KATHOLIEKE UNIVERSITEIT LEUVEN

Counterfactual impact evaluation of cohesion policy

Work package 2:

Examples from Support to Innovation and Research

Final Report

December 2011

Project team:

Prof. Dr. Dirk Czarnitzki (K.U.Leuven)

Cindy Lopes Bento (K.U.Leuven)

Thorsten Doherr (ZEW)

Contact:
Prof. Dr. Dirk Czarnitzki
K.U.Leuven
Dept. of Managerial Economics, Strategy and Innovation

Faculty of Business and Economics
Naamsestraat 69
3000 Leuven
Belgium

Phone: +32 16 326 906
Fax: +32 16 326 732

E-Mail: dirk.czarnitzki@econ.kuleuven.be

Contents

Figures

Tables

0 Executive Summary

The goal of this project was twofold. The first objective was to find out to what extent publicly available beneficiary data (from managing authorities and commercial databases) could be used for quantitative, econometric counterfactual analysis. This led to a second objective: in the countries and regions where data was most promising, a treatment effects analysis of the impacts of Cohesion Policy on innovation activities at the firm level has been conducted.

Data requirements and collection

Data from Belgium (Flanders), the Czech republic, France, Germany, Poland, Slovakia, Slovenia, Spain and the UK (Wales and London) were examined for two crucial data requirements:

i. it had to be possible to identify from the published beneficiary lists, which firms had been assisted (and which had not) . This is essential since counterfactual approaches compare the activities of beneficiary firms with an appropriate control group of firms that did not receive support. An appropriate control group contains a large enough sample of firms for a meaningful comparison;

ii. information on innovation (or other business activities) and further firm characteristics, such as size and sector, had to be collected.

Spain had to be eliminated from the analysis, since the published data did not specify the assisted firms clearly enough. Poland, Slovakia, Slovenia, Flanders, Wales and London were eliminated because of the relatively small number of innovation projects supported by Cohesion Policy.

Data for the Czech Republic, France and Germany were processed using two different strategies:

a) by linking the Cohesion Policy beneficiary lists to the Amadeus database that contains balance sheet data of firms from all EU countries. This was done for support recipients in France and the Czech Republic. Firms included in Amadeus that were not identified as Cohesion Policy grantees were used as control group. Both the identified recipients and the control group were then linked to a patent database, patent being used as a proxy variable for innovation activities at the firm level:

b) for the German Cohesion Policy recipients more complete data were available. These beneficiaries were linked to the German part of the Community Innovation Survey (the “Mannheim Innovation Panel" (MIP)). Similarly to the Amadeus database, this survey delivers a control group of firms (respondents within the survey that did not receive support from Cohesion Policy) as well as information on general firm characteristics. What is different, however, is the fact that the MIP contains several other measures of innovation activities. Instead of only analyzing patenting behavior, this database allows investigating broader facts such as R&D investment and R&D employment, total innovation expenditure of firms, as well as their innovation outcomes (process innovations, product innovations, etc.).

In the Czech Republic, 26,075 different grants were distributed within Cohesion Policy between 2006 and 2011 amounting to a total budget of almost € 11 billion. As from the grant recipient lists it is not clear how many different recipients have received the 26,075 projects nor how many firms have been supported (as the beneficiary lists only include names of the awardees but not the type of the entity, i.e. a firm, a not-for-profit institution, a municipality or private person, for instance), all names were searched in the Amadeus database which contained addresses of 14,609 different firms. Among those, 1,433 could be identified as Cohesion Policy recipients.

The equivalent procedure has been applied to the French data where 36,858 projects were listed in the beneficiary data that amounted to almost € 16 billion. In the Amadeus database, 1,231 firms were identified as Cohesion Policy beneficiaries (out of about 900,000 firms in total in the Amadeus data).

The German recipient data included 47,616 projects with a total amount of about € 9 billion. Although 1,904 different firms could be identified in the MIP survey as Cohesion Policy recipients, the subsequent econometric analysis could only use 623 different recipient firms, as the other firms did not participate in the survey in the relevant years for this study (the years 2007 to 2010 where the current round of Cohesion Policy has been active).

Econometric results – Czech Republic

For the data of the Czech Republic, we applied a difference-in-difference (DiD) estimator, i.e. patenting activity of recipient firms is compared over time and related to changes in patenting activity of the control group in the same time period. In this particular case, the pre-treatment phase was chosen to be patenting activity during the years 1993-2003 (time before the current round of Cohesion Policy) and the treatment period was chosen to be 2008 and 2009 (although the programme was formally active in 2006 and 2007, very few grants were distributed in these years). The overall patenting activity in the Czech Republic is characterized by a decline when the early 2000s are compared to the end of the 2000s. However, the econometric DiD estimation reveals that the treated firms actually suffered less from a reduction in patenting, as proxy for their overall innovation activities, than the control group of non-recipients in the same time period. While patent activity fell by 63% in control firms, it only fell by 14% in assisted firms (and, for reasons specified below, this probably understates the impact of Cohesion Policy). We thus conclude that Cohesion Policy had positive impacts on innovation in the Czech Republic.

Econometric results - France

For the French data, the same procedure as in the Czech Republic has been applied. However, we cannot confirm positive impacts of Cohesion Policy in France. This is most likely due to the fact that the French recipient data lack some important information that has been present in the Czech data: the French authorities did not provide detailed dates of the approval of each project. Thus it was impossible to be certain about the "before" and "after" periods which possibly led to misclassification of firms as treated or non-treated, and consequently bias in the regression results. This problem cannot be solved without improved data reporting by the regional authorities themselves. The recommendations at the end of this summary will follow up on data reporting issues.

Econometric results - Germany

In the German case, we do not have panel data at our disposal, i.e. the DiD method cannot be applied. Although we use four years of the survey, most firms are only observed once in the respective time period. Therefore, we apply cross-sectional methods for constructing counterfactuals, in particular nearest neighbor matching. Using this method, each treated firm is compared to the most similar firm in the control group. As our control group is large (about 20,000 firms), the application of matching is feasible in the sense that for each treated firms, a very similar "twin" firm can be found. After the matching has been performed, one can compare the differences in innovation variables between the treated firms and the selected twin firms. If differences in the innovation (or "outcome") variables are found, these can be attributed to the Cohesion Policy as two groups of firms do not differ in any other structural characteristic but the treatment receipt. After applying an initial matching estimation, it turns out that indeed the Cohesion Policy recipients score higher on a range of innovation indicators; for instance, R&D investment, R&D employment and total innovation investment, among others.

As we have detailed information on other activities of the firms, however, we also find that the Cohesion Policy recipients are also more likely to benefit from other subsidies, e.g. from the Federal German Government, than the firms in the selected control group. Thus, the identified effects might be confounded with impacts of other subsidies. Therefore, we took this into account in a subsequent analysis. After controlling for other subsidies, the results with respect to the estimated treatments effects become slightly less favorable, but stay positive. However, we can no longer identify a separate effect of Cohesion Policy on R&D employment. This might be due to relatively small amounts of money that are granted per project in Germany (especially Eastern Germany), but more research on this issue would be needed in order to validate this hypothesis.

In terms of magnitude of treatment effects, we find the following results after controlling for other subsidies that the firms might have gotten (more numbers are presented in the main body of the report):

	Selected control group	Subsidized firms	Treatment effect on the treated
Variable	Mean	Mean	in percentage points
R&D intensity	4.4%	6.2%	1.8%-points
innovation intensity	7.3%	9.5%	2.2%-points
investment intensity	31.9%	53.2%	21.3%-points

Note: R&D (innovation) intensity is calculated as R&D expenditure (total innovation expenditure) divided by sales times 100, and investment intensity is calculated as gross investment into tangible assets divided by stock of tangible assets (at the beginning of the period under review).

The subsidized firms in the German sample show an average R&D intensity of 6.2%. If they had not gotten a grant within the Cohesion Policy Programme, we estimate that they would have only achieved an R&D intensity of 4.4%. Thus the estimated treatment effect amounts to 1.8 percentage points. This number is not unreasonable. A representative subsidy recipient (the median firm in the sample) would have had R&D expenditure of 213,000 EUR and sales of 4,869,499 EUR (R&D intensity is about 4.4%) if it had not gotten the subsidy. As a response to the subsidy, it increases R&D expenditure to about 300,000 EUR according to our estimates. Thus, the treatment effect in terms of EUR amounts to 87,000 EUR, on average. This is plausible as the typical grant size in our German sample varies between 11,000 and 51,000 EUR. The estimated responses to a subsidy receipt are similar for total innovation investment and also for investment into tangible assets.

Policy Recommendations

The main lessons learned for policy actually deal with the goal concerning the feasibility of quantitative evaluations using the publicly available data on beneficiaries as published by the European Member States or regions respectively. During the data collection and preparation phase for the econometric application several shortcomings of the current reporting standards have been identified. This lack in data quality may result in measurement error and consequently in biased estimation results. Despite the fact that several countries or regions published their data in reasonable quality (examples in addition to the Czech Republic, France and Germany are Poland, Slovakia, Slovenia, Flanders, Wales, London, among others), none of them provided information at a level of detail desirable for an econometric evaluation. Typically, the names of the recipients, a project title, a date of approval and the granted amount in EUR or national currency are published. However, just minor improvements in reporting standards would facilitate the feasibility of future evaluations and their reliability enormously. For instance, in addition to the information already available future publications of beneficiary data on the Internet should include:

(i) not only the name of the recipient but also the location to ease identification within text field searches of recipients in other external data resources;

(ii) instead of an approval date of the grant, a start date and end date of the envisaged project should be provided. This information is essential for assigning the “treatment period” correctly in time;

(iii) the type of recipient should be provided at least in some rough categorical manner (e.g. firm vs. other entity);

(iv) the purpose of the grant should be provided, i.e. innovation project, general business support, environment, energy, tourism, culture etc.;

(v) and last but not least, the data should be published in database compatible formats and not as pdf documents or similar formats that cannot be imported into relational databases without further manipulation.

Given these shortcomings in data reporting structure, **the actual policy conclusions that can be drawn from the presented econometric analyzes are modest and should be interpreted with care.** At the very least, the quantitative results suggest that Cohesion Policy had a positive effect on innovation activity in recipient firms. The magnitude of these effects should not necessarily be stressed too much. In fact, **we believe our results underestimate the real effects as not all treatments could be accurately assigned to the correct timing of the actual grants and it could not be identified precisely which grants were meant for innovation projects and which for other purposes.** Thus, in the present study we might take into account grants that were not meant for innovation, but we relate these to innovation outcomes. This will obviously underestimate the true effect of innovation grants. Nevertheless, in the country-case study of Germany, we could already support the hypothesis that Cohesion Policy is not a substitute for other policies but offers some unique programme features that leads to complementarity with, for instance, the German Federal grants for innovation projects.

1 Introduction

European regional policy is designed to reduce the gap between the development levels of the various regions. From a scientific approach, regional policy brings added value to actions on the ground. The goal of this policy is to help to finance concrete projects for regions, towns and their inhabitants. The idea is to create potential so that the regions can fully contribute to achieving greater growth and competitiveness and, at the same time, to exchange ideas and best practices[1].

In this context, the Cohesion Policy is spending some €80 billion on enterprise and innovation support in the current period, representing a higher amount than the one spent on transport or human resources. In fact, innovation is the only field to be a key priority for Cohesion Policy in all Member States. Yet, evidence of impacts of the funds attributed to enterprises and innovation is very uneven throughout the regions. The evaluations vary in quality from serious to poor or simply non-existent. Even the Member States or regions which deliver serious evaluations of the impact of the current program produce only descriptive evaluations. Hence, there are very few examples of quantitative, causal assessments using counterfactuals or comparison groups. For such a key policy, being able to rely on quantitative results on top of qualitative evaluations is thus crucial.

In this vein, DG REGIO of the European Commission launched Work Package 6c of the ex post evaluation of cohesion policy 2000-2006, with the goal to pilot the use of such evaluations. The long term goal is to build up a body of evidence on enterprise support (including support for innovation and research) from the European Regional Development Fund (ERDF), and have evaluations done on a regular basis. To this end, DG Regional Policy has commissioned:

- An impact evaluation of ERDF support to enterprise (other than support specifically for innovation and research).

- An impact evaluation of ERDF support specifically for innovation and research in enterprises - the current study.

The two evaluations are conceived as complementary and parallel. The current study is divided into 2 main parts; (a) data preparation, (b) econometric analysis, in particular the estimation of treatment effects using counterfactual analysis.

As part of the contract, K.U.Leuven delivers hereby the final report five month after the interim meeting as agreed.

[1] http://ec.europa.eu/regional_policy/policy/why/index_en.htm

2 Goal of this project

The goal of the proposed research project is an evaluation study of ERDF support for R&D and innovation, in particular the application of treatment effects estimators on beneficiaries of ERDF support. More specifically, the goal is to undertake such an analysis without conducting a special survey or interviews to collect the necessary data, but to investigate to which extent data published by the Member States' benefiting regions can be used for such an analysis. More concretely, after assessing if, and to which extent, the published data by the Member States is usable for evaluation purposes, it will be explored to what extent the beneficiary firms of ERDF support would have engaged in innovation activities if they had not received public funding. The latter describes a counterfactual situation that cannot be observed, and thus has to be estimated with econometric techniques. The comparison of the actual innovation engagement of recipients with the estimated counterfactual situation then allows drawing conclusions on the effectiveness of the ERDF support on R&D and innovation. This exercise is highly interesting as the Member States select regions to be supported based on heterogeneous criteria and also favor different varieties of policy instruments. For instance, a country might favor policies for technological consultancy services whereas another country focuses on direct grants for proposed R&D projects. Thus, the variety of policy instruments applied across regions may have heterogeneous effects on enterprise innovation in the EU.

Conducting the proposed exercise involves linking the published beneficiary information to firm level data, such as the AMADEUS database, and external innovation data, such as patent databases. Thus, the ultimate goal of the project is twofold: on the one hand, it will be a pilot study on counterfactual impact analysis of the ERDF, on the other hand, it will lead to advice on future reporting standards for the Member States in order to facilitate and improve future econometric evaluations.

The following subsection will give a detailed overview of what has been done under task one. First we will present the data collection and merging exercise. Then, we will provide a detailed overview of the problems encountered and recommendations on how similar problems can be avoided in the future. Before going over to the econometric analysis of task two, we illustrate the steps of task one with an example using data from the Czech Republic. Subsequently, an analogous exercise is carried out using French data. Finally, we consider data from Germany.

2.1 Task 1: Data preparation

Information on beneficiaries of ERDF support has been collected from the following website:

http://ec.europa.eu/regional_policy/country/commu/beneficiaries/index_en.htm

The goal of the present project is to conduct a counterfactual impact analysis of current ERDF policies on Czech Republic, France and Germany. These countries have been selected for the following reasons: (i) there were many projects granted to recipients in these regions (ii) the beneficiary data has been of reasonable quality (iii) external firm level and innovation data has been available for drawing a control group of non-recipient firms and for performing an econometric treatment effects analysis. In order to make a decision on the selected countries, data on many more regions needed to be collected as their quality was unknown to the research team. Only after assessment of the data quality per country, the Czech Republic, France and Germany were chosen for the subsequent econometric exercise.

2.1.1 Examples of linked beneficiary data with firm level information

Before we document how the data collection was performed at large scale, we briefly discuss a few examples of the linked grant data with firm level information so that the reader gets an impression of what kind of information is processed in the subsequent econometric applications. We use examples from German grant recipients here, as we have the most comprehensive firm-level data available for this country.

Example 1:

The first example is small technology-oriented company working on the development and production of specialty polymers and adhesives. Founded in 1996 they have been developing high-temperature resistant adhesives .At its research and production facility, products are customized to meet various technical requirements of the clients by using the technology of interpenetrating networks. In total, this firm received three grants within the current ERDF program. In 2008, a project was granted for the development of a nano-composite adhesive with a total value of € 227,000. In 2009, € 2,500 were granted for developing the company’s strategic business concept for the Chinese market, as well as € 44,000 for entering the Chinese market for polymers.

The employment of the company grew slightly since the year 2000 where they had 7 employees. In 2010, the company had 10 employees, and except one all are R&D employees. According to our information the firm is permanently innovating products and processes, and spent on average about € 250,000 for innovation projects, on average. In 2005, the firm was nominated for an Innovation Prize of their regional government for a new product, a special fire prevention foam. Between 2002 and 2004, three patents were filed at the European Patent Office out of which two were granted.

Despite its high innovativeness, the company never managed to translates its inventions into products that generate sales sufficient for survival. Although the company’s sales doubled between 2001 and 2010, they are still very low. In the year 2010, the sales were not higher than € 200,000.

According to a German credit rating agency, the firm is in financial trouble. Last year’s balance sheet was characterized by negative equity, for instance. It is thus questionable whether the company can stay in the market.

Example 2:

The second company engages in manufacturing of high-quality endoscopic equipment for minimal-invasive surgery and exports both accessories and complete systems to about 30 countries world-wide. The products are being manufactured in an own production facility. The firm takes an active role in research. With its own R&D department it was able to develop products that have set standards in endo-surgery. Among other public support, several projects have been funded by the German government and are being tested by German university professors and clinics.

The total employment of the company declined over the last 10 years from 45 to 28 employees. On average, the firm spends about € 2 million of innovation projects per year and always had about 20 R&D employees in the last 10 years. In the last decade, the firm filed 9 patents at the European Patent Office out of which 2 were granted.

Within the current round of the ERDF, the firm received three grants for its research on non-invasive medical precision instruments. One in 2007 and one in 2008 had an amount of about € 2 million each, and the third grant amounted to € 109,000 in 2009.

Since the mid-2000s, the firm’s sales fluctuate around € 2 million per year.

