von Auer, Ludwig; Stepanyan, Andranik; Trede, Mark

Working Paper

Classifying industries into types of relative concentration

Research Papers in Economics, No. 13/16

Provided in Cooperation with:
University of Trier, Department of Economics

Suggested Citation: von Auer, Ludwig; Stepanyan, Andranik; Trede, Mark (2016) : Classifying industries into types of relative concentration, Research Papers in Economics, No. 13/16, Universität Trier, Fachbereich IV – Volkswirtschaftslehre, Trier

This Version is available at:
https://hdl.handle.net/10419/156258

Universität Trier

Classifying Industries Into Types of Relative Concentration

Ludwig von Auer
Andranik Stepanyan
Mark Trede

Research Papers in Economics
No. 13/16

Classifying Industries Into Types of Relative Concentration[☆]

Ludwig von Auer[a], Andranik Stepanyan[a], Mark Trede[b,*]

[a] *Universität Trier, Fachbereich IV-VWL, Universitätsring 15, D-54286 Trier, Germany.*
[b] *Universität Münster, Center for Quantitative Economics, Am Stadtgraben 9, D-48143 Münster, Germany.*

Abstract

When some industries are overrepresented in urban areas (urban concentration), some other industries must be overrepresented in rural areas (rural concentration). Existing measures of concentration do not distinguish between these different *types* of concentration. Instead, they rank industries according to their *degree* of concentration. However, knowing the concentration type is important, when investigating the forces of agglomeration that shape the geographical distribution of an industry. Therefore, the present paper proposes a new statistical approach that classifies each industry into one of seven different geographical patterns, five of which represent different types of concentration. The statistical identification of each industry's geographical pattern is based on two Goodman-Kruskal rank correlation coefficients. The power of our approach is illustrated by German employment data on 613 different industries in 412 regions.

Keywords: Geographical Concentration, Archetypes, Confidence Region, Goodman-Kruskal Coefficient

JEL classification: R10, R12

[☆] We thank Brian Bloch for his comprehensive editing of the manuscript.

[*] Corresponding author. Universität Münster, Institut für Ökonometrie und Wirtschaftsstatistik, Am Stadtgraben 9, 48143 Münster, Germany, Tel.: +49 251 8325006, fax: +49 251 8322012.

Email addresses: vonauer@uni-trier.de (Ludwig von Auer), stepanyan@uni-trier.de (Andranik Stepanyan), mark.trede@uni-muenster.de (Mark Trede)

23rd September 2016

1. Introduction

Around the globe, governments and managers have sought to create, preserve, and develop successful industrial clusters. The results of these efforts have been monitored by numerous empirical studies, many of which compare an industry's degree of concentration to that of other industries. For example, applying standard measures such as the Krugman index or the (relative) Gini coefficient, we can show that in Germany, the two industries "raising sheep and goats" and "radio broadcasting" are equally strongly concentrated. However, it would be misleading to describe the two industries as similarly concentrated, as illustrated in Figure 1.

The map of Germany is depicted in different shades of grey, indicating the density of overall employment. The grey is darkest in urban areas like Munich, Berlin, Cologne, Frankfurt, and Hamburg. The circles in the left panel depict the geographical distribution of employees in "raising sheep and goats". The area of each circle is proportional to the region's density of employees in that industry. The right panel of Figure 1 illustrates the geographical distribution of employees in "radio broadcasting".

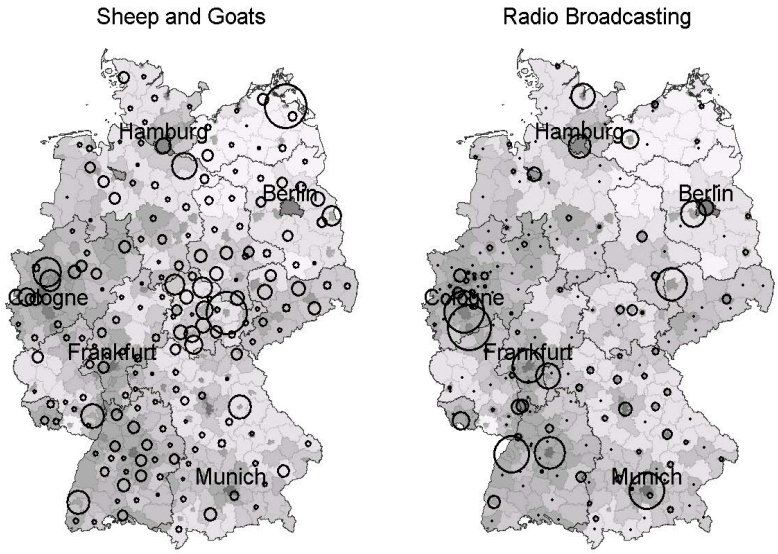

Figure 1: Geographical distribution of employees in the industries *Raising of Sheep and Goats* and *Radio Broadcasting* in Germany in 2010. Darker shades of grey indicate higher density of overall employment, larger circles indicate higher density of employees in the respective industry.

The Krugman index of "raising of sheep and goats" and "radio broadcasting" is 0.68, indicating that, relative to overall employment, both industries have the same *degree* of concentration.

Even a cursory inspection of Figure 1 reveals an important difference between the two industries: "radio broadcasting" is overrepresented in urban areas, whereas "raising of sheep and goats" is overrepresented in rural areas. Hence, although the degree of concentration is the same, the two industries exhibit different *types* of concentration.

Figure 2 depicts an example of two closely related industries. The left panel illustrates the geographical distribution of employees in "general medical practice activities" and the right panel that of "specialist medical practice activities". At first sight, the two panels look very similar and their respective Krugman indices are 0.15 and 0.11. However, a more careful statistical analysis reveals an important difference between the two distributions: "general medical practice activities" are overrepresented in rural areas whereas "specialist medical practice activities" are overrepresented in urban areas.

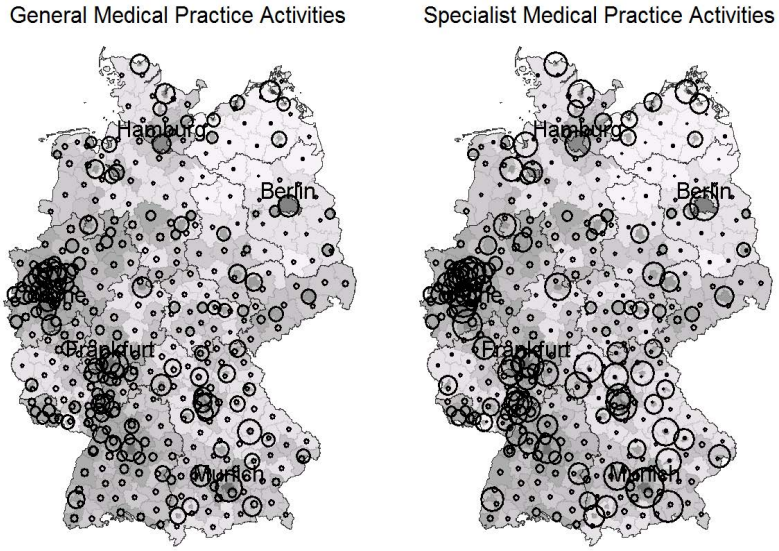

Figure 2: Geographical distribution of employees in the industries *General Medical Practice Activities* and *Specialist Medical Practice Activities* in Germany in 2010. Darker shades of grey indicate higher density of overall employment, larger circles indicate higher density of employees in the respective industry.

The existence of different concentration types follows from a simple logical consideration. When, relative to overall employment, some industries are overrepresented in urban areas, then some other industries must be underrepresented in urban areas, and therefore, overrepresented in rural areas. We can draw an important conclusion from the previous discussion. Even industries with similar *degrees* of concentration may exhibit different *types* of concentration.

Existing measures of concentration are not designed to identify different types of concentration. This is a problem, as the measurement of concentration is often motivated by the claim that it can teach us something about the *forces* responsible for a specific geographical distribution of economic activity. However, the coincidence of the Krugman indices of "radio broadcasting" and "raising of sheep and goats" tells us little about the forces that shape the geographical distribution of these industries that obviously exhibit completely different concentration types. "Radio broadcasting" is subject to an urbanisation force, whereas the geographical distribution of "raising of sheep and goats" is determined largely by natural conditions and the need to avoid urban areas where land is expensive.

This paper aims to identify different *types* of concentration, and its contribution is threefold. First, we define and characterize different geographical archetypes. Second, we develop an intuitive and powerful statistical procedure that classifies each industry into one of the geographical archetypes. The third contribution is empirical. We classify German industries into geographical archetypes, using a large administrative dataset with regionalized German employment data on 613 four-digit industries in 412 regions. As this is a measurement paper, we do not investigate or classify the underlying *forces* of geographical concentration and dispersion.

