 3 or higher than 6).

Table B1: Variable description

| Variable name | Name in AMADEUS | AMADEUS description | unit |
|---|---|---|---|
| capital | Tangible Fixed Assets | All tangible assets such as buildings, machinery etc. | thousands of Euros |
| employment | Number of employees | Total number of employees included in the company's payroll | |
| materials | Material's cost | Detail of the purchases of goods (raw materials + finished goods). No services. | thousands of Euros |
| value added | Added value | Profit for period + depreciation + taxation + interests paid + cost of employees | thousands of Euros |

*Note*: All monetary variables are later converted into real values, by using industry specific deflators obtained by EUROSTAT.

Table B2: Number of firms for calculating TFP in each year of the 7-year period

| year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| observations | 2,949 | 3,353 | 3,697 | 4,028 | 5,122 | 3,735 | 2,792 |

*Note*: The total number of observations is 25,676.

Table B3: Summary statistics

| Variable | Mean | St. Dev. | MAX | MIN | No. of observations | time period |
|---|---|---|---|---|---|---|
| ln(TFP) | 4.5003 | 0.4706 | 6.0429 | 2.9899 | 25,676 | 2009-2015 |
| Value added | 11,704.36 | 19,929.11 | 443,252.7 | 184.81 | 25,676 | 2009-2015 |
| Employment | 164.68 | 220.56 | 3,088 | 10 | 25,676 | 2009-2015 |
| Materials | 23,579.24 | 64,164.74 | 2,398,243 | 0.9381 | 25,676 | 2009-2015 |
| Capital | 6,754.56 | 18,706.95 | 646,448.20 | 0.9237 | 25,676 | 2009-2015 |

*Note*: All monetary variables have been converted into real values by using industry level deflators. Total number of firms is 7,317. Unbalanced panel.

### *B.2 Agglomeration variables*

For the calculation of the variables measuring MAR and Jacobs spillovers we use a larger set of firms. In this data set a firm is maintained if all of the following criteria are fulfilled:

- it has reliable location information,
- the consolidation code is different from C1 or C2,
- the firm has less than 10 branches,
- for the whole available period the number of employees is not smaller than 5 or larger than 3,500, and
- there have not been any large changes in the number of employees over the available time period.

For calculating the agglomeration variables, all manufacturing firms (including NACE rev. 2 sector 12, 15 and 19) with data on employment in a given year form 2009-2015 are considered. Missing values for number of employees between two years are interpolated. Furthermore, if other information (value added, tangible fixed assets, turnover) is available in a given year, but not the employment, then we set the number of employees to the number in closest year for which data is available. This data set contains 54,529 firms.

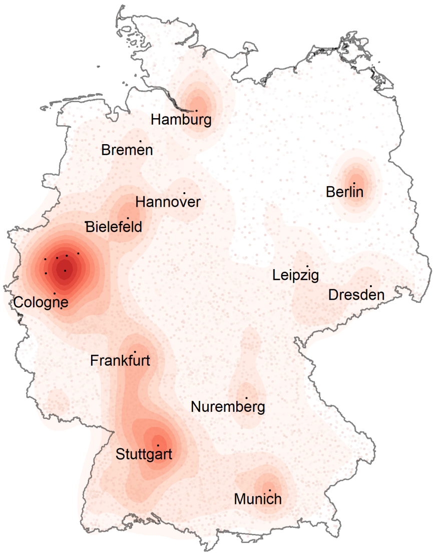


Figure B1: Heat map of the 54,529 firms

### *B.3 Control variables*

#### B.3.1 Market potential

Market potential is calculated by using the number of employees subject to social insurance based on their place of residence[13]. The municipality level data is obtained from the Federal Employment Agency of Germany for the period 2013–2015 and from the Database of the Federal Statistical Office of Germany for the period 2009–2012.

#### B.3.2 Transported goods

For the calculation of the control variable transported goods ($tg_{rt}$), data on all transported freight by airports, river and see ports, rail and highways is used.

For transported freight in airports, data from 24 airports in Germany (Berlin-Schönefeld, Berlin-Tegel, Bremen, Dortmund, Dresden, Düsseldorf, Erfurt, Frankfurt/Main, Friedrichshafen, Hahn, Hamburg, Hannover, Karlsruhe, Köln/Bonn, Leipzig/Halle, München, Münster/Osnabrück, Niederrhein, Nürnberg, Paderborn/Lippstadt, Rostock-Laage, Saarbrücken, Stuttgart, Zweibrücken) is considered. This includes data on loaded and unloaded freight (including mail).

For river and sea ports, data from 65 ports is taken, for which there is information over the period from 2009–2015.

For roads only highways are considered, as they are the most important for transportation of goods. However, data about the amount of freight transported on roads is available only at the national level. Therefore, to obtain a proxy of the variable for a region, we assume that the region's share of national amount of freight that is transported on its highways corresponds to the share of the region's highways in the national highway network, based on the length of the highways.