2.1.2 Collecting and merging the data of the retained regions

This subsection describes how the data was collected at large scale, assessed and merged to other datasets. After having downloaded the data of all the beneficiary regions of the Member States, the ones that had data where the quality was judged to be sufficiently good for further analysis have been retained for the next step of the data preparation exercise (see Annex A for a detailed overview of the data), consisting in the conversion of all the collected data into a harmonized, data compatible format like e.g. excel. After this step, the separate regional sheets have been converted into complete database tables. All excel-sheets have been exported into separate ASCII files using tabs as a column delimiter. All exported files were then concatenated into one file. The resulting file has been exported into file formats that could be used with statistical software.

The next step consisted in linking this publicly available data to an external dataset, i.e. the Amadeus database of Bureau Van Dijk in the case of the Czech Republic and France and the Mannheim Innovation Panel (MIP) for Germany. These linkages allow getting further information of the beneficiary firms, and further allow drawing a control group of non-beneficiaries for each selected region.

Subsequently, innovation data has been collected for both, the selected beneficiaries and the control groups. For the cases of the Czech Republic and France, the research team focused on patent data as innovation indicator. Although patents are admittedly a somewhat narrow measure of innovation (see e.g. Griliches, 1990, for a survey on the pros and cons of patent data for economic analysis), they have the advantage that data for the entire patentee population is available for a long time period (1978 until to date) for the whole EU27. Different sources of patent data have been used. First, the database of the European Patent Office (EPO) has been searched. Second, the PATSTAT database has been used. In comparison to the EPO database, the PATSTAT database does not only cover patent filing to the EPO but also to 40 different national patent offices. However, the quality of the applicant names and addresses is lower in the PATSTAT than in the EPO database.

For the German counterfactual exercise, the research team established a link of the beneficiary data to the Mannheim Innovation Panel. Unlike the patent databases that cover the population of patents, the MIP is an annual survey of German firms. Thus it does not cover the population of firms, but a randomly drawn, representative sample of the German business sector each year. It has the advantage that other innovation measures than only patent data can be investigated.

The links between the various data sources has been established by using a text field search engine that allows highly sophisticated string searches across databases. The search engine allows minimizing potential wrong matches due to different spellings of firm names or firm variations. This technique is outlined in Appendix C of this report. All potential hits of the text field search engine have been manually checked.

During the data preparation task, several problems and drawbacks were encountered. The following subsection intends to clarify what these caveats consisted in and gives recommendations on how they could be avoided in the future.

2.2 Caveats and recommendations

A first major drawback rendering the data collection exercise cumbersome is the way the data are reported by the Member States. The data of the various regions are published in many **different formats** (like e.g. html, excel, word, pdf), some of which are not database compatible formats. Hence, before being able to use the publicly available data (even for very basic exercises like mere descriptive statistics for example), the latter have to be converted into a database compatible format. For example, some countries provide easily accessible data on Microsoft Excel format which can be collected in one large beneficiary database. Others, however, are in various HTML formats which either requires a manual “copy-paste” collection or the development of some “web-crawling” software (or similar procedures) that identifies fields (such as beneficiary name, date of funding, amount of funding) in the HTML source code and translates it into a database-readable format. Finally, some data is just provided as pdf documents which will require a fully manual transformation of the provided information into a database. Hence, depending on the original format of the data, this conversion process can be very complex, time-consuming and requiring advanced IT skills. The table in appendix A provides a detailed overview of the different formats the data is available on the regional websites.

<u>**Recommendation 1**</u>: **Harmonized format of data publication**

All the Members States should publish their data in the **same, database compatible format** (or at least a database compatible format).

This would highly facilitate and accelerate the data collection exercise. It would allow to immediately export the different regional spreadsheets into ASCII files, enabling to export the data into almost any statistical software.

A second drawback is the use of **special characters** included in the alphabet of some EU countries (see e.g. Czech Republic). Those characters can render the above mentioned exercise of constructing spreadsheets like e.g. excel into ASCII files substantially more cumbersome as some of these special characters are not recognized as letters. Hence, some advanced IT skills allowing to circumvent this issue are needed.

On similar grounds, sometimes the published information is solely available in the **national language** of the concerned country. For certain languages, this can render the researcher's job substantially more difficult, as he or she might not be able to properly understand what the various projects/purposes of the regions are about.

<u>**Recommendation 2**</u>: **Avoidance of special characters and common language (optional)**

Since special characters used in some of the EU languages can render the data conversion exercise increasingly difficult (and might even cause the loss of some observations), it would be recommended that such characters be avoided to the largest extent, by e.g. publishing the data-related information in a **common language** like for instance English. Of course, special characters cannot be avoided in the beneficiary names in certain languages.

Having the information available in English would further allow having a better understanding of what the different projects are about, allowing for more precise evaluations (i.e. evaluations on a specific topic). As will be demonstrated in the following subsection, the lack of understanding the project categories might render it impossible to evaluate solely projects of a specific purpose, given that the evaluator might be unable to differentiate between the different projects categories.

<u>Note:</u> The research team is aware that these two recommendations are very sensitive issues and might not be realistic propositions for Member States (hence, optional). They are issues that can be dealt with. However, for reasons of completeness, the research team felt they should be mentioned as part of the recommendations in the present report.

The **lack of information published** by the managing authorities constitutes a further important shortcoming of the way in which Member States currently publish their data. In order to be able to use an observation for econometric analysis, more information about the beneficiary is needed than is published by the managing authorities. As already previously explained, the data is matched to other datasets. However, this exercise is often not possible because we do not have the necessary information to complete this match between two datasets. As a matter of fact, many of the websites only provide names of the recipients but not the full address. While this might seem sufficient, it can cause important caveats when trying to merge the beneficiary data to other databases like e.g. the Amadeus data or patent databases. If a firm name exists several times in a same region, which can easily happen, or many similar names exist, it will be impossible to identify which of those firms is the actual recipient of a subsidy when merging the data to external firm level data. This can lead to a substantial loss of observations in the treated as well as in the control group (see next section for an illustration).

Recommendation 3: More detailed information on beneficiaries I

It would be recommended to complete and harmonize the way regions report the information on the beneficiaries. The reported information should include:

- The full name of the recipient
- In the case of firms: the legal form of the firm
- The complete address (including zip-code) of the recipient

Additional information that is missing for many regions is the exact duration of the project. While some regions report start and end dates, this is not done systematically by all of them. Having information about yearly expenses would even be more useful, as one could take the distribution of money spent over time into account.

Recommendation 4: More detailed information on beneficiaries II

It would be recommended to have information on the exact duration of the project and the **amount of money spent per year**. Having information on yearly project expenditures would allow us to take the distribution of expenses over time into account. Hence, ideally, regions would report:

- Amount of money spent per year (in €)
- Start date of the project
- End date of the project

In case this would be too cumbersome for the Member States in terms if reporting, having the starting and finishing date would already be helpful:

- Start of project
- End of project

Finally, in order to avoid inaccuracies when converting national currencies into Euros, it would be recommended that all the amounts could be reported in **Euros**. As a matter of illustration, the Czech exchange rate had fluctuations of up to 20% during the period under review.

Finally, many of the ERDF beneficiaries are not firms, but local authorities or universities or other, non-profit organizations. Those beneficiaries cannot be found in the Amadeus database and as a consequence, no hit can be found for the latter. Furthermore, often it is not possible to distinguish the various purposes of the attributed grants because they are not reported by topic.

Recommendation 5: **Clearer structure in the reporting of the data**

In order to avoid ambiguities to the largest extent possible, it would be recommended that the beneficiaries be reported according to whether they are private firms or municipalities, public research centers / universities or other organization that would not be found in external firm datasets.

In a similar vein, and in line with has been suggested in the 1st recommendation, the projects should be **reported by topic**. This would allow having a clear overview of how many subsidies have been spent for what purposes. In our case, this would allow us to identify the beneficiaries of public support for innovation and R&D. If the reporting does not allow this, it may be possible to identify the purpose of the project through text field searches in the titles of the particular grants, or from other information on the different policy actions taken in the beneficiary regions (e.g. regions could be identified on basis of their proposed policy instruments so that the selection focuses on regions that included a large part of measures dedicated to innovation). However, the latter method is much more time consuming and less precise. Furthermore, if the information is published in a language unknown to the investigator, it might well be that the purpose of the grant might not be identified accurately.

Hence, ideally, the data would be organized as follows. The categorization below is based upon European Union, Directorate-General for Regional Policy (2010).

- Private firm beneficiaries
    - Innovation
    - Research and Development
    - Business support
    - Information and communication technologies (ICT)
    - Environment
    - Energy
    - Transport
    - Urban and rural development
    - Tourism and culture
    - Education and social
- Local / regional / national authority beneficiaries
    - Idem
- Public research center / university beneficiaries
    - Idem
- Other non-profit organization beneficiaries
    - Idem
- Etc.

Lastly, it has to be noted that one important recommendation is of course that the data reporting structure be the same in all the Member States. Appendix B provides a table suggesting how the reporting structure could be improved.

2.3 Illustration using data from the Czech Republic

In this subsection, we demonstrate how the above explained caveats impacted the data collection and merging exercises in the case of the Czech Republic, and what the consequence is for the data that will be used in the subsequent analysis.

We will start by giving the example of a successful match, meaning a successful link between the publicly available data and the Amadeus dataset. In other words, the beneficiary firm could successfully be linked to a firm of the Amadeus dataset using an automated search engine:

Example 1: Successful match

searched	found	identity	equal	beneficiary	city
3633			1	Lias Vintírov, lehkýstavební materiál k.s.	
3633	CZ46882324	99.87		lias vintirov, lehkystav. material k.s.	chodov u karlovych var 1

Given that the name of the searched firm and the found firm is exactly the same and that only one firm was found in the external dataset, it appears trustworthy to assume that the found firm is the actual beneficiary of the grant. Hence, this is a successful hit and we can include this firm in our sample of treated firms, merging it with all the additional information we could obtain form the external dataset. Note, however, that it could be the case in unfortunate situation that a hit is assigned mistakenly. The Amadeus database might not contain all firms of a country. Thus it could happen that the actual beneficiary is not included in the Amadeus database, but a firm with a similar (or the same name). Only information on the firm's address could further help to verify the match.

Example 2 provides an illustration where the success of the match is less straightforward and thus requires manual checking. As we can see in the table, the search engine found several firms containing the word "BEST" in their name, and hence suggests all of them as potential hits for the funded firm. In this case, after manual check, we can conclude that the first firm is the correct one, since this is the only one where the name coincides 100% and where the legal form is the same. As a consequence, we can include this firm in our sample of treated firms. Even though this is more time-consuming, no observations will be lost for a subsequent matching analysis.

Example 2: Usable match after manual verification

searched	Found	identity	equal	beneficiary	City
690			9	BEST, a.s.,	
690	CZ25201859	100	1	BEST, A.S.,	KAZNEJOV
690	CZ25328476	100		BEST TRANSPORT, A.S.	BRNO 34
690	CZ25573322	100		BEST - BUSINESS, A.S.	VYSKOV 1
690	CZ25769090	100		BEST HOLDING PRAHA, A.S.	PRAHA 614
690	CZ00505579	99.81		BEST I.A., A.S.	PRAHA 7
690	CZ45796360	99.81		BEST, S.R.O.	BENESOV U PRAHY
690	CZ46580743	99.81		BEST, S.R.O.	OPAVA 7
690	CZ60281022	99.81		METAL - BEST - LIBEREC, S.R.O.	LIBEREC 1
690	CZ60744995	99.81		BEST BOJKOVICE, S.R.O.	BOJKOVICE
690	CZ62029592	99.81		AGRO - BEST, S.R.O.	CHOCEN 1

Example 3 illustrates a case for which even after manual check, it was not possible to attribute a match to the concerned beneficiary firm. The title of the beneficiary firm is contained in all of the potential hits. Since we have no information on the exact location or the legal form, it is impossible to identify the firm in an external dataset. Hence, no further information about the firm (like e.g. size, sector etc) can be obtained and the observation cannot be used for econometric analysis. As a consequence, this firm will be taken out of the population of beneficiaries. Furthermore, as we are unable to tell which one of the potential hits is the actual funded firm, we do not know for sure which one did not get funding either. Hence, all of the potential hits have to be deleted as control observations as well. Otherwise we would run the risk of using an actual beneficiary as control observation.

Example 3: Un-usable match

searched	found	identity	equal	beneficiary	city
12553			9	Vysocina	
12553	CZ00112062	100		ZEMEDELSKE DRUZSTVO VYSOCINA ZELIV	ZELIV
12553	CZ00125202	100		ZEMEDELSKE DRUZSTVO VYSOCINA	HLINSKO V CECHACH 1
12553	CZ25250213	100		AGRO VYSOCINA BYSTRE AKCIOVA SPOLECNOST	BYSTRE U POLICKY
12553	CZ25573004	100		ZEMEDELSKA, A.S. VYSOCINA	HLINSKO V CECHACH 1
12553	CZ26272211	100		SERVISCENTRUM VYSOCINA S.R.O.	JIHLAVA 1
12553	CZ26297451	100		DRUBEZ - VYSOCINA, S.R.O.	MORAVSKE BUDEJOVICE 2
12553	CZ46992189	100		VYSOCINA, A.S.	TREST
12553	CZ47238381	100		VYSOCINA VYKLANTICE, A.S.	VYKLANTICE
12553	CZ49810162	100		VELKOOBCHOD VYSOCINA, S.R.O.	LEDEC NAD SAZAVOU
12553	CZ60850973	100		VYSOCINA DOLNI HRACHOVICE, SPOL. S R.O.	MLADA VOZICE

<u>Lost observations for further analysis and quality of the remaining data – Example of the Czech Republic</u>

In the case of the Czech Republic, which comparatively has data of good quality, many observations were lost for further analysis, mainly due to two reasons:

- A first reason is the lack of information, as illustrated by the example here above.
- A second reason the fact that not all ERDF recipients are firms, but some are municipalities, universities, hospitals etc., which cannot be found in external firm level datasets.

As an illustration of how this impacted the total number of observations, consider the following figures: **the total number of ERDF beneficiaries** amounts to **26,075**; the number of firms **successfully matched to the Amadeus dataset** amounts to **3,669**. Hence, **22,406** beneficiaries could **not be matched** to an external dataset.

To be able to have a more complete picture of how many beneficiaries are not found in external datasets because of lacking information (but are private firms and supposedly contained in the dataset) and how many beneficiaries cannot be found because are not contained in a firm level dataset (because they concern other beneficiary units than firms), it would be useful, as explained in the previous section, if the beneficiaries were reported according to the type of entity.

Furthermore, is has to be noted that the 3,669 firms that have been found in the Amadeus dataset and will be used for subsequent econometric analysis concern subsidies on **all the types of purposes cumulated**. In other words, with the data at hand, it was not possible to determine the purpose of the grants. Hence, this final dataset does not only concern firms that received support for innovation and R&D, but this dataset contains beneficiary firms for any kind of project that received EDRF support. As a consequence, it might be difficult to evaluate the effect of innovation subsidies on firms' innovation activity. One should not expect to find effects on innovation activity if grants are interpreted as a treatment, although the purpose of the project was not related to innovation at all.

It also has to be noticed that some "questionable" figures have been found in the data. For example, the four smallest amounts of ERDF support in the Czech Republic that have been reported by beneficiary regions range between 22 and 75 Euros. The four largest amounts allocated range between 110,862,300 and 154,182,860 Euros[2]. As one can see, these amounts differ immensely, to the point where some additional information on the data would be desirable. Do those 22 Euros concern a real ERDF contribution, and the concerned firms should stay in the sample, or is this merely the reimbursement of the delivery fees of unsuccessful project proposals, and the concerned firms should be taken out of the sample or serve in the control group? Or are we simply facing a reporting error and the concerned firms should be taken out altogether?

In the above sections, we used the Czech Republic as an illustration on what kind of problems we encountered when preparing the data for the analysis. It has to be noted though, that similar problems were encountered for the other countries as well and that these caveats are by no means specific to the Czech Republic.

Retained countries

During the inception phase, the feasibility of receiving the Spanish beneficiary data in a database compatible format had been assessed. As this was not possible, Spain was not retained for further analysis.

Other regions that had been considered during the inception phase were Poland, Slovenia, Slovakia, Flanders, Wales and London. Unfortunately, because of the low number of beneficiaries,

[2] The conversion from Czech crones to Euros has been made with the exchange rate of January 1st of each year under review.

these countries/regions could not be retained for further econometric analysis. Hence, after the assessment under task 1, it has been decided that the regions of the Czech Republic and France will be retained for task 2. Finally, since we managed to convert the German data into a database compatible format and to merge the Cohesion Policy beneficiaries with data of the Mannheim Innovation Panel (MIP), Germany has been retained for the accomplishment of task 3.

3 Cohesion Policy in the Czech Republic

3.1 Some descriptive statistics

Here below we will display some descriptive statistics on the Czech Republic data. Figure 1 displays the number of projects granted per year. In total, some 26,075 projects have been supported by the EU Cohesion Policy in the Czech Republic between 2006 and 2011 for a total amount of €10,747,210,000 (average amount per project: € 412,265). The bars (linked to the left axis) show the number of projects granted per year[3] and the curve (linked to the right axis) displays the percentage of projects granted per year out of the total number of accepted projects.

