Paper to be presented at the DRUID 2012

on

June 19 to June 21

at

CBS, Copenhagen, Denmark,

Testing for Clustering of Industries Evidence from micro geographic data

Tobias Scholl
Goethe Universität Frankfurt
Geography/Sociology
tobias.scholl@nephele-idea.de

Thomas Brenner
Philipps University Marburg
Faculty of Geography
brennert@uni-marburg.de

Abstract

We present a new statistical method that detects industrial clusters at a firm level. The proposed method does not divide space into subunits whereby it is not affected by the Modifiable Areal Unit Problem (MAUP). Our metric differs both in its calculation and interpretation from existing distance-based metrics and shows four central properties that enable its meaningful usage for cluster analysis. The method fulfills all five criteria for a test of localization proposed by Duranton and Overman (2005).

Jelcodes:C40,C60

Testing for Clustering of Industries

Evidence from micro geographic data[1]

Abstract

We present a new statistical method that detects industrial clusters at a firm level. The proposed method does not divide space into subunits whereby it is not affected by the Modifiable Areal Unit Problem (MAUP). Our metric differs both in its calculation and interpretation from existing distance-based metrics and shows four central properties that enable its meaningful usage for cluster analysis. The method fulfills all five criteria for a test of localization proposed by Duranton and Overman (2005).

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

JEL classifications: C40, C60, R12

[1] We thank Gilles Duranton, Henry Overman and Stefania Vitali for their helpful comments.

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 spatial econometric analysis to such data (Miller 2010: 182). Most of the papers still deal with the comparison of regions but do not concern the real spatial position of economic actors such as firms or research institutions. This is mainly due to the fact that the large majority of quantitative methods in spatial economics such as the LQ-, the Elison & Glaeser- or the Gini-Index base on the comparison of spatial-subunits in a considered area. In economics, these indices are wildly used to determine a region's specific industrial pattern and to check whether there is a high concentration that could indicate an industrial cluster. The today's common understanding of industrial clusters was mainly promoted by Michael Porter. Though Porter's cluster concept bases on individual firms and their interactions, the detection of clusters is usually conducted through the LQ-index that computes the ratio between the regional employment of industry *i* in region *r* and that industry's national share. These approaches can be criticized in two ways. First, as Woodward and Guimarães (2009) point out, the LQ-index is too simple to detect local clusters, as it only demonstrates a regions trend towards specialization (Woodward & Guimarães 2009: 77 f.). Second, which cluster are identified with such an index depends crucially on the choice of regional boundaries. Using only data on firm location and size, the first problem cannot be cured. Such data can only be used to identify specialization in space. Here we intend to tackle the second problem.

Though other more complex indices such as the Elison & Glaeser-index can circumvent several problems of the LQ-index, there exist three general problems of metrics that base on the comparison of regions:

1. Results are always sensitive to the chosen level of aggregation – e.g. counties, cities or states. Outcomes may vary to a large degree when changing from one aggregation level to another. Furthermore, spatial divisions normally do not depend on economic characteristics but on administrative classifications.
2. The indices do not provide a clear statement from which threshold on, a region's specialization indicates the presence of clusters.
3. The indices cannot identify the spatial dimension of a cluster but only regions with a high level of specialization. Analogous to point 1, clusters do normally not follow boundaries, especially not between regions that are located in the same country.

A solution to these problems is the usage of distance based methods that do not discretize an area under investigation into spatial subunits but concern it as a continuous space. There are only few papers and even less models that provide such a quantitative spatial analyses of

empirical economic activity. One of the first papers in this context was published by Duranton and Overman in 2002[2], in which the authors examine the concentration of manufacturing firms in the U.K. In comparison to spatial aggregated metrics, distance based indices allow the detection of the spatial dimension of clusters as they provide measures of significant spatial concentration/dispersion for single distance intervals (e.g. km).

Despite these advantages, distance based methods have rarely been used for empirical cluster analysis. While the availability of micro geographic data becomes less and less a difficulty, there are still two central problems of distance based methods. First, the available methods do not provide insight into the spatial location of clusters. The metrics allow for a detection of significant dispersion or concentration at specific distances but they cannot present the geographical location of highly clustered firms – in other words the metrics concentrate on distance intervals instead of firms. This is a shortcoming in comparison to spatial aggregated indices that can detect the localization of highly specialized regions. However, to our mind, the graver problem lies in the computational complexity of distance based methods. This is a central issue that is mentioned in almost every publication dealing with those indices. Kosfeld et al. (2011) point out that "even with high-speed computers, pure CPU time of estimating and testing the K function for a single branch of industry with several thousand plants by simulation is not a question of hours but of days." (Kosfeld et al. 2011: 312).

With regard to this situation, the aim of our paper is to present a new firm-level cluster index that is distance-based but differs both in its calculation and interpretation from existing distance-based metrics. Four central properties enable the metric's meaningful usage for cluster analysis:

1. Through three different statistical tests, our cluster index determines to what extent an industry is more concentrated (or more dispersed) in space in comparison to the overall industrial agglomeration.
2. The index reveals the spatial location of highly clustered firms and thus gives insight, both into the spatial dimension and position of firm-clusters.
3. The computational requirements of our method are comparatively low so that even large industries and large areas under investigation can be examined.
4. The metric computes a unique degree of concentration for each firm as an interval scaled variable. This enables an easy transfer to regression or correlation analysis with other firm-plant specific properties such as growth-levels or patent activity.

In the following sections, we will demonstrate our firm-level cluster index by means of theoretical considerations and by means of the German micro technology industry. Micro technology, or microsystems technologies (abbr. MST), is a high-tech industry that combines

[2] Working paper 2002, paper 2005

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. Thus, the MST industry should be an appropriate industry to test our method on the basis of real firms.

The rest of the paper is organized as follows: section 2 presents the data basis used in the empirical part. Section 3 defines the Modifiable Areal Unit Problem and presents the recently must popular distance-based index whereas section 4 outlines our new approach. In section 5 we show the results for the different methods used in this paper and discuss the advantages and disadvantages of our cluster-index. Finally, section 6 concludes and outlines new possibilities for further research.

2 Data

The dataset for the empirical part 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).

As our benchmark we used 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 all manufacturing firms' exact location (161,729 plants).