The paper is structured as follows. Section 2 is a brief review of the literature. Utilizing an artificial data set, we illustrate the different geographical archetypes of industries in Section 3. In Section 4, we explain how, in principal, an industry's employment data can be used to identify its geographical archetype. Real world data, however, require a more elaborate identification approach which we present in Section 5. We apply this approach to German employment data and present the results and some robustness checks in Section 6. The final Section 7 concludes with a summary of

our findings and an outlook on modifications that are necessary to adapt our approach to cases in which geo-referenced firm-level data are available.

2. Three Generations of Measures

Typically, measures of relative geographical concentration compare the industry's geographical employment pattern to the geographical employment pattern of the general economy. Well known measures include the Gini coefficient, Theil index, relative version of the Herfindahl index, and the Krugman (or Isard) index. These "first-generation" measures of concentration (terminology borrowed from Duranton and Overman, 2005, p. 1078) distinguish between "dispersion" and "concentration" and attempt to quantify an industry's degree of concentration, such that inter-industry comparisons are possible (for a comprehensive review, see Combes et al., 2008, pp. 255-275). The empirical basis of such measures are regionalized data sets where the total area is subdivided into regions, and regional employment (or some alternative measure of economic activity) is recorded for each industry.

Though simple to apply, the first-generation measures have some drawbacks. Ellison and Glaeser (1997) argue that the distinction between "dispersion" and "concentration" is insufficient. They introduce the notion of an industry's hypothetical random geographical distribution, conditional both on the overall geographical distribution of the economy, and on the industry's extent of internal economies of scale. Only when the industry's actual geographical distribution displays a significantly larger (lower) degree of concentration than the industry's hypothetical random distribution, should the industry be labelled as concentrated (dispersed). This adds "randomness" as a third type of geographical distribution, taking a middle position between "dispersion" and "concentration". This tripartition distinguishes the second-generation measures from the first-generation ones. Ellison and Glaeser (1997), Maurel and Sédillot (1999) and Devereux *et al.* (2004) propose second-generation measures that are based on regionalized firm-level data.

In some countries (e.g., France, Germany, U.K.) there are data that not only include the number of workers, but also the precise geographical coordinates of each firm. With such geo-referenced firm-level data at hand, distance-based measures of

geographical concentration – measures of the third generation – can be applied. Just like second-generation measures, third-generation measures distinguish between dispersion, randomness, and concentration. In addition, they provide information on the "spatial scale of concentration" (Duranton and Overman, 2005, p. 1077). For example, an industrial cluster covering an area of 500 square kilometers exhibits a larger scale of concentration than a cluster covering only 50 square kilometers. Third-generation measures were introduced into the economics literature by Marcon and Puech (2003) and Duranton and Overman (2005). A comprehensive review is provided by Marcon and Puech (2012). Bickenbach and Bode (2008) demonstrate that the first-generation measures can be augmented to incorporate information on distances between plants or regions.

A drawback of regionalized data is their dependence on the regions' sizes and their borders. With geo-referenced data, this "modifiable area unit problem" (Openshaw and Taylor, 1979; Arbia, 1989) can be solved. Therefore, the strong interest in geo-referenced firm-level data is justified. However, for the foreseeable future, regionalized instead of geo-referenced data will still be the rule rather than the exception. Improving the analysis of concentration when geo-referenced data are not available remains an important issue. It is a major strength of the approach suggested in this paper that it works not only with geo-referenced firm-level data, but also with regionalized data sets that neither contain firm-level information nor information on distances. We apply our approach to this less informative type of data, but describe how our approach can be adapted to geo-referenced firm-level data.

3. Geographical Archetypes

First-generation measures distinguish between two geographical archetypes, namely dispersion and concentration. Second and third-generation measures add randomness as a third geographical archetype, taking a middle position between the former two. However, this tripartition is still insufficient, as the broad category "concentration" can be divided into different, meaningful sub-types.

To introduce and define the geographical archetypes we utilize a simple artificial example in which employment is distributed in one dimension only. In the next sec-

tion, this simple example will also be used to visualize how, in principle, an industry's geographical archetype can be identified from its employment data.

Imagine a country that can be represented by a single straight road stretching from point 0 to point 1. The country's overall employment (its working population) is distributed along that road. The grey lines in diagrams (A) to (F) of Figure 3 depict this distribution. All six diagrams show the same grey line. The area below the line has a unitary measure, i.e. the line shows the density of overall employment. Therefore, the two spikes can be regarded as "urban districts" (high employment density) and the rest as "rural" ones (low employment density). Each diagram of Figure 3 depicts a different industry and the black lines capture the employment densities of the respective industries.

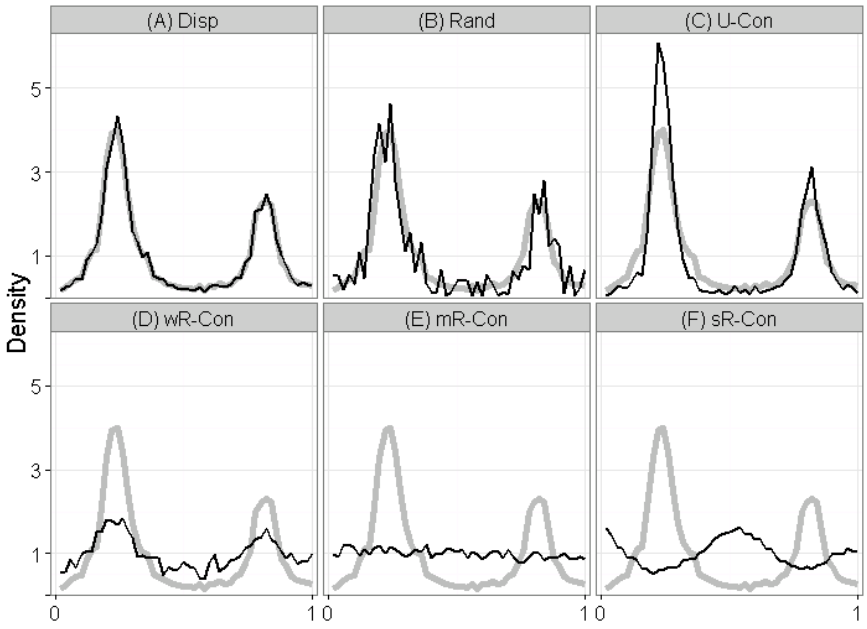

Figure 3: Different geographic archetypes. Grey lines depict the overall employment density (identical in all six diagrams), black lines depict the employment densities of the respective industries.

The employment of industry A (in diagram A) is almost perfectly positively correlated with overall employment. "In other words, plants in the same industry try to be

as scattered as possible" (Maurel and Sédillot, 1999, p. 582). This geographical archetype is usually denoted as *dispersion (Disp)*. Typically, basic services like restaurants or retail bakery sales are included in this type of industry.

However, basic service industries could also fit the geographical archetype depicted in diagram (B) of Figure 3. The diagram is similar to diagram (A), but the employment fluctuations in industry B around overall employment are more pronounced than in industry A. Therefore, the positive correlation between industry B's employment and overall employment is lower than that of industry A. Industry B represents the geographical archetype *randomness (Rand)*, which was emphasized in Ellison and Glaeser (1997) and in many subsequent studies on the measurement of concentration.

All of these studies distinguish between the three cases of dispersion, randomness, and concentration. The present study, however, argues that the category "concentration" is too wide, since it can display different sub-forms that need to be distinguished between, in order to provide a meaningful description and comparison of industries.

Diagrams (C) to (F) depict four archetypes, each representing a different type of concentration. Diagram (C) shows a positive correlation between industry employment and overall employment, but relative to overall employment, industry C's employment is underrepresented in rural areas, and therefore, overrepresented in urban areas. We denote this type of concentration as *urban concentration (U-Con)*. Likely candidates for *U-Con* are specialized service industries such as advertising agencies.

In diagram (D), it is still true that industry employment is positively correlated with overall employment. However, in contrast to industry C, industry D is overrepresented in rural areas, and therefore, underrepresented in urban areas. This type of concentration we denote as *weakly rural concentration (wR-Con)*. General practitioners or pharmacies can be expected to exhibit this type of concentration.

In Diagram (E), there is no longer a clear correlation between industry E's employment and overall employment. Therefore, the industry's overrepresentation in rural areas and underrepresentation in urban areas is even more pronounced. We label this type of concentration *moderately rural concentration (mR-Con)*. Possible candidates for this geographical archetype could be industries that process agricultural products.

