Center for
Quantitative
Economics

Urbanization in Industrialized Countries: Appearances Are Deceptive

Ludwig von Auer[†] und Mark Trede[#]

98/2022

[†] Department of Economics, University of Trier, Germany
[#] Department of Economics, University of Münster, Germany

wissen.leben
WWU Münster

Urbanization in Industrialized Countries: Appearances Are Deceptive[*]

Ludwig von Auer[†]
Universität Trier

Mark Trede[‡]
Universität Münster

April 29, 2022

Abstract: This study introduces the *urbanicity index* of employment. This density-based measure is derived from spatial point pattern analysis and, therefore, makes use of the complete spatial information contained in geo-coded sectoral employment data. The index accounts for both the scale and the concentration aspect of urbanization. Changes in concentration can be decomposed into intersectoral mobility of employment and spatial mobility of sectors and further into the contributions of each sector of the economy. The index is applied to a large industrialized country and reveals that strong urbanization trends have occurred that simpler measures would overlook.

Keywords: agglomeration, concentration, index, measurement, migration.
JEL classification: R12, J21, C43

[*] The authors are indebted to the *Institute for Employment Research IAB* at the *Bundesagentur für Arbeit* for granting access to their employment data. They also express their gratitude to Anne Otto and Wolfgang Dauth for their help and hospitality and to Michael Pflüger for his supportive advice. The authors presented their research at a staff seminar at Universität Würzburg. Helpful comments and suggestions from participants are gratefully acknowledged.

[†] Universität Trier, Fachbereich IV, Volkswirtschaftslehre, Universitätsring 15, 54296 Trier, Germany, `vonauer@uni-trier.de`

[‡] Institut für Ökonometrie und Wirtschaftsstatistik, Fakultät für Wirtschaftswissenschaften, Universität Münster, Am Stadtgraben 9, 48143 Münster, Germany, `mark.trede@uni-muenster.de`

1 Introduction

When economists occupy themselves with the issue of urbanization, they are often interested in its impact on productivity, economic growth, or income levels. Does urbanization increase productivity, growth and income, and how strong is the effect? Prominent empirical studies of such questions include Brülhart and Sbergami (2009), Castells-Quintana and Royuela (2014), Henderson (2003), and World Bank (2009). These studies suggest that growth enhancing agglomeration occurred in developing countries, but not in high-income countries. However, using a different data set and a different measure of urbanization, Frick and Rodríguez-Pose (2018) as well as Ganau and Rodríguez-Pose (2021) come to the opposite conclusion. Agglomeration has been beneficial for economic growth in high-income countries, but not in developing countries.

This is not the only controversy that highlights the strong demand for informative data sets and for reliable measures of urbanicity.[1] Such measures are also relevant for other related research areas within economics – for example, inequality, human capital formation or economic stability – as well as for scientific debates in other disciplines such as demography, geography, sociology, history, and political science. Pertinent examples are population growth, environmental sustainability, or social and political stability. In all of these fields, empirical investigations of urbanization can benefit from a measurement approach that makes full use of detailed location data. The present paper suggests such an approach.

Ideally, the economist is endowed with geo-coded sectoral employment data. When such data are available, the measure should satisfy a set of basic properties that are discussed below. Taking these desirable properties as a reference, the present paper introduces a novel measure of urbanicity that draws a comprehensive picture of a country's urbanization trends (in the following, "country" is used synonymously for "region of interest").

There is a lot of overlap between measures of urbanicity and measures of concentration be-

[1] According to Galea et al. (2007, p. 57), the term "urbanicity" characterizes the presence of conditions at a particular point in time that distinguishes urban areas from nonurban areas, while "urbanization" refers to changes in the size, density and heterogeneity of cities over time. In the following, urbanicity is simply used as a synonym for the *degree* of urbanization.

cause both types are concerned with the spatial distribution of employment or population.[2] For example, rural-urban migration should raise the measured degree of both, concentration and urbanization. However, the overlap is only partial because measures of the degree of urbanization also should account for a scale aspect. For example, quadrupling a country's employment or population while preserving its spatial distribution should not change the measured degree of concentration. However, it should increase the measured degree of urbanization because small cities have become large cities and large cities have become megacities. These considerations lead to the first requirement for a reliable measure:

Postulate 1 A measure of urbanicity should capture and identify both, the *concentration* and the *scale* of employment (or population).

Simple traditional measures usually take account of the concentration and the scale aspect, but cannot distinguish between the two. The urbanization rate is the most prominent example. It measures the share of the country's population living in "urban areas". Measures of *urban concentration* (as opposed to measures of *urbanicity*) usually do not even intend to satisfy Postulate 1. Such measures exclude the rural population from the analysis and focus on the spatial distribution of the urban population. Examples are the share of urban population living in the largest city (urban primacy), the share of urban population living in cities above a certain size threshold, and the sum of the individual cities' squared shares of total urban population (Hirschman-Herfindahl index).

A common question arising for the urbanization rate as well as for measures of urban concentration is: What are appropriate criteria for classifying the regions of a country either as "urban" or "non-urban"? Unfortunately, there is still no consensus on acceptable and operational criteria for this classification. This is the *urban-rural dichotomy* issue (e.g. Stewart Jr., 1958). It complicates international comparisons and causes reservations about the robustness of the empirical findings. However, when sufficiently rich data sets are at hand, the urban-rural dichotomy and its associated problems can be avoided. The following two postulates refer to such situations:

[2] A comprehensive survey of existing measures of concentration is Nakamura and Morrison Paul (2019).

Postulate 2 A measure of urbanicity should avoid the issues caused by the urban-rural dichotomy.

Postulate 3 A measure of urbanicity should allow for meaningful international comparisons.

The index by the demographer Arriaga (1970, p. 209) satisfies both postulates. It measures the average (or expected) employment of the country's municipalities as perceived by the country's employment. For example, an index value of 1000 would say that, on average, each employee in the country can expect to have 999 other employees in her "neighbourhood". It is to the credit of Lemelin, Rubiera-Morollón, and Gómez-Loscos (2016, pp. 594-595) that economists became aware of that index. The authors point out that the Arriaga index can be expressed as the product of the country's population and the Hirschman-Herfindahl index of the regional population shares. In the terminology of Postulate 1, the former accounts for the scale aspect, while the latter captures the concentration aspect. More specifically, the Arriaga index is the product of a *scale factor* (size of employment or population) and a *concentration factor* (Hirschman-Herfindahl index).

The urbanization rate, the Arriaga index, as well as the measures of urban concentration are "discrete" measures of urbanicity. Such measures are designed for the analysis of regionalized data. To obtain such data, the country's territory must be fragmented into well-defined regions or "neighbourhoods". Identifying adequate neighbourhoods can turn out to be a truly challenging task (e.g., Briant, Combes, & Lafourcade, 2010, pp. 288-289). This issue is known as the "modifiable area unit problem" (e.g., Arbia, 1989; Openshaw & Taylor, 1979). Even if the delineation of neighbourhoods were rather uncontroversial, another disadvantage of discrete measures would remain. Once the delineation is completed, the spatial dimension is fully eliminated, prompting a loss of information (e.g. Arbia, 2001, p. 273). The neighbourhoods are considered as fully comparable atomistic units without any interdependencies whatsoever.

In more and more countries, researchers are granted access to geo-coded sectoral employment data. Utilizing the spatial information contained in such data, one can avoid the modifiable area unit problem and the problem of atomistic neighbourhoods. This consideration leads to the fourth basic requirement that a measure of urbanicity should satisfy.

Postulate 4 A measure of urbanicity should not waste any spatial information contained in the data set.

When geo-coded employment data are available, the urbanization rate and the Arriaga index violate Postulate 4 because they are "discrete" measures of urbanicity. Preferable is a spatial point pattern analysis. This field of spatial statistics includes quadrat count methods, distance-based methods, and density-based methods. The spatial analysis of economic concentration is dominated by two classes of distanced-based methods. Duranton and Overman (2005) propose to compute the distribution of all pairwise distances between the observed firms. In a second step, the distribution of the pairwise distances is smoothed by kernel density estimation. The result is denoted as the K_d function. For statistical inference, the K_d function can be compared to the smoothed distribution of pairwise distances that would result under the complete spatial randomness hypothesis. The other popular class of distance-based methods elaborates Ripley's $K(d)$-function (e.g. Marcon & Puech, 2003, 2010).[3] This function shows for every radius d the average number of other firms covered by circles of radius d drawn around the observed firms. To conduct statistical inference, the $K(d)$ function can be compared to the corresponding function that would arise under the complete spatial randomness hypothesis.

Both, the K_d-function of Duranton and Overman (2005) and Ripley's $K(d)$ function allow for statistical inference. Since the seminal work of Ellison and Glaeser (1997), this property has been considered as indispensable for reliable measures of economic concentration. The same considerations apply to measures of urbanicity.

Postulate 5 A measure of urbanicity should allow for statistical inference.

Having a meaningful measure of a country's urbanicity is a welcome contribution. However, it would be even more useful to have a measure that also quantifies and compares the urbanicity of individual sectors. This requires sectoral employment data. With such data, it is seductive to simply measure each sector's urbanicity by its degree of concentration. However, this approach would be inappropriate. As emphasized by Auer, Stepanyan, and Trede (2019), there are many concentrated sectors that are distinctly rural.

[3] Lang, Marcon, and Puech (2020) and Marcon and Puech (2017) survey this class of methods.

Postulate 6 A measure of urbanicity should be able to quantify and rank the of urbanicity of the individual sectors.

From the mid-20th century until today, the urbanization rates of developing countries moved quickly upwards, while the urbanization rates of developed countries were rather flat (e.g. Jedwab & Vollrath, 2015). It is tempting to conclude that urbanization is no longer an issue in the developed world. However, a reliable economic analysis of a country's urbanization trends requires a decomposition of the measure into the two principal forces that drive the overall result.

Postulate 7 A measure of urbanicity should decompose changes in concentration into two components: *intersectoral mobility of employment* and *spatial mobility of sectors*.

