Identifying Innovative Regions: An Application of a Data Envelopment Analysis for German Regions

Tom Brökel and Thomas Brenner

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany

Abstract

The observation of variance in firms' innovation activities across regions has given rise to the research strand on regional innovation processes. This paper discusses the possibility of applying performance analysis to identify and evaluate specific regional phenomena that foster regional firms' innovativeness. It applies a semi-deterministic nonparametric frontier analysis on data of two industries for German labor market regions and two years.

1 Introduction

Recently, differences in regional innovativeness have been subject of intense research. Starting with the "knowledge production function" approach by Griliches (1979) much research has been done examining (mostly empirically) the co-locatedness of certain regional factor endowments with firms' innovation counts (see e.g. Jaffe, 1989; Feldman, 1994; Audretsch, 1998). It is argued that some empirically observable factors, such as universities, firms' headquarters and venture capital firms contribute to firms' innovativeness if located within the same region (see Broekel and Brenner, 2005).

Besides this 'co-locatedness' approach a rich literature emerged that analyses specific regional phenomena that relate to firms' innovativeness as well. For example in the well-known *innovative milieu* approach by (Camagni, 1991) the cooperative behavior among the regional actors is emphasized. The existence of an 'innovative milieu' enables regional actors to be more innovative because of higher resource sharing activities (see e.g. Asheim and Isaksen, 2002; Kostiainen, 2002). Other regional features that foster innovativeness are e.g. the role of institutions (Cooke et al., 1997), regional identities (Backhaus and Seidel, 1998), or the mobility of employees (Almeida and Kogut, 1999).

These approaches have in common that it is not only a regional factor endowment (as claimed in the 'co-location' approach) which enhances firms' innovativeness, but that unique (e.g. historic) developments have led to the creation of a specific regional feature. Under consideration of their rareness they are pooled under the label 'regional phenomena' in the following.

While both approaches aim at understanding the variance of regional innovativeness, they have been treated separately. This is due to the fact that the 'co-location' approach is mainly driven by empirical work that emphasizes the role of the measurable regional factor endowment. In contrast, the 'regional phenomena' approach is mainly based on case study research concentrating on hard to measure 'soft' factors influencing e.g. regional actors resource sharing activities. This paper emphasizes the importance of taking both approaches into account in order to generate a complete picture of the variance in regional innovativeness. We propose and apply efficiency analysis that allows us to make comparisons between regions on the basis of the non-measurable regional factors while controlling for the measurable factor endowment. It is argued that the hard to measure factors and processes that are emphasized by the 'regional phenomena' approach, cause differences in the use of measurable factor endowments, which is at center in the 'co-location' approach. In this paper, we will show how regions with such a regional phenomenon can be identified in the case of two industries in Germany.

To conduct the analysis of the deviations in a statistically sound way, the efficiency analysis is applied to the 270 German labor market regions using industry specific data on the measurable factor endowment that is commonly used in co-location studies. The innovative output of the electric / electrical / optical as well as the car industry is approximated by patent applications

for the years 1999 and 2000.

In order to cope with the trade-off between assumed determinism and ex-ante specification of a functional form, we employ a semi-deterministic nonparametric frontier approach: A standard output oriented Data Envelopment Analysis that allows for variable returns to scale is flagged by a prefixed outlier elimination as proposed by Simar (2003). Bias-corrected estimates and confidence intervals are computed for the obtained results based on the bootstrapping method presented in Simar and Wilson (2000*a*).

The paper proceeds as follows. Section 2 discusses the motivation and justification for the application of performance analysis as well as shortcomings of the present regional innovation research. The performance analysis and the employed model are presented in Section 3. Section 4 gives a description of the conception and components of the relevant regional knowledge / technology infrastructure. In Section 5 the results of the performance analysis are presented and followed by an extensive discussion of the obtained results and the employed method in Section 6. Section 7 concludes.

2 Analysis of regional innovation processes

2.1 The 'regional phenomena' approach

A. Marshall already noted that:

> "Sheffield and Solingen have acquired industrial "atmospheres" of their own; which yield gratis to the manufactures of cutlery great advantages, that are not easily to be had elsewhere: and the atmosphere cannot be moved" (Marshall, 1919, p. 284).

With the begin of the 1990s his phrase of an "industrial atmosphere" was taken up and transfered from a production to an innovation based context. Noticing that some regions seem to be highly innovative while other are not, a rich literature emerged that analyses (mostly with case study research) specific regional phenomena that relate to the innovativeness of regional firms. Most prominently the concept of an 'innovative milieu' by Camagni (1991), has gained wide attention in the literature. It is defined as:

> "... the set, or the complex network of mainly informal social relationships in a limited geographical area, often determining a specific external 'image' and internal 'representation' and sense of belonging, which enhance the local innovative capability through synergetic and collective learning processes" (Camagni, 1991, p. 3).

The most important features of the 'innovative milieu' concept may be seen in that it results from a collective learning process in which informal relationships play a crucial role. It enhances firms' innovativeness by reducing uncertainties that are inherent to the innovation processes (see Camagni, 1991).

Yet the 'innovative milieu' is just one concept that links the actors' knowledge sharing activities to the innovativeness of regional firms. While it highlights the cooperative behavior among regional actors that is expressed in intense informal social relationships, other (but similar) approaches in this context put a stronger focus on e.g. the role of institutions (Cooke et al., 1997), regional identity (Backhaus and Seidel, 1998), or the mobility of people (Almeida and Kogut, 1999).

These approaches have in common that it is not only a regional factor endowment which enhances firm's innovativeness, but that specific (e.g. historic) developments have led to the creation of special 'milieus' or 'atmospheres' within a few number of regions. Through various mechanisms such 'regional phenomena' add to the effect of the factor endowment in fostering the sharing and employment of resources among the actors, see e.g. the 'phenomenon' of Regional Innovation Systems (Cooke, 1992; Asheim and Isaksen, 2002).

The type of research in this field is mainly case study driven. This allows for in-depths analysis that take specific regional developments, historic events and special actors or circumstances into account. Furthermore the complexity of innovation processes is pictured more realistically. A good example is the well known work by Saxenian (1998) which sets regional innovativeness differences in relation to local cultures.

Such work clearly pictures regional innovation processes very realistic and accurate. The findings are yet difficult to transfer to a more general picture because of the mostly qualitative analysis. In contrast, the mainly quantitative 'co-location' approach is presented in the following subsection.

2.2 The 'co-location' approach

With the introduction of the regional dimension in innovation economics the question arose which regional factors contribute in what way to firms' innovativeness. Starting with Griliches (1979) 'knowledge production function' (KPF) approach, much research has been done examining (mostly empirically) the co-locatedness of certain regional factor endowments with firm's innovation counts (see e.g. Jaffe, 1989; Audretsch, 1998; Bode, 2004). It is argued that some quantitative measurable factors, such as universities, headquarters and venture capital firms, etc., contribute to the firms' innovativeness if located close by (Feldman, 1994; Feldman and Florida, 1994).

The KPF builds upon the standard economic understanding of a production function which describes the relationship between inputs of a transformation process and the corresponding maximal achievable output level. In a regional setting the KPF approach is based on the assumption that the co-location between regional factors and innovations of firms located within the region reflects the actual effective relationship between them. Thus the higher the observed co-location (correlation) the more likely a regional factor contributes to the firms' innovation processes.

In a regional innovation process context the inputs approximate the regional endowment with theoretically relevant factors. The output of the innovation process is represented by the sum of regional firms' innovations in a certain temporal frame. Exemplarily and most simple, one could argue that the innovation counts INN_r of region r are to be explained by different regional factors, e.g. students at universities $STUD_r$, venture capital firms VC_r, etc. Commonly the relationship between outputs and inputs is expressed as:

$$INN_r = STUD_r^{\beta_1} * VC_r^{\beta_2} ... \ . \tag{1}$$

In order to avoid non-linear regression techniques Equation 1 is transformed into a linear regression model by taking, exemplarily the logarithm to the basis 10. Due to the stochastic nature of the underlying process an error term is added.[1] The final model that is analyzed by regression techniques can be written as:

$$\log INN_r = \alpha + \beta_1 * \log STUD_r + \beta_2 * \log VC_r + ... + u_r \tag{2}$$

Such an approach has a number of advantages. It allows for the stochastic nature of innovation processes (even though there are some problems in practical implementation). A large number of variables can be tested for their significance in the modeled process. Furthermore the results represent the average statistical relationship between the dependent and independent variables for the sample of regions.

On the downside the ex-ante assumption of the functional form seems to be very critical. For example in the study by Broekel and Brenner (2005) addresses among other things the impact of the assumed functional form. They find serious differences in the outcomes for two dissimilar models (see Broekel and Brenner, 2005). Nevertheless, such approach seems to be especially valuable for evaluating the role of certain regional factors in regional innovation processes.

Regional factors which are hard to measure (if at all) and the 'regional phenomena' approach emphasizes on, are however not explicitly part of this approach. In fact, to the extent that they are not correlated to one or more independent variables they are part of the error term u_r. This represents a serious drawback for regional innovation research. A method that builds upon both approaches is presented in the following subsection.

2.3 Regional performance analysis

A third and very young strand in regional innovation economics argues in favor of a performance concept. In the first works relating the this issue of Fritsch (2006) and Broekel and Brenner (2006), it is argued that performance analysis can be used to evaluate the quality (efficiency with respect to the regional factor endowment) of regional innovation processes. The authors use the "relationship between the input and the output of the regional innovation activity

[1] Note that if this error term is already added in Equation 1 and still allow for linearization it takes the form $*10^{u_r}$ (depending on the type of logarithm used). This is rather hard to justify.

for assessing the efficiency" (Fritsch, 2006, p. 2).

Fritsch (2006) conducts the analysis in order to evaluate variances in the innovation performance of regions that result from differences in the organization and management of innovation processes or which are caused by all kinds of regional and sectoral factors, e.g. agglomeration advantages, clustering, or the presence of universities. A single input variable, the total number of R&D employees, and only one output variable, the total number of patents, are used in his work.

Instead of concentrating on slight differences in the performance scores of regions we rather suggest to employ the efficiency analysis as a method to join the 'co-location' approach and the 'regional phenomena' approach. The basic idea is presented in the following.

As discussed before, the 'co-location' approach sets innovation counts in relation to the measurable regional factor endowment. Regions with better factor endowments can be easily identified in this framework. In contrast the 'regional phenomena' literature emphasizes interactions between actors and the impact of institutions. This is, however, not easy to measure. As a consequence, it is difficult to identify which regions perform better than others in terms of these factors and dynamics. For each specific region we might study the strength of local interaction, knowledge exchange, and supportive institutional settings as well as other relevant aspects. On a more general level it is difficult to make a comparison between regions with respect to these hardly measurable features.

Both approaches argue and give evidence that the factors they emphasize influence the outcome of regional innovation processes. Thus the observed variance in innovation activities across regions is caused by both factors. While the effect of the existence of a 'regional phenomena' disappears in the error term in ordinary 'co-location studies', researchers of the 'regional phenomena' approach can hardly compare the factor endowment of the region under consideration to other regions. The latter is due to the very nature of case study research.

Thus both type of research strands look on the total effect of (measurable and not measurable) regional factors on firms' innovation activities. But they only take one side into consideration. This drawback can be overcome by efficiency analysis.

Not all regions that have a good factor endowment are able to generate high innovation output. On the other side many regions with comparatively poor factor endowment generate remarkable innovation counts. We argue that the processes described by the 'regional phenomena' approach cause differences in the use of the measurable factor endowment. This better use of the factor endowment results from some more helpful ways of interacting or some more helpful local institutional settings, as exist in e.g. regional innovation systems (Cooke, 1992) or 'innovative milieus' (Camagni, 1991).

The pure endowment of the region with (measurable) factors influences the innovation activities of local firms as well. For example the size and quality of local labor market has an effect on the ease with which firms find adequate personnel that might bring up innovative ideas (Kline

and Rosenberg, 1986).

