

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

Conference Paper

Regional Concentration and Confidence Regions

55th Congress of the European Regional Science Association: "World Renaissance: Changing roles for people and places", 25-28 August 2015, Lisbon, Portugal

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Stepanyan, Andranik; von Auer, Ludwig; Trede, Mark (2015) : Regional Concentration and Confidence Regions, 55th Congress of the European Regional Science Association: "World Renaissance: Changing roles for people and places", 25-28 August 2015, Lisbon, Portugal, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:

<https://hdl.handle.net/10419/124660>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Regional Concentration and Confidence Regions

Ludwig von Auer Andranik Stepanyan
Universität Trier Universität Trier

Mark Trede
Universität Münster

February 2015 (file: wp54v14.tex)

Abstract

Industries necessarily differ with respect to their type of geographical concentration. When some industries are overrepresented in urban areas (urban concentration), then some other industries must be overrepresented in rural areas (rural concentration). Unfortunately, the existing measures of concentration cannot distinguish between urban and rural concentration. They simply ignore the problem and rank industries with respect to their *degree* of concentration, even though these industries may exhibit completely different *types* of concentration. In the present paper we develop a new approach that avoids such misleading comparisons. Our approach distinguishes not only between urban and rural concentration but between seven different geographical patterns. The statistical identification of each industry's geographical pattern is based on two Goodman-Kruskal rank correlation coefficients and their bivariate confidence region. Using German employment data on 613 different industries, the power of our approach is demonstrated.

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 (relative) Gini coefficient, we can show that in Germany the two industries forestry and call centers are equally strongly concentrated. However, it would be a serious mistake to describe the two industries as being similarly concentrated.

This is illustrated in Figure 1. The circles in the left hand diagram depict the geographical distribution of the employees in forestry. The area of each circle is proportional to the region's share of Germany's forestry employees. The right hand part of Figure 1 illustrates the geographical distribution of employees in call centers. In both parts, the German map is depicted in different shades of grey. They indicate the density of overall employment, with the darkest greys in urban areas like Munich, Berlin, Cologne, Frankfurt and Hamburg.

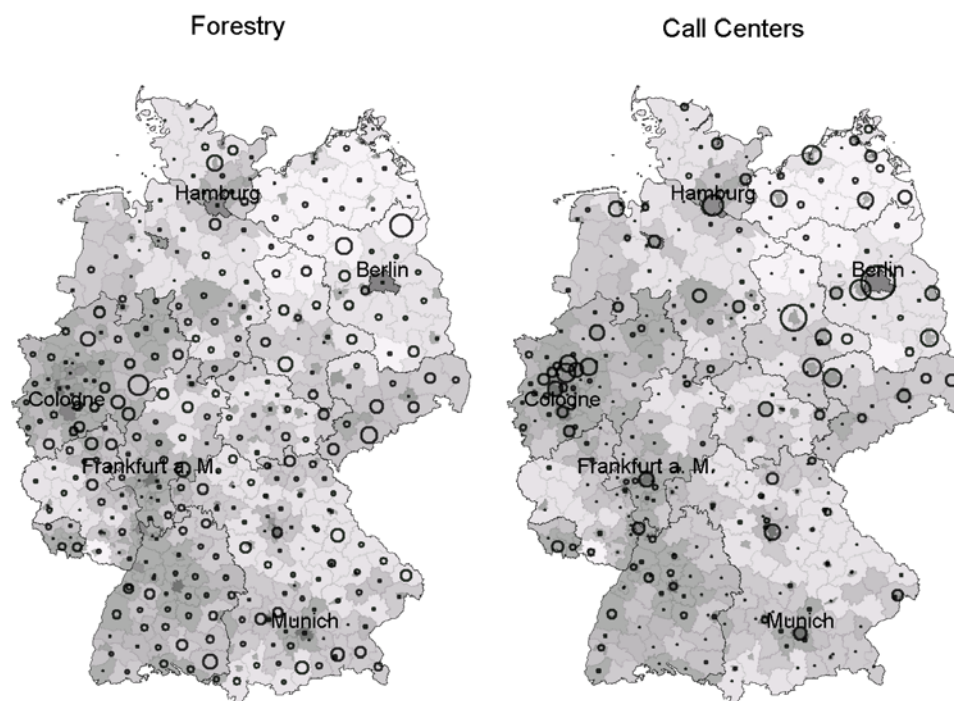


Figure 1: Geographical Distribution of Employees in the Industries *Forestry* and *Call Centers* in Germany in 2010.

Closer inspection of Figure 1 reveals an important difference between the two industries: call centers are overrepresented in urban areas, whereas forestry is overrepresented in rural areas, that is, the two industries exhibit different *types* of concentration. The existence of different types of concentration follows from a simple logical consideration. When, relative to overall employment, some industries are overrepresented in urban areas, then some other industries must exist that are underrepresented in urban areas, and therefore, overrepresented in rural areas. In other words, when some industries exhibit urban concentration (e.g., call centers), then some other industries must exhibit rural concentration (e.g., forestry).

Note that the Gini coefficient (and alternative measures such as the Theil index and the Krugman index) measure an industry's concentration relative to the geographical distribution of overall employment. The Gini coefficients of both, forestry and call centers are roughly 0.73, indicating that, relative to overall employment, both industries show the same degree of concentration. The visual impression, however, is different. Call centers look much more concentrated. In fact, using the Theil index instead of the Gini coefficient confirms this visual impression. However, the Krugman index yields the opposite result: forestry is more concentrated than call centers.