For example, when farmers leave their farms and switch to urban manufacturing sectors, this reflects intersectoral mobility of employment. Spatial mobility of sectors arises when the receiving manufacturing sectors absorb the additional workforce by expanding their rural production sites. Then, these manufacturing sectors become more rural. The two mobility aspects have opposing effects on the country's degree of concentration and, therefore, urbanization. Thus, even if the measures of urbanicity indicate no overall change (as in industrialized countries since the mid-20th century), it would be a mistake to conclude that no relevant urbanization trends occurred.

Since the urbanization rate and the Arriaga index are not designed for sectoral employment data, they violate Postulate 7. By contrast, distance-based measures of concentration such as the elaborations of the K_d and the $K(d)$ function often use sectoral employment data. Nevertheless, they make no attempt to decompose the changes in concentration into their two principal components.

Besides the decomposition into the principal components, a detailed and comprehensive picture of a country's urbanization trends requires a second dimension of decomposition. It would be very desirable to know which sectors are responsible for these trends. Therefore, a measure of urbanicity should also satisfy the following requirement.

Postulate 8 A measure of urbanicity should be able to identify the sectoral contributions to the overall change and to the principal components of this overall change.

For distance-based methods, including the K_d and $K(d)$ function, the decomposition properties specified by Postulates 7 and 8 pose serious problems. Moreover, distance-based methods start by transforming the observed point pattern into a pattern of observed pairwise distances d. The K_d and $K(d)$ function provide summary statistics for the degree of concentration and they represent the pattern of distances in a condensed form. Such forms give a rather sketchy impression of the underlying point pattern from which the distances were derived. Thus, when a detailed study of a country's urbanicity trends is required, a more direct utilization of the observed point pattern would be desirable.

Therefore, the present study leaves the field of distance-based methods and, instead, prefers a density-based approach. The paper makes two main contributions, one is theoretical and the other empirical. The theoretical one is the *urbanicity index* of employment. It is estimable from geo-coded sectoral employment data and satisfies Postulates 1 to 8. That is, it distinguishes between the scale aspect and the concentration aspect of urbanization, it avoids the urban-rural dichotomy, it allows for international comparisons, it makes full use of the available information in the data set, it is rooted in point pattern analysis and, thus, allows for statistical inference, it measures the urbanicity of each individual sector, it decomposes changes in concentration into intersectoral mobility of employment and spatial mobility of sectors, and it identifies each sector's contribution to these changes.

The urbanicity index of employment is a density-based method. More specifically, a kernel density estimation of the observed spatial distribution of employees transforms the country's map into a smooth surface with peaks in densely populated urban areas and lowlands in rural areas. The surface shows the estimated spatial density of employment and, therefore, provides a more direct and accurate visual impression of the country's employment pattern than would be possible by any inference from the distance-based K_d or $K(d)$ function. Analogous kernel density estimations are conducted for each individual sector. The resulting density estimations are the input to a rigorous statistical analysis that allows for statistical inference and an elucidating decomposition of the overall result.

The empirical contribution of the present paper is a detailed analysis of the urbanicity of employment in a large industrialized country. The high quality of its administrative employment data made Germany an ideal showcase for the scientific potential of the urbanicity index. The analysis uses sectoral micro-data on regional employment relating to the years 1995, 2000, 2005, 2010, and 2014.

Quite in line with the results of Jedwab and Vollrath (2015), the data show that between 1995 and 2014 the urbanicity in Germany as measured by the urbanicity index increased only slightly. However, decomposing the urbanicity index reveals that the apparent stagnation is deceptive. Beneath the calm surface, the two concentration factors moved in opposite directions: The intersectoral mobility of employment increased the overall index, that is, employees left rural sectors for more urban ones. However, this impact was largely offset by the spatial mobility of the German sectors. The sectors shifted their employment towards more rural regions.

These insights prompt a list of interesting follow-up questions. Which sectors can be considered as urban and which as rural? Which urban sectors received the employees that left the rural sectors? Are these the same urban sectors that shifted their employment towards more rural regions? Are the aggregated numbers driven by few influential sectors or do the changes occur across the board? The urbanicity index can also answer all of these follow-up questions.

The rest of the paper proceeds as follows. Section 2 starts out by defining and explaining the urbanicity index. Then it demonstrates how to compute the index using spatial employment data and how to perform statistical inference for changes of the urbanicity index over time. As its first empirical contribution, the paper describes the data set, calculates the urbanicity index for Germany using administrative employment data, and tests for intertemporal changes of the index. Section 3 introduces a measure of urbanicity of individual sectors along with a simple procedure for statistical inference. The measure and the associated inference are applied to all sectors in the German economy. In Section 4 it is shown that an intertemporal change of the urbanicity index number can be factorized into the scale effect and the concentration effect and that the latter can be additively decomposed into the intersectoral mobility of employment and the spatial mobility of sectors. The decomposition is then performed for the German employment data. Section 5 explains how the changes in the intersectoral mobility of employment and the spatial mobility of sectors can be further decomposed into the contributions of the individual sectors of the economy. Additional findings about the sectors' mobility in Germany are also presented. Section 6 concludes.

2 Measuring Urbanicity

2.1 Definitions

We consider some country with area G. The size of the area is $|G| = \int_G dx$ where the location variable x contains the longitudinal and latitudinal coordinates and varies over the area G.[4] The density function of the country's employment E is denoted by $f_E(x)$. Of course, $\int_G f_E(x)dx = 1$. A completely uniform distribution over G has a constant density of $f_E(x) = 1/|G|$ everywhere.

We suggest to measure the concentration aspect of urbanicity by the *concentration factor* which we define as the (normalized) expected density of total employment as perceived by a randomly drawn employee,

$$a_E = |G| \cdot \mathbb{E}(f_E(x)) = |G| \int_G f_E(x)^2 dx. \tag{1}$$

The integral $\int_G f_E(x)^2 dx$ can be regarded as a spatial continuous version of the Hirschman-Herfindahl index. It avoids the urban-rural dichotomy (Postulate 2). Since the concentration factor is normalized by the area $|G|$ it does not depend on the spatial unit of measurement (e.g. square kilometres or square miles). As a consequence, a uniform distribution of employment yields $a_E = 1$. The range of a_E is the interval $[1, \infty)$. The more concentrated the distribution of employment, the larger the value of a_E. It diverges to $+\infty$ if total employment is concentrated at a single point.

As long as the distribution function of employment does not change, scaling up or down the number of employees, E, should and would not change the value of the concentration factor, a_E. However, a meaningful measure of urbanicity, u_E, should reflect not only the concentration but also the scale of employment (Postulate 1). To obtain such a measure, the concentration factor, a_E, is multiplied by a *scale factor*. We propose the average number of

[4] The compact integral notation $\int_G (\cdot)dx$ for area G denotes the two-dimensional integral $\int_{\text{min\_lon}}^{\text{max\_lon}} \int_{\text{min\_lat}}^{\text{max\_lat}} (\cdot) G(\text{lon}, \text{lat}) d\text{lat}\, d\text{lon}$ where the function $G(\text{lon,lat}) = 1$ if the geo-coordinate with longitude lon and latitude lat belongs to the country area, and $G(\text{lon,lat}) = 0$ elsewhere.

employees per unit area, $E/|G|$, as scale factor. This factor makes countries of different size comparable (Postulate 3).

Accordingly, we define the *urbanicity index of total employment* as

$$u_E = \frac{E}{|G|}\, a_E. \tag{2}$$

Its range is $[E/|G|, \infty)$. Rewriting (2) as $u_E = E \int_G f_E(x)^2 dx$ reveals that the urbanicity index reflects the *expected number of employees per unit area as perceived by a randomly drawn employee*, while the concentration factor a_E is the normalized expected employment density.

Some parallels exist between the urbanicity index, u_E, and the Arriaga index (the product of total employment, E, and the spatial Hirschman-Herfindahl index). Both indices comprise a concentration factor and a scale factor. The scale factors of both indices are related to total employment size, and both concentration factors are related to the expected employment density as perceived by a randomly drawn employee. However, the Arriaga index is not designed for statistical inference (violation of Postulate 5) and for the comparison of individual sectors (violation of Postulate 6; though a modification of the index would solve this problem). Furthermore, the treatment of space differs strongly between the indices. Arriaga's approach requires an ex-ante definition of spatial units or neighbourhoods and, hence, is subject to the modifiable area unit problem and the problem of atomistic neighbourhoods (violation of Postulate 4). By contrast, the concept of space underlying the urbanicity index (2) is continuous, and while a unit of space (e.g. square miles) is, of course, still required, it does not enter the index. Another important advantage of the urbanicity index is its decomposability (Postulates 7 and 8), an issue which receives a more in-depth treatment in Sections 4 and 5 below.

2.2 Estimation

In empirical applications, the theoretical employment density $f_E(x)$ is not known, but needs to be estimated from employment data. For a given year, let the set of all companies be denoted by C. Let w_c denote company c's number of full-time equivalent employees and $x_c = (x_{c1}, x_{c2})$ its geo-coded location. This information allows us to estimate the employment density of total employment, $f_E(x)$, by the kernel density estimator

$$\widehat{f}_E(x) = \frac{1}{\sum_{c \in C} w_c} \sum_{c \in C} w_c K_h(x, x_c) \ , \tag{3}$$

where $x = (x_1, x_2)$ is the position at which the density is evaluated and h is the bandwidth. We will use the Gaussian product kernel

$$K_h(x, x_c) = \frac{1}{h^2} \, \phi\left(\frac{x_1 - x_{c1}}{h}\right) \phi\left(\frac{x_2 - x_{c2}}{h}\right)$$

with standard Gaussian density $\phi(z) = (1/\sqrt{2\pi}) \exp(-0.5z^2)$.

The impact of the type of kernel function $K_h(x, x_c)$ on the shape of the estimated density is relatively small. By contrast, the choice of the bandwidth h has a strong impact. If the bandwidth is very small, the density belonging to a single company is highly concentrated in its immediate neighbourhood and the densities of two companies do not overlap noticeably even if they are located relatively close to each other. One could interpret the bandwidth h as defining the "effective neighbourhood", or "impact region", of a company. About 86 percent of the impact occurs within a distance of $2h$, around 98.9 percent within a distance of $3h$. Silverman (1986, chap. 4.2.1) proposes a bandwidth of $\sigma|C|^{-1/6}$ where $|C|$ is the number of observations (firms) and σ is the average standard deviation of the longitudinal and latitudinal coordinates of the firms. In the empirical analysis, we will see that in the context of our employment data this rule generates an effective neighbourhood that is too large (see Figure 2).