Considering both effects we can argue that the deviations of the innovative output of regions from the value that is expected according to the measurable factor endowment have to be the result of the factors and processes that are not measurable or not measured. Thus we need to separate the effect of the factor endowment and the effect of the not measurable factors in order to gain a better understanding of the processes at work. This separation is achieved by performance analysis that evaluates the strength of the processes emphasized in the 'regional phenomena' literature while simultaneously controlling for the effect of the measurable factors put forward by the 'co-location' approach.

Concretely this means that regions which are identified as being highly innovative (being efficient) with respect to the measurable factor endowment, are likely to have a specific regional feature in the style of a 'regional phenomena'. Thus the analysis is not conducted in order to evaluate marginal differences, but rather to identify regions which are able to generate comparatively more innovations with respect to their measurable factor endowment. We claim that these regions are valuable objects for case study research that compares them to regions that have not been found effcient.

Conducting performance analysis however raises some problems in this context. The stochastic nature of innovation processes conflicts with the somewhat more deterministic view of the production processes to which efficiency analysis is generally applied. Due to its inherent uncertainty, any innovation creation process can have as 'quasi-output' only the likelihood of an innovation to take place. Nevertheless in practice, "inventions and innovations can be measured and counted even if each one is unique" (Fritsch, 2006, p. 2). The systematic influence of regional settings on innovation processes might be difficult to evaluate for a single innovation, but on an aggregate level (such as all innovations by regional firms) a carefully conducted performance analysis seems to be applicable.

Another problem relates to the unique and individual character of the way innovations are generated. Or in other words, for each innovation the elements and functional relationships in the creation processes can differ. Even when innovation processes are aggregated, e.g. to a regional level, there seems to be no way of translating it into an ex-ante specified functional relationship between inputs and outputs. This has to be considered in the choice of the correct analysis.

Because of the uniqueness of innovation processes furthermore an efficiency analysis in this field should not aim at evaluating marginal differences. Even though the analysis presented later in the paper includes up to thirteen input variables in addition to the industry specific R&D employees, we do not claim that our analysis represents the true picture of an industry specific 'regional innovation performance'. As has been pointed out before, it is just employed to identify the most innovative regions.

The proposed model and the employed method are described in detail in the next section.

3 Nonparametric efficiency analysis in a regional context

3.1 Efficiency analysis of innovation processes

The two approaches of Fritsch (2006) and Broekel and Brenner (2006) that aim at analyzing the efficiency of firms' innovation processes on a regional level with respect to the regional factor endowment differ significantly. Fritsch (2006) applies a variation of the standard knowledge production function (KPF):

$$ln(R\&D\ output)_r = ln\, a + b(ln\, R\&D\, input)_r + \epsilon_r \quad . \tag{3}$$

This however does not meet the requirement of the (no ex-ante) specification of a 'regional innovation production function' that has been emphasized in the previous section. This is explicitly taken into account by the non-parametric performance analysis approach by Broekel and Brenner (2006). Here the performance frontier is estimated from the observed data without any ex-ante specification of a functional form. Its disadvantage is however the 'assumption'[2] of a deterministic relationship between inputs and outputs. The KPF approach has a clear advantage due to its stochastic nature.

In order to cope with this trade-off between assumed determinism and ex-ante specification of a function form, we extent the deterministic approach by Broekel and Brenner (2006) by employing a semi-deterministic frontier approach: A standard output oriented Data Envelopment Analysis that allows for variable returns to scale is backed by an outlier identification method proposed by Simar (2003). Eventually, bias-corrected efficiency estimates and confidence intervals are computed based on the bootstrapping method presented in Simar and Wilson (1998, 2000*a*). The exact method is presented in the next subsections.

3.2 Remarks on the choice of estimator

Nowadays one has a wide range of non-parametric models to choose from when conducting performance analysis. One has to decide upon using new robust efficiency estimators or the traditional ones. The most commonly used robust estimator is the order-m frontier estimator introduced by Cazals et al. (2002). In contrast to traditional ones, it represent efficiency scores not in relation to the complete population of DMUs but rather to a (conditionally) randomly drawn sample. The order-m estimators have "several desirable properties" (Wheelock and Wilson, 2003), such as root-n consistency, low sensitivity to noise in the data and the absence of the "curse of dimensionality" (see e.g. Cazals et al., 2002; Wheelock and Wilson, 2003). This curse means that the number of observations in relation to the sum of input (p) and output (q)

[2] This is not really an assumption but relates more to the handling of statistical noise in these models. In the original DEA and in contrast to stochastic models all stochastic variance is included in the obtained efficiency estimates.

variables has to increase exponentially in order to achieve the same order of magnitude in estimation error as obtained with the order-m estimator (Simar and Wilson, 2000*b*; Wheelock and Wilson, 2003).

The theoretical basis for comparing a DMU only to a sup-sample of other DMUs rather than to the total sample, even though the sub-sample its drawn randomly, seems to be comparatively weaker. We solve the trade-off between statistical robustness and theoretical foundation by implementing a semi-deterministic mixture. We employ a standard DEA model, with a prefixed outlier-detection method and an estimation of bias-corrected efficiency scores and their confidence intervals afterwards. Because the traditional DEA estimator is used the obtained estimates as well as the confidence intervals suffer from the "curse of dimensionality". This has to be kept in mind for the interpretation of the results. At this stage we weight the theoretical foundation to be of higher importance.

Another important choice has to be made between using a Free Disposal Hull (FDH) or a Data Envelopment Analysis (DEA) approach. They differ in that there is an additional assumption of convexity regarding the underlying technology in the DEA case. Convexity implies that input (output) variables are assumed to be linearly substitutable with other input (output) variables. Unfortunately, there is no definite test of the technology's convexity (Simar and Wilson, 2007). But the DEA estimator is clearly preferred over the FDH estimator if the technology is convex, because the "curse of dimensionality" is weaker in this case for the DEA estimator (Simar and Wilson, 2000*b*). Regarding the employed input variables that are presented later in the paper the assumption of convexity seems to be valid. In other words, we assume linear substitutability between all of them, except for the R&D employees of the investigated industry. This is discussed in the next subsection in more detail.

3.3 The data envelopment analysis

The Data Envelopment Analysis (DEA), which has been developed by Charnes et al. (1978), is a method to measure decision making units' (DMU) efficiency. The analysis of efficiency of decision making units (DMU) is based on the performance concept of Farrell (1957). Thereby, an efficiency analysis compares the actual output of a DMU with the maximal output estimated by a production function. The *best-practice* units of a comparison group are used as reference for the evaluation of the other group units. With the assumption of convexity not only actual existing units are compared, but also not-negatively-weighted linear combinations of these "virtual units" (Kuhner, 2005). This requires the underlying technology T to be convex. Other assumptions regarding the technology are that all empirically observations and all non-negative convex combinations of them ("virtual" observations) are also part of T. Implicitly, it is assumed that input and output quantities are arbitrarily divisible. It is an 'optimistic' assumption concerning the transformation possibilities of DMUs (Scheel, 2000*a*, p. 47).

The attractiveness of nonparametric models results not only from these comparatively weak as-

sumptions on the technology, but also from the fact that no explicit production function (transformation function of inputs into outputs) is assumed. No ex-ante weighting of input or output factors is needed, because the DEA generates the most favorable weights from the data.[3]

This comprises that the weight of an input or output can drop for a DMU to almost zero. Thus it can be almost completely substituted by the other factors. While this does not seem to be problematic for the majority of the regional factors included, for the R&D employees of the investigated industry it clearly is. The R&D employees are the creative minds behind the innovations. They are the source of creativity and 'units' that recombine knowledge. We assume therefore that they cannot be substituted by other regional factors. In fact all regional factors that enhance the innovation processes only take effect through the R&D employees. Our regional innovation process model for region r and a group of size p of R&D employees (e.g. all R&D employees of a specific industry) can be written as:

$$INN_{r,p} = \sum_{i=1}^{p} C_i^{R\&D} * f(R_1, ..., R_e) \ , e = 1, ..., n, i = 1, ..., p \tag{4}$$

with $C_i^{R\&D}$ as the innovative capacity of the R&D employee i. $f(R_1, ..., R_n)$ reflects the regional influence on the R&D employee resulting from the n regional factors R_e. The innovative capacity of the R&D employees $C_i^{R\&D}$ is set to 1 because we cannot account for individual differences in the innovativeness. In practice, we simply multiply all input variables with the number of regional industry-specific R&D employees.

Regions with a 'regional phenomenon' generate an extra output from their input set which regions without this feature do not. This is the difference in the innovation counts resulting from the non-measurable factors. In comparison to other units with the same measurable factor set they are characterized by a higher output level. This seems to be best reflected by an output-oriented method in contrast to an often used input-orientation.

By using a *radial* measure (on the base of the vector to the origin) to estimate the distance between a DMU and the frontier function the obtained scores can be easily interpreted as the (proportional) increase in the output needed for that DMU to become efficient (Scheel, 2000*b*). We employ a variable returns to scale (VRS) model (Banker-Charnes-Cooper model, BCC-model (see Banker et al., 1984)) because of its lower sensibility to model misspecifications (Galagedera and Silvapulle, 2003). When the type of variables in the analysis is including ratios, VRS are the adequate option as well (Hollingsworth and Smith, 2003). Although we do not use pure ratios but ratios multiplied with number of R&D employees this argument still seems to hold. The use of such complex input variables however clearly prohibits the analysis of the variable returns to scale assumption and its reflection in the results itself. The assumption of VRS just yields the more appropriate model.

[3] For an introduction see Scheel (2000*a*).

The final DEA model that is employed in this paper shows as an output-oriented VRS model.

$$\max_{\Phi_l,\lambda_{il}} \Phi_l + \left(\sum_{j=1}^{m} \epsilon s_{jl}^{+} + \sum_{r=1}^{s} \epsilon s_{rl}^{-} \right) \tag{5}$$

$$\begin{aligned} s.t. \quad & \sum_{i=1}^{n} \lambda_{il} x_{ri} + s_{rl}^{-} = x_{rl} && \text{for all } r = 1, ..., s, \\ & \Phi_l y_{jl} - \sum_{i=1}^{n} \lambda_{il} y_{ji} + s_{jl}^{+} = 0 && \text{for all } j = 1, ..., m, \\ & \sum_{i=1}^{n} \lambda_{il} = 1, \\ & \lambda_{il}, s_{jl}^{+}, s_{rl}^{-} \geq 0. \end{aligned}$$

The data envelopment analysis assumes that all observed units belong to the attainable set. Thus by their very nature they are highly sensitive to extreme values and outliers in the data set. Therefore, the detection and elimination of such outliers is of uttermost interest. It is presented in the following.

3.4 Outlier detection - the order-m frontier approach

The approach by Simar (2003) is chosen because it "certainly offers an appealing and useful methodology in the exploratory data analysis phase of any efficiency analysis with real data" (Simar, 2003, p. 393). Its main advantage is its easy and fast to implementation irrespectively of the number of observations.[4]

The order-m frontier efficiency measures needed for this procedure might be explained as follows. The core idea is to set up a conditional frontier that does not envelop all DMUs in the population, but just a share of it. This share is determined by the integer value m which can be fixed by the researcher. Hereby, the condition for the input-oriented case[5] is that DMUs are considered with an output level of at least the same as the DMU of interest. The radial distance of a DMU (x_0, y_0) interior to the order-m frontier represents the proportional reduction in its input it needs, in order to become efficient to a randomly drawn sample of DMUs which have an output level of $Y \geq y_0$. Note that a FDH estimator is used here although the final model employs a DEA estimator. It has no effect on the applicability of the method (Simar, 2003).

