PAPERS on Economics & Evolution

0516

**The Regional Industry-size Distribution -
An Analysis of all Types of Industries in Germany**

by

Thomas Brenner

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

The Regional Industry-size Distribution – An Analysis of all Types of Industries in Germany

ABSTRACT. This paper studies the frequency of observing a certain number of firms or employees in a region for a given industry. Various predictions for these frequencies are deduced from theoretical considerations. Then, the empirical distributions of 198 industries in Germany are analysed. It is found that different kinds of industries show quite different distributions.

KEYWORDS: industry study, spatial distribution, local industrial clusters.
JEL classification: C12, L60, R12

1. Introduction

The spatial distribution of industries has attracted much attention in recent years. This was mainly triggered by the interest of researchers and policy makers in local clusters, industrial districts and the geographic concentration of industries. Research here has mainly focused on questions such as how and why local clusters emerge, why they are economically successful, how this changes over time, and how policies can influence the evolution of local clusters. The literature on these topics has grown tremendously in recent years.

The fact that local clusters and industrial districts exist, implies that industries are not uniformly distributed in space. This means that, by analysing one industry, some regions contain numerous firms while many other regions contain none or a small number of firms. Consequently, firms of one industry should not be expected to be randomly and independently distributed in space. This leads to the question of what forces drive the manner in which firms are distributed across space. This question has been only addressed in a few works in the literature.

First, New Economic Geography models can be found in the literature (see, e.g., Krugman 1996, Allen 1997 and Keilbach 2000). These approaches explicitly model an industry's spatial evolution. They show that by including positive and negative local externalities geographic concentration might be obtained. Through this they prove that local externalities might indeed by responsible for the existence of local clusters and industrial districts, as often claimed in the literature. However, they do not analyse in detail the characteristics of the spatial distribution that results from their models.

Second, there are also some works that identify all local clusters within a country (see Sforzi 1990, Isaksen 1996, Paniccia 1998, Braunerhjelm & Carlsson 1999, Brenner 2003, and Sternberg & Litzenberger 2004). However, except for Brenner's approach (2003), these methods assume a threshold for the number of firms or employees within a region and industry. All regions that contain a higher number and satisfy some additional conditions are declared to contain local clusters. They are not concerned with the explicit spatial distribution of industries. Brenner

(2003) uses an approach similar to the one used here, however it does not go into the details about the assumed distributions.

Finally, there is a small strand of literature that deals explicitely with the spatial distribution of industries. This literature presently consists of two papers and some related works (Ellison & Glaeser 1997 and Bottazzi, Dosi, Fagiolo and Secchi 2005). Ellison and Glaeser (1997) start from the assumption that firms are located randomly and independently in space. Then, they define an index that approximates the deviation from such a random distribution and show that firms in most industries are actually not randomly and independently distributed. Bottazzi, Dosi, Fagiolo and Secchi (2005) start from the assumption that start-ups when deciding where to locate taken into account the location of other firms in the same industry. They build a model and fit the parameters of the model to empirical data.

This paper adds to the existing literature by analysing the spatial distribution of industries in detail. We define the regional industry-size distribution as the distribution that assigns to each quantity of industry-specific activity a probability that it is observed in a randomly chosen region. The amount of industry-specific activity in a region is measured by either the number of firms or employees. This is done in relative terms, in relation to the total number of employees in the region. It is evident that larger regions should, on average, contain more economic activity.

In order to study the regional industry-size distribution, some plausible distribution shapes are deduced from theoretical considerations. The adequateness of these shapes is then empirically studied using data on 219 industries (3-digit level). The aim is to examine whether the postulated shapes are able to sufficiently describe the empirical situation and whether there are differences between industries. Hence, the paper goes beyond the approach by Ellison and Glaeser (1997) because it tries to find a mathematical formulation for the spatial distribution of industries instead of only proving that a random distribution is inadequate. It goes beyond the approach by Bottazzi, Dosi, Fagiolo and Secchi (2005) because it checks the adequateness of the distribution that is fitted to the empirical data.

The paper proceeds as follows. In Section 2 various alternative distributions are deduced

from theoretical considerations. The empirical data used and method applied for checking them and fitting the various theoretical distributions are presented in Section 3. In Section 4 this method is applied to 219 industries in Germany and the results are discussed. Section 5 concludes.

2. Theoretical considerations and predictions

It is not the intention of this paper to develop a model that explains the spatial distribution of industries. To develop such a model is a complex task that goes beyond the scope of this paper. Here, we intend to analyse the shape of regional industry-size distribution and compare the obtained shapes between different industries. For this analysis two different approaches can be taken: a non-parametric and a parametric approach. In order to compare and classify different industries's shapes, a parametric approach is more suitable. This implies that we have to set up some possible functional forms that can be tested. This section aims at identifying some plausible functional forms for regional industry-size distribution, which are derived from theoretical considerations.

2.1. Binomial distribution

Although Ellison and Glaeser (1997) have finally developed a somewhat different index, their initial argument is based on the assumption that firms are randomly and independently located in space. Ellison and Glaeser aim and succeed in rejecting this hypothesis for most industries. Nevertheless, such an assumption seems to be a natural starting point. It cannot be excluded that in some industries firms locate in a space almost independently of each other. If furthermore, there are no unequally distributed local resources involved in the location decision of these firms, we would expect firms to be randomly distributed in space.

In order to obtain a prediction for the regional industry-size distribution, we have to calculate how random firm location impacts the likelihood of observing a certain number of firms in a certain location. Let us denote the number of firms in an industry i by f_i and consider one region r. The likelihood of any of these f_i firms in the considered industry to be located in

region r depends on the size of the region. Even if firms were located randomly, large regions are more likely to be chosen. We denote the size of a region r by s_r, where s_r has to be defined such that the following identity holds:

$$\sum_r s_r = 1 \ . \tag{2.1}$$

This implies that the probability of each firm locating in region r is given by s_r. Consequently, the number of firms $f_{i,r}$ in industry i and region r is binomially distributed:

$$P(f_{i,r} = f) = Bnm(f_i, f, s_r) = \frac{f_i!}{f! \cdot (f_i - f)!} \cdot s_r^f \cdot (1 - s_r)^{f_i - f} \ . \tag{2.2}$$

Besides the distribution firm numbers, distribution of employees is also studied here. The above arguments apply much less to the number of employees. It seems unlikely that employees locate in space independently from each other because they work together in firms. Nevertheless, in order to use the same approach for firms and employees, the binomial distribution is also used for employees:

$$P(e_{i,r} = e) = Bnm(e_i, e, s_r) = \frac{e_i!}{e! \cdot (e_i - e)!} \cdot s_r^e \cdot (1 - s_r)^{e_i - e} \ . \tag{2.3}$$

e_i denotes the total number of employees in industry i and $e_{i,r}$ denotes the number of employees in industry i that are located in region r. The shape of the binomial distribution is depicted in Figure 3.