Figure 1: Number of projects granted per year

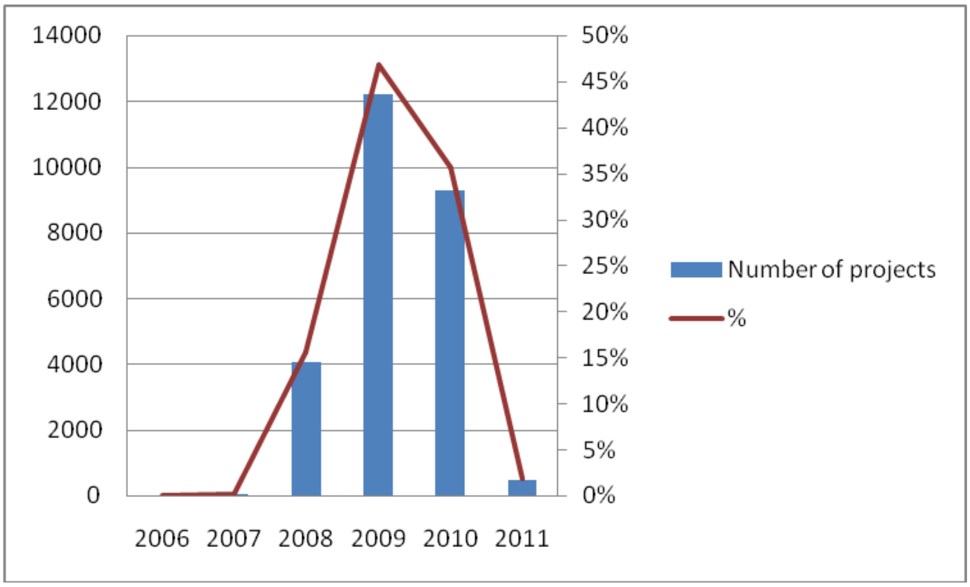
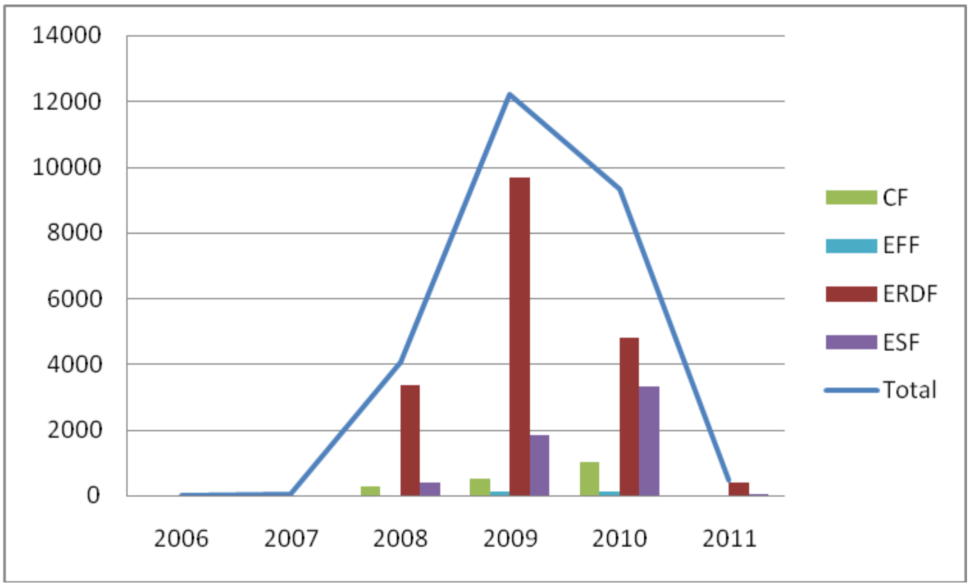


Figure 2 presents the repartition of the granted project by funding type.

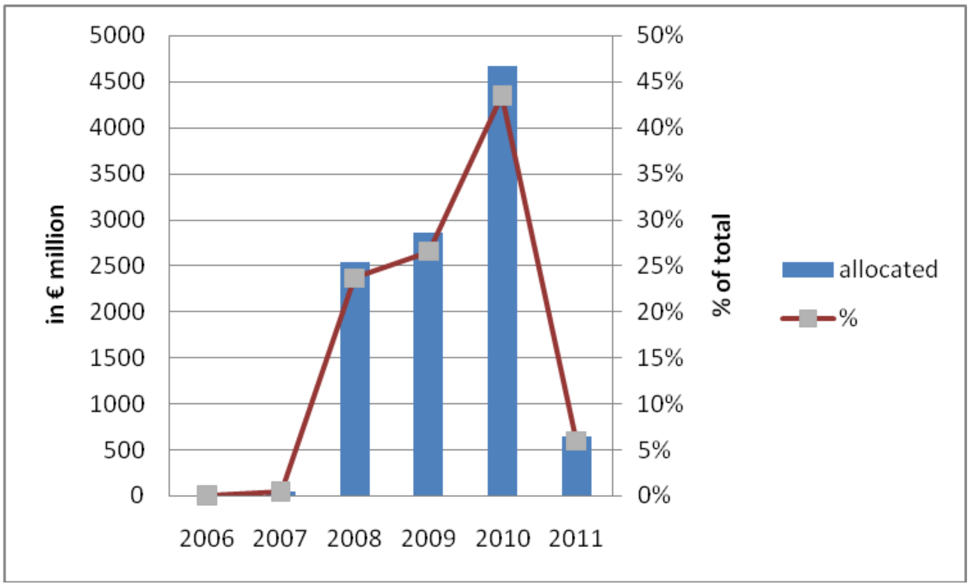
[3] Even though it cannot be recognized on the chart, in 2006 1 project was granted in the Czech Republic and in 2007 34 projects were granted.

Figure 2: Number of projects granted per year by funding type



Finally, Figure 3 presents the amounts allocated per year. As can be seen by this graph, the average amount attributed per year is very volatile. Hence, as previously explained, it would be useful to have the expenditures per year, allowing to take the duration and the money allocation over time per project into account. Being able to calculate monthly expenditures per project would enable to take the distribution of the grants over time into account. Indeed, one would expect that the most of the money gets spent after the kick-off period, and that there are less expenditures the beginning and the end of the project duration.

Figure 3: Amount allocated per year



3.2 Task 2: Econometric Application

The Amadeus database contains 14,609 firms. Out of those, 1,722 are dropped. These are firms that were suggested as potential hits by the text field search engine when the beneficiary data was linked to the Amadeus database. However, during the manual checks, we did not confirm these entries as a "hit", as the information could not be verified accurately (see the example 2 on page 12 of this report). As we want to avoid that the control group mistakenly contains actual recipients, we exclude the non-assigned, potential hits from the further analysis.

The remaining sample contains 1,433 firms that got a project grant, and 11,454 firms that can serve as control group for the estimation of a treatment effect, i.e. an effect of the grants on the innovation activity of firms.[4]

The Amadeus data provides information on firm size, sector of economic activity and various other characteristics, such as location, operating margin, cash flow, number of subsidiaries and the number of shareholders. For our initial match we used an old version of the Amadeus that covers the time period from 1997 to 2004. It seemed to be appropriate to use an old edition of Amadeus as this allows finding firms that possibly went out of business recently. If one would use an up-to-date Amadeus edition, it would not include information on firms that went out of business even if data for earlier periods existed. Using the old version of the Amadeus database, however, has the disadvantage that newly founded firms would not have been found when the beneficiary data has been linked to the firm level data.

The data on the Cohesion Policy grants cover the time period from 2006 to 2011. As, however, the patent data has only been available until 2009 at the time of the data preparation, the years 2010 and 2011 of the project grants cannot be considered. Note that patent databases typically suffer from a time lag concerning the data input. The most recent edition of the PatStat database includes patents from 2010, but it cannot be considered to be complete. At the earliest, the 2012 edition will have the complete information for the year 2010. The firms included in the Amadeus database cover about 24% of all Czech patenting activity in the world (the PatStat database covers the European Patent Office as well as US and Japanese patents and patents of about 40 other national patent offices).

The patent database covers all patent applications from 1978 to 2009 (at the time of data preparation). We use information starting in the year 1997 (the first year where we have information from the Amadeus database.

[4] Note that 3,669 projects could be linked to the Amadeus database. This does not correspond to 3,669 different firms as some firms got multiple grants within the program.

Our database is thus consisting of 4 major "data blocks". The data consists of the treatment group and the control group as well as two major time periods: the pre-treatment time, 1997 to 2005, and the time where the program has been active, 2006 to 2009. As this is panel data, that is, firms and their characteristics can be traced over time, we can apply a simple, yet very powerful estimator for program evaluation: the difference-in-difference estimator. For this, we initially only use the patent data and the treatment information.

The idea of the difference-in-difference (DiD) estimator is based on exploiting the panel structure. Consider a scenario, where we only have two time periods for simplicity: a pre-treatment period, and post treatment period, or more precisely in our case, a time period where the program to be evaluated is active. We can observe both the treated firms and control observations in both time periods.

The DiD estimator works as follows: We could calculate the difference in patenting activity over time for both the treated firms and the control group:

$\Delta_i^T = PAT_{i,t1} – PAT_{i,t0}$

$\Delta_j^C = PAT_{j,t1} – PAT_{j,t0}$

where T denotes the treatment group and C the control group. The treatment group receives a treatment, here the project grant, in period t1. We thus calculate the growth of patenting activity over time. As the growth of patenting activity may well be subject to economic shocks that concern the whole economy, one relates the growth of patenting of the treatment group to the growth of patenting of the control group. The underlying assumption is that both treated and control group would be affected by economic shocks in the same manner. We would thus be able to estimate the treatment effect α as difference in the both differences:

$\alpha^{DiD} = E(\Delta_i^T) – E(\Delta_j^C)$.

The expected value would simply be estimated as the sample average of patenting growth in the treatment and control group respectively. A test whether the treatment effect is positive, that is, the program increases innovation activity, could simply be implemented by a simple two-sample t-test on mean differences in this example.

As our database has not only two periods but multiple years, we implement this test by a simple fixed effects regression (within estimator). We use following time periods for this regression: we use 1997-2003 as pre-treatment period and observe the patenting activity in each year for the treatment and the control group. Although the program started in 2006, the descriptive statistics above show that grants were only distributed systematically in 2008. Therefore we omit all years between 2003 and 2008, and thus our treatment period corresponds to 2008 and 2009.

We are thus interested whether the average patenting activity in 2008 and 2009 of the treated firms is larger than in 1997 to 2003 relative to the control group of not-treated firms. Therefore we define two dummy variables: TREAT and CONTROL. These two dummy variables are equal to one in the years 2008 and 2009 for the TREATED firms and for the CONTROL firms, respectively. Using these two dummy variables, we run the following regression:

$$PAT_{it} = c_i + b_1 TREAT_{it} + b_2 CONTROL_{it} + e_{it}$$

The term *c* is the firm-specific effect, that is the average patenting activity of firm *i*. This captures the average patenting activity for all firms in the sample. The coefficient b_1 will capture the difference in the patenting activity of the treated firms between the pre-treatment periods and the treatment period (*TREAT* is equal to zero in 1997 to 2003, and then takes unit value in 2008 and 2009). CONTROL is defined accordingly for the control observations, and thus the coefficient b_2 indicates the difference in the corresponding time periods for the control firms. The treatment effect for this version of the difference in difference estimator is thus given as $\alpha^{DiD}= b_1 – b_2$, or in other words, we are interested whether $b_1> b_2$.

As patents are a count variable, an OLS regression is not the most appropriate estimation technique to be applied. Patents are a strictly non-negative integer variable, i.e. they takes values 0, 1, 2, 3, and so forth. In addition, the sample will contain many zeros as not all firms patent. Actually only a minority of firms typically files patents. Therefore, researchers typically apply count data models instead of linear regression as these are more efficient compared to OLS, that is, the estimates are more precise (smaller standard errors of the coefficients). Here we consequently apply a fixed effects Poisson regression (with fully robust standard errors).

3.2.1 Version 1: Estimation using a “wrong“ treatment indicator

As we described above the variable TREAT identifies the funding recipients in the period 2008 and 2009. Actually, this is not an accurate definition of the treatment indicator as there are firms in the data that only received the treatment after 2009. However, we would like to show this regression, as it nicely illustrates a bias arising from potentially poorly reported recipient data. The Czech recipient data actually contains information on the date of funding which we will use in the 2nd version of the estimation. We show this “wrong” definition of the treatment indicator, however, as in the case of France the authorities did not report the funding date systematically and thus our estimation reported in the subsequent section on the French case may be subject to bias due to poor data quality on the recipient firms.

In total the regression is performed with 12,887 firms and 9 year, resulting in 115,983 firm-year observations. The regressions results in following coefficients (and standard errors in brackets)

b1 = -0.310 (0.131) **

b2 = -1.025 (0.167) ***

*** indicate a significance level of 1% and ** of 5%.

As we see, the patenting activity in the sample of Czech firms declined in 2008/9 when compared to 1997-2003, as both coefficients have a negative sign. However, we also find that the patenting in the control group reduces more than in the treatment group. The actual growth rates are -63% [= exp(-1.025) -1] and -27% [= exp(-.31)-1]. The declining innovation activity might the result of a negative economic shock, e.g. the financial crisis.

In order to test whether the treatment effect, the difference in the coefficients, is significantly different from zero, we use a Wald test. The test statistic amounts to 11.35, which is significant at the 1% level. We thus reject the null hypothesis of equally sized coefficients and conclude that the treatment effect of the ERDF grants is positive in this setting.

3.2.2 Version 2: Estimation with a corrected treatment indicator

In this version of the estimation we take the timing of the treatment into account. Now the variable TREAT only switches from 0 to 1 in the year when the recipient firm got the grant. This is obviously a more accurate definition of the treatment variable as beforehand some recipient firms were considered as being treated although they received the grant only after the period that is covered by our data.

b1 = -0.149 (0.131)

b2 = -1.025 (0.167) ***

*** indicate a significance level of 1% and ** of 5%.

Now we actually find that the patenting activity of the treated firms did not significantly reduce in 2008/2009 when compared to the pre-treatment period (the reduction amounts to exp(-0.149) – 1 = 13.8%, but the coefficient b1 is not significantly different from zero). At the same time, however, the patenting activity of the control group did decline exactly by the same amount as in the previous version (this is of course expected as the control group did not change). We again use the Wald test on a significant difference of the two coefficients b1 and b2 and find that the Null hypothesis of equal coefficient is rejected at the 1% level (the test statistic amounts to 12.7). We thus conclude that the treatment effect is positive and that it is larger in the case where we can use the exact funding years for construction the treatment indicator when compared to the situation where we cannot use the information on timing of the grant.

Overall we conclude that the treated firms were thus better able to keep up their innovation activities in a time period where innovation activity as measured by patent counts reduced in the economy. This effect can be attributed to the ERDF program.

Although we find positive results here, it is still difficult to interpret the magnitude of the effect in terms of “policy significance” rather than statistical significance. Although we took the timing of the grants into account, we cannot really tell which grants were actually meant for innovation projects and which grants had other goals such as investments into physical assets, for instance. This could only be disentangled if this information were included in the beneficiary data as we recommended in Chapter 2 of this report. Thus, our positive findings on innovation may even underestimate the actual impact of Cohesion Policy as we now take treatments into account that were not at all directed at stimulating innovation in the Czech Republic.

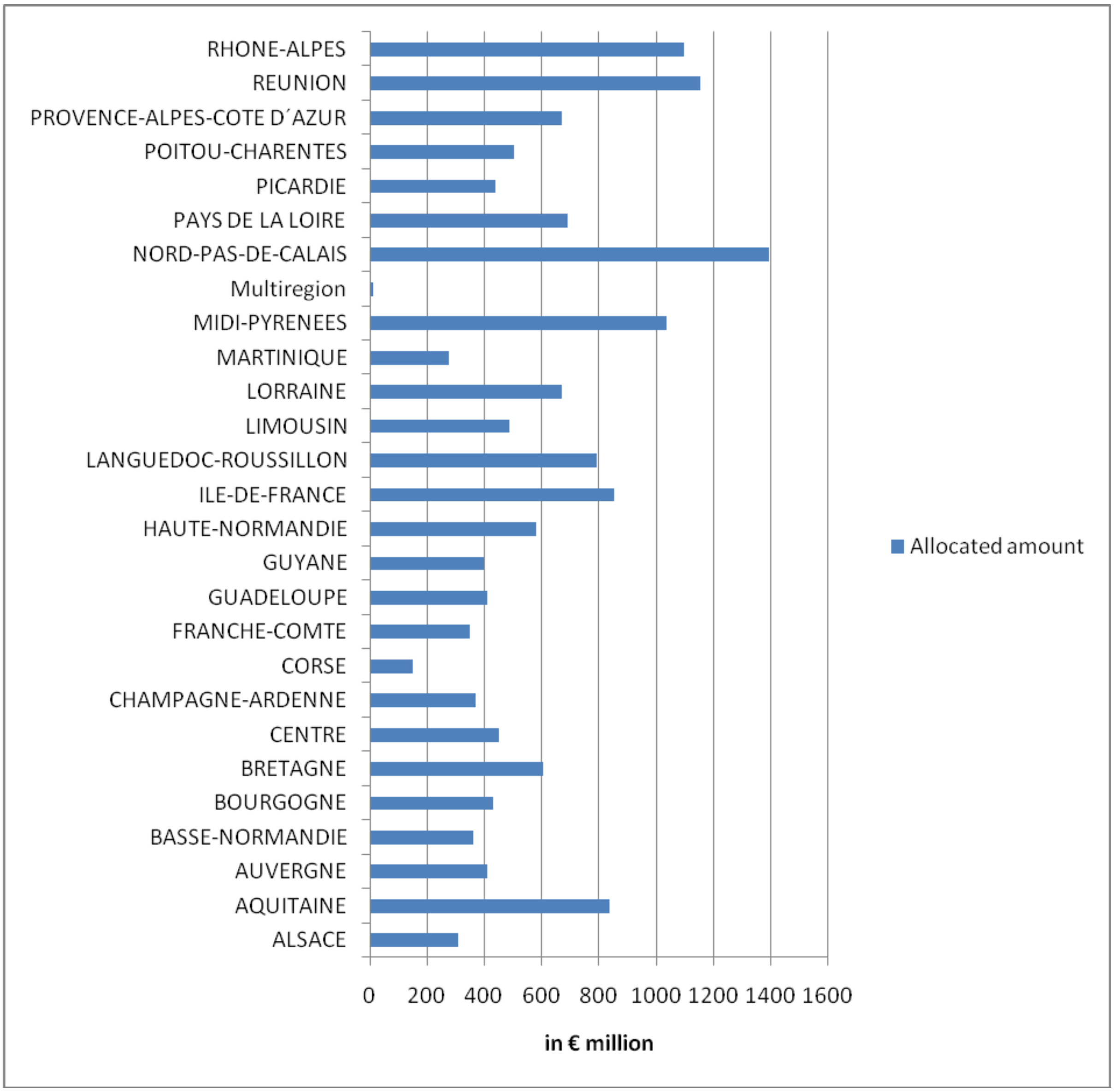
4 Cohesion Policy in France

4.1 Some descriptive statistics

A total amount of € 15,809,310,000 has been spent for 36,858 projects in France for the period under review. This amounts to an average amount of € 428,983 per project.[5]