For freight amounts transported on rail only data at the NUTS2 level is available. However, for 2009 and 2010 there is only data at the national level. Therefore, the annual change rates of the transported freight on rail are used to approximate the variable at the NUTS2 level for these two years. To obtain respective estimates of freight for the NUTS3 and artificial regions, a proportional approach based on the length of the rail is used, similarly as for the freight transport on roads. However, for regions without a railway stop the variable is set to equal zero.

It is assumed that there are no changes in the railway lines and the roads during the 7 year period.

All data for transported goods in 1,000t are obtained from the German Federal Statistical Office and the geographic data about airports, water ports, rail and roads are taken from EuroGeographics.

After approximating the regional amounts of freight for each of the four types of transportation, the total amount of transported goods in a region is calculated and then it is divide by the area of the region.

[13] First we considered looking at employees based on their place of work, however the data had many missing values.

## C Additional results

Table C1: Descriptive statistics of estimates of Model I

| | localisation | localisation 5% sign | urbanisation | urbanisation 5% sign |
|---|---|---|---|---|
| *LARGE regions* | | | | |
| N | 1,000 | 6 | 1,000 | 10 |
| Mean | -0.0165 | -0.0056 | -0.1077 | -0.3183 |
| SD | 0.0397 | 0.1299 | 0.0941 | 0.0491 |
| Min | -0.1365 | -0.1327 | -0.5357 | -0.3865 |
| Median | -0.0162 | -0.0038 | -0.1077 | -0.3195 |
| Max | 0.1207 | 0.1207 | 0.2787 | -0.2409 |
| *MEDIUM regions* | | | | |
| N | 1,000 | 12 | 1,000 | 35 |
| Mean | -0.0091 | -0.0277 | -0.0908 | -0.2337 |
| SD | 0.0280 | 0.0739 | 0.0746 | 0.0413 |
| Min | -0.1051 | -0.1051 | -0.3413 | -0.3413 |
| Median | -0.0102 | -0.0690 | -0.0884 | -0.2285 |
| Max | 0.0801 | 0.0760 | 0.1502 | -0.1435 |
| *SMALL regions* | | | | |
| N | 1,000 | 10 | 1,000 | 31 |
| Mean | 0.0059 | 0.0315 | -0.0327 | -0.0926 |
| SD | 0.0126 | 0.0247 | 0.0261 | 0.0119 |
| Min | -0.0370 | -0.0370 | -0.1193 | -0.1193 |
| Median | 0.0057 | 0.0377 | -0.0325 | -0.0911 |
| Max | 0.0485 | 0.0485 | 0.0780 | -0.0714 |