We can draw an important conclusion from the previous discussion: When two industries differ with respect to their *types of concentration*, then comparisons of their *degree of concentration* are questionable. To gauge the geographical concentration of industries, we have to know which industries belong to which type of concentration.

It is a serious failure of the existing measures of concentration that they cannot distinguish between different types of concentration. These measures simply ignore the problem and even rank industries that exhibit completely different types of concentration. A meaningful analysis of concentration patterns should start by first identifying each industry's type of concentration. Only then the analysis can proceed to measure and to compare the industries' degrees of concentration.

This paper is devoted to the first step, and the contribution that we make is threefold. First, we define and characterize seven different geographical archetypes. Second, we develop an intuitive and powerful statistical procedure that assigns each industry to one of the geographical archetypes. The third contribution is empirical. Using our new approach, we investigate the German industries' geographical archetypes. For this purpose we utilize a large administrative dataset with regionalized German employment data on 613 four-digit industries.

The paper is organized as follows. A brief review of the existing literature is provided in Section 2. 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 elaborated approach which we present in Section 5. The derivation of its statistical properties is provided in Section 6. We apply this approach to German employment data. The results are presented in Section 7. We conclude with a summary of our study’s findings.

2 Three Generations of Measures

Typically, measures of the degree of geographical concentration compare the industry’s geographical employment pattern to the geographical employment pattern of the general economy. Well known measures are the Gini coefficient, the Theil index, the 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 they also attempt to quantify an industry’s degree of concentration such that comparisons between industries 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 for each industry the regional employment (or some alternative measure of economic activity) is recorded.

Though simple to apply, the measures of the first generation exhibit 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 shows a significantly larger (lower) degree of concentration than the industry’s hypothetical random distribution, the industry should be tagged as geographically 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 measures. Ellison and Glaeser (1997) as well as Maurel and Sédillot (1999) propose second generation measures that are based on regionalized firm level data. Another second generation measure is the entropy approach of Brühlhart and Traeger (2005). It can be applied to regionalized data that contain no firm level information.

In some countries (e.g., France, Germany, U.K.) firm level data exist that contain not only the number of workers of each firm but also the firm’s precise geographical coordinates. With such geo-referenced firm level data at hand, distance-based measures of geographical concentration – measures of the third generation – can be applied. Just as 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 merely 50 square kilometers. The 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.

The strong interest in data sets with geo-referenced firm level data is justified. Such data can significantly improve the accuracy of measurement. A drawback of regionalized data is their dependence on the regions’ size and the course of their border. With geo-referenced data, this “modifiable area unit problem” (Openshaw and Taylor, 1979; Arbia, 1989) can be solved.

However, for the foreseeable future, regionalized instead of geo-referenced data will still be the rule rather than the exception. Therefore, improving the analysis of concentration when geo-referenced data are not available remains an important issue. A major strength of the approach suggested in this paper is that it works with regionalized data sets that neither contain firm level data nor information on distances.

3 Geographical Archetypes

First generation measures distinguish between two geographical archetypes: dispersion and concentration. Second and third generation measures add randomness as a third geographical archetype, taking a middle position between the former two archetypes. However, this tripartition is still insufficient, because within the broad category “concentration” different sub-types should be distinguished.

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 line in diagram (A) of Figure 2 depicts this distribution. The two spikes can be viewed as urban districts and the rest as rural districts. The area below the line is of

size 1. Therefore, the line shows the density of overall employment.

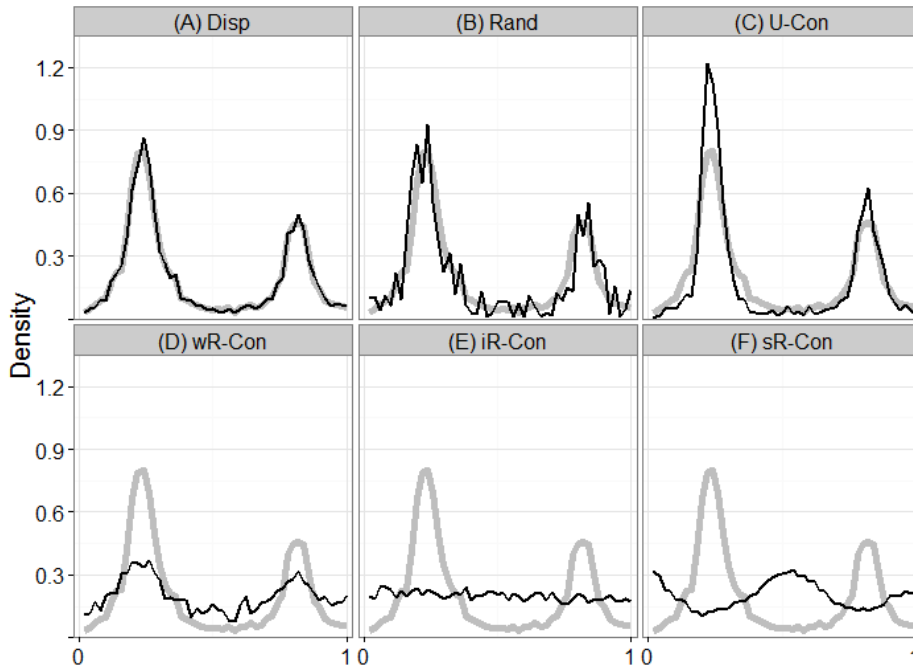


Figure 2: Different Geographic Archetypes.

Each diagram of Figure 2 depicts a different industry. The black lines capture the employment densities of the respective industries, the grey lines show the density of overall employment and, hence, do not vary.