2.2. Exponential distribution

The above model assumes that the probability of firms to locate in a certain region depends only on regional size. However, in reality other local factors are involved. Local resources, such as human capital, natural resources, and infrastructure are important. Ellison and Glaeser (1999) considered empirical data on such local resources in their analysis. Here, a different approach is used. It is unclear as to what kind of local resources are essential as their type and importance differs between industries. In the approach of Ellison and Glaeser (1999) one does not know whether all important local factors are considered. Hence, local resources are

not directly included in the approach taken here. This paper focuses rather on the shape of regional industry-size distribution.

However, empirical knowledge about local resources can be used to support the discussion of potential shapes of the regional industry-size distribution. One local resource that all economists agree is important is human capital. Part of this is given by the number of students trained in a region. The distribution of student numbers per inhabitant is depicted in Figure 1 for administrative districts in Germany. It is evident that this distribution does not have the form of a Binomial distribution. Instead, the distribution in Figure 1 seems to have the shape of an exponentially decreasing function (which fits the data better than a hyperbolical function).

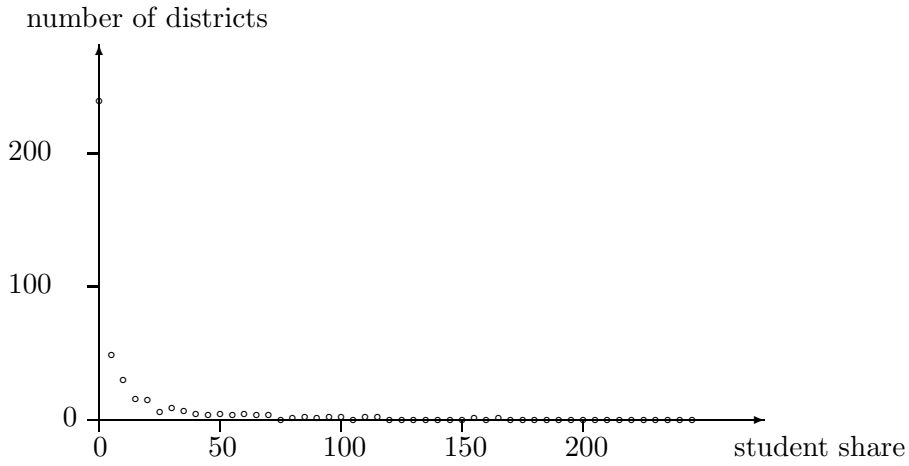

Figure 1: The horizontal axis depicts the number of student per 1000 inhabitants, while the vertical axis counts the number of administrative districts that approximately contain this number of students.

We use this as an argument that the exponential function, given by

$$P(m) = (1 - \xi) \cdot \xi^m \tag{2.4}$$

where m is the number of firms or employees and ξ is a parameter that might appropriately describe regional industry-size distribution. However, we do not claim that the number of students is the decisive factor for the location of an industry if we find the exponential function to be a suitable representation of region industry-size distribution. Other local factors might be similarly distributed in space. Furthermore, dependencies between firm locations in the same industry might have similar effects. Bottazzi, Dosi, Fagiolo, and Secchi (2005) find a similar functional form for specific parameter values of their firm location model. Our only task is to find plausible shapes of regional industry-size distribution that are then empirically tested.

2.3. Erlang-n-distribution

Bottazzi, Dosi, Fagiolo and Secchi (2005) find quite different distribution shapes in their analysis. Besides a nearly Binomial distribution and two distributions similar to an exponential distribution, there is a distribution that starts from a low value, which increases first and then decreases similar to an exponential function (Bottazzi et. al. 2005, Figure 2).

A similar distribution is also observed for the number of business service firms. These firms are also usually seen as a local factor that influences the location of manufacturing firms. Their distribution among the 441 regions in Germany is presented in Figure 2. Initially this distribution could also be interpreted as a Binomial distribution, however, the latter is rejected by the Kolmogorov-Smirnov test.

Therefore another function is proposed here: the Erlang-n-distribution ($n = 2$). Mathematically it is given by

$$P(m) = \frac{(1 - \xi)^2}{\xi} \cdot m \cdot \xi^m \tag{2.5}$$

where m is the number of firms found in a region and ξ is the parameter of this distribution.

The best fit to the distribution in Figure 2 is reached by $\xi = 0.99075$ which is not rejected by the Kolmogorov-Smirnov test.

