von Auer, Ludwig; Trede, Mark

Conference Paper

Decomposing the Urbanization of Employment: A New Measure

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2025: Revival of Industrial Policy

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: von Auer, Ludwig; Trede, Mark (2025) : Decomposing the Urbanization of Employment: A New Measure, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2025: Revival of Industrial Policy, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/325426

Decomposing the Urbanization of Employment: A New Measure*

Ludwig von Auer† Universität Trier

Mark Trede‡ Universität Münster

Februrary 2025

Abstract: Using sectoral spatial employment data from 1995 to 2014, this study identifies seemingly contradictory employment trends in Germany. While urban sectors (many of them service sectors) substantially expanded at the expense of rural ones, the overall urbanization level of employment increased only moderately. To address this "urbanization puzzle", the paper introduces a new density-based urbanization index. In contrast to existing measures, it computes the concentration and size of employment for each sector and aggregates these numbers additively into the degree of urbanization. This sectoral spatial analysis unravels the urbanization puzzle.

Keywords: agglomeration, concentration, index, structural change.

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 Statistische Woche 2022 in Münster as well as at a staff seminar at Universität Würzburg. Helpful comments and suggestions from participants are gratefully acknowledged. The usual disclaimer applies.

†Corresponding author: Universität Trier, Fachbereich IV, Volkswirtschaftslehre, Universitätsring 15, 54296 Trier, Germany, vonauer@uni-trier.de

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

1 Introduction

Over the past decades, industrialized countries have witnessed transformative shifts in their employment structure. Between 1970 and 2017, the employment share of services increased by 29 percentage points in the EU14, 27 percentage points in the UK, 27 percentage points in Japan, and 14 percentage points in the US (Duernecker and Sanchez-Martinez, 2023, Table 1). Germany was no exception (e.g., Dauth, Findeisen and Suedekum, 2017, pp. 338-339). The sectoral employment data used in the present study reveal that the employment share of services increased by 10.3 percentage points between 1995 and 2014, while the employment shares of manufacturing and agriculture & mining decreased by 10.2 and 0.1, respectively. The same employment data show that services are considerably more urban than manufacturing and agriculture & mining.[1]

Since urban sectors (i.e., services) have expanded and more rural sectors (i.e., manufacturing and agriculture & mining) have shrunk, one might expect that the overall degree of urbanization of German employment has substantially increased. However, the data tell a fundamentally different story: between 1995 and 2014, the overall degree of urbanization remained remarkably stable.[2] The present study attempts to solve this "urbanization puzzle".

To separate sectoral shifts from spatial developments, one needs spatial sectoral employment data and suitable statistical measurement methods. The paper's first contribution is a detailed discussion of the properties such a measure should possess.

None of the existing measures satisfies all of these properties. Therefore, the second contribution is methodological. This paper introduces a novel measure – the economic *urbanization index* – that satisfies all desirable properties. The measure is density-based. A

[1] In 2014, the expected number of employees (of all sectors) per unit area as perceived by a randomly drawn service sector employee was 331. For manufacturing and for agriculture & mining, the numbers are 233 and 137, respectively.

[2] The expected number of employees per unit area as perceived by a randomly drawn employee fluctuated around 283 without showing a stable upward trend (see Table 2 on p. 19).

kernel density estimation transforms the observed spatial distribution of employees into a smooth surface with peaks in urban areas and lowlands in rural areas. The surface provides an immediate and accurate visual impression of the economy's spatial employment pattern. The same process is applied individually to each sector, generating density estimations that become the basis for a rigorous statistical analysis. This analysis allows for statistical inference and an insightful decomposition of the overall urbanization trend.

The third contribution is empirical. We investigate the urbanization trends of employment in Germany using administrative micro-data on regional and sectoral employment from 1995, 2000, 2005, 2010, and 2014. The urbanization index does not show a distinct upward trend between 1995 and 2014.[3] However, an in-depth decomposition of the urbanization index reveals a nuanced narrative that existing measures could not provide. Behind the apparent stagnation lies a story of robust urbanization trends, providing valuable insights to disentangle the urbanization puzzle.

These insights prompt a list of interesting follow-up questions. Which sectors can be considered urban, and which rural? Which urban sectors received the employees that left the rural sectors? Do a few influential sectors drive the aggregated numbers, or do the changes occur across the board? Answering these follow-up questions marks the fourth contribution of this study.

The rest of the paper proceeds as follows. Section 2 delves into measurement issues, highlighting essential properties that urbanization measures need to disentangle sectoral and spatial employment shifts. This section also includes a review of the related literature. Section 3 defines and explains the urbanization index. We demonstrate how to compute the index using spatial employment data and perform hypothesis tests for changes in the index over time. As an empirical contribution, we calculate the urbanization index for Germany

[3]This is in line with urbanization trends in other industrialized countries. In their comprehensive study, Jedwab and Vollrath (2015) report that since the mid-20th century, the urbanization rates (share of the population living in "urban areas") of industrialized countries have been rather stagnant. In Germany, the urbanization rate increased from 73.9 in 1995 to 77.2 in 2014.

using administrative employment data, and test for intertemporal changes in the index. Section 4 introduces a measure of the degree of urbanization for individual economic sectors and a straightforward statistical inference procedure. The measure and the associated inference are applied to all sectors of the German economy. Section 5 shows that an intertemporal change in the urbanization index can be factorized into its scale and concentration effects. 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 6 explains how one can further decompose the two mobility aspects into the contributions of the individual sectors of the economy. Additional findings about the sectors' mobility in Germany are also presented. They help to resolve the urbanization puzzle. Section 7 concludes.

2 Measurement Issues and Related Literature

Explaining the urbanization puzzle requires spatial sectoral employment data and a suitable statistical approach to analyse them. We list the desirable properties of such an approach and relate them to the existing literature. Since none of the existing urbanization measures satisfies all properties, we will introduce (in Section 3) a new approach, the *urbanization index*.

2.1 Scale and Concentration

There is much overlap between measures of an economy's degree of urbanization and measures of an economy's degree of concentration because both types are concerned with the spatial distribution of employment.[4] For example, rural-urban migration should raise the measured degree of both concentration and urbanization. However, the overlap is only par-

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

tial because urbanization measures should also account for a scale aspect.[5] For example, quadrupling a country's employment 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. Simple traditional measures of the degree of urbanization usually take account of the concentration and the scale aspect but cannot distinguish between the two. The most prominent example is the urbanization rate (the share of the country's population in "urban areas").