Diagram (A) exhibits a situation in which industry A’s employment is almost perfectly positively correlated with overall employment. This geographical archetype is usually denoted as *dispersion* (*Disp*). Typically, basic services like restaurants or retail sales of bread and cakes will fit this type of industry.

However, basic service industries could also fit the geographical archetype depicted in diagram (B) of Figure 2. The diagram is similar to diagram (A), but the fluctuations of industry B’s employment around the overall employment are larger than that of 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*). This archetype has been emphasized in the work of Ellison and Glaeser (1997) and in many subsequent studies on the measurement of concentration. All of these studies distinguish between the three

archetypes dispersion, randomness, and concentration. The present study, however, argues that the category “concentration” is too wide since it can take on completely different sub-forms that need to be distinguished for a meaningful interpretation and comparison of industries.

Diagrams (C) to (F) depict four archetypes, each representing a different type of concentration. Diagram (C) shows again a positive correlation between the industry’s employment and overall employment. However, 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 the industry employment is positively correlated with the 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 *weak rural concentration (wR-Con)*. General practitioners or pharmacies could 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 as *intermediate rural concentration (iR-Con)*.

Diagram (F) depicts a situation in which the industry is grossly overrepresented in rural areas and grossly underrepresented in urban areas. As a result, a negative correlation between industry F’s employment and overall employment arises. This type of concentration we denote as *strong rural concentration (sR-Con)*. Industries like livestock farming are likely to exhibit *sR-Con*.

4 Assignment of Industries

Any comprehensive analysis of the geographical concentration of industries should proceed in two steps. In the first step, it should assign each industry to one of the geographical archetypes described in the previous section. Only in the second step, the analysis can attempt to rank the industries with respect to their degree of concentration. This paper is devoted to the first step.

We consider some country for which no firm level information is available and also no information on distances. Instead, we have regionalized employment data. That is, for each industry i ($i = A, B, \dots$) and each region r ($r = 1, 2, \dots, R$) we know the employment x_r^i . An industry’s total

employment is defined by

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

a region's overall employment by

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

and the country's overall employment by

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

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

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

Note that geographically very large rural regions can have relatively large S_r -values. Therefore, a reliable distinction between urban and rural areas requires a better indicator than the S_r -values. Fortunately, it is mostly easy to obtain the regions' geographical size, a_r (measured in square kilometers). 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 . The larger a region's E_r -value, the more urban this region. Correspondingly,

$$e_r^i = \frac{s_r^i}{a_r} \tag{2}$$

denotes the employment density of industry i in region r . 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, $\sum_r a_r E_r = 1$ and $\sum_r a_r e_r^i = 1$.

Several refinements are conceivable. For example, in the concentration analysis of some industry i , we could subtract x_r^i from x_r to get 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,

$$E_r^{(i)} = \frac{x_r^{-i}}{\left(\sum_{s=1}^R x_s^{-i}\right)} a_r. \quad (3)$$

For industries with small employment shares, x^i/x , the changes are negligible. However, for industries with a large employment share it is possible that the refinement matters.

How can we utilize the computed densities e_r^i and E_r for our assignment of industries to geographical archetypes? When some industry i is characterized by the geographical archetype *Disp*, then for every region r the data should yield $e_r^i \approx E_r$. This is illustrated in diagram (A) of Figure 3. The diagram corresponds to diagram (A) of Figure 2. We simply subdivided the “road” of Figure 2 into 50 equally small portions, each representing one region ($R = 50$). Each point in the scatterplot of diagram (A) of Figure 3 represents one region. The coordinates of each point (region) are given by its data pair (E_r, e_r^i) . With the geographical archetype *Disp*, the points are located very close to the 45°-degree line.

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 3. Again this diagram corresponds to its counterpart in Figure 2.

Industries that exhibit some type of concentration generate e_r^i -values that systematically deviate from their corresponding E_r -values. Urban concentration (*U-Con*) implies that, relative to overall employment, the employment of industry i should be 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 weak rural concentration (*wR-Con*) the opposite relationship holds, see diagram (D). The scatterplot of diagram (E) depicts intermediate rural concentration (*iR-Con*), that is, 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 strong rural concentration (*sR-Con*).

It is interesting to compare the archetype *U-Con* depicted in diagram (C) to the archetypes *Rand* and *sR-Con* depicted in diagrams (B) and (F). It turns out that the concentration archetype *U-Con* can have more in common with the archetype *Rand* than with the concentration archetype *sR-Con*. This observation reaffirms our claim that an industry’s concentration can take very different forms and that it is necessary to distinguish between these forms.

