Broekel, Tom

Working Paper

A concordance between industries and technologies: matching the technological fields of the Patentatlas to the German industry classification

Jena Economic Research Papers, No. 2007,041

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Broekel, Tom (2007) : A concordance between industries and technologies: matching the technological fields of the Patentatlas to the German industry classification, Jena Economic Research Papers, No. 2007,041, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/25608

JENA ECONOMIC RESEARCH PAPERS

2007 – 041

A Concordance between Industries and Technologies
Matching the technological fields of the Patentatlas to the German Industry Classification

by

Tom Broekel

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich-Schiller-University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact m.pasche@wiwi.uni-jena.de.

Impressum:

Friedrich-Schiller-University Jena
Carl-Zeiß-Str. 3
D-07743 Jena
www.uni-jena.de

Max-Planck-Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2007-041

A concordance between industries and technologies

Matching the technological fields of the *Patentatlas* to the German Industry Classification [‡]

Tom Broekel [*]

Max Planck Institute of Economics, Jena, Germany

20th July 2007

Abstract

The *Patentatlas* by Greif and Schmiedl (2002) represents an important source for patent data in Germany. Its use for industry-specific studies is however problematic because the correct assignment of patent data classified by technological fields to commonly used industry classifications is unclear.
This paper presents an application-oriented approach to this issue. In using industry-specific R&D employment numbers on a regional level, an approximate concordance is developed between the 31 technological fields of the *Patentatlas* and 21 manufacturing industries, as defined by the German Industry Classification.

JEL codes: O18, O34, R12
Keywords: Patentatlas, Industry Classification, IPC, NACE, Concordance

[‡] The author would like to thank Thomas Brenner for helpful comments and suggestions.
[*] Max Planck Institute of Economics, Evolutionary Economics Group, Kahlaische Strasse 10, D-07745 Jena, Germany. *Phone*: +49 3641 686811. *Fax*: +49 3641 686868. *E-mail*: broekel@econ.mpg.de

1 Introduction

The use of patent data has a long tradition in innovation research. Most commonly, patents or patent applications, serve as proxies for firms' innovative activities (see, e.g., Jaffe, 1989; Feldman, 1994). In addition, patents and their citation structure serve as a widely used data base for research on the transfer and diffusion of knowledge (see, e.g., Jaffe and Trajtenberg, 1996; Breschi and Lissoni, 2001).

It is a fact that patent data is often and extensively used. We refrain from discussing all (but one) advantages and disadvantages of using patent data in innovation research. In contrast to other indicators, one of the major advantages of patent data is that it is available: "Unfortunately, in most instances the choice is not between patent statistics and better data, but between patent statistics and no data" (Schmookler, 9966, p. 198). It is true that in most countries patent statistics are easily accessible and provide a reasonably coherent and reliable data base for innovation research.

In Germany, an often used source for patent data is the *Patentatlas*, provided by the German patent office ('Deutsches *Patent- und Markenamt*') (Greif and Schmiedl, 2002). It contains preprocessed patent application data: the numbers of patent applications for different regional levels in Germany. Furthermore, the patent applications are organized in 31 different technological fields which are based on the International Patent Classification (IPC). These IPC classes are aggregated to the 31 technological fields that have been put forward by the World Intellectual Property Organization (WIPO) (see Greif and Schmiedl, 2002, p. 18).

Most importantly, this is true for the German Industry Classification ('Klassifikation der Wirtschaftszweige') established by the German Statistical Office ('Statistisches Bundesamt') (DESTATIS, 2002). This classification is used for example, for the industry-specific organization of employee numbers.

For many applications industry-specific analyses are necessary or valuable. In these cases, it is of uttermost interest to assign patent applications only to those firms or industries that are truly responsible for them. This is especially true for regional analyses, if a researcher has to rely on data of the *Patentatlas* and wants to relate it to industry-specific data that is organized, for example by the German Industry Classification. In this case, he needs to assign the appropriate patent applications (technological fields) to the correct industries (industry classes). Thus, a concordance is needed between the technological fields and the industries. To the author's knowledge, no such concordance exists between the *Patentatlas*' technological fields and the German Industry Classification.

The paper develops such a concordance on the base of the spatial co-locatedness of R&D employees and patent applications. The employed matching procedure relies on the concordance by Schmoch et al. (2003) between three-digit IPC subclasses and the German Industry Classification. It is aggregated to correspond to the 31 technological fields of the *Patentatlas* and adapted to the matching between these technological fields (TF) and the two-digit indus-

tries of the German Industry Classification (GIC). Information on the extent to which R&D employees of an industry and patent applications in a technological field are co-located across German labor market regions are used in this task. However, the approach represents only a second-best one. A concordance based on firm-level data would seem to be the far better alternative. Such data is however not available to the author. Therefore, the concordance presented in the following should be seen as an "approximate" attempt.
The paper is structured as follows: Section 2 briefly describes the technology-oriented classifications by which the patent applications are organized in the *Patentatlas*. Furthermore, the German Industry Classification is introduced. Section 3 introduces the method used to set up the concordance. This is followed by the presentation of the results in Section 3.3. Section 4 concludes.