3 Spatial statistics and the MAUP

3.1 MAUP affected indices

The problems of spatial aggregated indices are a well-known issue in spatial econometrics and were first described by Openshaw (1984) with the term Modifiable Areal Unit Problem (MAUP). In the following, we will discuss the properties of the MAUP by means of the LQ-index that is the most common test to detect clusters. For industry *i* in region *r,* the LQ-index is defined as:

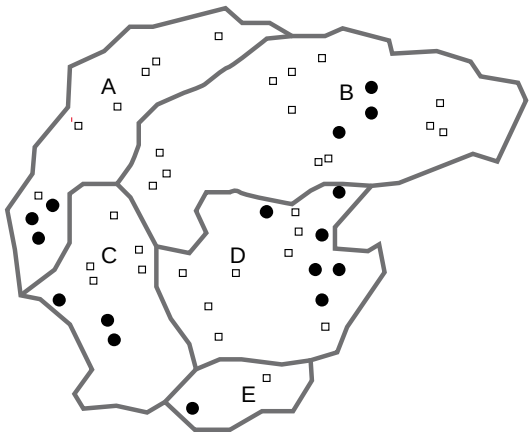
$$LQ = \frac{E_{i,r}/E_{i,n}}{E_r/E_n}, \tag{1}$$

where *E* stands for the number of employees and *n* for the country under investigation. For a theoretical demonstration, consider a country that exists of five regions A-E where every firm has one employee (see Figure 1 (a)). In our example, the share of the industry under investigation (dots) follows in all regions more or less the overall industrial localization (squares). The LQ-index for region A is 1 and 0.6 for region B. This outcome changes significantly when the aggregation level is lowered (see Figure 1 (b)) as most firms under investigation are now located in only a few regions. In this case, region A reaches a LQ of 0 while region B shows a high concentration (LQ of 3). This situation is called the scaling problem of the MAUP because results always depend on the chosen level of aggregation and there is no ex-ante correct level for a survey.

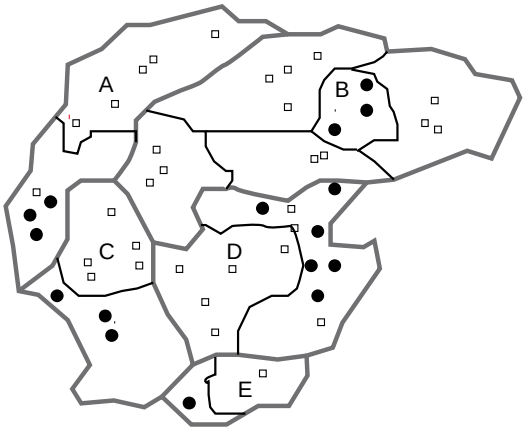
Beside the problem of the aggregation level, it is obvious that results also depend on the regions' boundary lines. Especially inside a country, it is not reasonable to assume that economic structures follow boundary lines. The arbitrariness of boundary lines is referred to as the zoning-problem of the MAUP.

A further shortcoming of MAUP-effected indices is their problem of giving an indication for the statistical significance of results. First, there is no agreement in the literature on a common threshold for the LQ value in order to detect the presence of clusters. Second, the index is not able to give a statistically validated statement about the overall localization pattern of the industry under investigation.

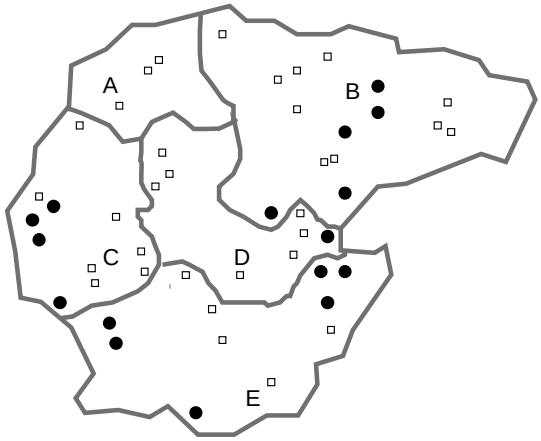
The mentioned problems demonstrate the need of distance-based methods that do not discretize the area under investigation into spatial subunits but take each single distance between the considered firms into account (see Figure 1 (d)). In the following we will discuss the index by Duranton and Overman (2005) that is so one of the most established metrics for MAUP-free investigations of economic activity.



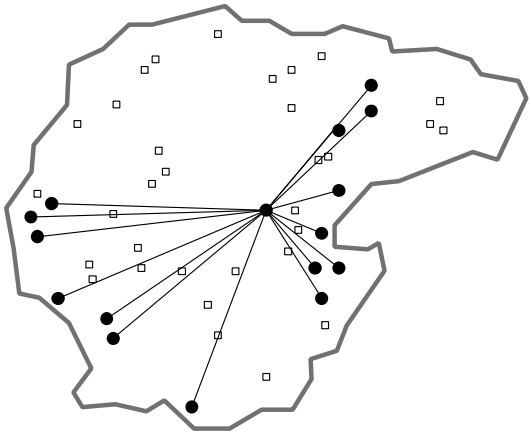
(a) Hypothetical country with 5 regions



(b) Scaling problem: Change of aggregation level



(c) Zoning problem: Change of boundaries



(d) MAUP-free distance based approach

Figure 1: Measuring spatial concentration: Aggregated and none aggregated approaches

3.2 The D&O-index

With respect to the mentioned problems of MAUP-effected indices, 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).

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.

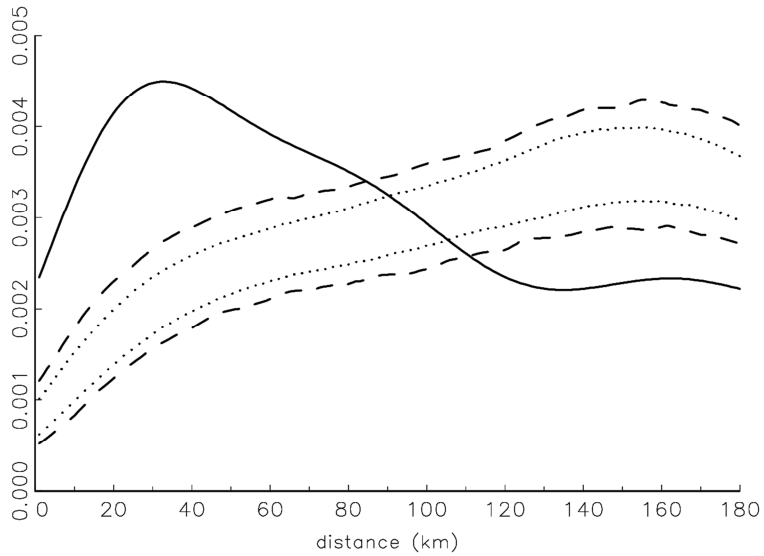


Figure 2: *K*-density, local confidence intervals and global confidence bands for an illustrative industry. Source: Duranton & Overman 2005

To this end, a smoothed density over all neighborhoods, expressed by the term *K(d)*, is used. The first step to compute *K(d)*-values is to build the geographical distances[3] between all possible pairs of firms so that one gains $N(N\text{-}1)/2$ unique bilateral distances. 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. Any high *K(d)*-value outside the distance of the median value could be interpreted as dispersion but Duranton and Overman see this information as redundant[4] (Duranton & Overman 2005:1086). The last step is smoothing the observed numbers using a Gaussian kernel function. Hence the formula is:

$$K(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{i,j}}{h}\right), \tag{2}$$

where h is the optimal bandwidth[5] and f stand for the kernel function.

The solid line in Figure 2 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.

[3] We computed orthodromic distances instead of Euclidian distances, proposed by Duranton & Overman (2005).
[4] We disagree with this statement as it is possible that an industry can show simultaneous concentration and dispersion (see section 4.1).
[5] Optimal bandwidth: $1.06sn^{-0,2}$, where n is the observed number and s is the standard deviation (Klier & McMillen 2006: 12).