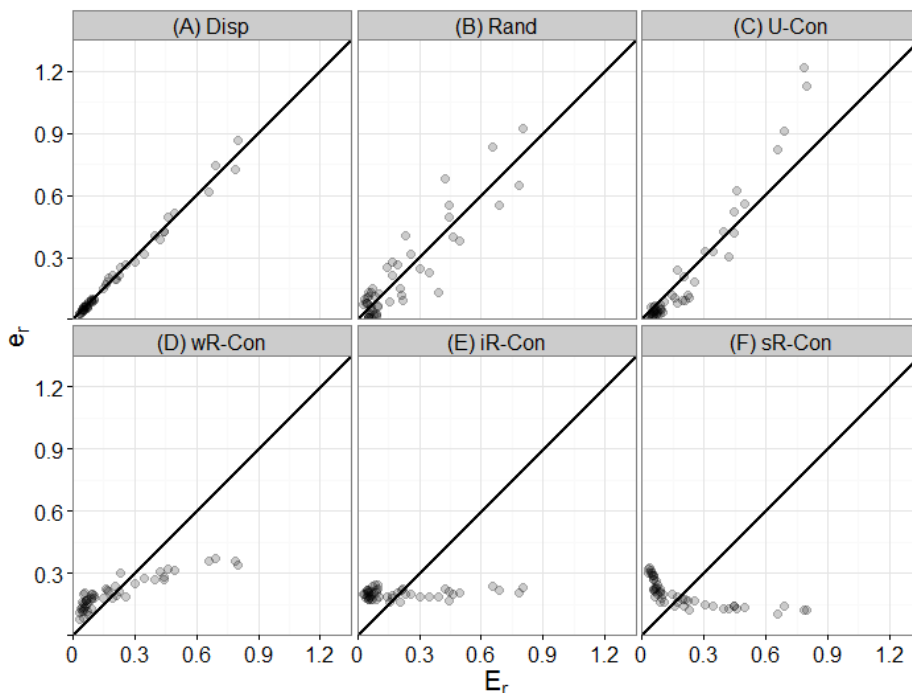


Figure 3: Scatterplots of Geographical Archetypes.

Each of the six scatterplots (A) to (F) has its own characteristic point pattern. Therefore, it should be possible to infer from an industry’s scatterplot its geographical archetype. For example, when an industry has a scatterplot resembling that of diagram (C), we can conclude that this industry is characterized by *U-Con*.

Figure 3 suggests to run for every industry a regression that parameterizes the different shapes of the lines fitted in the scatterplots. Using the coefficients of that regression, the industry can be assigned to the respective geographical archetype. For example, when the scatterplot of an industry generates a regression line with a slope greater than unity, as in diagram (C), the industry could be assigned to the archetype *U-Con*.

This line fitting approach works fine for the artificial data depicted in Figure 3. However, real world data will rarely generate “well behaved” scatterplots like those of Figure 3, because very few industries are present in all regions. For example, in the German employment data that we use in Section 7, almost half of the industries are present in less than half of the regions (see Figure 7 in Section 7). In other words, in diagrams like those of Figure 3 a substantial share of points is located on the horizontal axis. Therefore, the

assignment of industries to geographical archetypes is not accomplished by a simple line fitting exercise, but requires a more sophisticated approach.

5 Assignment of Real World Industries

One might be tempted to overcome 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 carry important information for the distinction between the archetypes. For example, for a *U-Con* industry one would expect that all points with $e_r^i = 0$ are located close to $E_r = 0$, that is, in rural areas, whereas with *sR-Con* such points could well be located at larger E_r -values. Consider an industry that has many such points at larger E_r -values (contradicting *U-Con*), but the other points exhibit a pattern like in diagram (C) of Figure 3 (supporting *U-Con*). If we deleted the points on the horizontal axis and then assigned the industry to some archetype, we would wrongly assign it to *U-Con*. To avoid such misassignments, the points on the horizontal axis must not be deleted.

A better approach to deal with the absence problem are regression techniques specifically designed for censored data (e.g., Tobit-family regressions). However, for many real world industries the share of censored data is so large (again, see Figure 7 in Section 7) that the regression approach cannot reliably identify an industry’s geographical archetype. Therefore, we propose a completely different approach that is based on the industry’s Goodman-Kruskal 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 $(E_r - E_s) \cdot (e_r^i - e_s^i) > 0$. Pairs of regions with $(E_r - E_s) \cdot (e_r^i - e_s^i) < 0$ are discordant. When $e_r^i = e_s^i$ or $E_r = E_s$, then the pair of regions is neither concordant nor discordant. Let C_I^i denote the proportion of concordant pairs and D_I^i the proportion of discordant pairs. Then the Goodman-Kruskal coefficient of industry i is defined as

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

with $C_I^i + D_I^i \leq 1$.

Figure 3 reveals that the archetype *sR-Con* corresponds to a negative coefficient γ_I^i , the archetype *iR-Con* to a coefficient γ_I^i close to 0, and the 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 that 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 ,

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

where C_{II}^i denotes the proportion of concordant pairs, i.e. pairs where $(E_r - E_s) \cdot (e_r^i/E_r - e_s^i/E_s) > 0$. Correspondingly, 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 Appendix).

Figure 4 shows why γ_{II}^i is a suitable instrument to distinguish between the four archetypes *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 *iR-Con*) to a negative coefficient.