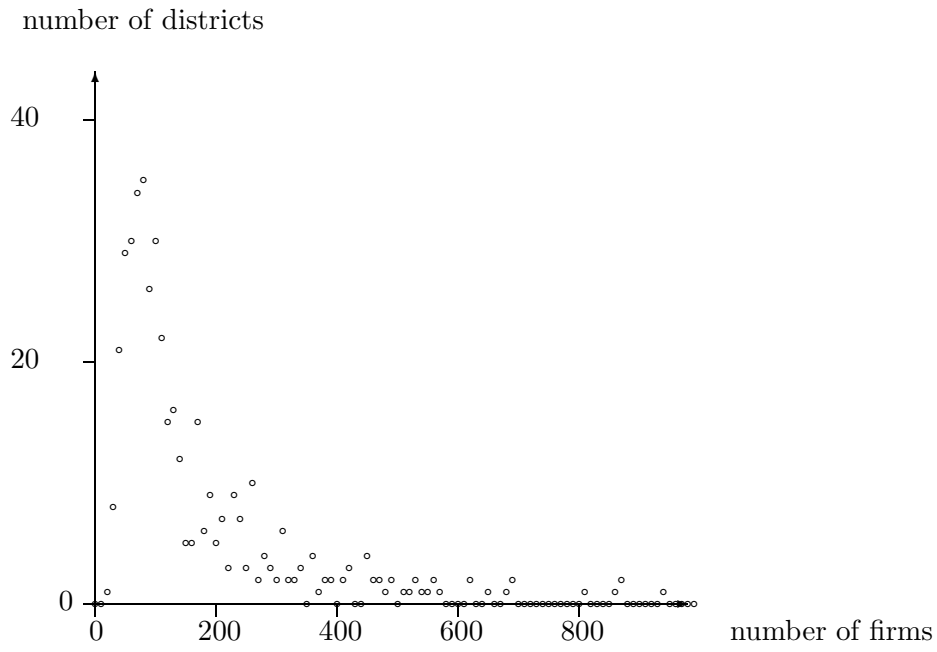

Figure 2: For each number of service firms that provide services to other firms, e.g. legal firms, marketing firms, PR consultants and so on, (horizontal axis) the number of administrative districts in Germany (vertical axis) that contain approximately this number of firms is depicted.

Again the above arguments should not be interpreted as claiming that specific factors or dependencies cause specific distributions. The idea is rather to find some distributions with a plausible explanation that they could occur. We could continue the process above and find more than the three functions that have been obtained so far. However, these three functions are plausible and represent various shapes that can be expected (see Figure 3). Therefore, only one specific shape is added.

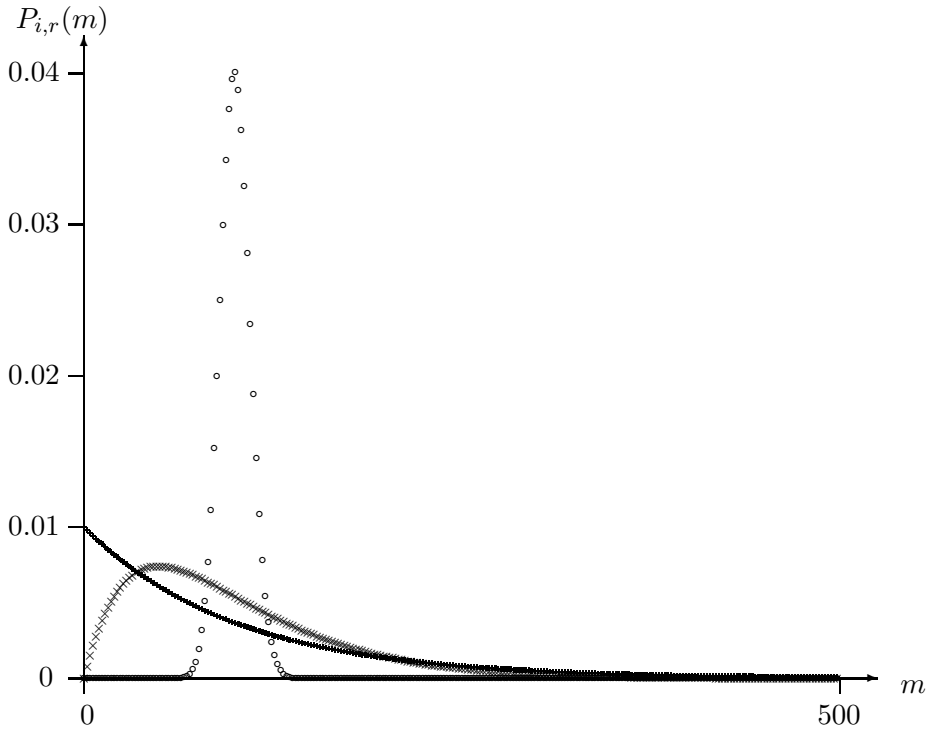

Figure 3: Theoretically predicted regional industry-size distribution for the number of firms or employees according to a Binomial distribution (o), an exponential distribution (+), and a Erlang-n-distribution (×). The parameters of the distributions are chosen such that they lead to the same average number.

2.4. Clustering

Again, we do not intend to find an exact modelling of local clusters. What we claim is that if local clusters exist, we should observe a small number of regions that contain a very high activity in the considered industry. The observed activity in these regions should be far beyond the activity in other regions (see Brenner 2001 and 2004 for a theoretical analysis) so that they do not fall into a regional industry-size distribution as described by the above functions. Therefore, it is argued (Brenner 2004) that the regional industry-size distribution should contain a second peak.

This second peak can be modelled in a variety of ways. It is, however, clear that an additional function is needed to create such a second peak, which represents the number of firms or employees substantially above the average number predicted by the other functions. The Erlang-n-distribution ($n = 2$) is perfectly suited for describing such a second peak because it starts from a value of zero and increases until it reaches its maximal value and then decreases exponentially.

Therefore, in order to model the second peak, we move the Erlang-n-distribution to higher values of the number m of firms or employees. Mathematically this can be done by

$$P(m) = \begin{cases} \frac{(1-\xi)^2}{\xi} \cdot (m - m_0) \cdot \xi^{m-m_0} & \text{if} \quad m > m_0 \\ 0 & \text{if} \quad m \leq m_0 \end{cases} \tag{2.6}$$

where ξ is a parameter and m_0 denotes the number of firms or employees at which the function starts.

2.5. Complete function

Four different functions have been formulated above that could describe empirical regional industry-size distribution. The first one, the Binomial distribution, is based on a clear modelling of the underlying processes. The other functional forms have been obtained on the basis of some plausible arguments. Subsequently, the obtained Binomial distribution does not contain a free parameter, while the other functions do. Hence, we change the Binomial distribution such that it also contains a free parameter: we define the probability of firms or employees to be located in a certain region r by $\xi_1 \cdot s_r$ instead of s_r.