Table C2: Descriptive statistics of estimates of Model II

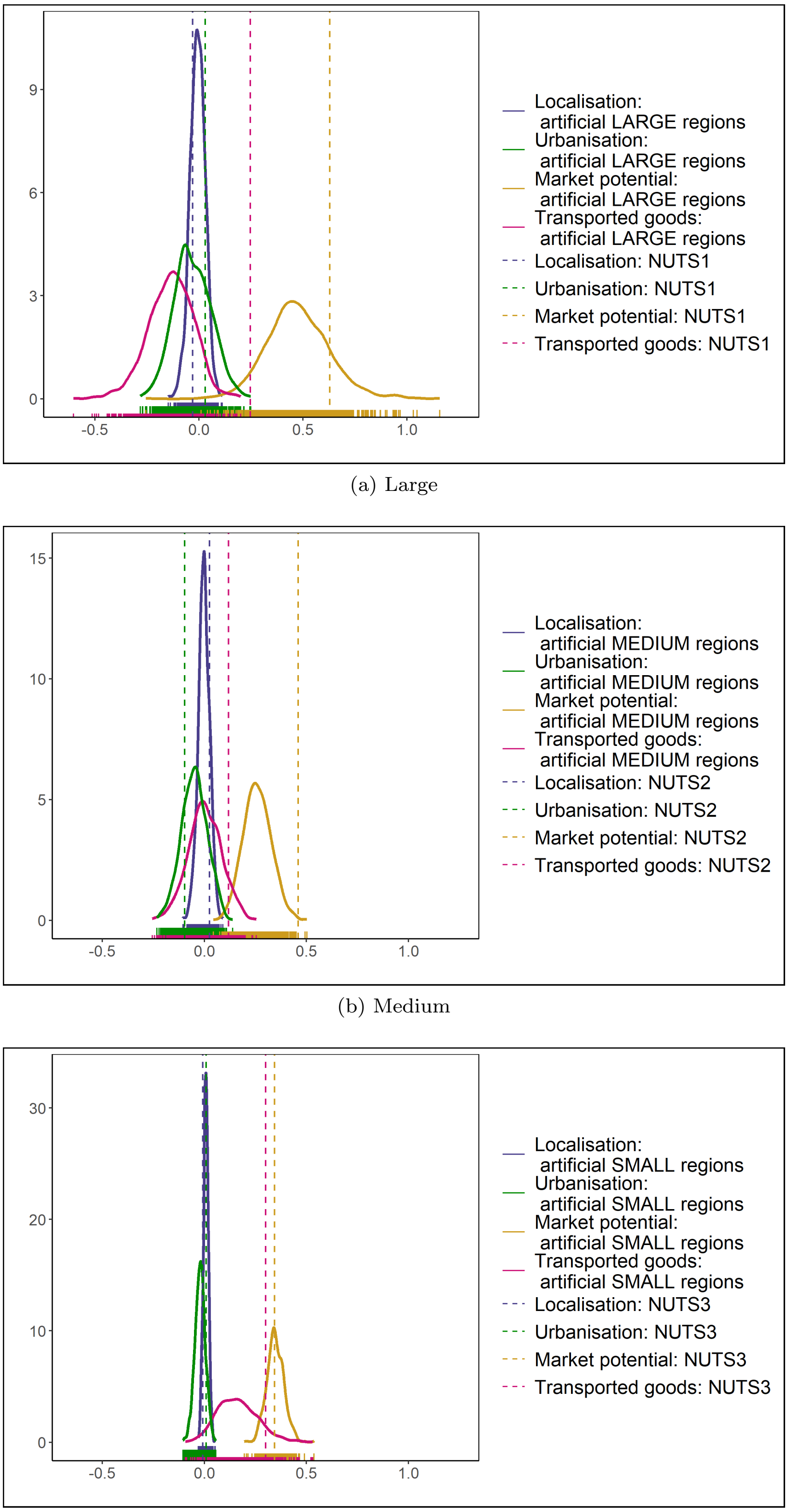
| | localisation | localisation 5% sign | diversity | diversity 5% sign |
|---|---|---|---|---|
| *LARGE regions* | | | | |
| N | 1,000 | 349 | 1,000 | 6 |
| Mean | -0.0481 | -0.0653 | -0.0057 | 0.0144 |
| SD | 0.0184 | 0.0114 | 0.0263 | 0.0667 |
| Min | -0.1031 | -0.1031 | -0.0927 | -0.0736 |
| Median | -0.0485 | -0.0645 | -0.0058 | 0.0439 |
| Max | 0.0050 | -0.0297 | 0.0781 | 0.0781 |
| *MEDIUM regions* | | | | |
| N | 1,000 | 254 | 1,000 | 25 |
| Mean | -0.0309 | -0.0488 | -0.0157 | -0.0635 |
| SD | 0.0154 | 0.0094 | 0.0206 | 0.0123 |
| Min | -0.0849 | -0.0849 | -0.0921 | -0.0921 |
| Median | -0.0305 | -0.0489 | -0.0146 | -0.0596 |
| Max | 0.0168 | -0.0290 | 0.0472 | -0.0488 |
| *SMALL regions* | | | | |
| N | 1,000 | 3 | 1,000 | 66 |
| Mean | -0.001 | 0.0075 | -0.0121 | -0.0282 |
| SD | 0.0091 | 0.0339 | 0.0087 | 0.0041 |
| Min | -0.0316 | -0.0316 | -0.0427 | -0.0427 |
| Median | -0.0003 | 0.0264 | -0.0118 | -0.0272 |
| Max | 0.0302 | 0.0279 | 0.0114 | -0.0216 |

Table C3: Descriptive statistics of estimates of Model III

| | | *LARGE regions* | | |
|---|---|---|---|---|
| | localisation | localisation 5% sign | urbanisation | urbanisation 5% sign |
| N | 1,000 | 10 | 1,000 | 4 |
| Mean | -0.0068 | -0.0549 | -0.0305 | -0.2432 |
| SD | 0.0363 | 0.0983 | 0.0894 | 0.0617 |
| Min | -0.1476 | -0.1476 | -0.2816 | -0.2816 |
| Median | -0.0064 | -0.1022 | -0.0361 | -0.2700 |
| Max | 0.1120 | 0.0873 | 0.2488 | -0.1510 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 724 | 1,000 | 111 |
| Mean | 0.4764 | 0.5225 | -0.1346 | -0.2950 |
| SD | 0.1532 | 0.1358 | 0.1100 | 0.0855 |
| Min | -0.2550 | 0.2020 | -0.6030 | -0.6030 |
| Median | 0.4678 | 0.5054 | -0.1296 | -0.2832 |
| Max | 1.1576 | 1.1576 | 0.2001 | -0.1405 |
| | | *MEDIUM regions* | | |
| | localisation | localisation 5% sign | urbanisation | urbanisation 5% sign |
| N | 1,000 | 16 | 1,000 | 19 |
| Mean | -0.0028 | 0.0009 | -0.0505 | -0.1855 |
| SD | 0.0279 | 0.0732 | 0.0624 | 0.0234 |
| Min | -0.1056 | -0.1056 | -0.2324 | -0.2154 |
| Median | -0.0024 | 0.0464 | -0.0499 | -0.1871 |
| Max | 0.0906 | 0.0906 | 0.1378 | -0.1418 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 546 | 1,000 | 5 |
| Mean | 0.2587 | 0.2974 | 0.0015 | -0.2162 |
| SD | 0.0686 | 0.0556 | 0.0820 | 0.0271 |
| Min | 0.0439 | 0.1585 | -0.2555 | -0.2454 |
| Median | 0.2562 | 0.2955 | 3e-04 | -0.2229 |
| Max | 0.5018 | 0.5018 | 0.2547 | -0.1775 |
| | | *SMALL regions* | | |
| | localisation | localisation 5% sign | urbanisation | urbanisation 5% sign |
| N | 1,000 | 15 | 1,000 | 1 |
| Mean | 0.0070 | 0.0380 | -0.0216 | -0.1009 |
| SD | 0.0124 | 0.0068 | 0.0252 | / |
| Min | -0.0300 | 0.0305 | -0.1043 | -0.1009 |
| Median | 0.0071 | 0.0368 | -0.0211 | -0.1009 |
| Max | 0.0529 | 0.0529 | 0.0555 | -0.1009 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 997 | 1,000 | 198 |
| Mean | 0.3476 | 0.3480 | 0.1625 | 0.2981 |
| SD | 0.0423 | 0.0417 | 0.0990 | 0.0684 |
| Min | 0.1964 | 0.2333 | -0.0928 | 0.1963 |
| Median | 0.3453 | 0.3453 | 0.1574 | 0.2823 |
| Max | 0.5368 | 0.5368 | 0.5308 | 0.5308 |