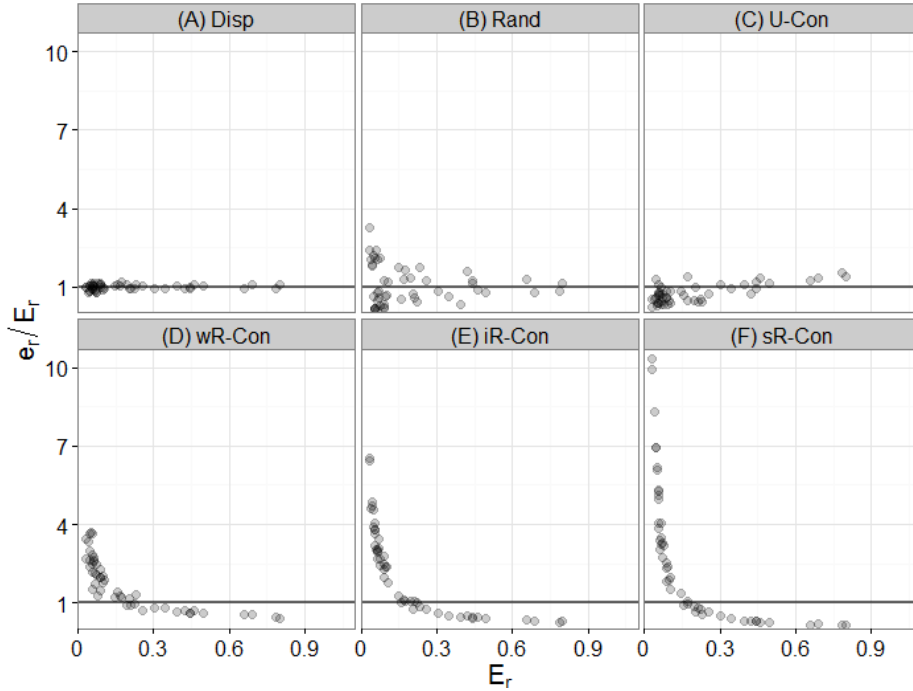


Figure 4: Geographical Archetypes from Another Perspective.

How can we distinguish between the archetypes *Disp* and *Rand*? We can overcome this indeterminacy by taking into account the statistical significance of the γ_I^i - and γ_{II}^i -values. For this purpose we derive for every industry

not only its γ_I^i - and γ_{II}^i -values, but also its bivariate confidence region. The basic idea of such confidence regions is illustrated in Figure 5. The horizontal axis depicts the value of $\gamma_I^i = \gamma(E_r, e_r^i)$ and the vertical axis the value of $\gamma_{II}^i = \gamma(E_r, e_r^i/E_r)$. A confidence region is an elliptic area with centre $(\gamma_I^i, \gamma_{II}^i)$. Figure 5 depicts the confidence regions of the same six industries that were displayed also in Figures 2, 3, and 4. Added to these six industries is a seventh confidence region (labelled with *Mis-Con*) which will be explained before long. Of course, the precise shape of a confidence region depends on the number of observations, R , and the significance level. The confidence regions of Figure 5 were computed on a 5 percent significance level and the number of observations was $R = 50$. The formula for computing an industry's confidence region is derived in Section 6.

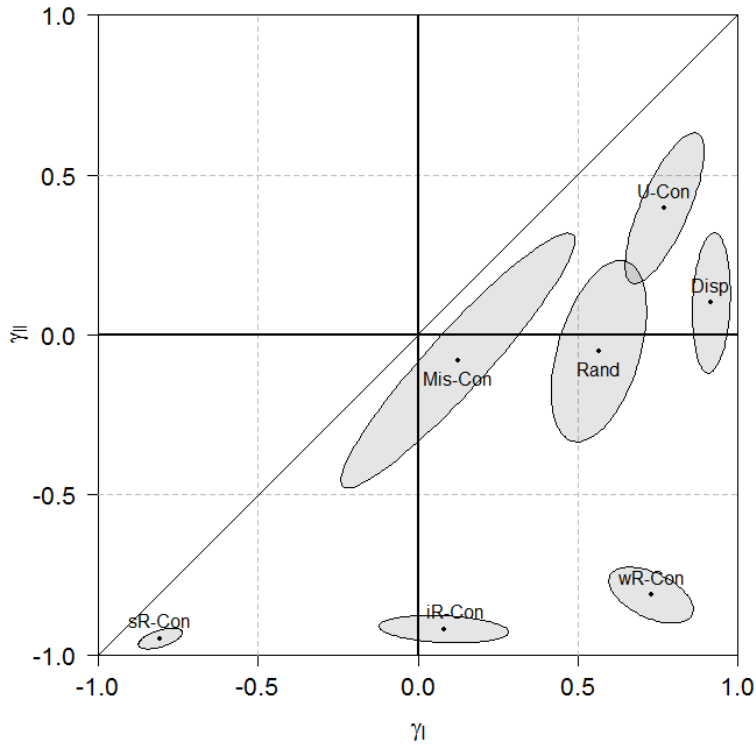


Figure 5: Assignment by Confidence Regions.

Once we know an industry's confidence region, we can assign the industry to one of the geographical archetypes. The assignment rule is straightforward and illustrated in Figure 5.

U-Con: The confidence region is completely above the horizontal axis and completely to the right of the vertical axis.

If the confidence region is completely below the horizontal axis, the industry is assigned to one of the three forms of rural concentration.

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

iR-Con: The confidence region is completely below the horizontal axis overlaps with the vertical axis.

wR-Con: The confidence region is completely below the horizontal axis and completely 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*.

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

Rand: The confidence region overlaps with the horizontal axis and is completely to the right of the vertical axis, but not completely 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 are significantly different from 0. Since a large γ_I^i -value is a signal of dispersion, a small γ_I^i -value is a signal for a 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)* and add to the assignment rule the following part:

Mis-Con: The confidence region overlaps with both, the horizontal and the vertical axis.

In total, the assignment rule distinguishes between seven geographical archetypes, namely five different types of concentration and the archetypes *Disp* and *Rand*. The distinction between the latter two archetypes relies on the choice of some limiting vertical line. In our assignment rule this vertical line was set at $\gamma_I = 0.5$. Notice that in Figure 5 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 2). The confidence region of industry A is completely

to the right of the vertical line at $\gamma_I = 0.8$. From simulations we know that even industries with confidence regions that are completely to the right of $\gamma_I = 0.5$ look very much like dispersed industries. Therefore, we propose to use the vertical line at $\gamma_I = 0.5$ as the reference for the archetype *Disp*. Of course, this is a somewhat arbitrary choice and researchers are free to choose a different reference.