2.2 Spatial Issues

The urbanization rate is a "discrete" measure in the sense that it analyzes regionalized data, which are obtained by fragmenting the country's territory into well-defined regions. Identifying adequate regions can be challenging (e.g., Briant, Combes and Lafourcade, 2010, pp. 288-289), and comparisons across country boundaries further aggravate the difficulties. This issue is known as the "modifiable area unit problem" (e.g., Arbia, 1989; Openshaw and Taylor, 1979).

Furthermore, the analysis of regionalized data often ignores the spatial arrangement of the regions and, instead, considers the regions as independent atomistic units. This may cause a loss of valuable spatial information.

Another spatial issue is the categorization of an economy's locations. For example, the measured urbanization rates depend on the delineation of urban and rural areas. This "urban-rural dichotomy" (e.g., Stewart Jr., 1958) can cause reservations about the robustness

[5]Measures of *urban concentration* ignore the scale aspect, 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 the total urban population (Hirschman-Herfindahl index).

of the empirical findings.[6] Introducing additional categories (e.g., "moderately urban" and "distinctly urban") mitigates the problem but does not solve it.

A closely related concern is the comparability of the results across national borders. For example, Roberts, Blankespoor, Deuskar and Stewart (2017) point out that, relative to their GPD per capita, Latin American countries appear to have much higher degrees of urbanization than comparable countries in other regions of the world. The authors argue that this observation is a statistical artifact caused by the differences in the categorization of areas.[7]

The listed spatial issues (modifiable are unit problem, atomistic regions, categorization of locations, comparisons across national borders) are relevant not only for measures of economic urbanization but also for other fields in urban economics.[8] Solutions proposed in these other fields could also be relevant in the present context. Duranton (2021) explains that the solutions involve many compromises and tradeoffs and that significant efforts have been made to reduce arbitrariness in the necessary decisions.

An early solution to the categorization issue has been proposed by the demographer Arriaga (1970, p. 209). He advocates an index which does without any categorization of locations. Instead, it measures the average employment of the economy's locations as perceived by its employees. For example, an index value of 1000 would say that, on average, each

[6]For example, the studies reviewed in Galdo, Li and Rama (2021) report India's urbanization rates ranging from 14.8 to 78.0 percent.

[7]In response to such problems, in March 2020, the United Nations (UN) Statistical Commission endorsed the *Degree of Urbanization*, which is not a measure of an economy's degree of urbanization, but a collection of rules that assign a country's regions to three different types: 1) cities, 2) towns and semi-dense areas, and 3) rural areas; for further discussions, see Dijkstra, Florczyk, Freire, Kemper, Melchiorri, Pesaresi and Schiavina (2021).

[8]Important examples are the identification of urban areas and the analysis of their internal structure. Both require a plausible categorization of locations and a proper delineation of space. The *Journal of Urban Economics* dedicated a complete special issue to these two problems with an introductory survey by Duranton (2021).

employee in the economy can expect to have 999 other employees in her location. One can interpret the Arriaga index as an early application of the concept of "experienced density" (Duranton and Puga, 2020, p. 5). Economists became aware of the Arriaga index to the credit of Lemelin, Rubiera-Morollón and Gómez-Loscos (2016, pp. 594-595). The authors point out that the Arriaga index can be expressed as the product of the economy's total employment and the Hirschman-Herfindahl index of the regional employment shares. The former accounts for the scale aspect of urbanization, while the latter captures the concentration aspect (see Section 2.1). More specifically, the Arriaga index is the product of a *scale factor* (size of employment) and a *concentration factor* (Hirschman-Herfindahl index).

For some research questions, the categorization of an economy's locations seems inevitable. For example, when analyzing the internal structure of cities, researchers distinguish between a city's core(s), more peripheral city areas, and areas that do not belong to the city. Meaningful categorizations require simple and transparent thresholds. Such thresholds are applied in Henderson, Nigmatulina and Kriticos (2021). They refer to the Arriaga index as "personal population density" and apply it to population grid cell data within urban areas.[9] Furthermore, they discuss a generalized version of the Arriaga index proposed by de la Roca and Puga (2017, p. 112) who calculate the average "experienced density" as perceived by the population within the urban area.[10] Henderson et al. (2021) propose a refined version involving a spatial discount factor such that the population in the outer grid cells of the neighborhood receive less weight than the population in the inner grid cells.[11]

[9]Instead of the grid cells of a given urban area, Arriaga (1970, pp. 209-213) includes all municipalities within the country in his original index. He argues that the summation terms relating to the rural municipalities can be omitted without affecting the index value.

[10]These authors trace a circle of radius 10 km around each person, count for each circle the population living within the circle, and compute the average of these numbers. This process is repeated for every grid cell within the urban area. Finally, the results are aggregated by a weighted arithmetic average where the weights are given by the grid cells' shares of the urban area's total population. This approach yields the average population as perceived by the population within the urban area.

[11]Combining such a discounting with a circular neighborhood would yield an index concept similar to the

Various innovative solutions have been proposed to assign population or employment to the correct locations. For example, in countries like Spain and France, high-resolution remote sensing data are available that show all buildings within a region (see e.g. de Bellefon, Combes, Duranton, Gobillond and Gorin, 2021). The location and shape of the buildings help to assign the population of a given region to smaller grid cells that can be aggregated to an "urban area", say.[12] Other approaches involve information on nighttime satellite images, cell phone data, commuting flow data and machine learning.[13]

2.3 Statistical Inference

Since the seminal work of Ellison and Glaeser (1997), it has been considered indispensable for reliable measures of economic concentration that they allow for statistical inference. Usually, this involves some bootstrapping approach.

2.4 Issues of Decomposition

It would be advantageous if an urbanization measure could quantify and rank individual economic sectors' degrees of urbanization. Obviously, the computation of such a measure requires sectoral employment data. With such data, it is seductive to simply measure each sector's degree of urbanization by its degree of concentration. However, this approach would be inappropriate. As Auer, Stepanyan and Trede (2019) emphasize, many concentrated sectors are distinctly rural.

---

empirical implementation we propose for our urbanization index; for details, see Section 3.

[12]Recent studies that apply this approach include Arribas-Bel, Garcia-Lopez and Viladecans-Marsal (2021) and Baragwanath, Goldblatt, Hanson and Khandelwal (2021).

[13]Recent studies that rely on nighttime light include Baragwanath et al. (2021), Ch, Martin and Vargas (2021), Dingel, Miscio and Davis (2021) and Harari (2020). Studies that use cell phone data include Louail, Lenormand, Cantu Ros, Picornell, Herranz, Frias-Martinez, Ramasco and Barthelemy (2014), Büchel and Ehrlich (2020). A recent paper that analyzes commuting flow data is Bosker, Park and Roberts (2021). Galdo et al. (2021) use a combination of human judgment and machine learning to identify urban areas.