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: Let *N* be the number of firms in the industry under investigation then we draw *N* firms out of the benchmark population. These 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 (Duranton 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 ascending order. The 5-th and 95-th percentile are selected to compute the *K(d)*-function according to formula (2). 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 2) (Duranton & Overman 2005:1086). The industry in Figure 2 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 50^{th} 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 $50\text{-}1^{st}$ (49^{th}) 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 $50\text{-}*^{th}$ 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 above 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

$$\Gamma(d) \equiv max\left(\hat{K}(d) - \bar{\bar{K}}(d), 0\right), \tag{3}$$

is the index of global localization at a distance d and

$$\Psi(d) \equiv \begin{cases} max\left(\underline{\underline{K}}(d) - \hat{K}(d), 0\right) \; if \sum_{d=0}^{d=362} \Gamma(d) = 0, \\ 0 \; otherwise, \end{cases} \tag{4}$$

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 $\Gamma_A = \sum_{d=0}^{d} \Gamma_A\,(d)$.

Compared to MAUP-affected indices, the D&O-index is a clear improvement. Using single distance intervals, the index is not affected by the zoning and scaling problem of the MAUP and its index of global dispersion/concentration gives a statistical validated value about the overall localization pattern of an industry. Despite these advantages, some minor problems remain. One problematic point are the enormous computational requirements of the index as calculations have to be conducted for each interval, both for the observed industry and for the 1000 benchmark simulations. This is also the reason why almost all papers (except of Duranton and Overman's own publications) only calculate an approximation of the D&O-index (e.g. Kosfeld et al. (2011), Glenn et al. (2010), Vitali et al. (2009), Klier & McMillen (2008)).

A possibility to reduce computational requirements is to lower the number of step-intervals. This is done for instance 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 arbitrariness suffers from the same problems as the mentioned MAUP-affected indices.

Hence, we see a secondary-MAUP-problem: Even if data provide point-localization of firms, high computational requirements or statistical needs might be solved in a subsequent division of space. This is not a specific shortcoming of the D&O-index but holds similar for the other existing distance-based indices as they all base on the evaluation of single distance intervals.

Nevertheless, the D&O-index is a very helpful method for the analysis of spatial concentration. It is based on some considerations that prove to be helpful also for our methodology. However, it does not allow for the identification of the location of clusters, which is the primary aim of our method.

4 Defining a firm-level cluster index

Keeping the mentioned problems of the existing MAUP-free indices in mind, we will now present our new firm-level cluster index that does not focus on single distance intervals but on spatial concentration values for single firms. In what follows, we will first present the function's mathematical background and show its behavior by means of theoretical tests. In section 5, we will discuss the empirical results for the German MST-industry and the advantages and disadvantages when comparing the metric to the D&O-index.

4.1 Mathematical formulation

The basic idea of our cluster index is to compute the sum of inverted distances D_i from one firm to all other firms of the same industry:

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

The term $\left(f(d_{i,j})\right)^{-1}$ stands for all possible functions that compute the inverted orthodromic distance between two points so that close neighborhoods have a high influence on a D_i value while the weight of large distances converges to zero. Obviously, the sum on the right-hand side of Equation (5) increases with the number of observations *J*. Therefore, an average is established to make values comparable across industries. The term $\frac{1}{J-1}$ makes the index independent of the number of firms or plants.

To give an example, consider 4 firms (A-D). For firm A in our example (Figure 2), its average inverted distance[6] D_A using the simple hyperbola function $\left(d_{i,j}\right)^{-1}$ is:

$$\frac{1}{3} \cdot \left(\frac{1}{10km} + \frac{1}{21km} + \frac{1}{55km}\right) = 0.055 \left[\frac{1}{km}\right]. \tag{6}$$

The higher its D_i value the more a firm is concentrated in space. In comparison to the other firms, A reaches the highest D_i, closely followed by B (0.052) and C, whereas D (0.02) is less concentrated. While every firm has its unique D_i value, this does not mean that a D_i value is the outcome of a unique set of distances. For example firm A would reach the same values if the other firms were located at a 500, 500 and 6.21 km distance. This however is not an unwanted bias but the general concept of our metric. In comparison to the D&O-index, our cluster index does not focus on specific distances but gives an approximation of

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

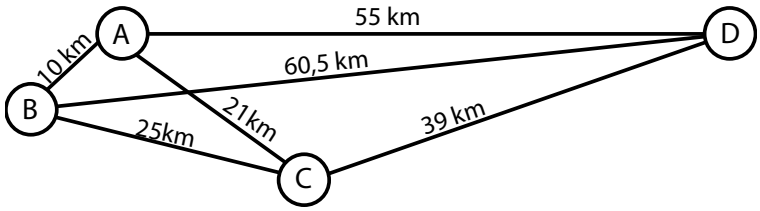


Figure 3: Distances between 4 illustrative MST-firms

the spatial concentration of firms. In the first case, firm A is characterized by two more or less close neighbours while the latter case represents a situation in which only one firm has an influence on the D_i value of the firm. The theoretical demonstration in the next section will reveal two central aspects of our metric: First, the influence of different combinations of distances is negligible when D_i values are built of large observations instead of just four firms. Second, although our index is always just an approximation, its capacity to detect spatial clustering or dispersion out of a random distribution is very high.

Obviously, the outcome of a D_i value does not only depend on the set of neighborhoods but also on the chosen distance function$\left(f(d_{i,j})\right)^{-1}$. Here, we will consider two functions: the simple hyperbola function $\left(d_{i,j}\right)^{-1}$ and the negative exponential function$e^{-\alpha d_{i,j}}$. Both variants show specific strengths and weaknesses. The hyperbola function is maybe the most intuitive one but is problematic when dealing with small distances. In some situations, very close neighborhoods can lead to distorting results for instance when two firms are located in the same building and therefore reach infinite D_i values. In order to deal adequately with small distances, we need a threshold that groups such values. We chose a threshold of 1 km so that the upper value range is the same for the hyperbola and the negative exponential function. Thus, formula (5) for the hyperbola function turns to:

$$D_i = \frac{1}{J-1} \sum_{j=1, j\neq i}^{J} \frac{1}{\max\{1km, d_{i,j}\}} \tag{7}$$

In contrast to the hyperbola function, the negative exponential function does not need a threshold as it converges to 1 for small values. However, this function needs a distance decay factor α. A small α will produce more values around 1 while a large α will increase the function's tendency to zero values. We chose α=-0.05 as this value seemed to be the most appropriate one to make results comparable between the two functions. Due to its exponential character, the negative exponential function has a shorter value range than the hyperbola function. In summary, the negative exponential function's advantage is its independence from thresholds, but it shows a tendency to more extreme results. Using the negative exponential function, formula (5) turns to:

$$D_i = \frac{1}{J-1} \sum_{j=1, j \neq i}^{J} e^{-0.05(d_{i,j})} \quad (8)$$

After having presented the basic calculation of D_i values, the next step is to demonstrate how the index can give statements about the significance of results. Analogue to the D&O-index, we draw random firms out of the population of all manufacturing firms. However, we do not take the same number of firms but a number that is usually much larger than the industry under investigation. The idea behind this procedure is that a large benchmark sample is not influenced by some firms that are located unusually dispersed or concentrated to each other, as a firm's D_i value is built by the distances to thousands of other firms. The number of benchmark firms has to be chosen with respect to the size of the area under investigation. For the size of Germany, this should be more than 1,000 firms. As we will discuss later on, a number of 4,000 seems to be an appropriate value.

