Testing for Clustering of Industries – Evidence from micro geographic data

02.11

Tobias Scholl and Thomas Brenner

Testing for Clustering of Industries – Evidence from micro geographic data[1]

Tobias Scholl[2]

Centre for Clusters and Competitiveness, EBS Business School, Hessen.

Thomas Brenner

Section Economic Geography and Location Research, Institute of Geography, Philipps-University, Marburg.

Abstract:

We present a new statistical method that describes the localization patterns of industries in a continuous space. The proposed method does not divide space into subunits whereby it is not affected by the Modifiable Areal Unit Problem (MAUP). Our method fulfils all five criteria for a spatial statistical test of localization proposed by Duranton and Overman (2005) and improves them with respect to the significance of its results. Additionally, our test allows inference to the localization of highly clustered firms. Furthermore, the algorithm is efficient in its computation, which eases the usage in research.

Keywords: Spatial concentration, localization, clusters, MAUP, distance-based measures.

JEL Classifications: C40, C60, R12

---

[2] Corresponding Author: Tobias Scholl, Centre for Cluster and Competitiveness, EBS Business School, Hessen, Germany. E-Mail: tobias.scholl@ebs.edu.

1   Introduction

Spatial data has experienced a recognizable growth, both in its daily usage and availability. Though more and more micro spatial data is freely accessible, there is a lack of applying such data to spatial econometric analysis (Miller 2010: 182). Most of the papers still deal with the comparison of regions and do not concern the real spatial position of economic actors such as firms or research institutions. Especially in geography there is a refusal of quantitative models that use spatial-aggregated data (Bathelt and Glückler 2003: 121). Many popular quantitative methods in spatial economics such as the Elison & Glaeser- or the Gini-Index base on the comparison of spatial-subunits in a research area (Marcon & Puech 2010: 746). Usually, this division does not depend on economic characteristics, but on administrative classifications that provide the data for these indices. Keeping this problem in mind, the criticism of quantitative spatial models becomes clear: Instead of concentrating on firms as economic actors, researchers compare regions – not because it is reasonable, but the easiest way to gain results.

The problems that derive from the usage of spatially aggregated data are not only recognized in "critical" economic geography, but also in statistics. The Modifiable Areal Unit Problem (MAUP) states that results of statistics that use spatial aggregated data always depend on the chosen level of aggregation[3]. MAUP effects can be serious to the extent that affected indices can produce contradicting results when changing from one aggregation level to another (Koh & Riedel 2009: 2).

There are only few papers and even less models that provide quantitative spatial analyses of empirical economic activity without the MAUP. The first paper in this context was published by Duranton and Overman in 2002[4], in which the authors examine the concentration of manufacturing firms in the U.K. They use a dataset that provides the postcode and the Standard Industrial Classification of all firms in the U.K. Given that postcodes in the U.K. typically refer to one property or a very small group of dwellings, the authors obtain the almost exact spatial localization for all firms. In their paper, Duranton and Overman formulate five criteria for a spatial statistical test of localization: "In summary any test of localization should rely on a measure which (i) is comparable across industries; (ii) controls for the overall agglomeration of manufacturing; (iii) controls for industrial concentration; (iv) is unbiased with respect to scale and aggregation. The test should also (v) give an indication of the significance of the results" (Duranton & Overman 2005: 1079). Duranton and Overman demonstrate that aggregated indices, such as the Elison-Glaser-Index, provide results that are too optimistic with regard to the extent of concentration in manufacturing (Duranton & Overman 2005: 1097).

---

[3] For a detailed analysis of the MAUP see Openshaw, S. (1984)
[4] Working paper 2002, paper 2005

As mentioned above, only a few publications have used the new index of Duranton and Overman (henceforth D&O-index). This is not a result of a refusal in the scientific community, but of two specific problems: The first can be found in the richness of their dataset, since only few investigations can provide an almost exact localization of firms in space. This, however, is not the main reason, as the index can also be applied to spatially aggregated data. The graver problem lies in its computational complexity. Despite applicable data, Vitali et.al (2009) partially abandoned the D&O-index due to its "tremendous computational requirements" (Vitali et.al 2009: 20). Ellison et al. (2010) simplify the D&O-index in several aspects in order to apply it on the whole population of manufacturing firms in the USA. Nevertheless, they state that the index "is much more computationally intensive vis-a-vis simpler discrete indices" and quantify its computing time to three months for their research (Ellison et al. 2009: 5). In computer science, algorithms are called inefficient if they show bad performance with an increasing number of observations.

The *M*-Function, an alternative approach to the D&O-index, was introduced by Marcon and Puech in 2007. Though their function has some interesting features, it has not attracted similar attention in the literature as the D&O-index.

Despite the few publications that deal with MAUP-free quantitative analyses, we see an increasing demand for new methods in this field, which can be applied in economics and economic geography. Therefore, the aim of our paper is to present a new statistical method that fulfills the 5 criteria of Duranton and Overman and is efficient in its computational requirements.

We will demonstrate our method by means of the German micro technology industry. Micro technology, or microsystems technologies (abbr. MST), is a high-tech industry that combines different microelectronics components in an embedded system in a very small measure. Its fields of application range from automobiles to medical technology. The MST is a young industry that evolved from microelectronics at the end of the eighties. There is a common sense in economics and economic geography that young high-tech industries tend to cluster in space, as they benefit from positive spatial externalities, such as local spillovers, local embeddedness and trust. Though the concept of local clusters is mentioned in countless publications, little research has been done to identify their spatial dimension. Porter, the founder of the cluster concept, uses the LQ-Index – a very simple Index that compares regions' share of employment, but is affected by the MAUP (Woodward & Guimarães 2009: 77 f.). However, the MAUP-free D&O-index and *M*-function are not applicable to cluster analysis, as they cannot state where clustered firms are located. Our new method also allows inferences to the spatial localization of highly concentrated firms and therefore delivers new insights into the debate of firm-clusters.

The rest of the paper is organized as follows: section 2 presents the data basis used in the methodical analysis. Section 3 describes recent MAUP-free statistical methods whereas section 4 outlines our new approach. In section 5 we show the results for the different me-

thods used in this paper and discuss the advantages and disadvantages of our method. Finally, section 6 concludes and outlines new possibilities for further research.