As we can see by Figure 4, Nord-Pas de Calais is the region that got most money, with a total of € 1,396,955,300 and Corsica the one that got the least money, with a total of € 149,788,110.[6]

Figure 4: Amount allocated per region

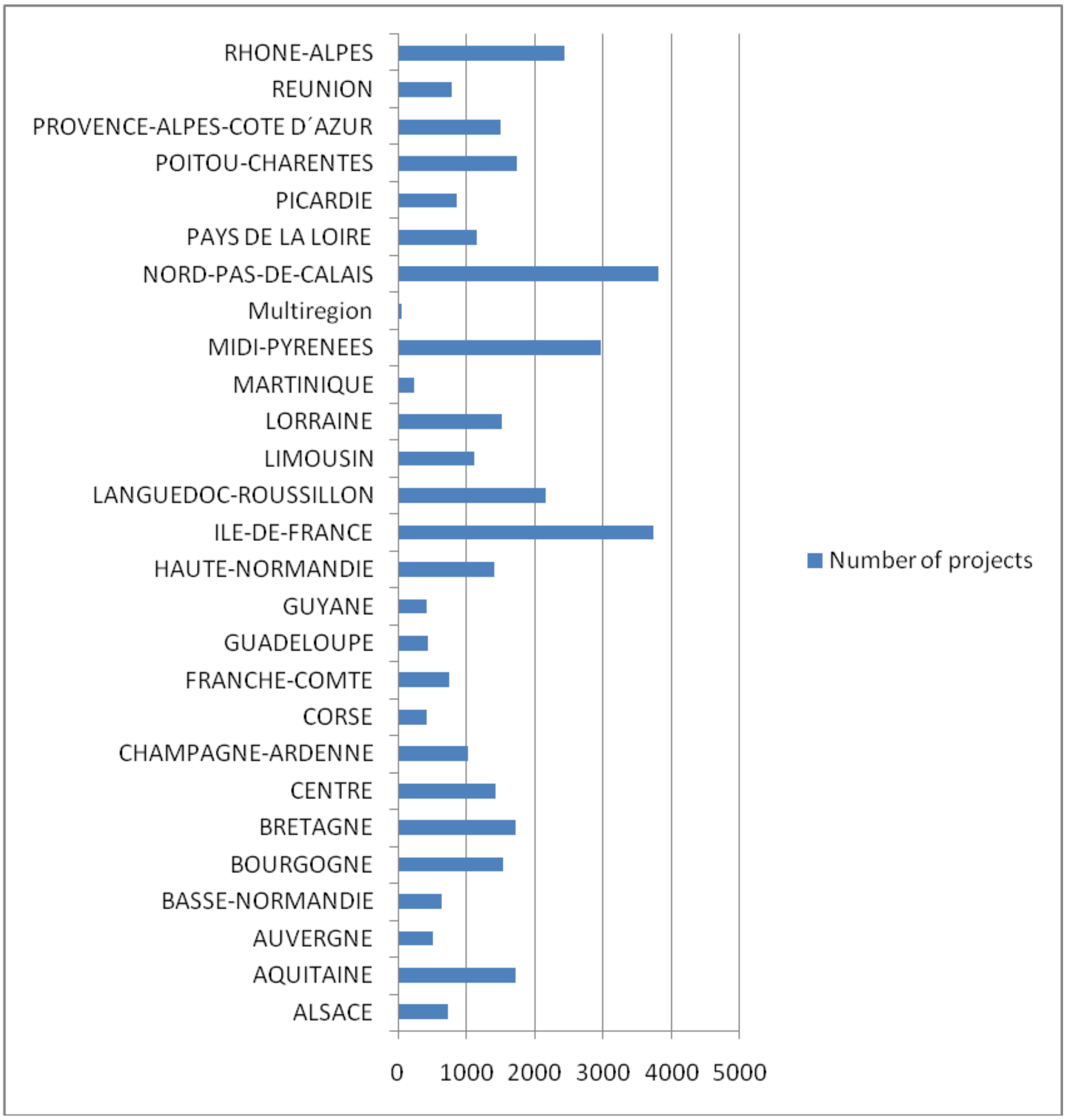


[5] Note that for 35 projects, neither a beneficiary region nor a funding source could be identified.

[6] Multiregion concerns money attributed to projects over several regions.

The number of projects per region also differs strongly, as can be seen by Figure 5.

Figure 5: Number of projects per region



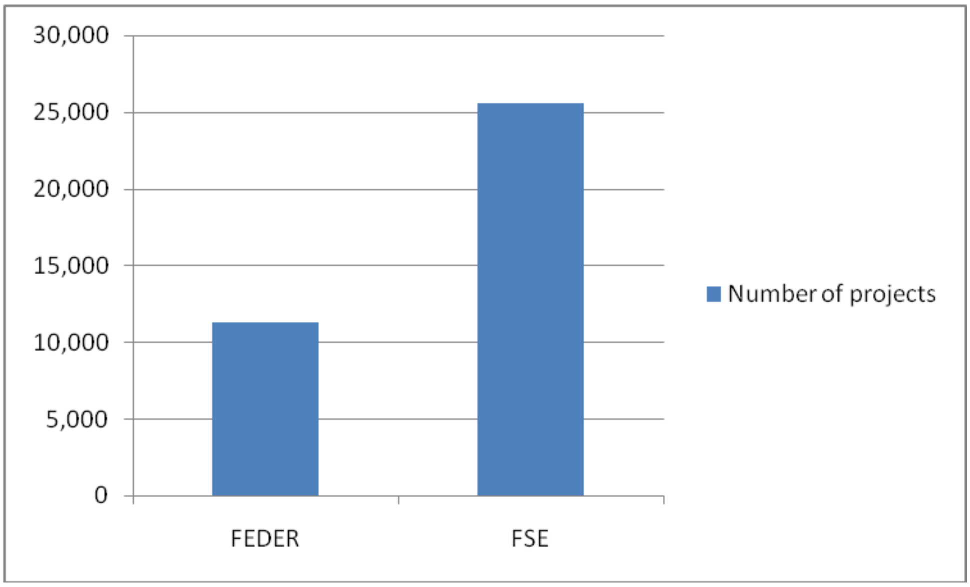
As we can see, most projects are done by Nord-Pas de Calais, with a total number of 3,814 projects and the region with the least number of porjects is Martinique, with a total number of 230 projects for the period under review.[7]

As is displayed be Figure 6, roughly 70% of all the supported projects have been financed by the European Social Fund (FSE).[8]

[7] Multiregion concerns projects covering several regions.

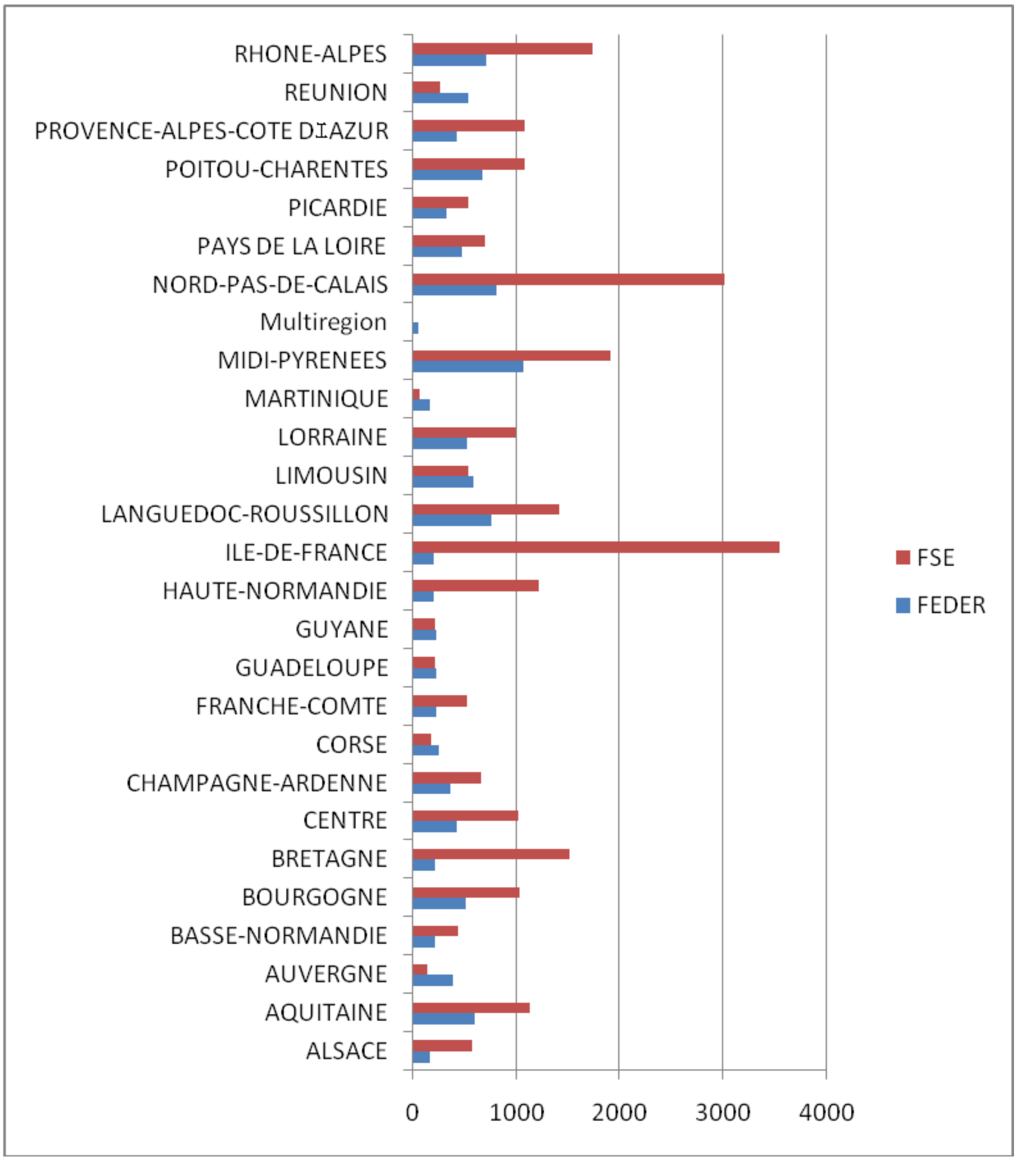
[8] FEDER stands for Fonds Européen de Développement Régional.

Figure 6: Number of supported projects by funding source



If we consider the funding source by regions, we see in Figure 7 that Ile-de-France gets the highest amount of financed projects from the FSE (with a total of 3,544 projects), and Midi-Pyrénée by the FEDER (with a total of 1,066 projects). The region with the fewsest supported projects is Alsace for the FEDER (161 projects) and Martinique for the FSE (64 projects).

Figure 7: Number of supported projects by funding source by region



4.2 Task 2: Econometric Application

In the French case, the Amadeus database contains 902,394 firms. Out of those 1,231 were identified as Cohesion Policy beneficiaries. 7,717 firms are dropped from the further analysis as they appeared as potential hit in the text field search of beneficiaries in the Amadeus database, but were not confirmed as hit after the manual checks. Note, however, that the vast majority of firms (890, 263) did never file any patent between 1978 and 2009.

The data structure is exactly the same as in the case of the Czech Republic. The pre-treatment phase refers to the years 1997 to 2003, and the treatment phase considered as 2007 and 2008. We cannot use 2009 here, as unlike in the Czech Republic's case, we restricted the patent search to the European Patent Office (EPO). The French number of firms turned out to the too large to be searched in global patent data. It turned out, however, that the EPO data is fairly incomplete for 2009 at the time of writing this report. Therefore, we restrict the analysis to the years 2007 and 2008 with respect to the time phase where the program has been active. The patent data that we collected for the firms included in the Amadeus database covers about 78% of all French patent activity at the EPO for the corresponding time period.

Analogously to the Czech Republic's data, we apply a DiD estimation using a Poisson fixed effects regression, and obtain the following coefficients for the TREAT (b1) and CONTROL (b2) .

b1 = -0.659 (0.290) **

b2 = -0.748 (0.127) ***

*** indicate a significance level of 1% and ** of 5%.

Here we also see that the overall patent activity declined in 2007/2008 (when compared to 1997 to 2003). Unlike in the Czech's Republic case, however, we do not find a significant difference between the two coefficients. We thus cannot conclude here that the Cohesion Policy programme resulted in more innovation as measured by EPO patents in the recipient firms.

In the case above, we again used a treatment indicator that is not fully accurate. There might be firms considered as treated firms in the period 2007/2008 although they only received a grant after this time period. Unfortunately the French recipient data does not include systematic information on the timing of the grants which makes a more meaningful evaluation impossible. We found out that in some project titles which are included in the recipient database, dates of funding were reported. We extract those dates from the titles' text fields and used that information to construct a more accurate treatment indicator that takes the timing of the grants into account. Unlike in the Czech case, this information is not reported systematically in the French data, however. We could only obtain information on the time of the grant for 319 firms out of the 1,231 that were identified as grant recipients in total. When using this more refined, but small treatment group the results did not improve. We still do not find a significantly positive treatment effect for the French data.

Thus we cannot conclude from our econometric analysis that the Cohesion Policy led to increased innovation activities in the recipient firms. However, this does not mean that there were no real effects in France. Instead, this inconclusive finding could be a result of the poor quality of the recipient data in France.

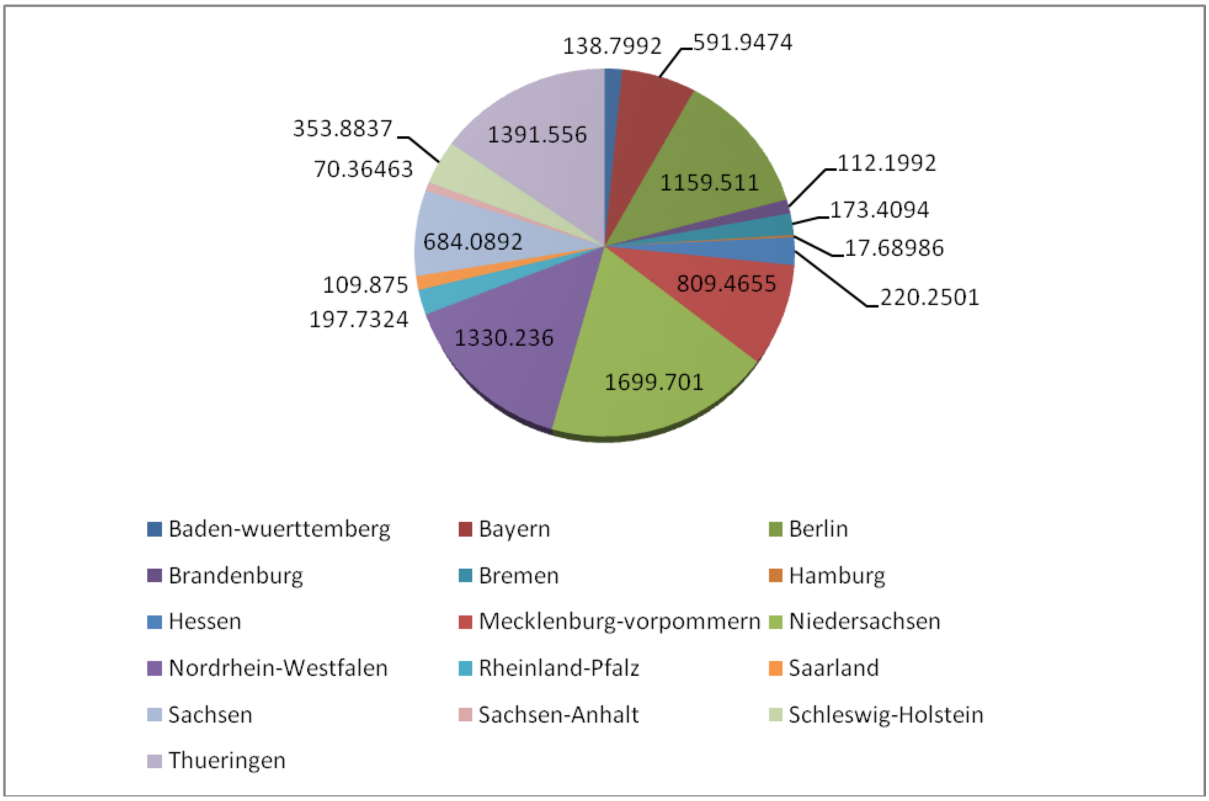
5 Cohesion Policy in Germany

5.1 Some descriptive statistics

Between 2005 and 2012, a total amount of € 9,060,653,000 has been spent on 47,616 projects in Germany.

As can be seen by Figure 8, the distribution of the attributed amounts varies greatly between the various German Laender, ranging from €1,699,701,000 in Sachen-Anhalt to €109,875,000 in Saarland.

Figure 8: Amount distribution by Bundesland (in € million)



In total, Eastern Germany received a total amount of € 3,067,674,000 for 33,201 projects (average amount per project of €92,400) and Western Germany € 5,993,036,000 for 14,415 projects (average amount per project of € 415,923).

Figure 9 displays the number of projects granted per year.[9]

[9] Even though it is not recognizable on the graph, in 2005, 4 projects were financed, in 2006, 13 projects and in 2010, 2 projects.