Finally, diagram (F) depicts a situation where the industry is grossly overrepresen-

ted in rural areas and grossly underrepresented in urban areas. As a result, employment in industry F and overall employment are negatively correlated. This type of concentration we denote as *strictly rural concentration (sR-Con)*. Industries like livestock farming are likely to exhibit *sR-Con*.

The literature distinguishes between different forces of concentration. Building on the work of Marshall (1890) and Hoover (1937), one can differentiate between internal economies of scale, external economies of scale related to the region's share of the industry's employment (*localisation*), and external economies of scale related to the region's share of overall employment (*urbanisation*). Of course, natural advantages and pure coincidence are also important forces.

The geographical archetypes provide important clues about the forces that shape the geographical concentration. Obviously, when an industry exhibits urban concentration (*U-Con*), the force of urbanisation is strong. If the industry were overrepresented in all urban regions, then the force of urbanisation would be the only relevant one. When the industry is present only in some of the urban regions, however, we can conclude that localisation and, possibly, natural advantages or coincidence are also relevant forces. Industries like farming need affordable land. This desire can be considered as "counter-urbanisation". Therefore, industries assigned to some type of rural concentration – *wR-Con*, *mR-Con*, *sR-Con* – are exposed to counter-urbanisation. For *Rand*-industries, none of these concentration forces is relevant. Dispersion (*Disp*) implies that employment in that industry is distributed almost proportionally to overall employment, suggesting a force that can be regarded as "counter-localisation".

Any comprehensive analysis of geographical concentration should classify each industry into one of the geographical archetypes. In a second step, can the industries be investigated with respect to the forces of concentration. Since this is a measurement paper, we focus on the first step.

4. Assignment of Industries

We consider a country for which neither firm-level information nor information on distances is available. Instead, we have regionalized employment data. For each industry $i = A, B, \ldots$ and each region $r = 1, 2, \ldots, R$ we know the level of employment

x_r^i. Total employment in industry i is

$$x^i = \sum_r^R x_r^i \ ,$$

overall employment in region r is

$$x_r = \sum_i^I x_r^i \ ,$$

and the country's overall employment is

$$x = \sum_i^I x^i = \sum_r^R x_r \ .$$

The overall employment share of region r is given by

$$S_r = \frac{x_r}{x}$$

and the employment share of region r with respect to industry i is defined by

$$s_r^i = \frac{x_r^i}{x^i} \ .$$

Our aim is to develop a statistical tool that classifies industries into geographical archetypes. For this purpose we distinguish between rural and urban regions and examine for each industry whether it shows some systematic over- and underrepresentation pattern in these regions. For example, if an industry is overrepresented in urban regions and underrepresented in rural regions, our classification approach should assign this industry to the geographical archetype urban concentration (*U-Con*). The identification of over- and underrepresentation patterns requires a *relative* measurement concept and a reliable distinction between urban and rural regions.

A region's S_r-value is a poor indicator of its degree of urbanity, because small urban regions might have smaller S_r-values than large rural regions. A better indicator can be obtained, when the regions' sizes, a_r (measured in square kilometers) are available (as is almost always the case). Dividing the overall employment share of region r by its geographical size, a_r, yields the region's overall employment density

$$E_r = \frac{S_r}{a_r} \ . \tag{1}$$

This density is the share of overall employment located within a square kilometer of region r. A region's overall employment density, E_r, is a reliable indicator of its degree of urbanity. The larger the region's E_r-value, the more urban it is.

Refinements of our benchmark, E_r, are conceivable. For example, in the concentration analysis of some industry i, we could subtract x_r^i from x_r to obtain the region's overall employment net of industry i: x_r^{-i}. Instead of the regions' overall employment densities (1), this would generate, for each industry, its own set of overall employment densities,

$$E_r^{-i} = \frac{x_r^{-i} \Big/ \left(\sum_{s=1}^{R} x_s^{-i}\right)}{a_r} \ . \tag{2}$$

For industries with small regional employment shares, x_r^i/x_r, the changes would be negligible. However, for industries with a large regional employment share, the refinement might matter. If we subtract from the overall employment of an urban region the employees of its dominant industry, the region's benchmark becomes more rural ($E_r^{-i} < E_r$). As a result, the industry's type of concentration becomes more rural. In our empirical analysis we have examined whether, for our comprehensive German employment data, the choice between E_r^{-i} and E_r matters. As documented at the end of Section 6, the differences are negligible.

If available, the regions' population (rather than employment) densities would be a possible alternative benchmark. In AppendixC we discuss this variant and present some empirical results. Other benchmarks such as the size of a region, a_r, are less appealing, as there is no clear monotonic relationship between a region's size and its degree of urbanity. Therefore, in our relative measurement concept, the overall employment densities, E_r, serve as the benchmark.

To learn whether an industry i is over- or underrepresented in some region r, we must know not only the region's overall employment density, E_r, but also the employment density of industry i in region r, that is,

$$e_r^i = \frac{s_r^i}{a_r} \ . \tag{3}$$

Note that the ratio of these two densities, e_r^i/E_r, is identical to the so-called location quotient, s_r^i/S_r. Furthermore, $\sum_r^R a_r E_r = 1$ and $\sum_r^R a_r e_r^i = 1$.

How can we utilize the densities e_r^i and E_r for our assignment of industries to geographical archetypes? Suppose some industry i belongs to archetype *Disp.* Then, for every region r, the data should yield $e_r^i \approx E_r$ as illustrated in diagram (A) of Figure 4. The diagram corresponds to diagram (A) of Figure 3 when the "road" of Figure 3 is subdivided into 50 equally large regions ($a_r = a$ for $r = 1, 2, \ldots, 50$). Each point in the scatterplot of diagram (A) in Figure 4 represents one region. The coordinates of each point (region) are given by (E_r, e_r^i). For an industry of archetype *Disp*, the points are located close to the 45 degree line.

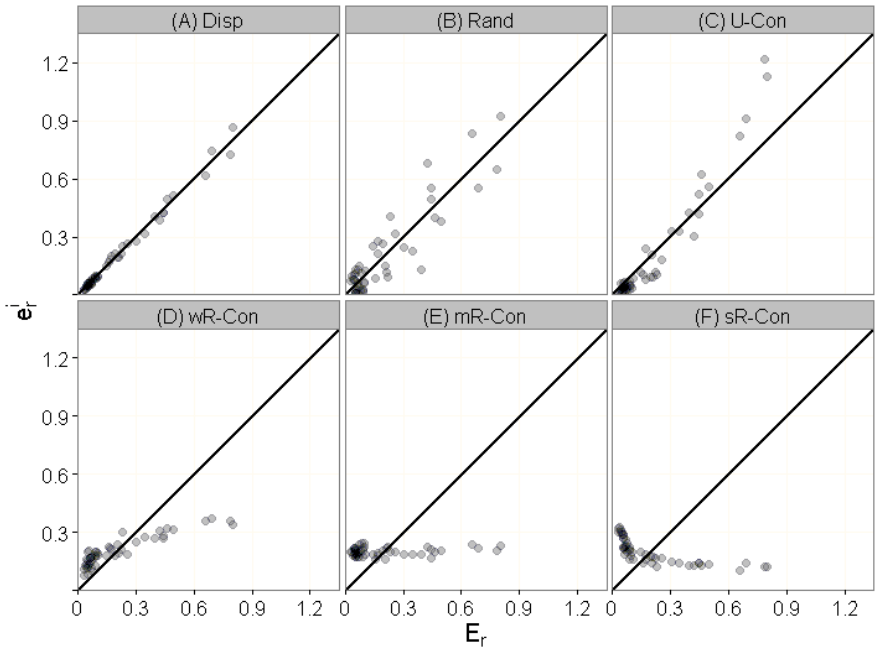

Figure 4: Scatterplots of six different geographical archetypes based on the densities E_r and e_r^i. The densities are derived from Figure 3, where the "road" is subdivided into $R = 50$ equally large regions. Each scatterplot of Figure 4 displays 50 points (regions).

The archetype *Rand* also leads to a point pattern that fluctuates around the 45 degree line, but the fluctuations are larger than with *Disp.* An example is shown in diagram (B) of Figure 4, again corresponding to its counterpart in Figure 3.