2 The *Patentatlas* and the German Industry Classification

2.1 The German 'Patentatlas'

There are different sources of patent data that are available for researchers in Germany. Most prominently, the German Patent Office offers access to their data. Patent data is also available at the European Patent Office. Both institutions provide mainly 'raw data,' i.e., the complete patent forms with all the information regarding inventor, applicant, technical details, etc. These can be acquired for complete sets of countries and years. Thus, it is up to the researcher to extract the relevant information from these data bases. However, he is often just interested in the total number of patents for a region or a specific industry. This type of data has to be generated with sophisticated queries.
The *Patentatlas* by the German Patent Office (Greif and Schmiedl, 2002) represents an alternative source for patent application data. Its main advantage is that it contains preprocessed data. The patent applications are collected from the European Patent Office as well as from the German Patent Office (without double counts). All applications by German but also by foreign applicants are considered if the inventor's place of residence is located in Germany. This assignment to regions according to the inventors' residence is one of the great services provided by the *Patentatlas*. Patent applications with more than one inventor are assigned on an equal basis to each inventor and thus to their region and institution.
On the base of this 'inventor principle', the patent applications are aggregated to different regional levels: districts ('Kreise'), labor market regions ('Arbeitsmarktregionen'), planning regions ('Raumordnungsregionen'), and federal states ('Bundesländer').
The applications are also published separately for private persons ('natürliche Personen'), public research institutes ('Wissenschaft'), and companies ('Wirtschaft').

Furthermore, Greif and Schmiedl (2002) organize the data into 31 different technological fields that have been put forward by the World Intellectual Property Organization (WIPO) (see Greif and Schmiedl, 2002, p. 18). This allows to gain an overview ('Gesamtübersichten') of the differences in the patent applications that stem from technological differences (Greif and Schmiedl, 2002, p. 18). These technological fields are based on two- to three-digit IPC subclasses. Each of these is precisely assigned to one technological field. The share of each class of the technological area is also published. An overview of the technological fields has be extracted from the *Patentatlas* and can be found in the Appendix in Table 2.

The organization of the patent applications according to regions, technologies, and type of institutions, makes the Patentatlas a valuable data source for regional researchers, saving them much time and guaranteeing high-quality data.

However, researchers (especially regional researchers) do not only rely on patent data. They also describe regions in terms of industrial structure, R&D capacities, etc. This data is provided by different sources than the patent data and is organized differently. In the following subsection, we present one major classification of industry-specific data that is commonly employed in studies that also use patent data.

2.2 Industry-specific data

In many studies, researchers rely on more than one source of data. For example, for calculating regional performance measures it is of uttermost interest to assign the correct inputs (e.g. R&D employees) to the innovative output (e.g. patent applications as in Broekel and Brenner (2007a,b)). The investigations are carried out separately for different industries, requiring that the input factors are industry specific and the output is assigned correctly to each industry.

In Germany, industry-specific data such as, e.g., employment numbers, are classified according to the German Classification of Economic Activities ('Klassifikation der Wirtschaftszweige') by the Federal Statistical Office ('Statistisches Bundesamt') (for more details, see, DESTATIS, 2002). It is based on the European Classification of Economic Activities ('Nomenclature statistique des Activités économiques dans la Communauté Européenne').[1] In its hierarchical order, 1041 subgroups exist, which can be aggregated to 513 classes, 222 groups, 60 divisions, 31 subsections, and 17 sections. This corresponds to 'five-digit levels.' For example, the subsection level corresponds to the 'single-digit level' and divisions to the 'two-digit level.' In the case of Germany, this classification is denoted as *German Industry Classification* (GIC) in the following.

Most of the industry-specific data, employment numbers, firm size, etc., is organized according to this classification. In many instances, it is important to match such data to the patent data provided by the *Patentatlas*. As for the data used to establish a concordance, we can disaggregate the industry-specific data employed down to the 60 divisions (two-digit level).

[1] In short: NACE Rev. 1.1.

However, because we only consider the manufacturing sector, for which patent activities are reported in (Schmoch et al., 2003), only 21 different industries are relevant. They are listed in Table 3.

As has been pointed out, the *Patentatlas* as well as the GIC represent important and frequently used data sources. In industry-specific contexts, the 31 technological fields of the *Patentatlas* are often argued to refer to specific industries (see, e.g., Broekel and Brenner, 2005). However, the relation between the technologically organized data to data organized in an industrial context is not straightforward. The nomenclature of the technological fields and industries offer little knowledge about the 'true' scope of the activities covered. Thus, it is not clear to which industry the technological field 2 'Food and tobacco' (TF2) should be assigned. In this case, one might expect that the patent applications in this field are mainly driven by innovations in food itself. However, our results suggest that most of the patent applications concern the food processing machines, so that these patent applications should be assigned to the machine building sector.

In order to avoid such misassignments, we argue in favor of establishing a concordance between these two different classifications. The procedure of how such a concordance can be established is described in the following section.

3 Method

3.1 General approach

The *Patentatlas* provides high-quality patent application data on a regional level for different types of institutions. What is missing however, is information about how this patent data can be matched to other data, as organized for example, by the GIC. In principle, setting up such a concordance is simple. If the number of patents each industry contributes to a technological field is known, one can easily estimate the relative importance of a technological field for an industry. Using this relative importance, the most important technological fields can be identified for an industry and vice versa.

However, in our case the number of patents an industry applies for in a specific technological field is also unknown. This makes the procedure more complex.

Thus, the importance of an industry for a certain technological field is unknown. Furthermore, to the author's knowledge the technological fields do not correspond to any other patent classification for which a concordance with the GIC exists. The patent data provided by the *Patentatlas* has also not been matched to a GIC before.

In this paper, we make use of a concept that has been presented in Broekel and Brenner (2007a,b): The correlation between industrial R&D employees and patent applications in space is used as an approximation for an industry's importance for the patent applications in a technological field (TF). Greif and Schmiedl (2002) show that the relation between to-

tal patent applications and the total number of R&D employees in a region proves to be very strong on the level of planning regions ('Raumordnungsregionen'). However, they do not disaggregate the R&D employees and patents into different industries and technologies, respectively. At the core of the matching procedure, we assign the strongest importance of a technological field to the industry whose R&D employees are correlated most strongly to its patent applications. Of course, we only consider industries of which we know that they apply for patents in a specific technological field. Information about this is found in Schmoch et al. (2003). The exact procedure is described in the next subsection.