Figure 9: Number of projects granted per year

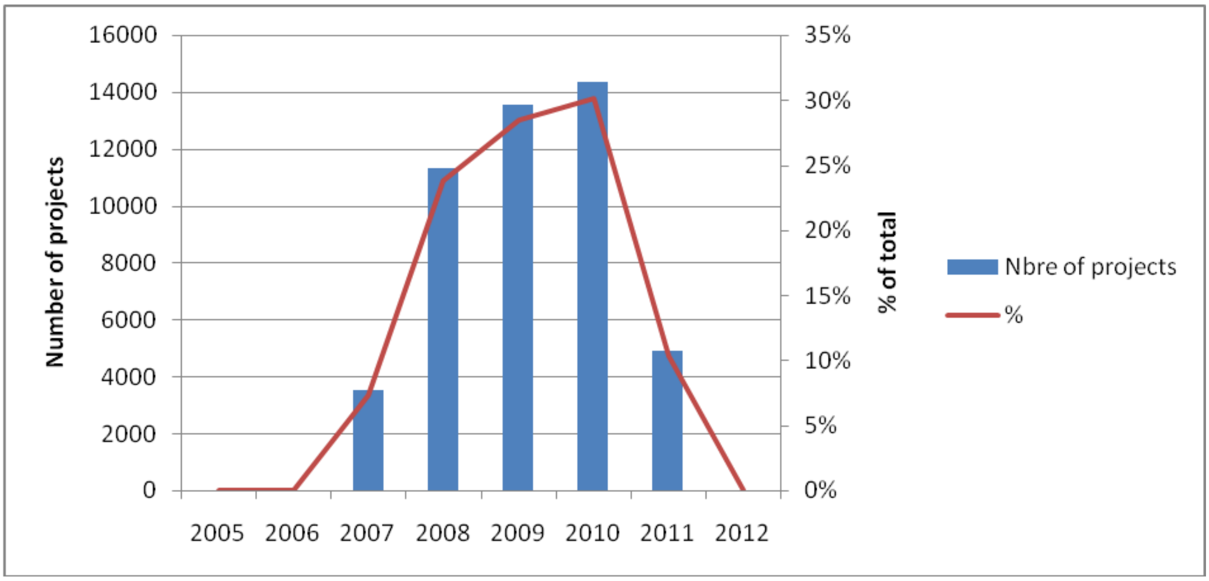
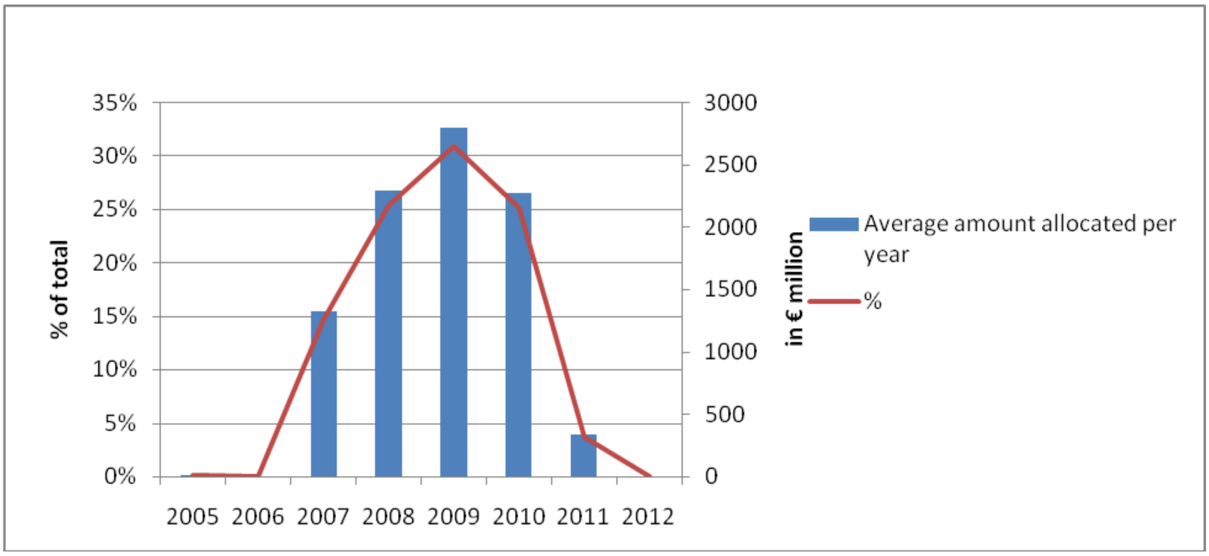


Figure 10 presents the average amount of money allocated per year.

Figure 10: Amount allocated per year



In the following sub-section, we will proceed to an econometric analysis using additional data retrieved from the German MIP, as agreed during the interim meeting.

5.2 Task 3: Econometric application using additional data sources

In addition to the Czech and French data, we consider Germany as a third country case study. This helps us to overcome some limitations of the data in the two countries considered above.

- In the recipient data, not all regions clearly indicate which grant was distributed within policy schemes that have the goal of fostering innovation in a region. Thus, it may occur that we consider a grant as a treatment towards innovation activity, but in fact this was not the main goal of the policy. Thus, we could underestimate the treatment effects, as we may not be able to distinguish accurately between innovation projects and others. In the German recipient data, we cannot identify either whether a grant was meant for innovation activity, but in the firm-level data we use, we can identify which firms did at all engage in innovation activities. Although still not ideal, we can at least exclude grant recipients from the analysis that did not at all conduct or even plan any innovation activities.

- Recipient firms may have used other sources of public funding in addition to Cohesion Policy. In this case, we might overestimate the treatment effects resulting from Cohesion Policy as the firms' budgets were actually supplemented by other public resources. In the German data, we have information for a subsample of firms on whether they have received other public subsidies for innovation projects. Consequently, we will investigate how the estimated treatment effects change if other public resources are taken into account.
- With the publicly available data in the French and Czech case study, we could only use patents as a proxy variable for innovation activity. As it is commonly known, patents are a narrow measure for innovation (see Griliches, 1990, for instance). Thus it would be desirable to investigate a region where other, more comprehensive innovation data could be used. The Community Innovation Surveys are a potential source, and the research team has access to non-anonymised micro data. Thus, we can link the Cohesion Policy data to the German part of the Community Innovation Survey, and can investigate a broader set of innovation variables, such as R&D investment and total innovation investment.

5.2.1 Linking the German recipient data to firm level information

For the German case, we do not use Amadeus and patent data like in the Czech and French case studies, but link the recipient data to the Mannheim Innovation Panel (MIP). The MIP is an annual innovation survey that has been carried out since 1993 by the Centre for European Economic Research (ZEW) in Mannheim, Germany. The MIP constitutes the German part of the Community Innovation Survey which is nowadays conducted every second year in all European Member States. In the MIP, each year about 5,000 companies report about their innovation behavior. This database has the advantage that we have more information about the recipient firms than in the Czech and French case. In particular, we can consider other measures than patents for the innovation efforts of recipient firms and a control group of non-recipient firms.

In total, there were 47,616 grants published on the Cohesion Policy websites of the German Länder. Out of those, 4,904 correspond to projects in the years 2011/2012 which cannot be used for any further analysis. Of the remaining projects, 5,606 could be assigned to firms that participated in the

Mannheim Innovation Panel at some point in time. These 5,606 were linked to 1,904 different recipient firms, i.e. program participants may have received multiple grants in the time period 2007-2010. Note that the MIP is conducted since 1993, but that we can only use the most recent four years for our analysis, the time when the Cohesion Program under review has been active (2007-2010). As it turns out several of the 1,904 firms that were identified participated in the MIP, but not in the period of interest for our study. Thus we cannot use all 1,904 firms for our analysis, but only the subgroup of those that answered to the survey between 2007 and 2010. After removing observations with missing values in our main variables of interest, we end up with a final sample of 623 'treated' observations. The number of observations that correspond to non-treated firms amounts to 21,226 firm-year observations.

5.2.2 The estimator used for the German analysis

As 50% of the firms are only observed once in our sample, we cannot apply panel data econometrics, like for instance the DiD estimator as we did in the cases of the Czech Republic and France. As a consequence, we can only apply estimators that can be applied to pooled cross-sections, including (parametric) selection models, instrumental variable estimators and matching techniques. In order to apply selection models and IV regressions, we would need a variable that determines the receipt of public funds, and that does not depend on outcome variables that are to be analyzed (here: innovation intensity, R&D intensity as well as R&D employment). Unfortunately our data does not contain variables that would be convincing candidates fulfilling these criteria. Therefore, we apply a matching estimator to the German data.

Matching estimators have been applied and discussed by many scholars, amongst which Angrist (1998), Dehejia and Wahba (1999), Heckman et al. (1997, 1998a, 1998b), and Lechner (1999, 2000). Generally, matching estimators are used to answer the question of what treated units with a given set of characteristics would have done if they would not have received the treatment. The objective is to compare the two outcomes – when receiving and when not receiving a treatment – for the same unit. The problem is of course that we can observe at most one of these outcomes because the observed unit has either received a treatment or not. Holland (1986) refers to this as the fundamental problem of causal inference. Hence, the counterfactual situation of a treated firm (i.e. an untreated firm) is not directly observable and has to be estimated.

Our fundamental evaluation question can be illustrated by an equation describing the average treatment effect on the treated individuals or firms, respectively:

$$E\left(\alpha_{TT}\right)=E\left(Y^{T}\mid S=1\right)-E\left(Y^{C}\mid S=1\right) \qquad (1)$$

where Y^T is the outcome variable. The status S refers to the group: S=1 is the treatment group and S=0 the non-treated firms. Y^C is the potential outcome which would have been realized if the treatment group (S=1) had not been treated. As previously explained, while $E(Y^T|S=1)$ is directly observable, it is not the case for the counterpart. $E(Y^C|S=1)$ has to be estimated. In the case of matching, this potential "untreated outcome" of treated firms is constructed from a control group of firms that did not receive innovation subsidies. The matching relies on the intuitively attractive idea to balance the sample of program participants and comparable non-participants. Remaining differences in the outcome variable between both groups are then attributed to the treatment.

Because of a potential selection bias due to the fact that the receipt of a subsidy is not randomly assigned, $E(Y^C|S=1) \neq E(Y^C|S=0)$ and the counterfactual situation cannot simply be estimated as average outcome of the non-participants. Rubin (1977) introduced the conditional independence assumption (CIA) to overcome this selection problem, that is, participation and potential outcome are statistically independent for individuals with the same set of exogenous characteristics X. Thus, the critical assumption using the matching approach is whether we can observe the crucial factors determining the entry into the programme. If this assumption is valid, it follows that

$$E\left(Y^C \mid S=1, X\right) = E\left(Y^C \mid S=0, X\right) \tag{2}$$

Provided that there no systematic differences in the observed characteristics between both groups, the treatment effect can be written as:

$$E\left(\alpha_{TT}\right) = E\left(Y^T \mid S=1, X=x\right) - E\left(Y^C \mid S=0, X=x\right) \tag{3}$$

In the present analysis, we conduct a nearest neighbour matching. More precisely, we pair each subsidy recipient with the single closest non-recipient. The pairs are chosen based on the similarity in the estimated probability of receiving such a subsidy, meaning the propensity score stemming from a probit estimation on the dummy indicating the receipt of subsidies S. Matching on the propensity score has the advantage not to run into the "curse of dimensionality" since we use only one single index as matching argument (see Rosenbaum and Rubin, 1983). In addition of matching on the propensity score, we also require the observations of firms in the selected control group to belong to the same year and to have a similar patent stock than the firms in the treatment group.

Last but not least, it is essential that there is enough overlap between the control and the treated group. We thus calculate the minimum and the maximum of the propensity scores of the potential control group, and delete observations on treated firms with probabilities larger than the maximum and smaller than the minimum in the potential control group.

The detail of our matching protocol is summarized in Table 1.

Table 1: The matching protocol

Step 1	Specify and estimate a probit model to obtain the propensity score $\hat{P}(X)$.
Step 2	Restrict the sample to common support: delete all observations on treated firms with probabilities larger than the maximum and smaller than the minimum in the potential control group. (This step is also performed for other covariates that are possibly used in addition to the propensity score as matching arguments.)
Step 3	Choose one observation from the subsample of treated firms and delete it from that pool.
Step 4	Calculate the Mahalanobis distance between this firm and all non-subsidized firms in order to find the most similar control observation. $MD_{ij} = \left(Z_j - Z_i\right)' \Omega^{-1} \left(Z_j - Z_i\right)$ where Ω is the empirical covariance matrix of the matching arguments based on the sample of potential controls. If only the propensity score is used, there is no need to calculate a multidimensional distance. In that case, e.g. a Euclidian distance is sufficient.
Step 5	In this application of the matching, we restrict the group of potential neighbors to firms active in the same industry as the particular treated firm. Select the observation with the minimum distance from the remaining sample. (Do not remove the selected controls from the pool of potential controls, so that it can be used again.)
Step 6	Repeat steps 3 to 5 for all observations on subsidized firms.
Step 7	Using the matched comparison group, the average effect on the treated can thus be calculated as the mean difference of the matched samples: $\hat{\alpha}_{TT} = \frac{1}{n^T}\left(\sum_i Y_i^T - \sum_i \widehat{Y_i^C}\right)$ with $\widehat{Y_i^C}$ being the counterfactual for *i* and n^T is the sample size (of treated firms).
Step 8	As we perform sampling with replacement to estimate the counterfactual situation, an ordinary *t*-statistic on mean differences is biased, because it does not take the appearance of repeated observations into account. Therefore, we have to correct the standard errors in order to draw conclusions on statistical inference. We follow Lechner (2001) and calculate his estimator for an asymptotic approximation of the standard errors.

5.2.3 Data and variable description

The data used for this more in-depth evaluation of the EU Cohesion Policy stem from the Mannheim Innovation Panel (MIP), which is the German part of the Community Innovation Survey (CIS).[10] More precisely, we use the data covering the years 2007 to 2010.

Our sample concerns innovative as well as non-innovative firms and covers manufacturing as well as business related services sectors.[11] In total, the sample consists of 21,849 observations, out of which 11,443 are innovative firms and 623 received a public R&D subsidy from the EU Cohesion Policy. Table 2and Table 3 here below show the industry structure as well as the firm size distribution of our sample.

Table 2: Size class distribution of the firms of the MIP sample

Size class distribution	Number of observations
Min. - 4	1,366
5 - 9	3,172
10 - 49	8,480
50 - 249	5,979
250 – max.	2,915
Total	**21,912**

[10] The CIS covers all of the EU Member States, Norway and Iceland using a largely harmonized questionnaire throughout participating countries. The CIS databases contain information on a cross-section of firms active in the manufacturing sector and in selected business services.

[11] According to the 3rd edition of the Oslo Manual – which is the definition followed by the CIS - an innovative firm is one that has implemented an innovation during the period under review. An innovation is defined as the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations (see Eurostat and OECD, 2005).

Table 3: Industry distribution of the firms composing the MIP sample

Industries	**Number of observations**
Food, beverages and tobacco	903
Textiles, clothing and leather	662
Paper, Wood, Publishing, Furniture	1,767
Chemicals and refining	749
Rubber, plastics, and non-metal mineral products	1,230
Metal	1,571
Machinery	1325
Office equipment and electronics	733
Communication technologies and precision instruments, optics	1,191
Vehicles	614
Other manufacturing industries incl. energy, water supply, construction	1,823
Trade	1,217
Transport	1,664
ICT and R&D services	1188
Other Business services	3,280
Other services incl. banking, insurance, renting	1,995
Total	**21,912**

The receipt of a subsidy form the Cohesion Policy is denoted by a dummy variable equal to one for firms that received public R&D funding and zero otherwise.

Outcome variables

As outcome variables, we consider the internal R&D investment, *RDINT*, being the ratio of internal R&D expenditures[12] to sales (multiplied by 100) as well as R&D employment, *RDEMP*, being the ratio of R&D employment to total employment (multiplied by 100). Further, we consider total innovation intensity (ratio of total innovation expenditure to sales multiplied by 100), investment intensity (investment/tangible assets * 100), and four dummy variables that indicate whether the firm has introduced at least one new product in (*PD*), one new process (*PC*), has abandoned one or more innovation projects (*PA*) or has ongoing innovation projects (*PN*).

Control variables

We use several control variables in our analysis likely to impact the fact of whether or not a firm applies and receives public support for its R&D activities. The number of employees takes into

[12] The CIS definition of R&D expenditure follows the Frascati Manual (OECD, 1993).

account possible size effects. As the firm size distribution is skewed, the variable enters in logarithms (*LOGEMP*).

In addition, we include a dummy variable capturing whether or not a firm is part of a group (*GP*), and if so, whether it has its headquarters on national or foreign territory (*FOREIGN*). Firms that belong to a group with the parent company on national territory might be more likely to receive subsidies because they presumably have better information about governmental programmes due to their network linkages. From a governmental point of view, the funding decision might be more positively assessed if a firm is part of a group because of potential incoming knowledge spillovers that can result from foreign branches. On the other hand, firms belonging to a group with a foreign parent company, might be more susceptible to file subsidy applications in their home country. Furthermore, as explained in the previous section, the Flemish government maintains a special policy instruments for small and medium-sized firms, being eligible only if they are not part of a group. If a small firm is majority-owned by a large parent company, it would no longer qualify for the SME-programs. The dummies *GP* and *FOREIGN* thus also control for this type of company profile, and an *a-priori* judgement of whether the effect is positive or negative is complicated because of the two opposing arguments outlined above.

We also control for the degree of international competition by including an export dummy in our analysis (*EXPO*). Firms that engage more heavily in foreign markets may be more innovative than others and, hence, more likely to apply for subsidies.

We further include the labour productivity as a covariate, measured as sales per employee, *LABPROD*, since high labour productivity is often an important determinant for receiving public funds.