Now, an intuitive step would be to calculate D_i values for all drawn benchmark firms according to formula (5). However, for the benchmark values, we are confronted with the fact that this procedure would imply independence between the bilateral distances of the benchmark firms. This is not given as even if firms are independently located, the bilateral distances between them will not be independent (Duranton and Overman 2005: 1084). The existing MAUP-free indices solve this problem by using a bootstrapping approach in order to build their confidence-intervals. As bootstrapping in this context implies the usage of distance intervals, we propose an alternative procedure that considers the dependence of bilateral distances in an analogous manner:

First, we draw a sample of random firms, denoted by I. For each firm $i \in I$ an independent randomly drawn set J_i containing $|I|$-1 firms is built. This allows to calculate for each firm i_1 its D_i according to formula (5) using all firms $j_1 \ldots j_{|N|-1} \in J_1$. This procedure results in a benchmark set of D_i values that are independent to each other as the D_i value of each firm i is built by another set of random firms.

After that step, we can compare the D_i values of our benchmark with the D_i values of the industry under investigation. 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. Therefore we estimate the density distribution of the two populations through a kernel density estimation. For an industry *I* this is given by:

$$g_I(D) = \frac{1}{nh}\sum_{n=1}^{N} f\left(\frac{D - D_i}{h}\right), \tag{9}$$

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.

Just as the *K(d)*-function, we obtain two density curves whose intersections can be interpreted. Figure 4 plots the D_i densities of an illustrative industry (solid line) and a benchmark industry (dashed line). The density can be easily interpreted with respect to spatial concentration. On the one hand, the illustrative industry shows clearly more concentrated firms because large D_i values have a higher probability as in the benchmark case (horizontally striped area). On the other hand, there are also some firms that are more dispersed (vertically striped area), showing higher probabilities for small D_i values in comparison to the benchmark. Therefore, the illustrative industry shows both global dispersion and concentration. Note, that in contrast to the *K(d)*-function, our cluster index is able to detect simultaneous dispersion and concentration as all distances are considered.

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 median of its values (dotted line in Figure 4). The value of dispersion Θ_{disp} 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 the median *m* of the benchmark sample (horizontally striped area). Mathematically this is expressed by the indefinite integral:

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

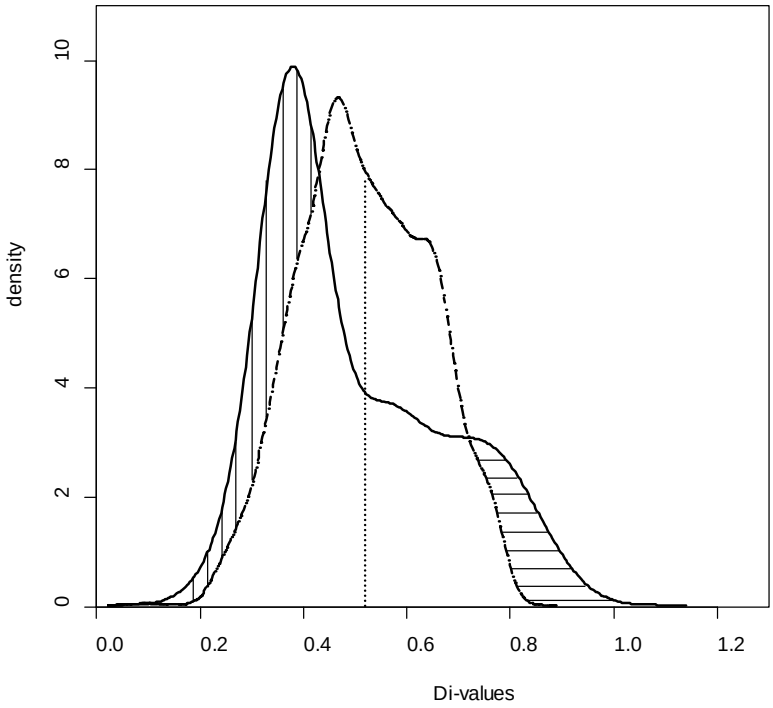


Figure 4: D_i-density for an illustrative industry

The value of concentration Θ_{conc} is computed in the same way as Θ_{disp} using values that lie above *m*:

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

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

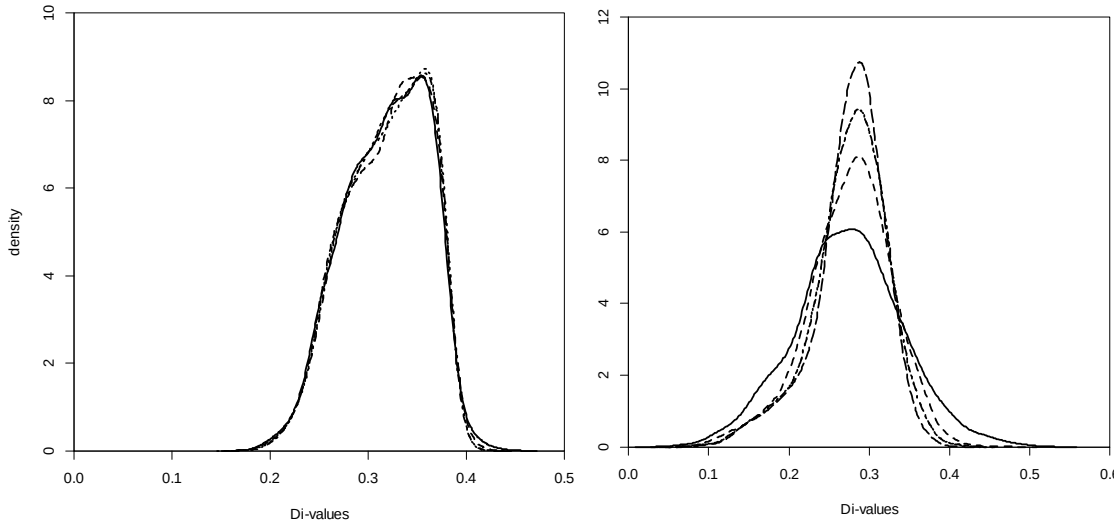
$$\Theta\ =\ \Theta_{\text{conc}} - \Theta_{\text{disp}}. \tag{12}$$

As the area of a density functions sums up to 1, Θ can reach values from -1 (not a single firm is more concentrated than any random firm ≙ 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 4, 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. In comparison to metrics like the LQ-index, Θ enables a clear statement about the overall tendency of an industry to cluster in space.

The most important feature of our metric is its ability to give insight into the spatial localization and dimension of clusters. We are able to detect the firms that are located inside of clusters without the use of any predefined boundaries. The easiest way to detect these firms is to order the firms by their D_i values and select those that lie, e.g., in the first quartile of the

industry under investigation. Another option is using the confidence interval from the benchmark population and identify all firms that have a higher D_i value than the 95-th percentile of the benchmark distribution of D_i values. This procedure allows the detection of statistically validated cluster-cores without any spatial discretization. In the following, we will show what detecting firms in clusters with the help of our method means in practice.