For a multivariate setting consider $X^1, ..., X^m$ are m (p-dimensional) random DMUs drawn from the conditional distribution function of X given $Y \geq y_0$. The random variable

$$\tilde{\theta}_m(x_0, y_0) = \min_{i=1,...,m} \left\{ \max_{j,...,p} \left(\frac{X^{i,j}}{x^{i_0}} \right) \right\} \tag{6}$$

with $X^{i,j}(x_0^j)$ as the jth component of X^i (of x_0 respectively). It measures the distance between point x_0 and the free disposal hulls $X^1, ..., X^m$. The latter are generated from the conditional

[4] We used the software package **FEAR** by Wilson (2006), which can be implemented in the statistical software **R**.

[5] For the outlier-detection both orientations are needed. To save space only the input-oriented model is present here. The output-oriented notation can be found in Simar (2003).

distribution function of X given $Y \geq y_0$. The order-m efficiency measure of $DMU_0(x_0, y_0)$ is then defined as

$$\theta_m(x_0, y_0) = E[\tilde{\theta}_m(x_0, y_0)|Y \geq y_o] \quad . \tag{7}$$

Because the distribution of the population is unknown the calculation of the order-m frontiers requires the use of the empirical distribution functions. In a multivariate case this calculation involves a numerical integration which is easier to solve by Monte-Carlo approximation. Therefore the number of Monte-Carlo replications (B) has to be chosen. Following Simar (2003) B is set to 200.

The principle of the outlier-detection procedure is simple: Any DMU is a potential outlier, even when m increases, if its super-efficiency[6] order-m input efficiency measure $\hat{\theta}^s_{m,n}(x_0, y_0)$ is greater than one plus a certain threshold α (respectively, for output orientation $\hat{\lambda}^s_{m,n}$ is smaller than one minus α). The routine has to be conducted for the input as well as for the output-direction. While this is fairly easy the main problem relates to the definition of the relevant threshold above which the level of super-efficiency characterizes an outlier while below it, a unit produces the output just "super" efficient. In the absence of a definite rule a sensitivity analysis has to be conducted.

In order to find the appropriate threshold level α and parameter m the $\hat{\theta}^s_{m,n}(x_i, y_i)$ are computed for different values of m in a first step. Here we chose $m = (30, 60, 90, 120, 150, 180, 210)$. As the sample size decreases due to the elimination of outliers this vector is adapted accordingly.

Next a vector of threshold levels $\alpha = (0.25, 0.35, 0.45, 0.55, 0.65)$ is chosen. For each of the combinations of m and α the percentage of points with $\hat{\theta}^s_{m,n}(x_i, y_i) \geq 1 + \alpha$ for the input direction and $\hat{\lambda}^s_{m,n}(x_i, y_i) \leq 1 - \alpha$ for the output direction is plotted against the according m values. Because B is finite the order-m efficiency estimates have to be adjusted for the Monte-Carlo imprecision. In the input case only points are considered for which $\hat{\theta}^s_{m,n}(x_i, y_i) - 1.645 * STD_{MC}(\hat{\theta}^s_{m,n}(x_i, y_i)) \geq 1 + \alpha$ holds. In the output oriented case the correction is: $\hat{\lambda}^s_{m,n}(x_i, y_i) + 1.645 * STD_{MC}(\hat{\lambda}^s_{m,n}(x_i, y_i)) \geq 1 - \alpha$.[7]

The obtained graphs show the share of DMUs outside the order-m frontier with respect to the chosen α and m values. Besides decreases with increasing m these curves should converge (at least approximatively) linearly to the percentage of points in sample with a true FDH efficiency score of one or above (Simar, 2003). The presence of potential outliers is indicated if the curves show "elbow-effects", i.e. depart strongly from linearity. The corresponding values of m and α at the strongest departure from linearity then is used as m value and α threshold for the individual investigation of potential outlier.

Furthermore because the sample is finite, the conditional distribution function $S_c(x|Y \geq y_i)$ is estimated by its empirical counterparts. In practice all units with an output value $y \geq y_i$ and

[6] Simar (2003) speaks of "Leave-one-out" efficiency which is equivalent to the super-efficiency concept proposed by Andersen and Petersen (1993), (see Walden, 2006, p. 33). It is denoted with s.

[7] 1.645 is the 0.95 percentile of the standard normal distribution.

with an input value $x \leq x_i$ are counted. This number is denoted as $N_{input}(x_i, y_i)$ for $y \geq y_i$ and $N_{output}(x_i, y_i)$ for $x \leq x_i$ (Simar, 2003). For the input value in this multivariate setting this implies that we count all units whose input vector dominates (at least weakly) the input vector of unit i. Thus we follow Roudaut and Vanhems (2004) in that the order relation between the two vectors x_j and x_i is performed element wise: $x_j = (x_{j1}, ..., x_{jp}) \leq x_i = (x_{i1}, ..., x_{ip})$ if and only if $x_{jp} \leq x_{ip}$ for all $p = (1, ..., P)$. $N_{output}(x_i, y_i)$ shows as:

$$N_{output}(x_i, y_i) = \sum_{j=1, i \neq j}^{n} 1I(x_{j,p} \leq x_{i,p}) \ , \ p = (1, ..., P) \tag{8}$$

These values provide information about the location of DMU_0 within the sample. Small values indicate that a DMU_0 is at the fringe of the data points which is likely to characterize an outlier. Applying the main idea of this method, an outlier is characterized by very low N_{input} and N_{output} values as well as large order-m input and small order-m output estimates at the chosen m and α combination. These DMUs are then excluded. Wilson (1993, 1995) points towards the phenomena that outliers might 'mask' other outliers, i.e. outliers might exist as clusters of points. Due to this "masking effect" the analysis has to be redone after the first potential outliers are excluded from the data.

While the outlier-detection is conducted before the standard DEA, the sensitivity analysis has to be done afterwards.

3.5 Sensitivity analysis for efficiency estimates

One of the main differences to stochastic models may be seen in the parametric models' implicitness of the data generating process (Simar and Wilson, 1998). This led to the criticism that they lack a statistical basis. The identification of outliers in the data already weakens this criticism. Nevertheless the robustness of the obtained estimators to potential noise in the data is still unclear.

Filling this gap, Simar and Wilson (1998, 2000*a*) present a sensitivity analysis based on the bootstrap methodology that allows for the computation of confidence intervals for efficiency measures. The complete algorithm of the method is described detailedly in Simar and Wilson (2000*a*), but we also draw on the nice presentation in Walden (2006). In the following just a very brief description is provided.

The confidence interval of the efficiency estimates $\hat{\theta}_1, ..., \hat{\theta}_n$ is to be calculated for all n DMUs.[8] Next, using a 'smoothed bootstrap' method a b-th $(b = 1, ..., B)$ random sample of size n from $\hat{\theta}_i$ is drawn with replacement. This is denoted as $\theta_{b1}, ..., \theta_{bn}$. Then the reflection method proposed by Silverman (1986) is applied. In order to 'smooth' the bootstrap the bandwidth h is estimated by minimizing the mean weighted integrated square error (MWISE). The resulting

[8] Note that θ represents here the output-oriented efficiency estimated while in Simar and Wilson (2000*a*) it indicates the input-oriented efficiency score.

'smoothed' θ-values are further adjusted for an asymptotically correct variance of the bootstrap sequence (see Simar and Wilson, 1998, p. 56) and denoted as θ_i^*.
θ_i^* is now used to generate a pseudo data set $(x_{i,b}^*, y_{i,b}^*)$ with $x_{i,b}^*$ as the original input level and $y_{i,b}^*$ as the adjusted pseudo output level using the ratio $\frac{\hat{\theta}_i}{\theta_i^*}$.[9]
For these pseudo data set the original DEA is resolved obtaining $\tilde{\theta}_k^*$. Except for the estimation of the efficiency estimates, $\hat{\theta}_i$, all steps are repeated B times for the n estimates. This means that each of the DMUs is allowed to replace a DMU_0 (x_0, y_0) sequentially. In doing so the sensitivity of distance from a fixed point (x_0, y_0) to the estimated frontier $\delta\widehat{X(y_0)}$ is analyzed. The latter is relative to the estimator's sampling variation of the frontier. Naturally the number of repetitions B conflicts with limited computational resources. Simar and Wilson (2006) suggest $B = 2000$ to ensure adequate coverage of the confidence intervals which is followed here.
For each DMU B sets of estimates are provided. Once they are generated a bias-corrected estimator $\hat{\theta}_k^{**}, k = 1, ..., n$ can be computed by:

$$\hat{\theta}_k^{**} = \hat{\theta}_k - bias\,\hat{\theta}_k \;\text{ with }\; bias\,\hat{\theta}_k = B^{-1}\sum_{b=1}^{B}\tilde{\theta}_{kb}^* - \hat{\theta}_k \;\;, \tag{9}$$

with $\hat{\sigma}_k^2$ as the sampling variance of the bootstrap values $\tilde{\theta}_{kb}^*$. Because $\hat{\theta}_k^{**}$ has a higher mean-square error than the original estimator $\hat{\theta}_k$ and contains only estimates of variance and bias, Simar and Wilson (2006) recommend to use it only if the ratio $bias\,\hat{\theta}_k/\hat{\sigma}_k \leq 1/\sqrt{3}$.
The $\tilde{\theta}_k b^*$ can be used to calculate confidence intervals. Since the true distribution is unknown for the $(1-\alpha)$ confidence interval a_α^* and b_α^* are chosen such that:

$$1 - \alpha = Prob(-b_\alpha \leq \tilde{\theta}_k^* - \hat{\theta}_k b \leq -a_\alpha) \;\;. \tag{10}$$

This is acquired by sorting the difference between $\tilde{\theta}_{kb}^*$ and $\hat{\theta}_k$ in ascending order and then deleting $(\alpha/2 * 100)$-percent of the elements from both ends of the sorted list. $-b_\alpha^*$ and $-a_\alpha^*$ are set as the endpoints of the sorted array, with $a_\alpha \leq b_\alpha$. The confidence interval of the bootstrap estimator is then given by:

$$\hat{\theta}_k + a_\alpha^* \leq \theta \leq \hat{\theta}_k + b_\alpha^* \;\;. \tag{11}$$

Except for $a_\alpha^* = 0$, the original value of $\hat{\theta}_k$ will be outside the confidence interval, since $\hat{\theta}_k$ is a downward biased estimator of θ.
On the one hand, the bias-corrected estimator in combination with the confidence intervals increase the reliably of the obtained results. But on the other hand overlapping confidence intervals make it difficult to rank DMUs. The remaining question is when do we declare a DMU as being efficient. Here we decide upon using the confidence interval of the DMU_1 with the smallest bias corrected estimator (if it lies within its confidence interval) as base. Now all DMUs whose lower confidence interval bound is smaller than the upper confidence interval

[9] In the input-orientation the pseudo input level $x_{i,b}^*$ is adjusted while the output level stays unchanged.

bound of DMU_1 and whose bias corrected estimator lies within the DMU's confidence interval are declared as being efficient.
Having introduced the employed all steps of the efficiency analysis, next the employed data is presented.