Usually, the geographical size of the regions, a_r , is known. However, what can be achieved, if this information is not available? Is it still possible to identify the industries' geographical archetypes? Fortunately, our approach works also with employment shares ($S_r = x_r/x$ and $s_r^i = x_r^i/x^i$) instead of employment densities (E_r and e_r^i). The coefficients are $\gamma_I^i(S_r, s_r^i)$ and $\gamma_{II}^i(S_r, s_r^i/S_r)$ and the confidence region is still centred around $(\gamma_I^i, \gamma_{II}^i)$. The assignment rule remains unaltered. It generates reliable assignments unless the variance and the range in the geographical size of the regions, a_r , are extremely large.

6 Derivation of Bivariate Confidence Regions

The confidence regions depicted in Figure 5 were computed from a formula that we now derive. The regional observations (E_r, e_r^i) , $r = 1, \dots, R$, may be interpreted as a random sample from a superpopulation (E, e^i) .¹ Let (E_1, e_1^i) and (E_2, e_2^i) be independent draws from (E, e^i) and define the following probabilities of concordances and discordances:

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

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

$$\Gamma_I^i(E, e^i) = \frac{\pi_{C,I}^i - \pi_{D,I}^i}{\pi_{C,I}^i + \pi_{D,I}^i} \quad (6)$$

$$\Gamma_{II}^i(E, e^i) = \frac{\pi_{C,II}^i - \pi_{D,II}^i}{\pi_{C,II}^i + \pi_{D,II}^i} \quad (7)$$

¹See Särndal, Swensson and Wretman (2003), chap. 14.5, for the concept of superpopulations.

of the superpopulation.

In order to construct joint confidence intervals for Γ_I^i and Γ_{II}^i , we draw on the asymptotic theory for multivariate U -statistics and the delta method. As shown in Hoeffding (1948) and Kowalski and Tu (2008), the proportions of concordances and discordances are asymptotically normally distributed as $R \rightarrow \infty$,

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

with a covariance matrix Σ that can be estimated consistently by $\hat{\Sigma}$ from the data (see the appendix for details). 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 given by

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

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

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} : \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \gamma_I^i \\ \gamma_{II}^i \end{bmatrix} \right)' [J\hat{\Sigma}J/R]^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \gamma_I^i \\ \gamma_{II}^i \end{bmatrix} \right) \leq q_{1-\alpha} \right\} \quad (8)$$

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

7 Geographical Distribution of German Industries

We apply our approach to regionalized German employment data from 2010. They are provided by the *Institute for Employment Research IAB* at the *Bundesagentur für Arbeit*. The data contain the complete full-time employed population 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 outperforms survey data by far.

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. On the four-digit level, the WZ 2008 distinguishes between $I = 613$ different industries. In 2010, Germany was partitioned into $R = 412$ administrative NUTS 3 regions, 102 of which are cities. The size of each region, a_r , was computed from freely available online data of the *Bundesamt für Kartographie and Geodäsie*.

For every four-digit industry i ($i = 1, \dots, 613$) and every region r ($r = 1, \dots, 412$) we know the employment, x_r^i . From these numbers, we computed for every region its overall employment share, $S_r = x_r/x$, and its overall employment density, $E_r = S_r/a_r$. We also calculated the industries’ overall employment shares, x^i/x . The largest share is below 1 percent. Nevertheless, we used the refined formula (3). 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 kilometer in Munich ($E_r = 7.94470 \times 10^{-5}$). The histogram of the employment densities, E_r , depicted in Figure 6, reveals a very asymmetric distribution.

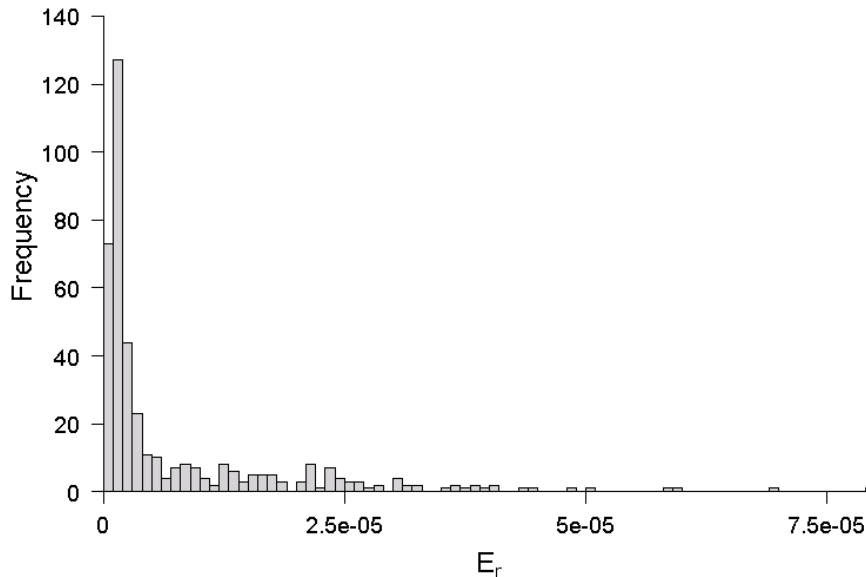


Figure 6: Histogram of the Employment Densities E_r .

Furthermore, for every industry i and every region r , formula (2) gave us the industry’s employment density, e_r^i . For each industry we computed

the share of regions with $e_r^i > 0$ and denoted it by z^i . Figure 7 depicts these shares in an empirical cumulative distribution function with z^i on the horizontal axis and the cumulated number of industries (that is, the industries are not weighted by their employment x^i) on the vertical axis. The figure shows that roughly 42 percent of the 613 industries have a share z^i below 50 percent. Only 13 percent of the industries are present in all regions ($z^i = 1$). These results reconfirm the aforementioned “absence problem” in real world employment data. This problem prompted us to discard the regression approach and to utilize the Goodman-Kruskal coefficients.