In Section 2.1 we advocated measures of economic urbanization that distinguish between the scale and the concentration aspect of urbanization. The concentration aspect can be further decomposed into two principal forces: *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. These two mobility aspects (intersectoral mobility of employment and spatial mobility of sectors) have opposing effects on the degree of the economy's concentration and, therefore, urbanization. Thus, a detailed analysis of urbanization trends in employment would benefit from a measure that decomposes changes in concentration into the two mobility aspects and further into the contributions of the individual sectors.

2.5 Spatial Point Pattern Analysis

In more and more countries, researchers are granted access to geocoded sectoral employment data. The spatial information in such data can be analyzed by spatial point pattern analysis. The field of spatial statistics includes quadrat count methods, distance-based methods, and density-based methods.

Two classes of distance-based methods dominate the spatial analysis of economic *concentration*. One approach, proposed by Duranton and Overman (2005), involves computing the distribution of pairwise distances between observed firms, which is subsequently smoothed using kernel density estimation, yielding the K_d function. Statistical inference is conducted by comparing the K_d function with the smoothed distribution of pairwise distances under the complete spatial randomness hypothesis. The other popular class revolves around Ripley's $K(d)$-function (e.g., Marcon and Puech, 2003,1).[14] This function shows for every radius d the average number of other firms covered by circles of radius d drawn around the observed firms. For statistical inference, the $K(d)$ function can be compared to the corresponding

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

function that would arise under the complete spatial randomness hypothesis.

The K_d and $K(d)$ functions satisfy all of the desirable properties discussed in Sections 2.2 and 2.3. In addition, they allow for comparisons between economic sectors. However, concentration measures are not designed to distinguish between the concentration aspect and the scale aspect of urbanization (see Section 2.1). Furthermore, the K_d and $K(d)$ functions are less suitable for the decompositions outlined in Section 2.4. Finally, distance-based methods transform the observed point pattern into a pattern of observed pairwise distances d. The K_d and $K(d)$ functions provide summary statistics for the degree of concentration and represent the pattern of distances in a condensed form. However, such forms give a rather vague impression of the underlying point pattern from which the distances were derived.

2.6 Implications for the Present Study

Since we strive for a detailed analysis of an economy's spatial and sectoral urbanization trends, we explore a particularly direct utilization of the observed point pattern. Therefore, the present study leaves the field of distance-based methods and, instead, explores a density-based approach that avoids the listed spatial issues (modifiable area unit problem, atomistic regions, categorization of locations, comparisons across national borders). We seek a measure that distinguishes between the scale aspect and the two components of the concentration aspect of urbanization (intersectoral mobility of employment and spatial mobility of sectors). In addition, the measure should quantify and rank the economic sectors' degree of urbanization and it should identify the sectors' contributions to the overall result. Finally, the measure should allow for statistical inference. The following section develops such a measure.

3 Measuring the Degree of Economic Urbanization

3.1 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.[15] 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 perfectly uniform distribution over G has a constant density of $f_E(x) = 1/|G|$ everywhere.

We suggest measuring the concentration aspect of the degree of urbanization 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 continuous spatial version of the Hirschman-Herfindahl index. It avoids the urban-rural dichotomy, the delineation of locations, and the regionalization of data. In short, it fully exploits the spatial information contained in geo-coded data. Since the concentration factor is normalized by the area $|G|$, it does not depend on the spatial unit of measurement (e.g., square kilometers 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 the degree of urbanization, u_E, should reflect not only the concentration but also the scale of employment (see Section 2.1). To obtain

[15] $\int_G (\cdot)dx$ 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(lon,lat) = 1 if the geo-coordinate with longitude lon and latitude lat belongs to the country area, and G(lon,lat) = 0 elsewhere.

such a measure, the concentration factor, a_E, is multiplied by a *scale factor*. We propose the average number of employees per unit area, $E/|G|$, as the scale factor. This factor also ensures that economies of different sizes are comparable.

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

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

Its range is $[E/|G|, \infty)$. Rewriting (2) as $u_E = E \cdot \int_G f_E(x)^2 dx$ reveals that the urbanization 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 urbanization 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. Their scale factors 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, unlike the urbanization index, the Arriaga index is neither designed for statistical inference nor for comparing individual sectors (though modifying the index could solve this problem). Furthermore, the treatment of space differs between the indices. Arriaga's approach requires an ex-ante definition of spatial units or neighborhoods and, hence, is subject to the modifiable area unit problem and the problem of atomistic neighborhoods (see Section 2.2). By contrast, the concept of space underlying the urbanization 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.[16] Another essential advantage of the urbanization index is its decomposability (see Section 2.4), an issue that receives a more in-depth treatment in Sections 5 and 6 below.

[16] As described in Section 2, de la Roca and Puga (2017) and Henderson et al. (2021) propose generalized versions of the Arriaga index. Their underlying concept of space can be considered as "quasi-continuous".

3.2 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 C denote the set of all companies. Let w_c denote company c's number of full-time equivalent employees and $x_c = (x_{c1}, x_{c2})$ its geo-coded location. The employment density of total employment, $f_E(x)$, is estimated 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. More important is the choice of the bandwidth h (the "tuning" parameter). If the bandwidth is too small, the density of a single company is highly concentrated in its immediate neighborhood, 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 a company's "effective neighborhood" or "impact region".

Computing the kernel density (3) at a single point x requires evaluating the kernel function K_h for every firm. Hence, if the number of firms is large, it is critically important to apply computationally efficient algorithms. Gramacki (2018) suggests using fast Fourier transforms (FFT) to speed up the kernel density estimations in big data settings.

The estimated counterpart of the concentration factor (1) is

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

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 $\tilde{x}_m$

$(m = 1, \ldots, M)$ 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.

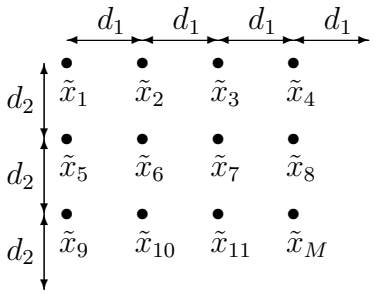

Figure 1: Grid points

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

$$\widehat{u}_E = \frac{E}{|G|}\widehat{a}_E \approx E \sum_{m=1}^{M} [\widehat{f}_E(\tilde{x}_m)]^2 d_1 d_2. \tag{5}$$

3.3 Inference

When estimating the urbanization index for two periods, a routine question is whether the change in the index is statistically significant. A natural test statistic is the difference between the estimated urbanization indices of some reference period 0 and some later period t: $T = \widehat{u}_{E,t} - \widehat{u}_{E,0}$. The null hypothesis of no change is rejected if the test statistic is larger (smaller) than the upper (lower) critical value.