4 The employed data

4.1 Patents and industry definition

We apply the analysis to data of labor market regions which have been argued by Broekel and Binder (2006) to be the adequate regional unit for investigating regional innovation processes. The latter and the propensity to patent differ largely between industries (see e.g. DTI, 2003; Broekel and Brenner, 2005). These differences have to be taken into account by using industry specific input and output data. In contrast to Fritsch (2006) who uses total R&D employees and total patent applications, the our analysis only investigates two industries: the manufacturer of transport equipment (CAR) and the manufacturer of electrical & optical equipment (EOTEC). Both represent highly innovative industries for which patents are an important property rights protection mechanism (see Arundel and Kabla, 1998).
As an approximation of a region's innovativeness we use the patent applications, published by the *Deutsches Patent- und Markenamt (German Patent Office).* The applications are industry specific and do not include the applications by public research institutes, such as universities and research societies (e.g. Max Planck Society) as well as the patent applications by private inventors.[10]
The industry-specific patent data has to be linked to the industry classification in order to assign the appropriate R&D employees to the patent classes. The data on R&D personnel of private firms is obtained from the German labor market statistic. The number of R&D personnel (R&D) includes all engineers and natural scientists. The R&D personnel is defined as the sum of the occupational groups agrarian engineers (032), engineers (60), physicists, chemists, mathematicians (61) and other natural scientists (883) (Bade, 1987, p. 194ff.).
Lacking a definite concordance between the International Patent Classification (IPC) (by which the patent data is organized) and the industry classification (WZ03)[11] (by which the R&D employees are categorized) we employ the concordance between IPC and NACE by Schmoch et al. (2003). However, this concordance is not of definite character as well.
A *stepwise* linear regression models on district (German "Kreise")-level between patent data and R&D employees is used to find the best fitting combinations. The underlying assumption is that on a regional level the number of R&D employees of an industry as the primary resource

[10] For a detailed discussion for the use of this patent class see Broekel and Brenner (2005).
[11] The German industry classification - "Wirtschaftszweigklassifikation" of the year 2003 corresponds to the European NACE classification.

in the innovation process should correspond to the patents applications of this industries' firms located in the region. The procedure is simple: The patent application data (PAT) is organized

Industry	Code	IPC	WZ03
Manufacturing of transport equipment	CAR	10	34,35
Manufacturing of electrical and optical equipment	EOTEC	4, 26, 27, 28, 30, 31	30,31,32,33

Table 1: Definition of industries

by the 31 technological classes presented in Greif and Schmiedl (2002) and the R&D employees on the 31 2-letter WZ03 levels. Firstly, the patents of the IPC classes that are assigned to a technological field by Greif and Schmiedl (2002) are summed up and used as dependent variable. Secondly, the R&D employees of each WZ03 class that have been assigned (not definite) by Schmoch et al. (2003) to an IPC belonging to this technological field are included as separate independent variables. The stepwise mechanism of the linear regression now estimates the combination of R&D employee categories with the highest significant fit to the technological field (patent data). For the two industries presented here the best fitting combination is presented in Table 1.[12]

Recall that only one output variable, the industry specific patent applications, is included in the model while the corresponding R&D employees serve as most important input variable. The concordance between PAT and R&D is therefore likely to have a large impact on the results. This has to be evaluated later in the paper.

4.2 Regional factors

In section 2 we argued that 'co-location' studies are very important for evaluating the relevance of measurable factors which can serve as input variables here. As discussed elsewhere, (Broekel and Brenner, 2005), a great number of variables have been considered in these studies to represent important aspects of this regional endowment. Nevertheless, there are some 'core' factors that are commonly employed in such analyses. Most of them have been proposed by Feldman and Florida (1994) in defining a region's 'technological infrastructure' and by Griliches (1979) as the ingredients of the 'knowledge production function'. Building upon these findings, we define the German regional technological infrastructure consisting of factors with a purely regional effect and factors with an inter-regional effect. The first are argued to affect firms' innovation activities only if they are located in the same region. They are presented in this subsection. In contrast the effects on firms' innovativeness by inter-regional factors are to a lesser extent locally bounded. Firms might benefit from the presence of these factors in a region close

[12] This method is far from being exact. Yet a concordance between NACE or WZ03 and IPC is still subject to ongoing research.

by.[13]

In order to account for urban agglomeration advantages the population density is added (POP_DEN). The financial situation of the region as well at its economic activity is approximated by the gross domestic product (GDP) per capita. The importance of business services is well accepted in the literature (see e.g. Feldman, 1994). Therefore the variable SERVICE is added as the share of employees working in business services (WZ03 code 74) on total employment. Furthermore the potential impact of the share of employees with high qualification (EMP_HIGH) is considered, because it is a common measure of the quality of local human capital (Weibert, 1999).

As a proxy for a region's over-all endowment with private technological research facilities the share of the total R&D employment on total employment is chosen (ALL_RD). In contrast to public research institutions which are said to have an inter-regional spill-over effect (see next section), we consider their spill-over potential to be regionally bounded.

As has been argued before, the assignment of the R&D employees to the patent classes is likely to be incorrect. Therefore we include the the sum of all R&D employees of those industries that are technologically related to the investigated industry (REL_...) as share on total employment.[14] These are defined as all different industries' R&D employees that are assigned to the IPC of the investigated industry by Schmoch et al. (2003), except the ones that are identified as this industry's R&D employees. They also account for the potential of technological spill-overs by local firms. Their exact definition is shown in Table 2. In contrast to technologically related

Industry	Patent classes	Related industries' WZ03 code	Variable
CAR	10	19, 25, 31, 36	REL_CAR
EOTEC	4, 26, 27, 28, 30, 31	23, 24, 25, 27, 29, 34, 36	REL_EOTEC

Table 2: Technologically related industries

industries, supplier and customer industries may enhance the innovativeness of local firms by direct orders, or process and product innovations that are inter-industrially transfered, e.g. by the knowledge embedded in products (see e.g. Schibany and Schartinger, 2001*b*; Arndt, 2000; Kohlbecher, 1997). This is likely to be independent from these industries' own R&D activities in the region. Rather their total presence seems to be the better proxy. The variables SUP_CAR for the CAR industry and SUP_EOTEC as well as CUS_EOTEC are added. They are defined as the share on total employment of the most important customer / supplier industry of the industry investigated. From the German national accounting of the year 2000 we derived for EOTEC as most important supplier industry (apart from EOTEC itself) the 'Other business activities' (WZ03 74) which are mostly business service firms. But they are included as separate variable

[13] All variables are only presented very briefly due to the space limitations. A more detailed description and literature references of most of the employed variables can be found in Broekel and Brenner (2005).

[14] The three dots (...) are placeholders for the name of the industry that is analyzed.

SERVICE anyway. Next is the industry 'Wholesale on a free or contract basis' (WZ03 51) which is unlikely to be source of technological spill-over. We include the third most important industry 'Manufacture of fabric metal products, except machinery and equipment' (WZ03 28) as important supplier industry EOTEC_SUP.
As most important customer industry (apart form EOTEC itself) clearly the 'Manufacture of machinery and equipment n.e.c.' (WZ03 29) are chosen.
A similar situation is found for the CAR industry. On the supply side the CAR industry itself is most important followed by the SERVICE industry. For other reasons both are already included. As variable CAR_SUP the total employment of the third most important supplier industry is chosen instead, the 'Manufacture of fabric metal products, except machinery and equipment' (WZ03 28). We are not able to detect an important customer industry for CAR because it seems to be a real end-user industry. The concordance of the supplier and customer

Industry	WZ03	created variable
Manufacturing of metal goods	28	CAR_SUP
Manufacturing of metal goods	28	EOTEC_SUP
Manufacturing systems engineering	29	EOTEC_CUS

Table 3: Top customer and supplier industries

industries is presented in Table 3.
As most important characteristics concerning the presence of the industry in a region, we include the average firm size as employees per firm (SIZE_...) and the share of the industry's employees (EMP_...) on total employment.
These eleven variables are argued to affect only firms if they are located within the same region. The factors presented in the next section are regionally less bounded and therefore deserve a special treatment.

4.3 Inter-regionally effective factors

One problem occurs when approximating regional factors and resources: How to deal with resources that are locally created but whose effect is not totally restricted to that region? For example university graduates which are argued to be important for firms' innovativeness do not stay within the region where they obtain their degree (Mohr, 2002). Similarly the knowledge spill-overs of research institutes are not totally bounded to the region they are located in (Bottazzi and Peri, 2003).
In order to assign such interregional resources with an adequate share to all of the regions they are effective in, we employed an exponential distribution procedure. In the case of the graduates we define two parameters that determine their likelihood to move to a certain location. First, the distance from the location of the university they obtained their degree, and second the population of the destination region. There are two assumptions being made: The greater the

distance the less likely a graduate moves to that region. Second, the larger the population in the destination region the higher is his likelihood to move. An exponential function is assumed to calculate the probability of the movement. The parameters of the function are fitted by a maximum likelihood calculation using geographic coordinates and population counts for 8.196 German five digit postal code areas. Insights from the literature on the mobility of graduates are used additionally for the calculation (see Legler et al., 2001; Mohr, 2002). Because the mobility parameters are only available for the graduates of technical colleges and universities separately, we had to weight them with their shares in the sum of all graduates with a degree in the same subject. Note that the α value which determines the slope of the exponential function is estimated using the combined data for 1999/2000 because no year specific data on the mobility parameters is available. Therefore the calculation has to be as general as possible.

To control for size effects regarding a region's industry endowment, the distributed graduate counts enter the analysis as ratios to the region's total employment. For technological innovations the graduates of engineering and natural science faculties of technical colleges as well as of universities are of special interest and included as variables GRAD_ENG and GRAD_NAT.

Similarly to universities public research institutes are important actors in innovation processes (Nicolay and Wimmers, 2000). Beise and Stahl (1999) show that a significant part of their influences takes inter-regionally effect. While their research output is highly codified (Schibany and Schartinger, 2001a) and effected only weakly by geographic distance their employees' mobility which constitutes an important knowledge transfer mechanism is surely spatially restricted (Witt and Zellner, 2005; Broekel and Binder, 2006). Thus we argue that their positive effect on firms' innovativeness decreases with growing distance. For the analysis the effect of public science institutions is modeled similar to the graduates' one.

In order to approximate their influence, the structural factor SCIENCE is chosen. It consists of the summed up number of employees working at the different regional (public) research institutions. These are mainly the 'big four' institutions in Germany: the *Helmholtz Association*, the *Max Planck Society*, the *Fraunhofer Society* and the *Leibnitz Association*.[15]

Identically to the graduates we assume that the effect of the public research institutes decreases exponentially with growing distance. Assigned to postal codes they are distributed with the same procedure as the graduates. The relevant parameters for the distribution procedure are taken from the findings of Beise and Stahl (1999). Table 6 summarizes the distribution parameters. Eventually SCIENCE, GRAD_ENG and GRAD_NAT are computed as ratios to total employment to control for differences in the regions' size.

To sum up we have 14 regional input variables, besides R&D employment, that can enter the efficiency analysis. They are presented in Table 7. Before the results are presented and evaluated in detail the procedure of the analysis is described in the next section.

[15] All institutes are included in this number, even those that are not technological relevant such as institutes doing research on language or history.

5 Procedure of the analysis

5.1 Data preparation

In a first step all input variables are multiplied with the R&D employees of the according industry, see Section 3. As a by-product this raises the correlation between the output and the input variables (because the R&D employees are highly correlated with the according patent data). This increases the stability and robustness of the efficiency results: "High covariance between inputs therefore allows DEA to yield better estimates of efficiency" (Smith, 1997, p. 247). The multiplication can also be considered as a safeguard against the exclusion of relevant variables. Their omission biases the results less when they are correlated (which is true through the multiplication) with other variables included (see Smith, 1997; Ruggiero, 2005). While the model becomes more realistic, its statistical properties are enhanced as well.

The correlation structure of the variables has an effect on the estimated efficiency scores as well (Smith, 1997). Recently, some researchers tackled the question if there is a correlation coefficient threshold that beyond which omission of one or more input vectors have no statistically significant effect on the efficiency mean (see Pedraja-Chaparro et al., 1999; Smith, 1997; Saen et al., 2005; Ruggiero, 2005). We follow Saen et al. (2005) which derive clear cut benchmarks for Banker-Charnes-Cooper DEA model which is employed in this paper: If two input variables are correlated with or above 0.96, the omission of one of these input vectors has no statistically significant effect on the efficiency mean.

Applying this criterion to the data set after the multiplication with the R&D employees the variables that have been excluded are listed in Table 4. From the two variables that are correlated with 0.96 or above with each other the one with smaller theoretical relevance is excluded. In case of EOTEC EMP_HIGH is excluded in favor of leaving GDP in the analysis. This decision is based on the results of Broekel and Brenner (2005) that found variables approximating agglomerations and the region's financial situation of higher relevance for the electrical and optical industry.