2   Data

The dataset of our paper contains the exact location (street, house number and zip-code) of all German MST-firms. The dataset was provided by the German-based IVAM, an international association of companies and institutes in the field of micro technology. The dataset included 873 firms that fulfill at least one or more of the following prerequisites:

- (Former) Members of the IVAM or another associations in the field of micro technology
- Firms that are listed in specific databases (e.g. www.mst-online.de)
- Participants of fairs or conferences that deal with micro technology
- Participants of public/federal projects covering micro technology
- Firms that are mentioned in trade journals
- Firms that are listed in the German Commercial Registry under the headword "micro"

For all firms the IVAM checks via the company's homepage whether they are really active in the MST-sector. Additionally, we double-checked the data with the Germen Commercial Registry, in order to obtain the firms date of inception and to check whether they still exist or have relocated. Finally, 861 MST-firms were included in the statistical analysis.

We computed the longitude and latitude of the firms' exact location (street, house number and postcode) whereby we gain data that is even more detailed than that of Duranton and Overman.

As our benchmark we used a random sample of 20,000 German manufacturing firms, drawing them randomly from the list of all manufacturing firms in the Creditreforms' database (MARKUS; most comprehensive database on German firms). In the same way to the MST-firms, we computed the easting and northing of the firms' exact location.

3   Existing distance-based methods

Section 1 has mentioned that only few papers and even less models deal with MAUP-free quantitative analysis. Though similar methods have a longer tradition in ecology, they were not used in economics or economic geography, as they are not applicable to economic activity (Marcon & Puech 2010: 747, 750). To our knowledge, there are only two distance-based methods that fulfill all of the 5 above mentioned requirements: the D&O-index by Duranton

and Overman (2005) and the *M*-function by Marcon and Puech (2010). Though at first glance the D&O-index and *M*-function seem to be quite similar, they have a different mathematical background and provide different results. In order to keep the focus on our new approach, we will not discuss them in great detail, as they both suffer from the same problems that can be solved with our new method[5]. Furthermore, we will concentrate on the basic model of all three indices (D&O-index, *M*-function and our approach): Intra-industrial concentration without weighting the distances by a firm's share of employment[6]. Nevertheless all three indices can be applied to measure co-concentration between two industries and can account for a firm's share of employment.

3.1 D&O-index: a density function

In the following we will present the three indices considering the German MST industry that consists of N=861 firms, located over the entire federal territory. The basic idea of the D&O-index is to check whether the number of neighborhoods at a specific distance between firms is significantly higher or lower than expected by random. However, the empirical number of neighborhoods is not considered, but its smoothed density over all neighborhoods, expressed by the term *K(d)*. The first step to compute *K(d)*-values is to build the geographical distances[7] between all possible pairs of firms so that one gains $N(N-1)/2$ unique bilateral distances (370.230 in our example). In the next step, one counts the number of firm pairs that have a certain distance. Duranton and Overman (2005) use a step interval of 1km and consider only those distances that are below the median distance between manufacturing firms in the entire UK. For Germany we calculated a median distance of 362 km[8]. This distance is split at each km so that we gain 362 intervals. Any high *K(d)*-value outside the distance of 362 km could be interpreted as dispersion but Duranton and Overman see this information as redundant (Duranton & Overman 2005:1086). The last step is smoothing the observed numbers using a Gaussian kernel function. Hence the formula is:

$$K(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{i,j}}{h}\right), \tag{1}$$

where h is the optimal bandwidth[9] and f stand for the kernel function.

---

[5] For a detailed comparison of the D&O-index and the *M*-function see Marcon & Puech (2010).

[6] We had to modify Marcon and Puech's formula towards the non-observance of employment data.

[7] We computed orthodromic distances instead of Euclidian distances, proposed by Duranton & Overman (2005).

[8] Due to high computational requirements we draw 4000 out of the 20.0000 firms

[9] Optimal bandwidth: $1.06sn^{-0,2}$, where n is the observed number and s is the standard deviation (Klier & McMillen 2006: 12).
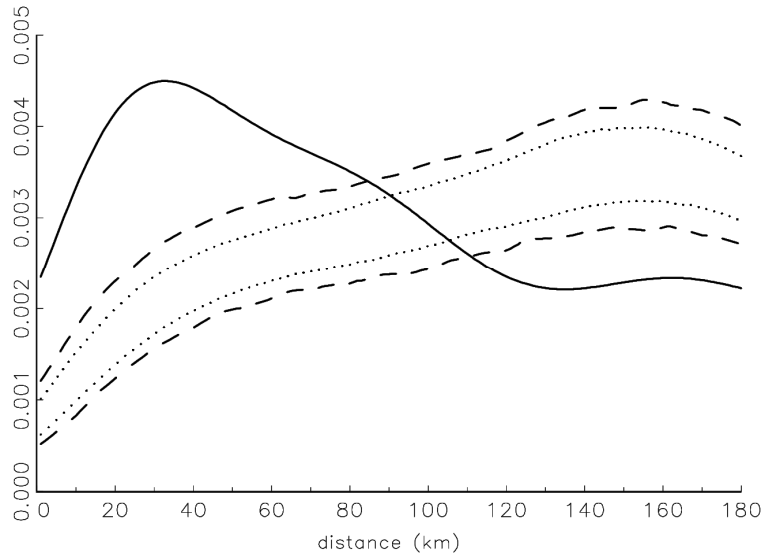
Figure 1: *K*-density, local confidence intervals and global confidence bands for an illustrative industry. Source: Duranton & Overman 2005.

The solid line in Figure 1 plots the *K(d)*-values for an illustrative industry (source: Duranton & Overman 2005). The dashed and dotted lines refer to the local and global confidence intervals that will be explained now.

We want to control whether the *K(d)*-values of our industry of interest show significant spatial concentration or dispersion at specific distances. At this stage we need confidence intervals that are constructed by a Monte-Carlo approach: From the 20,000 German manufacturing firms we draw the same number of firms (861) without replacement. These 861 firms represent a random industry localization, whose bilateral distances are computed.

The basic idea behind this procedure is that the spatial localization of industries does not follow a pure random schema, as industries cannot settle anywhere in a country. It is obvious that natural barriers (lakes, rivers, mountains) or political restriction (nature reserves, residential areas) limit the location choice of entrepreneurs (Duranton & Overman 2005:1085). Consequently, a purely stochastic pattern (e.g. a Poisson distribution) as a benchmark would provide too optimistic results. A better way is to build random samples of real company locations and use them as a benchmark (Durnaton and Overman call it counterfactuals).