Typically, the number of firms (that is, the number of observations) is very large. Computing the kernel density (3) at a single point x requires to evaluate the kernel function K_h for every firm. Since we have to compute the kernel density at a large number of points, it is critically important to apply computationally efficient algorithms. Gramacki (2018) suggests to use fast Fourier transforms (FFT) to speed up the kernel density estimations in big data settings.[5]

[5] The FFT based kernel density estimation method of Gramacki (2018) is implemented in the R package `ks`. The computations in the empirical part of the paper have been conducted using version 1.13.2 of the package.

The estimated counterpart of the concentration factor (1) is

$$\widehat{a}_E = |G| \int_G [\widehat{f}_E(x)]^2 dx. \tag{4}$$

In the appendix we demonstrate the close relationship between our density-based measure (4) and the distance-based approach of Ripley's $K(d)$ function. In addition, we explain the more intricate relationship to the K_d function of Duranton and Overman (2005).

For empirical applications, the integral in (4) has to be computed numerically. The simplest way is to approximate it by finite sums over a fine grid with equidistant points on G. Let $\tilde{x}_m$ $(m = 1, \dots, M)$ denote the grid points, that is, each $\tilde{x}_m = (\tilde{x}_{m1}, \tilde{x}_{m2})$ is a pair of coordinates. Denote the grid's longitudinal and latitudinal step sizes by d_1 and d_2 as shown in Figure 1.

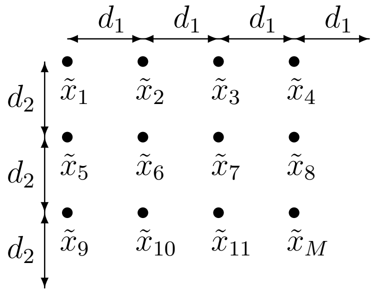

Figure 1: Grid points

If the grid is sufficiently fine (and hence M sufficiently large), the integral in (4) can be accurately approximated by a sum, and the point estimator of the urbanicity index is

$$\widehat{u}_E = \frac{E}{|G|}\widehat{a}_E \approx E \sum_{m=1}^{M} \widehat{f}_E(\tilde{x}_m)\widehat{f}_E(\tilde{x}_m) d_1 d_2. \tag{5}$$

2.3 Inference

When estimating the urbanicity index for two periods, a routine question to ask is whether the change in the index is statistically significant. A natural test statistic is the absolute difference of the estimated urbanicity indices in the two periods $T = |\widehat{u}_{E,1} - \widehat{u}_{E,2}|$. The null hypothesis of no change is rejected if the test statistic is larger than the critical value. The distribution of the test statistic under the null hypothesis (and hence the critical value) is determined by a bootstrap method.

Under the null hypothesis the spatial distribution of employment in period 1 is the same as the spatial distribution in period 2. Bootstrap resamples can therefore be generated in the following way. First, all company locations of both periods are merged. Let $|C_1|$ and $|C_2|$ denote the number of observations in period 1 and 2. Then the merged distribution contains $|C_1| + |C_2|$ locations. If a company is part of the sample in both periods without moving its location, this location appears twice in the merged distribution. Second, $|C_1|$ locations are randomly drawn without replacement from the merged distribution. These observations constitute the bootstrap resample for period 1. The remaining locations form the resample for period 2. For each bootstrap replication, the urbanicity indices of both resamples are computed as well as their absolute difference,

$$T^{(b)} = |\widehat{u}_{E,1}^{(b)} - \widehat{u}_{E,2}^{(b)}|$$

for $b = 1, \ldots, B$ where B is the number of bootstrap replications. The B realizations of the test statistic approximate the distribution of the test statistic T under the null hypothesis. The null hypothesis is rejected at significance level α if the actually observed value of the test statistic is larger than the $(1-\alpha)$-quantile of $T^{(1)}, \ldots, T^{(B)}$. Alternatively, on can calculate the p-value of the test as the fraction of bootstrapped test statistics that are larger than the actually observed test statistic.

In the following two subsections, we apply these concepts to administrative German employment data.

2.4 Data

German employment data for 1995, 2000, 2005, 2010, and 2014 have been provided by the *Institute for Employment Research IAB* at the *Bundesagentur für Arbeit.* The data set contains information about all companies with employees subject to social security contributions. Since social security contributions and benefits are calculated on the basis of these data, their reliability far outperforms survey data. Information about each company includes the number of employees in different employment types (full-time, part-time, apprentices etc.), its location, and the sector in which the company mainly operates (Schmucker, Seth, Ludsteck, Eberle, & Ganzer, 2016).

We aggregate the number of employees in the different employment types in company c to the number of full-time equivalent employees w_c. Table 1 reports some descriptive statistics about the companies and employment numbers. The number of companies has increased by about 50 percent over the 19-year observation period (that is, on average about 2.2 percent annually). Total employment (E) fluctuated around 25 million (full-time equivalent employees). From 2005 to 2014 it increased by almost 10 percent. The mean and median number of full-time equivalent employees per company has decreased. The number of employees is very skewed, the largest 0.1 percent of companies have more than 300 times as many full-time equivalent employees as the median company.

	1995	2000	2005	2010	2014
# companies	1 953 521	2 522 771	2 668 859	2 887 117	2 927 359
Total empl. (E)	24 127 918	25 249 566	23 519 820	24 781 540	25 868 698
Full-time equivalent employees per company					
Mean	12.35	10.01	8.81	8.58	8.84
Median	3.00	2.00	1.75	1.50	1.50
Q(0.99)	160.00	131.25	118.75	118.00	122.00
Q(0.999)	856.00	666.25	608.00	590.00	619.50

Table 1: Number of companies, total number of full-time equivalent employees, and descriptive statistics for the number of full-time equivalent employees per company for 1995 to 2014.

The sectors are categorized according to the German WZ Classification Code (Statistisches Bundesamt, 2008). This code mimicks the United Nations *International Standard Industrial Classification (ISIC)* and the *Nomenclature statistique des activités économiques dans la Communauté européenne (NACE)*. As the classifications are subject to periodic revisions, the number of sectors and their composition change over time. To ensure comparability of sector classifications in different years, the data set contains a consolidated 3-digit classification, based on WZ 1973, that does not change between 1995 and 2014. We eliminate as outliers very small sectors with less than 10 companies or less than 100 employees. The number of sufficiently large sectors is always above 215.[6]

[6] A detailed table listing the number of companies and the number of full-time equivalent employees for

Concerning the companies' locations, we know the municipality ("Gemeinde") where each company $c \in C$ is located. The number of municipalities or regions is roughly 11000.[7] As our measurement approach is based on geo-coded locations, we assign geo-coordinates $x_c = (x_{c1}, x_{c2})$ to each company.[8] The geo-coordinates of each company are randomly sampled from a uniform distribution over the area of the region where the company is located. Of course, this approach entails a slight loss of information compared to a situation where the exact geo-coded locations of all companies are known.

2.5 Urbanicity Index of Germany

We proceed to estimate the urbanicity index of Germany for the years 1995, 2000, 2005, 2010, and 2014. As argued above, the choice of the bandwidth is relevant. Figure 2 shows two heat maps of the estimated density of total employment in Germany. The left map is constructed using a bandwidth of $h = 5000$ (that is, 5 kilometres), while the right map uses $h = 15000$. The value $h = 15000$ results from the formula provided by Silverman (1986) (see Section 2.2 above) rounded to the nearest 1000. However, this bandwidth causes the urban agglomerations in the Rhine-Ruhr area (large dark area in the far West of Germany) to be merged into a single metropolitan area. We prefer to be able to distinguish between urban centres and the less urban regions between them and, therefore, choose a bandwidth of $h = 5000$.

The grid points for computing the values of $\widehat{u}_E$ and $\widehat{a}_E$ by formula (5) have longitudinal and latitudinal distances $d_1 = d_2 = 1000$ metres resulting in $M = 612\,234$ grid points. The employment numbers, E, are taken from Table 1. The size of Germany is $|G| = 357\,839$ km^2. The results are listed in Table 2. The expected number of employees in

each sector in each observation year is available as a csv file on request from the authors. The table can also be viewed in the web appendix of the paper.

[7] There are minor changes in the number of regions due to occasional reshapings, e.g. mergers of municipalities.

[8] The geo-coordinate system is UTM32, and the geo-data about the municipalities are provided by the "Bundesamt für Kartographie und Geodäsie". The coordinates are measured in metres, they extend from 280371.1 in the East to 921292.4 in the West (longitude) and from 5235856.0 in the South to 6101443.7 in the North (latitude).

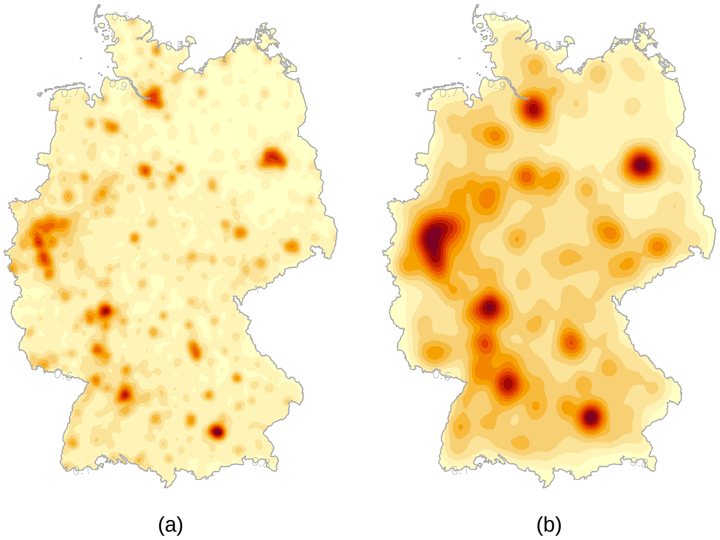

Figure 2: Comparison of kernel estimated densities (a) with bandwidth $h = 5000$ and (b) with $h = 15000$. The colour scales are not the same.

the neighbourhood, $\widehat{u}_E$, is roughly 280 per km^2. A completely uniform distribution would result in about 70 employees per km^2. Intertemporal changes in $\widehat{a}_E$ and $E/|G|$ translate into changes of $\widehat{u}_E$. Table 2 shows that, between 1995 and 2014, the urbanicity index, $\widehat{u}_E$, increased non-monotonically from 269.6 to 297.9, that is, by 10.5 percent. More specifically, the concentration factor, $\widehat{a}_E$, increased by 3.1 percent and the scale factor, $E/|G|$, by 7.2 percent. Note that $1.072 \cdot 1.031 = 1.105$. The numbers reveal that the urbanization trend is largely driven by a positive scale effect rather than a trend towards increasing concentration of employment.