For CAR 1999 and CAR 2000 GRAD_ENG clearly seem to be more important for this heavily engineering based industry than the graduates of natural science faculties (GRAD_NAT). Since

Industry / year	excluded variables
CAR	
1999	GDP (EMP_HIGH), GRAD_NAT (GRAD_ENG)
2000	GDP (EMP_HIGH), GRAD_NAT (GRAD_ENG)
EOTEC	
1999	EMP_HIGH (GDP)
2000	EMP_HIGH (GDP)
In parenthesis the variable that caused the exclusion.	

Table 4: Excluded variables

this industry is less attracted by agglomerations (Broekel and Brenner, 2005), the endowment with highly qualified persons (EMP_HIGH) stays within the analysis while GDP is excluded.
In a second step we exclude all regions in which the R&D employees or the patent data show zero values. An efficiency analysis including such cases is obviously senseless. In Table 8 these regions are marked by an "NA".
The resulting data set is than investigated with the outlier detection method which has been introduced in Section 3.4. Its application and outcomes are presented in the next section.

5.2 Outlier identification

As a result of the input variables' multiplication with the R&D employees the model is strongly influenced by the relationship between the number of R&D employees and the corresponding patent data. This relationship is however strongly determined by the underlying concordance between patent data and industry classification. As has been pointed out before this concordance is not unproblematic. Indeed, when looking at the ratio between patent applications (PAT) and R&D employees of the industry under consideration (R&D) and sorting it in decreasing order of the number of R&D employees two patterns can be observed: While on the one hand there exists massive disturbance at the lower end of R&D, it is strongly reduced as the number of R&D employees increases. Because the scatterplots do not change much for the two years, Figure 1 shows exemplarily only the scatterplots for CAR_2000 and EOTEC_1999. Note that for better illustration the x-axis represents the logarithm R&D. This observation indicates that

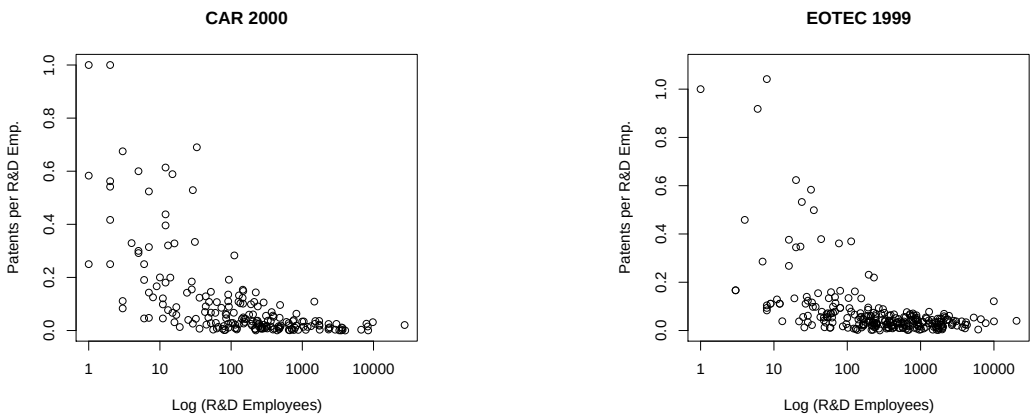

Figure 1: Patents per R&D employee before outlier elimination

the model (and especially the concordance of IPC and WZ03) seems to work well for regions with comparatively large R&D in a specific industry. In contrast in the case of small R&D the assignment of patent data to industries did not yield acceptable results.
This suggests itself to cut off all regions with low R&D from the sample. But at which R&D employment number the line should be drawn and how to define this breaking point? Lacking an adequate method we rely on the outlier-detection method by Simar (2003) which has been

introduced in Section 3.4. Due to the very nature of the model regions with a high ratio between R&D and PAT are likely be identified as potential outliers. As can be seen in Figure 1 they are however not scattered randomly across the sample, but rather concentrate on the lower bound of R&D. For the outlier detection method this means that such a clustering backs the likelihood of observing "masking effects" (Wilson, 1993), i.e. an outlier masks another. Therefore the method needs to be run several times.

The outlier detection method 'attacked' the cluster of regions with high PAT / R&D ratio and small R&D right from the start. Furthermore the masking was confirmed as well. Taken together, a clear pattern can be observed that becomes weaker and weaker with more and more rounds of the outlier detection method being run and potential outliers being excluded: The procedure identifies regions as potential outliers that show two characteristics, small R&D (N_{output} close to zero) and a high PAT / R&D ratio (very low order-m efficiency scores in the output orientation).

Concerning the chosen parameters for most rounds and industries the parameter combination of $m = 60$ and $\alpha = 0.55$ was most suitable. Besides choosing the best parameters another problem arose in the input-orientation. Due to the subtraction of the $1.645*$ Monte-Carlo imprecision from the estimate negative order-m scores are obtained. This forced us to rely stronger on the output orientation results. The effect however disappears as more and more outliers are excluded.

For EOTEC_1999 we had to run the procedure eleven times until the described pattern disappeared. At his point the identified regions are not anymore characterized by small R&D as well as output and input-orientation order-m efficiency scores above any acceptable range. For EOTEC_2000 even fourteen rounds were needed. In the cases of CAR_1999 and CAR_2000 ten rounds were sufficient. As a result the variance of the ratio between patents and R&D employees flattened. The scatterplot of EOTEC_1999 and CAR_2000 are presented exemplarily in Figure 2. Note that we excluded far more observations than the upper bound of $\frac{\sqrt{n}}{n}$ proposed

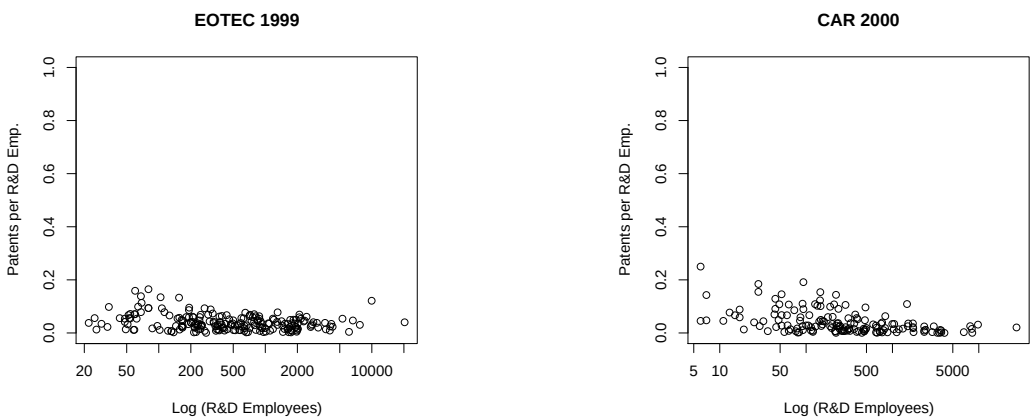

Figure 2: Patents per R&D employee after outlier elimination

by Barnett and Lewis (1995) and in this context suggested by Simar (2003). This is justified by the specific structural problem of the data resulting from the assignment of the patent data to industries' R&D employment. In fact for EOTEC_1999 we excluded 33 of the 42 smallest regions in terms of R&D employees. Similar relations are obtained for the other industries, see Table 8. The justification for this is that the outlier-detection method has been employed not only to detect random outliers, but also to cut off the extreme lower end of the sample without having to choose a cut-off point ex-ante. Without the regions that have been cut off at the lower end of R&D employment the suggested maximal share of outliers holds for all four industries.
In Table 8 excluded outliers which are characterized by small R&D (below 50) are marked with OL. Outliers with high R&D are denoted by OH. All other excluded outliers are marked by O.

6 Results and discussion

6.1 Results and inter-temporal stability

Applying the analysis to data of two consecutive years seems to be not appropriate for analyzing dynamic effects. But it certainly allows to investigate the inter-temporal stability of the results and their sensitivity to slight changes in the data. This is of uttermost importance because the outlier detection method especially is not objective but involves the (informed) decision of the researcher. Therefore we will concentrate on the (stable) findings for the two years and discuss year specific findings only very briefly.
As a first impression the bias corrected estimates of the two industries for 1999 are compared with the ones of 2000. Figure 3 depicts the bias-corrected efficiency score of each region in the year 1999 on the x-axis and the ones in 2000 on the y-axis. For better illustration the logarithm has been taken on both axis. In the plot three patterns deserve attention: A potential clustering in the south west corner[16], a potential linear trend of the data as well as points that are arranged parallel but remote to the axes.
The latter represents all points that have been excluded in one year, either because they have been identified as outliers or because of zero values. The number of points within these structures hints at the performance of the analysis on the extreme ends of the data (small values of R&D). For CAR these structures contain 37 points. Thus 37 regions are part of the final efficiency analysis in one year but not in the other. Note that 110 regions are not part of the analysis in both years. They do not show up in the plot. At the other end, 23 regions have been identified as efficient in both years (in contrast to 41 in each year). Note that the region with the lowest value in one or more input variables is always identified as efficient in VRS models. Such regions are left aside here, because they gain their efficiency only from a methodological effect.

[16] Note that bias-corrected estimates are used, i.e. they show values larger than one.

For the evaluation of the time stability we subtract 12 regions from the 41 efficient. They are excluded in one year as outliers with low or zero R&D or zero PAT. This seems to be reasonable in the light of the previously mentioned disturbance at the lower bound of R&D. A ratio of 23 efficient regions in both years to 29 regions being efficient in the single year case is derived, which is roughly eighty percent. It means that eighty percent of the regions identified as efficient in one year are also efficient in the consecutive year. In the light of the sensibility of the analysis to changes in the data this is a very high value.

About 100 points varies roughly around a linear trend. Taking the 123 cases into account that are valid for both years round about 18 percent of these regions are claimed to be efficient. If one is willing to take all 270 regions as base the percentage drops below ten percent. Note also that some of the 23 regions are also neighbors which could indicate that the claimed 'regional phenomenon' that fosters innovativeness, might span across more than one labor market regions. This has to be investigated further.

The evaluation of the outlier detection method's performance in relation to the regions being identified as efficient can be done simultaneously. We take the number of regions that have been excluded as outliers in one year but which have been identified as efficient in the other. Without the cases in which the efficient region has a minimum in the input values and the cases in which the region is excluded due to zero values 7 cases are found. Only one region is not located at the lower bound of R&D employment. But still this region (172) has only 72 R&D employees which is comparatively low as well.

From a qualitative view the regions that have been identified in both years as likely hosts of 'regional phenomena' do not seem to be totally 'abstruse' cases. Not surprisingly the most prominent locations of the German car manufacturing industry are identified as efficient regions. For example Wolfsburg and Braunschweig host large facilities by VW, the same holds for Stuttgart and Munich with the headquarter of $Daimler - Chrysler$ / $Porsche$ and respectively BMW. Such large firms often centralize their patent applications at the region of their headquarter. These patents are not generated in that region, but will show up as the headquarters region's output. Although the average firm size is included as input variable, a closer look at the patent data and size of local firms should flag any decision to conduct a case study for one of these regions. Therefore we argue that regions identified as efficient and show extreme large R&D numbers should be handled with care. Likely an effect that has nothing to do with 'regional phenomena' processes, as e.g. the centralization of patents by large firms, causes outstanding patent application counts.