Industries that exhibit concentration generate e_r^i-values that deviate systematically

from their corresponding E_r-values. Urban concentration (*U-Con*) implies that, relative to overall employment, employment of industry i is underrepresented in rural regions ($e_r^i < E_r$ when E_r is small) and overrepresented in urban regions ($e_r^i > E_r$ when E_r is large). In diagram (C), the points are below (above) the 45 degree line for small (large) E_r-values. For weakly rural concentration (*wR-Con*) the opposite relationship holds, see diagram (D). The scatterplot of diagram (E) depicts moderately rural concentration (*mR-Con*), where the e_r^i-values are no longer correlated with the E_r-values. In diagram (F), the e_r^i-values decrease as the E_r-values increase. This plot corresponds to strictly rural concentration (*sR-Con*).

It is interesting to compare the archetype *U-Con* in diagram (C) to the archetypes *Rand* and *sR-Con* in diagrams (B) and (F). It turns out that the concentration archetype *U-Con* is more closely related to the archetype *Rand* than to the concentration archetype *sR-Con*, confirming our claim that an industry's concentration can comprise very different types and that it is necessary to distinguish between them.

Each of the six scatterplots (A) to (F) has its own characteristic point pattern. Therefore, it should be possible to infer the geographical archetype of an industry from its scatterplot. For example, when a scatterplot resembles diagram (C), the industry belongs to *U-Con*.

Figure 4 suggests a need to run a nonlinear regression of e_r^i on E_r for every industry i. Using the estimated regression coefficients, the industry can be assigned to an archetype. For example, when a regression line has a positive and increasing slope, as in diagram (C), the industry is assigned to archetype *U-Con*.

Such a line-fitting approach works well for artificial data. However, real world data will rarely generate "well behaved" scatterplots like those in Figure 4, because very few industries are present in all regions. For example, in the German employment data used in our empirical illustration, almost half of the industries are present in less than half the regions (see Figure 8 in Section 6). In other words, a substantial share of points is located on the horizontal axis, so that the assignment of industries to geographical archetypes requires a more sophisticated approach than a nonlinear regression.

5. Assignment of Real World Industries

One might be tempted to eliminate the "absence problem" of real world industries by deleting all points on the horizontal axis and then fitting a regression line through the remaining points. However, the deleted points convey important information for the distinction between the archetypes. For example, for a *U-Con* industry, one would expect the points with $e_r^i = 0$ to be located at small E_r-values: The industry is absent in rural areas. By contrast, for a *sR-Con* industry, points with $e_r^i = 0$ tend to be located at larger E_r-values: The industry is absent in urban areas. Consider an industry with many points on the horizontal axis at larger E_r-values (contradicting *U-Con*), while the other points are located as in diagram (C) of Figure 4 (supporting *U-Con*). If the points on the horizontal axis were discarded, one would wrongly assign the industry to *U-Con*. To avoid such a misclassification, one should keep the points on the horizontal axis.

Another approach to dealing with the absence problem are regression techniques specifically designed for censored data (e.g., Tobit-, Cragg-, or Heckit-regressions). However, for many real world industries, the share of censored data is so large (see Figure 8) that such regressions cannot reliably identify the archetypes.

We propose a different approach that is based on the Goodman-Kruskal rank correlation coefficient of E_r and e_r^i. The Goodman-Kruskal coefficient considers all $R(R-1)/2$ pairs of regions. A pair of regions r and s is concordant (for industry i) if $(E_r - E_s) \cdot (e_r^i - e_s^i) > 0$, and it is discordant if $(E_r - E_s) \cdot (e_r^i - e_s^i) < 0$. When $e_r^i = e_s^i$ or $E_r = E_s$, the pair of regions is then neither concordant nor discordant. Let C_I^i denote the proportion of concordant pairs and D_I^i the proportion of discordant pairs. The Goodman-Kruskal coefficient of industry i is defined as

$$\gamma_I^i = \gamma(E_r, e_r^i) = \frac{C_I^i - D_I^i}{C_I^i + D_I^i} \ , \tag{4}$$

with $C_I^i + D_I^i \leq 1$. The γ_I^i-coefficient can take on values between -1 to 1, where $\gamma_I^i > 0$ signals a positive correlation and $\gamma_I^i < 0$ a negative one.

Figure 4 reveals that the archetype *sR-Con* corresponds to a negative coefficient γ_I^i, the archetype *mR-Con* to a coefficient γ_I^i close to 0, and the remaining four archetypes

wR-Con, *U-Con*, *Rand*, and *Disp* to a positive coefficient γ_I^i. How can we distinguish between the latter four archetypes? For this purpose, we compute a second Goodman-Kruskal coefficient that is based on the location quotients, e_r^i/E_r, instead of the densities e_r^i,

$$\gamma_{II}^i = \gamma(E_r, e_r^i/E_r) = \frac{C_{II}^i - D_{II}^i}{C_{II}^i + D_{II}^i} \ , \tag{5}$$

where C_{II}^i is the proportion of concordant pairs, $(E_r - E_s) \cdot (e_r^i/E_r - e_s^i/E_s) > 0$ and D_{II}^i is the proportion of discordant pairs, $(E_r - E_s) \cdot (e_r^i/E_r - e_s^i/E_s) < 0$. Note that always $\gamma_I^i \geq \gamma_{II}^i$ (see proof in AppendixA).

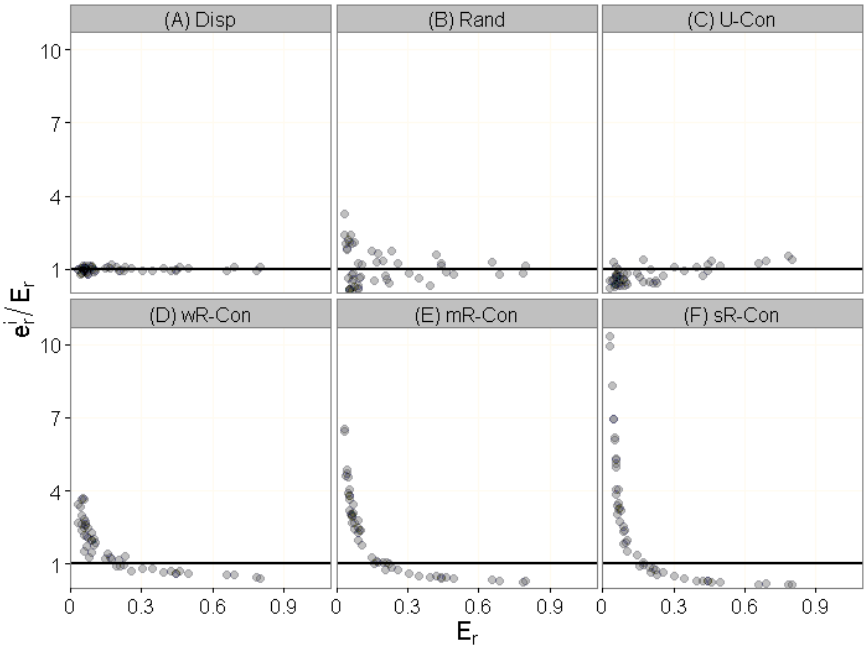

Figure 5: Scatterplots of six different geographical archetypes based on the densities e_r^i and the location quotients e_r^i/E_r corresponding to Figure 3. Each scatterplot of Figure 5 displays 50 points (regions).

Figure 5 shows how γ_{II}^i can help distinguish between *wR-Con*, *U-Con*, *Rand*, and *Disp*. The archetype *U-Con* corresponds to a positive coefficient, *Disp* and *Rand* to a coefficient close to 0, and *wR-Con* (as well as *sR-Con* and *mR-Con*) to a negative coefficient.

Table 1 restates the relationship between the geographical archetypes and the γ_I^i- and γ_{II}^i -values. As an example, consider an industry with $\gamma_I^i = 0.04$ and $\gamma_{II}^i = -0.7$. Since $\gamma_{II}^i < 0$, it is classified as rural concentration. However, as γ_I^i is very close to zero, it is not clear whether the industry is weakly or moderately rural-concentrated. We solve this indeterminacy by taking into account the statistical significance of the estimated γ_I^i- and γ_{II}^i-values. We consider values not significantly different from zero as "≈ 0". However, since the estimators of γ_I^i and γ_{II}^i are correlated, separate hypothesis tests are inappropriate. Taking into account that correlations are linked, we derive, for each industry, not only the γ_I^i- and γ_{II}^i-values, but also their bivariate confidence region.

	$\gamma_I^i = \gamma(E_r, e_r^i)$	$\gamma_{II}^i = \gamma(E_r, e_r^i/E_r)$
Disp	$\gg 0$	≈ 0
Rand	> 0	≈ 0
U-Con	> 0	> 0
wR-Con	> 0	< 0
mR-Con	≈ 0	< 0
sR-Con	< 0	< 0

Table 1: Geographical archetypes and their γ_I^i- and γ_{II}^i-values.