Finally, a set of industry and time dummies control for unobserved heterogeneity across industries and time.

Descriptive statistics

Table 4 displays the descriptive statistics of the variables of our sample. As we can see, almost all the variables are significantly different between the treated and the non-treated firms.

Table 4: Descriptive statistics

	Non-subsidized firms **N=21,226**		**Subsidized firms** **N= 623**		**p-value of t-test on mean difference**
Variable	Mean	Std. Dev.	Mean	Std. Dev.	
			Covariates		
LOGEMP	3.633	1.664	3.993	1.239	p<0.001
LABPROD	0.245	2.347	0.146	0.215	p<0.001
EAST	0.313	0.464	0.868	0.338	p<0.001
GP	0.284	0.451	0.274	0.447	p=0.591
FOREIGN	0.064	0.245	0.061	0.24	p=0.748
EXPO	0.549	0.498	0.778	0.416	p<0.001
			Outcome variables		
RDINT	1.527	5.785	5.43	12.278	p<0.001
RDEMP[13]	3.608	10.584	11.957	19.154	p<0.001
INNOINT	3.358	9.119	9.216	15.917	p<0.001
INVINT[14]	37.72	93.863	51.554	124.60	p=0.047
PD	0.351	0.477	0.584	0.493	p<0.001
PC	0.294	0.456	0.435	0.496	p<0.001
PA	0.139	0.346	0.108	0.31	p=0.013
PN	0.397	0.489	0.632	0.483	p<0.001

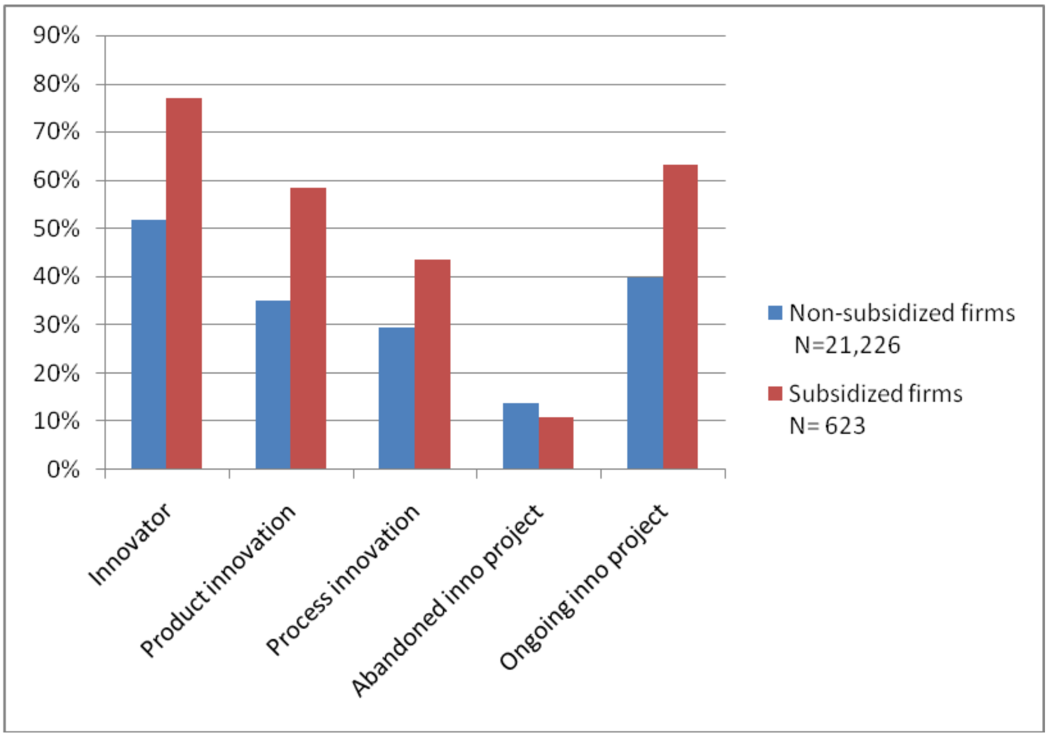
For instance, firms receiving subsidies from the Cohesion Policy are on average larger, more export oriented, belong to Eastern Germany more often and have a lower labour productivity. The descriptive statistics hence give us already a first indication that the selection criteria are correctly applied when choosing beneficiaries. While the beneficiary firms are more export oriented and larger than the non-beneficiaries, which are typical characteristics of R&D subsidy recipients, the Cohesion Policy recipients tend to belong to Eastern Germany more frequently and being less labour productive (hence in line with the Cohesion Policy selection criteria of brining depressed regions up to speed).

With regards to the outcome variables, funded firms have on average higher internal R&D intensity, more R&D employees, higher total innovation expenditures, more product and process innovations, more ongoing innovation projects and less abandoned innovation projects, and, as can be seen by Figure 11 here below, are overall more often innovators.

[13] Due to missing values, the number of observations is of 16,748 for the non-subsidized firms and of 488 for the subsidized firms for R&D employment.

[14] Due to missing values, the number of observations is of 9,291 for the non-subsidized firms and of 330 for the subsidized firms for innovation intensity.

Figure 11: Innovation behavior by subsidy status



The econometric analysis in the next section will reveal to which extent these differences can be attributed to the treatment.

5.2.4 Econometric results for the baseline specification

In order to apply the matching estimator as presented in the previous section, we first have to estimate a probit model so as to obtain the predicted probability of receiving support from the EU Cohesion Policy, to be employed as matching argument subsequently.

We can see by the probit estimation in Table 5, with the exception of the coefficients of the foreign parent and the labor productivity variables, all the other coefficients are significantly different from zero and hence are important in driving the selection of into the EU funding scheme.

Table 5: Probit estimation

	Coef.	Std. Err.
LOG EMPLOYMENT	0.140***	0.016
FOREIGN	-0.072	0.093
EXPO	0.383***	0.050
GP	-0.191***	0.056
OST	1.299***	0.050
LABOR_PRO	-0.059	0.077
CONS	-4.015***	0.165
Test on joint significance on industry dummies		χ^2 (15) = 175.71***
Test on joint significance on time dummies		χ^2 (3) = 46.22***
Number of observations		21,849

As explained in the previous section, a necessary condition for the validity of the matching estimator is common support. In our case, zero observations are lost because no common support could be found. Hence, the matching estimation could be done on our entire sample. The results are displayed in Table 6.

Table 6: Matching results

	Selected control group, N=623		**Subsidized firms, N=623**		**p-value on the t-test on mean difference**
Variables	Mean	Std.dev.	Mean	Std.dev.	
			Covariates		
LOGEMP	3.971	1.561	3.99	1.239	p=0.798
EAST	.870	.336	.868	.338	p=0.937
FOREIGN	.072	.259	.061	.239	p=0.457
EXPO	.775	.418	.778	.415	p=0.898
GP	.274	.447	.274	.447	p=1.000
LABPROD	.148	.228	.146	.215	p=0.853
			Outcome variables		
RDINT	3.442	7.882	5.430	12.278	p<0.001
INNOINT	6.903	13.585	9.215	15.917	p=0.009
RDEMP	9.177	16.275	11.957	19.154	p=0.028
PD	.495	.499	.584	.493	p=0.003
PC	.373	.493	.435	.496	p=0.039
PA	.167	.373	.107	.310	p=0.004
PN	.526	.501	.632	.482	p<0.001
INVINT	37.45	108.030	47.091	109.286	p=0.410

As shown by Table 6, all our covariates are well balanced after the matching. Hence, we can conclude that our matching was successful and that we found a neighbor for each treated firm. Note that the two samples are also balanced with respect to the industries and year (not shown in table).

The only variables where there still is a significant difference after the matching are our outcome variables. This difference could be attributed to the subsidy. With the exception of investment intensity, where no significant differences are found between treated and non-treated firms after the matching, all our other variables display significant differences after the matching. In other words, we can conclude that the null hypothesis of total crowding out can be rejected and that treated firms have higher R&D and innovation intensity, higher R&D employment intensity, are more likely to introduce new products and processes and are also more likely to have ongoing innovation projects and are less likely to have abandoned an innovation project than similar non-recipients, that is, the treated companies in the counterfactual situation of receiving no money from the Cohesion Policy.

5.2.5 Further results concerning only firms that show some innovation efforts

As the published data on grant recipients does not allow identifying those projects that were meant for innovation exactly, we attempt to address this problem to some extent by limiting the analysis to firms that show some innovation activities or at least plans to innovate. We use the information from the MIP in order to restrict the analysis to such firms. Now we re-run the analysis from above using only those firms that indicated that they introduced a new product or process, or have reported that they either abandoned an innovation project or have at least one ongoing innovation project.

For this estimation we use 480 innovating firms that received a grant from the Cohesion Policy, and 10,963 firms that did not receive such grants.

As we restrict our sample to innovators now, we can only investigate the continuous outcome variables meaningfully, that is, innovation intensity, R&D intensity, R&D employment intensity and investment intensity.

After matching the recipient firms to the controls as done above but only using the innovating companies, we confirm that the results reported above hold with respect to R&D intensity and innovation intensity. The estimated treatment effects are 2.1% and 3.7% respectively. These are both significant at the 1% level. With respect to the R&D employment intensity, we find a similar treatment effect as in the table above which amounts to 2.7%, but this is no longer significant, though. This might be due to the reduced sample size, though. The investment intensity is insignificant as reported in the table above.

5.2.6 Accounting for subsidies received from other sources

Now we consider the robustness test mentioned in the introduction to the German country case. We are interested to what extent the estimated treatment effects are confounded with the receipt of innovation subsidies from other sources than the Cohesion Policy.

Unfortunately, it actually turns out that the results reported above are confounded with other subsidy receipts. For this analysis, we can only use 2 out of the 4 years from the MIP, as only these years of the survey include question on other subsidies. In particular, we now consider a dummy variable indicating whether a firm has received public innovation funding from the German Federal Government (GOV). For this analysis we have 332 firms that received Cohesion Policy (CP) money and a control group of 10,620 firms that did not receive grants from the CP.

When we perform the analysis as conducted in Table 6 with the reduced sample (the years 2008 and 2010, as these include the subsidy variable), we find similar results as reported above. If we check however to what extent the firms have received other subsidies, it turns out that out of the 332 CP firms, 58% have also received money from the German Federal Government. This share, however, only amounts to 37% percent in the 332 firms that were picked as nearest neighbors during the matching procedure. Thus, the results above are subject to an omitted variable bias. Therefore, we take the other subsidy receipt into account by adding the receipt of other subsidies as a matching criterion. Now we repeat the analysis reported in Table 6 but only match CP firms that also have received subsidies from the German government to non-CP firms that have received subsidies from the German government. Accordingly, CP firms that did not receive other subsidies are only matched to firms that did not receive other funding either.

As we now use an additional matching criterion, that is, the receipt of other subsidies, but only have the sub-sample of 332 treated firms (the subsidy information is only available in two out of the four years), we now draw two nearest neighbor instead of a single control firm for each treated firm to avoid small sample problems. The quality of the matching estimator does not suffer from drawing two nearest neighbors instead of one, as the control group is still very rich in terms of potential candidates (more than 10,000 observations).

The matching results are presented in Table 7. Again we see that the two samples are well balanced with respect to the employed covariates. In addition, we still find that the treatment effects are positive and significant for the R&D intensity, innovation intensity (at the 10% level) as well as the intensity of investment with respect to physical assets. However, compared to the previous analysis where the other subsidies were not taken into account, the magnitudes of the estimated treatment effects reduce slightly. Furthermore, we no longer find a positive impact of Cohesion Policy on R&D employment intensity. This may be due to the fact that most of the Cohesion Policy grants are relatively smaller than the average grant of the Federal Government, so that Cohesion Policy grants alone cannot finance an entire new R&D workplace in a firm. This is however speculative to a certain extent and would require attention in future research.

In addition, we also see that the innovation indicator variables PD, PC, PA and PN do not differ anymore between the treated firms and their controls. This may be a data artifact as only innovating firms had to respond to the subsidy question in the survey affirmatively. Therefore, this results should not be causally interpreted in this particular setting.

Table 7: Matching results taking other subsidies into account

	Selected control group, N=664		**Subsidized firms, N=332**		**p-value on the t-test on mean difference**
Variables	Mean	Std.dev.	Mean	Std.dev.	
		Covariates			
LOGEMP	3.94	1.41	3.93	1.20	p=0.854
EAST	0.88	0.33	0.88	0.32	p=0.755
FOREIGN	0.05	0.23	0.06	0.23	p=0.859
EXPO	0.77	0.42	0.77	0.42	p=0.923
GP	0.26	0.44	0.27	0.45	p=0.646
LABPROD	0.15	0.18	0.15	0.17	p=0.599
		Outcome variables			
RDINT	4.41	9.69	6.20	13.80	p=0.045
INNOINT	7.34	14.06	9.54	16.70	p=0.057
RDEMP	11.05	17.72	12.48	19.49	p=0.388
PD	0.55	0.50	0.60	0.49	p=0.186
PC	0.39	0.49	0.43	0.50	p=0.321
PA	0.16	0.36	0.13	0.34	p=0.382
PN	0.54	0.50	0.62	0.49	p=0.028
INVINT	31.92	63.76	53.19	122.85	p=0.021

Note: The nearest neighbours are also matched on industry and year.

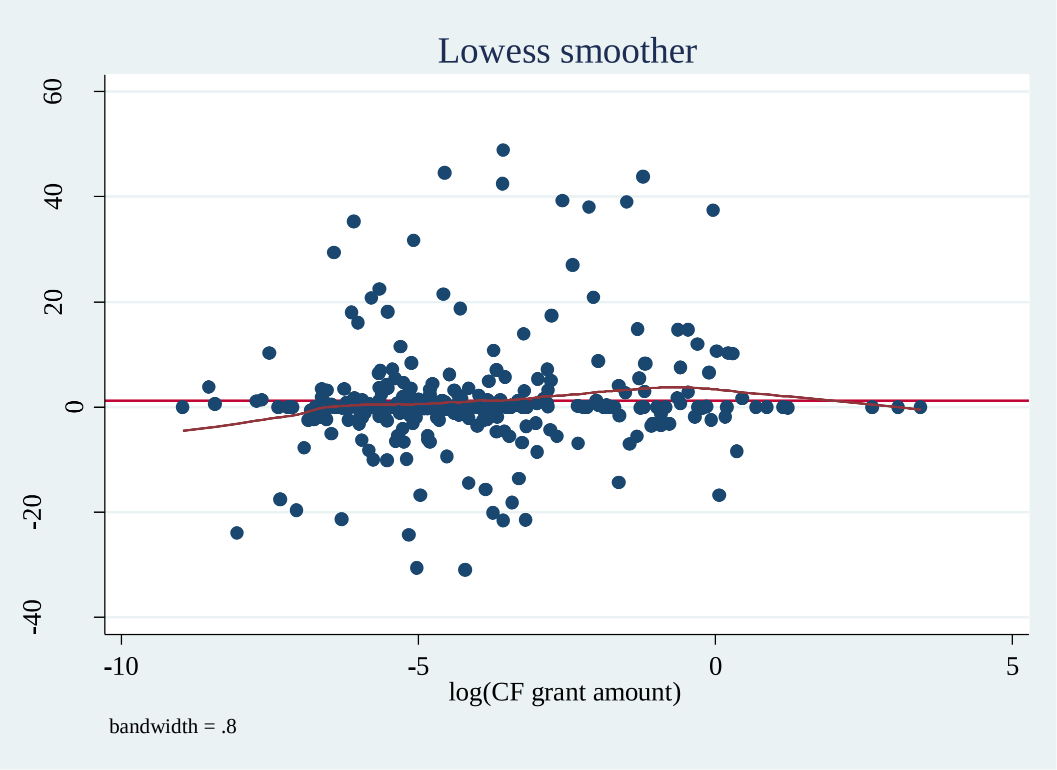
5.2.7 Taking the amount of the subsidy into account

Finally, we had a closer look at the individual grants that were distributed in Germany. Among those that went to firms, there is a large heterogeneity in types and sizes of grants. While several apparently address innovation projects, many grants are very small grants, for instance support to visit a trade fair. Only few are larger (or very large grants), for instance for setting up a new production facility.

Consequently, we investigated whether the identified treatment effect varies with the amount of funding the CP firms received. In order to do so, we take the estimated treatment effects from the previous subsection and regress those on the amount of funding.

Unfortunately, we do not find strong correlations between the estimated treatment effect and the amount of the grant. As an example, we illustrate a result in the figure below. On the vertical axis, we plot the average treatment effect on the treated, and on the horizontal axis is the size of the grant. The straight, flat horizontal line displays the estimated average treatment effect on the treated we got out of our matching routine as reported above. In this case, we use the two years of the survey where we can control for the other German Federal subsidies. If the grant size would matter, we would expect that the treatment effect increases as the grant size increases. As an illustration, we plot the result of a non-parametric regression of the treatment effect on the grant amount. We would expect that this curve has an upward slope, but unfortunately we do not find such pattern. Instead the estimated curve peaks before the end of the grant-size distribution and then decreases again.

Figure 12: Correlation between treatment effect and size of CP grants



If we, however, regard the few very large grants at the right tail of the distribution as outliers (the three observations on the far right end of the horizontal axis) and base our analysis on the remaining data, we find an upward sloping curve. This result is also confirmed when we use parametric regressions. For instance, for R&D intensity we find a positive slope in an OLS regression of the treatment effect on the log of grant size which is significant at the 10% level. Thus we find at least some evidence that the innovation impacts increase with the amount distributed to the firms, on average.

6 Conclusions

The goal of this project was to demonstrate to what extent researchers can conduct quantitative evaluations of the EU Cohesion Policy at the firm level using publicly available information that had been provided by the EU Member States. In particular, enterprise support for innovation and research has been the focus. We chose to conduct treatment effects estimations for grant recipients located in the Czech Republic, France and Germany.