4.2 Theoretical testing



(a) Hyperbola function (threshold=1 km) (b) Negative exponential function (α=-0.05)

Figure 5: KDE for 2000 (solid line), 4000 (dashed line), 6000 (dotted line) and 8000 (dotted-dashed line) CSR-points.

After having presented the mathematical model of our metric, we will now discuss its properties on the basis of theoretical tests. In order to have a realistic setting, consider a 640x876 km rectangle whose dimensions represent the maximum expansion of Germany in geographical latitude and longitude.

The first question that arises is whether the expected D_i value of a firm is invariant to the number of observed plants. This is an important point for the benchmarks because we normalize D_i values to the number of observed plants in order to compare industries and benchmarks of different sizes. Given a known spatial distribution of firms, results should not be affected by scaling the number of observations up or down. Such a known distribution is the Poisson process that generates randomly located points of intensity p in an area. The homogenous Poisson process is used here to simulate complete spatial randomness (CSR).

In order to test for the invariance of distributions to the number of observations, we generate four independent Poisson processes with p=28 (≈2000 points). We start with the first distribution and then concatenate the second the third and the fourth Poisson process. Thus, we

have four sets of CSR distributions with 2000, 4000, 6000 and 8,000 points that have an intersection to each other. D_i values are calculated separately for each set according to the procedure for benchmark firms (see section 4.1). Figure 5 plots the KDEs for the four distributions according to the two distance functions. The hyperbola function shows good results both with respect to the shape of the KDEs and their similarity between the different sets. The curves of the density estimations are bell-shaped and are almost identical for the different point sets. The negative exponential function also shows bell-shaped curves but reveals a higher difference between the four sets and a wider value range. This indicates the functions tendency to more extreme results in comparison to the hyperbola function and a higher dependence on specific locations of firms.

Table 1 (see appendix) also confirms the similarity between the four sets. The Mann-U-Test indicates a high probability that the sets belong to the same population. The KS-Test is a little bit more pessimistic but also indicates high similarity for the majority of the sets. Again, results are better for the hyperbola function in comparison to the negative exponential function. Finally, we can evaluate the changes in the outcome of the D_i values of single firms (points) when the numbers of observations are scaled up. In other words, we mark for instance the firms of the first set (2000 CSR points) and then compute the differences of their D_i values to the values that the same firms reach in the 4000, 6000 and 8000 points set. The average change of the values lies between 3.61 and 0.71 % for the hyperbola function and between 20 and 4 % for the negative exponential function. Again, the outcomes are better for the hyperbola function and results are stable for the last three CSR sets ($\geq$ 4000 firms). Note that the D_i values are built by drawing new independent sets for each point (see section 4.1). This means that for instance the D_i value of a firm in the second set is built by completely different distances to other firms than the same firm in the third or fourth set. Despite the independent calculations, the D_i value of a firm is not highly influenced even if the number of observations is scaled up by a factor of 400 %.

The discussed results show that D_i values are relatively invariant to the number of observations. Thus, normalization is a proper approach of making results comparable between different industries and the overall manufacturing population. The number of benchmark firms has to be chosen with respect to the size of the area under investigation. For the size of Germany a Poisson process with p=2.8 ($\approx$200 points) produces too fluctuating result as the influence of different settings is comparatively high (see Figure 12). For the theoretical tests, a number of 4,000 observations seem to be an appropriate value, as results only slightly differ between a set of 4000, 6000 and 8000 points. This number seems also suitable for the empirical distribution of manufacturing firms in Germany. Figure 11 and Table 2 (see appendix) demonstrate that independent drawings of 4000 firms out of the MARKUS-database do not differ significantly.

The second aspect that we want to test is whether the metric detects spatial clustering in a correct way. Therefore, we generate spatial clusters through a Matérn process (434 points)

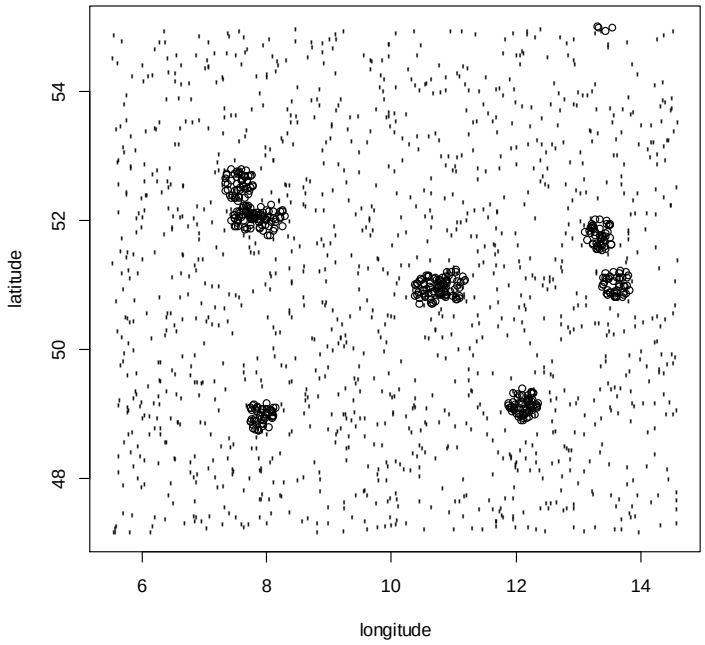


Figure 6: Distribution of a Poisson- (dots) and a Matérn process (circles)

and add 1302 CSR points to this distribution (see Figure 6). Thus, we have a theoretical industry where 3/4 of all firms are randomly distributed and 1/4 is located in a cluster. When applying the metric to this distribution, we can test whether the upper quartile of the D_i values is mainly represented by the clustered firms. In other words, we construct a known training set and then test the classification accuracy of the cluster index. For the hyperbola function, results are again very good. 386 points of the upper quartile are those generated by the Matérn process, so that the hyperbola function has an accuracy of 89 %. The negative exponential function reaches an accuracy of 72 %.