The basic idea behind such confidence regions is illustrated in Figure 6. It depicts the confidence regions for the six industries already shown in Figures 3, 4, and 5. In addition, there is a seventh confidence region (labelled *Mis-Con*) which will be explained below. The horizontal axis depicts $\gamma_I^i = \gamma(E_r, e_r^i)$ and the vertical axis $\gamma_{II}^i = \gamma(E_r, e_r^i/E_r)$.

A confidence region is an elliptic area centred at $(\gamma_I^i, \gamma_{II}^i)$. It has the usual statistical interpretation: given a significance level of 10 per cent, say, the share of repeated samples that produce confidence regions for industry i that cover the industry's pair of "true" Goodman-Kruskal coefficients is 90 per cent. The shape of a confidence region depends on the number of observations (i.e. regions), R, and the significance level. The confidence regions of Figure 6 were computed at a 10 per cent significance level, and

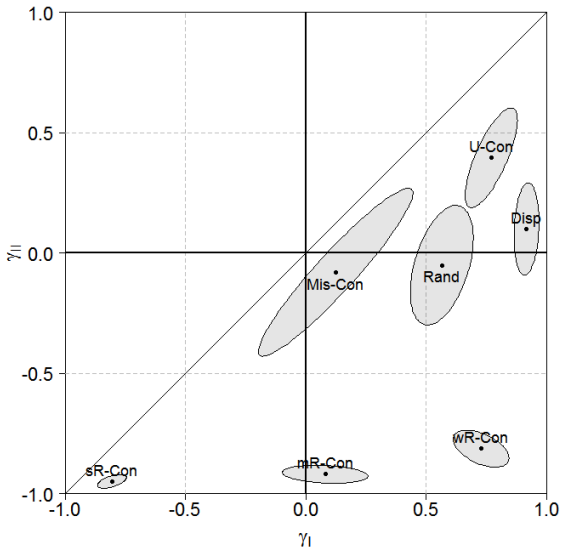

Figure 6: Bivariate confidence regions for the assignment of industries to geographical archetypes.

the number of regions is $R = 50$. The formal derivation of the confidence regions can be found in AppendixB.

If the confidence region of industry i is entirely to the right (left) of the vertical line at $\gamma_I = 0$, the industry's estimated value of γ_I^i is significantly larger (smaller) than zero. Analogous reasoning applies to the horizontal line at $\gamma_{II} = 0$ and the industry's estimated value of γ_{II}^i.

Once we know an industry's confidence region, we can assign the industry to one of the geographical archetypes.

U-Con: The confidence region is entirely above the horizontal axis and entirely to the right of the vertical axis (see Figure 6). In other words, there is strong empirical evidence that $\gamma_I^i > 0$ and $\gamma_{II}^i > 0$, i.e., strong evidence of urban concentration.

If the confidence region is entirely below the horizontal axis, the industry is assigned to one type of rural concentration.

sR-Con: The confidence region is entirely below the horizontal axis and entirely to the left of the vertical axis.

mR-Con: The confidence region is entirely below the horizontal axis and overlaps with the vertical axis.

wR-Con: The confidence region is entirely below the horizontal axis and entirely to the right of the vertical axis.

When the confidence region overlaps with the horizontal axis, but not with the vertical one, the industry is assigned either to *Rand* or to *Disp*. The larger the γ_I^i-value, the stronger the case for *Disp*. We set the boundary at $\gamma_I = 0.5$.

Disp: The confidence region overlaps with the horizontal axis and is entirely to the right of the vertical line drawn at $\gamma_I = 0.5$.

Rand: The confidence region overlaps with the horizontal axis and is entirely to the right of the vertical axis, but not entirely to the right of the vertical line drawn at $\gamma_I = 0.5$.

One case is not covered by these six geographical archetypes. A confidence region may cover both the horizontal and the vertical axis. Such a confidence region would suggest that for this industry, neither γ_I^i nor γ_{II}^i is significantly different from 0. Since a large γ_I^i-value is a signal of dispersion, a small γ_I^i-value signals strong concentration. However, the small value of γ_{II}^i implies that this concentration exhibits neither a pronounced urban nor a pronounced rural pattern. Therefore, we denote this type of concentration as the geographical archetype *miscellaneous concentration (Mis-Con)*.

Mis-Con: The confidence region overlaps with both axes.

In total, the industries are classified into seven geographical archetypes, namely *Disp*, *Rand*, and five different types of concentration. The distinction between *Disp* and *Rand* depends on the cut-off value γ_I. In our assignment rule, this vertical line is set to $\gamma_I = 0.5$. Notice that in Figure 6, the confidence region of industry A does not reach the vertical line at $\gamma_I = 1$, even though the employment of this industry is almost perfectly correlated with overall employment (see Figure 3). The confidence region of industry A is entirely to the right of the vertical line at $\gamma_I = 0.8$. Simulations show that

even industries with confidence regions that are entirely to the right of $\gamma_I = 0.5$, look very much like dispersed industries. Therefore, we propose using the cut-off $\gamma_I = 0.5$ for the archetype *Disp*.

The only other modifiable parameter of our assignment approach is the significance level of the confidence regions. Does the choice of the significance level and the cut-off value for γ_I affect the assignment results? To answer this question we perform some robustness checks. The results are presented at the end of the following section.

6. Geographical Concentration of German Industries

We apply our approach to regionalized German employment data from 2010, provided by the *Institute for Employment Research IAB* at the *Bundesagentur für Arbeit*. The data contain the complete full-time employed population that is subject to social security contributions. As a consequence, self-employed individuals and civil servants are not included. Since social security contributions are calculated on the basis of these data, their reliability far outperforms survey data.

In 2010, Germany was partitioned into $R = 412$ administrative NUTS 3 regions, 102 of which are cities. Size data for each region, a_r, is provided by the *Bundesamt für Kartographie und Geodäsie*. The industries are categorized according to the German WZ 2008 Code. This code mimicks the United Nations "International Standard Industrial Classification (ISIC)" of 2007 and the "Nomenclature statistique des activités économiques dans la Communauté européenne (NACE)" of 2008. At the four-digit level, the WZ 2008 distinguishes between 615 different industries. $I = 613$ of these industries are included in the data set available to us.

For each four-digit industry $i = 1, \ldots, 613$ and each region $r = 1, \ldots, 412$ we know the employment, x_r^i. For each region, we compute its overall employment share, $S_r = x_r/x$, and its overall employment density, $E_r = S_r/a_r$. We also calculate the industries' overall employment shares, x^i/x. The largest share is below 5 per cent. Nevertheless, we use the refined formula (2) to compute the employment densities.

The regions' overall employment ranges from 7.6 employees per square kilometer in Mecklenburg-Strelitz ($E_r = 0.02983 \times 10^{-5}$), to 2030.6 employees per square kilo-

meter in Munich ($E_r = 7.94470 \times 10^{-5}$). The histogram in Figure 7 (grey bars) reveals that the employment densities have a highly skewed distribution.

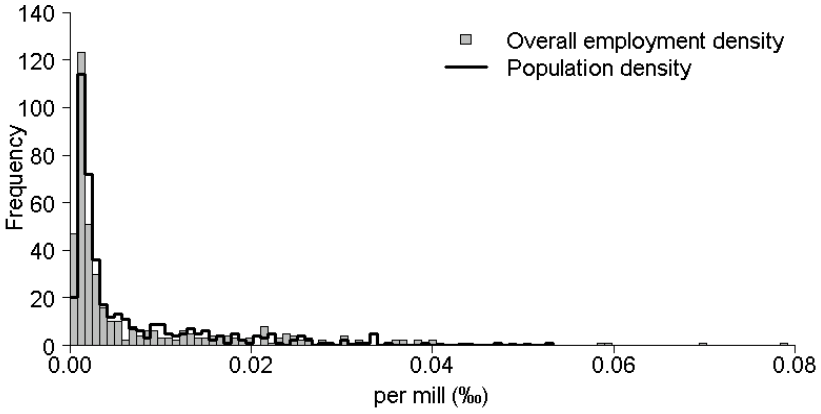

Figure 7: Histograms of the employment densities E_r (grey bars) and population densities (black line) in 412 NUTS 3 regions.