The data on R&D employment is obtained from the German labor market statistics. The R&D personnel is defined as the sum of the occupational groups of agrarian engineers (032), engineers (60), physicists, chemists, mathematicians (61), and other natural scientists (883) (Bade, 1987, p. 194ff.). Further, the R&D employees are assigned to the industries they work in. Since these employees are likely to represent industrial R&D capacities, we use only patent applications by companies, i.e., only the category 'Wirtschaft' in the *Patentatlas*. By excluding the patents of private persons ('natürliche Personen'), we ensure that the patents of private inventors do not bias the results, because they are not likely to show up in the used employment statistics. However, in doing so, we also tend to exclude patent applications by smaller firms. At this point, this trade-off cannot be conclusively resolved. Further, we do not consider the patent applications by public science institutions (category 'Wissenschaft') as they would bias the results because their employment data is not included in the R&D employee data.

Besides choosing the appropriate patent applications, a choice has to be made regarding the used level of spatial units. We decided on the use of labor market regions ('Arbeitsmarktregionen') as an appropriate level. This is justified by a number of reasons. The district level is too small when using patent data that is assigned according to the place of residence of the inventor. The reason for this is that in too many cases an inventor's region of residence is different than the region of his workplace (Greif and Schmiedl, 2002, p. 10). In this case, the patent would be assigned to the wrong region (region of residence) since the invention originates at the inventor's workplace.

In contrast to districts, planning regions (PR) and labor market regions (LMR) are useful levels for such investigations (Greif and Schmiedl, 2002). Compared to planning regions, LMR offer three additional advantages. First, they are smaller in size than PR. This results in a larger number of LMR than PR: 270 to 97. This improves the reliance of the statistical analysis and is more likely to reveal the true relationship between R&D employees and patent applications. Second, LMR are set up even more functionally oriented than PR by taking only daily commuting behavior into account. Third, labor market regions are under constant revision (see Eckey et al., 2007). Thus, they reflect up-to-date research on the structure of economic activities in space.

The latter point is of importance also for another reason. The *Patentatlas* reports the patents for a system of 225 labor market regions. In contrast, we use the more recent delineation with 270 labor market regions as used by the Joint Task "Improvement of the Regional Economic Structure" (see Eckey et al., 2007). Therefore, we aggregated the district data provided by the *Patentatlas* to this delineation.

In using these data, we are able to develop an "approximate" concordance between the technological fields of the *Patentatlas* and the German Industry Classification. The quality of the approximation depends on the extent to which the correlation in space reflects the true relationships between R&D employees and the number of patent applications. Thus, the weakness of such an approach lies in the fact that many different industries are strongly geographically co-located. This co-location is likely to bias our results. There is a good chance that we assign high importance to an industry (which may in truth be of little importance) for a TF because it is strongly co-located to an industry that is truly important for this TF. With respect to the data available, we have no possibility to rule out such aspects. However, this is likely to be a problem when the observed correlations are rather low. In these cases, the results have to be interpreted with care.

Following the presentation of the data used, we now turn to the description of the procedure of matching the technological fields to the industry classification.

3.2 The development of an 'approximate' concordance

3.2.1 First step: a two-digit IPC subclasses to two-digit GIC industries matrix

Patent applications and R&D employees are strongly related. Without industrial R&D employees no (industrial) patents are applied for. Both are, however, organized differently. In the previous sections, we presented the *Patentatlas* and its 31 technological fields by which the patent applications are organized. In contrast, the R&D employees are classified according to the German Industry Classification (GIC) with its 21 different industries in the manufacturing sector.

Based on the assumption that the place of residence of the inventor, which is documented in the patent application, is located in the same labor market region as the R&D employee's workplace, we use this spatial co-location to match the technological fields to the different industries.

Figure 1 illustrates the procedure for the relation between TF19, TF9 and industries DF21, DK29. In a first step, we extract the relations between three-digit IPC subclasses and three-digit GIC classes from Annex 2 of Schmoch et al. (2003). They relate the 625 three-digit IPC subclasses to 44 sectoral fields they defined ex ante. For the same 44 sectoral fields they provide a definite concordance with the 60 three-digit GIC industries. Thus, it is easy to match

the IPC subclasses to the GIC industries, using the 44 sectoral fields as links.[2] The matching is represented in the middle of Figure 1 by the arrows emerging from the 2-digit GIC classes to the three-digit IPC classes. For example, in the case of industry DF21.0 the arrows point to all IPC subclasses its firms patent into.

Before this information can be used for our purpose, we have to make these concordances concerning relationships correspond to our data. This is achieved by translating the concordance between IPC subclasses and GIC industries to a higher level of aggregation.

The 31 technological fields of the *Patentatlas* are developed on the base of two-digit IPC subclasses. Furthermore, the data on regional R&D employees are available on the two-digit GIC industry level. Hence, the three-digit IPC subclasses to three-digit GIC industries matrix needs to be transformed into a two-digit IPC subclasses to two-digit GIC industries matrix. This is easy on the industry side. We just have to translate the relations between three-digit GIC industries to the higher level of two-digit GIC industries. In Figure 1 this is represented by the arrows departing the three-digit GIC industries to the right, connecting them to the two-digit GIC industries (DF21.0 $\rightarrow$ DF21). Note that we only develop this matrix for the 21 manufacturing industries and not for all 31 two-digit GIC industries.

On the side of the patent application data all three-digit IPC subclasses are easily aggregated to the two-digit level. Hence, the original three-digit IPC subclasses to three-digit GIC industries matrix is transformed into a two-digit IPC subclasses to two-digit GIC classes matrix. It is represented in Figure 1 by all in between the two bold dashed lines. Since in principle it is just a transformation of the information obtained from Schmoch et al. (2003) to a different aggregation level, it is labeled with Schmoch et al. (2003) in the figure.