Table C4: Descriptive statistics of estimates of Model IV

| | | *LARGE regions* | | |
|---|---|---|---|---|
| | localisation | localisation 5% sign | diversity | diversity 5% sign |
| N | 1,000 | 342 | 1,000 | 21 |
| Mean | -0.0462 | -0.0630 | 0.0065 | 0.0481 |
| SD | 0.0181 | 0.0119 | 0.0297 | 0.0677 |
| Min | -0.1146 | -0.1146 | -0.1015 | -0.1015 |
| Median | -0.0468 | -0.0621 | 0.0059 | 0.0672 |
| Max | 0.0195 | -0.0355 | 0.1344 | 0.1344 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 764 | 1,000 | 153 |
| Mean | 0.4947 | 0.5363 | -0.1432 | -0.2878 |
| SD | 0.1552 | 0.1379 | 0.1077 | 0.0804 |
| Min | -0.2407 | 0.2068 | -0.6198 | -0.6198 |
| Median | 0.4871 | 0.5230 | -0.1378 | -0.2787 |
| Max | 1.1540 | 1.1540 | 0.1980 | -0.1161 |
| | | *MEDIUM regions* | | |
| | localisation | localisation 5% sign | diversity | diversity 5% sign |
| N | 1,000 | 219 | 1,000 | 18 |
| Mean | -0.0302 | -0.0499 | -0.0148 | -0.0568 |
| SD | 0.0158 | 0.0103 | 0.0196 | 0.0081 |
| Min | -0.0856 | -0.0856 | -0.0738 | -0.0738 |
| Median | -0.0302 | -0.0478 | -0.0152 | -0.0571 |
| Max | 0.0150 | -0.0296 | 0.0451 | -0.0443 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 617 | 1,000 | 7 |
| Mean | 0.2698 | 0.3005 | -0.0091 | -0.2077 |
| SD | 0.0666 | 0.0558 | 0.0817 | 0.0297 |
| Min | 0.0378 | 0.1477 | -0.2449 | -0.2449 |
| Median | 0.2669 | 0.2980 | -0.0094 | -0.1979 |
| Max | 0.5272 | 0.5272 | 0.2442 | -0.1652 |
| | | *SMALL regions* | | |
| | localisation | localisation 5% sign | diversity | diversity 5% sign |
| N | 1,000 | 6 | 1,000 | 37 |
| Mean | -0.0002 | -0.0051 | -0.0098 | -0.0283 |
| SD | 0.0088 | 0.0294 | 0.0090 | 0.0039 |
| Min | -0.0270 | -0.0270 | -0.0391 | -0.0391 |
| Median | -0.0002 | -0.0219 | -0.0098 | -0.0269 |
| Max | 0.0342 | 0.0342 | 0.0155 | -0.0243 |
| | market potential | market pot. 5% sign | transported goods | transp. goods 5% sign |
| N | 1,000 | 997 | 1,000 | 202 |
| Mean | 0.3505 | 0.3509 | 0.1639 | 0.2984 |
| SD | 0.0421 | 0.0414 | 0.0990 | 0.0685 |
| Min | 0.1977 | 0.2378 | -0.0889 | 0.1907 |
| Median | 0.3484 | 0.3485 | 0.1590 | 0.2837 |
| Max | 0.5310 | 0.5310 | 0.5317 | 0.5317 |