Furthermore, for each industry i and each region r, formula (3) gives us the industry's employment density, e_r^i. For each industry, we compute the share of regions with $e_r^i > 0$ (i.e. the proportion of regions where industry i is present) and denote it by z^i. We order the industries by their z^i-value. Figure 8 depicts the results, with the industries on the horizontal axis and z^i on the vertical axis. The figure shows that 41 per cent of the 613 industries have a share z^i below 50 per cent. Only 13 per cent of the industries are present in all regions ($z^i = 1$). These results confirm the abovementioned "absence problem" in real world employment data.

For each industry, we compute the Goodman-Kruskal coefficients $\gamma_I^i = \gamma(E_r, e_r^i)$ and $\gamma_{II}^i = \gamma(E_r, e_r^i/E_r)$ together with their joint confidence region (B.3), and compare the confidence region to the cut-off lines $\gamma_I = 0$ (vertical axis in Figure 6), $\gamma_{II} = 0$ (horizontal axis), and $\gamma_I = 0.5$ (vertical line at $\gamma_I = 0.5$). We use a significance level of 10 per cent. Applying the approach described in Section 5, we classify 608 of the 613 industries into one of the seven geographical archetypes.[1]

[1]Five industries (e.g., "raising camels") could not be assigned to an archetype, because the computation of confidence regions requires that at least one pair of regions is discordant and another pair concordant. The

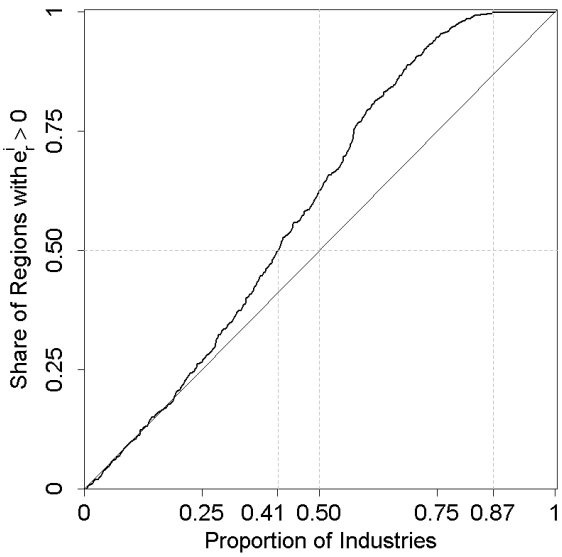

Figure 8: Visualization of the "absence problem".

For the 608 industries, Figure 9 shows their Goodman-Kruskal coefficients γ_I and γ_{II} and the geographical archetypes. Each point in the diagram represents one industry. The location of the point indicates the industry's values of γ_I^i and γ_{II}^i. The geographical archetypes are indicated by the symbols. Empty symbols stand for industries with rural concentration, with circles indicating the archetype *sR-Con*, triangles indicating *mR-Con*, and squares indicating *wR-Con*. Crosses symbolize the archetype *Mis-Con*, and the filled symbols stand for the archetypes *Rand* (filled circles), *Disp* (filled triangles), and *U-Con* (filled squares).

32 of the 613 industries are assigned to strictly rural concentration (*sR-Con*). These industries represent a mere 0.8 per cent of total employment. *sR-Con* is dominated by the agricultural sector (e.g., growing grain; raising cattle, sheep, goats, pigs and poultry; mixed agriculture).

The agricultural sector (e.g., growing of vegetables and potatoes) also plays an important role in the geographical archetype of moderately rural concentration (*mR-*

total employment of these five industries was 180 (out of 25,561,128) employees.

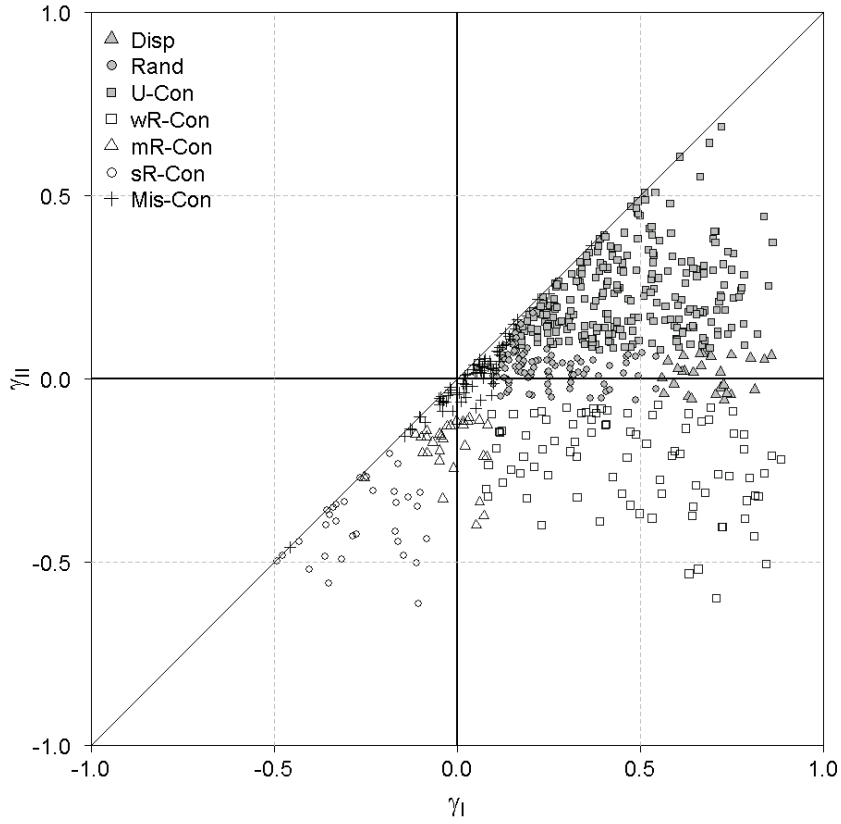

Figure 9: Geographical archetypes of German industries in 2010.

Con). However, several food processing industries (e.g., processing of fish, production of juices, processing of milk) and a couple of other industries (e.g., tyre remolding, production of spirituous beverages) are also classified as *mR-Con*. This archetype contains 30 of the 613 industries, representing an employment share of 1.2 per cent.

87 industries, or 31.1 per cent of total employment, are assigned to weakly rural concentration (*wR-Con*). Agriculture is largely absent from *wR-Con*. The composition of this archetype is more heterogeneous than the compositions of *sR-Con* and *mR-Con*. Many industries in the construction sector belong to this archetype (e.g., construction of buildings and roads, electrical installation, roofing, tiling, plastering). Furthermore, there are many basic retail sale industries (e.g., filling stations, food stores, butchers,

pharmacies) and industries related to basic services (e.g., general practitioners, dentists, hotels, hairdressers, driving schools, funeral parlours) in this archetype. Some manufacturing industries are assigned to *wR-Con* as well. Most of them, however, are related to construction (e.g., manufacturing of office furniture; production of fresh concrete; production of elements made of concrete, cement and sand-lime brick).

Most manufacturing industries and most wholesale ones can be found in the geographical archetype urban concentration (*U-Con*). 255 industries are assigned to this archetype. They cover 45.9 per cent of total employment. The archetype's composition is extremely heterogeneous, ranging from manufacturing, wholesaling, and retailing to a wide range of services (e.g., pubs, taxis, cinemas, life insurance, advertising agencies, security firms, hospitals, universities).

The archetype dispersion (*Disp*) is dominated by the retail industry and by services (e.g., bakeries, retailing of fruits and vegetables, retailing of cosmetic products and toiletries, restaurants, nursery schools, and churches). 27 industries with a combined employment share of 9.5 per cent are assigned to this archetype.

The archetype randomness (*Rand*) comprises 96 industries with a combined employment share of 9.9 per cent. Manufacturing has the largest share within this archetype. However, wholesale (e.g., sugar, sweets, bakery products, flowers, fruits, and vegetables), a few retail sale industries (e.g., fish), and some services (e.g., event-caterer, renting of aircrafts, amusement and theme parks, laundry) are also present in *Rand*.

Only 1.6 per cent of total employment are assigned to the archetype of miscellaneous concentration (*Mis-Con*). Since 81 industries belong to this archetype, the employment per industry is low. On average, the *Mis-Con*-industries are present in only one fifth of all regions. In fact, none of the industries assigned to this archetype is present in more than half the regions. Manufacturing dominates this archetype (e.g., production of sugar, sanitary ware, shoes, bright steel, arms and munitions, ships, toys, kitchens). There are only a few industries from agriculture (e.g., growing of grapes) and some service industries, many of which are related to shipping (e.g., repair of ships, inland navigation, coastal shipping).

As pointed out before, our assignment rules utilize only two parameters. The first one is the significance level for the confidence regions. The second parameter is the

cut-off value of γ_I that distinguishes between the archetypes *Disp* and *Rand*. As a robustness check, we examine in how far varying these parameters alters the assignment results.