Concerning statistical inference, the null hypothesis of no change in urbanicity cannot be rejected (at a significance level of 5 percent) for any consecutive years, as shown in the bottom row of Table 2. The p-values are all above the 5% level, the smallest p-value (0.09) is recorded for the test of no change between 2010 and 2014. The test for the total 19-year period from 1995 to 2014 rejects the null hypothesis at the 5% level but not at the 1% level (the p-value being 0.04).

Overall, Table 2 conveys the impression that Germany did not go through remarkable urbanization trends. However, such a conclusion would be premature. A careful decomposi-

	1995	2000	2005	2010	2014
$\widehat{u}_E$	269.614	281.607	264.826	278.335	297.911
$E/\|G\|$	67.427	70.561	65.727	69.253	72.291
$\widehat{a}_E$	3.999	3.991	4.029	4.019	4.121
p-value		0.48	0.71	0.55	0.09

Table 2: Evolution of urbanicity indices $\widehat{u}_E$ (perceived employees per km^2), scale factors $E/|G|$, concentration factors $\widehat{a}_E$ of employment in Germany, and p-values for the test of no change in urbanicity between the preceding and the current observation year. Differences between $\widehat{a}_E \cdot E/|G|$ and $\widehat{u}_E$ are due to rounding errors.

tion of the concentration factor, a_E, reveals that below the deceptive surface of the relatively stable numbers listed in Table 2 important shifts occurred. The decomposition is conducted in two stages as suggested by Postulates 7 and 8. In the first stage, the change in the concentration factor is decomposed into the intersectoral mobility of employment (IME) and the spatial mobility of sectors (SMS). The underlying theory and the application to the German employment data are presented in Section 4. Section 5 is devoted to the second stage, that is, the IME and SMS are further decomposed into the contributions of the individual sectors. However, before we turn to the two stages of the decomposition analysis, we derive from the urbanicity index, u_E, a measure of the urbanicity of individual sectors.

3 Urbanicity of Individual Sectors

3.1 Definitions

To compare the urbanicity of individual sectors, a reliable measure must be derived. If we were concerned with a sector's degree of concentration, we could use the measure $|G| \int_G f_i(x)^2 dx$, quite in analogy to the concentration factor defined in (1). The measure is the normalized expected density of sector i employment as perceived by a randomly drawn employee of sector i. However, this ist not what we need for measuring the sector's urbanicity. Instead, we need the normalized expected density of *total employment* experienced by a randomly drawn employee of sector i. This is given by

$$a_i = |G| \int_G f_E(x) f_i(x) dx. \tag{6}$$

Just like a_E, this expression can be interpreted as a measure of concentration of total employment. The only difference is the perspective. While a_E is the density as perceived by the employees of all sectors, a_i is the density as perceived by sector i. Therefore, a_i is the concentration factor of our measure of sectoral urbanicity. Again, $E/|G|$ offers itself as scaling factor. Multiplication of the two factors gives

$$u_i = \frac{E}{|G|} a_i. \tag{7}$$

This is the *expected number of employees (of all sectors) per unit area as perceived by a randomly drawn employee of sector i.* The sector with the largest u_i-value exhibits the largest degree of urbanicity. Therefore, we denote u_i as the *urbanicity index of sector i.*

Generally, the values of a_i and u_i increase as employees from sector i move from regions with low total employment density to regions with high total employment density. The same effect occurs if total employment shifts from regions with low sector i employment density to regions with high sector i employment density. Formally, this symmetry follows from the fact that

$$|G| \int_G f_i(x) f_E(x) dx = |G| \int_G f_E(x) f_i(x) dx.$$

This implies that the expected density of sector i employment of a randomly selected employee of all sectors is equal to the expected density of total employment of a randomly selected employee from sector i.

The range of the urbanicity index of total employment, u_E, is $[E/|G|, \infty)$, whereas the range of the urbanicity index of sector i, u_i, is $(0, \infty)$. If sector i is located in places where no other employees are, the index reaches its minimum value: $u_i \approx 0$. The index can become infinitely large when some sector i is concentrated in a location where the density of total employment is very high. If sector i is uniformly distributed across the country, then $f_i(x) = 1/|G|$ everywhere and, therefore, $a_i = 1$ and $u_i = E/|G|$, regardless of the distribution of total employment, E.

The urbanicity index of sector i, u_i, is a relative measure in the sense that its value depends on the distribution of both, sector i employment and total employment. To distinguish between rural and urban sectors, we define the *coefficient of urbanicity* of sector i as

$$U_i = u_i - u_E.$$

When $U_i > 0$, the expected number of employees (of all sectors) per unit area as perceived by sector i employees is larger than the expected number of employees (of all sectors) per unit area as perceived by all employees. Therefore, the sector can be considered as (relatively) urban. Conversely, if $U_i < 0$, sector i is (relatively) rural. Note that $U_i > 0$ if and only if $a_i - a_E > 0$.

3.2 Estimation

In Section 2.2, we described how $\widehat{u}_E$ can be computed by formula (5). A perfectly analogous formula can be used for the computation of $\widehat{u}_i$:

$$\widehat{u}_i = \frac{E}{|G|}\widehat{a}_i \approx E \sum_{m=1}^{M} \widehat{f}_i(\tilde{x}_m)\widehat{f}_E(\tilde{x}_m) d_1 d_2. \tag{8}$$

The point estimator of the coefficient of urbanicity of sector i is simply

$$\widehat{U}_i = \widehat{u}_i - \widehat{u}_E = E \sum_{m=1}^{M} \left[\widehat{f}_i(\tilde{x}_m) - \widehat{f}_E(\tilde{x}_m)\right] \widehat{f}_E(\tilde{x}_m) d_1 d_2. \tag{9}$$

3.3 Inference

We consider two types of hypothesis. First, for the static case, we develop a hypothesis test about the urbanicity of a sector: Is a given sector significantly urban or rural? Second, we suggest a procedure to test hypotheses about the evolution of urbanicity over time: Is the intertemporal change of a sector's coefficient of urbanicity statistically significant?

As to the static case, the natural null hypothesis states that employment in sector i follows the same spatial distribution as total employment (resulting in $U_i = 0$). The corresponding alternative hypothesis postulates that sector i has a different spatial distribution, that is, it is either more rural or more urban than overall employment. One-sided alternative hypotheses are, of course, also possible, but disregarded here as it is straightforward to adapt the procedure.

The obvious test statistic is the coefficient of urbanicity $\widehat{U}_i$ and the null hypothesis is rejected if the test statistic is larger than an upper critical value (for a significantly urban sector) or smaller than a lower critical value (for a significantly rural sector). The distribution of the test statistic and the critical value can be determined by bootstrapping. Under the null hypothesis the spatial distribution of employment in sector i equals the spatial distribution of total employment. To preserve the company size distribution we generate the pseudo-samples in a three-step procedure. In the first step, a set of $|C_i|$ company locations of sector i are randomly drawn from the C company locations of total employment with sampling weights proportional to the number of employees at each location, w_c for $c \in C$. In the second step, the observed employment shares of each company of sector i are computed: $\breve{w}_c = w_c/(\sum_{c' \in C_i} w_{c'})$ for $c \in C_i$. These shares are randomly assigned to the set of company locations of sector i. In the final step, the individual employees of sector i are distributed over the locations with $\breve{w}_c$ as the locations' sampling weights.

For each bootstrap resample, the coefficient of urbanicity is computed, say $\widehat{U}_i^{(1)}, \ldots, \widehat{U}_i^{(B)}$ where B is the number of bootstrap replications. The null hypothesis is rejected at significance level α if the observed value $\widehat{U}_i$ is less than the $\alpha/2$-quantile of $\widehat{U}_i^{(1)}, \ldots, \widehat{U}_i^{(B)}$, or if it is greater than the $(1-\alpha/2)$-quantile. Alternatively, one can calculate the p-value of the test. It is the minimum of two proportions, namely (i) the proportion of $\widehat{U}_i^{(b)} < \widehat{U}_i$, and (ii) the proportion of $\widehat{U}_i^{(b)} > \widehat{U}_i$.

The procedure for testing hypotheses about the development over time of the coefficient of urbanicity is similar to the method described in Section 2.3. In fact, since the difference of the coefficients of urbanicity, $\widehat{U}_{i,1} - \widehat{U}_{i,2}$, is closely related to the difference of the urbanicity indices, $\widehat{u}_{i,1} - \widehat{u}_{i,2}$, the test approach is virtually identical.

3.4 Sectoral Urbanicity in Germany

The sectors' urbanicity indices, $\widehat{u}_i$, and their coefficients of urbanicity, $\widehat{U}_i$, are computed according to (8) and (9). Figure 3 displays the cumulative distribution functions of the urbanicity indices, $\widehat{u}_i$, for all sectors in all available years. The plots demonstrate large differences between the sectors. The most rural sectors have an urbanicity index below 100 (employees per km^2), whereas the index exceeds 600 for the most urban sectors. Half of the sectors have an urbanicity index of about 250 or less in all years. None of the sectors has an urbanicity index larger than 710 in any year.