3.2.2 Second step: the estimation of maximum weights

However, the patent application data is not organized by two-digit IPC subclasses but according to 31 technological fields. In order to correspond to these technological fields of the *Patentatlas*, the two-digit IPC subclasses to two-digit GIC industries matrix needs to be further aggregated. Although the technological fields of the *Patentatlas* are developed on the base of two-digit IPC subclasses, the *Patentatlas* does not provide direct patent application data on this level. The only information available about the set up of the technological fields is the weight of the two-digit IPC subclasses patent application on a technological field's total number of patent applications. In Figure 1 this is illustrated by the arrows emerging from the two-digit IPC subclasses to the left, connecting them to the technological fields of the *Patentatlas*. In the chosen example the shares of the patent application of each of the IPC subclasses in the TF are given in percentages. In order to aggregate the **IPC-GIC matrix** to make them correspond to the 31 TF of the *Patentatlas*, this information is used below. All two-digit IPC subclasses an industry patents into are known (see above). For each of these IPC

[2] This concordance can be requested from the author.

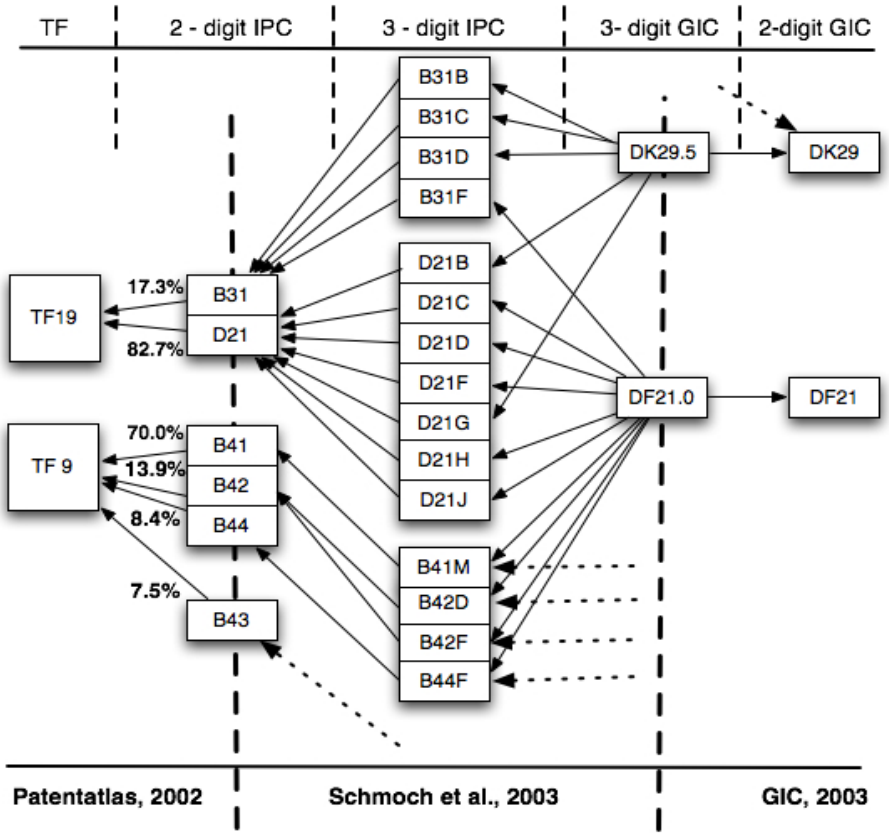

Figure 1: Representation of relationships between *Patentatlas* and GIC

subclasses, the *Patentatlas* provides weight on one particular TF. Thus, in order to estimate the weight of an industry's patent applications in a TF, the weight of all IPC subclasses this industry patents into and which are part of this particular TF, have to be summed up. This is done for all 21 industries, resulting in a 31×21 (TF $\times$ GIC industries) matrix. Its entries represent the (in principle) maximum share of patent applications in a TF that an industry can be responsible for. Note, however, that it does not represent the share of an industry's patent applications compared to a TF's total patent applications, but the maximum value this share can reach. Again, Figure 1 may be used to illustrates this procedure in the case of industry DF21.

Starting from the right side of the Figure, it shows that the two-digit GIC industry DF21 is not disaggregated on the three-digit GIC level. Hence, there is only one three-digit industry (DF21.0) that needs to be taken into consideration. Schmoch et al. (2003) give us all three-digit IPC subclasses this industry's firms patent into. In the figure this is represented by the arrows departing from this industry to the left. For example, DF21.0 patents in to IPC B41M, B42D, B42F, and B44F. Their superior two-digit IPC subclasses are B41, B42 and B44. The *Patentatlas* provides us with the information that B41 accounts for about 70.0 % of the patent applications categorized into TF9. In the case of B42, about 13.9 % of TF9's patent applications fall into this two-digit IPC-subclass. Finally, B44 accounts for additional 8.4 % of the

patent applications of TF9. Since there is no other connection between the industry DF21.0 and TF9 (see Figure 1), this tells us that at maximum DF21.0 accounts for 92.3 % (70 %+13.9 %+8.4 %) of all patent applications in TF9. This is the case if all patent applications filed in the three-digit IPC subclasses B41M, B42D, B42F, B44F stem from DF21.0. However, all we know is that firms from DF21.0 apply for patents into this subclasses, but not to what extent.[3] However, this procedure gives us the maximum share of applications in TF9 for which it can theoretically account for. Note that 7.5 % of the patent applications in TF9 are classified into the two-digit IPC subclass B43 which no firm of DF21.0 patents into (indicated by the dashed arrow pointing towards B43 in Figure 1). In a similar manner the maximal shares of patent applications an industry is responsible for in a TF are estimated for all other industries.
This resulting matrix is denoted as *max-weights matrix* in the following. It shows positive values in the case of an industry patenting into a TF and zero values if there is no relation between an industry and a TF.