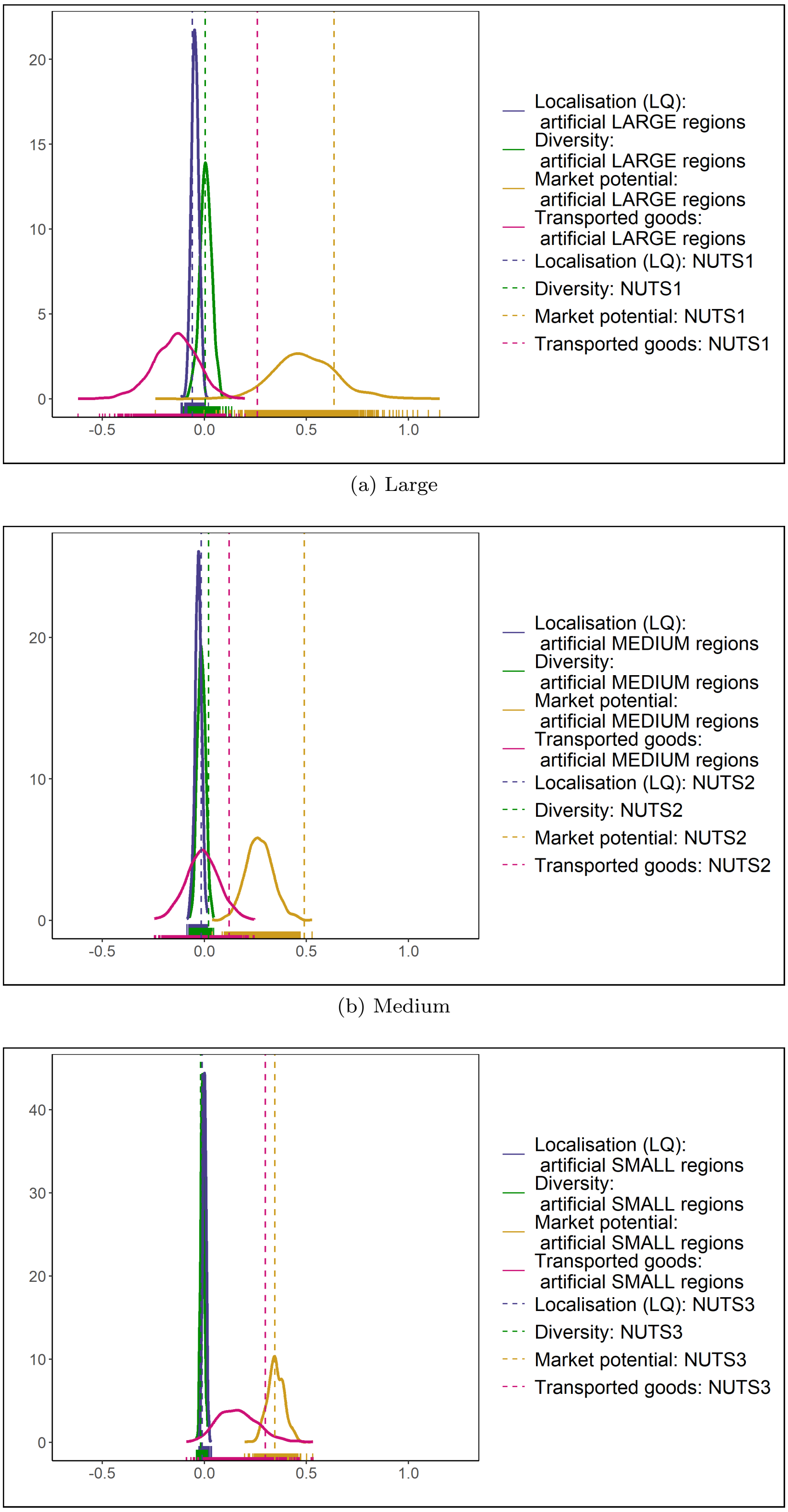
(a) Large

(b) Medium

(c) Small

Figure C1: Model III based on simulated regions with population restriction

*Note*: Estimates for equation 11 with both time and firm fixed effects. The densities of the parameter estimates rely on 1,000 settings of artificial regions.