The step of drawing random firms and computing their bilateral distances is done 1000 times. For the 1000 benchmark simulations the number of neighborhoods for each interval is sorted in ascendingorder. The 5-th and 95-th percentile are selected to compute the *K(d)*-function according to formula (1). We obtain a lower 5% and an upper 5% confidence interval that Duranton and Overman call local confidence intervals or $\overline{K_A}(d)$ and $\underline{K_A}(d)$ respectively, (dotted lines in Figure 1) (Duranton & Overman 2005:1086). The industry in Figure

1 lies between 0 and 90 km over the upper local confidence interval, stating that this industry shows significantly more neighborhoods at small distances.

Due to the fact that the *K(d)*-function is built separately for each km, an industry will probably hit the local bands once. In order to test whether an industry is generally more concentrated, Duranton and Overman propose the computation of global confidence intervals. By means of the thousand simulations, the upper global confidence interval $\overline{\overline{K}}(d)$ is computed in such way that only 5 % of the thousand simulations hit the global confidence interval; the same is performed for the lower interval (Duranton & Overman 2005:1087). The computation of global confidence intervals is somewhat tricky and we will explain it through the lower global band: For the lower band, we begin by selecting the 50th lowest values for each of the 362 intervals (interval step: 1 km) out of all 1000 simulations. This step is in line with the computing of the local band but now, we additionally count how many different benchmark simulations were used to build this band. If this number Ω exceeds 50 (5 %), we have to select the 50-1st (49th) lowest values and so on until we reach a set of values that contains $\Omega^* \leq 50$ different simulations. The band that is built of the 50-*[th] lowest values is the global lower confidence band.

Duranton and Overman define an industry as globally concentrated if their *K(d)*-function at least once lies over the global confidence interval. Respectively, an industry is globally dispersed if their *K(d)*-function once lies under and for all distances never lies over the global band. Using the global bands, Duranton and Overman propose two global parameter Γ and Ψ that represent an index of global localization/dispersion, where

$$\Gamma(d) \equiv max\big(\hat{K}(d) - \overline{\overline{K}}(d), 0\,\big), \tag{2}$$

is the index of global localization at a distance d and

$$\Psi(d) \equiv \begin{cases} max\left(\underline{\underline{K}}(d) - \hat{K}(d), 0\,\right) \; if \sum_{d=0}^{d=362} \Gamma(d) = 0\,, \\ \qquad\quad 0 \quad otherwise, \end{cases} \tag{3}$$

is the index of global dispersion. Note that an industry can only show global localization or dispersion and that the value of Γ and Ψ refers to a specific distance interval. In order to compare the two indices between industries, one can sum up its values over all distances such that Γ for industry A is $\Gamma_A = \sum_{d=0}^{362} \Gamma_A(d)$.

3.2 Weaknesses of the D&O-index

Compared to MAUP-affected indices the D&O-index is a clear improvement, yet it still faces problems that affect its usage in research:

(1) High computational requirements: As mentioned in section 1, Vitali et al. complain about the "tremendous computational requirements" of the D&O-index. Considering the

function's mathematical background allows insights into this problem: A computation has to be conducted for each interval, both for the observed industry and for the 1000 benchmark simulations. If industries with many plants or large areas are observed, computational requirements reach the limit of common computing power.

(2) Risk of wrong benchmarks: One might argue that computational problems can be solved by reducing the number of benchmark simulations. However, this is not feasible for the D&O-index, as all distance intervals are regarded separately. There is a huge risk that few random samples will show abnormal high or low *K(d)*-values at some intervals, thus, leading to a false interpretation of the concentration or dispersion of the observed industry at these distances. For that reason, 1000 benchmark simulations have to be computed, to gain 1000 independent values for each interval.

(3) Re-division of space/secondary MAUP: As the number of benchmark simulations cannot be modified, decreasing the number of intervals is a possibility to reduce computational requirements. This is done in the paper by Vitali et al. (2009) and probably in the paper by Klier & McMillen (2008). In their study of manufacturing localization in different European countries, Vitali et al. use 40 evenly spaced intervals (Vitali et al. 2009: 11). The choice of 40 intervals is arbitrary, and it is obvious that the size of the intervals differs among countries such as Germany and Belgium. Thus this attempt suffers from the same problems as the mentioned MAUP-affected indices. The second concern refers to the computation of global confidence intervals. The finding of Ω^* heavily depends on the number of intervals. Since its computation becomes impossible when a lot of intervals are used, larger areas under investigation need a subsequent grouping of values. In summary, these mentioned problems can be called a secondary-MAUP: Even if data provide point-localization of firms, high computational requirements or statistical needs might be solved in a subsequent division of space.

(4) Unresolved lack of significance: In comparison to prior indices, a central strength of the D&O-index is its ability to give an indication of the significance using the confidence bands as the null hypothesis. Γ and Ψ as parameters of global concentration/dispersion may be useful to compare different industries because in most cases, the values should differ clearly. However, when Γ- or Ψ-values of two samples are very similar, the D&O-index cannot detect whether these differences are significant or not. For instance, this situation may appear when subunits of one industry are compared (see Klier & McMillen 2008: 254). As mentioned above, the finding of Ω^* depends on the number of intervals. Duranton and Overman face this problem by interpolating values if even the highest/lowest band is built by more than 5 % of all simulations (Duranton & Overman 2005:1087). To our mind this is an improper approximation because one cannot interpolate to unknown values. A further lack of significance is that the D&O-index cannot detect non-random spatial distribution patterns that do not involve significant localization or dispersion at some distances (Duranton & Overman 2005:1088).

3.3 *M*-function: a cumulative function

So far, we have described the D&O-index whose basic concept is the usage of a density function. Before we introduce our cumulative-density method, we will discuss the *M*-function as an example for a cumulative function.