Lowering the significance level from 10 per cent to 5 per cent (or even 1 per cent) increases the size of the confidence regions. However, the inflation is modest as shown in Figure 10 for seven different industries. The grey ellipses correspond to a significance level of 10 per cent. They are surrounded by two larger ellipses corresponding to the 5 per cent and 1 per cent level. The small inflation implies that only few assignments are altered as we lower the significance level. In fact, less than 7.2 per cent of the industries (representing a total employment share of 5.1 per cent) change their assignment when the significance level is lowered from 10 per cent to 5 per cent. Most of them move from *U-Con* to *Rand* and from *Rand* to *Mis-Con*.

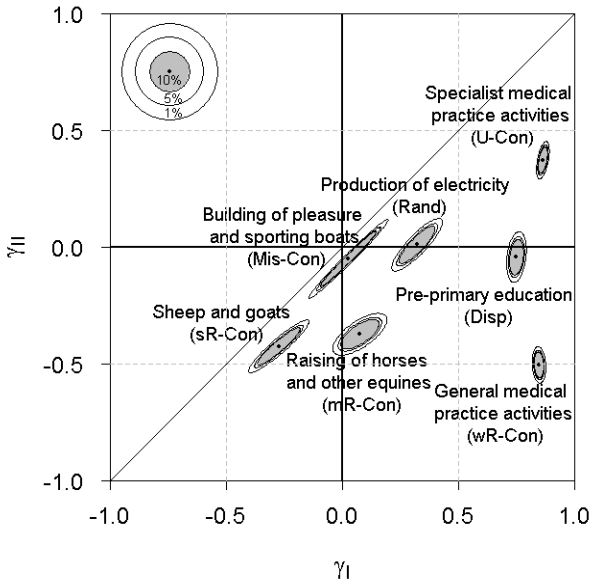

Figure 10: The inflation of confidence regions when the significance level is lowered.

Next, we vary the original cut-off value $\gamma_I = 0.5$ to 0.6 and 0.4. The only assignment changes that can occur are between *Disp* and *Rand*. For $\gamma_I = 0.6$, the number of *Disp*-industries drops from 27 to 17, while for 0.4 it increases to 35. Apparently, the

cut-off value matters for the distinction between *Disp* and *Rand*.

As a final robustness check, we investigate the impact of changing the benchmark distribution. Our empirical analysis of the German employment data is not based on overall employment, E_r, but on overall employment net of the relevant industry's own employment, $E_r^{-i} = E_r - x_r^i$. However, the choice between these benchmarks turns out to be irrelevant, as only three industries with a combined employment share of 2.1 per cent change their archetype when overall employment is used as the benchmark distribution. Another possible benchmark is population density. In how far this benchmark results in different assignments is investigated in AppendixC.

7. Concluding Remarks

Investigating the geographical concentration of industries should start by identifying each industry's geographical archetype. We define seven archetypes, five of which represent different types of concentration. Within the latter group, we emphasize the distinction between rural and urban concentration. If all industries were present in all regions, assigning the industries to the most appropriate archetype would be a rather straightforward regression exercise. In the real world, however, few industries are present in all regions. Therefore, we develop a new statistical approach that can deal with such data.

Our approach is based on two Goodman-Kruskal rank correlation coefficients and their respective confidence region. Depending on the position and size of the confidence region, the industry is classified into a geographical archetype, and each assignment is associated with a specific level of statistical significance. It is useful, but not essential, to know the geographical size of the regions.

We apply our approach to a rich and reliable data set on employment in Germany. Our empirical findings reveal that the 613 German industries exhibit very different types of concentration. All seven geographical archetypes are relevant. We identify clear differences between the geographical patterns of agriculture, manufacturing, retail sale, wholesale, basic services, and other services.

It is another virtue of our assignment approach that it can cope with regionalized data sets that neither contain firm-level data, nor information on distances. In most

countries, only this type of data exists. However, in exceptional circumstances, empirical researchers may have geo-referenced firm-level data. Fortunately, our assignment approach can be adapted to such cases. One can use geo-referenced firm-level data to compute the kernel density distributions of overall employment and the employment of each industry. The estimated kernel densities can then be used to compute the two Goodman-Kruskal coefficients.

References

Arbia, G., 1989. Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems. Kluwer: Dordrecht, The Netherlands.

Bickenbach, F., E. Bode, 2008. Disproportionality Measures of Concentration, Specialization, and Localization. International Regional Science Review, 31(4), 359-388.

Combes, P.-P., T. Mayer, J.-F. Thisse, 2008. Economic Geography. Princeton University Press: Princeton (New Jersey).

Devereux, M.P., R. Griffith, H. Simpson, 2004. The Geographic Distribution of Production Activity in the UK. Regional Science and Urban Economics, 35, 533–564.

Duranton, G., H.G. Overman, 2005. Testing for Localization Using Micro-Geographical Data. Review of Economic Studies, 72, 1077-1106.

Ellison, G., E.L. Glaeser, 1997. Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. Journal of Political Economy, 105, 889-927.

Hoeffding, W., 1948. A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics, 19, 293-325.

Hoover, E.M., 1937. Location Theory and the Shoe and Leather Industries. Harvard University Press: Cambridge, MA.

Kowalski, J., X.M. Tu, 2008. Modern Applied U-Statistics. Wiley and Sons: Hoboken, New Jersey.

Marcon, E., F. Puech, 2003. Evaluating the Geographic Concentration of Industries using Distance-Based Methods. Journal of Economic Geography, 3, 409-428.

Marcon, E., F. Puech, 2012. A Typology of Distance-Based Measures of Spatial Concentration, halshs-00679993v3.

Marshall, A., 1890. Principles of Economics. Macmillan: London.

Maurel, F., B. Sédillot, 1999. A Measure of the Geographic Concentration in French Manufacturing Industries. Regional Science and Urban Economics, 29, 575-604.

Openshaw, S., P.J. Taylor, 1979. A Million Or So Correlated Coefficients: Three Experiments On the Modifiable Areal Unit Problem. In: Statistical Applications in the Spatial Sciences, N. Wrigley and R.J. Bennet, 127-144. Pion: London.

Särndal, C.-E., B. Swensson, J. Wretman, 2003. Model Assisted Survey Sampling. Springer: New York.

AppendixA. Proof that $\gamma_I^i \geq \gamma_{II}^i$

Consider an industry i. For every region with $s_r^i = 0$ we obtain $e_r^i = 0$ and $e_r^i/E_r = 0$. Therefore, the number of ties is identical in $\gamma_I(E_r, e_r^i)$ and $\gamma_{II}(E_r, e_r^i/E_r)$: $C_I^i + D_I^i = C_{II}^i + D_{II}^i$.

Next, consider the coefficient $\gamma_{II}(E_r, e_r^i/E_r)$ and some concordant pair of regions r and s. Since $E_r < E_s$ and $e_r^i/E_r < e_s^i/E_s$, it follows that

$$\begin{aligned} &\left(e_r^i/E_r\right)E_r < \left(e_s^i/E_s\right)E_s \\ \Rightarrow \qquad &e_r^i < e_s^i \, . \end{aligned}$$

Hence, every pair of regions that is concordant with respect to E_r and e_r^i/E_r is also concordant with respect to E_r and e_r^i.

Now consider a pair of regions that is discordant with respect to E_r and e_r^i/E_r: $E_r < E_s$ and $e_r^i/E_r > e_s^i/E_s$. When $e_s^i = 0$, this discordance then implies that the pair of regions is also discordant with respect to E_r and e_r^i. However, when $0 < e_r^i < e_s^i$ and $E_r \ll E_s$, we then have concordance with respect to E_r and e_r^i, but possibly $e_r^i/E_r > e_s^i/E_s$, that is, discordance with respect to E_r and e_r^i/E_r.

In sum, we obtain $C_I^i \geq C_{II}^i$ and $D_I^i \leq D_{II}^i$, and therefore, $\gamma_I(E_r, e_r^i) \geq \gamma_{II}(E_r, e_r^i/E_r)$. The share of potential pairs of regions that are concordant with respect to E_r and e_r^i,

but discordant with respect to E_r and e_r^i/E_r, increases with z^i (the share of regions with $e_r^i > 0$) and also with the variance of E_r among this group of regions. In other words, the larger the share z^i, the more $\gamma_I(E_r, e_r^i)$ can exceed $\gamma_{II}(E_r, e_r^i/E_r)$. A second influencing factor is the value of $\gamma_{II}(E_r, e_r^i/E_r)$. A large positive value implies that there are few discordant pairs which can turn into concordant ones with respect to E_r and e_r^i. A large negative value (in absolute terms) indicates that there are many discordant pairs which can turn into concordant ones with respect to E_r and e_r^i.