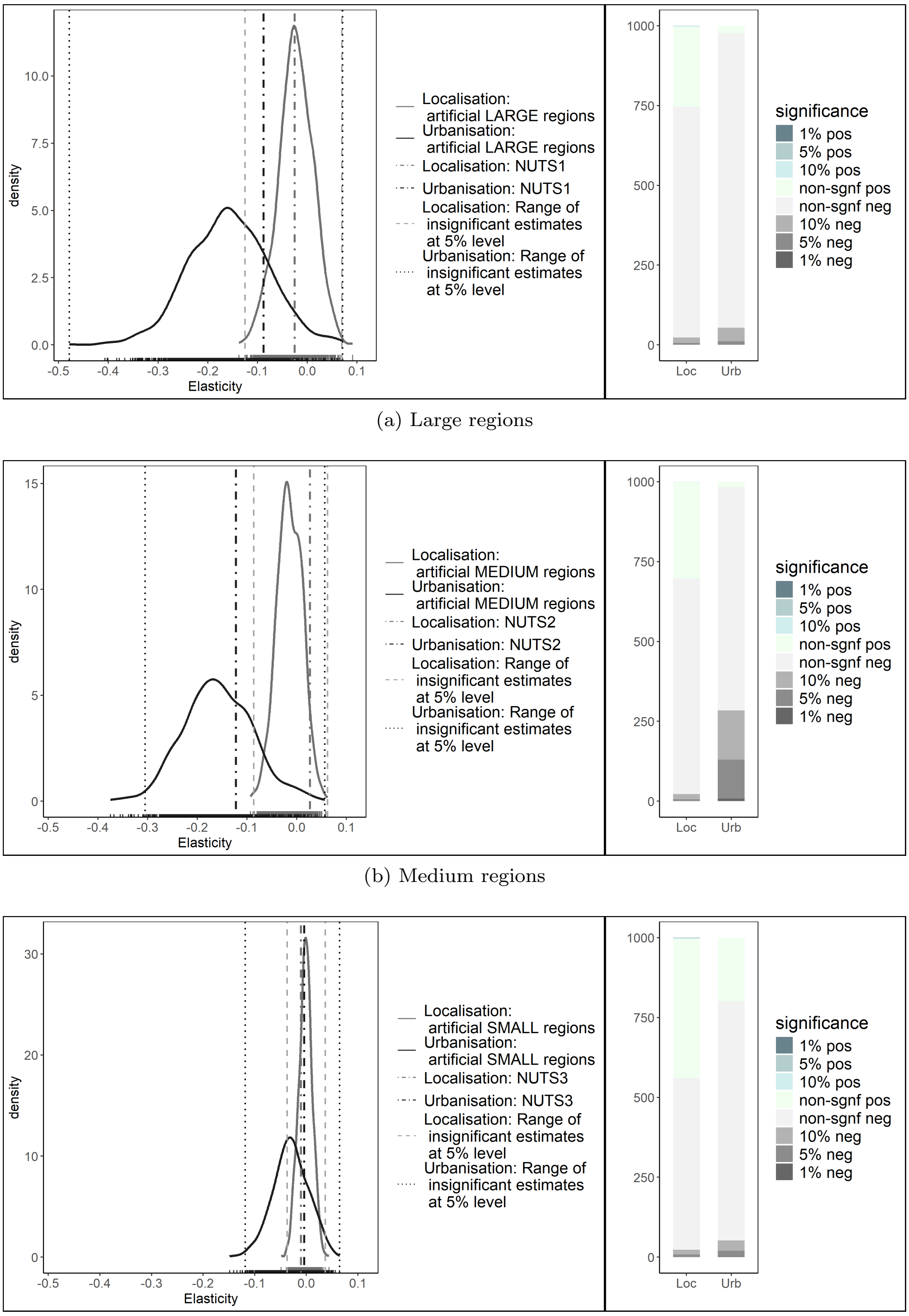
(a) Large

(b) Medium

(c) Small

Figure C2: Model IV based on simulated regions with population restriction

*Note*: Estimates for equation 12 with both time and firm fixed effects. The densities of the parameter estimates rely on 1,000 settings of artificial regions.