Furthermore, the Binomial distribution above depends on the regional size under consideration. The same can be expected to hold for the entire regional industry-size distribution. Therefore, we define the parameters of all other functions also dependent on the regional size. Using all functions simultaneously regional industry-size distribution in terms of firm numbers ($m = f$) or employee numbers ($m = e$) is given by

$$
\begin{aligned}
P_{i,r}(m) \quad = \quad & \mu_1 \cdot Bnm(M_i, m, \xi_1 \cdot s_r) + \mu_2 \cdot (1 - \xi_2(s_r)) \cdot \xi_2(s_r)^m \\
& + \mu_3 \cdot \tfrac{[1-\xi_3(s_r)]^2}{\xi_3(s_r)} \cdot m \cdot \xi_3(s_r)^m \\
& + \begin{cases} \mu_4 \cdot \tfrac{[1-\xi_5(s_r)]^2}{\xi_5(s_r)} \cdot (m - \xi_4(s_r)) \cdot \xi_5(s_r)^{m-\xi_4(s_r)} & \text{if} \quad m \geq \xi_4(s_r) \\ 0 & \text{if} \quad m < \xi_6(s_r) \end{cases} .
\end{aligned} \tag{2.7}
$$

μ_1, μ_2, μ_3, and μ_4 determine to what extent each of the functions contributes to the total distribution. These have to sum up to one. ξ_1, $\xi_2(s_r)$, $\xi_3(s_r)$, $\xi_4(s_r)$, and $\xi_5(s_r)$ are the parameters of the various functions. As mentioned above, the Binomial function is defined such that its average value increases linearly with the regional size s_r. It seems plausible that the whole distribution should have the same characteristic. The average value predicted by Equation (2.7) increases linearly in s_r if each term increases linearly in s_r. In the case of the Binomial distribution this is already given. Hence, we have to examine the other three terms in Equation (2.7).

The average expected number for exponential distribution, the second term in Equation (2.7), is given by

$$\frac{\xi_2(s_r)}{1 - \xi_2(s_r)} . \tag{2.8}$$

We would like this average expected number to increase linearly with regional size s_r, meaning that we would like to obtain an average number of

$$\xi_2 \cdot s_r . \tag{2.9}$$

Equating (2.8) and (2.9) we obtain

$$\xi_2(s_r) = \frac{\xi_2 \cdot s_r}{1 + \xi_2 \cdot s_r} . \tag{2.10}$$

Hence, we define $\xi_2(s_r)$ by Equation (2.10). The same will also be done with the third term in Equation (2.7), the Erlang-n-distribution. The expected average number according to this term results to be $\frac{2 \cdot \xi_3(s_r)}{1-\xi_3(s_r)}$. Using the same calculation as above we obtain

$$\xi_3(s_r) = \frac{\xi_3 \cdot s_r}{2 + \xi_3 \cdot s_r} . \tag{2.11}$$

The average expected firm number according to the last term in Equation (2.7) is given by $\xi_4(s_r) + \frac{2 \cdot \xi_5(s_r)}{1-\xi_5(s_r)}$. There are several ways to make this value linearly depending on s_r. The most obvious is to assume that the first term, $\xi_4(s_r)$, depends linearly on s_r,

$$\xi_4(s_r) = \xi_4 \cdot s_r \ , \tag{2.12}$$

and that the second term also depends linearly on s_r, which is calculated by

$$\xi_5(s_r) = \frac{\xi_5 \cdot s}{2 + \xi_5 \cdot s_r} \ . \tag{2.13}$$

Hence, we finally obtain an Equation (2.7) with nine parameters that are independent of regional size, although this distribution shows such a dependence. Four parameters represent the share that each functional form contributes, μ_1, μ_2, μ_3, and μ_4, and five parameters determine the exact shape of functions, ξ_1, ξ_2, ξ_3, ξ_4, and ξ_5.

3. Data and empirical method

Having developed a general distribution function, it will now be fitted to the empirical data on the spatial distribution of industries within Germany. The analysis proceeds in several steps that are subsequently explained below, following a description of the empirical data used.

3.1. Empirical data

The data used in this approach has been collected by the German Federal Institute for Labour. The dataset contains the number of firm sites and employees for each 3-digit industry[1] and each of the 97 'Raumordnungsregionen'[2] in Germany for the 30th of June in 2003. The number of firm sites is the number of firm sites at which at least one person is employed.

[1] Industries are classified according to the WZ93-classification, which has been the standard industry classification in Germany at the time considered.

[2] 'Raumordnungsregion' are the type of regions used in German statistics that come nearest to labour market areas. They take commuters into account. However, they do not split the 440 German administrative regions ('Kreise') and do not include areas from different states, which makes them different from what real labour market areas would be in some cases.

Firms that only consist of the owner are not included in the data. This is not relevant in most industries, but might lead to a bias in a few industries, especially some service industries in which self-employed, one-person firms are frequent. If a firm has several sites in the same municipality, it is counted as one firm site.

The study conducted here is restricted to 198 of the 222 industries on the 3-digit level. We exclude all industries that contain less than 100 firm sites because the distribution of less than 100 firm sites is not well represented. Furthermore, we excluded those industries that represent single households and general state expenditure. Ten of the considered industries belong to agriculture and mining, 95 are manufacturing industries, and 93 are service industries, with industries denoted by i ($\in \{1, 2, ..., N_i\}$, $N_i = 198$), and regions denoted by r ($\in \{1, 2, ..., R\}$, $R = 97$).

The size of the 97 regions is calculated on the basis of total employment numbers in these regions. We define

$$s_r = \frac{\sum_{i=1}^{N_i} e_{i,r}}{\sum_{r=1}^{R} \sum_{i=1}^{N_i} e_{i,r}} \ . \tag{3.1}$$

Using the number of employees as a measure for regional size implies that the conducted analysis with $m = e$ is conducted in relative terms such that the regional industry-size distribution represents how much more or less employees regions contain in an industry compared to the assumption that industrial structure is exactly the same in each region. In the case of the firm number analysis ($m = f$), this does not hold. However, we decided to use the same measure for regional size in both analyses.

3.2. Fitting to the data

Firstly, the complete distribution (2.7) is fitted to the empirical data. Thus, the parameter set that maximises the negative log-likelihood value is calculated:

$$L_i^{(f)} = -\ln\left[\prod_{r=0}^{R} P_{i,r}(f_{i,r})\right] \qquad \text{and} \qquad L_i^{(e)} = -\ln\left[\prod_{r=0}^{R} P_{i,r}(e_{i,r})\right] \ . \tag{3.2}$$

The parameter sets that maximise $L_i^{(f)}$ and $L_i^{(e)}$, respectively, have to be calculated numerically, using the direction set method (according to Powell). Because local maxima exist for the log-

likelihood value, 1000 different randomly chosen starting values are used for the parameters. A repetition shows that the resulting values are reliable. The maximal log-likelihood values are denoted by $\hat{L}_i^{(f)}$ and $\hat{L}_i^{(e)}$, respectively.