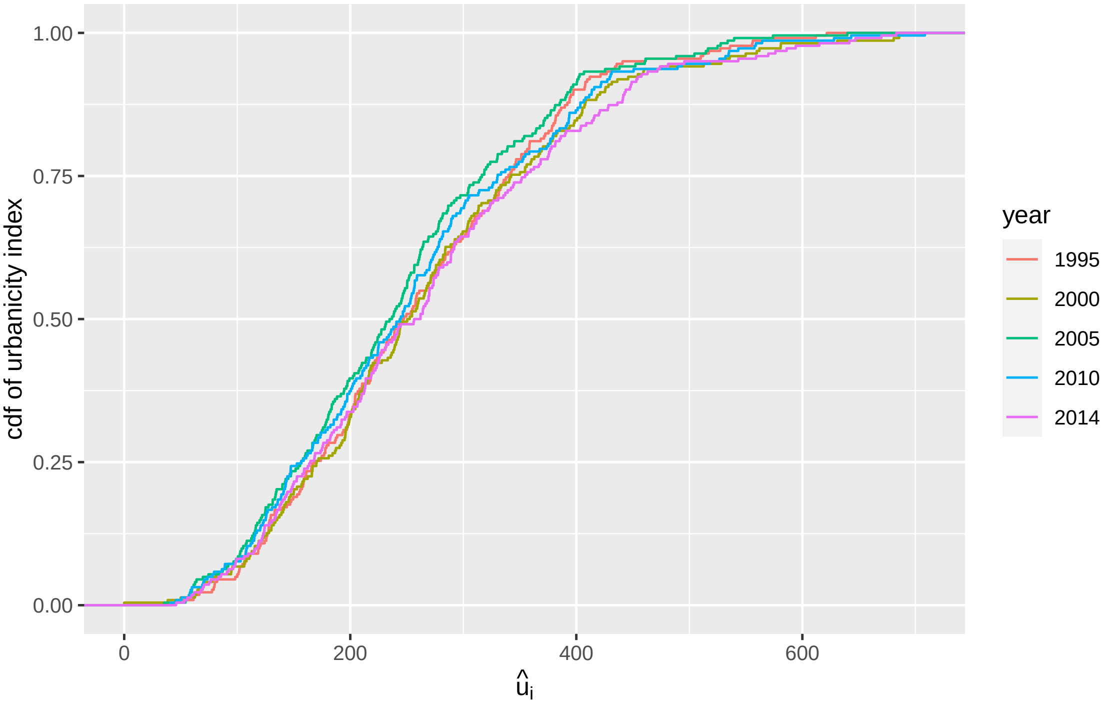

Figure 3: Cumulative distribution functions of the urbanicity indices $\widehat{u}_i$ of all sectors for all years.

How many sectors are significantly rural or urban, and which sectors are neither? We tested for each sector the null hypothesis that the employment in that sector has the same spatial distribution as total employment. Since the computation time is considerable, we restrict the analysis to the year 2014. The null hypothesis could be rejected for most sectors. At the 5 percent level, a share of only 22.5 percent was neither significantly rural nor urban. Among these borderline sectors are "primary education" (kindergartens) and "sale of motor vehicles". Further examples are "wholesale of food, beverages and tobacco", "human health activities", and "post and courier activities". Slightly more than half of all sectors are significantly rural (50.9%), among them – unsurprisingly – the agriculture and mining sectors.

The share of significantly urban sectors is 26.6%, for example restaurants, many consulting sectors, and higher education.[9]

We turn to the change in urbanicity over time. For each sector, we tested the null hypothesis that the urbanicity index did not change between the first and last observation years (that is between 1995 and 2014). Out of all sectors, only 17 (that is 7.7%) changed their urbanicity index significantly at the 5 percent level, 4 of them became more rural (for example, "farmings of animals" and "labour recruitment"), 13 more urban (for example, "hotels" and "transport via railways").[10]

Table 3(a) shows that within each year the cross-sectional standard deviation of the concentration factors $\widehat{a}_1, \ldots, \widehat{a}_I$ is relatively large (at roughly 1.8) and that it increased (non-monotonically) over time. An increase can also be observed for the correlation between the concentration factor and the employment share of the sectors. Larger sectors tend to be more concentrated than smaller ones. While this association is weak in 1995 (with a correlation coefficient of only 0.047), it has increased notably (to 0.124) until 2014.

Turning to Table 3(b) we find that the correlation between the sectoral urbanicity indices across different years is very high. Even over the entire 19-year horizon the correlation coefficient still exceeds 0.85. Hence, a sector with a high level of urbanicity tends to remain highly urban two decades later. Koh and Riedel (2014) also report a high level of persistence of agglomeration patterns in Germany.

4 The Mobility Components IME and SMS

4.1 Decomposition into IME and SMS

The urbanicity index of total employment, u_E, is defined as the product of the scale factor $E/|G|$ and the concentration factor a_E. Changes of the latter are driven by the intersectoral

[9] A table showing the urbanicity indices $\widehat{u}_i$ and coefficients of urbanicity $\widehat{U}_i$ for all sectors and all years is available as a csv file from the authors on request. The table is also included in the web appendix.

[10] For all sectors, the test statistics and p-values of both the test of urbanicity/rurality and the test of changes over time are available as a csv file on request from the authors. They are also listed in the web appendix.

	1995	2000	2005	2010	2014
(a) Sectors					
Std.dev($\widehat{a}_i$)	1.735	1.836	1.772	1.843	1.849
Corr($s_i, \widehat{a}_i$)	0.047	0.070	0.107	0.103	0.124
(b) Intertemporal correlations of $\widehat{u}_i$					
1995		0.966	0.936	0.879	0.893
2000			0.966	0.905	0.922
2005				0.944	0.971
2010					0.947

Table 3: Part (a): Standard deviation of concentration factors of all sectors, coefficients of correlation between employment shares s_i and concentration factors $\widehat{a}_i$. Part (b): Intertemporal correlations of urbanicity indices of all sectors.

mobility of employment, IME, and the spatial mobility of sectors, SMS. In the real world, such shifts in employment are likely to occur at the same time. Therefore, a meaningful measure of urbanicity should identify and quantify both shifts (Postulate 7). To this end, we suggest an additive decomposition of the change of the concentration factor, a_E, into the IME and the SMS.

The density function of employment in sector i is denoted by $f_i(x)$. The total employment density is the weighted sum of the sector densities,

$$f_E(x) = \sum_{i=1}^{I} s_i f_i(x)$$

where I is the number of sectors and $s_i = E_i/E$ is the employment share of sector i.

Since a_E is invariant with respect to the area unit, we can simplify the notation. Without loss of generality, we assume that the total area is normalized to $|G| = 1$. Let

$$a_E^{0,0} = \int_G f_E^0(x)^2 dx = \int_G \left(\sum_{i=1}^{I} s_i^0 f_i^0(x) \right)^2 dx$$

and

$$a_E^{t,t} = \int_G f_E^t(x)^2 dx = \int_G \left(\sum_{i=1}^{I} s_i^t f_i^t(x) \right)^2 dx$$

denote the concentration factors of total employment in some reference period 0 and some later period t. The change in the concentration factor is $\Delta a_E = a_E^{t,t} - a_E^{0,0}$. Further, define the counterfactual concentration factors

$$a_E^{t,0} = \int_G \left(\sum_{i=1}^{I} s_i^t f_i^0(x) \right)^2 dx \qquad \text{and} \qquad a_E^{0,t} = \int_G \left(\sum_{i=1}^{I} s_i^0 f_i^t(x) \right)^2 dx.$$

The first one is the concentration factor that would have occurred if the densities of period 0 had prevailed also in period t, but the employment shares had changed to s_i^t. The second one describes the opposite scenario, that is, it combines the employment shares of period 0 with the densities of period t.

Using the Bennet decomposition approach (Bennet, 1920, p. 457), the change in a_E can now be decomposed as

$$\Delta a_E = \Delta_{\text{IME}} + \Delta_{\text{SMS}} \ , \tag{10}$$

where

$$\Delta_{\text{IME}} = [(a_E^{t,0} - a_E^{0,0}) + (a_E^{t,t} - a_E^{0,t})]/2 \tag{11}$$

is the contribution of the intersectoral mobility of employment (IME) to the change in the concentration factor and

$$\Delta_{\text{SMS}} = [(a_E^{0,t} - a_E^{0,0}) + (a_E^{t,t} - a_E^{t,0})]/2 \tag{12}$$

is the contribution of the spatial mobility of sectors (SMS). The term $(a_E^{t,0} - a_E^{0,0})$ in expression (11) is the counterfactual change in a_E that would have occurred if the employment shares had shifted but the densities had remained as in period 0. The neighbouring term $(a_E^{t,t} - a_E^{0,t})$ has a perfectly analogous interpretation but with densities fixed at their period t values. The terms $(a_E^{t,t} - a_E^{t,0})$ and $(a_E^{0,t} - a_E^{0,0})$ in expression (12) measure the change in a_E due to the changing sector density functions, holding employment shares constant. Collecting terms, the two contributions can be rewritten as

$$\Delta a_E = \int_G \sum_i \sum_j \left(s_i^t s_j^t - s_i^0 s_j^0\right) \frac{f_i^t(x) f_j^t(x) + f_i^0(x) f_j^0(x)}{2} dx$$
$$+ \int_G \sum_i \sum_j \frac{s_i^t s_j^t + s_i^0 s_j^0}{2} \left[f_i^t(x) f_j^t(x) - f_i^0(x) f_j^0(x)\right] dx, \tag{13}$$

where the first integral is the term Δ_{IME} and the second integral is the term Δ_{SMS}.