The distribution of the test statistic under the null hypothesis (and hence the critical values) are determined by the percentile bootstrap method (Efron and Hastie, 2016, section 11.2). Let $|C_0|$ and $|C_t|$ denote the number of observed companies in periods 0 and t, respectively. Bootstrap resamples are generated in the same way as the counterfactural spatial distributions in Duranton and Overman (2005) who argue that, due to ubiquitous zoning and planning restrictions, the set of all existing locations is a good proxy for the set of possible locations.[17] Hence, to generate a bootstrap resample, we first draw $|C_0|$ companies

[17]If the focus is on resident population or buildings, rather than workers and firms, the set of potential

from the set C_0 with replacement and store their locations and number of employees. In the same way, a resample is drawn for period t, that is, $|C_t|$ companies from the set C_t. For each bootstrap replication $b = 1, \ldots, B$, the urbanization indices of both resamples $\widehat{u}_{E,0}^{(b)}$ and $\widehat{u}_{E,t}^{(b)}$ are computed, where B is the number of bootstrap replications. Some companies may appear multiple times in the bootstrap resamples, others not at all. Hence, as a statistical artifact, there is an upward bias in the concentration factor of the urbanization index. Using the boostrapped bias correction terms

$$\widehat{v}_0 = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{u}_{E,0}^{(b)} - \widehat{u}_{E,0}\right) \qquad \text{and} \qquad \widehat{v}_t = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{u}_{E,t}^{(b)} - \widehat{u}_{E,t}\right),$$

the bias-corrected urbanization indices of the resamples are $\widehat{u}_{E,0}^{(b)} - \widehat{v}_0$ and $\widehat{u}_{E,t}^{(b)} - \widehat{v}_t$. Moreover, to ensure that the distribution of the test statistic is approximated under the null hypothesis of no change, the difference of the bias-corrected urbanization indices $(\widehat{u}_{E,t}^{(b)} - \widehat{v}_t) - (\widehat{u}_{E,0}^{(b)} - \widehat{v}_0)$ has to be shifted by the difference of the urbanization indices of the original sample, $\widehat{u}_{E,t} - \widehat{u}_{E,0}$. Hence, the bootstrapped test statistics are

$$\begin{aligned} T^{(b)} &= (\widehat{u}_{E,t}^{(b)} - \widehat{v}_t) - (\widehat{u}_{E,0}^{(b)} - \widehat{v}_0) - (\widehat{u}_{E,t} - \widehat{u}_{E,0}) \\ &= \left(\widehat{u}_{E,t}^{(b)} - \frac{1}{B}\sum_{a=1}^{B}\widehat{u}_{E,t}^{(a)}\right) - \left(\widehat{u}_{E,0}^{(b)} - \frac{1}{B}\sum_{a=1}^{B}\widehat{u}_{E,0}^{(a)}\right). \end{aligned}$$

The B realizations of the test statistic $T^{(1)}, \ldots, T^{(B)}$ 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 T is less than the $\alpha/2$-quantile of $T^{(1)}, \ldots, T^{(B)}$ or greater than the $(1 - \alpha/2)$-quantile. Alternatively, one can calculate the p-value of the test as

$$p\text{-value} = 2\min\left(\frac{1}{B}\sum_{b=1}^{B}1(T \leq T^{(b)}), \frac{1}{B}\sum_{b=1}^{B}1(T > T^{(b)})\right)$$

with indicator function $1(A) = 1$ if A is true and $1(A) = 0$ if A is false, and reject the null hypothesis if the p-value is less than the significance level α.

contrafactual locations can be wider, e.g. all locations except non-buildable ones due to water, elevation or slope (de Bellefon et al., 2021).

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

3.4 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*. This confidential data set contains information about all companies with employees subject to social security contributions. Since social security contributions and benefits are calculated based on these data, their reliability 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 and 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.

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 distribution 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.

The sectors are categorized according to the German WZ Classification Code (Statistisches Bundesamt, 2008) that 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 comparable sector classifications in different years, the data set contains a consolidated 3-digit classification based on the WZ Classification Code of 1973 that does not change between 1995 and 2014.

	1995	2000	2005	2010	2014
# companies	1 953 521	2 522 771	2 668 859	2 887 117	2 927 359
Total empl. (E)	24 128	25 250	23 520	24 782	25 869
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 (in thsnd.), and descriptive statistics for the number of full-time equivalent employees per company for 1995 to 2014.

We eliminate very small sectors with less than 10 companies or less than 100 employees as outliers. The number of sufficiently large sectors is constantly above 215.[18]

Concerning the companies' locations, we know where each company $c \in C$ is located. Its geo-coordinates are denoted by $x_c = (x_{c1}, x_{c2})$.[19]

[18]A detailed table listing the number of companies and full-time equivalent employees for each sector in each observation year is provided in the web appendix.

[19]For confidentiality reasons, we were given the firms' municipalities instead of their precise geo-coordinates. The number of municipalities or regions is roughly 11000. There are minor changes in the number of regions due to occasional reshapings, e.g., mergers of municipalities. Our measurement approach is based on geocoded locations, so we assign geo-coordinates $x_c = (x_{c1}, x_{c2})$ to each company. 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 exact geo-coded locations. The data are provided by the "Bundesamt für Kartographie und Geodäsie". The coordinate system is UTM32. The coordinates 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).

3.5 Urbanization Index of Germany

We proceed to estimate the urbanization index for Germany in 1995, 2000, 2005, 2010, and 2014. As argued above, the choice of the bandwidth is relevant. Silverman (1986, section 4.3.2) 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. It is, however, known that this rule of thumb gives too large a bandwidth and leads to over-smoothing (Chacón and Duong, 2018, chap. 3), such that distinct agglomerations in densely populated areas are merged. To mitigate this problem, we determined the bandwidth for the most densely populated German state, North Rhine-Westphalia, yielding a bandwidth of $h = 5{,}000$ (rounded to the nearest 1,000). We then used this bandwidth for the entire area. Figure 2 shows Germany's estimated density of total employment.