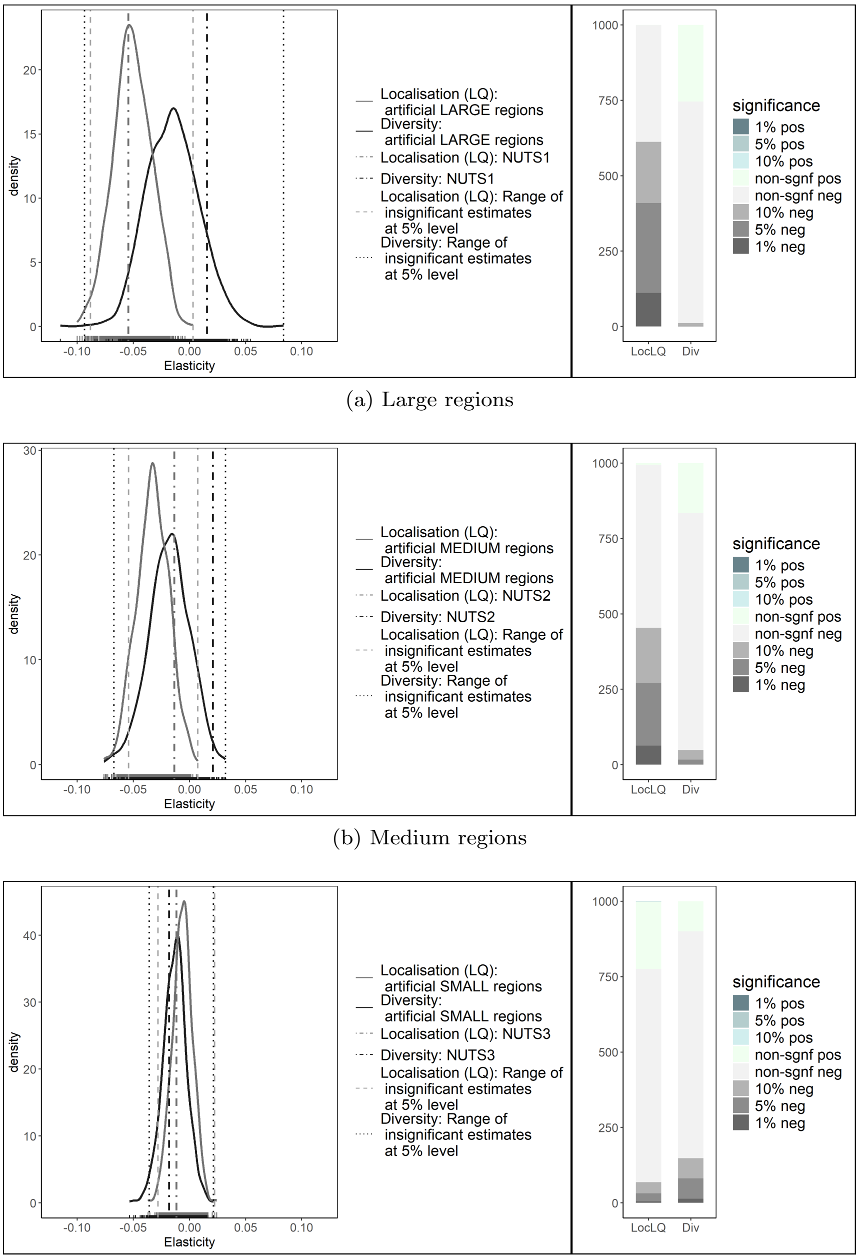
(a) Large regions

(b) Medium regions

(c) Small regions

Figure C3: Parameter estimates of the localisation and urbanisation variables in the Model I and the distribution of their significance (no population restriction simulation model)

*Note*: The densities of the parameter estimates rely on 1,000 settings of artificial regions. The range of insignificant estimates at 5 % level covers all parameter estimates that are not statistically significant at 5 % significance level.



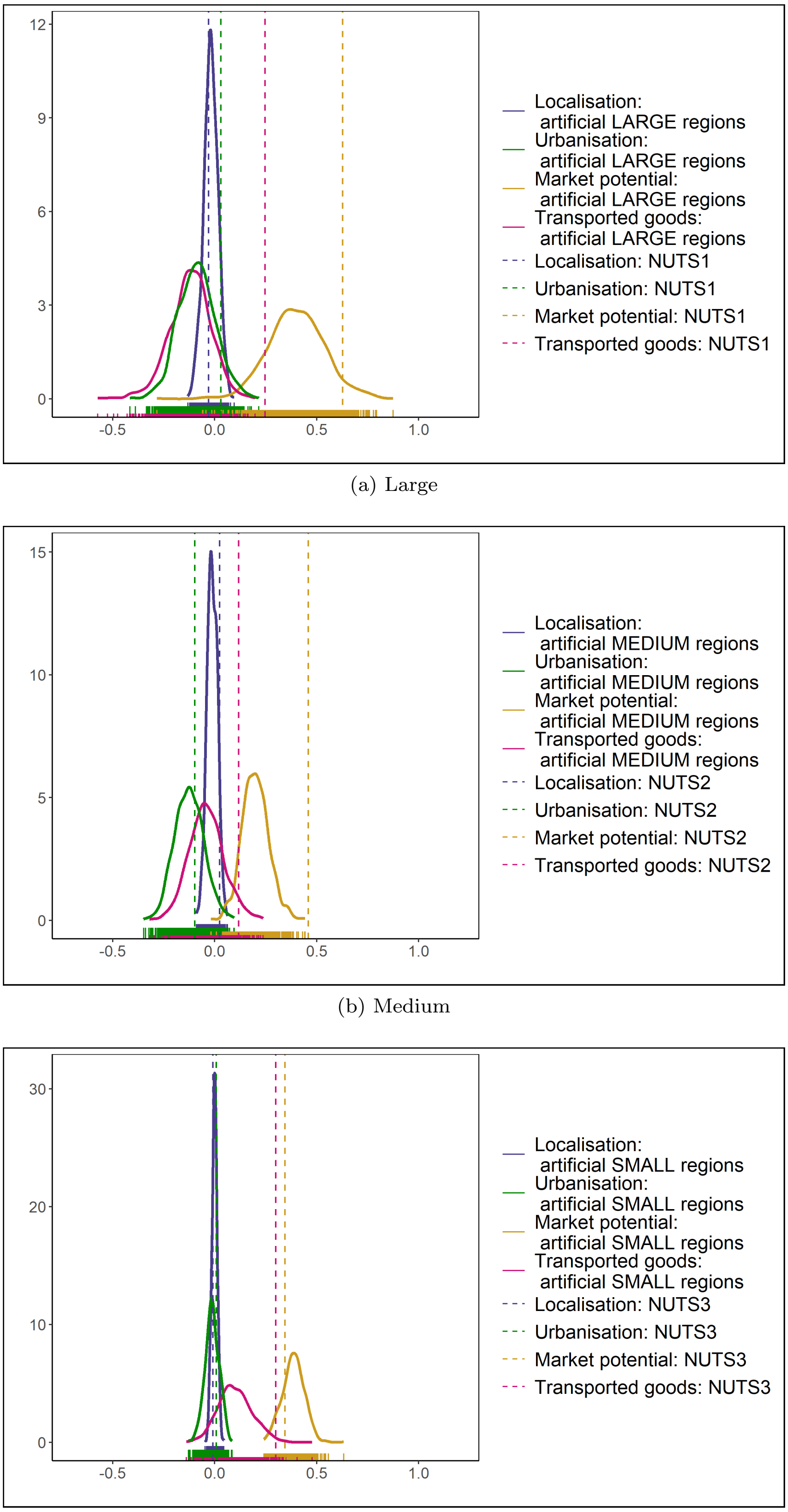
(a) Large regions

(b) Medium regions

(c) Small regions

Figure C4: Parameter estimates of the localisation and urbanisation variables in the Model II and the distribution of their significance (no population restriction simulation model)

*Note*: The densities of the parameter estimates rely on 1,000 settings of artificial regions. The range of insignificant estimates at 5 % level covers all parameter estimates that are not statistically significant at 5 % significance level.