While fitting the parameters their ranges have to be restricted. Obviously, all parameters have to be positive. Furthermore, $\mu_1 + \mu_2 + \mu_3 + \mu_4 = 1$ has to be satisfied. In addition, the last term in Equation (2.7) requires specific consideration. If we restrict the parameters related to this term by only requiring $\mu_4 > 0$, $\xi_4 > 0$, and $\xi_5 > 0$, then the second peak of the distribution, which is described by this term, might move into the range of the first peak. In contrast, the intention of including this term was to produce a second peak. Therefore, two restrictions are necessary.

First, $\xi_4 \cdot s_r$ determines the minimal number of firms that must be located in region r if this region falls into the second peak of the distribution and is said to contain an industrial agglomeration. It has been argued above that this number has to be high enough so that no region would contain such firm numbers under normal conditions. For each value of ξ_4, the number of regions that can be expected without industrial agglomeration (according to the first three terms) to contain more than $\xi_4 \cdot s_r$ firms can be calculated. This number is denoted by $n_{norm}(\xi_4)$ here. It will not be zero because the other parts of the distribution exponentially decrease and therefore never reach a value of zero. However, according to the above argument it should be small compared to the number of regions that contain industrial agglomerations according to the clustering part of distribution. The condition

$$\frac{n_{norm}(\xi_4)}{R} < 0.2 \cdot \mu_4 \tag{3.3}$$

is used here. This implies that share of the distribution explained by the last term in Equation (2.7) has to be five times greater than the share of regions that are expected to have also more than $\xi_4 \cdot s_r$ firms or employees.

Second, if local agglomerations are a distinctive phenomenon, they have to be the exception. We would not talk about 'local industrial clusters' if they occurred in most of the regions. Hence, the share of regions that contain a cluster has to be small. 10% is the maximum share accepted here. This implies that $\mu_4 \leq 0.1$ has to be satisfied.

3.3. Adequateness of the theoretical distribution

The theoretical distribution (2.7) has been set up rather ad-hoc, although based on some plausible considerations. Therefore, before we further investigate the fit of this theoretical distribution to the empirical situation in different industries, we should check whether it adequately describes the empirical data. To this end, the Kolmogorov-Smirnov test is used. The Kolmogorov-Smirnov test compares the cumulative distribution functions of a theoretical and an empirical distribution and checks whether the differences between distributions fall within the range that would be expected on the basis of randomness.

To obtain an empirical distribution, the numbers of firms or employees for different regions have to be aggregated. In the case of firm numbers the empirical distribution $P_{i,emp}(f)$ for a certain industry i is given by

$$P_{i,emp}(f) = \frac{1}{R} \cdot \sum_{r=1}^{R} \partial_{f_{i,r},f} \tag{3.4}$$

where ∂ is the Kronecker delta defined by

$$\partial_{f_{i,r},f} = \begin{cases} 1 & \text{if} \quad f_{i,r} = f \\ 0 & \text{if} \quad f_{i,r} \neq f \end{cases} . \tag{3.5}$$

Consequently, when using theoretical distribution $P_{i,theo}(f)$, the total sum over all regions has to be considered. It is given by

$$P_{i,theo}(f) = \frac{1}{R} \cdot \sum_{r=1}^{R} P_{i,r}(f) \; . \tag{3.6}$$

In the case of employee numbers, f is replaced by e in Equations (3.4) and (3.6). The Komologorov-Smirnov test is applied to these functions with a significance level of 0.01.

3.4. Importance of the various functional forms

The theoretical distribution used here consists of four different functions: the Binomial distribution, the exponential distribution, the Erlang-n-distribution and the cluster part. However, it is unclear whether all these functional forms are necessary to describe empirical distribution. It might be the case that empirical distribution is already sufficiently described by only one,

two or three of these functional forms. In order to test this we eliminate one functional form each time and fit the remaining theoretical distribution to the empirical data. If the eliminated part is necessary, the fit should become worse.

The likelihood ratio test is used to check whether eliminating one part of the theoretical distribution (2.7) significantly worsens the fit. The log-likelihood ratio is given by

$$\lambda_k = 2\ln[\hat{L}_i^{(f)}(\mu_k = 0)] - 2\ln[\hat{L}_i^{(f)}] \tag{3.7}$$

in the case of firm numbers, if the k-th functional form is eliminated. The respective log-likelihood ratio for employee numbers is calculated by replacing f by e. λ_k measures this difference between the models in fitting the data and can be used to check the significance of the difference (see Mittelhammer 1996). If one of the first three functions in Equation (2.7), namely the Binomial, the exponential, or the Erlang-n-distribution, is eliminated, this reduces the number of parameters fitted to the empirical data by two. In this case the likelihood ratio test rejects the hypothesis which states that eliminating the respective part does not change the theoretical distribution's suitability if $\lambda_k >= \lambda_{c,1} = \lambda_{c,2} = \lambda_{c,3} = 2.99$ (significance level of 0.05). Regarding the cluster part, the number of parameters is reduced by three, so that the hypothesis of this being unnecessary can only be rejected for $\lambda_k > \lambda_{c,4} = 3.91$. Hence, the likelihood ratio test allows for each part of the theoretical distribution to check whether this is necessary. If $\lambda_1 > 2.99$, $\lambda_2 > 2.99$, $\lambda_3 > 2.99$, and $\lambda_4 > 3.91$, the complete theoretical distribution describes the empirical data significantly better than any reduced form. This is separately tested for each industry.

If not all parts of the theoretical distribution are proven necessary, the theoretical distribution is successively reduced. First the part with the lowest value of λ_k and $\lambda_k < \lambda_{c,k}$ is eliminated. Then, the procedure is repeated with the remaining parts of the theoretical distribution: One part is eliminated, fitting the remaining function to the empirical data and the likelihood ratio is calculated. Again, it might result that all remaining parts contribute significantly to the goodness of fit. In this case, the procedure stops, otherwise the insignificant part with the lowest value of the likelihood ratio is eliminated. This procedure is repeated until every elimination leads to a significant fit reduction or only one part; or one part and the

cluster part remain. We obtain a theoretical distribution that sufficiently describes the empirical data. This theoretical distribution might consist of all four or some functional forms. The results might differ between industries.