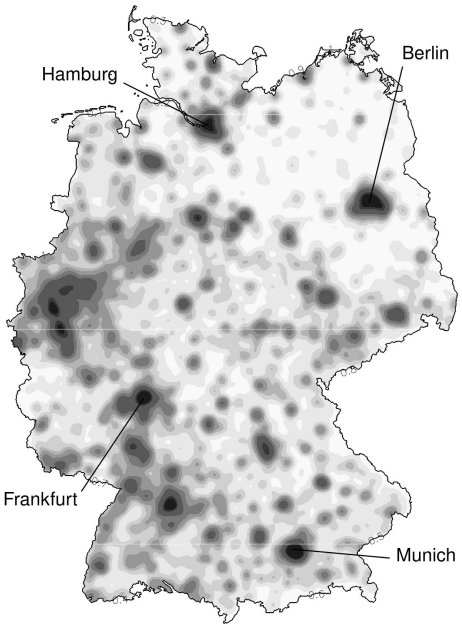

Figure 2: Estimated density of total employment in Germany ($h = 5000$).

The grid points for computing the integrals in $\widehat{u}_E$ and $\widehat{a}_E$ by formula (5) have longitudinal and latitudinal distances $d_1 = d_2 = 1000$ meters resulting in $M = 612\,234$ grid points. The employment numbers, E, are taken from Table 1. The area size of Germany is $|G| = 357\,839$

km^2.

The results are listed in Table 2. The expected number of employees in the neighborhood, $\widehat{u}_E$, is roughly 280 per km^2. A perfectly uniform distribution would result in about 70 employees per km^2. Intertemporal changes in $\widehat{a}_E$ and $E/|G|$ translate into changes in $\widehat{u}_E$. Table 2 shows that, between 1995 and 2014, the urbanization 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.[20] The numbers reveal that the urbanization trend is primarily driven by a positive scale effect rather than a trend toward increasing concentration of employment.

Concerning statistical inference, the bottom row of Table 2 shows that urbanization decreased significantly (at a level of 5 percent and $B = 1000$ bootstrap replications) between 2000 and 2005, then increased significantly between 2005 and 2010 and between 2010 and 2014. Comparing non-adjacent observation periods, the null hypothesis of no change in urbanization cannot be rejected for 1995-2005, 1995-2010, and 2000-2010. Over the total 19-year period from 1995 to 2014, the change in urbanization is highly significant, with a p-value less than 0.001 (not shown in the table).

	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		0.096	0.026	0.036	0.004

Table 2: Evolution of urbanization 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 urbanization between the preceding and the current observation year. The number of bootstrap draws is $B = 1000$. Differences between $\widehat{a}_E \cdot E/|G|$ and $\widehat{u}_E$ are due to rounding errors.

[20] Note that $1.072 \cdot 1.031 = 1.105$.

Even though many changes in urbanization are statistically significant, Table 2 conveys the overall impression that Germany did not go through remarkable urbanization trends. In particular, the concentration factor, a_E, is remarkable stable over time. However, a careful decomposition of the concentration factor reveals that substantial shifts occurred below the deceptive surface of the relatively stable numbers listed in Table 2. The decomposition is conducted in two stages. In the first stage, the change in the concentration factor is decomposed into the intersectoral mobility of employment and the spatial mobility of sectors. The underlying theory and the application to the German employment data are presented in Section 5. Section 6 is devoted to the second stage, that is, the intersectoral mobility of employment and the spatial mobility of sectors 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 urbanization index, u_E, a measure of the urbanization of individual sectors.

4 Urbanization of Individual Sectors

4.1 Definitions

In order to compare the urbanization 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 is not what we need to measure the sector's urbanization. Instead, we need the normalized expected density of *total employment* experienced by a randomly drawn employee of sector i. This is given by

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

$i = 1, \ldots, I$ where I is the number of sectors. 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 employees. Therefore, a_i is the concentration factor of our measure of sectoral urbanization. Again, $E/|G|$ offers itself as the scale 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 urbanization. We name u_i the *urbanization 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, the 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 urbanization index of total employment, u_E, is $[E/|G|, \infty)$, whereas the range of the urbanization index of sector i, u_i, is $(0, \infty)$. If sector i is located 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 where total employment density 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 urbanization 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 urbanization* 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 $a_i - a_E > 0$.

4.2 Estimation

In Section 3.2, we described how $\widehat{u}_E$ can be computed by formula (5). A perfectly analogous formula can be used for the computation of $\widehat{u}_i$:

$$\widehat{u}_i = \frac{E}{|G|}\widehat{a}_i \approx E \sum_{m=1}^{M} \widehat{f}_i(\tilde{x}_m)\widehat{f}_E(\tilde{x}_m)d_1d_2. \tag{8}$$

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

$$\widehat{U}_i = \widehat{u}_i - \widehat{u}_E = E \sum_{m=1}^{M} \left[\widehat{f}_i(\tilde{x}_m) - \widehat{f}_E(\tilde{x}_m)\right] \widehat{f}_E(\tilde{x}_m)d_1d_2. \tag{9}$$

4.3 Inference

We consider two types of hypotheses. First, for the static case, we develop a hypothesis test about the urbanization of a sector: Is a given sector significantly urban or rural? Second, we suggest a procedure to test hypotheses about the evolution of urbanization over time: Is the intertemporal change of a sector's coefficient of urbanization 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, 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 urbanization $\widehat{U}_i$, and the null hypothesis is rejected if the test statistic is greater than an upper critical value (for a significantly urban sector) or less than a lower critical value (for a significantly rural sector). The distribution of the test statistic and the critical values can be determined by bootstrapping. Under the null hypothesis, the spatial distribution of employment in sector i equals the spatial distribution of total employment. We generate the pseudo-samples in a three-step procedure to preserve the company size distribution. 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 drawn in the preceding step. 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 urbanization 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 urbanization is similar to the method described in Section 3.3. In fact, since the difference of the coefficients of urbanization, $\widehat{U}_{i,0} - \widehat{U}_{i,t}$, is closely related to the difference of the urbanization indices, $\widehat{u}_{i,0} - \widehat{u}_{i,t}$, the test approach is virtually identical.

4.4 Sectoral Urbanization in Germany

The sectors' urbanization indices, $\widehat{u}_i$, and their coefficients of urbanization, $\widehat{U}_i$, are computed according to (8) and (9). Figure 3 displays the cumulative distribution functions (cdf) of the urbanization indices, $\widehat{u}_i$, for all sectors in 1995 and 2014. The graphs for 2000, 2005 and

2010 are very similar. The plot demonstrates considerable differences between the sectors but relatively minor changes over the years. The most rural sectors have an urbanization index below 100 (employees per km^2), whereas the index exceeds 600 for the most urban sectors. Half of the sectors have an urbanization index of about 250 or less in all years. None of the sectors has an urbanization index larger than 710 in any year. Overall, the cdf slightly shifts to the left, indicating that a majority of sectors experienced a decline in urbanization.