As shown in section 3.1, the D&O-index asks whether the density of neighborhoods at a certain km is significantly below or above a random distribution. The *M*-function, in contrast, does not regard neighborhoods *at* but *up to* a certain distance and compares them to a random distribution of firms. Again, we build geographic distances between all possible pairs of firms in our MST industry (denoted by N_{mst}) and consider only distances from 0 to 362 km. The benchmark is given by *N*, a number of firms, built by a random population of firms plus the MST-firms. Now a circle is lain around each MST-firm with a radius r that grows in 362 steps from 1 to 362 km. Consider a dummy variable $c_{mst}(i,j,r)$ that is equal to 1 if the distance between two MST-plants *i* and *j* is less than or equal to the radius *r*, otherwise $c_{mst}(i,j,r)$ is 0. For a given radius, the number of neighborhoods for plant *i* is thus $\sum_{j=1,\,i\neq j}^{N_{mst}} c_{mst}(i,j,r)$. In the same way we can define $\sum_{j=1,\,i\neq j}^{N} c(i,j,r)$ as the number of neighborhoods between plant *i* and firms that belong to the firm population of our benchmark *N*. The *M*-function explains the ratio of neighborhoods in the MST-industry and a random industry for a given radius *r* as:

$$M_{mst}(r) = \frac{N-1}{(N_{mst}-1)*N_{mst}} \sum_{i=1}^{N_{mst}} \frac{\sum_{j=1,i\neq j}^{N_{mst}} c_{mst}(i,j,r)}{\sum_{j=1,i\neq j}^{N} c(i,j,r)}. \tag{4}$$

M-values are 1 if the number of neighborhoods in the MST-industry does not differ from those of a random firm population. If they are above 1, MST firms show a higher concentration up to a certain radius (Marcon & Puech 2010:749). In comparison with the D&O-index, the number of benchmark firms does not need to be equivalent to the number of the observed MST-firms. The first fraction takes account of this property. In order to check whether results are significant, local and global confidence intervals are built according to the D&O-index (Marcon & Puech 2010:750).

As the *M*-function is a cumulative index, it cannot state at which exact distances dispersion or concentration occurs. This is an obvious shortcoming for spatial analysis. However the *M*-function also has some clear advantages:

1. The risk of wrong benchmarks is reduced because the *M*-values of succeeding intervals are highly correlated; however until now this feature has not been quantified (Marcon & Puech 2010:750).
2. Marcon and Puech demonstrate that their function can detect whether clusters are located randomly or repulsively to each other (Marcon & Puech 2010:755).
3. To some extent, *M*-values are easier to interpret as they present the more intuitive

ratio of firms instead of densities (Marcon & Puech 2010:757).

Due to the lack of papers, we could not find any information whether the *M*-function is able to solve one of the problems of the *K(d)*-function mentioned in section 3.2. In our opinion, with exception of point 2, it does not. Central aspects, such as the computation of benchmarks, the usage of step-intervals and the exclusion of distances above the median, are similar for the *M*- and the *K(d)*-function. By virtue of this fact, we will not include the *M*-function in our empirical results – on the one hand because the D&O-index is more established, on the other hand to keep the focus on the solution of the mentioned problems that affect both indices to our knowledge.

4   Defining a cumulative density function

Keeping the mentioned problems of the D&O-index in mind, we will now present our new method that consists of a cumulative and a density part. In what follows, we will first present the function's mathematical background. In section 5, we will discuss the empirical results and the advantages and disadvantages when comparing the two existing indices.
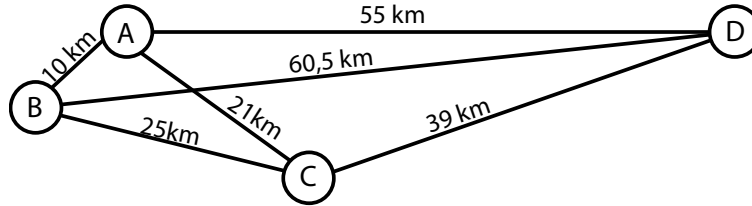
Figure 2: Distances between 4 illustrative MST-firms

Consider 4 MST-firms (A-D). For each firm, an average inverted distance D_i is built as follows:

$$\widetilde{D}_i = \frac{1}{J-1} \sum_{j=1, j \neq i}^{J} \left(d_{i,j}\right)^{-1}. \tag{5}$$

Similar to the *M*-function, an average is established to make values comparable across industries because the term $\frac{1}{J-1}$ makes the index independent of the number of firms or plants. For firm A in our example (Figure 2), its average inverted distance[10] $\widetilde{D}_A$ is: $\frac{1}{3} \cdot \left(\frac{1}{10km} + \right.$

---

[10] A similar computation has been conducted by Sorenson & Audia (2000) but not in a context of index-based test statistics.

$\frac{1}{21km} + \frac{1}{55km}) = 0.055 \left[\frac{1}{km}\right]$. In order to obtain a more easily interpretable value, we might re-invert the $\tilde{D}_i$ value:

$$D_i = \left(\frac{1}{J-1} \sum_{j=1, j \neq i}^{J} (d_{i,j})^{-1} \right)^{-1}, \tag{6}$$

what leads to a weighted average of the distances between firm i and all other firms. Let us consider again firm A that shows a $\tilde{D}_A$ value of $0.055[\frac{1}{km}]$. We now obtain:

$$D_i = \left(0.055[\frac{1}{km}] \right)^{-1} = 18.18 \ km. \tag{7}$$

This means that the weighted average of the distance of other firms to firm A is 18.18 km. The lower its D_i-value the higher a firm is concentrated in space. In comparison to the other firms, A reaches the lowest D_i, closely followed by B (19.23 km) and C, whereas D (50 km) is less concentrated.

The method of average inverted distances has two advantages: It detects local clustering and allows including all firms in a given area, as the inverted value $(d_{i,j})^{-1}$ for long distances becomes zero and thus has little impact on a firm's D_i. Notwithstanding there is also a disadvantage: D_i values become zero when distances are very small. Consider a firm E, located 0.2 km east of D. Then, firm D and E, both reach D_i values that are smaller than that of firm A. This is an unwanted bias because firms shall only reach a small D_i when they are generally clustered and not just because they show one single close neighborhood.