It turns out that the current reporting standards of the Member States enable researchers to apply some basic treatment effects models after the recipient data has been supplemented with firm-level information on innovation activities and other characteristics. In addition a control group of non-funded firms has to be constructed. In our example, we used a combination of the recipient data with the Amadeus database which also allows drawing a control group and patent data that we collected from the PATSTAT database.

In chapter 2 of this report, we recommend several improvements concerning the reporting standard of the Member States. For instance, not only publishing the name of the grant recipients but also their location (e.g. the name of the city) would ease the process of identifying the treated entities in external databases such as Amadeus. In addition, it would be essential that the Member States publish the project start and end dates in the beneficiary databases. Our applications to the French and Czech cases clearly show that it is essential to determine the exact timing of treatment receipts.

Our policy conclusions can only be tentative as it would be desirable to have better quality information on the recipient data. For the Czech Republic and for France, we conducted difference-in-difference estimations on the patent application activity of firms. In the Czech case, we identify a positive treatment effect of Cohesion Policy. Although the overall patent activity of Czech firms is declining during the program period, we find that the reduction in patenting is significantly less for the grant recipient firms when compared to a control group of firms that did not receive funding with the Cohesion Policy program. For France, we cannot identify a positive treatment effect with the same methodology. This does not mean, however, that there is no impact of Cohesion Policy on firms' innovation activities in France. Instead we attribute this result to poorer data quality in the French case. As we point out in chapter 4 of the report, the French authorities did not report funding dates systematically. Thus, for many recipient firms, we cannot determine exactly when the grant has been received. This might bias our estimation of potential treatment effects.

For the German case, we linked the recipient data not to the Amadeus data base and patent data, but used the Mannheim Innovation Panel (MIP) which is the German part of the Community Innovation Survey (CIS). Using this database allows investigating other, broader measures of innovation instead of just patenting. Applying non-parametric matching methods, we do actually find positive treatment effects for a number of different variables in a first estimation when we do not control for other subsidies the firms have received. Firms benefitting from Cohesion Policy program grants are more likely to introduce product innovations, are less likely to abandon innovation projects, spend more on R&D and innovation in general when compared to the counterfactual situation where these firms would not have obtained such a grant. Once we account for subsidies received from the German Federal Government, the estimated positive effects decrease slightly in magnitude, but we still find positive impacts that are statistically significant. However, we can no longer identify a separate effect of Cohesion Policy on R&D employment. In a final setting, we analyze to what extend the size of the grant matters, and we find some support – although somewhat weaker in terms of statistical significance levels – for larger treatment effects, for instance concerning R&D investment, as the grant size increases.

This first exercise of a quantitative evaluation of the Cohesion Policy program grants to firms shows that applications of econometric treatment effects estimators are possible with the publicly reported recipient data by the Member States. However, improving and harmonizing the reporting standards of the recipient data is certainly desirable, as it would ease the systematic linking of this information to other external databases. In addition, it is advisable that future evaluations allow more time between the programme completion and the evaluation itself. As external databases are typically only available with a certain lag in time, data on variables of interest can at the earliest be obtained two years after the program grant. If output measure such as employment or sales growth should be considered, even longer time lags are necessary.

7 Appendix A: Data information per region of beneficiary country

	Country	Region	Format of the data	Variables available	Comments
1.	**Luxembourg**		Pdf	* Beneficiary institution * Project name * Amount	
2.	**Belgium**	Wallonia	Pdf	* Beneficiary institution * Portfolio * Project * Amount (total and detail)	
		Prov. Hainaut	Pdf	* Beneficiary institution * Portfolio * Project * Amount (total and detail)	
		Flanders	Excel	* Project name * Beneficiary institution * Project description * Amount (total and detail)	
		Bruxelles	Pdf	* Beneficiary institution * Project name * Year * Amount	
3.	**Germany**	Bayern	Pdf	* Beneficiary institution * Project name * Year * Amount	No categories per project type for most of the regions.

Saarland	Pdf	* Beneficiary institution * Project name * Year * Amount	
Rheinland-Pfalz			ERROR ON THE PAGE*
Baden-Württemberg		* Beneficiary institution * Project name * Year * Amount	
Hessen			ERROR ON THE PAGE*
Thüringen	Pdf	* Beneficiary institution * Project name * Year * Amount	
Sachsen	html	* Beneficiary institution * Project name * Year * Amount	Separate sites for ESF and EFRE Programmes.
Sachsen-Anhalt	Pdf	* Beneficiary institution * Project name * Year * Amount	Separate sites for ESF and EFRE Programmes.
Schleswig-Holstein	Pdf	* Beneficiary institution * Project name * Year * Amount	
Mecklenburg-Vorpommern			ERROR ON THE PAGE*
Hamburg	Pdf	* Beneficiary institution * Project name * Year * Amount (granted/paid)	

		Brandenburg	Pdf	* Beneficiary institution * Project name * Year * Amount (granted/paid)	
		Berlin			Website difficult to assess. No data information found.
		Niedersachsen	Pdf	* Beneficiary institution * Project name * Year * Amount (granted/paid)	Separate sites for ESF and EFRE Programmes /separate pdfs for Konvergenz& RWB.
		Bremen	Pdf	idem Niedersachsen	
		Lüneburg	Pdf	idem Niedersachsen	
4.	**Denmark**		html	* Project name * Start date *Description	
5.	**Spain**		Pdf	* Beneficiary institution * Project name * Year * Amount (granted/paid)	Harmonized for all regions. Same structure/content/format.
6.	**Italia**	Veneto	Pdf	* Beneficiary institution * Project name * Year * Amount	
		Lombardia	Pdf	* Beneficiary institution * Project name * Year * Amount	
		Piemonte	Excel	* Beneficiary institution * Project name * Year * Amount	

	Valle d'Aosta	Pdf	* Beneficiary institution * Project name * Year * Amount granted/paid	Axis and activity.
	Provincia autonoma di Bolzano - Alto Adige	Pdf	* Beneficiary institution * Project name * Year * Amount granted/paid	Axis and activity.
	Sardegna	Pdf	* Beneficiary institution * Project name * Year * Amount	
	Sicilia	Pdf	* Beneficiary institution * Project name * Year * Amount granted/paid	Axis and activity.
	Calabria	Pdf	* Beneficiary institution * Project name * Year * Amount	Only 2 projects.
	Basilicata		ERROR ON THE PAGE*	
	Puglia	Pdf	* Beneficiary institution * Project name * Amount	
	Campania	Pdf	* Beneficiary institution * Project name * Amount (total/paid)	Containsactivitycodes.
	Molise	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Containsactivitycodes.

	Abruzzo	Pdf	* Project name * Year * Amount	No list of beneficiaries available. The only beneficiary "the region of Abbruzzo.
	Lazio		ERROR ON THE PAGE*	
	Marche	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Activity code / beneficiary list available by activity.
	Umbria	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Activity code / beneficiary list available by activity.
	Toscana	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Activity code / beneficiary list available by activity.
	Emilia Romagna	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Activity code / beneficiary list available by activity.
	FriuliVeneziaGiulia			Website difficult to assess. No data information found.
	Trento	Excel	* Beneficiary institution * Project name * Year * Amount (detail)	
	Liguria			Website difficult to assess. No data information found.
7. Portugal	Centro	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Classification per project, per activity, per project nomber.

		Norte			Website difficult to assess. No data information found.
		Algarve	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	
		Lisboa			ERROR ON THE PAGE*
		Alentejo	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Classification per project, per activity, per project nomber.
		Açoras	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	
		Madeira			ERROR ON THE PAGE*
8.	**Malta**		Pdf	* Beneficiary institution * Competent ministry * Project description * Project name * Year * Amount (detail)	
9.	**Nederland**	Noord	html	* Beneficiary institution * Project name * Year * Amount	
		Oost	html	* Beneficiary institution * Project name * Year * Amount	
		Zuid			ERROR ON THE PAGE*

	West	html	* Beneficiary institution * Project name * Year * Amount	
10. Ireland	Border, Midland, Western	html	* Beneficiary * Project name * Amount	Years per file.
	Southern/Eastern		ERROR ON THE PAGE*	
11. France		pdf html excel	* Beneficiary institution * Project description * Project name * Year * Amount (detail)	
12. UK	North East		ERROR ON THE PAGE*	
	West North	html	* Beneficiary institution * Project name * Amount (detail)	
	York Shire and the Humber	Pdf	* Beneficiary institution * Project name * Amount (detail)	
	East Midlands	Pdf	* Priority * Fin start and end date * Sponsor * Project title * Project description *Amount	
	West Midlands	pdf	* Project name * Applicant * date * amount	
	Cornwall And The Isles Of Scilly	pdf	* Beneficiary institution * Project name * Amount (detail)	

	East of England	pdf	* Beneficiary institution * Project name * Amount (detail)	
	London	Excel	* Name of the beneficiary organization * Name of the project * Geographical coverage * Description (in project's own words) * Amount (detail)	
	South East	html		Link to different website: http://www.seeda.co.uk/what-we-do/european-investment/erdf/existing-erdf-projects.
	South West	pdf	* Beneficiary institution * Project name * Amount (detail)	Separate pdf file per activity.
	West wales and the valleys	pdf	* Priority * Sponsor * Project title * Project description *Amount	
	East Wales	pdf	* Priority * Sponsor * Project title * Project description *Amount	
	Scotland (Highlands and Island)	html	* Beneficiary institution * Project name * Amount	
	Nothern Ireland	html	Searchengine	

		Gibraltar	html	* Beneficiary institution * Project name * Amount	
13.	**Bulgaria**	Severozapaden	pdf	* Beneficiary institution * Project name * Amount	Separate pdf file per project type: only data on environment OP available in English.**In Bulgarian, more data seems to beavailable**.
		Severentsentralen	pdf	* Beneficiary institution * Project name * Amount	Separate pdf file per project type: only data on environment OP available in English. **In Bulgarian, more data seems to beavailable**.
		Severoitztochen	pdf	* Beneficiary institution * Project name * Amount	Separate pdf file per project type: only data on environment OP available in English. **In Bulgarian, more data seems to beavailable.**
		Yugoiztochen	pdf	* Beneficiary institution * Project name * Amount	Separate pdf file per project type: only data on environment OP available in English. **In Bulgarian, more data seems to beavailable.**
		Yugozapaden	pdf	* Beneficiary institution* Project name* Amount	Separate pdf file per project type: only data on environment OP available in English. **In Bulgarian, more data seems to beavailable.**
		Yuzhuntsentralen	pdf	* Beneficiary institution * Project name * Amount	Separate pdf file per project type: only data on environment OP available in English.**In Bulgarian, more data seems to beavailable.**
14.	**Estonia**		html		**Information only available in Estonian**
15.	**Sweden**		html		Same structure for all regions. Search engine ESF Council website : http://www.esf.se/sv/Rotsida-for-topmeny/In-english/.

16. Austria	Burgenland	pdf	* Beneficiary institution * Project name * Amount	
	Niederösterreich	pdf	* Beneficiary institution * Project name * Amount	Only information found on regional competitiveness and employment projects.
	Wien	pdf	* Beneficiary institution * Project name * Amount	Only information found on regional competitiveness and employment projects.
	Kärnten			ERROR ON THE PAGE*
	Steiermark	pdf	* Beneficiary institution * Project name * Amount	
	Oberösterreich	pdf	* Beneficiary institution * Project name * Amount	
	Salzburg	pdf	* Beneficiary institution * Project name * Amount	
	Tirol	pdf	* Beneficiary institution * Project name * Amount	
	Vorarlberg	pdf	* Beneficiary institution * Project name * Amount	
17. Cyprus		pdf		Links empty or not working in English. **They seem to work in Greek and/or Cypriots.**
18. Lithuania				Website difficult to assess. No data information found.
19. Lavia		excel		**No data could be found in English.**
20. Hungry	same site for all regions			**No data could be found in English.**

21. CzechRepublic	Praha	excel		
	StredniCechy	pdf	* Beneficiary institution * Project name * Amount	**Detailed information available only in Czech.** The titles are also available in English.
	Jihozapad	pdf	* Beneficiary institution * Project name * Amount	**Detailed information available only in Czech.** The titles are also available in English.
	Severozapad	Excel	* Beneficiary institution * Project name * Amount	**Detailed information available only in Czech.** The titles are also available in English.
	Severovychod		ERROR ON THE PAGE	
	Jihovychod	Excel Pdf	* Beneficiary institution * Project name * Amount	**Detailed information available only in Czech**. The titles are also available in English.
	Stredni Morava	word	* Beneficiary institution * Project name * Amount	**Detailed information available only in Czech.** The titles are also available in English.
	Moravskoslezsko			Website difficult to access. No data information found.
22. Poland	Lodskie	RAR	* Beneficiary institution * Project name * Amount	Classified per activity.
	Mazowieckie	RAR Excel	* Beneficiary institution * Project name * Amount	Classified per activity.
	Malopolskie	RAR Excel	same for all regions	
23. Romania	Nord-Vest	searchengine		Same structure for all regions.
24. Slovenia		html	* Beneficiary institution * Project name	

			* Amount	
25. Slovakia	Bratislavsky	excel pdf	* Beneficiary institution * Project name * Amount	**Information only available in Slovak.** Same for all regions.
26. Finland	all regions	html	* Beneficiary institution * Project name * Amount	Same structure for all the regions.
27. Greece	all regions	pdf	* Beneficiary institution * Project name * Amount	**Information only found in Greek.**

* “ERROR ON THE PAGE” indicates that the region’s website did not work at the time when we tried to access it.

8 Appendix B: Example of a data reporting structure

Format: excel

Reportinglanguage: English

						European Union funding							
Name of beneficiary, complete address and legal form	**ID**	**Operation**	**Operational Programme**	**Fund EU**	**Duration of the projects in months (including exact start and end date)**	**Date of allocation**	**Amountsallocated (in €)**	**Date of interim payment (in €)**	**Total Amounts paid from the start of the Project (in €)**	**Total amount spent at the end of year 1 (in €)**	**Total amount spent at the end of year 2 (in €)**	**Total amount spent at the end of year 3 (in €)**	**Status**

Private firm beneficiaries

Purpose of the grant: R&D and innovation

Beneficiary 1

Beneficiary 2

Beneficiary 3

…

Purpose of the grant: Infrastructure

Beneficiary 1

Beneficiary 2

Beneficiary 3

…

Purpose of the grant: Environment

Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Etc.**
National/regional/local authorities
Purpose of the grant: R&D and innovation Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Purpose of the grant: Infrastructure** Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Purpose of the grant: Environment** Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Etc.**

Universities and research centers
Purpose of the grant: R&D and innovation Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Purpose of the grant: Infrastructure** Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Purpose of the grant: Environment** Beneficiary 1 Beneficiary 2 Beneficiary 3 … **Etc.**

9 Appendix C: Linking databases via text field searches

The following text describes the matching of the beneficiaries of the Cohesion Policy programme of the European Union with the company data of the Amadeus database provided by Bureau van Dijk.

One major drawback encountered during this exercise was linked to missing information of the programme beneficiaries. Because the managing authorities are only obliged to report the names of the beneficiaries but not their entire address, the only additional information about the locations of the beneficiaries can be implicitly determined by the NUTS-3 code. Because the NUTS-3 code is not represented in the Amadeus database, it is necessary to append this information using the postal code and/or the city name of the beneficiary. For some countries, i.e. France, this can easily be achieved by using the postal code per NUTS 3 unit, because the department is a hierarchical part of it. For other countries, like Germany, the connection between postal codes and the NUTS-3 regions is much more ambiguous, resulting in multiple NUTS-3 regions per postal code and vice versa.

The inclusion of additional information besides the beneficiary name would thus be important to reduce false positive matches. The latter are usually generated by heuristically matching algorithms in combination with weak specifications (beneficiary name only versus full address). Having the full address of each beneficiary per region would thus help to avoid this kind of error.

The regional beneficiary data is matched to Amadeus at the country level, that is, we perform separate searches for all selected countries. The programme that is used for the matching is called SearchEngine (developed by Thorsten Doherr for the Centre of European Economic Research). In a first step, this tool creates a kind of dictionary, containing all the words of company names of the country specified in the Amadeus data, along with an occurrence counter for every entry and supporting tables that link the words back to the original records of the Amadeus table (companies). This counter is the base for the heuristics of the matching algorithm. Secondly, each beneficiary name will be separated into words. Each word is associated with the occurrence counter of the corresponding dictionary entry or with the average occurrence if no entry was found. The occurrence reflects the identification potential of the word. A relatively low occurrence has a high potential to identify the company, because the resulting list of potential hits is small. Legal forms as part of a companies' names have for instance a very high occurrence and because of this reason a very large list of potential hits. Each list represents a percentage of the whole identity of a beneficiary name. This percentage is inversely proportional to the size of the list. By combining all potential hit lists, starting with the smallest, it is possible to rank the combined hits by the summarized percentage of the lists they are part of.

The main advantage of this algorithm is its potential to ignore non-identifying filler words, changes of the legal status and different positions of the words. The algorithm automatically adjusts itself to the specific combination of words of a beneficiary name. If a name consists of a few words with a high identification potential among some words with a low potential, the algorithm ignores the lesser important words. If all words are equally important, the algorithm is more restrictive and ignoring words is not an option.

Because the algorithm is word based, it is very prone to typing errors that can transform a common word into a very unique word or into a word that is not represented by the dictionary. This problem can be circumvented by introducing n-grams to the algorithm. An n-gram is a part of a word with the length n. The complete representation of a word with the length L consists of max(L-n,0)+1 n-grams. For example, the name "Thorsten" is represented by five 4-grams: Thor, hors, orst, rste, sten. The dictionary of the SearchEngine then contains n-grams instead of complete words. The impact of typing errors on an n-gram based dictionary is much smaller than on a word based dictionary because one will only loose a couple of chunks of a word but not the entire word. Because this more tolerant approach can result in too many false positive hits, it should always be done after the normal matching and only for beneficiaries without a match.