AppendixB. Derivation of Bivariate Confidence Regions

The confidence regions in Figure 6 are computed as follows. The observations (E_r, e_r^i), $r = 1, \ldots, R$, may be interpreted as a random sample from a superpopulation (E, e^i).[2] Let (E_1, e_1^i) and (E_2, e_2^i) be independent draws from (E, e^i). We define the following probabilities of concordances and discordances:

$$\begin{aligned}
\pi_{C,I}^i &= P\left((E_1 - E_2)(e_1^i - e_2^i) > 0\right) \\
\pi_{D,I}^i &= P\left((E_1 - E_2)(e_1^i - e_2^i) < 0\right) \\
\pi_{C,II}^i &= P\left((E_1 - E_2)(e_1^i/E_1 - e_2^i/E_2) > 0\right) \\
\pi_{D,II}^i &= P\left((E_1 - E_2)(e_1^i/E_1 - e_2^i/E_2) < 0\right) \ .
\end{aligned}$$

The sample proportions C_I^i, D_I^i, C_{II}^i and D_{II}^i are estimators of these probabilities, and the Goodman-Kruskal coefficients (4) and (5), calculated from the regional sample data, are point estimators for the values

$$\Gamma_I^i(E, e^i) = \frac{\pi_{C,I}^i - \pi_{D,I}^i}{\pi_{C,I}^i + \pi_{D,I}^i} \tag{B.1}$$

$$\Gamma_{II}^i(E, e^i) = \frac{\pi_{C,II}^i - \pi_{D,II}^i}{\pi_{C,II}^i + \pi_{D,II}^i} \tag{B.2}$$

of the superpopulation.

In order to construct joint confidence intervals for Γ_I^i and Γ_{II}^i, we draw on asymptotic theory for multivariate U-statistics and the delta method. As shown in Hoeffding

[2] See Särndal, Swensson and Wretman (2003), chap. 14.5, for the concept of superpopulations.

(1948) and Kowalski and Tu (2008), the proportions of concordances and discordances are asymptotically normally distributed as $R \to \infty$,

$$\sqrt{R}\left(\begin{bmatrix} C_I^i \\ D_I^i \\ C_{II}^i \\ D_{II}^i \end{bmatrix} - \begin{bmatrix} \pi_{C,I}^i \\ \pi_{D,I}^i \\ \pi_{C,II}^i \\ \pi_{D,II}^i \end{bmatrix}\right) \sim N(0, \Sigma).$$

The covariance matrix Σ can be estimated consistently from the data in the following way (Hoeffding, 1948). The univariate statistic $\pi_{C,I}$ is estimable by a U-statistic of degree 2, since

$$E\left(\varphi_C\left((E_1, e_1^i), (E_2, e_2^i)\right)\right) = \pi_{C,I}$$

where the kernel φ is defined as

$$\varphi_C\left((E_1, e_1^i), (E_2, e_2^i)\right) = 1(E_1 < E_2, e_1^i < e_2^i) + 1(E_1 > E_2, e_1^i > e_2^i)$$

with indicator function $1(A) = 1$, if A is true, and 0 otherwise. The kernel for discordances is

$$\varphi_D\left((E_1, e_1^i), (E_2, e_2^i)\right) = 1(E_1 < E_2, e_1^i > e_2^i) + 1(E_1 > E_2, e_1^i < e_2^i).$$

The estimator of $\pi_{C,I}$ is the U-statistic

$$C_I = \binom{R}{2}^{-1} \sum_{r<s} \varphi_C\left((E_r, e_r^i), (E_s, e_s^i)\right)$$

where the summation extends over all pairs of regions and R is the number of regions. The U-statistic has a normal asymptotic distribution, since the second moment of the kernel $E(\varphi_C^2(\cdot,\cdot))$ exists. For a large R, the variance is approximately

$$Var(C_I) = \frac{4}{R}\zeta$$

with

$$\zeta = E\left(\varphi_{C,1}^2\left((E_1, e_1^i)\right)\right) - \pi_{C,I}$$

and

$$\varphi_{C,1}(\cdot) = E\left(\varphi_C\left(\cdot, (E_2, e_2^i)\right)\right).$$

In order to estimate the variance, we need a consistent estimator for ζ. The empirical counterpart of $\varphi_{C,1}(E_r, e_r^i)$ is

$$\hat{\varphi}_{C,1}(E_r, e_r^i) = \frac{1}{R-1} \sum_{s=1}^{R} \varphi_C \left((E_r, e_r^i), (E_s, e_s^i) \right).$$

Then

$$\hat{\zeta} = \frac{1}{R} \sum_{r=1}^{R} \left(\frac{1}{R-1} \sum_{s=1}^{R} \varphi_C \left((E_r, e_r^i), (E_s, e_s^i) \right) \right)^2 - (C_I)^2$$

and the estimated variance of C_I is $4\hat{\zeta}/R$.

When two U-statistics are considered jointly (e.g. C_I and D_I), the derivations proceed in the same way. Their covariance $Cov(C_I, D_I)$ can be estimated by

$$\widehat{Cov}(C_I, D_I) = \frac{4}{R} \hat{\zeta}^{C,D}$$

with

$$\begin{aligned} \hat{\zeta}^{C,D} = & \frac{1}{R} \sum_{r=1}^{R} \left(\frac{1}{R-1} \sum_{s=1}^{R} \varphi_C \left((E_r, e_r^i), (E_s, e_s^i) \right) \right) \\ & \times \left(\frac{1}{R-1} \sum_{s=1}^{R} \varphi_D \left((E_r, e_r^i), (E_s, e_s^i) \right) \right) \\ & - C_I D_I. \end{aligned}$$

The estimator $\hat{\Sigma}$ of the covariance matrix Σ is built from these estimated variances and covariances.

Since (4) and (5) are differentiable functions of the proportions, the delta method applies, and hence, the random vector $(\gamma_I^i, \gamma_{II}^i)'$ is asymptotically normally distributed with expectation vector $(\Gamma_I^i, \Gamma_{II}^i)'$ and covariance matrix $J\Sigma J'$, where the Jacobian matrix, J, is

$$J = \begin{bmatrix} \frac{2D_I^i}{(C_I^i + D_I^i)^2} & -\frac{2C_I^i}{(C_I^i + D_I^i)^2} & 0 & 0 \\ 0 & 0 & \frac{2D_{II}^i}{(C_{II}^i + D_{II}^i)^2} & -\frac{2C_{II}^i}{(C_{II}^i + D_{II}^i)^2} \end{bmatrix}.$$

A $(1-\alpha)$-confidence region for $(\Gamma_I^i, \Gamma_{II}^i)'$ is given by the elliptically shaped set

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} : \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \gamma_I^i \\ \gamma_{II}^i \end{bmatrix} \right)' [J\Sigma J/R]^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \gamma_I^i \\ \gamma_{II}^i \end{bmatrix} \right) \leq q_{1-\alpha} \right\} \tag{B.3}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2-distribution with 2 degrees of freedom.

AppendixC. Using Population Densities Instead of Employment Densities

If we use the regions' population densities as a benchmark and if the share of commuters living in rural regions, but working in urban regions, is large, there are then fewer distinctly urban and fewer distinctly rural regions than in the benchmark cases E_r (overall employment) or E_r^{-i} (net overall employment). The data depicted in Figure 7 confirm this conjecture. The black line shows the total population density. It starts at a lower level than the first grey bar, i.e. there are fewer distinctly rural regions, and it does not spread out as far to the right as the grey bars, i.e. there are fewer distinctly urban areas.

How this will affect the assignment results, can be inferred from Figure 3. The black lines do not change since they are computed from employment data and not from population data. However, the grey lines will become more uniformly distributed in the $[0, 1]$-interval. As a result, some of the former *Disp*- and *Rand*-industries become *U-Con*-industries. Some former *wR-Con*-industries move to *Disp*-, *Rand*-, or even *U-Con*. However, only few of the former *mR-Con*-industries and even fewer of the former *sR-Con*-industries will change their assignment.

Our German employment data confirm all of these predictions. All *Disp*-industries, 20.8 per cent of the *Rand*-industries, and 11.5 per cent of the *wR-Con*-industries turn into *U-Con*-industries. Furthermore, all *sR-Con*-industries remain in that archetype. In total, 17.9 per cent of the industries change their archetype.