In order to deal adequately with small distances, we need a threshold that groups such values. In our empirical work we tested three thresholds[11] from which the 5km threshold performed best. We suggest that the choice of the threshold should always depend on the object of research. In our example, a 5km threshold is a reasonable choice because the costs and ability for communication and interaction between MST-firms should not differ that much between 0 and 5 km. So formula (6) turns to:

$$D_i = \left(\frac{1}{J-1} \sum_{j=1, j \neq i}^{J} \frac{1}{\max\{5km, d_{i,j}\}} \right)^{-1} \tag{8}$$

With the purpose of testing whether the D_i values of the MST industry are significantly higher than those values expected for a random distribution we need to build a benchmark. Out of the 20,000 manufacturing firms, 4,000 plants are drawn whose D_i values are computed according to formula (8). Now we have two samples with 861 and 4,000 single values

---

[11] we tested a 0 km, 5 km and 10 km threshold

that represent the MST- and a random/benchmark industry[12]. Since every D_i stands for a firm's degree of spatial concentration as an interval-scaled variable, standardized statistical tests can be applied. There are three options, which all provide different information:

(1) We can compare the distribution of the D_i values calculated for the studied firm population and the benchmark firm population. A standard Kolmogorov-Smirnov-test can be applied, answering the question of whether the studied firm population deviates in its spatial distribution from the benchmark case.

(2) We can check whether the mean value or median of D_i for the studied firm population is different from the benchmark value. Since usually D_i values are not normally distributed, a Mann-U-test can be applied. This provides information of whether the studied firms are, on average, more or less concentrated than the total firm population. However, a firm population might be at the same time more concentrated and more dispersed, as we will show below, so that the average has to be interpreted carefully.

(3) We can study each level of localization and its frequency separately. Up to now, we have discussed the cumulative part of our function. Its density part is similar to the $K(d)$-function using the D_i values to build a kernel density estimations. For an industry I this is given by:

$$g_I(D) = \frac{1}{nh} \sum_{n=1}^{N} f\left(\frac{D - D_i}{h}\right),$$

(9)

where h is the optimal bandwidth and f the Gaussian kernel function. In the same way the density function $g_B(D)$ can be calculated for the benchmark population.

In the same way to the $K(d)$-function, we obtain two density curves whose intersections can be interpreted. Figure 3 plots the D_i-densities of an illustrative industry (solid line) and a random industry (dashed line). Note that in contrast to the $K(d)$-function all distances are considered. The density can be easily interpreted with respect to spatial concentration. On the one hand, the illustrative industry shows clearly more concentrated firms because smaller D_i-values have a higher probability as in our benchmark case (lengthwise striped area). On the other hand, there are also some firms that are more dispersed (horizontally striped area), showing higher probabilities for large D_i-values in comparison to the benchmark. Therefore, the illustrative industry shows both global dispersion and concentration.

To state whether an industry is more characterized by dispersion or localization we need to compare the areas of intersection of the two curves. Let $g_B(D)$ be the function that describes the density curve of our benchmark and let m be the mean of its values (dotted line in Figure 3). The value of concentration Θ_{conc} is the sum of all areas of intersection where the density curve of the investigated industry $g_I(D)$ lies above $g_B(D)$ and whose D_i values are below

---

[12] Our approach does not depend on the number of firms considered, so that this number can be different for the studied firm sample and the benchmark firm sample.
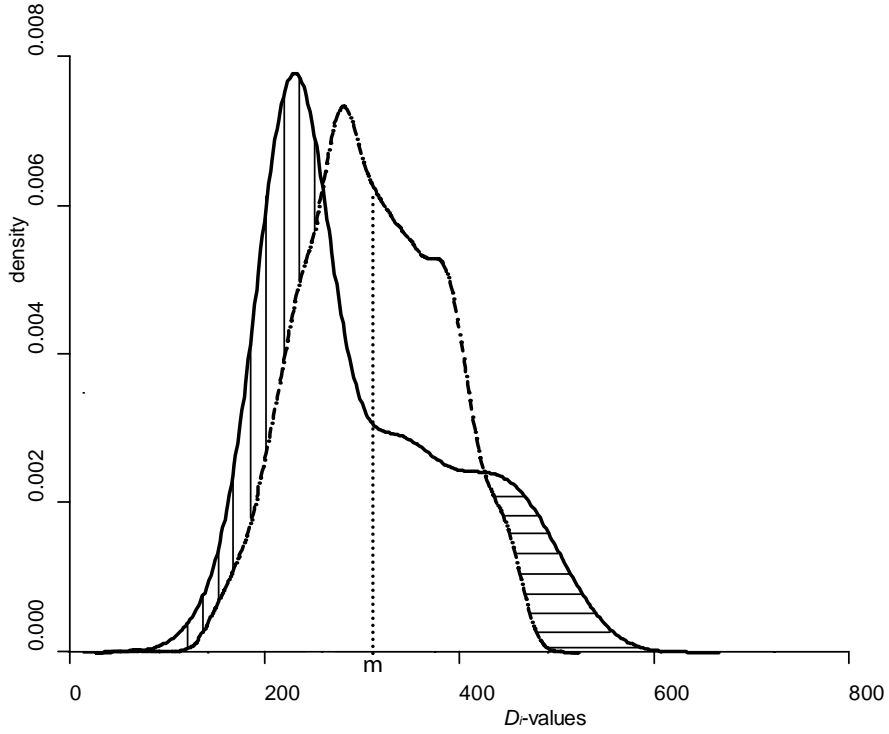
Figure 3: D_i-density for an illustrative industry

the mean m (vertically striped area). Mathematically this is expressed by the indefinite integral:

$$\Theta_{\text{conc}} = \int_0^m max\{0, g_I(D) - g_B(D)\}\ dD. \tag{10}$$

The value of dispersion Θ_{disp} is computed in the same way to Θ_{disp} using values that lie above m:

$$\Theta_{\text{disp}} = \int_m^\infty max\{0, g_I(D) - g_B(D)\}\ dD. \tag{11}$$

Finally, we can define Θ as a conjoint index of dispersion and localization:

$$\Theta = \Theta_{\text{conc}} - \Theta_{\text{disp}}. \tag{12}$$

As the area of a density functions sums up to 1, Θ can reach values from -1 (not one firm is more concentrated than any random firm $\hat{=}$ absolute dispersion) to 1 (absolute concentration). A value of zero indicates that an industry is neither characterized only by dispersion nor by localization. However, this does not automatically imply that its localization pattern is analog to the random industry. An industry, such as the illustrative industry in Figure 3, can reach Θ values around zero, but the Kolmogorov-Smirnov-test applied to the two distributions would state that the studied firm population's localization pattern clearly differs from the distribution for the total firm population.