3.2.3 Third step: the spatial relation between R&D and patent applications

At this point, the spatial relation between R&D employees and patent applications come into play. The idea is simple: the higher the correlation in space between an industry's R&D employees and a TF's patent applications, the more likely this industry's R&D employees are an important source of patent applications assigned to this TF. Hence, one could estimate the correlation in space between each industry's R&D employees and the patent application data in the technological fields the particular industry patents into.
However, in most cases more than one industry usually patents into a TF. This implies that a bivariate correlation model is inadequate and a model that accounts for multiple industries - TF relations is required.
In this paper, we apply a standard multiple OLS regression. In this a TF's patent applications in a region serve as dependent variable. As independent variables, the R&D employees of those industries that patent into a particular TF are used (which show a positive value in the *max-weight matrix with this TF*).
The importance of each independent variable is approximated by the standardized beta values. These are the results of a standardized regression. Before this multiple regression equation is fitted, all variables (dependent and independent) are standardized by subtracting the mean and dividing by the standard deviation. The obtained standardized regression coefficients, thus represent the change in a dependent variable for a change of one standard deviation in an independent variable (Backhaus et al., 2000).
For an easier interpretation of the beta values we normalize them such that they sum up to one across all industries patenting into a particular TF. The resulting matrix, which is again

[3] The lack of this information is the motivation for this rather complex matching procedure in the first place.

of 31×21 dimensions, is denoted as *beta matrix* in the following.[4]

The *beta matrix* is the most important result of this procedure, because the normalized beta values can be interpreted as the weight each industry has on a TF's patent applications. However, it is possible that these weights exceed the maximum values estimated before and represented by the *max-weights matrix*. This is caused by the fact that the co-location approach is only an approximation and hence allows for over- and underestimations of the industries' effects. The obtained regression results can be biased by all kinds of factors as, e.g., performance differences of R&D employees, strong co-location between two industries, etc. This needs to be corrected for in a fourth step.

3.2.4 Fourth step: correcting for overestimation using the maximum weights

In the present context, the overestimation is especially problematic because it implies that the weights obtained on the base of the OLS regression are larger than the theoretical maximum weights. Since these are result of a fairly exact matching between Schmoch et al. (2003) and the *Patentatlas* (see above), they are certainly of superior reliability. Therefore, it seems reasonable to adjust the weights obtained by the regression (normalized beta values) such that they do not conflict with the theoretical maximum weights.

This is achieved by a two-step procedure. In a first step, all OLS weights that exceed the maximum weights are cut back to the value of the maximum weights. In a second step, the excess that has been cut off is summed across all industries for each TF. We know that all patent applications in a TF stem from the industries assigned to it with a positive weight. Hence, an overestimation (excess) in one industry's weight corresponds to an underestimation of another (or more than one) industry's weight that also patents into this TF. Therefore, it is straightforward to raise those industries' weights that patent into this TF, and whose weight does not exceed (or is identical to) the maximum weight, by the same value that has been cut off. In order not to introduce to much disturbance into the OLS regression results, these industries' weights are raised in such a manner that their size proportions are not changed.

For 16 out of 18 TFs in which such a phenomenon is observed this procedure yields sufficient results. In the case of TF13, there is only one industry (DG24) patenting into this field with a maximum share below one. This is (theoretically) impossible, but it is a result of imprecision in the underlying concordance by Schmoch et al. (2003). Lacking other options, we raise the weight of industry DG24 to one.

A similar phenomenon is observable in technological field TF15. Although there is more than one industry patenting into this field, their maximum weights do not add up to one in this TF.

[4] Note that we do not want to answer the question whether there exists a relationship between an industries R&D employees and the patent applications in a TF. This we already know from Schmoch et al. (2003). We only perform the OLS regression in order to get a first-best approximation for the weight an industry has on the patent applications in a TF. This implies that we do not have to worry about the significance of the observed relationships.

As a solution, we raise only that industry's weight (only one has a weight below its maximum weight) that has not reached the maximum weight. Furthermore, we cut off the excess in two industries whose weights exceed the maximum weight. However, in the case we cut it not to the corresponding maximum level, but such that the sum of all weights in this industry is one.[5] Thus, as a result all weights are at their maximum and add up to one.

3.2.5 Result: two concordance matrices

Table 5 (which is denoted as *technology matrix* in the following) shows the normalized – and for over-estimation – corrected beta values. The matrix allows the investigation of each industry's contribution to the patent applications in a TF.

The values of the *technology matrix* can be read as the approximate percentages of patent applications by an industry that are assigned to a technological field. For each technological field their values add up to one. The *technology matrix* represents the relationships between patent applications and industries only from the technology perspective. It is also worthwhile to set up a relationship matrix that implements the industry perspective. In contrast to the *technological matrix*, the *industry matrix* shows the percentages of an industry's patent applications that are assigned to the TFs. The values are easily estimated by using the *technology matrix* and each TF's shares of the total number of patents. The latter are provided by the *Patentatlas* and can be found in Table 2. The resulting *industry matrix* is presented in Table 4. Both, *technology matrix* and *industry matrix*, incorporate the same information. The difference lies only in the normalization. Depending on the purpose (industry or technology perspective), they represent a concordance between the 31 TFs of the *Patentatlas* and the 21 GIC manufacturing industries.

The two matrices provide the information needed to match the technological fields to the industry classes. However, note once more that, for example, strong co-locatedness of industries and differences in innovation performance across regions are likely to bias the results substantially. But as long as firm-level data is not available, this concordance can serve as a usable down-to-earth approach. Unfortunately, the description of the TFs in Greif and Schmiedl (2002) is rather unclear (and skimpy). Therefore, we cannot check the results for their contextual plausibility.