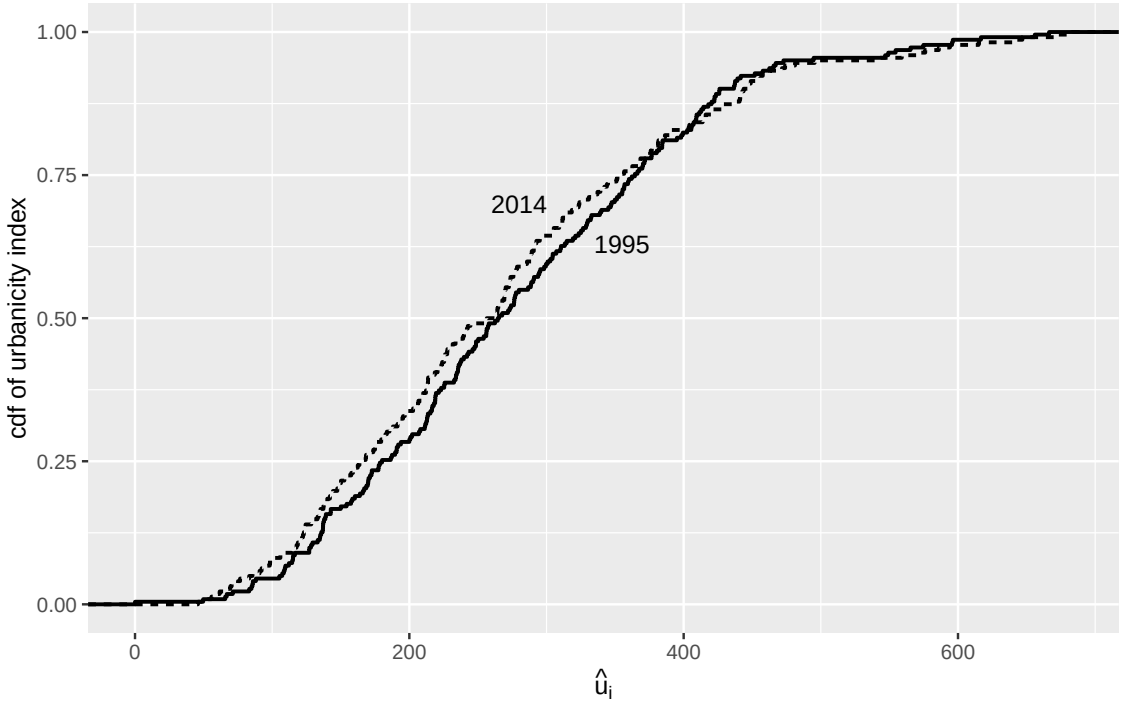

Figure 3: Cumulative distribution functions of the urbanization indices $\widehat{u}_i$ of all sectors for 1995 and 2014.

How many sectors are significantly rural or urban, and which 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 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.[21]

We turn to the change in urbanization over time. For each sector, we tested the null hypothesis that the urbanization index did not change between the first and last observation years (1995 and 2014). Out of all sectors, 41 (that is 18%) changed their urbanization index significantly at the 5 percent level, 12 of them became more rural/less urban (for example, "farmings of animals" and "labor recruitment"), 29 more urban/less rural (for example, "hotels" and "transport via railways"). Of these 29 sectors, two turned from being rural to being urban ("provision of services to the community" and "adult education").[22]

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 in 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 was weak in 1995 (with a correlation coefficient of only 0.047), it increased until 2014 (to 0.124).

Turning to Table 3(b), we find that the correlation between the sectoral urbanization 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 urbanization tends to remain highly urban two decades later. Koh and Riedel (2014) also report a high level of persistence of agglomeration patterns in Germany.

[21] A table showing the urbanization indices $\widehat{u}_i$ and coefficients of urbanization $\widehat{U}_i$ for all sectors and all years is included in the web appendix.

[22] For all sectors, the test statistics and p-values of both the test of urbanization/rurality and the test of changes over time are 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 urbanization indices of all sectors.

5 The Mobility Components IME and SMS

5.1 Decomposition into IME and SMS

The urbanization index of total employment, u_E, is defined as the product of the scale factor, $E/|G|$, and the concentration factor, a_E. Changes in the latter are driven by the intersectoral mobility of employment, IME, and the spatial mobility of sectors, SMS. In the real world, such shifts in employment are likely to co-occur. Therefore, we argued in Section 2.4 that our measure of urbanization should identify and quantify both shifts. 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 also prevailed 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 neighboring 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

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

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

5.2 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 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}:[23]

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

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 (see also Figure 3). 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.

[23]To avoid overburdening the notation, we 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.

6 Sectoral Contributions

As described in Section 2.4, our analysis requires a measure of urbanization that quantifies the contributions of the I individual sectors to the overall change as well as to the changes in its components. For the urbanization 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.

6.1 Sectoral Decomposition

We start with a randomly drawn sector and replace its period 0 number of employees with its period t number of employees. In contrast, 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, which 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 with 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 concentration value, a''_E, say, and the previous value, a'_E, is the contribution of the second selected sector. Also, this contribution can be split into 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 $1,\dots,I$ and let s_j^i denote the counterfactual employment share of sector j if the number of employees of sectors $\sigma(1),\dots,\sigma(i)$ are taken from period t while the number of employees of the remaining sectors are taken from period 0, that is, $E^t_{\sigma(1)},\dots,E^t_{\sigma(i)},E^0_{\sigma(i+1)},\dots,E^0_{\sigma(I)}$. Correspondingly, the counterfactual employment share s_j^{i-1} is obtained when the number of employees of the sectors are $E^t_{\sigma(1)},\dots,E^t_{\sigma(i-1)},E^0_{\sigma(i)},\dots,E^0_{\sigma(I)}$. 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 randomly drawing 1000 permutations, say, and averaging 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:

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

where

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

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

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: $\sum_i \Delta^i_{\text{IME}} = \Delta_{\text{IME}}$. 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: $\sum_i \Delta^i_{\text{SMS}} = \Delta_{\text{SMS}}$. Thus, the decomposition (10) can be further refined to the following decomposition of the economy's concentration factor:

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

The decomposition of the economy'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 $\Delta E_i = E^t_i - E^0_i$.