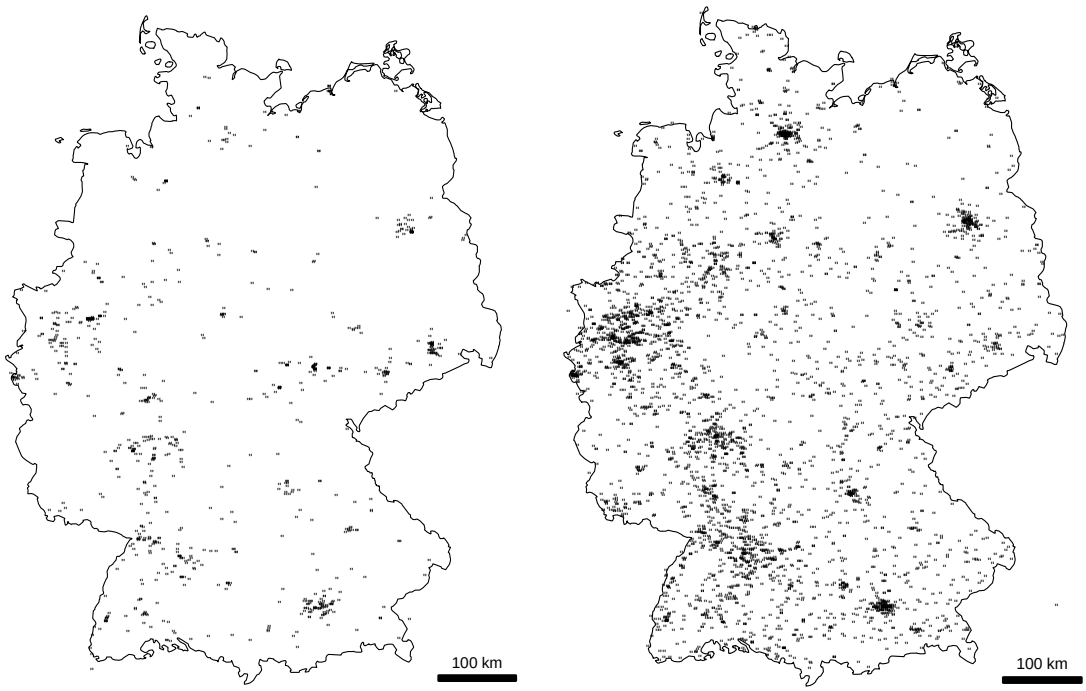
The last feature to be discussed is the metric's capacity to detect simultaneous concentration and dispersion of a point pattern. Therefore, we first generate a Poisson process with 2000 points and select the 600 less concentrated firms by comparing their D_i values. To these points, we add 300 firms, generated by a Matérn process so that the point pattern simulates an industry that shows both concentration and dispersion. Figure 13 plots the KDEs for the two distance functions. Both detect the simultaneous concentration and dispersion as the density curve of the investigated industry (solid line) lies above the benchmark industry (dashed line) for D_i values on the left hand and the right hand of the median. Both, the Mann-U and the KS-Test indicate a high probability that the two samples do not originate from the same population.

Concerning the three mentioned aspects, we can state that our cluster index meets its expectations and that the hyperbola function performs better than the negative exponential one. However, this does not mean that our proposed cluster index can only be run with this distance function but that the choice of the function depends on the research topic. For exam-

ple, the investigation of centrality with the focus on commuters could be modeled by a negative logistic function as this function reflects the willingness of commuting the best (see Vries et al. 2009). For our investigation, there is no ex-ante preferable function and the hyperbola function is more intuitive and shows better results. Thus, in order to the focus on our index general properties, empirical results for the German MST industry will only be presented for the hyperbola function.

5 Empirical testing and results

When considering Figure 7, 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 cluster-index.



(a) MST-firms (N=861) (b) benchmark-firms (N=4000)

Figure 7: Distribution of the MST-firms and the benchmark-firms in the area under investigation

5.1 *K(d)*-function

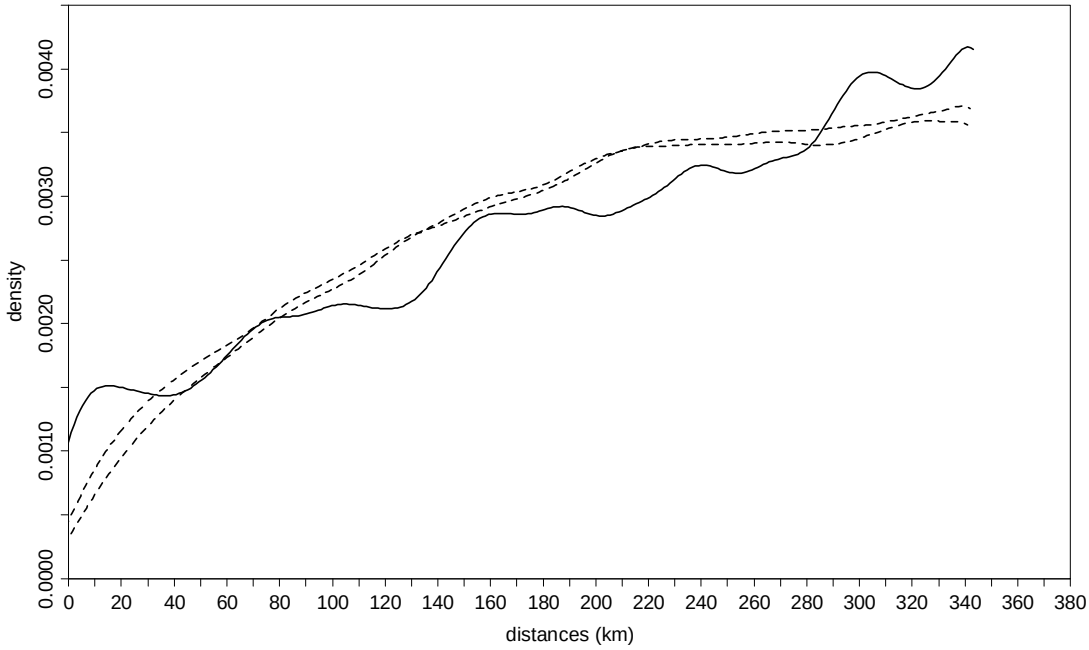


Figure 8: *K*-density and global confidence bands for the MST industry (step-interval: 5km)

We start with the results of the D&O approach for the MST industry in Germany. With respect to Figure 8, 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.

5.2 Firm-level cluster index

Now we discuss the results that we obtain by applying our new approach. The D_i value distributions for the MST industry and the benchmark case are given in Figure 9. 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 3 and 4in the appendix). The median and mean of the MST industry are approximately 30 % and 60 %, respectively lower than those of the whole firm population in Germany (see Table 5 in the appendix). In line with