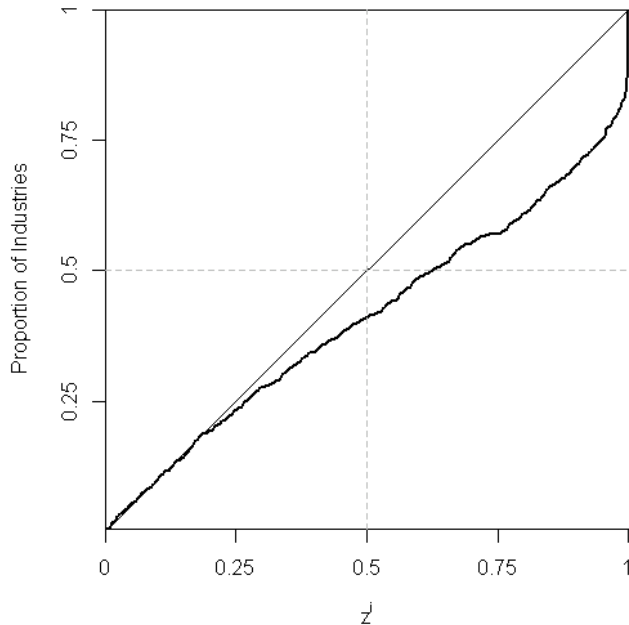


Figure 7: Visualization of the “Absence-Problem”.

From our data we also computed the location quotients, e_r^i/E_r , and each industry’s 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 confidence region (8). Then we compared each industry’s confidence region to the values $\gamma_I = 0$ (vertical axis of Figure 5), $\gamma_{II} = 0$ (horizontal axis of Figure 5), and $\gamma_I = 0.5$ (vertical line at position $\gamma_I = 0.5$). Following the rule described in Section 5, we were able to assign 606 of the 613 industries to one of the seven geographical archetypes.²

²Seven industries (e.g., “raising camels”, “growing of sugar cane”) could not be assigned to an archetype, because the respective industry was present in less than five of the 412 regions and the computation of confidence regions requires a presence in at least five regions. The total employment of these seven industries was less than 200 employees.

Figure 8 shows for every industry i not only the values of its two Goodman-Kruskal coefficients, but also its geographical archetype. 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 . To distinguish between different geographical archetypes, we use different symbols. Empty symbols stand for industries with rural concentration, with circles indicating the archetype $sR-Con$, triangles indicating $iR-Con$, and quadrats indicating $wR-Con$. The crosses symbolize the archetype $Mis-Con$, and the filled symbols stand for the archetypes $Rand$ (filled circles), $Disp$ (filled triangles), and $U-Con$ (filled quadrats).

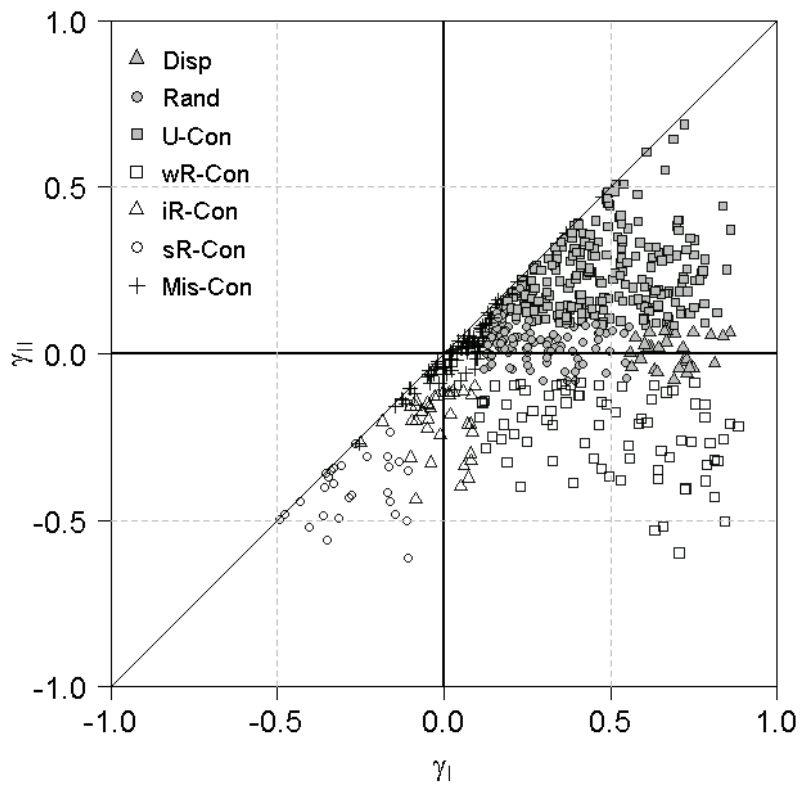


Figure 8: The Geographical Archetypes of German Industries.

28 of the 613 industries (about 4.6 percent) were assigned to strong rural concentration ($sR-Con$). These industries represent merely 0.8 percent of total employment. $sR-Con$ is dominated by the agricultural sector (e.g., growing grain; raising cattle, pigs and poultry; mixed agriculture).