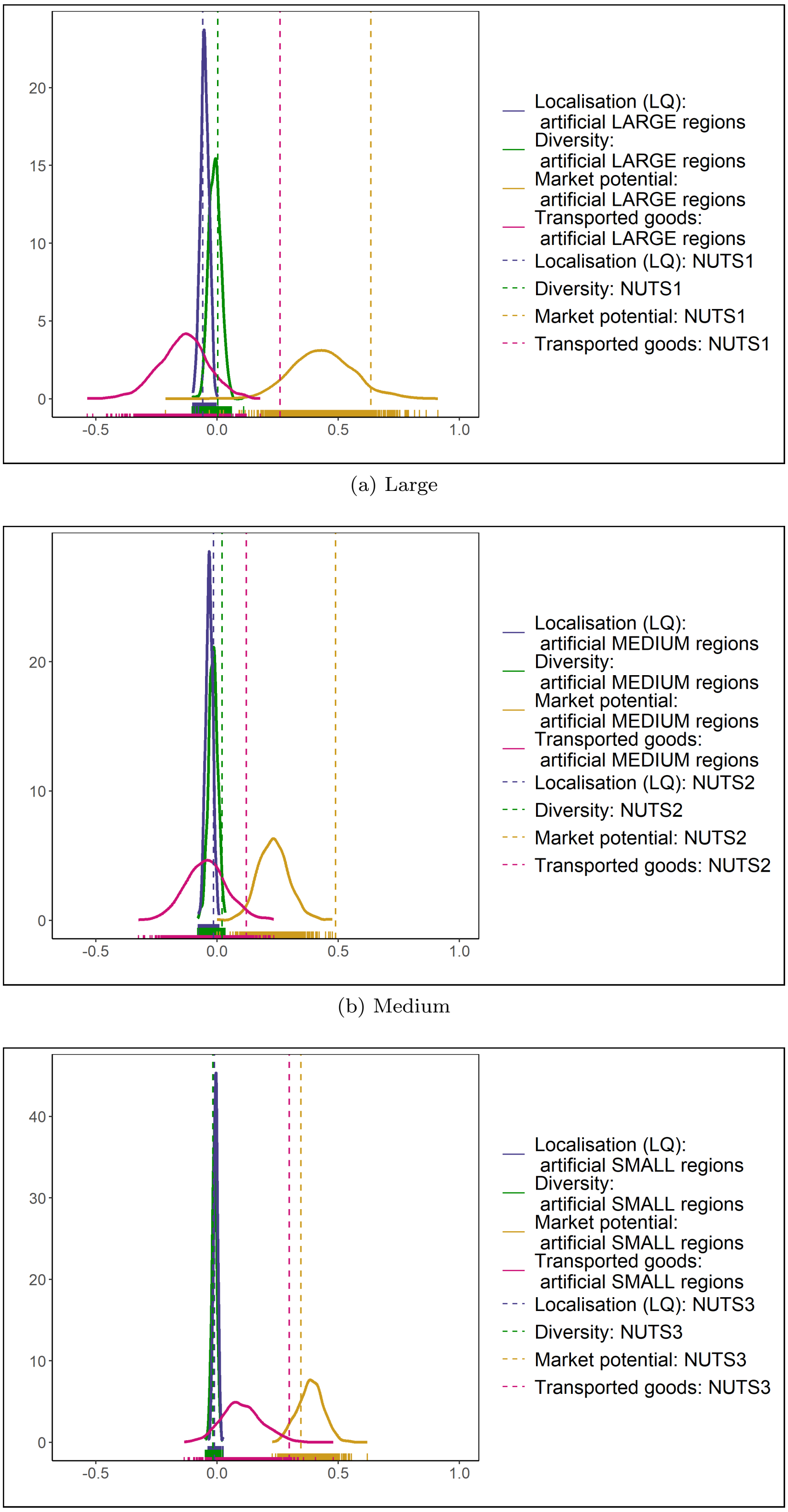
(a) Large

(b) Medium

(c) Small

Figure C5: Model III based on simulated regions with no population restriction

*Note*: Estimates for equation 11 with both time and firm fixed effects. The densities of the parameter estimates rely on 1,000 settings of artificial regions.



(a) Large

(b) Medium

(c) Small

Figure C6: Model IV based on simulated regions with no population restriction

*Note*: Estimates for equation 12 with both time and firm fixed effects. The densities of the parameter estimates rely on 1,000 settings of artificial regions.