In 'identifying' these regions the analysis does not provide a value added, it rather proves its validity. More interesting are such cases with reasonable high R&D employment as e.g. Coburg (741) [17], Regensburg (826), Reutlingen / Tübingen (487), Aalen (1725) and Bielefeld (309). In pointing at these regions, the true value added of the analysis is illustrated. In these regions,

[17] R&D employment in 2000 in parenthesis.

certain structures may develop / exist that allow actors to employ regional resources more efficiently than in other regions. A case study comparing these regions with each other and with randomly chosen other (but similar) regions that have not been identified as efficient seems therefore valuable. The analysis of the results for EOTEC show very similar patterns with even

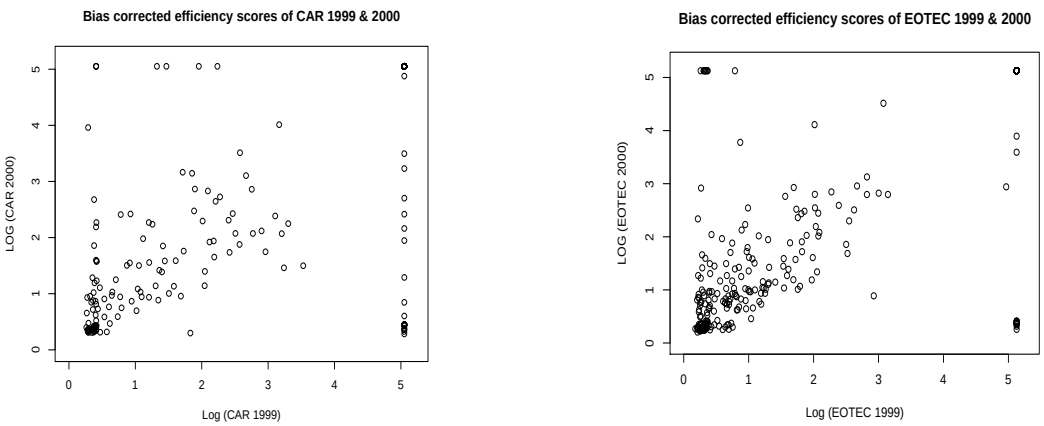

Figure 3: Year to year comparison of bias corrected efficiency scores.

better relations as for CAR. In Figure 3, only 27 regions are arranged parallel but remote to the axis, i.e. are part of the analysis in only one year. Additional 40 regions are excluded in both years (they do not show up in the plot) leaving 203 valid regions. The number of regions that have been identified as outliers in one year and as efficient region in the other year, without the cases in which the efficient region has a minimum input value, and without the exclusion due to zero values, is 12. In only five cases this is true for regions with large or moderate R&D employee numbers. All of them show extreme high ratios PAT and R&D. This has to be investigated separately in the future.

27 regions are identified as efficient in both years (in contrast to 51, respectively 55 in each year). Furthermore 7 regions are efficient in one year and are excluded as outliers with low or zero R&D , or zero PAT in the other year. The latter results from the disturbance at the lower bound of R&D employee numbers, see Section 5 and therefore should not be part of the evaluation. Nevertheless the time stability of the results is weaker to CAR: about 50 percent of the regions identified as efficient in one year are also identified as efficient in the consecutive year (27 of 44 respectively 48 regions). Taking the high sensibility of the analysis to changes in the data into account this still seems to be a very reasonable ration. What can be said for sure is that clearly a pure single year analysis does not seem to be adequate for this context.

Similar to CAR the mere quantitative evaluation of the results needs to be backed by a qualitative evaluation. Compared to CAR EOTEC is less dominated by large corporations. *Siemens* certainly plays a very large role. But the region hosting its headquarter (München (Munich)) is only found to be efficient in 2000. This has to be checked in detail.

Efficient regions with large R&D employment in EOTEC are Erlangen (6705), Nürnberg

(8054), Regensburg (2759), Hildensheim (2311) and Hannover (2065). Such large R&D numbers likely result from the presence of large corporations. Again we rather argue that this are not the interesting cases. Rather regions with medium R&D employment seem to be valuable cases for case study research aiming at the corporative behavior and networks among the regional actors of the EOTEC industry. Such regions are: Essen (869), Krefeld (459), Mönchengladbach (579), Bonn (1270), Borken (515), Bielefeld (817), Paderborn (972), Aalen (1169), Traunstein (944), Weilheim (394), Passau (430), Jena (944) and Potsdam-Brandenburg (371).
Thirteen regions are still too many for conducting comparative case study research, so another selection method needs to be applied. But checking the locations of the largest corporations in this industry will probably help to reduce the number of interesting cases.
Summarizing we find reasonable regions being identified as efficient for CAR as well as for EOTEC. Yet some points concerning the applied analysis and data need to be taken into consideration when taken these results as basis for further research. They are discussed in the following subsections.

6.2 Discussion of analysis and results

The model proposed in Section 3 is characterized by a very strong emphasis on the non-substitutive character of the R&D employees. Thus the question is to what extent the relationship between R&D employees and patent data drives the results. For CAR 1999 and EOTEC 2000 Figure 4 exemplarily shows the position of efficient regions in comparison to all the other regions ratio of PAT to R&D. The solid points mark the efficient regions and the circles all other regions except outliers and regions with zero values. From the figures which look similar for

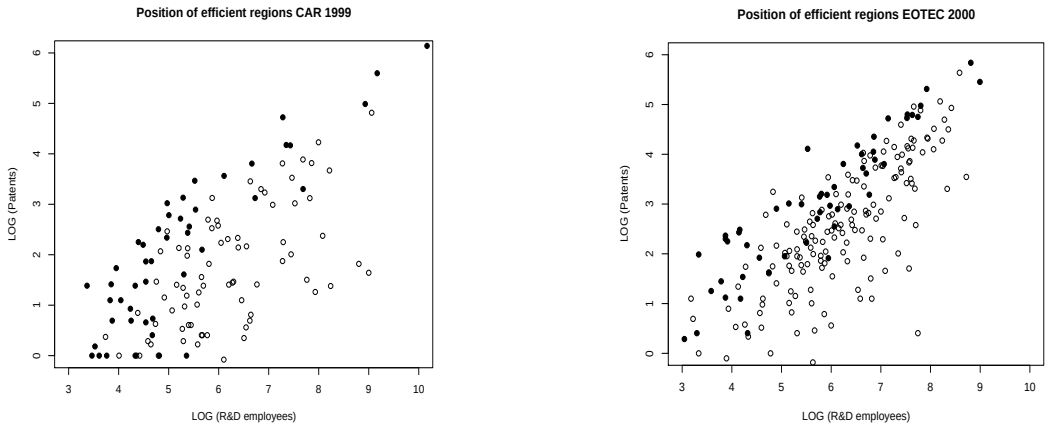

Figure 4: Efficient and not-efficient regions ratio of R&D employees and patents

CAR 2000 and EOTEC 1999, two points can be derived. Firstly, the solid dots (efficient regions) cluster closely to the upper frontier of the points in both plots. It indicates that the likelihood of identifying a region as being efficient is high if this region shows a large ratio between patent applications and R&D employment. This is not coming as a surprise, it just confirms the model

specifics. Secondly, relying only on the ratio and declaring regions with the largest ratio as the ones that likely contain a 'regional phenomena' is insufficient: Not all regions with a high PAT / R&D ratio are identified as efficient regions. Quite a number of regions shows additionally comparatively weak ratios and are still found to be efficient. The other input variables influence the results also significantly. The latter represents the true value added of the analysis.

Nevertheless the model's specification with the strong emphasis on the PAT / R&D ratio might be valuable to be relaxed in the future. This however includes to drop the non-substitutability of R&D employment which is not unproblematic. The question of how to define the extent to which substitutability is allowed, is far from being easy to answer theoretically and empirically. As long as this is not solved assuming non-substitutability seems to be the better choice.

While the results of the DEA are reasonable and fairly reliable some general problems concerning the data and analysis need to be discussed at the end. Most importantly for many regions that were not found to be efficient we derived very large bias-corrected efficiency scores and negative confidence intervals. This can be caused by the too small sample with respect to the employed number of input and output variables, i.e. the 'curse of dimensionality' problem. While there is no way of increasing the number of observations (labor market regions are already comparatively small) decreasing the number of employed variables is problematic as well. From a theoretical perspective using only 10 to 12 variables represents a strong reduction of regional innovation processes complexity. For future work, the methods by Simar and Wilson (2000*c*) and Mouchart and Simar (2002) may offer ways of reducing the number of input variables. New robust nonparametric methods based on the order-m frontier idea might also resolve this issue as well (see e.g. Daraio and Simar, 2005, 2006). Also the relaxation of the convexity assumption that is inherent to DEA models needs to be further discussed and possibly relaxed, too. In the case of a relaxation, a free disposable hull (FDH) estimator can be employed, either in the traditional way or as robust order-m estimator. Certainly there are enough subjects for future research.

The great variance in the efficiency estimates might also results from the large heterogeneity of the regions under examination. This seems to be a specific problem of regional researchers. In contrast to researchers interested in firm level data, regional research are always faced with very low sample size (due to the limited nature of geographic space) which does not allow for very much sub-sampeling. Again the approaches cited above might help to overcome this problem in the future. Note however, that some of this variance is caused by the assignment of the patent application data to industry specific variables. The employed concordance is far from exact and is subject to ongoing research that will increase its performance and by this the applicability of this type of analysis.

Apart from these problems the analysis seems to offer a very helpful way of gaining new insights in regional innovation research. Most importantly we found some stable patterns in the innovativeness structure of German regions with respect to time and industry under observation.

Taking the very limited number of ex-ante specifications and assumptions underlying the analysis into account the obtained results seem to be reliable. Thus while it is very bold to say that all the identified regions are very likely be characterized by a 'regional phenomenon', it seems to be acceptable to suggest not to search for such phenomena in regions not identified as efficient with an analysis similar to the one in this paper.

7 Conclusion

In regional innovation research two very different approaches have been put forward to investigate the variance in regions' innovativeness. On the one hand the 'co-location' approach concentrates on analyzing the influence of regional factor endowments on the local firms' innovation activities. In general an empirical framework is chosen that sets the measurable regional factor endowment in relation to local firms' innovation counts. Most prominently a (regional) 'knowledge production function' model is employed in these kind of studies.

On the other hand researchers have argued that some regions have acquired a specific feature that allows regional actors to be more innovative. The development of such a feature results from region specific historic, social and geographical characteristics and processes. The well-known 'innovative milieu' can be considered as such a 'regional phenomena' that exists in a very limited number of regions.

This paper argues that both approaches have to be combined in order to gain a more realistic understanding of the variance in regions' innovativeness'. We propose the use of efficiency analysis as method that takes both approaches into account. Furthermore we present a semi-deterministic nonparametric data envelopment analysis as a promising method to take on this task.

By analyzing data of German labor market regions for two different industries and years the advantages and drawbacks of this type of analysis are discussed and evaluated. While there exist certainly grave problems that need to be overcome in future work, first results seem to support the application of efficiency analysis in this context.

Appendix

A Employed variables and results

Distance	<50km #	50<200 km #	share 1999	share 2000
GRAD_ENG (university)	46.6%	23.6%	41.6%	40.7%
GRAD_ENG (tc)	41.5%	34.7%	58.4%	59.3%
GRAD_NAT (university)	57.3%	13.9%	89.7%	89.8%
GRAD_NAT* (tc)	44.1%	35.0%	10.3%	10.2%
* No data available, the shares of the overall technical graduates are used. # Data from Legler et al. (2001) p. 120-121				

Table 5: Graduates Mobility

	empirical values		estimation		
spill-over source	**<50km**	**200km<**	**<50km**	**200km<**	α
GRAD_ENG	43,6%	30.2%	56.0%	43.5%	0.0336
GRAD_NAT	56.0%	16.1%	69.8%	30.1%	0.0445
SCIENCE #	36.0%	34.0%	49.5%	49.4%	0.0287
Estimation based on sum of 1999 and 2000 data. # Numbers are approximated from figures in Beise and Stahl (1999).					