The main problem of heuristic matches is false positives, resulting from matches with a relative high identification. There is a trade-off between getting the most correct hits and getting too many false hits. It is like searching the internet using Google. Only the user knows the context of the search and sometimes the first result page has no matching entry. This analogy leads to the conclusion that manual control of the matches is required to distinguish between correct and false positives. Additional external information, like the NUTS-3 regions of searched beneficiaries and found companies, can further reduce the false positives to minimize the manual effort.

10 Appendix D: An overview on treatment effects estimators

In this section econometric models that tackle the problem of endogeneity of the treatment in the evaluation of public grants are discussed.[15] As treatment effects models usually consider discrete treatments we start with such methodologies, and briefly mention possible extensions for multiple or continuous treatments afterwards.

10.1 Discrete Treatments

In this subsection, we focus on methods that are applicable to cross-sectional data, and second those that require panel data. Models allow estimating different kinds of treatment effects: the average treatment effect, the local average treatment effect, the marginal treatment effect, the average treatment effect on the treated and the treatment effect on the untreated (see e.g. Heckman et al., 2001, for a discussion of treatment effects commonly used in programme evaluation literature). Here, we focus on the treatment effect on the treated (*TT*). Suppose we consider subsidies for R&D activities. Thus our basic evaluation question would be: "How much would a firm that has received a subsidy have spent on R&D activities if it would not have been subsidized, on average?", or expressed as equation:[16]

$$\alpha_{TT} = E\left(Y^T \middle| S=1\right) - E\left(Y^C \middle| S=1\right), \tag{3}$$

whereY^T refers to the potential outcome (e.g. R&D expenditure) of firms that receive subsidies, and Y^C to the situation where they do not. *S* indicated the treatment status. It is equal to 1 for treated firms and zero otherwise. Thus, the *TT* results from comparing the actual outcome of subsidized firms with their outcome in case of not receiving a grant. The approach of measuring potential outcomes goes back to Roy (1951). The outcome $E\left(Y^T \middle| S=1\right)$ can be estimated by the sample mean of *Y* in the group of subsidized firms. In order to identify $E\left(Y^C \middle| S=1\right)$ one needs to make further assumptions. The latter cannot simply be calculated from non-subsidized firms as

$$E\left(Y^C \middle| S=1\right) \neq E\left(Y^C \middle| S=0\right) \tag{4}$$

due to non-random assigned treatments. This would only be valid in an experimental setting where subsidies are granted randomly to firms, which is obviously not the case in current innovation policy practice.

[15] This section draws heavily from the surveys of Heckman et al. (1999) and Blundell and Costa-Dias (2000, 2002), and Imbens and Wooldridge (2009).

[16] All variable are measured at the firm level *i* (with *i* = 1,...,*N*), but we omit the index *i* for convenience.

Suppose the outcome equation has following form

$$\begin{array}{lll} Y = X\beta + S\alpha + U & \text{if} & S = 1 \\ Y = X\beta + U & \text{if} & S = 0 \end{array} \qquad (5)$$

where X denotes a set of exogenous variables, βtheir parameters and α is the impact of the treatment. *U* is the error term with mean zero and is assumed to be uncorrelated with *X*.

Since *S* is not randomly assigned - as this is most likely not the case when subsidies are the subject of the analysis - *U* will be correlated with *S*. This happens because the grant decision is expected to be related to firm characteristics that may well affect *Y* as well. If this is the case, and one is unable to control for all the characteristics affecting Y and S simultaneously, some correlation between *S* and *U* is expected. Therefore, standard econometric approaches that regress *Y* on *X* and *S* are not valid.

In order to solve this problem, one assumes that the subsidy receipt can be written as

$$S^* = Z\gamma + V\,, \qquad (6)$$

where *D** is an index depending on a set of variables *Z* and parameters γ, as well as an error term *V*. The receipt of a subsidy happens when *D** is larger than zero:

$$S = \begin{cases} 1 \ \text{ if } \ S^* > 0 \\ 0 \ \text{ otherwise} \end{cases} . \qquad (7)$$

In the following we refer to this as selection equation.

10.1.1 The Heckman Selection Estimator

The application of the Heckman estimator requires the existence of one regressor that is not included in the outcome equation, but that has a non-zero coefficient in the selection equation, and is independent of *V*. Moreover, the joint distribution of *U* and *V* either has to be known or one has to able to estimate it. This estimator directly controls for the part of the error term *U* that is correlated with *S*. Typically, scholars assume that *U* and *V* follow a joint normal distribution, which leads to the conditional outcome equation:

$$\begin{aligned} E\left(Y \mid S = 1\right) &= X\beta + \alpha + \rho\phi\left(\frac{Z\gamma}{\sigma_V}\right)\Phi\left(\frac{Z\gamma}{\sigma_V}\right)^{-1} \\ E\left(Y \mid S = 0\right) &= X\beta - \rho\phi\left(\frac{Z\gamma}{\sigma_V}\right)\left[1 - \Phi\left(\frac{Z\gamma}{\sigma_V}\right)\right]^{-1} \end{aligned} \qquad (8)$$

where the last term in each equation represents the error term conditional on *S*. This separates the true impact of *S* from the selection process, which accounts for differences among funded and non-funded firms. The TT can be obtained by regressing S* on Z, and running a least squares estimation on equation (6).

Note that one would assume that the parameters of *X* are the same for subsidized and non-subsidized firms in this case. One can easily relax that assumption: then we would omit *S* in eq. (8) and estimate the two equations separately with least squares. In order to obtain TT, we calculate

$$\alpha = X\left(\beta_1 - \beta_0\right) + \left(\rho_1 - \rho_0\right)\phi\left(\frac{Z\gamma}{\sigma_V}\right)\Phi\left(\frac{Z\gamma}{\sigma_V}\right)^{-1} \qquad (9)$$

where subscript 1 refers to the parameters of the treated group's equation, and subscript 0 to the non-treated (see e.g. Heckman et al., 2003).

This model has often been criticized as it is quite demanding on assumptions about the structure of the model. Several generalizations of the fully parametric model have been suggested in the literature. Among others, semiparametric variants of the Heckman model include Gallant and Nychka (1987), Cosslett (1991), Newey (1999), or Robinson's (1988) partial linear model. Note, however, that in such models the intercept in the outcome equation is no longer identified. A precise estimate of the intercept is required for deriving TT, though. Heckman (1990) and Andrews and Schafgans (1998) developed estimators for the identification of TT.

10.1.2 Instrumental variable regressions (IV)

In contrast to the Heckman model, the IV regression does not involve estimating a selection equation. Suppose *Z** is a valid instrument, i.e. it is (highly) correlated with the treatment dummy *S*, we can find a transformation, g, such that *g(Z*)* is uncorrelated with *U* conditional on *X*, and *Z** is not completely determined by *X*. This amounts to standard instrumental variable regression. [17]

Although this is a very simple estimator as it does not require estimating the selection equation, it has a major drawback: it is not easy to think about a variable that could serve as a valid instrument. Recall that it should, for instance, determine the subsidy receipt but not R&D, i.e. a simultaneous requirement of "participation determination" and "non-influence on the outcome of participation". As there are usually no straightforward candidates for instrumental variables available, a convincing application of this estimator is rare. Even if longitudinal data are available, the common practice to use lagged values does not necessarily solve the problem as lags are often highly correlated with future values of the variable.

[17] Alternatively one could, of course, estimate a simultaneous equation model with 2SLS or 3SLS for example.

10.1.3 Matching estimators

The matching estimator is a non-parametric method and has one main advantage: no particular functional form of equations has to be specified. The disadvantages are strong assumptions and heavy data requirements.

The main purpose of the matching estimator is to re-establish the conditions of an experiment. The matching estimator attempts to construct a correct sample counterpart for the treated firms' outcomes if they had not been treated by pairing each treated firm with members of a comparison group. Under the matching assumption, the only remaining difference between the two groups is the actual subsidy receipt.

Rubin (1977) introduced the so-called conditional independence assumption (CIA) to solve the problem arising in eq.(4). This condition means that the receipt of subsidies and potential outcome are independent for firms with the same set of exogenous characteristics

$$Y^T, Y^C \perp S \mid X = x. \tag{10}$$

The condition helps to overcome the problem that $E\left(Y^C \mid S = 1\right)$ is unobservable. If the conditional independence assumption is valid, then $E\left(Y^C \mid S = 0, X = x\right)$ can be used as a measure of potential outcome for the subsidy recipients. However, the CIA is only fulfilled if all variables that influence the outcome and selection status *S* are known and available in the dataset. In that case the equation

$$E\left(Y^C \mid S = 1, X = x\right) = E\left(Y^C \mid S = 0, X = x\right) \tag{11}$$

holds, and the average outcome of subsidized firms in the absence of a subsidy can be calculated from a sample of comparable ("matched") firms. Note, however, that matching requires a further assumption, which is $0 < \Pr\left(S = 1 \mid X\right) < 1$ in order to guarantee that all treated firms have a counterpart in the non-treated population, and that every firm constitutes a possible subsidy recipient. However, this does not ensure that this happens in every sample. Thus, matching requires a common support restriction. If the samples of treated and non-treated firms would have no or only little overlap in *X*, matching is not applicable to obtain consistent estimates.

If the CIA holds and common support is given, the treatment effect on the treated would consequently amount to

$$\alpha = E\left(Y^T \mid S = 1, X = x\right) - E\left(Y^C \mid S = 0, X = x\right) \tag{12}$$

which can be estimated using the sample means of both groups.

Usually *X* contains a large number of variables, so that matching can be very difficult due to the high dimensionality of *X*. Rosenbaum and Rubin (1983, 1984) have shown that conditioning the matching on the propensity score (the probability to receive a subsidy) Pr(X) instead of *X* is a valid procedure. This reduces the curse of dimensionality, and makes matching feasible as one can use a single index. Lechner (1998) suggested a hybrid matching, that is, one conditions on Pr(*X*) and a subset of *X*; for example, industry dummies if one wants to ensure that a matched control observation is in the same industry as the treated firm.

The comparison group for each treated firm is chosen to a predefined criterion of proximity. Having defined the neighborhood for each treated firm, the next issue is the choice of appropriate weights for non-treated observations within the neighborhood, such that TT is obtained as

$$\hat{\alpha} = \sum_{i \in T} \left(Y_i - \sum_{j \in C} w_{ij} Y_j \right). \tag{13}$$

Common procedures are nearest neighbor matching, that is, the weight is set to unit value for the closest match, and zero otherwise. So, one ends up with one single non-subsidized twin firm for each treated one. If one picks more than one neighbor, one could, for instance, set the weights to equal value for each control observation. Kernel matching uses all firms in the control group for each treated firm, and assigns kernel weights according to proximity in *X* or Pr(*X*) to each control observation.

10.1.4 Difference-in-difference estimators

The difference-in-difference (DiD) estimator uses the idea that a good guess for the outcome in the absence of a treatment, would be an observation of a treated firm in an earlier period where it did not receive a subsidy. In order to control for macroeconomic changes over time, DiD relates the development of treated firms over time to a control group of non-treated firms to eliminate effects that are due to changes over time. Thus, the DiD estimator compares subsidized firms and a control group of non-subsidized firms before (t_0) and after (t_1) the treatment:

$$\alpha_{TT}^{DiD} = \Big(E\big(Y_{t_1} \mid S=1\big) - E\big(Y_{t_0} \mid S=1\big) \Big) - \Big(E\big(Y_{t_1} \mid S=0\big) - E\big(Y_{t_0} \mid S=0\big) \Big) \tag{14}$$

The obvious disadvantage of this estimator is that panel data are required. For studies on R&D subsidies, this actually amounts to a heavy data requirement, as not only two periods have to be available at least, but in particular observations in the case of subsidy receipts and observations on previous periods where the same firm did not receive a subsidy. As subsidies are often longer term research projects, and firms get multiple grants over time, it actually turns out to be difficult to construct a database that allows an appropriate application of DiD in practice.

One underlying assumption in the DiD estimator is that treated and non-treated firms react similar to shocks that occur over time (aside of the treatment). However, as evidence shows treated and non-treated firms are often very different in characteristics, which would suggest that they may also react differently to macroeconomic shocks. In order to overcome this potential bias the conditional difference-in-difference estimator (CDiD) can be applied. It is a combination of matching and DiD. There one does not employ a general control group, but matches comparable firms to the treated firms in the period before receiving the treatment, and compares the evolution of two comparable groups over time. Blundell and Costa Dias (2000) suggest employing CDiD for repeated cross-sections if no panel data is available. This requires matching three times: find the controls for the subsidized firms before the treatment, and controls before and after the treatment.

10.2 Continuous Treatments

As mentioned earlier the previous estimators focus on binary treatments, that is, one distinguishes only the subsidy receipt and no subsidy receipt. However, in the R&D context, the size of the treatment may play an important role for the treatment effects, of course. We just briefly refer to extensions of the binary treatment case.

Lee (1994) and Honoré et al. (1997) provide semiparametric selection models when the treatment is not only a binary variable, but of Tobit-type, i.e. it is zero for the non-treated firms but positive continuous for treated firms (the value is the amount of the subsidy).

IV regressions are not limited to discrete treatment. The same procedure would also be valid if the amount of funding is available. See e.g. Wooldridge (2000) for a comprehensive discussion on how to obtain treatment effects with IV regressions.

Imbens (2000) has introduced a treatment effects estimator that allows to account for heterogeneous but still discrete treatments. The multiple treatments could either be different programmes, e.g. a subsidy of a local government versus an EU subsidy, or the size of a subsidy could be grouped into different classes, e.g. low, medium, high subsidy. Similarly, Gerfin and Lechner (2002) present a matching approach for heterogeneous treatments.

Recently, Hirano and Imbens (2004) suggested estimating dose-response functions using a generalized propensity score method. This is, like matching, a non-parametric method but is suitable for continuous treatments.

11 References

Andrews, D.W.K. and M.M.A. Schafgans (1998), Semiparametric estimation of the intercept of a sample selection model, *Review of Economic Studies* 65, 497-518.

Blundell, R. and M. Costa Dias (2000), Evaluation methods for non-experimental data, *Fiscal Studies* 21(4), 427-468.

Blundell, R. and M. Costa Dias (2002), Alternative approaches to evaluation in empirical microeconomics, *Portuguese Economic Journal* 1, 1-38.

Cosslett, S.R. (1991), Nonparametric and semiparametric estimation methods in econometrics and statistics, in: W.A. Barnett, J. Powell and G. Tachen (eds.*), Semiparametric estimation of a regression model with sample selectivity*, Cambridge: Cambridge University Press, 175-197.

Czarnitzki, D., B. Ebersberger and A.Fier (2007), The Relationship between R&D Collaboration, Subsidies and R&D performance: Empirical Evidence from Finland and Germany, *Journal of Applied Econometrics* 22(7), 1347-1366.

Czarnitzki, D. and C. Lopes Bento (2011), Innovation Subsidies: Does the Funding Source Matter for Innovation Intensity and Performance? Empirical Evidence from Germany, ZEW Discussion Paper No. 11-053, Mannheim.

European Union, Directorate-General for Regional Policy (2010), *Investing in our regions: 150 examples of projects co-funded by European regional policy*, Brussels.

Gallant, R. and D. Nychka (1987), Semi-Nonparametric Maximum Likelihood Estimation, *Econometrica* 15, 363-390.

Gerfin, M. and M. Lechner (2002), A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland, *Economic Journal* 112(482), 854-93.

Heckman, J.J. (1990), Varieties of Selection Bias, *American Economic Review* 80, 1121-1149.

Heckman, J.J., R.J. Lalonde and J.A. Smith (1999), The economics and econometrics of active labor market programmes, in: A. Ashenfelter and D. Card (Hrsg.), *Handbook of labor economics* 3, Amsterdam, 1866-2097.

Heckman, J.J., J.L. Tobias, and E. Vytlacil (2001), Four parameters of interest in the evaluation of social programmes, *Southern Economic Journal* 68(2), 210-223.

Heckman, J., J.L. Tobias, and E. Vytlacil (2003), Simple Estimators for Treatment Parameters in a Latent-Variable Framework, *Review of Economics and Statistics* (85)3, 748-55.

Hirano K. and G.W. Imbens (2004), The propensity score with continuous treatments, in: A. Gelman and X.L. Meng (eds.): *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives*, Wiley and Sons.

Honorè, B.E., E. Kyriazidou and C. Udry (1997), Estimation of Type 3 Tobit Models Using Symmetric Trimming and Pairwise Comparison, *Journal of Econometrics* 76: 107-128.

Imbens, G.W. (2000), The role of the propensity score in estimating dose-response functions,*Biometrika* 83, 706-710.

Lechner, M. (1998), *Training the East German labour force: microeconometric evaluations of continuous vocational training after unification*, Heidelberg.

Lee, L.F. (1994), Semiparametric Two-Stage Estimation of Sample Selection Models Subject to Tobit-Type Selection Rules, *Journal of Econometrics* 61(2), 305-344.

Newey, W.K. (1999), Two-step series estimation of sample selection models, *MIT Working Papers No. 99-04*, Cambridge, MA.

Robinson, P. (1988), Root-N-Consistent semiparametric regression, *Econometrica* 56(4), 931-954.

Rosenbaum, P.R. and Rubin, D.B. (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika* 70(1), 41-55.

Rosenbaum, P.R and D.B. Rubin, (1984), Estimating the Effects Caused by Treatments: Comment [On the Nature and Discovery of Structure], *Journal of the American Statistical Association* 79(385), 26-28.

Roy, A.D. (1951), Some Thoughts on the Distribution of Earnings, *Oxford Economic Papers* 3(2), 135-146.

Rubin, D.B. (1977), Assignment to treatment group on the basis of covariate, *Journal of Educational Statistics* 2, 1-26.

Wooldridge, J.M. (2000), *Econometric analysis of cross-section and panel data*, Cambridge: