

No. 16-2009

**Reinhold Kosfeld, Hans-Friedrich Eckey, and
Jørgen Lauridsen**

Spatial Point Pattern Analysis and Industry Concentration

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/index_html%28magks%29

Coordination: Bernd Hayo • Philipps-University Marburg
Faculty of Business Administration and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Spatial Point Pattern Analysis and Industry Concentration

Reinhold Kosfeld · Hans-Friedrich Eckey · Jørgen Lauridsen

Abstract. Traditional measures of spatial industry concentration are restricted to given areal units. They do not make allowance for the fact that concentration may be differently pronounced at various geographical levels. Methods of spatial point pattern analysis allow to measure industry concentration at a continuum of spatial scales. While common distance-based methods are well applicable for sub-national study areas, they become inefficient in measuring concentration at various levels within industrial countries. This particularly applies in testing for conditional concentration where overall manufacturing is used as a reference population. Using Ripley's K function approach to second-order analysis, we propose a subsample similarity test as a feasible testing approach for establishing conditional clustering or dispersion at different spatial scales. For measuring the extent of clustering and dispersion, we introduce a concentration index of the style of Besag's (1977) L function. By contrast to Besag's L function, the new index can be employed to measure deviations of observed from general spatial point patterns. The K function approach is illustratively applied to measuring and testing industry concentration in Germany.

Keywords: Spatial concentration, clustering, dispersion, spatial point pattern analysis, K function

JEL : C46, L60, L70, R12

1. Introduction

The spatial distribution of economic activity is an important issue in regional economic theory and policy. Rationales for benefits from agglomerations already date back to Marshall (1920). Arrow and Romer seized on Marshall's reasoning on technological externalities by pointing to localisation economies due to agglomeration of firms in the same branch of industry (Neffke et al., 2008). Marshall-Arrow-Romer (MAR) externalities are founded in benefits from a specialised labour pool, scale economies of input suppliers and knowledge spillovers within industries. Jacobs (1970, 1986) worked out that externalities may additionally arise when firms of different industries agglomerate (Neffke et al., 2008). Jacobs' externalities or urbanisation economies may result from a large and varied labour market pool, scale economies in infrastructure provision, a variety of business services and knowledge spillovers across industries.

Reinhold Kosfeld
Institute of Economics, University of Kassel, 34109 Kassel, Germany
e-mail: rkosfeld@wirtschaft.uni-kassel.de

Hans-Friedrich Eckey
Institute of Economics, University of Kassel, 34109 Kassel, Germany

Jørgen Lauridsen
Institute of Public Health, University of Southern Denmark, Odense M 5230, Denmark

New Economic Geography (NEG) explains the emergence of a core-periphery structure by the tense of centripetal and centrifugal forces (Krugman, 1991; Helpman, 1998; Fujita et al., 1999). Savings of transport costs favour the development of agglomerations as both intermediary and final goods become cheaper for firms and consumers. With decreasing prices, real wages will increase thereby attracting additional consumers. Because of larger sales markets, firms benefit from moving from periphery towards economic centres. Dispersive forces can be picked up in the demand of immobile workers living in peripheral regions as well as congestion. While standard NEG models treat the whole industry uniformly, recent research focuses on identifying sector-specific clusters and their impact on regional growth and development (Feser et al., 2008).

A cluster is a geographically concentrated group of companies and associated institutions sharing local resources, using associated technologies, forming linkages and alliances, as well as co-operating in complementary relationships (Porter, 2008). Porter points to the role of clusters in regional competition and explains how clusters can positively affect competition by increasing productivity and innovation as well as stimulating the formation of new businesses. Thus, the EU commission, national and regional governments have designed and implemented different types of instruments of cluster policy (Oxford Research, 2008). The European Commission has launched new initiatives to encourage national and regional governments to develop regional clusters. However, although there is a consensus that economic sectors benefit differently from spatial clusters, evidence on the efficacy of clusters on development and growth of regions is not unambiguous (Litzenberger, 2007; Menzel, 2008). Depending on the factors viewed as particularly relevant for the formation and development of clusters, localisation or urbanisation economies attain a greater weight (Beaudry and Schiffauerova, 2009).

For policy makers in Europe, clusters can be relevant on a number of levels. In order to get an insight in economic effects from clusters, information is necessary on the formation of potential cluster at different regional scales. Moreover, the degree of concentration is expected not to be independent on the reach of grouping of firms. Traditional concentration indices fail to provide such information.

The spatial Gini coefficient and Herfindahl index are elementary instruments for measuring spatial concentration of economic activity (Feser, 2000; Bickenberger and Bode, 2008). Although the spatial Gini coefficient is preferable from the viewpoint of data requirements, (Südekum, 2006), Ellison and Glaeser (1997) have revealed some distortive effects coming along with this measure. They derive an index of concentration on the basis of a probabilistic model of plant location decisions. The Ellison-Glaeser index “corrects” the Gini coefficient by eliminating distortions from industry structure with the aid of the Herfindahl index. When the null hypothesis of a perfectly random location process is rejected, spatial concentration is driven by spillover forces, natural advantages or a mixture of both factors.

Location choice by firms may not only lead to a clustering patterns on the industry or sector level. By conditioning on the industry as a whole, dispersed patterns of plants within industries may arise. In this case, plants belonging to different sectors tend to co-locate with a higher probability than plants from the same sector. While dispersive plant patterns cannot be revealed by the Gini coefficient and the Herfindahl index, they can be detected on the basis of the Ellison-Glaeser index. However, clustering and dispersion patterns may vary across spatial scales. For example, clustering may occur at small spatial scales, whereas complete random or dispersed patterns may be prevalent at larger distances.

Spatial point pattern analysis is an appropriate approach to deal with these issues. It provides a toolbox of evaluating industry concentration by analysing the spatial distribution of plants in a study region. In particular we make use of Ripley's K function (Ripley, 1976, 1977) that allows measuring of clustering and dispersion simultaneously at all relevant spatial scales. Although the K function is a powerful analytic tool in measuring the covariance structure of the location process of plant decisions, its application on real economies is extremely time-consuming.¹ This is one of the reasons of the restrained use of this approach in assessing industrial concentration. While Barff (1987) applies the K function approach to a single city, Sweeney and Feser (1998) extend it to the state level. Both papers differentiate between different size classes of establishments, not between sectors.

Marcon and Puech (2003) utilise methods of point pattern analysis to evaluate sector concentration of firms in the greater Paris area and an idealised area of France. Like Sweeney and Feser (1998), they make use of Diggle and Chetwynd's (1991) D function to establish clustering or dispersion of a branch relative to the industry as a whole. Duranton and Overman (2005) test for localisation of British branches of industry on the basis of a K density function. In referring to an earlier draft of the paper, Marcon and Puech (2003) compare the Duranton and Overman's K density with Ripley's K function approach. In spite of some advantages, the construction of the K density function is not without problems. In particular clustering and dispersion can only be detected but not quantified. Arbia et al. (2008) use the bivariate K function approach to identify co-location across different industries.²

The contribution of this paper to the literature is threefold. First, we introduce a concentration index of the style of Besag's (1977) L function that is based on the concept of the K function. While Besag's L function is intended to measure deviations from the CSR process, the new index can be applied to measure deviations from more general spatial processes. The index is also used for identifying the importance of sector-specific and more general industry-specific forces inducing clustering. Secondly, we propose a feasible testing procedure enabling a usage of the K function approach efficiently for large study regions. For this, Diggle and Chetwynd's (1991) D function approach is replaced by a spatial similarity test based on subsamples drawn from the industry under analysis and the reference population. Third, up to now, concentration of the branches of industry in Germany is only available at spatial scales given by more or less arbitrary defined regions. While we illustrate our K function approach by selected industries, we additionally provide concentration numbers for sixteen German industries within different distance bands.

The paper is organised as follows. In section 2, we introduce the methods of spatial point pattern analysis for evaluating industry clustering and dispersion. Section 3 deals with issues regarding the construction of geographical coordinates from the available source of data. Estimation and testing results on unconditional concentration of branches are discussed in section 4, while section 5 illustrates the application of our K function approach in assessing conditional clustering and dispersion. The empirical analysis is in both cases performed for mining and manufacturing industries in Germany. Section 5 draws conclusions and points to directions of future research in this field.

¹ 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. A single simulation run takes several hours.

² Arbia et al. (2008) do not analyse the spatial point pattern of plants as a realisation of firms' location decisions, but show how the K function approach can be applied to investigate co-agglomeration or repulsion of economic events. In particular they analyse spatial nearness and remoteness between locations of inventions across sectors of industry in Italy.

2. Spatial Point Processes

Testing for unconditional concentration and dispersion

The spatial approach in measuring industry concentration investigates the point pattern of industrial establishments. Firms' decisions on where to locate industrial production is viewed as a spatial point process $\{N(A), A \subseteq R\}$ where the random variable $N(A)$ renders the number of plants in the area A as part of the whole study area R . For a stationary process, the intensity λ , defined by the number of plants per unit area, is constant over the whole study area R . Often it is sensible to assume that the spatial point process is not only stationary but isotropic. In this case, the second-order intensity $\gamma(s_i, s_j)$ measures the dependence between two plants at locations s_i and s_j solely as a function of their distance d : $\gamma(s_i, s_j) = \gamma(d)$.

Unfortunately, the second-order intensity $\gamma(d)$ is of little practical use as it cannot directly be estimated from sample data. Explorative tools for analysing the spatial point patterns are the cumulative distribution functions of nearest neighbour event-event distances and nearest neighbour point-event distances (see e.g. Bailey and Gatrell, 1995; Martinez and Martinez, 2008). However, these functions only give insight into pattern characteristics over the smallest scales. Ripley's K function (Ripley, 1976, 1977) is a much more powerful tool in investigating second-order properties with real data. Grounded on the close connection between second order properties and the distances between pairs of occurrences of plants, this tool provides insight on clustering and dispersion of a point pattern at a range of relevant scales. It can be meaningfully interpreted in relation to the K function of a benchmark like the complete spatially random (CSR) process. Moreover, we test for significance from hypothesised patterns using the bounds of confidence intervals derived from Monte Carlo simulation. The power of the K function rests most notably in its use as a graphical tool to provide detailed insight into the extent of concentration of industrial sectors in dependence of the regional scale.

As a reduced second-order moment measure, the K function $K(d)$ is closely related to the second-order intensity $\gamma(d)$. It measures the normalised expected number of occurrences of plants within a distance d of an arbitrary establishment. The normalisation is accomplished by dividing the expected value $E[N(A_d)]$ by the intensity λ where A_d is the area of a circle with the radius d around an arbitrary plant located at a point s_i in R .³

$$(1) \quad K(d) = \frac{1}{\lambda} \cdot E[N(A_d)].$$

Without a relation to the intensity λ the expectation of $N(A_d)$ cannot be meaningfully compared across different populations. In defining the K function, the "density effect" is eliminated from the absolute measure of occurrences of additional plants in a well-defined neighbourhood of an arbitrarily chosen establishment.

Although the expected value $E[N(R)]$ for the whole study region is always given by $\lambda \cdot R$, the expected number of plants in A_d is not generally equal to $\lambda \cdot A_d$. The relation $E[N(A_d)] = \lambda \cdot A_d$ only holds for a completely spatially random (CSR) point process. While the expectation of $N(A_d)$ is larger than $\lambda \cdot A_d$ in case of concentration at scale d , it is lower than $\lambda \cdot A_d$ in case of dispersion. The latter case reflects a regular distribution of plants at the considered scale. Both tendencies are mirrored in the K function.

³ For the sake of simplicity, we use the same symbols for the labels of the areas (R, A, A_d) as for the areas themselves.

In measuring unconditional or absolute spatial concentration, the intensity λ is assumed to be constant across the study region R . Let $I_d(d_{ij})$ be an indicator function that takes the value of 1 if the distance between two plants i and j is lower or equal to d and 0 otherwise. Then a preliminary non-parametric estimate of the K function is given by

$$(2) \quad \hat{K}(d) = \frac{1}{\hat{\lambda}^2 \cdot R} \sum_{i=1}^n \sum_{i < j} I_d(d_{ij}).$$

In (2), the expected number of ordered pairs of plants at most d units away from one another, $\lambda^2 \cdot R \cdot K(d)$,⁴ is estimated by the double sum of the indicator function $I_d(d_{ij})$. In order to compute the estimate of $K(d)$ for a series of distances d , it is suggested to use the ratio n/R as an estimator $\hat{\lambda}$ for the intensity λ with n as the number of observed plants in R .

As the expected number of plants tends to be underrated by (2) due to border effects, the estimator $\hat{K}(d)$ is generally not unbiased. This problem especially becomes serious at large scales. The border effects result from the ignorance of possible occurrences of plants outside the study region R when counting these entities within concentric circles around the locations of critical plants. An edge correction can be accomplished by introducing correction factors $1/w_{ij}$ where the weights w_{ij} is chosen as the proportion of the circumferences of the circles lying in region R (Martinez and Martinez, 2008).⁵ With this adjustment, a feasible edge-corrected estimate for the K function is given by

$$(3) \quad \hat{K}(d) = \frac{R}{n^2} \sum_{i=1}^n \sum_{i < j} \frac{I_{ij}(d)}{w_{ij}}.$$

In order to interpret the values of the K function of an observed point pattern, one has to look for a benchmark. In measuring absolute spatial concentration, the K function of a complete spatially random (CSR) process serves a natural benchmark. For this process, $K(d)$ is simply given by the area $\pi \cdot d^2$. Hence, $K(d)$ generally measures a hypothetical area A_d for the spatial point process under investigation. In the case of concentration, $K(d) > \pi \cdot d^2$ measures the area that is expected under the CSR hypothesis given the increased number of plants. Conversely, in the case of dispersion, $K(d) < \pi \cdot d^2$ reflects the area that is expected in view of the decreased number of plants for a CSR process.

In order to test for dispersion and clustering on the basis of the K function, lower and upper sets of critical values within a range of relevant distances, $\hat{K}_{\alpha/2}^l(d)$ and $\hat{K}_{\alpha/2}^u(d)$, are needed. Thus, $\hat{K}_{\alpha/2}^l(d)$ and $\hat{K}_{\alpha/2}^u(d)$ define the bounds of a confidence interval suitable with a significance level of α . As the distribution of $\hat{K}(d)$ is unknown, the bounds have to be determined by Monte Carlo methods. For each industry we simulate B random patterns of size n . In case of $\hat{K}_{\alpha/2}^l(d) \leq \hat{K}(d) \leq \hat{K}_{\alpha/2}^u(d)$, the CSR hypothesis cannot be rejected for the distance d at a significance level of α . If $\hat{K}(d)$ is outside the confidence interval, the following testing decisions result:

⁴ This expectation is obtained by multiplying the expected number of plants in R , $\lambda \cdot R$, by the left-hand side of $\lambda \cdot K(d) = E[N(A_d)]$ which is implied by equation (1).

⁵ Alternatively, the area of the circle can be used for calculating the correction factor (cf. Marcon and Puech, 2003).

$$\hat{K}(d) < \hat{K}_{\alpha/2}^1(d) \Rightarrow \text{Significant dispersion at scale } d$$

and

$$\hat{K}_{\alpha/2}^u(d) > \hat{K}(d) \Rightarrow \text{Significant clustering at scale } d.$$

Given the computational expense of the testing procedure, we restrict ourselves to determine the confidence band for a special case. Let $\hat{K}_{\text{CSR}}^b(d)$ denote the estimated K function of the simulated CSR process in the b th run. Then the lower and upper envelopes, $L_{\text{CSR}}^B(d)$ and $U_{\text{CSR}}^B(d)$, of the estimates $\hat{K}_{\text{CSR}}^b(d)$ are defined by

$$L_{\text{CSR}}^B(d) = \min\{\hat{K}_{\text{CSR}}^b(d), b = 1, 2, \dots, B\}$$

and

$$U_{\text{CSR}}^B(d) = \max\{\hat{K}_{\text{CSR}}^b(d), b = 1, 2, \dots, B\}.$$

For $B=20$, the bounds $\hat{K}_{\alpha/2}^1(d)$ and $\hat{K}_{\alpha/2}^u(d)$ of the confidence interval coincide with the lower and upper envelopes, $L_{\text{CSR}}^B(d)$ and $U_{\text{CSR}}^B(d)$, of the estimated K function $\hat{K}_{\text{CSR}}^b(d)$ for a significance level α of 0.05⁶

Some authors advice to use one half of the maximum distance between the pairs of events as an upper bound for d (e.g. Smith, 2008; Marcon and Puech, 2003). However, at large scales, edge effects increasingly dominate the estimator for the K function. In particular for irregular shaped study areas, with this rule of thumb, serious interpretation problems may arise. Therefore we restrict the maximum radius by one fourth of the maximum pairwise distances between locations of plants (cf. Duranton and Overman, 2005; Arbia et al., 2008).

Testing for conditional concentration and dispersion

In measuring concentration of manufacturing sectors relative to the industry as a whole, location decisions of firms are taken as given. On this account the null hypothesis of complete spatial randomness of plant locations is no longer effective. More specifically we replace the CSR hypothesis by the hypothesis of spatial similarity as a benchmark. The spatial point patterns of an industrial sector, $S_1 = (s_{11}, s_{12}, \dots, s_{1n_1})$, and all other manufacturing sectors $S_2 = (s_{21}, s_{22}, \dots, s_{2n_2})$, are called spatially similar, when they are generated by the same spatial point process. Let $m_i \in \{1, 2\}$ be a label denoting whether a location is a manufacturing sector (1) or all other industrial sectors (2). Under the null hypothesis the labels m_i can be exchanged such that S_1 and S_2 consist of both types of plants. In all, there exist $n!$ permutations of labels that are all equiprobable under the spatial similarity hypothesis. One speaks of a marked point process that assigns the n labels m_i randomly to the observed n industry locations s_i . By conditioning on observed set of locations, a wide variety of point patterns can be compared without the need to identify alternative locations.

In principal, conditional concentration or dispersion of manufacturing sectors could be measured by investigating the difference of the empirical K functions of both patterns S_1 and

⁶ The value of B is in line with Marcon and Puech's (2003) choice of the number of simulations for the idealised area of France.

S_2 that defines the so-called D function (Diggle and Chetwynd, 1991). Critical values for the test of the spatial similarity hypothesis can be derived by Monte Carlo simulations. This type of test for relative concentration is proposed by Marcon and Puech (2003). Because the number of locations in S_2 is in general many times over that of S_1 , the usual test of the spatial similarity hypothesis is not very efficient. In cases of economies of a special size, it is not feasible without imposing restrictions.⁷

A more efficient and feasible test on conditional concentration can be based on a subsample similarity hypothesis. First, we use the observed point pattern,

$$S_1^0 = (s_{11}, s_{12}, \dots, s_{1n_1}),$$

to construct an estimate $\hat{K}_1(d)$ of the K function for the industry under analysis. Then, we simulate B random permutations $(p_1(b), p_2(b), \dots, p_n(b))$, from the order of natural numbers $N_n = (1, 2, \dots, n)$ and use the first n_1 numbers $(p_1(b), p_2(b), \dots, p_{n_1}(b))$ to define a sample of locations from all industrial sites (IND):

$$S_{\text{IND}}^b = (s_{p_1(b)}, s_{p_2(b)}, \dots, s_{p_{n_1}(b)}), b=1,2,\dots,B.$$

Under the null of spatial indistinguishability, both S_1^0 and S_{IND}^b are subsamples from the same spatial point process. Therefore the estimated K functions from S_{IND}^b , $\hat{K}_{\text{IND}}^b(d)$, can be used to construct a confidence interval for $\hat{K}_1(d)$. The lower and upper bounds of this confidence interval, $\hat{K}_{1,\alpha/2}^l(d)$ and $\hat{K}_{1,\alpha/2}^u(d)$, provide critical values for the test of the subsample similarity hypothesis. The following testing decisions are obtained with respect to conditional dispersion and clustering:

$$\hat{K}_1(d) < \hat{K}_{1,\alpha/2}^l(d) \Rightarrow \text{Significant conditional dispersion at scale } d$$

and

$$\hat{K}_{1,\alpha/2}^u(d) > \hat{K}_1(d) \Rightarrow \text{Significant conditional clustering at scale } d.$$

In case of $\hat{K}_{1,\alpha/2}^l(d) \leq \hat{K}_1(d) \leq \hat{K}_{1,\alpha/2}^u(d)$, the subsample similarity hypothesis cannot be rejected for the distance d at a significance level of α .

As in testing for unconditional concentration we determine the confidence band on the basis of the envelopes of the simulated K functions:

$$L_{\text{IND}}^B(d) = \min\{\hat{K}_{\text{IND}}^b(d), b = 1,2,\dots,B\}$$

and

$$U_{\text{IND}}^B(d) = \max\{\hat{K}_{\text{IND}}^b(d), b = 1,2,\dots,B\}.$$

⁷ Marcon and Puech (2003) note that it was impossible to perform K function analysis for the whole area of France. Instead they apply Diggle and Chetwynd's D function to an idealised French rectangular area.

For a significance level α of 0.05 the bounds $\hat{K}_{1,\alpha/2}^l(d)$ and $\hat{K}_{1,\alpha/2}^u(d)$ are given by the lower and upper envelopes $L_{IND}^B(d)$ and $U_{IND}^B(d)$ with $B=20$.

Although the testing procedure draws subsamples of size n_1 from the reference population for calculating confidence bands, it becomes infeasible for large samples sizes n_1 . In this case, subsampling is applied as well to the industry under consideration. In order to ensure feasibility and efficiency, we restrict the subsample size n_1 to 500.

Indices of clustering and dispersion

Concentration indices for single industries can be constructed from K function analysis in form of difference measures. In principal, a spatial concentration index could be defined by relating the K function to its expectation under CSR. In order to avoid comparisons of areas, Besag (1977) proposed an L function as access radii of the circles around the locations necessary to capture the observed number of events under the assumption that the point pattern had been generated from a CSR process:

$$(4) \quad L(d) = \sqrt{\frac{K(d)}{\pi}} - d$$

As the dividing value is zero, from $L(d) > d$ clustering and from $L(d) < 0$ dispersion is inferred at distance d .

Despite its vivid interpretation, Besag's L function fails to detect insignificant deviations from the CSR process. It indicates clustering or dispersion even if the observed point pattern is a realisation from a CSR process. Moreover, $L(d)$ is not conceived to measure conditional concentration.

In order to establish the extent of significant clustering (dispersion), we define a concentration index $L^*(d)$ that measures the excess radii with respect to the upper (lower) confidence band. Let $\hat{K}(d)$ be the observed K function of the industry under analysis and $\hat{K}_{\alpha/2}^u(d)$ ($\hat{K}_{\alpha/2}^l(d)$) the upper (lower) confidence band. With this, the concentration index $L^*(d)$ is built up as follows.

$$(5) \quad L^*(d) = \begin{cases} \sqrt{\frac{\hat{K}(d)}{\pi}} - \sqrt{\frac{\hat{K}_{\alpha/2}^u}{\pi}} & \text{for } \hat{K}(d) > \hat{K}_{\alpha/2}^u \text{ (sign. clustering)} \\ \sqrt{\frac{\hat{K}(d)}{\pi}} - \sqrt{\frac{\hat{K}_{\alpha/2}^l}{\pi}} & \text{for } \hat{K}(d) < \hat{K}_{\alpha/2}^l \text{ (sign. dispersion).} \\ 0 & \text{for } \hat{K}_{\alpha/2}^l \leq \hat{K}(d) \leq \hat{K}_{\alpha/2}^u \text{ (acceptation of CSR)} \end{cases}$$

While $L^*(d)$ becomes positive in case of significant spatial clustering, significant dispersion is indicated by a negative L^* value at distance d . At spatial scales where the observed K function runs between the lower and upper confidence bands, the L^* function takes the value zero.

With the notation used above, L^* is defined as an index of unconditional concentration. Regular patterns are usually not observed in economics. Moreover, a completely random point pattern will only occur by exception. Thus, in the unconditional case, L^* is suited to measure the extent of spatial concentration of industries and sectors. By replacing the observed K function $\hat{K}(d)$ by the $\hat{K}_1(d)$ function along with the respective confidence bands, the L^* function can be employed for identifying clustered or dispersed industry patterns relative to an arbitrary reference population. Usually the industry as whole is used as a benchmark.

3. Data

In this study we make use of regional and sectoral disaggregated data on German industrial establishments. While the number of employees is the preferred variable with traditional measures of concentration like the Gini index and the Ellison-Glaeser index (see e.g. Südekum, 2006), spatial methods preferably make use of location data on plants. They directly reflect firms' decisions on sites of production. From the viewpoint of spatial statistics, decisions of enterprises where to locate industrial production define a spatial point process, whereby the scale of spatial analysis cannot *a priori* be fixed.

The regional database of the Federal Statistical Office Germany Data includes data on the number of plants in 439 German districts the latest for the year 2006. The industry is defined by the sections Mining and Quarrying (C) and Manufacturing (D) of the German Classification of Economic Activities (WZ 2003). Up to the four-digit sectors, this classification matches in terms of content with the NACE Rev. 1.1 classification⁸ which is based on the International Standard Industrial Classification of all Economic Activities (ISIC Rev. 3.1) of the United Nations. Two out of sixteen two-letter industries pertain to section C, while fourteen are manufacturing industries belonging to section D (Appendix). The industries are subsections of the sections C and D. Some analyses are additionally performed with two-digit sectors that are called divisions in the NACE classification.

In all, the industry comprises 45611 establishments in 2006. Principally these are all plants with 20 and more employees.⁹ With a share of 97.4 per cent, the overwhelming majority of plants belong to the manufacturing sector (section D). The district level is the finest level of regional disaggregation for which data are available. Coordinates of cities and centres of rural districts are available. However, industrial plants can be located in any municipality of a rural district. Thus, in order to capture dispersion of plants in districts, we replace central locations by randomly distributed points within the areal units.

From sampling surveys it is known that plants of the same branch are usually dispersed across districts (IAB, 2008). Thus, location of plants within districts should not be represented by coordinates of central places. In view of this information, a random allocation of plant locations within districts will be best approximate their real distribution. In urban districts on average from each point all plants are covered by circles with a radius of 5 km. Although some fuzziness is introduced by larger-sized rural regions, we use this threshold as the lower bound of the spatial scale in the analysis of industry concentration. However, as plant density is much sparser in rural regions, the general tendency can be modified but not completely

⁸ Nomenclature des Activités Economiques dans les Communautés Européennes (NACE).

⁹ In selected sectors, for instance, manufacture of food products and beverages and manufacture of glass and ceramics, establishments with 10 to 19 employees are additionally included (StatBA, 2008).

reversed at low spatial scales between 5 and 15 km in case of opposite arrangements in rural regions. In case of short-distance clustering despite a random distribution of plant location within districts no interpretation problems occur. Some degree of uncertainty at low spatial scales arises in case of acceptance of the null hypothesis of randomness.

4. Unconditional industry concentration in space

Here we investigate spatial industry concentration against the hypothesis of complete spatial randomness (CSR). While testing the assumption of spatial homogeneity is not of particular interest, it enables us to detect the intensity of clustering at different spatial scales within and across industries. Moreover, we are interested in revealing the extent to that industry concentration can be attributed to forces effective at the level of industry under consideration or subordinated sectors.

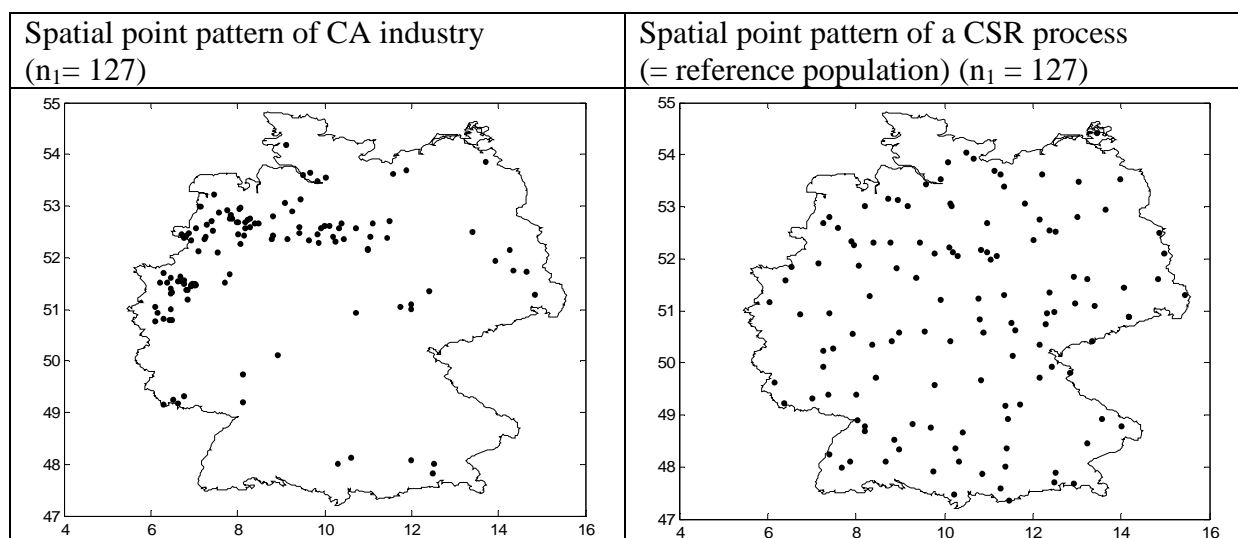
We discuss these issues exemplary for three industries:

- CA: Mining and quarrying of energy producing materials,
- DA: Manufacture of food products, beverages and tobacco,
- DB: Manufacture of textiles and textile products.

In order to determine the relevant spatial scale we apply the $d_{\max}/4$ rule to the industry as a whole. According to this rule, spatial point patterns of industries are analysed for all distances from 5 to 215 km.

For understanding testing for spatial concentration on the basis of the K function, a preceding exploratory data analysis of spatial point patterns may be helpful. In the left panel of Figure 1, the observed spatial point pattern for the CA industry is plotted. It shows a strong clustering of coal mines and quarrying plants in the western part of North-Rhine Westphalia and the western and middle part of Lower Saxony. Using the same number of plants, the observed point pattern is compared with a completely random point pattern.¹⁰ The right panel of Figure 1 exhibits a CSR point pattern showing neither clustering nor regularity. In testing and

Figure 1: Spatial point patterns of the CA industry and a CSR process

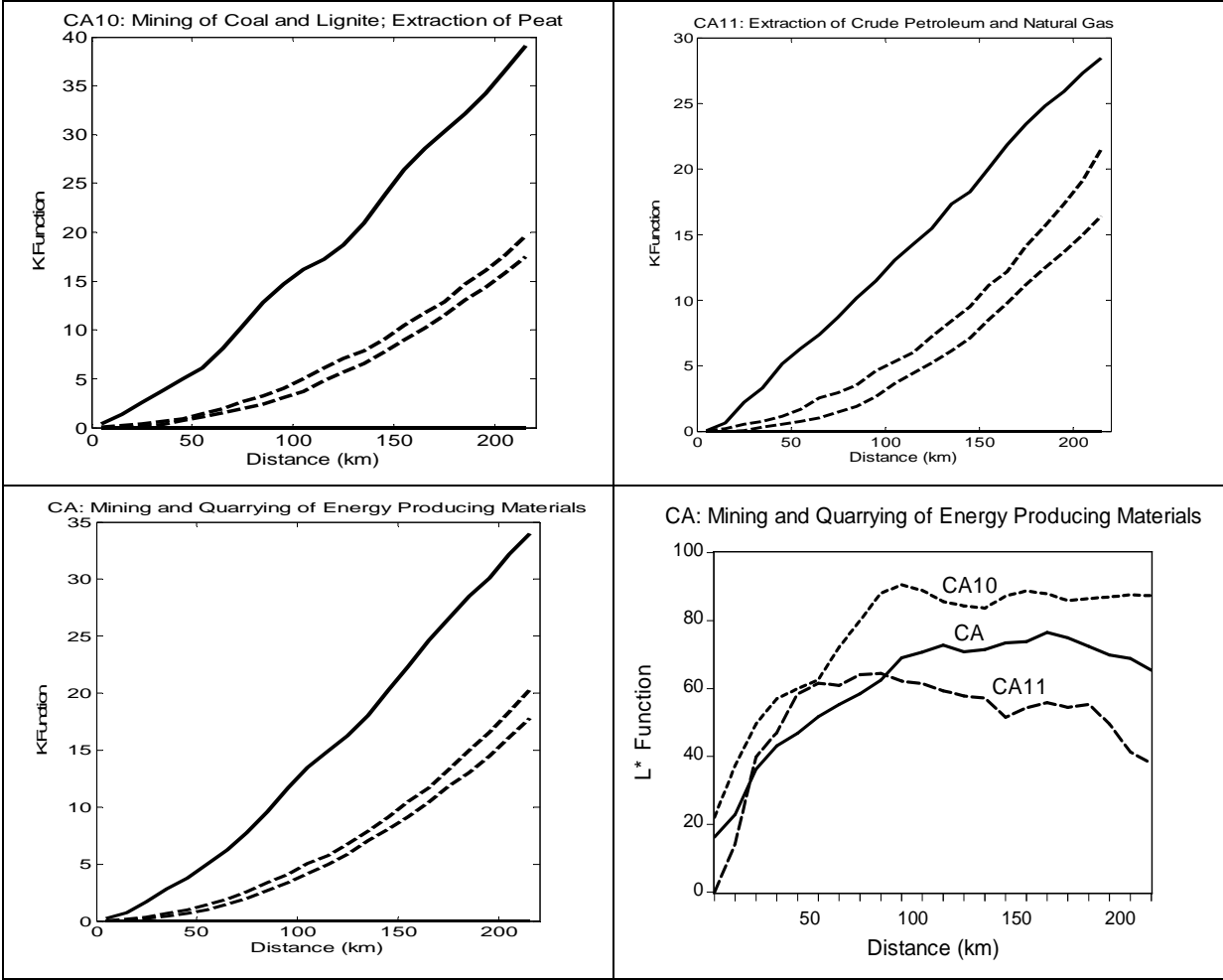


¹⁰ For a fixed intensity rate λ , a CSR process is given by a homogenous Poisson process. When one conditions on the number of plants, the CSR process is usually termed binomial process (cf. Martinez and Martinez, 2008).

measuring spatial concentration using the K function approach, hypothetical spatial point patterns for an industry are simulated from a CSR process by conditioning on the observed number of plants.

The observed K function for the entire CA industry and the subordinated CA10 and CA11 sectors are plotted along with the lower and upper confidence bands in Figure 2.¹¹ As all three K functions lie above the upper confidence bands, significant clustering is established. The intensity of clustering at different spatial scales is measured by the L* functions that reflect the gaps between the observed K functions and the upper confidence bands.

Figure 2: K and L* functions of the CA industry for testing unconditional concentration



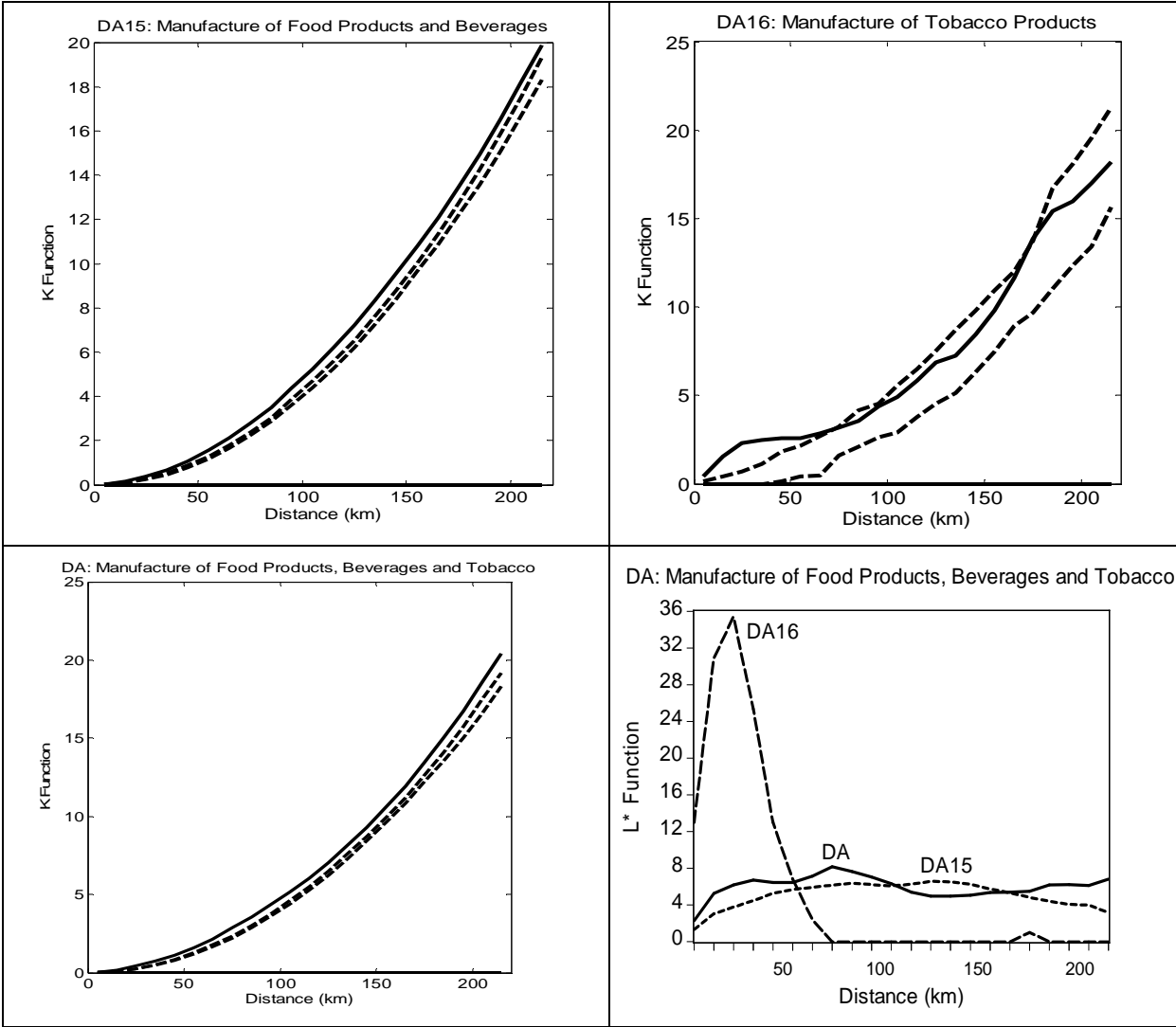
The lower right panel of Figure 2 shows an increasing spatial concentration of the CA industry up to a distance of 160 km. The L* function indicates that the radii of circles around the plants are up to 75 km larger than acceptable under the CSR hypothesis. The CA10 sector is marked by an increasing degree of concentration as far as 80 km after that it remains relatively constant. Coal mines are stronger concentrated than the petroleum and gas factories over the full spatial scale. The run of the L* function of the CA industry below both sector functions in the interval between 20 km and 80 km reflects a lack of mixed industry-specific clustering at lower and medial regional scales. Thus, within this distance band, concentration of the CA industry is attributable to clustering inside both sectors. At larger distances the high

¹¹ The ordinates of the K function diagrams have to be multiplied by a factor 10,000.

degree of concentration of the CA industry is mainly driven by CA10-specific agglomeration forces.

In the DA industry, sector concentration is not uniform (Figure 3). For the entire DA industry as well as the DA15 sector, significant clustering occurs at a level of 5% for all distances. However, plants belonging to the DA16 sector only cluster significantly up to a distance of 75 km. Beyond this threshold, the K function of the tobacco sector lies within the confidence band indicating complete spatial randomness.

Figure 3: K and L* functions of the DA industry for testing unconditional concentration

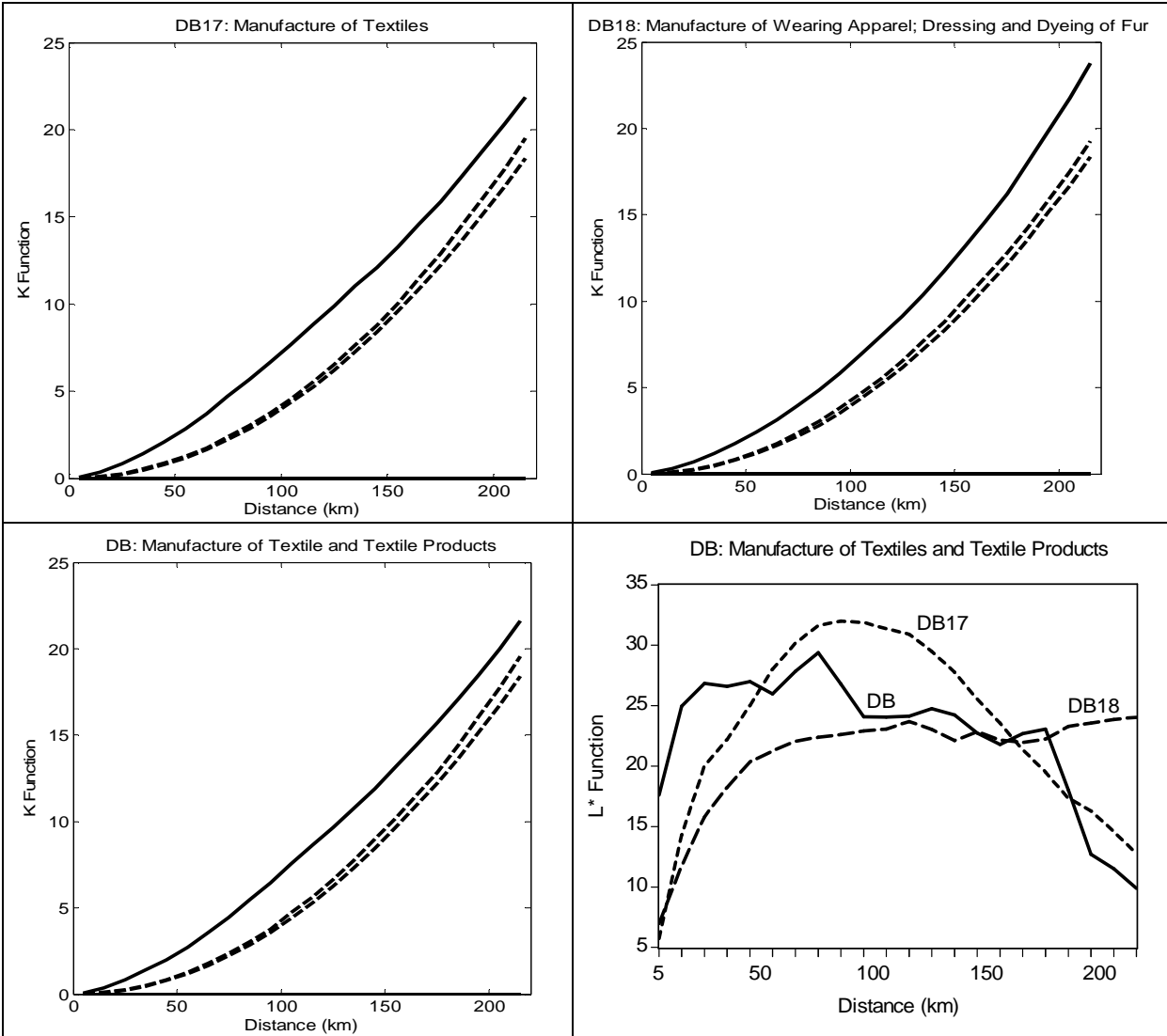


The DA industry as a whole reaches its maximum concentration at 75 km. In the interval of 20 to 100 km its L* function shows a mean excess radius between 6 and 8 km. While the L* functions of the DA industry and the DA16 sector run very similar over a long distance band, they drift apart at a large spatial scale. At distances above 160 km, industry clustering is no more only sector-specific but driven by forces effective across both DA sectors.

Because the K functions are well above the upper confidence bands at all distances in the DB industry, clear clustering structures emerge (Figure 4). At low spatial scales, the run of L* curve of the DB industry above the sector curves indicates the presence of mixed spatial clustering of textile and clothing plants. Industry concentration declines only slightly for

distances between 50 and 175 km, but sharply at larger spatial scales. The highest concentration is measured at a distance of 75 km with an L^* value of nearly 30.

Figure 4: K and L^* functions of the DB industry for testing unconditional concentration



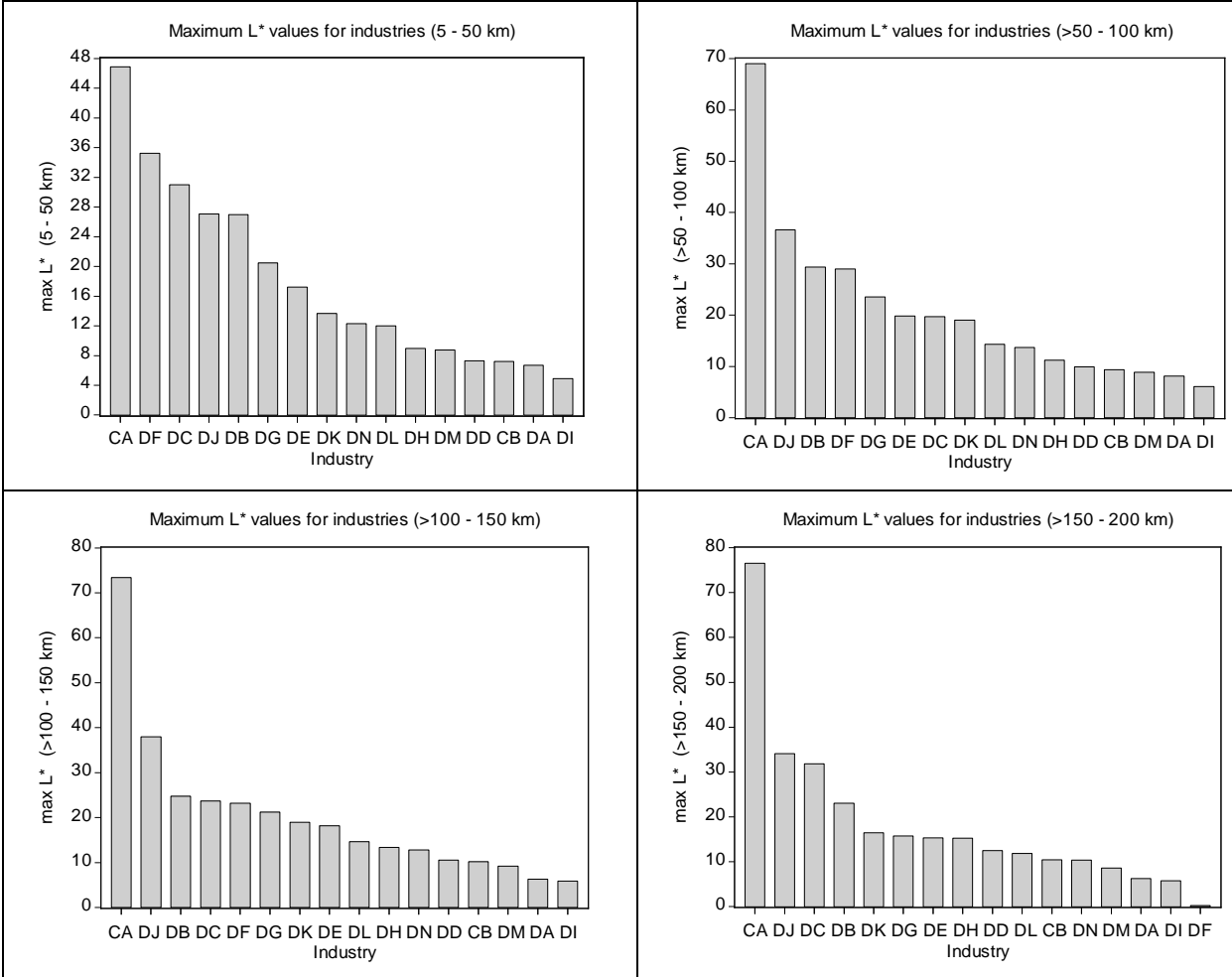
The sector profiles turn out to be very different. After a steep slope up to a distance of 50 km, the DB18 L^* curve remains relatively constant at a high level. This is in contrast to the U-shaped L^* curve of the DB17 sector. For this sector the degree of clustering is highest at distances between 70 and 110 km. Beyond this threshold, the curve declines with a rate of about 0.2 km per unit distance.

The L^* curves are as well comparable across industries. From our examples it can be inferred that the CA industry is considerably stronger concentrated than the DA and DB industries over the whole scale. Out of these industries, the DA industry shows the lowest degree of concentration. In Figure 5 the 16 industries are ranked according to their degree of concentration within four distance bands. For this purpose, we use for each industry the maximum L^* value within a distance band as a concentration index.

Figure 5 confirms the extremely high concentration of coal mines and quarrying plants (CA). At smaller spatial scales, however, the difference to the second-placed industry is by far lesser

than at median and high distances. Manufacture of textiles (DB), leather (DC), coke, refined petroleum and nuclear fuel (DF) and fabricated metal products (DJ) are highly concentrated within different distance bands. But beside the CA industry only the DB industry belongs to the five strongest concentrated branches at all distances. Particularly conspicuous is the last rank of the DF industry in the case of large distances.

Figure 5: Unconditional concentration of industries at different spatial scales

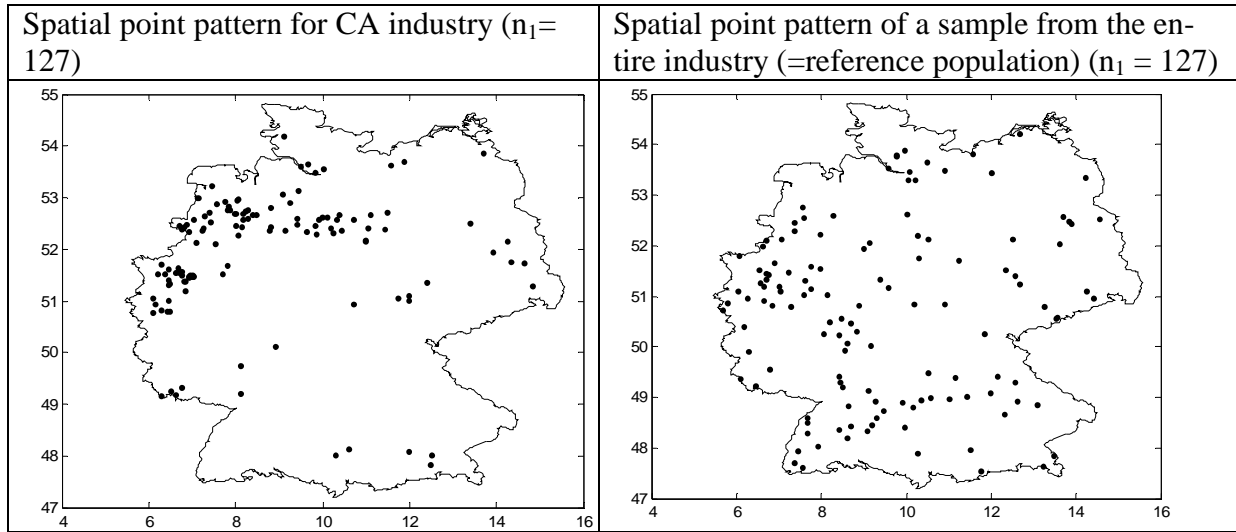


Over the entire spatial scale low concentration is found for mining of metal ores and other mining and quarrying (CB), manufacture of food products (DA), metallic mineral products (DI) and transport equipment (DM). With some qualifications this also applies for the manufacture of wood and wood products (DD). In all, ranking differences at different spatial scales are much greater in the case of high than of low concentration.

5. Conditional industry concentration in space

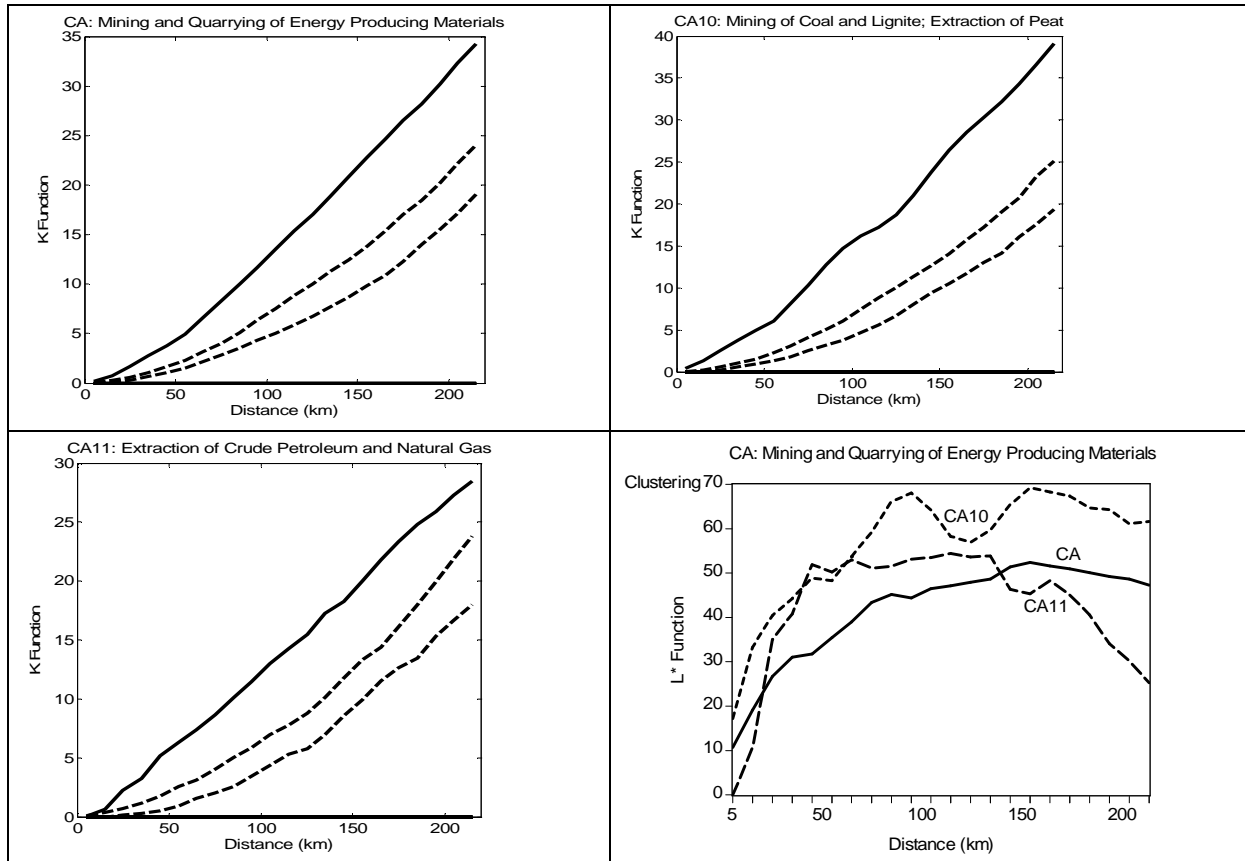
After having established the degree of spatial concentration for 16 industries at the subsection level of NACE, we aim at identifying clustering and regularity by abandoning the assumption of homogenous space. In regional economics, advantages of sites are attributed to natural and economic features. We assume that locational advantages are reflected in firms' decisions on the sites of production. In testing for conditional concentration and dispersion, we refer to plant locations of the whole industry as the reference population.

Figure 6: Spatial point patterns of the CA industry and a sample from the entire industry



As in the case of unconditional concentration, K function analysis is based on comparisons of an industry-specific and hypothetical spatial point patterns. By using the industry as a whole as the benchmark, spatial clustering is present under the null hypothesis. The hypothetical point pattern is specifically generated by randomly labelling all plants of the entire industry. Subsamples of equal size are generated from both samples. Under the null hypothesis of spatial similarity, the point pattern of the industry-specific subsample is just a realisation from the set of all industrial establishments.

Figure 7: K and L* functions of the CA industry for testing conditional concentration

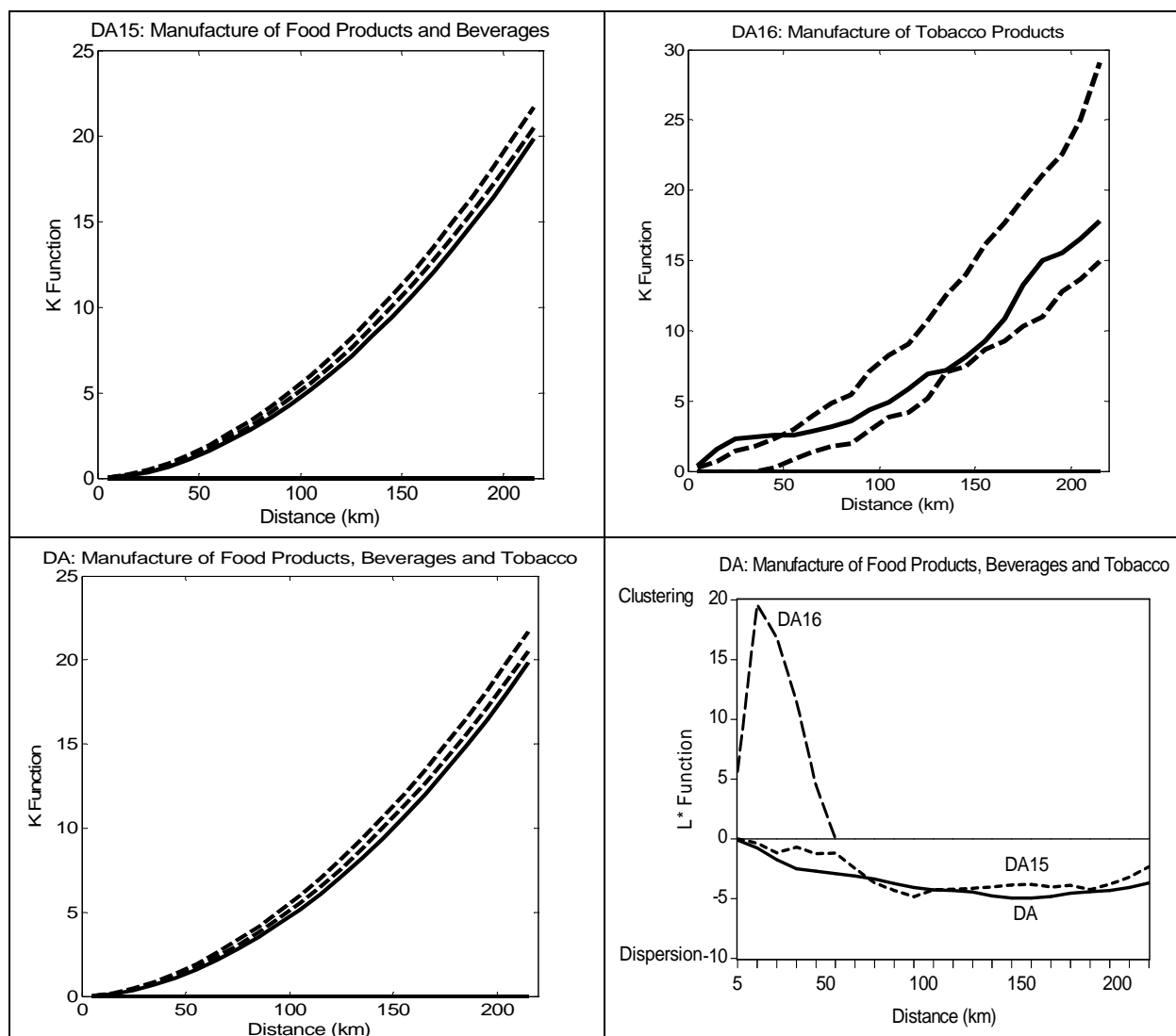


Because of the relatively small number of coal mines and quarrying plants, we use the full sample in testing for conditional concentration and dispersion of the CA industry.¹² A visual inspection of Figure 6 reveals that the CA industry is considerably more concentrated than the industry as a whole. The extent of conditional concentration is measured with the aid of the L^* function.

K function analysis for the CA industry confirms the visual finding. As all three observed K functions run well above the upper confidence bands, conditional clustering is clearly found for the CA industry as well as its CA10 and CA11 sectors at all distances. The lower right panel of Figure 7 exposes that the L^* functions run similarly to the case of unconditional concentration. Note, however, that the index of concentration for the entire CA industry has dropped to about two third.

In contrast to the CA industry, the DA industry is significantly less concentrated than other manufacturing sectors. Its K function runs below the lower 5% confidence band over the whole spatial scale. As the spatial distribution of all industrial plants is used as the reference population, the testing outcome indicates conditional regularity or dispersion.

Figure 8: K and L^* functions of the DA industry for testing conditional concentration

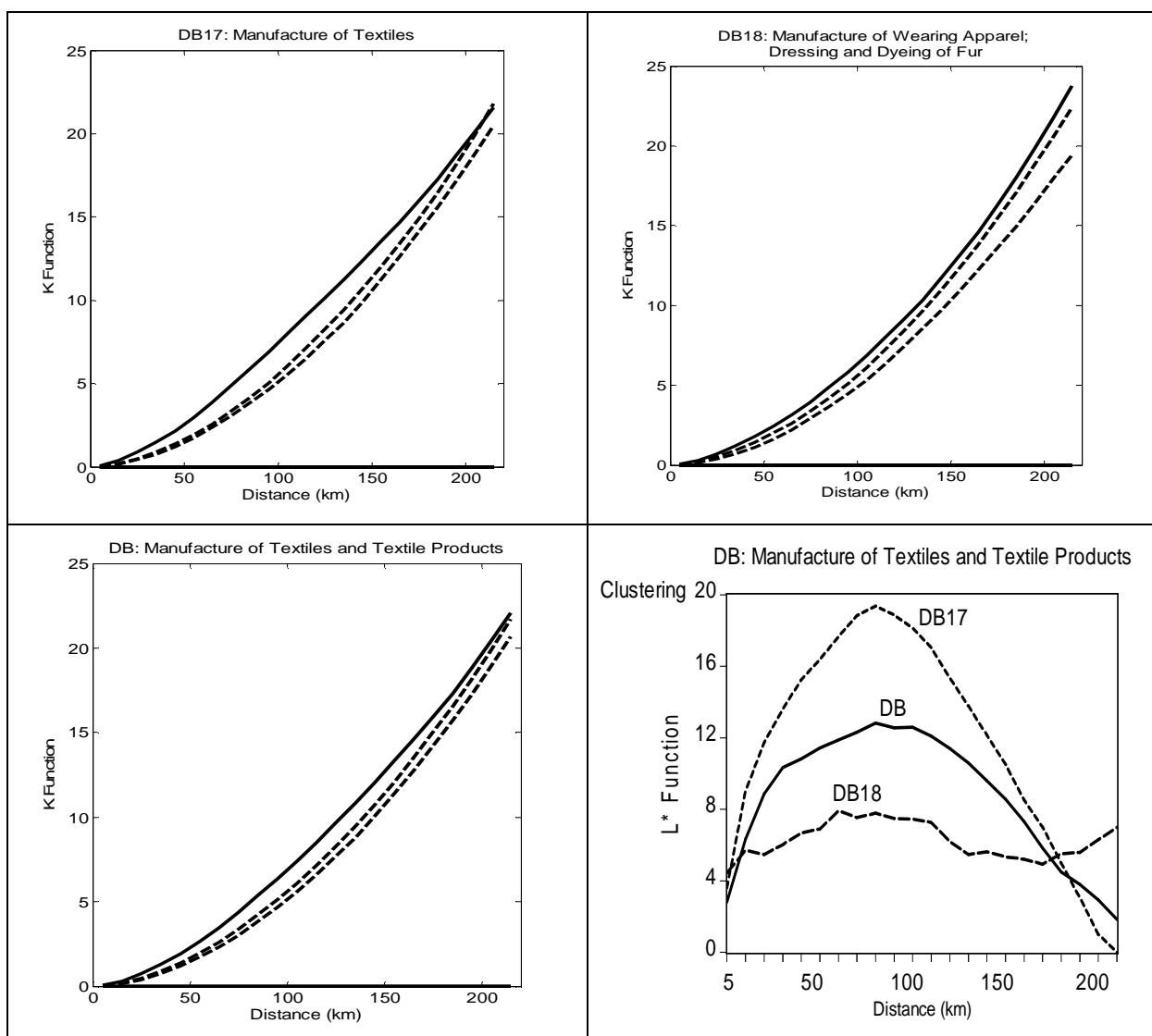


¹² In order to ensure feasibility, the test is conducted with a maximum subsample size of 500 ($= n_1$).

The same result applies for the DA15 sector. In particular the plot of the L^* functions reveals nearly identical dispersive point patterns in the hierarchically related DA and DA15 branches. By contrast, tobacco-producing plants are clustered up to a distance of 50 km, while no significant differences from the null hypothesis occur at larger spatial scales. As the tobacco sector is small compared to the food and beverages sector, its completely different type of point pattern does not substantially affect the overall tendency.

In the DB industry, conditional clustering is predominant over the whole spatial scale (Figure 9). The observed L^* function of this industry exhibits an inverted U-shaped form that runs between the sector curves. This points to a lack of substantial clustering of plants belonging to different sectors. Maximum concentration is reached at a distance of 85 km. Beyond this threshold, concentration steadily declines.

Figure 9: K and L^* functions of the DB industry for testing conditional concentration

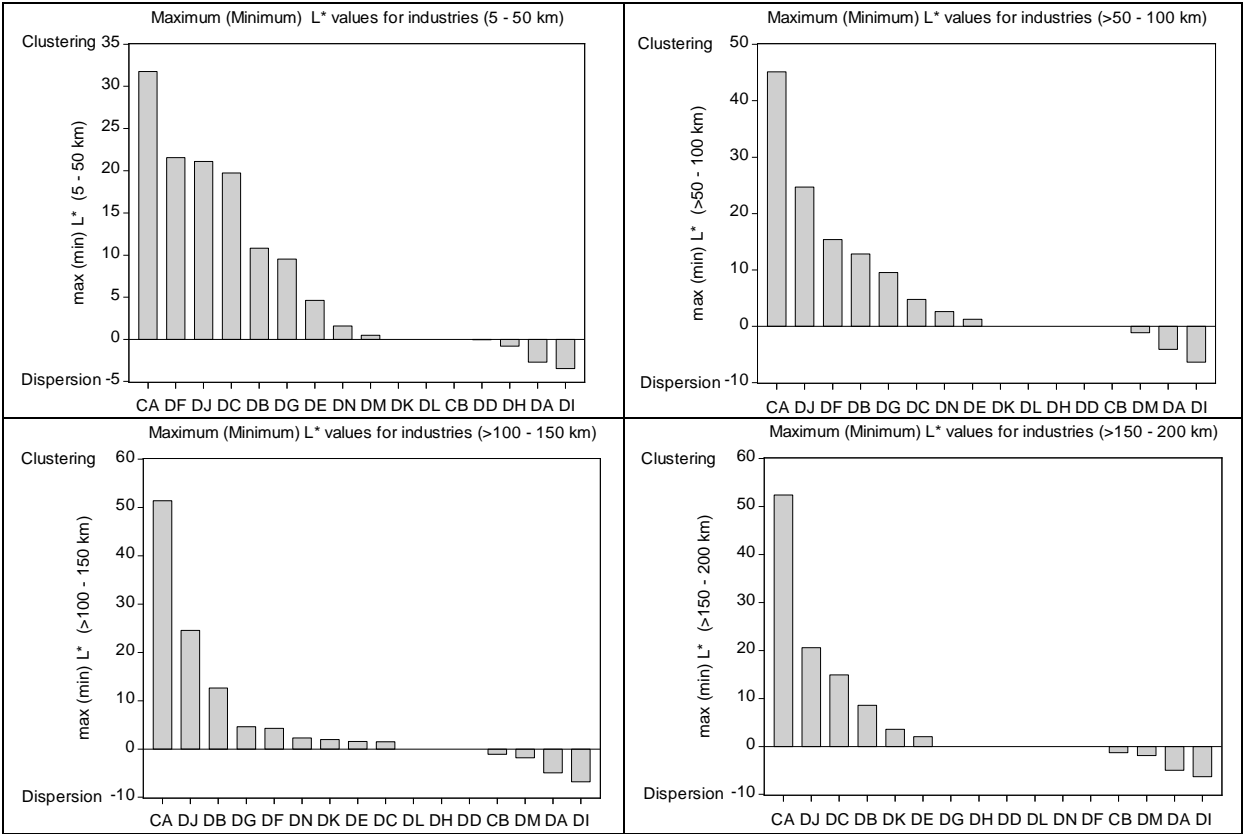


In the DB17 sector, the U-shaped L^* curve is even more pronounced than in the entire DB industry. By contrast, the extent of conditional concentration in the DB18 sector stays relatively constant over the whole spatial scale.

Because some sort of clustering is present in reference population, the concentration index L^* is dropped compared to the case of the null of independency. Moreover, when conditioning on the spatial point pattern of the industry as a whole, not only clustering but also dispersion can emerge. Figure 10 displays the testing results for the 16 industries for the same four distance bands as before. The extent of conditional clustering (dispersion) in a distance band is measured by the maximum (minimum) value of the L^* function.

Figure 10 exhibits that conditionally clustering occurs more frequently than dispersion. Clustering structures are not uniform but vary with distances. In particular, more industries are relatively strongly concentrated at short spatial scales than at larger distances. However, five out of sixteen industries, mining and quarrying (CA), manufacture of fabricated products (DJ), manufacture of leather (DC), manufacture of textiles (DB) and manufacture of pulp and paper and publishing and printing (DE) show always some degree of conditional clustering. Moreover, manufacture of non-metallic mineral products (DI) and manufacture of food products, beverages and tobacco (DA) are dispersed at any spatial scales. In contrast, both clustering and dispersion can be found in the manufacture of transport equipment (DM). Conditional clustering of manufacture of machinery and equipment (DK) and manufacture n.e.c (DN) depends on the distance at which concentration is considered. The same holds for manufacture of rubber and plastic products (DH) and other mining and quarrying (CB) with respect to conditional dispersion.

Figure 10: Conditional clustering and dispersion of industries at different spatial scales



6. Conclusions

This paper introduces a concentration index of the style of Besag's (1977) L function that is based on the concept of the K function. The index aims at measuring the extent of substantial clustering and dispersion at different spatial scales. While Besag's L function is intended to measure deviations from the CSR process, the new index can be applied to measure deviations from more general spatial processes. We also used the measure for identifying the importance of sector-specific or more general industry-specific forces inducing clustering of industries.

In testing for conditional concentration, previous papers mainly relied on Diggle and Chetwynd's (1991) D function approach. However, this approach is not efficient and feasible for evaluating clustering and dispersion in medium and large economies. We have outlined a spatial similarity test based on subsamples drawn from the industry under analysis and the entire industry as the reference population. It is illustrated how the subsample similarity test can be efficiently employed in measuring conditional concentration of German industries.

We found that some industries like coal mines and quarrying plants, manufacture of fabricated metal products and other mining and quarrying are highly concentrated at any spatial scale, while the extent of concentration and the relative positions of industries generally varies with distance. For example, while manufacture of coke, refined petroleum products and nuclear fuel is significantly clustered at a low and medium scale, no clustering at all is found for distances beyond 150 km. Coal mines and quarrying plants are highest concentrated, but the gaps to other industries increase considerably at medium and large distances. Manufacture of textiles and textile products as well a manufacture of rubber and plastic products are always dispersed compared to the industry as a whole. By contrast, evidence for dispersion is found for manufacture of food products and manufacture of transport equipment only within some distance bands.

The K function approach can as well be advantageous employed for analysing co-location between plants of different industries. Such an extension of spatial point pattern analysis could provide interesting insights on inter-industry clustering. We made a first step in this direction by assessing clustering and dispersion between hierarchical branches. With regard to the identification of Jacobs spillovers, non-hierarchical comparisons are additionally necessary. For this, univariate point pattern analysis hit the wall. In identifying attraction and repulsion of establishments across sectors, bivariate point pattern analysis seems to be a promising approach.

References

- Arbia, G., Espa, G., Quah, D. (2008), A class of spatial econometric methods in the empirical analysis of clusters of firms in the space, *Empirical Economics* 34, 81-103.
- Bailey, T.C., Gatrell, A.C. (1995), *Interactive spatial Data Analysis*, Prentice Hall, Harlow, London.
- Barff, R.A. (1987), Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio, *Annals the Association of American Geographers* 77, 89-103.

- Beaudry, C., Schiffauerova, A. (2009), Who's right, Marshall or Jacobs? The localization versus urbanization debate, *Research Policy* 38, 318-337.
- Besag, J. (1977), Contribution to the discussion of Dr. Ripley's paper, *Journal of the Royal Statistical Society B* 25, 294.
- Bickenbach, F., Bode, E. (2008), Disproportionality Measures of Concentration, Specialization, and Localization, *International Regional Science Review* 31, 359-388.
- Diggle, P.J., Chetwynd, A.G. (1991), Second-order analysis of spatial clustering for inhomogeneous populations, *Biometrics* 47, 1155-1163.
- Ellison, G., Glaeser, E. (1997), Geographic concentration in U.S. manufacturing industries: A dartboard approach, *Journal of Political Economy* 105, 879-927.
- Feser, E. J. (2000), On the Ellison-Glaeser geographic concentration index, Discussion Paper, University of North Carolina.
- Feser, E., Renski, H, Goldstein, H. (2008), Clusters and Economic Development Outcomes: An Analysis of the Link Between Clustering and Industry Growth, *Economic Development Quarterly* 22, 324-344.
- Fujita, M., Krugman, P., Venables, A.J. (1999), *The Spatial Economy: Cities, Regions and International Trade*, MIT Press, CambridgeMA.
- Helpman, E. (1998), The Size of Regions, in: Pines, D., Sadka, E., Zilcha, I. (eds.), *Topics in Public Economics*, Cambridge University Press, Cambridge, 33-54.
- Jacobs, J. (1970), *The Economics of Cities*, Penguin, London.
- Jacobs, J. (1986), *Cities and Wealth of Nations*, Random House, New York.
- Kauffman, R.J., and Kumar, A. Scale-and-scope externalities in the growth of IT industries in India: An agglomeration perspective. In: R. Sprague (ed.), *Proc. 40th Hawaii Intl. Conf. Sys. Sci.*, IEEE Comp. Soc. Press, Los Angeles.
- Krugman, P. (1991), *Geography and Trade*, MIT Press, Cambridge, MA.
- Litzenberger, T. (2006), *Cluster und die New Economic Geography*, Lang, Frankfurt/M.
- Marshall, A. (1920), *Principles of Economics*, 8th ed., Macmillan, London.
- Martinez, W.L., Martinez, A.R. (2008), *Computational Statistics Handbook with MATLAB*, 2nd ed., Chapman & Hall, Boca Raton, London.
- Menzel, M.P. (2008), Zufälle und Agglomerationseffekte bei Clusterentstehung, *Zeitschrift für Wirtschaftsgeographie* 52, 114-128.
- Neffke F.M.H., Svensson Henning M., Boschma R.A., Lundquist K.-J., Olander L.-O., 2008, Who Needs Agglomeration? Varying Agglomeration Externalities and the Industry Life Cycle, *Papers in Evolutionary Economic Geography (PEEG)* 0808, Utrecht University, Netherland.

Oxford Research (2008): Cluster Policy in Europe. A Brief Summary of Cluster Policies in 31 European Countries, Kristiansand.

Porter, M.E (2000), Location, competition, and economic development: Local clusters in a global economy, *Economic Development Quarterly* 14, 15-34.

Porter, M. (2008), Clusters and Economic Policy: Aligning Public Policy with the New Economics of Competition, Discussion Paper, Harvard Business School, Cambridge, Mass.

Porter, M., Delgado, M., Stern, S. (2006), Convergence, Clusters and Economic Performance,

Ripley, B.D. (1976), The second-order analysis of stationary point processes, *Journal of Applied Probability* 13, 255-266.

Ripley, B.D. (1977), Modelling spatial point patterns, *Journal of the Royal Statistical Society* B39, 172-212.

Smith, T.E. (2008), Notebook on Spatial Data Analysis, Web Book, <http://www.seas.upenn.edu/~ese502/#notebook>.

Südekum, J. (2006), Concentration and Specialization Trends in Germany since Re-unification, *Regional Studies* 40, 861-873.

Sweeney, S.H., Feser, E.J. (1998), Plant size and clustering of manufacturing activity, *Geographical Analysis* 30, 45-64.

Appendix: Industries and Sectors

German Classification of Economic Activities (WZ 2003)

Source: Statistisches Bundesamt, Wiesbaden

Subsections (two-letter industries)	Divisions (2-digit sectors)
CA: Mining and quarrying of energy producing materials	CA10: Mining of coal and Lignite; extraction of peat C11: Extraction of crude petroleum and natural gas
CB: Mining and quarrying, except of energy producing materials	
DA: Manufacture of food products, beverages and tobacco	DA15: Manufacture of food products and beverages DA16: Manufacture of tobacco
DB: Manufacture of textiles and textile products	DB17: Manufacture of textiles DB18: Manufacture of wearing apparel; dressing and dyeing of fur
DC: Manufacture of leather and leather products	
DD: Manufacture of wood and wood products	
DE: Manufacture of pulp, paper and paper products; publishing and printing	DE21: Manufacture of pulp, paper and paper products DE22: Publishing, printing, reproduction of recorded media
DF: Manufacture of coke, refined petroleum products and nuclear fuel	

DG: Manufacture of chemicals, chemical products and man-made fibres	
DH: Manufacture of rubber and plastic products	
DI: Manufacture of non-metallic mineral products	
DJ: Manufacture of basic metals and fabricated metal products	DJ27: Manufacture of basic metals DJ28: Manufacture of fabricated metal products, except machinery and equipment
DK: Manufacture of Machinery and Equipment	
DL: Manufacture of electrical and optical instruments	DL30: Manufacture of office machinery and computers DL31: Manufacture of electrical machinery and apparatus DL32: Manufacture of radio, television and communication equipments and apparatus DL33: Manufacture of medical, precision and optical instruments, watches and clocks
DM: Manufacture of transport equipment	DM34: Manufacture of motor vehicles, trailers and semi-trailers DM35: Manufacture of other transport equipment
DN: Manufacturing n.e.c.	DN36: Manufacture of furniture; manufacturing n.e.c. DN37: Recycling