Table 6: Range of spill-overs of R&D institutes

Variable	Description	Source
EOTEC	patents	(Greif and Schmiedl, 2002)
CAR	patents	(Greif and Schmiedl, 2002)
RD_EOTEC	R&D employees EOTEC	
RD_CAR	R&D employees CAR	
SIZE_EOTEC	average size of firms EOTEC	
SIZE_CAR	average size of firms CAR	
EMPLOY	total employees	
EMP_HIGH	share of employees with high qualification	(INKAR, 2002)
GDP	gross domestic product per inhabitant	(INKAR, 2002)
EOTEC_EMP	share of employees working in EOTEC	
CAR_EMP	share of employees working in CAR	
EOTEC_SUP	share of employees working in manufacturing of metal goods industry	
CAR_SUP	share of employees working in manufacture of fabric metal products	
EOTEC_CUS	share of employees working in manufacturing systems engineering industry	
POP_DEN	habitants per $kilometer^2$ land area	(INKAR, 2002)
SERVICE	share of employees working in 'other business activities' (business services) industry	
ALL_RD	share of R&D employees	
REL_EOTEC	share of related industries' R&D employees	see Table 2
REL_CAR	share of related industries' R&D employees	see Table 2
GRAD_ENG	engineering graduates per employee	
GRAD_NAT	natural and science graduates per employee	
SCIENCE	researchers at public research institutions per employee	
All shares refer to total employment.		

Table 7: Variables, estimation base and sources

		CAR		EOTEC	
AMR	AMR_Nr	1999	2000	1999	2000
Husum	1	NA	NA	OL	
Heide	2	NA	NA	**	
Itzehoe	3	OL	**		OL
Flensburg	4	OL	**		
Lübeck	5	*	*		
Kiel	6				
Ratzeburg	7	OL	OL	OL	*
Hamburg	8				
Braunschweig	9	*	*		
Salzgitter	10				
Wolfsburg	11	*	*	O	O
Göttingen	12	OL	*		

Goslar	13	OL	OL	*	
Helmstedt	14	NA	NA	OL	OL
Einbeck	15				
Osterode	16	NA	NA		
Hannover	17			*	*
Sulingen	18			OL	OL
Hameln	19	OL	**		
Hildesheim	20	OL	OL	*	*
Holzminden	21	OL	OL		
Nienburg	22	OL	OL		
Stadthagen	23	OL	OL	*	
Celle	24	OL	*	OL	OL
Lüneburg	25	OL	OL		
Zeven	26	NA	OL		
Soltau	27	OL	*	*	
Stade	28	*	*	OL	O
Uelzen	29	OL	NA	OL	OL
Verden	30	NA	NA	OL	OL
Emden	31	NA			
Westerstede	32	NA	NA	OL	OL
Oldenburg	33		OL	O	*
Osnabrück	34			*	O
Wilhelmshaven	35	**		OL	*
Cloppenburg	36	NA		OL	NA
Lingen	37	**	*	OL	
Nordhorn	38	NA	NA	OL	OL
Leer	39	OL	OL		
Vechta	40			NA	OL
Nordenham	41	NA	*	O	O
Bremen	42				
Bremerhaven	43	*		*	**
Höxter	44				
Düsseldorf	45				*
Duisburg	46	*	*		
Essen	47	OL	OL	*	*
Krefeld	48	*		*	*
Viersen	49				
Mönchengladbach	50	OL	NA	*	*
Heinsberg	51	NA	NA		
Wuppertal	52	*	*	*	
Schwelm	53	*	*		
Remscheid	54			*	O
Kleve	55	OL	NA	*	*
Aachen	56				*
Köln	57				*
Leverkusen	58	**		OL	OL
Bonn	59			*	*
Düren	60	*		OL	OL
Euskirchen	61	OL	OL	*	OL
Gummersbach	62		*		
Gelsenkirchen	63				

Münster	64	*			
Borken	65	OL	OL	*	*
Steinfurt	66		*		
Bielefeld	67	*	*	*	*
Gütersloh	68				*
Detmold	69				
Minden	70				
Paderborn	71			*	*
Bochum	72				
Dortmund	73	O	O		
Hagen	74	OL	OL		*
Lüdenscheid	75				
Meschede	76				
Siegen	77	*		*	
Olpe	78	*			
Soest	79	OL	*		
Korbach	80	OL	**		*
Kassel	81				
Eschwege	82	NA	OL	OL	OL
Schwalm-Eder	83	OL	OL		*
Hersfeld	84	OL	OL		
Marburg	85			O	
Lauterbach	86	OL	OL		
Fulda	87	*			
Wetzlar	88				
Gieŝen	89				
Limburg	90	*			
Wiesbaden	91	*	*		*
Frankfurt/Main	92			*	
Hanau	93	O	O		
Darmstadt	94	OL	OL		
Erbach	95	NA	OL		
Altenkirchen	96	NA			
Montabaur	97	*		OL	*
Neuwied	98	OL		*	*
Ahrweiler	99	*		*	*
Koblenz	100				
Bad Kreuznach	101	NA	*		
Idar-Oberstein	102	*	NA	NA	OL
Cochem	103	NA	NA	NA	NA
Simmern	104			NA	*
Trier	105			*	*
Bernkastel-Wittlich	106	OL	NA	OL	*
Daun	107	NA	NA		NA
Bitburg	108	OL	OL	**	*
Kaiserslautern	109				
Landau	110			*	
Mainz	111	OL	OL	O	O
Alzey-Worms	112	OL	OL		
Pirmasens	113	OL	OL		*
Ludwigshafen	114				

Germersheim	115	NA			
Merzig	116	NA	NA	NA	OL
St. Wendel	117	OL	OL		
Saarbrücken	118				
Homburg/Saar	119	NA	NA		
Stuttgart	120	*	*		*
Göppingen	121				
Heilbronn	122				
Schwäbisch Hall	123	*	*		
Tauberbischofsheim	124				
Heidenheim	125		*		
Aalen	126	*	*	*	*
Baden-Baden	127				
Karlsruhe	128				
Heidelberg	129				
Mannheim	130				
Mosbach	131	*			
Pforzheim	132				
Calw	133	*	*		
Freudenstadt	134	OL	OL		
Freiburg	135	*	*		
Offenburg	136	*	*		*
Rottweil	137				
Villingen-Schwenningen	138	O	**		
Tuttlingen	139	NA	NA		
Konstanz	140	OL	OL		
Lörrach	141	OL	OL		
Waldshut	142	OL	OL		
Reutlingen/Tübingen	143	*	*		
Balingen	144	OL	OL		
Ulm	145				
Biberach	146				
Friedrichshafen	147				
Ravensburg	148				
Sigmaringen	149	OL	OL		
Bad Reichenhall	150	NA	NA	O	O
Traunstein	151	NA	OL	*	*
Burghausen	152	OL	NA	OL	OL
Mühldorf	153				
Rosenheim	154	OL	OL	*	*
Bad Tölz	155	OL	OL	O	*
Garmisch-Partenkirchen	156	OL	NA	OL	OL
Weilheim	157	*		*	*
Landsberg	158	NA	NA	O	O
München	159	*	*		*
Ingolstadt	160				
Kelheim-Mainburg	161			*	
Landshut	162				
Dingolfing	163				
Eggenfelden/Pfarrkirchen	164	OL	OL	OL	OL
Passau	165			*	*

Freyung	166	OL	OL		
Regen-Zwiesel	167	NA	OL		
Deggendorf	168	*			*
Straubing	169	**	*		
Cham	170	OL	OL		
Regensburg	171	*	*	*	*
Schwandorf	172	O	*	*	*
Amberg	173	OL	OL		
Neumarkt	174	OL	OL	*	
Weiden	175				
Marktredwitz	176	NA	NA		
Hof	177	OL	OL		
Bayreuth	178	**			
Bamberg	179				
Kulmbach	180	NA		OL	OL
Kronach	181	OL	OL	OL	
Coburg	182	*	*	*	*
Lichtenfels	183	NA	NA	*	
Erlangen	184	OL	OL	*	*
Nürnberg	185			*	*
Weiš̈enburg-Gunzenhausen	186	NA	NA		*
Ansbach	187	O	O		
Neustadt/Aisch	188	NA	*	*	
Kitzingen	189	NA	NA		OL
Würzburg	190	*	*		
Schweinfurt	191	O	O	*	*
Haš̈furt	192	NA	NA		
Bad Neustadt/Saale	193	**			
Bad Kissingen	194	*	*	OL	OL
Lohr am Main	195	NA	NA		*
Aschaffenburg	196	*			
Donauwörth-Nördlingen	197				
Dillingen	198	OL	OL	OL	*
Günzburg	199	OL	OL	*	
Augsburg	200	*	*		
Memmingen	201	*			
Kaufbeuren	202	OL	OL		
Kempten	203	NA	NA		
Lindau	204			*	
Pasewalk	205	NA	NA	NA	OL
Greifswald	206	NA		*	*
Stralsund	207	NA	*	NA	NA
Bergen	208	NA	NA	NA	NA
Neubrandenburg	209	NA			
Waren	210	NA	NA	NA	OL
Güstrow	211	NA	NA	OL	**
Rostock	212			OL	OL
Wismar	213		NA		
Schwerin	214	NA	OL	*	
Parchim	215	OL	OL	OL	OL
Salzwedel	228	**	NA	OL	OL

Stendal	229	NA	NA	OL	NA
Burg	230	NA	NA	NA	NA
Magdeburg	231		*		
Halberstadt	232	OL	*	*	
Staš̈furt	233	NA	OL	OL	**
Schönebeck	234	NA	NA	NA	OL
Dessau	235	*	NA	*	
Wittenberg	236	OL	NA		NA
Sangerhausen	237	NA	NA		NA
Halle	238	*		O	*
Bitterfeld	239	NA	NA	OL	OL
Naumburg	240	NA	NA	*	*
Erfurt	241	NA	NA		
Weimar	242	NA	OL	*	O
Gera	243		NA		
Jena	244	NA	NA	*	*
Suhl	245	*	*		
Eichsfeld	246	NA	NA	NA	
Nordhausen	247		NA		
Eisenach	248				
Mühlhausen	249	NA	NA	NA	
Sondershausen	250				
Meiningen	251	NA			
Gotha	252	NA			
Arnstadt	253	NA	OL		
Sonneberg	254	NA	NA		
Saalfeld	255	NA	NA		
Pös̈neck	256	NA	NA		
Altenburg	257	NA	**	OL	OL
Leipzig	258	OL	**		
Torgau/Oschatz	259	NA			
Grimma	260	NA	NA	*	
Freiberg	261				
Chemnitz	262				
Annaberg	263				
Zwickau	264				
Plauen	265				
Dresden	266			*	
Riesa	267	NA	NA		
Pirna	268		NA		
Bautzen	269				
Görlitz	270			*	
Löbau-Zittau	271	OL	NA	**	OL
Berlin (n)	300	*	*	OH	OH
Potsdam-Brandenburg (n)	301	OL	*	*	*
Cottbus (n)	302	NA	OL		
Frankfurt/Oder (n)	303	OL	OL	*	
Eberswalde (n)	304	NA	OL	OL	OL
Luckenwalde (n)	305				
Finsterwalde (n)	306	NA	NA		
Oranienburg (n)	307				

Neuruppin (n)	308	NA	NA		**
Perleberg (n)	309	NA	NA		NA
Prenzlau (n)	310	NA		**	**
	Regions	270	270	270	270
	Valid Cases	196	212	257	260
	Outlier	66	59	44	41
	Lowest Input	7	7	4	5
	Efficient	41	41	51	55
	Not Efficient	82	5	159	159
*=Efficient , NA=zero input / output, OH=Outlier with high R&D, OL=Outlier with low R&D **=Efficient but minimum input					

Table 8: Final results

References

Almeida, P. and Kogut, B. (1999), 'Localization of Knowledge and the Mobility of Engineers in Regional Networks', *Management Science* **45**(7).

Andersen, P. and Petersen, N. C. (1993), 'A Procedure for Ranking Efficient Units in Data Envelopment Analysis', *Management Science* **39**(10), 1261–1264.

Arndt, O. (2000), *Innovative Netzwerke als Determinante betrieblicher Innovationstätigkeit. Das Beispiel Süd-Wales/UK*, Selbstverlag im Wirtschafts- und Sozialgeographischen Institut der Universität zu Köln, Köln.

Arundel, A. and Kabla, I. (1998), 'What Percentage of Innovations are Patented? Empirical Estimates for European Firms', *Research Policy* **27**(2), 127–141.