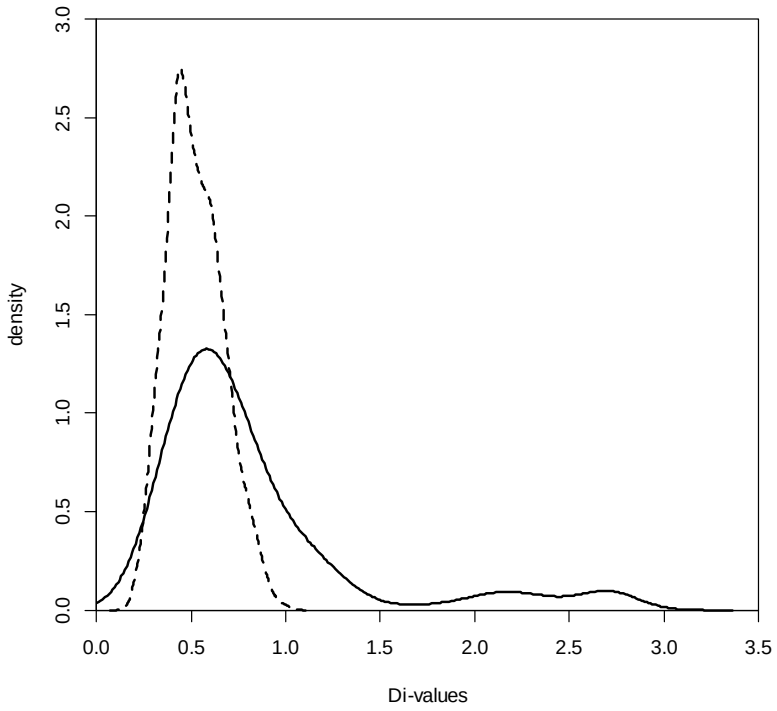
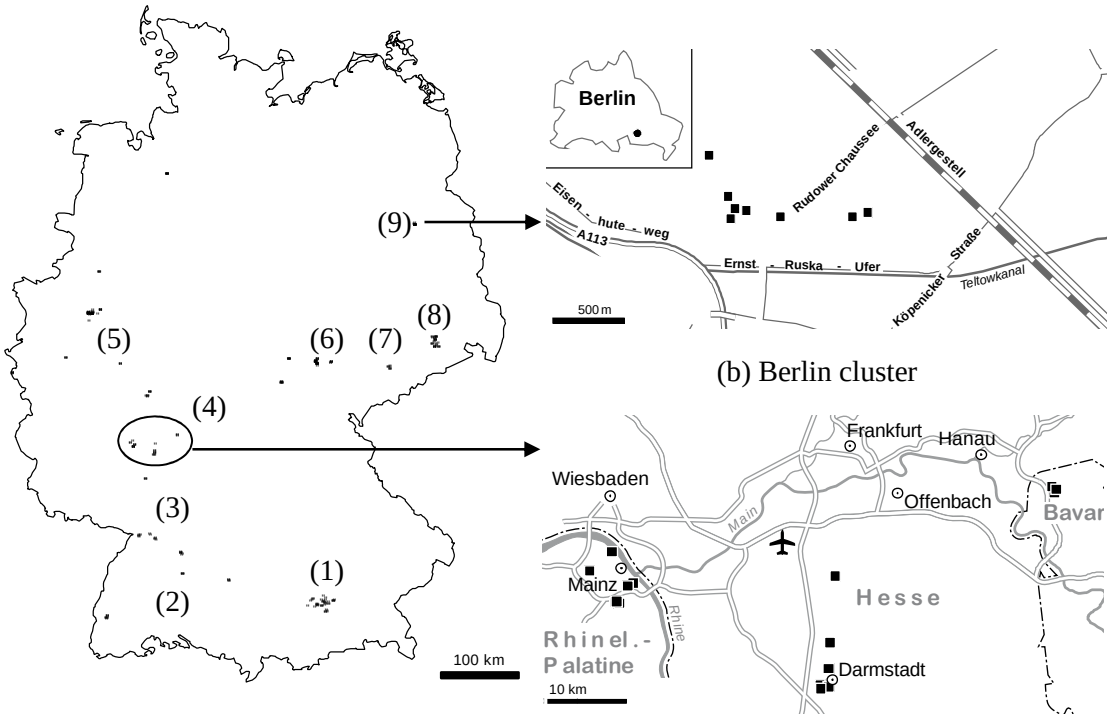


Figure 9: D_i-density for the MST industry and randomly drawn firms

the D&O-index, we can state that the localization pattern of the MST industry differs from that of the total firm population and that MST-firms are more concentrated in space. The intersections of the kernel density estimations confirm these findings: For the D_i values from 0.7 to 3.2 the MST industry (solid line) reaches higher density values than the random firm population (dashed line). Hence, we have many firms that are located unusually near to other firms but by the same token, the MST industry has also comparatively more firms that are dispersed. These are however just a few firms as the conjoint index of concentration and dispersion Θ reaches a value of 0.224. Though the density curves of *K(d)* and D_i values are clearly different, 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 8).

As mentioned in section 4, our method also allows for identifying the localization of highly clustered firms. Here, we identify the cluster-cores of the German MST industry by selecting those firms that exceed the 95-th percentile of the benchmark distribution of D_i values (290 firms). Figure 10 (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 suggestion of the D&O-index that there are several MST clusters located at a larger distance to each other.



(a) MST Clusters (N=290)

(b) Berlin cluster

(c) Rhine-Main cluster

Figure 10: Localization of German MST clusters (distance function: hyperbola function)

Taking a closer look at the clusters reveals an interesting difference in their geographical scope. The Berlin MST cluster-core consists of eight firms within a distance of approximately 1 km, while the Rhine-Main cluster-core contains 15 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 firm-level cluster index

After having presented the mathematical background and the empirical results of our new firm-level cluster index, we will now discuss its advantages and disadvantages compared to the existing 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. With respect to the five criteria of Duranton and Overman, we can state that our index fulfills all requirements.

(2) Inference to localization: Our method is able to deliver insights into the spatial localization of a firm and its degree of spatial clustering. By selecting firms that exceed the 95-th percentile of the benchmark distribution of Di values, we can identify the localization of statistically validated cluster-cores. To our knowledge, this feature has not yet been introduced to MAUP-free methods as the other indices focus on distance intervals but cannot state where firms that show close neighborhoods are located.

(3) Low risk of secondary-MAUP: In contrast to the existing indices, our cluster index 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 when a hyperbola function is used.

(4) Low Computational requirements: This central feature derives from the two last points: As the research area is not divided into intervals, computations have 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 metric 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 even with a standard personal computer. In terms of computational complexity theory our metric is bounded by

$$O\left(\frac{n*(n-1)}{2}\right) \text{ and } O\,(m^2), \text{ respectively,} \tag{13}$$

where *n* is the size of the industry under investigation and *m* is the number of benchmark firms.

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 and the existing MAUP-free methods. 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. However, finally the choice between the methods should depend on the research question that is to be addressed.

6 Conclusions

In this paper, we have introduced a firm-level cluster index as a new MAUP-free statistic method that fulfills the five criteria of Duranton and Overman, 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 in the context of spatial concentration 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.

The ability to give insight into the spatial localization and dimension of firm clusters makes our index also applicable to other investigations and scientific disciplines. To our mind, an implementation could be the detection of focuses of infection in epidemiology.

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 that analyze firm characteristics, such as firm growth. 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

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.

Kosfeld, Reinhold; Eckey, Hans-Friedrich Lauridsen, Jørgen (2011): Spatial point pattern analysis and industry concentration. In: The Annals of Regional Science (47): 311–328.

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.

Miller, Harvey J. (2010): The data avalanche is here. Shouldn't we be digging? In: Journal of Regional Science 50 (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.

De Vries, Jacob; Nijkamp, Peter; Rietveld, Piet (2009): Exponential or power distance-decay for commuting? An alternative specification. In: Environment and Planning A 41: 461-480.