4. Study of Germany

The empirical method, described above, is applied to firm numbers and employment numbers of 198 3-digit industries in Germany by using three steps. First, the theoretical distribution parameters (2.7) are fitted to the empirical data. Second, it is checked whether the empirical data is sufficiently described by the resulting theoretical distribution. If the theoretical distribution sufficiently fits the empirical data, a third step is taken: The theoretical distribution is simplified as much as possible without significantly decreasing the fit. The results are presented and discussed in the following.

4.1. Adequacy of the theoretical distribution

After separately fitting the complete theoretical distribution to the empirical data for each industry, we use a Kolmogorov-Smirnov test to check whether the empirical data is adequately described. The results are given in Table 1. The industries are presented in classes. To simplify the presentation, industy classes are defined and for each class the number of studied industries and the Kolmogorov-Smirnov test results are given.

Several observations can be made. First, the theoretical distribution does quite well. It adequately describes 159 out of 198 firm distributions and 168 out of 198 employment distributions. The fit is especially good for manufacturing industries. The theoretical distribution is rejected among the 95 manufacturing industries in only two cases for firm numbers and in only three cases for employment numbers. Hence, this distribution seems to be well-suited to describe the situation, especially in the manufacturing sector.

There are some classes of industries that are not fittingly described by the theoretical distribution used here. Such industries include agriculture, construction, hotels & restaurants, and social services & education. The theoretical above distribution has been deduced from

arguments regarding random location decisions, local resource distributions and clustering effects. A force that causes economic activity to be uniformly distributed in space has not been considered. This means that industries in which local activity correlates with local population is not adequately described by theoretical distribution. This seems to be the case for service sector industries including construction, hotels & restaurants, and social services & education; plus agriculture and to some extent for wholesale, transport, financial intermediation, and unions & organisations.

Hence, the results in Table 1 confirm the theoretical distribution as set up above. They show that this distribution does not fit every empirical situation, but most industries, especially the manufacturing sector. The theoretical distribution has problems describing industrial distribution in industries where the spatial distribution of economic activity closely follows the number of inhabitants. Only industries that are suitably described by theoretical distribution are included further in this paper.

There are slight differences between the theoretical distribution fit to the empirical distribution of firms and employees. The fit is, on average, somewhat better for employee numbers, especially in cases of agriculture, transport, and social services & education. However, no structural differences are detected.

4.2. Choice of distribution

The theoretical distribution has been set up above in a quite general form. It particulary contains four different functional forms that may not all be necessary to describe the distribution of economic activity in each industry. Therefore, the necessary parts are identified for each industry and for firm and employee numbers. The results are given for each industry class in the appendix (see Tables 4 and 5).

Here we mainly discuss the aggregated results for the manufacturing and service sectors. The first crucial question is whether all four parts of the theoretical distribution are necessary. Table 2 shows that all four parts are necessary in quite a number of industries. Each part is necessary in at least 50% of the studied cases. If we consider the total number of cases in

which these different parts play a role, the numbers are even quite similar, although the total numbers are somewhat higher for the Erlang-n-distribution and the Binomial distribution. These are the distributions which assign the highest probability to a number close to the average. The Binomial distribution results from a random location of firms and employees, which are predominant in many industries. However, we can also conclude from Table 2 that all four functional forms that have been included in the theoretical distribution are necessary.

The Erlang-n-distribution and the Binomial distribution are quite similar, so it could be assumed that they are substitutes. This is not the case. The results listed in Tables 4 and 5 show that a combination of these two functions is very frequent (178 of 327 cases). In the case of firm numbers in manufacturing and service industries and employee numbers in service industries, a combination of these two functions is a likely result. Next comes a combination of these two functions and the cluster term. This importance of the just named combinations does not hold for employee numbers in manufacturing industries, as here, a combination of all four functional forms is most frequent. Following is a combination of the exponential distribution, the Erlang-n-distribution, and the cluster term. These two combinations appear, excluding wood, paper and machinery industries, within all classes of manufacturing industries for, at least, half of the subindustries.

The importance of the four functional forms varies between industry sectors, between firm and employee numbers. The Erlang-n-distribution and the Binomial distribution play a similar important role for all sectors and numbers as they are necessary for between 56% and 85% of industries. In contrast, the necessity of exponential distribution and cluster terms varies tremendously. The exponential distribution is necessary for 88% of industries in the manufacturing sector if employee numbers are considered, while it is only necessary for 26% of service industries if firm numbers are considered. The exponential distribution is much more important for manufacturing industries than for service industries. It is also more important for the number of employees than for the firm numbers. It represents a distribution with a very high number of regions that contain no or very little economic activity and a small number of regions that contain a high economic activity. Hence, exponential distribution describes

geographic concentration[3], which seems mainly to be given in employment numbers of manufacturing industries. To a much lesser extent, it is also given in firm numbers of manufacturing industries and in employment numbers of service industries. The distribution of firm numbers within service industries can be described in 74% of cases without an exponential distribution. This means that most service industries, especially considering firm numbers, are not geographically concentrated.

For the cluster term results similar to those for the exponential distribution are obtained. The cluster term is also necessary for many (84%) manufacturing industries if employee numbers are considered and only for a few (21%) service industries if firm numbers are considered. It represents an extreme geographic concentration with a small number of regions with an extraordinary high number of firms or employees. If this term is necessary for the empirical data description, there are some regions that contain many more firms or employees than expected according to the rest of the theoretical distribution. One might be tempted to call these regions a local industry cluster. However, the approach used here only identifies these regions as an agglomeration of industry-specific activity. It makes no statements as to the causes of this agglomeration, so that further studies would be necessary to examine whether the identified agglomerations are local clusters (for a detailed discussion see Brenner 2004).