A brief description of how this concordance can be used is presented in the next subsection.

3.3 Application of the concordance

As a result of the matching procedure, two 31×21 matrices are obtained. They link the 31 technological fields of the *Patentatlas* (rows) to the 21 manufacturing industries, as defined by the GIC (columns).

[5] Please note this changes the results only marginally.

In the case of the *technology matrix* , its entries represent the percentages of patent applications that a specific industry contributes to a technological area. Thus, one might use these values to estimate the number of patent applications that an industry accounts for in different TFs. Acting straightforwardly, one could multiply the shares with the observed number of patent applications in order to obtain the industry's patent applications. While this would certainly be the ultimate goal of such a concordance, it would clearly overstretch the capacity for accuracy of this approach. We rather disadvise to use the concordance for such issues.
However, the concordance can be used to find TFs that are most relevant for an industry and vice versa. This is to say that we hope to identify 'clusters' of technologies and industries which show strong within relations, but low outside relations. We speculate that certain technologies and industries belong to the same superior field of R&D activity. This may result from the existence of a certain base technology employed by a number of industries. Or, which is more likely, on the one hand there exists a number of industries that use more or less the same technologies, while on the other hand these technologies are rather seldom used by the other industries. Hence, the aim is to find industries that are somewhat homogeneous in their technological profile. At the same time, we want to identify technologies that are somewhat homogeneous in their industrial profile. These 'clusters' are denoted as 'sectors' in the following.
In order to find such structures we rearrange the concordance matrices manually by changing the order of rows and columns such that the highest values are concentrated in the diagonal.[6] A sector is characterized by high values (strong relations) between industries and technologies that are proximate neighbors and low values (weak relations) between distance industries and technologies in both matrices. Of course, this is a matter of thresholds. We apply a "fifty percent rule" which seems intuitive in this context.

> A technological field is assigned to a sector if at least fifty percent of its patent applications originate from this sector's industries.

Similarly, the threshold in the other direction is defined as follows:

> An industry is assigned to a sector if at least fifty percent of its patents applications are classified to the technological fields of this sector.

Since the aim is to assign as many industries and technologies to homogeneous sectors, the chosen thresholds are comparatively weak: they allow for up to fifty percent mispecifications! Still, we failed to assign one TF to a sector. TF6, "Separation and alloying," seems to be a 'perfect intersection' of the chemistry, machine building, and medical & optical equipment building sectors. Since it represents about 3.7 percent of all patent applications, it cannot be

[6] The use of principal component analysis did not provide acceptable results: the construction of the components was strongly driven by negative relations which do not make sense in this context.

simply ignored. Rather, one is advised to control for the impact of this TF's patent applications when investigating one of these sectors.

Another problematic case is TF26, "Measurement, testing, optics, photography." In the *technological matrix*, its weight in the medical and & optical equipment building sector stays below fifty percent as well. However, the values in the *industry matrix* suggest to assign it to this sector. Nevertheless, when investigating the chemistry sector or the medical & optical equipment sector, sensitivity analyses should be conducted as to whether the results are sensitive to the inclusion and exclusion of these technological fields' patent applications.

With these two exceptions the assignment seems to be fairly robust: in most cases the shares of patent applications stay far above the fifty percent threshold. In fact, there are only some exceptions in which this share drops below seventy percent.

On the base of this procedure, we can identify five sectors that cover all industries and 30 out of 31 technological fields: Chemistry (CHEM), Machine building (MACH), Transport equipment (TRANSP), Electrics & electronics (ELEC) and Medical & optical equipment (MED/OPT). Their composition is listed in Table 1 and depicted as gray shaded areas in the *technology matrix* (Table 5) and the *industry matrix* (Table 4). We advise to conduct sensitivity analyses

Sector	**Technological fields***	**Industries****	**Control** ***
Chemistry	TF5, TF12, TF13, TF14, TF15	DG24, DI26	TF6 ,TF20, DF23
Machine building	TF1, TF2, TF3, TF7, TF8, TF9, TF11, TF17, TF18, TF19, TF20,TF21, TF23, TF24, TF25	DA15, DA16, DB17, DB18, DC19, DC20, DE21, DE22, DH25, DJ27, DJ28, DK29, DN36	TF6, TF22, DM34
Transport equipment	TF10, TF22	DM34, DM35	TF23, TF20
Electrics & electronics	TF27, TF28, TF29, TF30, TF31	DL30, DL31, DL32	DL33
Medical & optical equipment	TF4, TF16, TF26	DL33, DF23	TF6, TF15, DL30
* As defined in Greif and Schmiedl (2002) ** According to the GIC DESTATIS (2002) *** Technological fields of industries which have to be controlled for			

Table 1: Overview technological fields

for the TFs and industries that are listed in the column "control" because they show strong relations to this sector.

With respect to the 31 TFs and 21 industries, the final number of only five sectors is somewhat disappointing. In the case of **Machine building** a further breakdown would have been desirable. Only in the case of TF6, DF23, and to some extent DI26, the definitions are rather vague. Hence, this concordance seems to be rather robust with respect to the definition of these five sectors.

Industry DF23, "Manufacture of coke, refined petroleum products and nuclear fuel," shows an almost equal importance for the **Chemistry sector** and the **Medical and optical equipment sector**. Therefore, we cannot assign it to either. The name of this industry, "Manufacture of coke, refined petroleum products and nuclear fuel," suggests to assign it to the **Chemistry sector**. Since we do not know to which extent the names represent the actual industrial activities, we also suggest to control for this industry if one of the two sectors are investigated.

In summarizing, we find five sectors that can be defined fairly clearly in terms of technological fields and industries. The 'noise' in these definitions is rather small which indicates that they are robust if one controls for a limited number of ambiguous TFs and industries.