4.2 IME and SMS in Germany

We apply the decomposition (13) to the German data. According to Table 2, the change in Germany's concentration factor between the years 1995 and 2014 is $\Delta\widehat{a}_E = \widehat{a}_E^{2014} - \widehat{a}_E^{1995} = 4.121 - 3.999 = 0.122$. Applying decomposition (13), this change can be split into the intersectoral mobility of employment, Δ_{IME}, and the spatial mobility of sectors, Δ_{SMS}:[11]

$$\Delta\widehat{a}_E = \Delta_{\text{IME}} + \Delta_{\text{SMS}} = 0.414 + (-0.292) = 0.122.$$

The value $\Delta_{\text{IME}} = 0.414$ can be interpreted as the increase in the concentration factor that would have occurred if the employment shares had shifted, but the densities of all sectors had remained constant. The positive value implies that Germany has experienced an employment shift towards more urban sectors. If, on the other hand, the densities had changed and the shares had remained constant, the concentration factor would have been reduced by $\Delta_{\text{SMS}} = -0.292$. This negative value indicates that the sectors shifted their employment towards more rural regions. In sum, the two forces driving the concentration factor offset each other to a large part. Observing only the relatively small net effect, one would overlook the substantial underlying shifts.

5 Sectoral Contributions

Postulate 8 demands a measure of urbanicity that quantifies the contributions of the I individual sectors to the overall change as well as to the changes of its components. For the

[11] To avoid overburdening the notation we simply write Δ_{IME} and Δ_{SMS} rather than the more precise $\widehat{\Delta}_{\text{IME}}$ and $\widehat{\Delta}_{\text{SMS}}$ when referring to estimated quantities.

urbanicity index, $u_E = (E/|G|)a_E$, this postulate implies that a sectoral decomposition of the changes of the scale factor and the two components IME and SMS of the concentration factor must be accomplished.

5.1 Sectoral Decomposition

We start with some randomly drawn sector and replace its period 0 number of employees by its period t number of employees, while the employment of all other sectors remains at its period 0 level. The new employment of the selected sector changes not only its own employment share and the employment share of all other sectors, but also its own employment density and the employment density of total employment. The new shares and densities yield a new value for the economy's concentration factor that we denote by a'_E, say. The difference $(a'_E - a_E^{0,0})$ is the selected sector's contribution to the total change, Δa_E. Applying the above Bennet decomposition, the sector's contribution can be split into the contribution to the sector's intersectoral mobility of employment (IME) and the sector's contribution to the spatial mobility of sectors (SMS).

Then, the same process is repeated for another randomly drawn sector. Its period 0 employment is replaced by its period t employment, while the employment of all other sectors remains at its current status, that is, at its period 0 employment except for the first selected sector which has already attained its period t employment. The difference between the new value of concentration, a''_E, say, and the previous value, a'_E, is the contribution of the second selected sector. Also this contribution can be split into the contributions to IME and SMS. This incremental process is repeated until all sectors have been selected and, therefore, have attained their period t employment.

More formally, let E_j^0 and E_j^t be the number of employees in sector j in periods 0 and t, respectively. Let $\sigma(i)$ denote the i-th element (sector) of some permutation σ of the sectors $1, \ldots, I$ and let s_j^i denote the counterfactual employment share of sector j if the number of employees of sectors $\sigma(1), \ldots, \sigma(i)$ are taken from period t while the number of employees of the remaining sectors are taken from period 0, that is, $E_{\sigma(1)}^t, \ldots, E_{\sigma(i)}^t, E_{\sigma(i+1)}^0, \ldots, E_{\sigma(I)}^0$. Correspondingly, the counterfactual employment share s_j^{i-1} is obtained when the number of employees of the sectors are $E_{\sigma(1)}^t, \ldots, E_{\sigma(i-1)}^t, E_{\sigma(i)}^0, \ldots, E_{\sigma(I)}^0$. In the same manner, we define the densities $f_j^{i-1}(x)$ and $f_j^i(x)$.

For given permutation σ, the change in the economy's concentration factor attributable to sector i is defined by

$$\Delta a_E^i = a_E^i - a_E^{i-1},$$

where

$$a_E^i = \int_G \left(\sum_j s_j^i f_j^i(x) \right)^2 dx \quad \text{and} \quad a_E^{i-1} = \int_G \left(\sum_j s_j^{i-1} f_j^{i-1}(x) \right)^2 dx. \tag{14}$$

Furthermore, we define the economy's counterfactual concentration factors

$$a_E^{i,i-1} = \int_G \left(\sum_j s_j^i f_j^{i-1}(x) \right)^2 dx \quad \text{and} \quad a_E^{i-1,i} = \int_G \left(\sum_j s_j^{i-1} f_i^i(x) \right)^2 dx. \tag{15}$$

In theory, each of the four concentration factors compiled in (14) and (15) should be averaged over all possible permutations σ. In practice, this is computationally infeasible if the number of sectors is large because the number of permutations is $I!$. Therefore, we propose to randomly draw 1000 permutations, say, and to average over them. Let the results be denoted by $\bar{a}_E^i$, $\bar{a}_E^{i-1}$, $\bar{a}_E^{i,i-1}$ and $\bar{a}_E^{i-1,i}$. These numbers can be used for the following Bennet decomposition of Δa_E^i:

$$\Delta a_E^i = \Delta_{\text{IME}}^i + \Delta_{\text{SMS}}^i,$$

where

$$\Delta_{\text{IME}}^i = [(\bar{a}_E^{i,i-1} - \bar{a}_E^{i-1}) + (\bar{a}_E^i - \bar{a}_E^{i-1,i})]/2 \tag{16}$$

$$\Delta_{\text{SMS}}^i = [(\bar{a}_E^i - \bar{a}_E^{i,i-1}) + (\bar{a}_E^{i-1,i} - \bar{a}_E^{i-1})]/2. \tag{17}$$

In analogy to the Bennet decomposition represented by expressions (11) and (12), expression (16) measures the contribution of the change of the employment share of sector i to Δa_E, while expression (17) measures the contribution of the change in the density distribution of sector i.

For each sector i ($i = 1, \ldots, I$), the value of expression (16) can be computed. Adding these I values yields the same result as expression (11). This equivalence says that the I values compiled by (16) represent the sectoral decomposition of the economy's measured intersectoral mobility of employees: $\sum_i \Delta_{\text{IME}}^i = \Delta_{\text{IME}}$. Analogously, summing over the I values compiled by (17) produces the same number as expression (12), because the I values (17) are the sectoral decomposition of the economy's measured change in the spatial mobility of its sectors: $\sum_i \Delta_{\text{SMS}}^i = \Delta_{\text{SMS}}$. Thus, the decomposition (10) can be further refined to the following decomposition of the country's concentration factor:

$$\Delta a_E = \Delta_{\text{IME}} + \Delta_{\text{SMS}} = \sum_i \left(\Delta_{\text{IME}}^i + \Delta_{\text{SMS}}^i\right). \tag{18}$$

The decomposition of the country's scale factor is straightforward as the area $|G|$ remains constant over time:

$$\frac{\Delta E}{|G|} = \frac{1}{|G|} \sum_i \Delta E_i, \tag{19}$$

with $\Delta E = E^t - E^0$ and $\Delta E_i = E_i^t - E_i^0$.

Applying the Bennet decomposition, the change in the urbanicity index, Δu_E, can be expressed in the form

$$\Delta u_E = \frac{\Delta E}{|G|} \frac{a_E^0 + a_E^t}{2} + \Delta a_E \frac{E^0 + E^t}{2|G|}.$$

The first term is the contribution of a change in E while the second term is the contribution of a change in a_E. Inserting (18) and (19) yields the sectoral contributions to Δu_E:

$$\Delta u_E = \sum_i \frac{1}{2|G|} \left[\Delta E_i \left(a_E^0 + a_E^t\right) + \left(E^0 + E^t\right) \left(\Delta_{\text{IME}}^i + \Delta_{\text{SMS}}^i\right)\right]. \tag{20}$$

For two reasons, statistical inference of the decompositions (18) and (20) is problematic. First, the bootstrap approach requires repeated computations of the decomposition a large number of times. Computing the Bennet decomposition of each repetition is time-consuming for it is calculated by averaging over a large number of permutations, as explained above. Hence, in practice the bootstrap method is not implementable with reasonable computing resources. Second, due to the additivity of the decompositions (18) and (20), looking at the marginal distribution of the contribution of a single sector would not be informative. Instead, it is imperative to consider the joint distribution of all contributions simultaneously. Even though the bootstrap approach would deliver the joint distribution in a natural way, there is no transparent and easily comprehensible way to report or visualise the joint distribution. Therefore, we only report point estimates for the sectoral decomposition.

5.2 Sectoral Decomposition for Germany

How much did each sector contribute to the overall change in urbanicity in Germany? To answer this question we decompose the changes of the concentration factor a_E according to (18). The second equation in expression (18) decomposes the values of Δ_{IME} and Δ_{SMS} into the sectoral contributions Δ^i_{IME} and Δ^i_{SMS}, respectively. These sectoral contributions are computed by expressions (16) and (17).

Figure 4 visualizes the findings. Each point represents one sector. The point sizes represent the employment shares of the sectors in 1995, while the colours indicate their coefficient of urbanicity $\widehat{U}_i$ in 1995. The colour scale ranges from green (very rural sector) to red (very urban sector). For example, the sector "labour recruitment" was distinctly urban in 1995, while the sector "civil engineering" was distinctly rural.

The ordinate depicts the sector's spatial mobility between 1995 and 2014, Δ^i_{SMS}. The majority of the points is below the abscissa indicating that, on average, the sectors became more rural, confirming the previous finding $\Delta_{\text{SMS}} = -0.292$. The negative values of Δ^i_{SMS} contributed to the overall negative impact of Δ_{SMS} and, therefore, to the fall in $\widehat{a}_E$.

The abscissa indicates the sector's intersectoral mobility of employment between 1995 and 2014, Δ^i_{IME}. This value measures each sector's contribution to Δ_{IME}. Therefore, all points to the right of the ordinate represent sectors that increased the value of the concentration factor $\widehat{a}_E$ via their positive contribution to Δ_{IME}.

A high proportion (about 45 percent) of sectors is located in the bottom right quadrant: $\Delta^i_{\text{IME}} > 0$ and $\Delta^i_{\text{SMS}} < 0$. The changing employment shares of these sectors had a positive impact on the concentration factor, whereas the changing densities of those same sectors had a negative impact. It should be noted that the sign of Δ_{IME} alone does not indicate an increase or decrease of the employment share. Whether a growing employment share results in $\Delta^i_{\text{IME}} > 0$ or $\Delta^i_{\text{IME}} < 0$ depends on the urbanicity of the sector. For example, the employment share of the rural sector "civil engineering" declined, whereas the employment share of the more urban sector "labour recruitment" increased.[12] Both shifts resulted in

[12] The full names of the example sectors are: "building of complete constructions or parts thereof; civil engineering" (short: *civil engineering*), "monetary intermediation" (short: *monetary intermediation*), "legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion

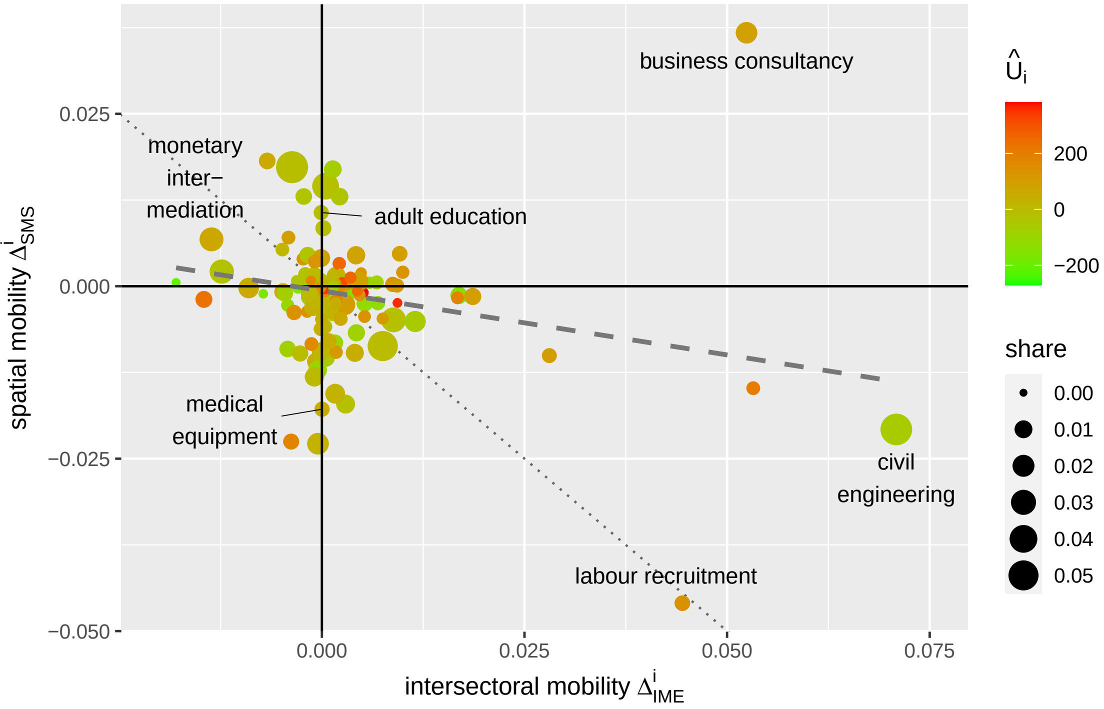

Figure 4: Decomposition of the changes in the sectoral concentration factors between 1995 and 2014 into the sectors' changing employment shares (intersectoral mobility, Δ^i_{IME}) and densities (spatial mobility, Δ^i_{SMS}). Point sizes reflect the employment shares of the sectors in 1995, point colours (from green = very rural to red = very urban) reflect the coefficient of urbanicity $\widehat{U}_i$ in 1995. The dashed line is fitted by a linear regression of Δ^i_{SMS} on Δ^i_{IME} weighted by employment shares. The dotted line has slope −1 and intercept 0.

$\Delta^i_{\text{IME}} > 0$, contributing to an increase of the concentration factor, $\widehat{a}_E$. At the same time, both sectors became more rural. Thus, $\Delta^i_{\text{SMS}} < 0$, contributing to a fall in $\widehat{a}_E$. For the sector "labour recruitment" this offsetting effect was so large that the overall effect on $\widehat{a}_E$ was close to zero.

Such almost completely offsetting effects apply to all sectors located close to the dotted line in Figure 4 (e.g., the sector "monetary intermediation"). The sector "civil engineering"

polling; business and management consultancy; holdings" (short: *business consultancy*); "labour recruitment and provision of personnel" (short: *labour recruitment*), "Manufacture of medical and surgical equipment and orthopaedic appliances" (short: *medical equipment*), "adult and other education" (short: *adult education*).

is clearly located above this line. This indicates a positive overall effect of this sector on the concentration factor, $\widehat{a}_E$. The sector “business consultancy” shows the strongest positive overall effect.

The spread of the points in Figure 4 indicates that the values of Δ_{IME} and Δ_{SMS} are not driven by very few influential sectors, but by many small contributions of a wide range of sectors. The dominant role of the sectors in the lower right quadrant is confirmed by the dashed regression line. It regresses Δ^i_{SMS} on Δ^i_{IME} weighted by employment shares.

The value $\Delta_{\text{IME}} = 0.414$ says that an overall employment shift towards more urban sectors occurred. At the same time, the value $\Delta_{\text{SMS}} = -0.292$ implies that, on average, the sectors became more rural. Since the correlation between $\widehat{u}_i$ (in 1995) and Δ^i_{SMS} is merely -0.072, the shift towards rurality is not limited to urban sectors but also applies to rural sectors. However, this does not contradict the conjecture that, among the the urban sectors, it was the subgroup of *expanding* sectors that strengthened their rural locations to absorb the new employees from the rural sectors.

To examine this conjecture, we compute the correlation between the change in employment shares, $(s_i^{2014} - s_i^{1995})$, and Δ^i_{SMS} for the subgroup of urban sectors ($U_i > 0$ in 1995). The correlation is -0.280. The larger the employment gain of an urban sector, the stronger its shift towards rural regions. We repeat the same exercise for the group of rural sectors ($U_i \leq 0$ in 1995). The correlation between $(s_i^{2014} - s_i^{1995})$ and Δ^i_{SMS} is 0.424. The latter correlation implies that, on average, the (few) expanding rural sectors became more urban and the (many) shrinking rural sectors became more rural, that is, kept their particularly rural production sites and closed those in less rural locations. Overall, we find the conjecture confirmed. Rural sectors lost employees in their more urban production sites and urban sectors gained employees in their more rural production sites. Quite likely, many employees switched to more urban sectors without changing their location of employment.

6 Concluding Remarks

The urbanicity index of employment is a powerful statistical measure of the degree of urbanization. It is density-based and, therefore, avoids the modifiable area unit problem. Furthermore, it distinguishes between the scale aspect and the two components of the concentration aspect (intersectoral mobility of employment and spatial mobility of sectors). Finally, it can consistently factorize the overall numbers into the contributions of the individual sectors of the economy. As a result, the urbanicity index can detect urbanization trends that simpler measures would fail to notice.

In the empirical application, this paper finds that strong urbanization trends occurred in Germany between 1995 and 2014. Employment shifted from rural sectors to more productive urban sectors and many of the growing urban sectors absorbed the new employees by expanding their rural production sites. Thus, many employees were able to switch from rural to urban sectors without changing their location of employment.

The applicability of the urbanicity index is not restricted to economic questions. For example, in demography, geography, political science, sociology, social medicine, or history, employment data on economic sectors could be replaced by population data that distinguish between different population groups. In biology or zoology, the urbanicity index could be used to study the spatial pattern of different species.

Appendix

We demonstrate within a unifying framework how distance-based tools of concentration measurement, in particular Ripley's $K(d)$ and Duranton-Overman's K_d functions, are related to our approach. To simplify the notation we ignore weighting and put equal mass of $1/n$ on each of the $n = |C|$ companies. All concentration measures can eventually be traced back to the integral

$$\int_G \widehat{f}_E(x) dH(x), \tag{21}$$

where G is the area under consideration, $\widehat{f}_E(x)$ the estimated spatial density of employment evaluated at location x and $H(x)$ the cumulative distribution function (cdf) of the estimated spatial distribution. Both, distance-based and density-based concentration measures can be derived from (21) by choosing $\widehat{f}_E(x)$ and $H(x)$ in a suitable way.

We begin with our density-based concentration factor a_E. As (1) shows, a_E is based on the (unobservable) spatial density $f_E(x)$ of overall employment. The spatial density $f_E(x)$

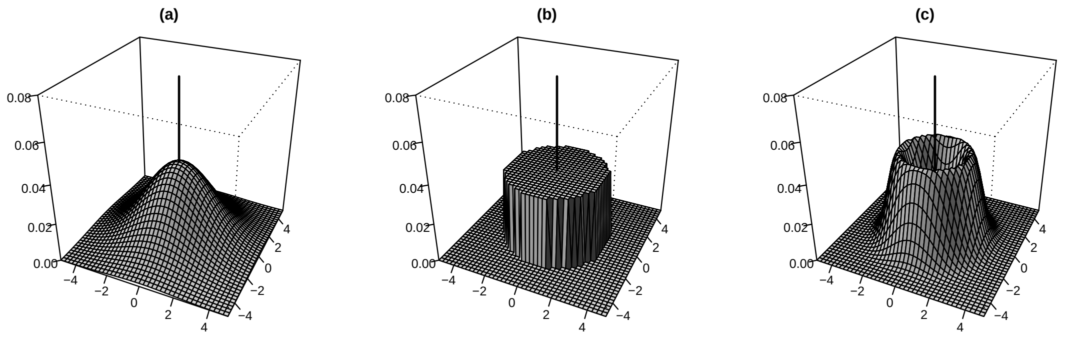

Figure 5: (a) Gaussian product kernel density with bandwidth $h = 2$, (b) density of radial symmetric uniform kernel with distance $h = 2$, (c) density of radial symmetric Gauss kernel with distance $d = 2$ and bandwidth $h = 0.7$.

is estimated by kernel density estimation with a Gaussian product kernel and bandwidth h, resulting in the estimated spatial density function $\widehat{f}_E(x)$ defined in (3). Figure 5(a) illustrates how in this estimation the density belonging to a company is "smeared" by the Gaussian product kernel in the neighbourhood of the company. One obtains for each point of G the company's contribution to the density. Of course, points far away from the location of the company only receive a negligible contribution. Conducting this process for each company and then adding at each point of G the density contributions of all firms yields for each point of G the estimated density of overall employment. This estimated density is embodied in the function $\widehat{f}_E(x)$. The integral in (1) can be written in the form (21) if we set $H(x) = \widehat{F}_E(x)$, that is, the cdf belonging to the estimated density $\widehat{f}_E(x)$,

$$\int_G \widehat{f}_E(x) d\widehat{F}_E(x) = \int_G \widehat{f}_E(x)\widehat{f}_E(x) dx.$$

The estimated concentration factor $\widehat{a}_E$ is just a rescaling of this integral.