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

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

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

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

For two reasons, the statistical inference of the decompositions (18) and (20) needs to be revised. First, the bootstrap approach requires repeated computations of the decomposition

many times. As explained above, computing the Bennet decomposition of each repetition is time-consuming, for it is calculated by averaging over many permutations. Hence, 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 visualize the joint distribution. Therefore, we only report point estimates for the sectoral decomposition.

6.2 Sectoral Decomposition for Germany

How much did each sector contribute to the overall change in urbanization in Germany? And in which ways did these changes contribute to the urbanization paradox? To answer these questions, 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 circle represents one sector. The circle sizes represent the employment shares of the sectors in 1995, while the colors indicate their coefficient of urbanization $\widehat{U}_i$ in 1995. The color scale ranges from green (very rural sector) to red (very urban sector). For example, the sector "labor recruitment" was distinctly urban in 1995, while the sector "civil engineering" was distinctly rural. The figure reveals that the changes underlying the urbanization paradox are complex but decipherable.

The ordinate depicts the sector's spatial mobility between 1995 and 2014, Δ^i_{SMS}. The majority of circles are below the abscissa, indicating that, on average, the sectors became more rural, confirming the previous finding $\Delta_{\text{SMS}} = -0.292$. This result is not caused by formerly urban service sectors but by a wide range of industries that include formerly rural farming and manufacturing sectors. The negative values of Δ^i_{SMS} contributed to the overall

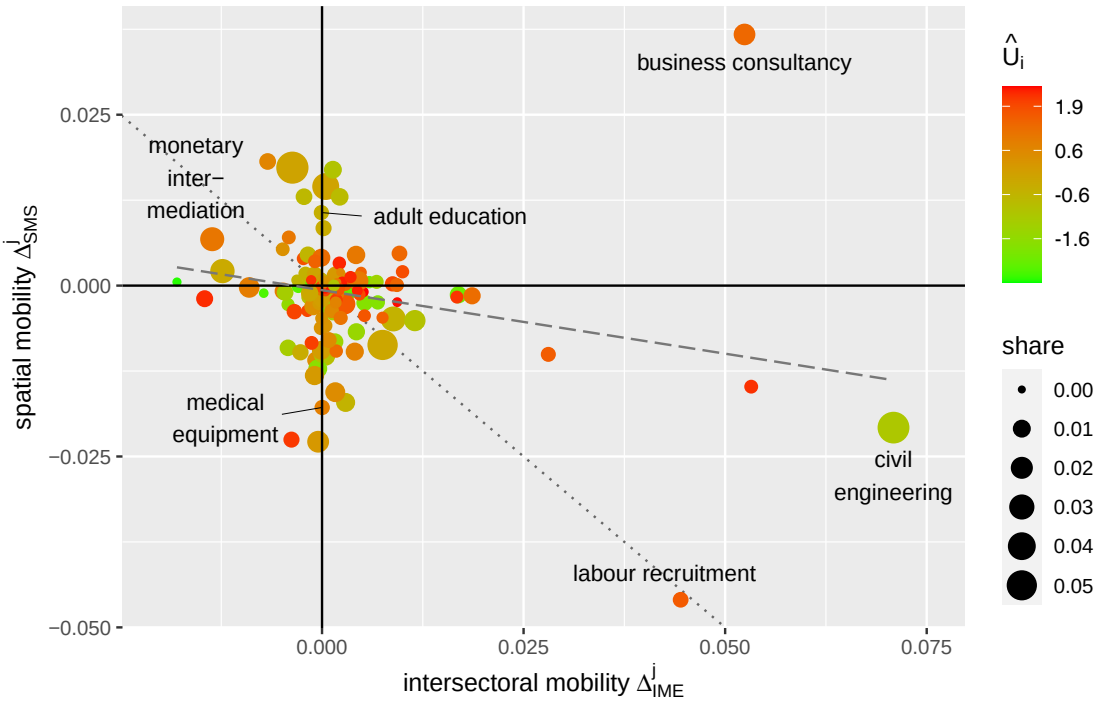

Figure 4: Decomposition of the changes in the sectoral concentration factors between 1995 and 2014 into the sectors' changing employment shares (intersectoral mobility, Δ^i_{IME}) and densities (spatial mobility, Δ^i_{SMS}). Circle sizes reflect the employment shares of the sectors in 1995, circle colors (from green = very rural to red = very urban) reflect the coefficient of urbanization $\widehat{U}_i$ 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.

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 circles 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}. Such a positive contribution arises when a rural sector's employment share falls or when an urban sector's employment share increases.

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. For example, the employment share of the rural sector "civil engineering" declined, whereas the employment share of the more urban sector "labor recruitment" increased.[24] Both shifts resulted in $\Delta^i_{\text{IME}} > 0$, contributing to an increase in 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 "labor recruitment", this offsetting effect was so large that the overall effect on $\widehat{a}_E$ was close to zero.

Such offsetting effects apply to all sectors close to the dotted line in Figure 4 (e.g., the sector "monetary intermediation"). The sector "civil engineering" is clearly located above this line, indicating an overall positive effect of this sector on the concentration factor, $\widehat{a}_E$. The sector "business consultancy" shows the most decisive positive overall effect. This service sector expanded and became considerably more urban.

The spread of the circles 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 dashed regression line confirms the dominant role of the sectors in the lower right quadrant. 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 toward 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 urban sectors, the subgroup of *expanding* sectors strengthened their rural locations to absorb the new employees from the rural sectors.

[24]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 polling; business and management consultancy; holdings" (short: *business consultancy*); "labor recruitment and provision of personnel" (short: *labor recruitment*), "Manufacture of medical and surgical equipment and orthopedic appliances" (short: *medical equipment*), "adult and other education" (short: *adult education*).

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, they kept their mainly 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. Many employees likely switched to more urban sectors without changing their employment location. In other words, the jobs move towards the employees rather than vice versa.

This finding nicely complements the empirical results in Dauth et al. (2017, pp. 338-339). Their analysis reveals that the "switches" from manufacturing sectors to service sectors are not necessarily smooth but often interrupted by spells of joblessness. Furthermore, their findings imply that young entrants usually take their first job in expanding service sectors. Our spatial analysis shows that the expansion of such service sectors was often accompanied by a shift towards more rural locations.