4 Conclusion

The *Patentatlas* by Greif and Schmiedl (2002) represents an important source for patent data in Germany. It does not only provide disaggregated patent application data on different regional levels, but also divides the applications into 31 technological fields. While it seems straightforward to use these technological fields in industry-specific studies, the correct assignment of technological fields to industries is problematic. The reason is that there is no concordance that matches these technological fields to a commonly used industry classification, as, e.g., the NACE or the German Industry Classification ('Wirtschaftszweigklassifikation'). Thus, the usability of the *Patentatlas* is seriously reduced for studies that use industry-specific data defined by one of these industry classifications.

This paper presents an application-oriented approach to this issue. In using industry-specific R&D employment numbers on a regional level, a concordance between the 31 technological fields and 21 manufacturing industries of the German Industry Classification is developed. At its core, the matching procedure translates the basic concordance by Schmoch et al. (2003) (which is aggregated to the 31 technological fields of the *Patentatlas*) with the weights an industry has on the technological field's patent applications. The estimated weights reflect the extent to which R&D employees of an industry and patent applications in a technological field are co-located across German labor market regions.

Five sectors are identified that can be comparatively clearly defined in terms of technological fields and industries. These sectors represent groups of industries that are quite homogeneous with respect to their technological profile. They are the "Chemistry sector," "Machine building sector," "Transport equipment sector," "Electrics and electronics sector, " and "Medical and optical equipment sector."

For only two of the 21 industries this assignment is rather weak. In the case of one technological field, an assignment seems to be even impossible with the data and method employed here. Also, a further breakdown of the machine building sector would have been desirable.

The use of firm-level data would certainly increase the reliability and breakdown of the concordance matrix. In any case, this attempt is to be seen as an invitation to motivate more work on this issue. Nevertheless, in the light of the availability of appropriate data, our approach yields robust results and provides a helpful tool for researchers using these data sources.

Appendix

Code	**Share*** patents	**Technological fields****
TF1	1.2%	Agriculture
TF2	0.8%	Food and tobacco
TF3	3.1%	Convenience goods, household articles
TF4	4.5%	Healthcare (without P5), entertainment
TF5	1.9%	Medical, dental, cosmetic products
TF6	3.7%	Separation and alloying
TF7	3.1%	Metalworking, foundry, machine tools
TF8	3.7%	Grinding, companding, tools
TF9	1.4%	Printing
TF10	10.3%	Vehicles, ships, airplanes
TF11	4.5%	Conveying, lifting, saddlery
TF12	1.6%	Inorganic chemistry
TF13	3.1%	Organic chemistry
TF14	2.1%	Organic macromolecular compounds
TF15	1.8%	Dyestuffs, petroleum industry, oils, fats
TF16	1.2%	Fermentation, sugar, molting
TF17	1.3%	Metallurgy
TF18	1.5%	Textiles, ductile materials
TF19	0.5%	Paper
TF20	5.7%	Construction industry
TF21	0.3%	Mining
TF22	5.1%	Power and work machines
TF23	6.4%	Machine building
TF24	3.3%	Illumination, heating
TF25	0.6%	Weapons, explosives
TF26	7.2%	Measurement, testing, optics, photography
TF27	4.3%	Horology, controlling, checking, computing
TF28	1.6%	Instruction, acoustics, information processing
TF29	0.2%	Nuclear physics
TF30	9.0%	Electrotechnics
TF31	5.0%	Electronics, communication industry

*Figures for 2000 from Greif and Schmiedl (2002).
**Translation by author based on German nomenclature by Greif and Schmiedl (2002)

Table 2: Overview technological fields

Code	Industry
DA15	Manufacturing of food products, beverages, and tobacco
DA16	Manufacturing of tobacco products
DB17	Manufacturing of textiles
DB18	Manufacturing of wearing apparel, dressing and dyeing of fur
DC19	Tanning and dressing of leather, manufacture of luggage, handbags, saddlery, harness and footwear
DC20	Manufacture of wood and products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
DE21	Manufacture of pulp, paper, and paper products; publishing and printing
DE22	Publishing, printing, and reproduction of recorded media
DF23	Manufacture of coke, refined petroleum products, and nuclear fuel
DG24	Manufacture of chemicals and chemical products
DH25	Manufacture of rubber and plastic products
DI26	Manufacture of other nonmetallic mineral products
DJ27	Manufacture of basic metals
DJ28	Manufacture of fabricated metal products, except machinery and equipment
DK29	Manufacture of machinery and equipment n.e.c.
DL30	Manufacture of office machinery and computers
DL31	Manufacture of electrical machinery and apparatus n.e.c.
DL32	Manufacture of radio, television, and communication equipment and apparatus
DL33	Manufacture of medical, precision, and optical instruments, watches and clocks
DM34	Manufacture of motor vehicles, trailers, and semitrailers
DM35	Manufacture of other transport equipment
DN36	Manufacturing of furniture, manufacturing n.e.c.