Table 2 shows that industry-specific agglomeration mainly appears with respect to employee numbers, while firms show a lower tendency to agglomerate. However, this might also be caused by lower firm numbers compared to employees which makes statistical results less significant, so that the cluster term's importance is not so apparent. The necessity of the cluster term is more often given for manufacturing industries than for service industries. Nevertheless, the difference is less significant than for the exponential distribution. Hence, although service industries show a much lower geographic concentration, they also show a large proportion of agglomeration, especially in terms of employee numbers. 39 (out of 68) service industries show agglomeration in employment numbers, of which 21 do not show geographic concentration (meaning an involvement of the exponential function).

[3] It leads, for example, to a high value of the Herfindahl index.

Besides the necessity of each part of the theoretical distribution (2.7), it is also interesting to examine the share of empirical data that is described by each distribution part. These shares are given by the parameters μ_1, μ_2, μ_3, and μ_4 that result from fitting the theoretical distribution to empirical data. Again, we do not consider each industry separately but aggregate the results to the manufacturing and service sector.

The results are given in Table 3 which appear to be similar to those in Table 2. However, the results in Table 3 are more pronounced. The Erlang-n-distribution and the Binomial distribution clearly dominate the firm number distribution in the manufacturing and service sector. The employee number distribution in manufacturing industries is mainly described by a combination of the exponential and Erlang-n-distribution. Regarding employee numbers in service industries, the Erlang-n-distribution dominates. In total, the Erlang-n-distribution highly contributes to the description of the actual data. This distribution has been included in the analysis only because it also describes the distribution of some local resources quite well. There has to be a reason for the fact that industries seem to locate in places according to this distribution or, at least, similar to the predictions of this distribution. This calls for further research.

The cluster term contributes very little to the overall distribution, although it is a significant term in around half of all analysed distributions. This is not a surprise because the cluster term only represents those regions that contain an industry-specific agglomeration. Therefore, the cluster term numbers of around 2% in Table 3 have to be interpreted such that on average around 2% of all regions contain an agglomeration in a randomly chosen industry.

5. Conclusions

This paper studies the regional industry-size distribution. While, for example, firm-size distributions have been extensively studied in the economic literature, how the distribution of the industrial activity among regions is shaped has not been addressed. While in the literature on the firm-size distribution, specific functions are tested and fitted, which is a proposed approach that combines various functional forms. It is argued that different theoretical considerations

lead to different shapes of regional industry-size distribution. It is checked with the help of empirical data whether these different shapes are involved in reality.

We find that the empirical regional industry-size distribution is not adequately described by one specific shape. Rather, for most industries it is a result of a combination of shapes. This implies that different forces are involved in the spatial distribution of industries.

Furthermore, we find that there quite some differences between different kinds of industries. We particularly distinguished between the manufacturing and service sector as well as between an analysis of firm and employees. It proves that geographic concentration and local agglomerations play a very strong role for employment numbers in manufacturing industries.

6. Appendix

References

ALLEN, P.M. (1997) **Cities and Regions as Self-Organizing Systems. Models of Complexity**. Amsterdam: Gordon and Breach.

BOTTAZZI, G., DOSI, G., FAGIOLO, G. & SECCHI, A. (2005) **Sectoral and Geographical Specificities in the Spatial Structure of Economic Activitie**. Sant' Anna School of Advanced Studies, Pisa.

BRAUNERHJELM, P. & CARLSSON, B. (1999) Industry Clusters in Ohio and Sweden, 1975-1995, **Small Business Economics**, 12, pp. 297-293.

BRENNER, T. (2001) **Self-organisation, Local Symbiosis of Firms and the Life Cycle of Localised Industrial Clusters**. Papers on Economics & Evolution #0103, Max-Planck-Institute Jena.

BRENNER, T. (2003) **An Identification of Local Industrial Clusters in Germany**. Papers on Economics & Evolution #0304, Max-Planck-Institute Jena.

BRENNER, T. (2004) **Local Industrial Clusters: Existence, Emergence and Evolution**. London: Routledge.

ELLISON, G. & GLAESER, E.L. (1997) Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach, **Journal of Political Economy**, 105, pp. 889-927.

ELLISON, G. & GLAESER, E.L. (1999) The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?, **American Economic Review**, 89, pp. 311-316.

ISAKSEN, A. (1996) Towards increased Regional Specialization? The Quantitative Importance of New Industrial Spaces in Norway, 1970-1990, **Norsk Geografisk Tidsskrift**, 50, pp. 113–123.

KEILBACH, M. (2000) **Spatial Knowledge Spillovers and the Dynamics of Agglomeration and Regional Growth**. Heidelberg: Physica Verlag.

KRUGMAN, P. (1996) **The Self-Organizing Economy**. Cambridge: Blackwell Publishers.

MITTELHAMMER, R.C. (1996) **Mathematical Statistics for Economics and Business**. New York: Springer.

PANICCIA, I. (1998) One, a Hundred, Thousands of Industrial Districts. Organizational Variety in Local Networks of Small and Medium-sized Enterprises, **Organizational Studies**, 19, pp. 667-699.

SFORZI, F. (1990) The Quantitative Importance of Marshallian Industrial Districts in the Italian Economy, in F. Pyke, G. Becattini & W. Sengenberger (eds.) **Industrial Districts and Inter-Firm Co-operation in Italy**, pp. 75–107. Geneva: International Institute for Labour Studies.

STERNBERG, R. & LITZENBERGER, T. (2004) Regional Clusters in Germany: Their Geography and their Relevance for Entrepreneurial Acitivities, **European Planning Studies**, 12, pp. 767–792.

		number of sub- of industries	model describes adequately the distribution of	
			firms	workers
A	Agriculture, etc.	6	1	4
Q	Mining and quarrying	4	3	4
	Primary sector	10	4	8
F	Food products, etc.	8	7	7
T	Textiles and leather	11	11	11
W	Wood and products	5	5	5
P	Paper, etc.	5	5	5
H	Chemicals, etc.	9	9	9
N	Non-metalic products	7	7	6
M	Metals and products	12	11	11
Y	Machinery and equipment	7	7	7
E	Electrical and optical equip.	15	15	15
V	Transport equipment	8	8	8
O	Manufacturing n.e.c.	8	8	8
	Manufacturing sector	95	93	92
G	El., gas and water supply	4	3	4
C	Construction	5	2	2
S	Wholesale, etc.	19	13	13
R	Hotel and restaurants	5	3	2
X	Transport, etc.	12	8	10
I	Finanial intermediation	5	3	3
B	Business activities, etc.	22	21	22
Z	Social services and education	11	1	4
U	Unions and organisations	3	2	2
L	Services to leisure activities	7	6	6
	Service sector	93	62	68
	total	198	159	168