The agricultural sector (e.g., growing of vegetables and potatoes) plays an important role also in the geographical archetype intermediate rural concentration ($iR-Con$). However, several food processing industries (e.g., pro-

cessing of fish, production of juices, processing of milk) and a couple of other industries (e.g., remolding tires, production of spirituous beverages) are also classified as *iR-Con*. This archetype contains 35 of the 613 industries, representing an employment share of 1.7 percent.

78 industries, or 28.0 percent of total employment, were assigned to weak 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 *iR-Con*. Many industries of the construction sector belong to this archetype (e.g., construction of buildings and roads, electrical installation, roofing, tiling, plastering). Furthermore we find in this archetype 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). Also some manufacturing industries are assigned to *wR-Con*. 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 can be found in the geographical archetype urban concentration (*U-Con*). 237 industries are assigned to this archetype. They cover 44.3 percent of total employment. The archetype's composition is extremely heterogeneous, ranging from manufacturing, wholesale, and retail sale 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, retail of fruits and vegetables, retail of cosmetic products and toiletries, restaurants, nursery schools, and churches). 29 industries with a combined employment share of 11.3 percent are assigned to this archetype.

The archetype randomness (*Rand*) comprises 106 industries with a combined employment share of 11.9 percent. Manufacturing has the largest share within this archetype. However, wholesale (e.g., sugar, sweets, bakery products, flowers, fruits, and vegetables), 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 2.0 percent of total employment are assigned to the archetype miscellaneous concentration (*Mis-Con*). Since 93 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 of 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 few industries from agriculture (e.g., growing of grapes) and some service industries, many of which are somehow related to shipping (e.g., repair of ships, inland navigation, coastal shipping).

8 Concluding Remarks

The measurement of 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, it would be a rather straightforward regression exercise by which we could assign the industries to the most appropriate archetype. In the real world, however, most industries are present only in some of the regions. Therefore, we develop a new statistical approach that can deal with such data.

Our approach is based on two Goodman-Kruskal correlation coefficients. For each industry these two coefficients together with a confidence region are computed. Depending on the position and size of the confidence region, the industry can be assigned to one geographical archetype, and each assignment comes with a statistical significance. For the reliability of the assignment it is useful, but not essential, to know the geographical size of the regions.

We applied our approach to an extremely 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 identified clear differences between the geographical patterns of agriculture, manufacturing, retail sale, wholesale, basic services, and other services.

It is another virtue of our approach that it works 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. How should the industries be assigned to the various geographical archetypes in that case? Of course, ignoring the information on distances and aggregating the firm level data to regionalized data, our confidence-region-approach could still be utilized. Such a procedure, however, would be unsatisfactory, since valuable information would be wasted. Therefore, in future research, one could try to develop an assignment approach that makes efficient use of geo-referenced firm level data.

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.
- Brühlhart, M., R. Traeger, 2005. An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics*, 35, 597-624.
- Combes, P.-P., T. Mayer, J.-F. Thisse, 2008. *Economic Geography*. Princeton University Press: Princeton (New Jersey).
- 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.
- 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.
- 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.

Appendix

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

Consider some industry i . For every region with $s_r^i = 0$ we get $e_r^i = 0$ 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 : $E_r < E_s$ and $e_r^i/E_r < e_s^i/E_s$. Therefore

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

This says that every pair of regions that is concordant with respect to the two variables E_r and e_r^i/E_r is also concordant with respect to the two variables E_r and e_r^i .

Now consider some 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 $e_s^i = 0$, then this discordance 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 $E_r \ll E_s$, then we 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 get $C_I^i \geq C_{II}^i$ 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 few discordant pairs exist that can turn into concordant pairs with respect to E_r and e_r^i . A large negative value (in absolute terms) indicates that many discordant pairs exist that can turn into concordant pairs with respect to E_r and e_r^i .

Asymptotics of concordance and discordance proportions

In this appendix we reproduce results of Hoeffding (1948) for the specific cases considered in this paper. For ease of notation we drop the industry superscripts and let $X_r = (E_r, e_r)'$. The univariate statistic $\pi_{C,I}$ is estimable by a U -statistic of degree 2 since

$$E(\varphi_C(X_1, X_2)) = \pi_{C,I}$$

where the kernel φ is defined as

$$\varphi_C(X_1, X_2) = 1(E_1 < E_2, e_1 < e_2) + 1(E_1 > E_2, e_1 > e_2)$$

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

$$\varphi_D(X_1, X_2) = 1(E_1 < E_2, e_1 > e_2) + 1(E_1 > E_2, e_1 < e_2)$$

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

$$C_I = \binom{R}{2}^{-1} \sum_{r < s} \varphi_C(X_r, X_s)$$

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 large R , the variance is approximately

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

with

$$\zeta = E(\varphi_{C,1}^2(X_1)) - \pi_{C,I}$$

and

$$\varphi_{C,1}(x_1) = E(\varphi_C(x_1, X_2)).$$

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

$$\hat{\varphi}_{C,1}(X_r) = \frac{1}{R-1} \sum_{s=1}^R \varphi_C(X_r, X_s).$$

Then

$$\hat{\zeta} = \frac{1}{R} \sum_{r=1}^R \left(\frac{1}{R-1} \sum_{s=1}^R \varphi_C(X_r, X_s) \right)^2 - (C_I)^2$$

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 $\text{Cov}(C_I, D_I)$ can be estimated by

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

with

$$\hat{\zeta}^{C,D} = \frac{1}{R} \sum_{r=1}^R \left(\frac{1}{R-1} \sum_{s=1}^R \varphi_C(X_r, X_s) \right) \left(\frac{1}{R-1} \sum_{s=1}^R \varphi_D(X_r, X_s) \right) - C_I D_I.$$