We now show that distance-based tools of concentration measurement, in particular Ripley's $K(d)$ and Duranton-Overman's K_d functions, are also special cases of (21). To this end, we define the spatial empirical distribution function at location $x = (x_1, x_2)$ as $F_n(x) = (1/n)\sum_{i=1}^{n} 1(x_{1c} \leq x_1 \text{ and } x_{2c} \leq x_2)$, where x_{1c}, x_{2c} are the geo-coordinates of company c and $1(\cdot)$ is an indicator function, that is, $1(A) = 1$ if A is true and $1(A) = 0$ otherwise. When the number of firms is large, the distribution function $F_n(x)$ with its stepwise ascent can be hardly distinguished from the estimated distribution function $\widehat{F}_E(x)$ with its smooth ascent. In fact, according to the Glivenko-Cantelli theorem (van der Vaart, 1998, chap. 19.1) the maximal distance between $F_n(x)$ and $\widehat{F}_E(x)$ converges to 0 as the number of firms tends to infinity. For a given distance d, both Ripley's $K(d)$ and Duranton-Overman's K_d functions can be written in the form (21) if we replace $H(x)$ by $F_n(x)$:

$$\int_G \widehat{f}_E(x) dF_n(x), \tag{22}$$

Since the empirical distribution $F_n(x)$ puts point mass $1/n$ on each spatial observation, the estimated density $\widehat{f}_E(x)$ in the integral (22) corresponds to $1/n$ in the discrete case. Therefore, the integral (22) can be rewritten as a mean (or weighted mean if the number of employees differs between companies),

$$\int_G \widehat{f}_E(x) dF_n(x) = \sum_{i=1}^{n} \frac{1}{n} \widehat{f}_E(x_i).$$

First, consider Ripley's $K(d)$ evaluated at distance d (see e.g. Baddeley, Rubak, & Turner, 2016, chap. 7.3). In the kernel density estimator (3), we substitute the Gaussian product kernel by a radial symmetric uniform kernel with bandwidth (radius) $h = d$,

$$K(d)(x, x_c) = \frac{1}{\pi d^2} \, 1((x_1 - x_{c1})^2 + (x_2 - x_{c2})^2 < d^2),$$

with $1(\cdot)$ again representing the indicator function. Figure 5(b) depicts this kernel function for distance $d = 2$. When the number of companies is very large, the choice between the uniform kernel and the Gaussian kernel hardly makes a difference for the estimation of f_E by (3). Regardless of the applied kernel function, estimating f_E by (3) and calculating (22) returns the average number of neighbouring companies (per unit area) where Ripley's approach reckons two companies as neighbours if their distance is less than d. Thus, apart from edge-correction and scaling, the computation of Ripley's distance-based $K(d)$-value at a given distance d differs from the computaton of our density-based concentration measure $\widehat{a}_E$ only in two respects. When computing the integral (21), the $K(d)$-approach uses F_n and a radial symmetric uniform kernel rather than $\widehat{F}_E$ and a Gaussian product kernel. The value of Ripley's $K(d)$ function is usually calculated for a range of distances d. In our own density approach, the formal analogue would be to compute $\widehat{a}_E$ for a range of bandwidth values h. Though this is not difficult to do, it would not be particularly enlightening.

Second, we turn to the K_d function of Duranton and Overman (2005) evaluated at a fixed distance d with bandwidth h. As a radial symmetric Gauss kernel consider

$$K_h(x, x_c) = \frac{1}{2\pi h d}\phi\left(\frac{\sqrt{(x_1 - x_{c1})^2 + (x_2 - x_{c2})^2} - d}{h}\right). \tag{23}$$

Figure 5(c) depicts this kernel function for distance $d = 2$ and bandwidth $h = 0.7$. It is shaped like a ringfort (or an eggcup). Estimating the spatial density f_E using the kernel (23) and computing the integral (22) yields the (rescaled) value of the K_d function of Duranton and Overman (2005) at distance d with bandwidth h.

To summarise, distance-based approches are closely related to the density-based approach. Both approaches construct concentration measures by rescaling the integral (21). There are two differences. First, for a given distance, the common distance-based measures are computed using the (discrete) empirical spatial distribution whereas the density-based approach is based on the (continuous) kernel estimated spatial density. Second, the choice of the kernel function differs. While distance-based methods work with radial symmetric kernels of rather peculiar shapes, the density-based approach uses a Gaussian product kernel. Obviously, density-based measures could also be defined for any other kind of kernel function. In this sense, density-based measures can be regarded as more general.

References

Arbia, G. (1989). *Spatial data configuration in statistical analysis of regional economic and related problems.* Dordrecht: Springer. doi: 10.1007/978-94-009-2395-9

Arbia, G. (2001). The role of spatial effects in the empirical analysis of regional concentration. *Journal of Geographical Systems*, *3*, 271-281. doi: 10.1007/PL00011480

Arriaga, E. E. (1970). A new approach to the measurements of urbanization. *Economic Development and Cultural Change*, *18*(2), 206-218. doi: 10.1086/450419

Auer, L. v., Stepanyan, A., & Trede, M. (2019). Classifying industries into types of relative concentration. *Journal of the Royal Statistical Society, Series A*, *182*(3), 1017-1037. doi: 10.1111/rssa.12441

Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial point patterns: Methodology and applications with R*. Boca Raton: CRC Press.

Bennet, T. (1920). The theory of measurement of changes in cost of living. *Journal of the Royal Statistical Society*, *83*(3), 455-462. doi: 10.1111/j.2397-2335.1920.tb00631.x

Briant, A., Combes, P.-P., & Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, *67*, 287-302. doi: 10.1016/j.jue.2009.09.014

Brülhart, M., & Sbergami, F. (2009). Agglomeration and growth: Cross-country evidence. *Journal of Urban Economics*, *65*, 48-63. doi: 10.1016/j.jue.2008.08.003

Castells-Quintana, D., & Royuela, V. (2014). Agglomeration, inequality and economic growth. *Annals of Regional Science*, *52*, 343-366. doi: 10.1007/s00168-014-0589-1

Duranton, G., & Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, *72*(4), 1077-1106. doi: 10.1111/0034-6527.00362

Ellison, G., & Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy*, *105*(5), 889-927. doi: 10.1086/262098

Frick, S. A., & Rodríguez-Pose. (2018). Change in urban concentration and economic growth. *World Development*, *105*, 156-170. doi: 10.1016/j.worlddev.2017.12.034

Ganau, R., & Rodríguez-Pose, A. (2021). Does urban concentration matter for changes in country economic performance? *Urban Studies*. doi: 10.1177/0042098021998927

Gramacki, A. (2018). *Nonparametric kernel density estimation and its computational aspects*. Cham: Springer.

Henderson, V. (2003). The urbanization process and economic growth: The so-what question. *Journal of Economic Growth*, *8*, 47-71. doi: 10.1023/A:1022860800744

Jedwab, R., & Vollrath, D. (2015). Urbanization without growth in historical perspective. *Explorations in Economic History*, *58*, 1-21. doi: 10.1016/j.eeh.2015.09.002

Koh, H.-J., & Riedel, N. (2014). Assessing the localization pattern of German manufacturing and service industries: A distance-based approach. *Regional Studies*, *48*(5), 823-843. doi: 10.1080/00343404.2012.677024

Lang, G., Marcon, E., & Puech, F. (2020). Distance-based measures of spatial concentration:

introducing a relative density function. *Annals of Regional Science*, *64*, 243-265. doi: 10.1007/s00168-019-00946-7

Lemelin, A., Rubiera-Morollón, F., & Gómez-Loscos, A. (2016). Measuring urban agglomeration: A refoundation of the mean city-population size index. *Social Indicators Research*, *125*, 589-612. doi: 10.1007/s11205-014-0846-9

Marcon, E., & Puech, F. (2003). Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, *3*, 409-428. doi: 10.1093/jeg/lbg016

Marcon, E., & Puech, F. (2010). Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, *10*, 745-762. doi: doi:10.1093/jeg/lbp056

Marcon, E., & Puech, F. (2017). A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, *62*, 56-67. doi: 10.1016/j.regsciurbeco.2016.10.004

Nakamura, R., & Morrison Paul, C. J. (2019). Handbook of regional growth and development theories. In R. Capello & P. Nijkamp (Eds.), (p. 386-412). Edward Elgar.

Openshaw, S., & Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical applications in the spatial sciences* (p. 127-144). London: Pion.

Schmucker, A., Seth, S., Ludsteck, J., Eberle, J., & Ganzer, A. (2016). *Establishment history panel 1975-2014* (FDZ Datenreport. Documentation on Labour Market Data No. 201603). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung (IAB).

Silverman, B. (1986). *Density estimation for statistics and data analysis.* London: Chapman and Hall.

Statistisches Bundesamt. (2008). *Klassifikation der Wirtschaftszweige.* Retrieved from `https://www.destatis.de/static/DE/dokumente/klassifikation-wz-2008-3100100089004.pdf`

Stewart Jr., C. T. (1958). The urban-rural dichotomy: Concepts and uses. *American Journal of Sociology*, *64*(2), 152-158. doi: 10.1086/222422

van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge UK: Cambridge University

Press.

World Bank. (2009). *Reshaping economic geography.* World Development Report. doi: 10.1596/978-0-8213-7640-9