Asheim, B. T. and Isaksen, A. (2002), 'Regional Innovation Systems: The Integration of Local 'Sticky' and Global 'Ubiquitous' Knowledge', *Journal of Technology Transfer* **27**, 77–86.

Audretsch, D. (1998), 'Agglomeration and the Location of Innovative Activity', *Oxford Review of Economic Policy* **14**(2).

Backhaus, A. and Seidel, O. (1998), 'Die Bedeutung der Region fur den Innovationsprozess', *Raumforschung und Raumordnung* **56**, 264–274.

Bade, F.-J. (1987), *Regionale Beschäftigungsentwicklung und produktionsorientierte Dientsleistungen*, Sonderhelft 143. Deutsches Institut für Wirtschaftsforschung, Berlin.

Banker, R. D., Charnes, A. and Cooper, W. W. (1984), 'Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis', *Management Science* **30**(9), 1078–1092.

Barnett, V. and Lewis, L. (1995), *Outliers in Statistical Data*, Chichester, Wiley.

Beise, M. and Stahl, H. (1999), 'Public Research and Industrial Innovations in Germany', *Research Policy* **28**(4), 397–422.

Bode, E. (2004), 'The Spatial Pattern of Localized R&D Spillovers: An Empirical Investigation for Germany', *Journal of Economic Geography* **4**(1).

Bottazzi, L. and Peri, G. (2003), 'Innovation and Spillovers in Regions: Evidence from European Patent Data', *European Economic Review* **47**(4), 687–710.

Broekel, T. and Binder, M. (2006), 'The Regional Dimension of Knowledge Transfers - A Behavioral Approach', *Papers on Economics & Evolution* .

Broekel, T. and Brenner, T. (2005), 'Local Factors and Innovativeness - An Empirical Analysis of German Patents for Five Industries', *Papers on Economics & Evolution* **509**.

Broekel, T. and Brenner, T. (2006), 'Exploring the Innovation Performance of German Regions: An Application of a Data Envelopment Analysis on the Instruments Industry', *Paper presented at the "Firm-Level Longitudinal Data on Economic Performance and their Determinants"- Workshop, Volterra, I, 5.-8. June* .

Camagni, R. (1991), Local "Milieu", Uncertainty and Innovation Networks: Towards a New Dynamic Theory of Economic Space, *in* R. Camagni, ed., 'Innovation Networks: Spatial Perspectives', Belhaven Stress. London, UK and New York, USA.

Cazals, C., Florens, J.-P. and Simar, L. (2002), 'Nonparametric Frontier Estimation: A Robust Approach', *Journal of Econometrics* **106**(1), 1–25.

Charnes, A., Cooper, W. W. and Rhodes, E. (1978), 'Measuring the Efficiency of Decision Making Units', *European Journal of Operational Research* **2**(6), 429–444.

Cooke, P. (1992), 'Regional Innovation Sytems: Competitive Regulation in the New Europe', *GeoForum* **23**, 356–382.

Cooke, P., Uranga, M. G. and Etxebarria, G. (1997), 'Regional Innovation Systems: Institutional and Organisational Dimensions', *Research Policy* **26**(4-5), 475–491.

Daraio, C. and Simar, L. (2005), 'Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach', *Journal of Productivity Analysis* **24**, 93–121.

Daraio, C. and Simar, L. (2006), 'A Robust Nonparametric Approach to Evaluate and Explain the Performance of Mutual Funds', *European Journal of Operational Research* **175**, 516–542.

DTI (2003), 'Competing in the Global Economoy - The Innovation Challenge', *DTI Economics Paper* **7**.

Farrell, M. J. (1957), 'The Measurement of Productive Efficiency', *Journal of the Royal Statistical Society* **120**(A), 253–290.

Feldman, M. (1994), *The Geography of Innovation*, Economics of Science, Technology and Innovation, Vol. 2, Kluwer Academic Publishers, Dordrecht.

Feldman, M. P. and Florida, R. (1994), 'The Geographic Sources of Innovation: Technological Infrastructure and Product Innovation in the United States', *Annals of the Association of American Geographers* **84**(2), 210–229.

Fritsch, M. (2006), 'Measuring the Efficiency of Regional Innovation Systems: An Empirical Assessment', *Freiberger Arbeitspapiere* **6**.

Galagedera, D. and Silvapulle, P. (2003), 'Experimental Evidence on Robustness of Data Envelopment Analysis', *Journal of the Operational Research Society* **54**, 654–660.

Greif, S. and Schmiedl, D. (2002), *Patentatlas 2002 Dynamik und Strukturen der Erfindungstätigkeit*, Deusches Patent- und Markenamt, München.

Griliches, Z. (1979), 'Issues in Assessing the Contribution of R&D to Productivity Growth', *Bell Journal of Economics* **10**, 92–116.

Hollingsworth, B. and Smith, P. (2003), 'Use of Ratios in Data Envelopment Analysis', *Applied Economics Letters* **10**, 733–735.

INKAR (2002), *Aktuelle Daten zur Entwicklung der Städte, Kreise und Gemeinden, CD-Rom, Berichte*, Vol. 14, Bundesministerium fur Bauwesen und Raumplanung, Berlin.

Jaffe, A. (1989), 'Real Effects of Academic Research', *American Economic Review* **79**(5), 957–970.

Kline, S. and Rosenberg, N. (1986), An Overview on Innovation, *in* R. Landau and N. Rosenberg, eds, 'The Positive Sum Strateg', National Academy Press. Washington, D.C., USA, pp. 275–305.

Kohlbecher, S. (1997), *Forderung betrieblicher Innovationsprozesse - Eine empirische Erfolgsanalyse*, Deutscher Universitats Verlag, Wiesbaden.

Kostiainen, J. (2002), 'Urban Economic Development Policy in the Network Society', *Tekniikan akateemisten liitto, Tampere, Finland* .

Kuhner, M. (2005), *Ein Verfahren zur Analyse prozessualer Logistikleistung auf Basis der Data Envelopment Analysis*, IPA-IAO - Forschung und Praxis, Vol. 416, Jost Jetter Verlag, Heimsheim.

Legler, H., Licht, G. and Egeln, J. (2001), *Zur Technologischen Leistungsfahigkeit Deutschlands - Zusammenfassender Endbericht 2000*, Gutachten im Auftrag des Bundesministeriums fur Bildung und Forschung.

Marshall, A. (1919), *Industry and Trade*, MacMillan, London.

Mohr, H. (2002), Räumliche Mobilitat von Hochschulabsolventen, *in* 'Arbeitsmarkte fur Hochqualifizierte', L. Bellmann & J. Velling (Hrsg.), Nurnberg, pp. 249–281.

Mouchart, M. and Simar, L. (2002), 'Efficiency Analysis of Air Controllers: First Insights', *Consulting Report, Institute de Statistique, Universitè Catholique de Louvain, Belgium* **0202**.

Nicolay, R. and Wimmers, S. (2000), *Kundenzufriedenheit der Unternehmen mit Forschungseinrichtungen*, Deutscher Industrie- und Handelstag (DIHT), Berlin, Bonn.

Pedraja-Chaparro, F., Salinas-Jimenez, J. and Smith, P. (1999), 'On the Quality of the Data Envelopment Analysis Model', *The Journal of the Operational Research Society* **50**(6), 636–644.

Roudaut, N. and Vanhems, A. (2004), 'Measure of Firms Inefficiency and Heterogeneity using Nonparametric Frontier Models: The Cote d'Ivoire Case', *Paper presented at the "E.A.R.I.E. 2004, Berlin September 2.-5.* .

Ruggiero, J. (2005), 'Impact Assessment of Input Omission on DEA', *International Journal of Information Technology & Decision Making* **4**(3), 359–368.

Saen, R. F., Memariani, A. and Lotfi, F. H. (2005), 'The Effect of Correlation Coefficient Among Multiple Input Vectors on the Efficiency Mean in Data Envelopment Analysis', *Applied Mathematics and Computation* **162**(2), 503–521.

Saxenian, A. (1998), *Regional Advantage - Cluture and Competition in Silicon Valley and Route 128*, Cambridge, Mass.

Scheel, H. (2000*a*), *Effizienzmaße der Data-Envelopment-Analysis*, Deutscher Universitats-Verlag, Schriftenreihe: Gabler Edition Wissenschaft, Wiesbaden.

Scheel, H. (2000*b*), 'EMS: Efficiency Measurement System User's Manual', *http://www.wiso.uni-dortmund.de/lsfg/or/scheel/ems/ems.pdf* .

Schibany, A. and Schartinger, D. (2001*a*), 'Interactions Between Universities and Enterprises in Austria: An Empirical analysis at the Micro- and Sector Level', *Innovative Networks, Co-Operation in National Innovation Systems, OECD* pp. 235–252.

Schibany, A. and Schartinger, D. (2001*b*), *Interaktions Between Universities and Enterprises in Austria: An Empirical Analysis at the Micro and Sector Levels*, OECD, Paris, pp. S. 235 – 252.

Schmoch, U., Laville, F., Patel, P. and Frietsch, R. (2003), 'Linking Technology Areas to Industrial Sectors', *Final Report to the European Commission, DG Research, Karlsruhe, Paris, Brighton* .

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

Simar, L. (2003), 'Detecting Outliers in Frontier Models: A Simple Approach', *Journal of Productivity Analysis* **20**, 391–424.

Simar, L. and Wilson, P. W. (1998), 'Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models', *Management Science* **44**(1), 49–61.

Simar, L. and Wilson, P. W. (2000*a*), 'A General Methodology for Bootstrapping in Nonparametric Frontier Models', *Journal of Applied Statistics* **27**(6), 779–802.

Simar, L. and Wilson, P. W. (2000*b*), 'Statistical Inference in Nonparapetric Frontier Models: The State of the Art', *Journal of Productivity Analysis* **13**, 49–78.

Simar, L. and Wilson, P. W. (2000*c*), 'Testing Restrictions in Nonparametric Frontier models', *Communications in Statistics: Simulation and Computation* **30**, 161–186.

Simar, L. and Wilson, P. W. (2006), Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives, *in* C. L. H. Fried and S. Schmidt, eds, 'The Measurement of Productive Efficiency', Oxford University Press, chapter 4.

Simar, L. and Wilson, P. W. (2007), 'Estimation and Inference in Two-Stage Semi-Parametric Models of Production Processes', *Journal of Econometrics* **136**, 31–64.

Smith, P. (1997), 'Model Misspecification in Data Envelopment Analysis', *Annals of Operations Research* **73**, 233–252.

Walden, J. B. (2006), 'Estimating Vessel Efficiency using a Bootstrapped Data Envelopment Analysis Model', *Paper presented at the "NOAA Fisheries Economics & Social Science Workshop, San Francisco, CA, 18.-20. April 2006* .

Weibert, W. (1999), *Regionale Determinanten der Innovation*, Nomos Verlagsgesellschaft, Baden-Baden.

Wheelock, D. C. and Wilson, P. W. (2003), 'Robust Nonparametric Estimation of Efficiency and Technical Change in U.S. Commercial Banking', *The Federal Reserve Bank of St. Louis Working Paper Series* **37A**, 1–34.

Wilson, P. W. (1993), 'Detecting Outliers in Deterministic Nonparametric Frontier Models with Multiple Outputs', *Journal of Business & Economic Statistics* **11**(3), 319–323.

Wilson, P. W. (1995), 'Detecting Influential Observations in Data Envelopment Analysis', *The Journal of Productivity Analysis* **6**, 27–45.

Wilson, P. W. (2006), 'FEAR 1.0: A Software Package for Frontier Efficiency Analysis with R', *http://www.clemson.edu/economics/faculty/wilson/Software/FEAR/fear.html* .

Witt, U. and Zellner, C. (2005), 'Knowledge-based Entrepreneurship: The Organizational Side of Technology Commercialization', *Papers on Economics & Evolution* **504**.