Table 1: Classes of industries and the number of industries within these classes where the Komolgorov-Smirnov test confirms the adequateness of the theoretical distribution.

sector	number of	Distribution exp.	Erlang	Binom.	cluster
manufacturing	firms	41 (44%)	66 (71%)	62 (67%)	38 (41%)
	employees	81 (88%)	74 (80%)	59 (64%)	77 (84%)
service	firms	16 (26%)	53 (85%)	51 (82%)	13 (21%)
	employees	26 (38%)	57 (84%)	38 (56%)	39 (57%)
total		164 (50%)	250 (76%)	210 (64%)	167 (51%)

Table 2: Number of industries in which the different distribution shapes are necessary to describe the empirical situation.

sector	number of	Distribution exp.	Erlang	Binom.	cluster
manufacturing	firms	17.4%	41.9%	38.3%	2.4%
	employees	47.4%	36.0%	14.2%	2.3%
service	firms	7.8%	49.8%	40.6%	1.7%
	employees	17.7%	68.1%	12.2%	1.9%

Table 3: Average share of different distributions in the description of empirical data.

	Combination of functions													
	e	t	b	et	eb	tb	etb	ec	tc	bc	etc	ebc	tbc	etbc
A	-	-	-	-	-	-	-	-	-	-	-	-	-	1
Q	-	-	-	-	-	1	-	-	-	-	-	1	1	-
Primary	-	-	-	-	-	1	-	-	-	-	-	1	1	1
F	1	-	1	2	-	2	-	-	-	-	1	-	-	-
T	-	-	-	2	-	1	-	1	-	1	1	1	4	-
W	2	-	-	-	-	1	-	-	1	-	-	-	1	-
P	-	-	-	1	-	3	-	-	-	-	-	1	-	-
H	1	-	-	1	-	1	3	1	-	-	-	1	-	-
N	-	-	-	1	-	2	-	-	-	-	1	-	2	-
M	-	-	-	1	-	3	-	1	-	-	1	1	3	-
Y	-	-	2	1	-	3	-	-	1	-	-	-	-	-
E	-	-	-	2	-	4	1	-	-	1	-	4	3	-
V	1	-	-	3	-	3	-	-	-	-	-	-	1	-
O	-	-	1	-	-	2	-	-	-	-	-	2	2	1
Manufact.	5	-	4	14	-	25	4	3	2	2	4	10	16	1
G	-	-	-	-	-	3	-	-	-	-	-	-	-	-
C	-	-	-	1	-	1	-	-	-	-	-	-	-	-
S	-	-	-	-	2	8	-	-	-	-	-	1	2	-
R	-	-	-	-	-	1	-	-	-	-	1	-	1	-
X	1	-	-	1	-	3	-	-	-	-	1	1	1	-
I	-	-	-	1	-	2	-	-	-	-	-	-	-	-
B	-	3	-	1	1	13	1	-	1	-	-	1	-	-
Z	-	-	-	-	-	1	-	-	-	-	-	-	-	-
U	-	-	-	-	-	-	-	-	-	1	-	-	1	-
L	-	-	-	-	1	3	1	-	-	-	-	-	1	-
Service	1	3	-	4	4	35	2	-	1	1	2	3	6	-
total	6	3	4	18	4	61	6	3	3	3	6	14	23	2

Table 4: Number of industries in each class of industries for which a certain combination of functions (e=exponential distribution; t=Erlang-n-distribution; b=Binomial distribution; c=cluster term) is necessary to describe the empirical distribution of firms.

	Combination of functions													
	e	t	b	et	eb	tb	etb	ec	tc	bc	etc	ebc	tbc	etbc
A	-	-	-	1	-	-	-	-	1	-	2	-	-	-
Q	-	-	-	-	-	-	-	-	-	-	1	-	-	3
Primary	-	-	-	1	-	-	-	-	1	-	3	-	-	3
F	-	-	-	-	-	-	-	1	-	-	-	1	1	4
T	1	-	-	2	-	-	-	-	1	-	-	-	-	7
W	1	-	-	-	-	-	-	1	-	-	1	1	1	-
P	1	-	-	-	-	-	-	-	1	-	1	1	1	-
H	-	-	-	1	-	-	1	-	-	-	2	1	1	3
N	-	1	-	-	-	-	-	-	-	-	1	1	-	3
M	-	-	-	-	-	-	2	-	-	-	-	3	1	5
Y	-	-	-	-	-	-	-	-	1	-	2	1	-	3
E	1	-	-	-	-	-	-	1	-	-	3	1	1	7
V	1	-	-	-	-	-	1	-	-	-	2	-	-	4
O	-	-	-	-	-	-	1	-	2	-	3	1	-	1
Manufact.	5	1	-	5	-	-	5	3	5	-	15	11	6	37
G	1	1	-	-	1	-	-	-	-	-	-	-	-	1
C	1	-	-	-	-	-	-	-	1	-	-	-	-	-
S	-	-	-	-	-	5	-	1	3	-	-	-	4	-
R	-	1	-	-	-	1	-	-	-	-	-	-	-	-
X	1	-	-	-	-	1	2	-	1	-	1	1	1	2
I	-	-	-	-	-	-	-	1	-	-	1	-	-	1
B	-	1	-	-	-	9	-	3	2	-	2	1	4	-
Z	-	1	-	-	-	-	-	-	1	-	-	-	2	-
U	-	-	-	-	-	-	-	-	1	-	-	-	-	1
L	-	1	-	1	-	-	1	-	1	-	2	-	-	-
Service	3	5	-	1	1	16	3	5	10	-	6	2	11	5
total	8	6	-	7	1	16	8	8	16	-	24	13	17	45

Table 5: Number of industries in each class of industries for which a certain combination of functions (e=exponential distribution; t=Erlang-n-distribution; b=Binomial distribution; c=cluster term) is necessary to describe the empirical distribution of employees.