5 Empirical testing and results

(a) MST-firms (N=861)   (b) benchmark-firms (N=4000)
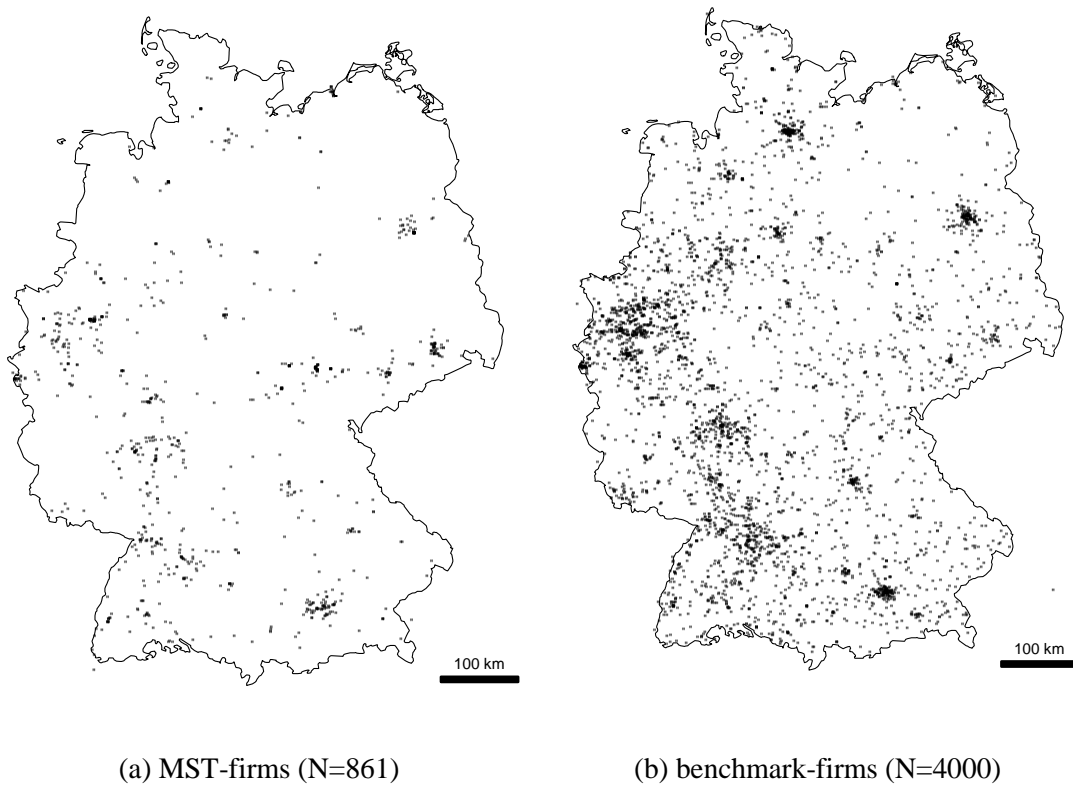
Figure 4: Distribution of the MST-firms and the benchmark-firms in the area under investigation

When considering Figure 4, MST- and bechmark firms show a similar localization pattern at first glance. Firms are clearly concentrated in the west and south of Germany, while the east (former GDR) shows less firms. However the MST industry seems to be less localized outside conurbations. Whether these differences are significant or not shall now be tested by the *K(d)-* and our cummulative-density function.

5.1 *K(d)*-function

Figure 5: *K*-density and global confidence bands for the MST industry

With respect to Figure 5, we can state that the MST industry is globally concentrated as their density curve lies above the upper global confidence interval for the distances of 0-30 km and 290-360 km. For most of the other distances, the MST shows fewer neighborhoods than expected according to the benchmark calculation. The data suggests that there are several clusters that are located at larger distance to each other. Γ reaches to a value of 0.183.

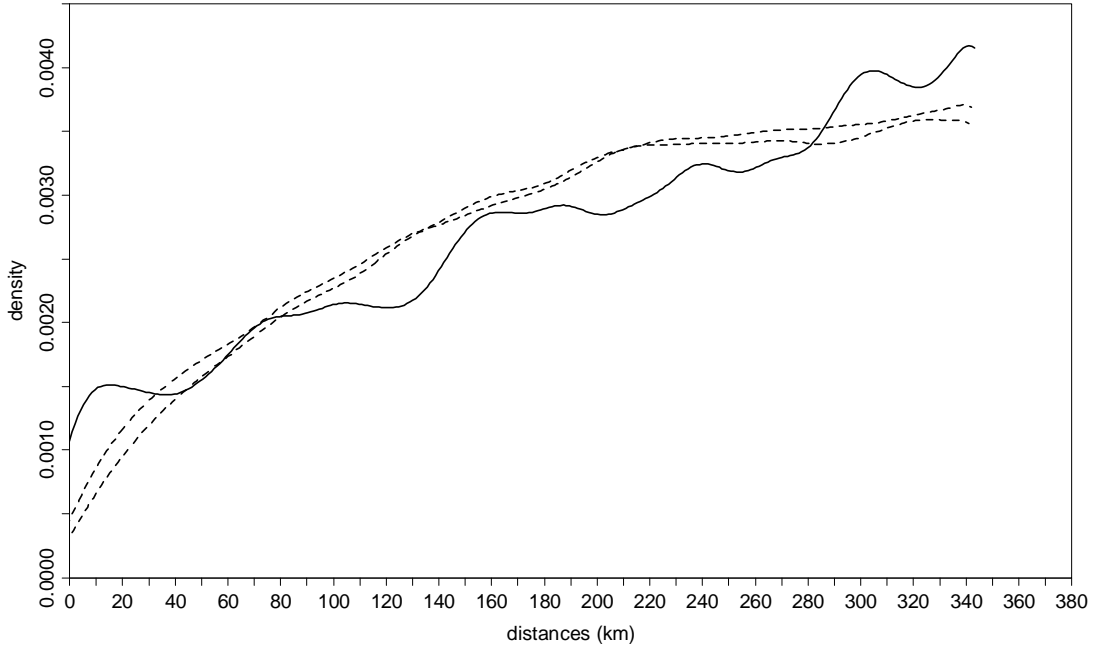For all intervals, the distances between the upper and the lower band are quite small, but this confirms the findings of Koh and Riedel (see Koh & Riedel 2009: 9). Although the data is smoothed, the *K(d)*-density of the MST industry exhibits considerable fluctuations. This might be owed to the relatively small sample of 861 firms, associated with a large area under investigation. The size of the area also required a subsequent grouping of values in order to compute the global bands (see section 3.2). With a step size of 1 km, even the highest band was built of more than 300 simulations. We had to reduce the step size to 5 km so that the D&O-test is faced with a secondary MAUP for our analysis.

5.2 Cumulative-density function

The results of the Mann-*U*-test and the Kolmogorov-Smirnov-test show that MST and benchmark firms clearly have a different localization level (see Table 1 and 2 in the appendix). The median and mean of the MST industry are approximately 20 % lower than those of the whole firm population in Germany (see Table 3 in the appendix). In line with the D&O-index, we can state that the localization pattern of the MST industry differs from that
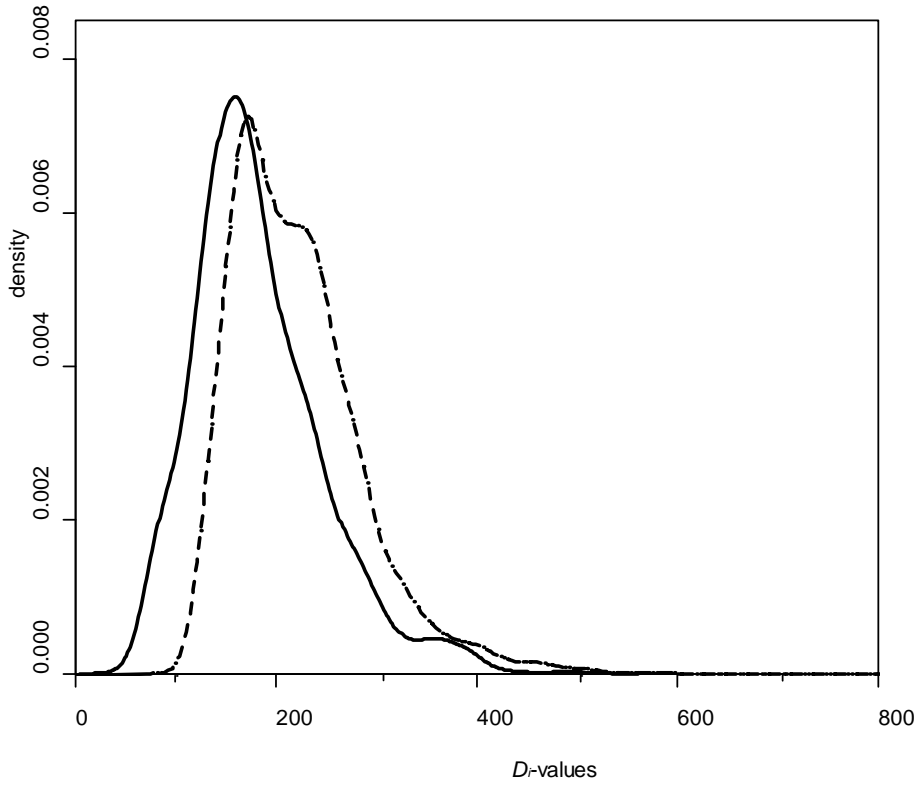
Figure 6: D_i-density for the MST industry and randomly drawn firms

of the total firm population and that MST-firms are more concentrated in space. The inter-sections of the kernel density estimations confirm these findings: For the average inverted distance from 25 to 180 km the MST industry (solid line) reaches higher D_i-density values than the random firm population (dashed line). Hence, we have many firms that are located unusually near to other firms. The conjoint index of concentration and dispersion Θ reaches a value of 0.224. Though the density curves of the $K(d)$ and the D_i values are clearly differ-ent, both functions give similar statements about the degree of spatial concentration of the MST industry. The fact that Γ is slightly lower than Θ is due to non-observance of values above the median where the MST-industry shows global concentration (see Figure 5).

As mentioned in section 1, our method also allows for identifying the localization of highly clustered firms, as we obtain a localization measure for each single firm. For this purpose, we simply select the firms that lie in the first quartile of the MST industry with respect to their D_i values. Figure 7(a) shows these firms. Most of the MST clusters are located in the south-western part of Germany (1-4). Furthermore, we find clusters in the Ruhr area (5) and in Eastern Germany: Jena, Chemnitz, Dresden and Berlin (6-9). This confirms the sugges-tion of the D&O-index that there are several MST clusters located at a larger distance to each other.

(a) First quartile of MST-firms (N=216)   (c) Rhine-Main cluster

(b) Berlin cluster

Figure 7: Localization of German MST clusters

Taking a closer look at the clusters reveals an interesting difference in their geographical scope. The Berlin MST cluster consists of eight firms within a distance of approximately 1 km, while the Rhine-Main cluster contains 19 firms in a much larger area (distances up to 70 km). An interesting aspect of further research is to investigate whether communication or sense of belonging are sensitive to the geographical scope of these clusters.

5.3 Features of our cumulative-density function

After having presented the mathematical background and the empirical results of our new function, we will now show its advantages and disadvantages compared to the existing two indices.

(1) Significance of results: We use three different methods that allow for a comprehensive test for localization patterns. The Kolmogorov-Smirnov-test checks whether the two samples originate from the same population. In comparison to the D&O-index, this enables us to detect even patterns that do neither show clear dispersion nor clear concentration, but nevertheless differ from the distribution of the total firm population. The comparison (Mann-U-test) of the median and mean gives an indication about the differences in average values. The conjoint index Θ represents the strength of concentration/dispersion over all distances. This index is not affected by the size of the area under investigation and can be easily compared between different industries.

(2) Inference to localization: Our method is able to deliver insights into the spatial localization of a firm and its degree of spatial clustering. As every firm has its own D_i value, e.g. the first quartile can be selected to show firms that are highly concentrated. To our knowledge this feature has not yet been introduced to MAUP-free methods as the other two indices do not allow inference to localization of single firms. The D&O-index only regards the density of neighborhoods at a certain distance but cannot detect whether firms, showing neighborhoods at small distances, are generally clustered (consider the example of firm D and E). The *M*-function allows more insight into this question. For example, one could count the numbers of neighborhoods at a 20 km radius and then present the upper quartile of firms with the highest number. However this attempt suffers from its dependence on the choice of the observed radius: In highly concentrated industries, certain firm might be included in the quartile for one radius but not for another. Increasing the radius, some firms might even jump in and out of this group repeatedly.