7 Concluding Remarks

The urbanization index of employment is a powerful statistical measure of the degree of urbanization. It is density-based and, therefore, avoids the issues of categorizing and delineating locations. 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 urbanization 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. The detailed sector-by-sector decomposition analysis shows that employment shifted from rural sectors to a wide range of more productive urban sectors, and many of the growing urban sectors absorbed the new employees by expanding their rural production sites. This explains the urbanization puzzle.

Disclosure statement: The authors report there are no competing interests to declare.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data statement: The data are accessible for researchers through the research data centre of the Institute for Employment Research IAB at the Bundesagentur für Arbeit.

References

Arbia, G. (1989), *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Advanced Studies in Theoretical and Applied Econometrics, Springer, Dordrecht.

Arriaga, E. E. (1970), 'A new approach to the measurements of urbanization', *Economic Development and Cultural Change* **18**(2), 206–218.

Arribas-Bel, D., Garcia-Lopez, M.-A. and Viladecans-Marsal, E. (2021), 'Building(s and) cities: Delineating urban areas with a machine learning algorithm', *Journal of Urban Economics* **125**, 103217.

Auer, L. v., Stepanyan, A. and Trede, M. (2019), 'Classifying industries into types of relative concentration', *Journal of the Royal Statistical Society, Series A* **182**(3), 1017–1037.

Baragwanath, K., Goldblatt, R., Hanson, G. and Khandelwal, A. K. (2021), 'Detecting urban markets with satellite imagery: An application to India', *Journal of Urban Economics* **125**, 103173.

Bennet, T. (1920), 'The theory of measurement of changes in cost of living', *Journal of the Royal Statistical Society* **83**(3), 455–462.

Bosker, M., Park, J. and Roberts, M. (2021), 'Definition matters. metropolitan areas and agglomeration economies in a large-developing country', *Journal of Urban Economics* **125**, 103275.

Briant, A., Combes, P.-P. and 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.

Büchel, K. and Ehrlich, M. v. (2020), 'Cities and the structure of social interactions: Evidence from mobile phone data', *Journal of Urban Economics* **119**, 103276.

Ch, R., Martin, D. A. and Vargas, J. F. (2021), 'Measuring the size and growth of cities using nighttime light', *Journal of Urban Economics* **125**, 103254.

Chacón, J. E. and Duong, T. (2018), *Multivariate Kernel Smoothing and Its Applications*, Taylor and Francis.

Dauth, W., Findeisen, S. and Suedekum, J. (2017), 'Trade and manufacturing jobs in germany', *American Economic Review: Papers and Proceedings* **107**(5), 337–342.

de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillond, L. and Gorin, C. G. (2021), 'Delineating urban areas using building density', *Journal of Urban Economics* **125**, 103226.

de la Roca, J. and Puga, D. (2017), 'Learning by working in big cities', *Review of Economic Studies* **84**, 106–142.

Dijkstra, L., Florczyk, A. J., Freire, S., Kemper, T., Melchiorri, M., Pesaresi, M. and Schiavina, M. (2021), 'Applying the degree of urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation', *Journal of Urban Economics* **125**, 103312.

Dingel, J. I., Miscio, A. and Davis, D. R. (2021), 'Cities, lights, and skills in developing economies', *Journal of Urban Economics* **125**, 103174.

Duernecker, G. and Sanchez-Martinez, M. (2023), 'Structural change and productivity growth in Europe — past, present and future', *European Economic Review* **151**, 104329.

Duranton, G. (2021), 'Classifying locations and delineating space: An introduction', *Journal of Urban Economics* **125**, 103353.

Duranton, G. and Overman, H. G. (2005), 'Testing for localization using micro-geographic data', *Review of Economic Studies* **72**(4), 1077–1106.

Duranton, G. and Puga, D. (2020), 'The economics of urban density', *Journal of Economic Perspectives* **34**(3), 3–26.

Efron, B. and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press, New York.

Ellison, G. and Glaeser, E. L. (1997), 'Geographic concentration in U.S. manufacturing industries: A dartboard approach', *Journal of Political Economy* **105**(5), 889–927.

Galdo, V., Li, Y. and Rama, M. (2021), 'Identifying urban areas by combining human judgment and machine learning: An application to India', *Journal of Urban Economics* **125**, 103229.

Gramacki, A. (2018), *Nonparametric Kernel Density Estimation and Its Computational Aspects*, Springer, Cham.

Harari, M. (2020), 'Cities in bad shape: Urban geometry in India', *American Economic Review* **110**(8), 2377–2421.

Henderson, J. V., Nigmatulina, D. and Kriticos, S. (2021), 'Measuring urban economic density', *Journal of Urban Economics* **125**, 103188.

Jedwab, R. and Vollrath, D. (2015), 'Urbanization without growth in historical perspective', *Explorations in Economic History* **58**, 1–21.

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

Lang, G., Marcon, E. and Puech, F. (2020), 'Distance-based measures of spatial concentration: introducing a relative density function', *Annals of Regional Science* **64**, 243–265.

Lemelin, A., Rubiera-Morollón, F. and Gómez-Loscos, A. (2016), 'Measuring urban agglomeration: A refoundation of the mean city-population size index', *Social Indicators Research* **125**, 589–612.

Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez,

E., Ramasco, J. J. and Barthelemy, M. (2014), 'From mobile phone data to the spatial structure of cities', *Scientific Reports* **4**, 5276.

Marcon, E. and Puech, F. (2003), 'Evaluating the geographic concentration of industries using distance-based methods', *Journal of Economic Geography* **3**, 409–428.

Marcon, E. and Puech, F. (2010), 'Measures of the geographic concentration of industries: Improving distance-based methods', *Journal of Economic Geography* **10**, 745–762.

Marcon, E. and Puech, F. (2017), 'A typology of distance-based measures of spatial concentration', *Regional Science and Urban Economics* **62**, 56–67.

Nakamura, R. and Morrison Paul, C. J. (2019), Measuring agglomeration, *in* R. Capello and P. Nijkamp, eds, 'Handbook of regional growth and development theories', Edward Elgar, chapter 19, pp. 386–412.

Openshaw, S. and 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', Pion, London, pp. 127–144.

Roberts, M., Blankespoor, B., Deuskar, C. and Stewart, B. (2017), 'Urbanization and development: Is Latin America and the Caribbean different from the rest of the world?', World Bank Policy Research Working Paper 8019.

Schmucker, A., Seth, S., Ludsteck, J., Eberle, J. and Ganzer, A. (2016), Establishment history panel 1975-2014, FDZ Datenreport. Documentation on Labour Market Data 201603, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Statistisches Bundesamt (2008), 'Klassifikation der Wirtschaftszweige'.
URL: *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.