Table 3: Overview industries

	DG24	DI26	DK29	DH25	DA15	DJ28	DJ27	DA16	DC20	DE21	DB17	DB18	DC19	DN36	DE22	DM34	DM35	DL31	DL32	DL30	DF23	DL33	Sector
TF13	0.24																						**Chemistry**
TF14	0.16																						
TF5	0.15																						
TF12	0.07	0.57																					
TF15	0.09		0.01																		0.44	0.09	
TF6	0.09		0.06																				
TF21			0.01							0.46													**Machine building**
TF19			0.02											0.046									
TF25	0.01		0.02				0.01							0.04									
TF24	0.01		0.10			0.03												0.10					
TF2			0.02		0.25			1															
TF11			0.12	0.38		0.04							0.29	0.047									
TF23	0.00		0.17	0.15		0.02										0.14							
TF8	0.02	0.03	0.09	0.23		0.09			0.49														
TF18	0.02		0.03				0.01				1												
TF7			0.06			0.17	0.46																
TF3			0.06	0.11		0.11						1	0.71	0.47									
TF1			0.02		0.59	0.04																	
TF17	0.02		0.02			0.00	0.28													0.10			
TF20		0.40	0.04	0.071		0.47	0.20		0.51							0.04	0.23			0.00			
TF9			0.02			0.01				0.54				0.22	1								
TF22			0.07			0.01								0.09		0.24	0.02	0.21				0.03	**Transport equipment**
TF10			0.03	0.03												0.50	0.75						
TF29	0.00	0.00	0.00			0.00												0.02	0.34	0.13		0.00	**Electrics & electronics**
TF30				0.03			0.05									0.00		0.54	0.15	0.32		0.12	
TF27														0.02				0.11	0.083	0.022			
TF28																		0.019	0.38	0.19		0.05	
TF31																			0.05	0.24	0.56	0.28	**Medical & optical equipment**
TF26	0.04		0.01													0.09							
TF16	0.02		0.00		0.16								0.08									0.07	
TF4	0.04		0.00																			0.36	
Shares	**0.71**	**0.57**	**0.80**	**0.94**	**0.84**	**0.99**	**0.95**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.81**	**1.00**	**1.00**	**0.73**	**0.77**	**0.69**	**0.95**	**0.66**	**0.56**	**0.71**	

Table 4: Industry matrix: GIC and technological fields of the *Patentatlas*

	DC24	DI26	DK29	DH25	DA15	DJ28	DJ27	DA16	DC20	DE21	DB17	DB18	DC19	DN36	DE22	DM34	DM35	DL31	DL32	DL30	DF23	DL33	Shares	Sector
TF13	1																						**1.00**	Chemistry
TF14	1																						**1.00**	
TF5	1																						**1.00**	
TF12	0.58	0.42																					**1.00**	
TF15	0.67		0.17																		0.17		**0.67**	
TF6	0.32		0.42																			0.26	**0.42**	
TF21			1																				**1.00**	Machine building
TF19			0.97							0.03													**1.00**	
TF25	0.17		0.75											0.08									**0.83**	
TF24	0.04		0.72				0.01							0.01				0.18					**0.78**	
TF2			0.68		0.26	0.04		0.06					0.01										**1.00**	
TF11			0.66	0.28		0.04								0.01									**1.00**	
TF23	0.00		0.64	0.08		0.01										0.27							**0.72**	
TF8	0.08	0.01	0.60	0.20		0.10			0.01														**0.92**	
TF18	0.21		0.52				0.01				0.26												**0.79**	
TF7			0.49			0.23	0.28																**1.00**	
TF1			0.43		0.43	0.14																	**1.00**	
TF17	0.21		0.37			0.01	0.41									0.09	0.10			0.10			**0.79**	
TF20		0.08	0.16	0.04		0.35	0.07		0.01														**0.62**	
TF3			0.47	0.11		0.16						0.06	0.04	0.16						0.01			**1.00**	
TF9			0.31			0.01				0.01				0.16	0.49								**0.99**	
TF22			0.35	0.01		0.01										0.58	0.01					0.05	**0.59**	Transport equipment
TF10			0.07											0.01		0.60	0.18	0.12					**0.79**	
TF29	0.17	0.01	0.01			0.01												0.71	0.38	0.08		0.08	**0.71**	Electrics & electronics
TF30				0.01			0.01											0.36	0.37	0.46		0.15	**0.83**	
TF27																0.01		0.16	0.37	0.08			**0.99**	
TF28														0.01				0.07	0.53	0.08		0.31	**0.68**	
TF31																			0.77	0.23			**1.00**	
TF26	0.07		0.04													0.15			0.07	0.20	0.05	0.41	**0.46**	Medical & optical equipment
TF16	0.24		0.01		0.12																	0.63	**0.63**	
TF4	0.12		0.02											0.02								0.84	**0.84**	

Table 5: Technology matrix: GIC and technological fields of the *Patentatlas*

References

Backhaus, K., Erichson, B., and Plinke, W. (2000). *Multivariate Analysemethoden*. Springer - Verlag Berlin, Heidelberg.

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

Breschi, S. and Lissoni, F. (2001). Knowledge Spillovers and Local Innovation Systems: A Critical Survey. *Liuc Papers, Serie Economia e Impresa*, 84(27):1–30.

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. (2007a). Identifying Innovative Regions: An Application of a Data Envelopment Analysis for German Regions. *Paper presented at the DRUID Winter Conference 2007, Aalborg, Denmark, January 25-27, 2007*.

Broekel, T. and Brenner, T. (2007b). Measuring Regional Innovativeness - Methodological Aspects and an Application to the German Electrics Industry. *Paper presented at the "Interdependencies of Interactions in Local and Sectoral Innovation Systems" workshop at the Friedrich-Schiller-University and the Max Planck Institute for Economics, in Jena, March 22-24, 2007*.

DESTATIS (2002). *Klassifikation der Wirtschaftszweige, Ausgabe 2003 (WZ2003)*. Statistisches Bundesamt, Wiesbaden.

Eckey, H.-F., Schwengler, B., and Türck, M. (2007). Vergleich von deutschen Arbeitsmarktregionen. *IAB Discussion Paper*, 3.

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

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

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

Jaffe, A. B. and Trajtenberg, M. (1996). Flows of Knowledge from Universities and Federal Laboratories: Modeling the Flow of Patent Citations over Time and Across Institutional and Geographic Boundaries. *Proceedings of the National Academy of Sciences of the United States of America*.

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*.

Schmookler, J. (19966). *Invention and Economic Growth*. Harvard University Press, Cambridge.