(3) Low risk of secondary-MAUP: In contrast to the existing indices, our cumulative-density function does not divide the research area into intervals, thus avoiding the risk of a secondary-MAUP. Furthermore, the median-distance of the population is not needed as all distances are included. The only restriction of our function in this aspect is its threshold that groups small distances so that it is a right-continuous function.

(4) Low risk of wrong benchmarks: As each D_i-value of our random-industry is built of the weighted distances to 3999 firms, abnormal small or large distances between random firms do not affect results when a threshold is used. Thus the computation of benchmarks can be reduced from 1000 to 1 iteration. The number of 4000 firms seems to be an appropriate value for our purpose. We generated several benchmark simulations and tested them for their equality using the Mann-U-test that showed significance equality. A number of 6000 firms did not change results.

(5) Low Computational requirements: This central feature derives from the two last points: As the research area is not divided into intervals, the computation has to be performed only once and not for each interval. Thus the run-time of our function only depends on the observed numbers of firms but is independent from the research area's size. Moreover, the computation of benchmarks can be reduced from 1000 to 1 iteration. In our empirical work, the computation of our function was around 85 times faster than that of the *K(d)*-function. This advantage becomes even more obvious, when multiple industries in one area under investigation are considered because the same random D_i-values can be used as the benchmark for all industries. The computation for a test of all German manufacturing industries should take less than one day.

Besides the mentioned advantages, our function also shows a central weakness: Each D_i-value represents the average inverted distance from one to all other firms, but it cannot state at which exact distances concentration or dispersion occurs. This feature is the clear strength of the D&O-index in comparison to the *M*-function and our method. However, in contrast to

the *M*-function, our cumulative-density function is able to give solutions to the problems of the D&O-index. Therefore, the choice among these methods might well depend on the observed number of firms and the area under investigation. When both parameters become huge, our new method has clearly many advantages due to its fast computation and its rigorous test for localization patterns.

6   Conclusions

In this paper, we have introduced a cumulative-density function as a new MAUP-free statistic method that fulfils the first four criteria of Duranton and Overman, improves the fifth criteria, is efficient in its computational requirements and allows for identifying clustering and clusters. Our approach offers a number of indices. First, it provides an interval-scaled value of concentration for each firm. Second, we can test for differences in the distribution of these values. By this, our method provides indices for excess concentration and dispersion of an industry in comparison to the total economy but we can also detect non-random patterns that do neither show clear dispersion nor clear concentration. Third we defined a conjoint index as the difference between concentration and dispersion. Hence, our approach provides a number of indices that can be used for different purposes in further studies.

Both the D&O- and our new index have shown that the German MST industry is concentrated in space, especially at small distances. The localization of the most clustered MST firms revealed significant differences in the geographical scope of the clusters. An analysis of the different scopes of clusters might be an interesting object for further research.

Another starting point concerns the D_i-values as the basic concept of our index. In contrast to all other distance-based methods, our index assigns to every firm a unique D_i-value that represents the firm's degree of spatial concentration as an interval-scaled variable. This does not only enable the usage of significance tests (such as Kolmogorov-Smirnov-test and Mann-U-test), but D_i-values can also be applied in regression models. By means of this transfer, distance-based methods leave their restriction on measuring (co-)localization only and enable us to investigate the diverse nature of firm-localization choice from a micro-geographic perspective.

7 References

Bathelt, Harald and Glückler, Johannes (2003): Toward a relational economic geography. In: Journal of Economic Geography 3 (2): 117-144.

Duranton, Gilles; Overman, Henry G. (2005): Testing for Localization Using Micro-Geographic Data. In: Review of Economic Studies 72: 1077–1106.

Ellison, Glenn; Glaeser, Edward; Kerr, William (2009): Data and Empirical Appendix to "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." (http://econ-www.mit.edu/files/3200).

Ellison, Glenn; Glaeser, Edward; Kerr, William (2010): What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. In: American Economic Review 100 (2010): 1195–1213.

Klier, Thomas; McMillen, Daniel P. (2008): Evolving Agglomeration in the U.S. Auto Supplier Industry. In: Journal of Regional Science 48 (1): 245–267.

Koh, Hyun-Ju; Riedel, Nadine (2009): Assessing the Localization Pattern of German Manufacturing & Service Industries – A Distance Based Approach. In: Oxford University Centre for Business Taxation Working Papers 09 (13): 1-30.

Marcon, Eric; Puech, Florence (2010): Measures of the geographic concentration of industries: improving distance-based methods. In: Journal of Economic Geography 10 (5): 745-762.

Miller, Harvey J. (2010): The data avalanche is here. Shouldn't we be digging? In: Journal of Regional Science50 (1): 181–201.

Openshaw, S. (1984): The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography 38.

Sorenson, Olav; Audia, Pino G. (2000): The Social Structure of Entrepreneurial Activity: Geographic Concentration of Footwear Production in the United States, 1940–1989. In: The American Journal of Sociology 106 (2): 424-462.

Vitali, Stefania; Mauro, Napoletano; Fagiolo, Giorgio (2009): Spatial Localization in Manufacturing: A Cross-Country Analysis. In: LEM Working Paper Series 2009 (04): 1-37.

Woodward, Douglas; Guimarães, Paulo (2009): Porter's cluster strategy and industrial targeting. In: Goetz, Stephan J.; Deller, Steven C.; Harris Thomas R. (eds.): Targeting Regional Economic Development: 68-84.

8 Appendix

	IS_MST	N	Mean-rank	Rank-sum
Di	BENCHMARK	4000	2586,36	10345456,00
	MST	861	1705,55	1466774,00
	Total	4861		

	Di
Mann-Whitney-U	1096544,000
Wilcoxon-W	1466774,000
Z	-16,701
Asymptotic significance (2-sided)	,000

Table 1: Mann-Whitney-Test

Most Extreme Differences	Absolute	,288
	Positive	,000
	Negative	-,288
Kolmogorov-Smirnov-Z		7,654
Asymptotic significance (2-sided)		,000

Table 2: Kolmogorov-Smirnov-Test

	N	Minimum	Maximum	Mean	Median	Standard deviation	Variance
BENCHMARK	4000	121,952	1454,423	220,365	209,469	68,330	4669,031
MST	861	73,988	490,943	181,824	169,412	63,769	4066,548

Table 3: Descriptive statistics