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

		2000 | 4000	2000 | 6000	2000 | 8000	4000 | 6000	4000 | 8000	6000 | 8000
Hyperbola	Mann-U-Test	0.564	0.314	0.446	0.615	0.866	0.680
	KS-Test	0.650	0.438	0.869	0.771	0.728	0.742
	Av.Change (%)	3.61	3.45	3.34	1.36	1.29	0.73
Negative exponential	Mann-U-Test	0.047	0.005	0.005	0.446	0.446	1
	KS-Test	0.000	0.000	0.000	0.000	0.000	1
	Av.Change (%)	20.39	18.88	18.37	7.32	6.70	4.01

Table 1: p-values for Mann-Whitney-Test and KS-Test and average change of D_i values for 2000, 4000, 6000 and 8,000 CSR-point sets

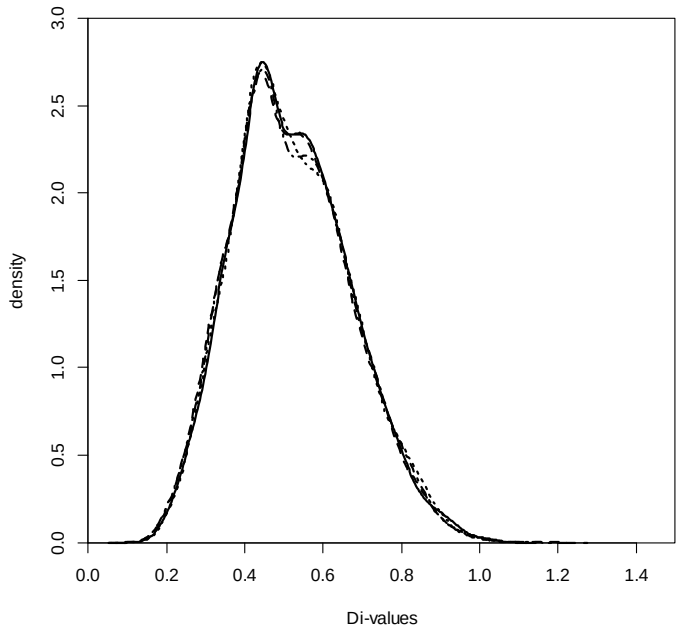


Figure 11: KDE for four independent drawings out of the MARKUS-database (4000 firms). Distance function: hyperbola function

Drawings	1 | 2	1 | 3	1 | 4	2 | 3	2 | 4	3 | 4
Mann-U-Test	0.1089	0.9617	0.3948	0.1204	0.4562	0.4161
KS-Test	0.1641	0.925	0.4005	0.4163	0.7944	0.6099

Table 2: p-values for Mann-Whitney-Test and KS-Test for four independent drawings of the MARKUS-database (4000 firms). Distance function: hyperbola function

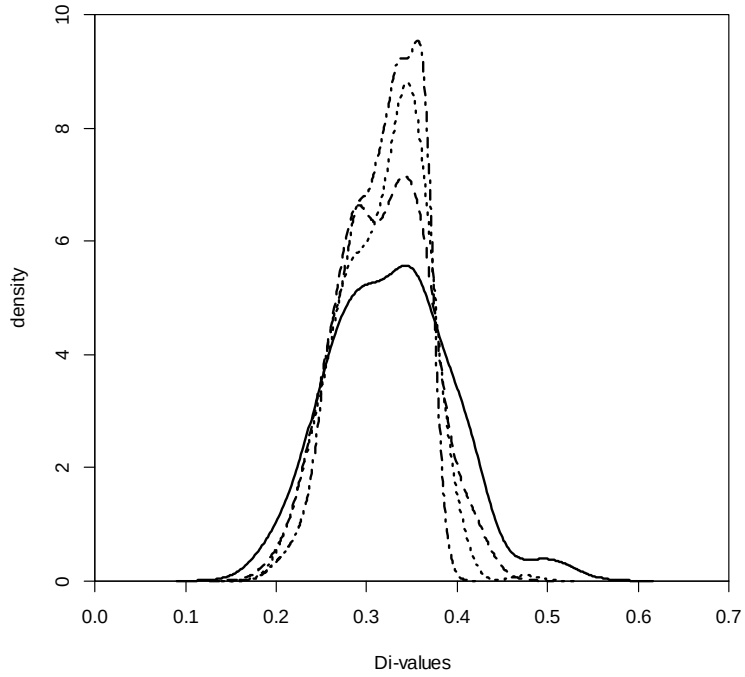
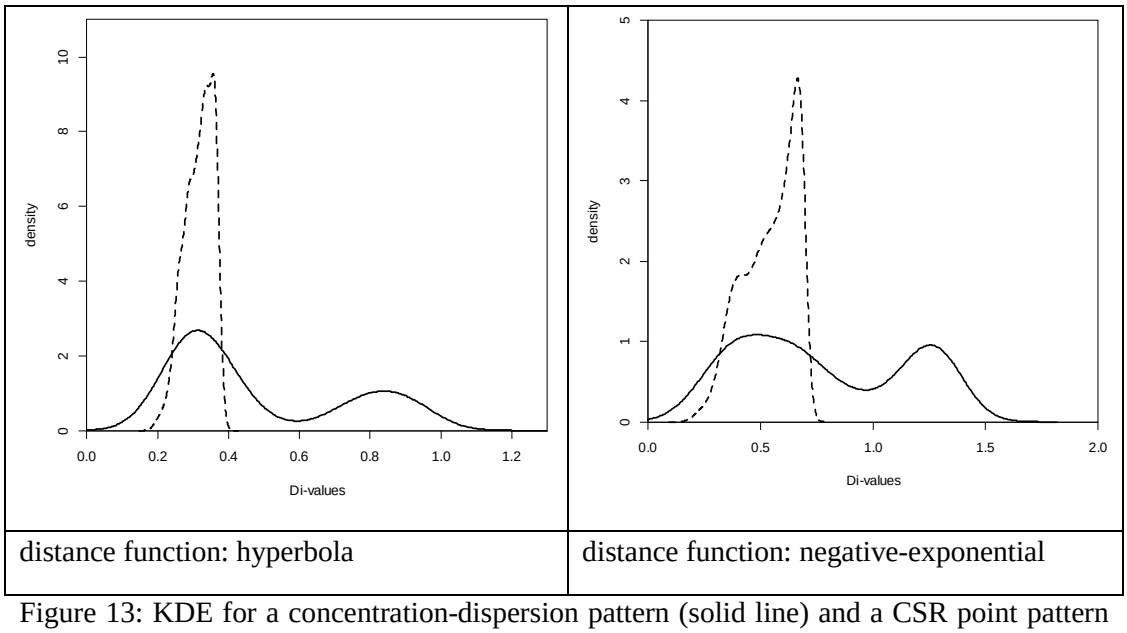


Figure 12: KDEs for 200 (solid line), 400 (dashed line), 600 (dotted line) and 800 (dotted-dashed line) CSR-points (hyperbola-function)



distance function: hyperbola	distance function: negative-exponential

Figure 13: KDE for a concentration-dispersion pattern (solid line) and a CSR point pattern (dashed line)

	IS_MST	N	Mean-rank	Rank-sum
Di	BENCHMARK	4000	2252,23	9008914.00
	MST	861	3261.53	2808177.00
	Total	4861		

	Di
Mann-Whitney-U	1006914.000
Wilcoxon-W	9008914.000
Z	-19.143
Asymptotic significance (2-sided)	.000

Table 3: Mann-Whitney-Test

Most Extreme Differences	Absolute	.317
	Positive	.317
	Negative	.000
Kolmogorov-Smirnov-Z		8,449
Asymptotic significance (2-sided)		.000

Table 4: Kolmogorov-Smirnov-Test

	N	Minimum	Maximum	Mean	Median	Standard deviation	Variance
BENCHMARK	4000	.17427	1.100	.519251	.504962	.14931	.022
MST	861	0.2036	2.752	.825544	.654013	.55406	.307

Table 5: Descriptive statistics