KATHOLIEKE UNIVERSITEIT LEUVEN

Counterfactual impact evaluation of cohesion policy

Work package 2:

Examples from Support to Innovation and Research

Interim Report

July 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

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 is planning:

• 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 an interim report six month after the kick-off meeting as agreed. The interim report should focus on task 1 and provide a draft of what will be done under task 2. This report is intended to fulfill these requirements. Furthermore, the current report will

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

give concrete suggestions on how to address task 3 (a more detailed analysis for one region) which has been added during the inception meeting.

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 allow drawing conclusions about the effectiveness of different instruments. This may enable local policy makers to improve the selection of different instruments in the future.

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. Finally, we suggest a region and an approach for task 3, added after the inception meeting, to evaluate one beneficiary region in more detail.

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 **at least two beneficiary regions**. In order to be able to fulfill this requirement, 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 region, regions for future analysis could be chosen.

The following subsection describes how the data was collected, assessed and merged to other datasets.

2.1.1 Collecting and merging the data of the retained regions

The data collection exercise started during the inception phase. 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 statistical file formats or databases.

The next step consisted in linking this publicly available data to an external dataset, i.e. the Amadeus database of Bureau Van Dijk, to get further information of the beneficiary firms. This link to an external database provides more information on the recipient firms, and further allows drawing a control group of non-beneficiaries for each selected region.

Subsequently, innovation data has been collected for both, the selected beneficiaries and the randomly drawn controls. As the analysis will cover multiple countries, 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.

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 has been outlined in more detail in 3.3 in the inception report. All potential hits of the text field search engine have been manually checked.

Table 1 presents the activities that have been dealt with during the inception phase as well as the current state of these activities.

Table 1: Current state of activities that started during the inception phase

Activity	Description	Status
Download of regional data	The publicly available data on the regions of the 27 Member States has been downloaded.	Completed
Assessment of data quality	Manual checks of data availability, quality and content have been made in order to judge which regions could be used for further analysis.	Completed
Choice of regions	According to the data quality (data format, information contained, …), geographical balance and equilibrium between old and new Member States, several regions have been retained for further analysis, namely: France (all regions), Wales and London, Czech Republic (all regions), Slovenia (all regions), Slovakia (all regions), Poland (all regions) and Spain (all regions). • For Spain, the data has not been provided in excel or word format by the Member State. Consequenly, Spain is not considered for further analysis. • The Slovenian data does not appear usable for further analysis (not enough cases for a quantitative study) and thus Slovenia will be dropped from the selected regions. • The project team also considered the data from Flanders but these were also too few projects for conducting a meaningful quantitative analysis.	Completed
Conversion into a database compatible format	The data of the various regions are published in different formats (e.g. html, excel, word, pdf, …). In order to be able to work with these data, it has to be converted into a database compatible format. The complexity of these conversions is directly linked to the original format the data is published in.	Completed
Conversion into complete database tables	Once the data has been converted into a database compatible format (e.g. excel), the separate regional sheets have to be converted into complete database tables. All excel-sheets have to be exported into separate ASCII files using tabs as a column delimiter. All exported files are then concatenated into one file. The resulting file can be exported into different	Completed

	statistical file formats or databases. The completion of this exercise can be cumbersome if special characters are involved, as is the case for e.g. the Czech Republic. This exercise has been completed for France, the Czech Republic, Flanders, Slovakia, Poland and London.	
Merger of regional data with other datasets	For the retained regions, the data of the beneficiaries have been matched with companies of the Amadeus dataset in order to get the needed firm level information. This has been done by automated text field searches by beneficiary name and manual checks (see section 4 of this report). For France, this exercise has been done at the country level, then on a regional level. As the data does not contain the beneficiaries address and as many beneficiaries have very unspecific names, a lot of potential hits have been received during this exercise. By applying regional information, the ratio between searched and found was narrowed down and got much more precise. Nonetheless, a manual check of the potential hits has been undertaken. The same exercise took place for the Czech Republic. After the manual checks of the potential hits issued form the merger of the 2 first databases, the pooled datasets have been linked to patent data along with sample of firms drawn from the Amadeus database (control group).	Completed

During the accomplishment of the above tasks, 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.

2.1.2 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 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.

> **Recommendation 1**: **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>N.B.</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

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).

<u>Recommendation 3</u>: **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.

<u>Recommendation 4</u>: **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 if 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.

<u>Recommendation 5</u>: 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 would look like in an ideal case.

2.1.3 Illustration using data from the Czech Republic

In the following subsection, we will 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, lehký stavební 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

Lost observations for further analysis and quality of the remaining data

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

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

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?

Some descriptive statistics

Here below we will display some descriptive statistics on the Czech Republic data. Table 1 displays the number of projects granted per year. 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.

Table 1: Number of projects granted per year

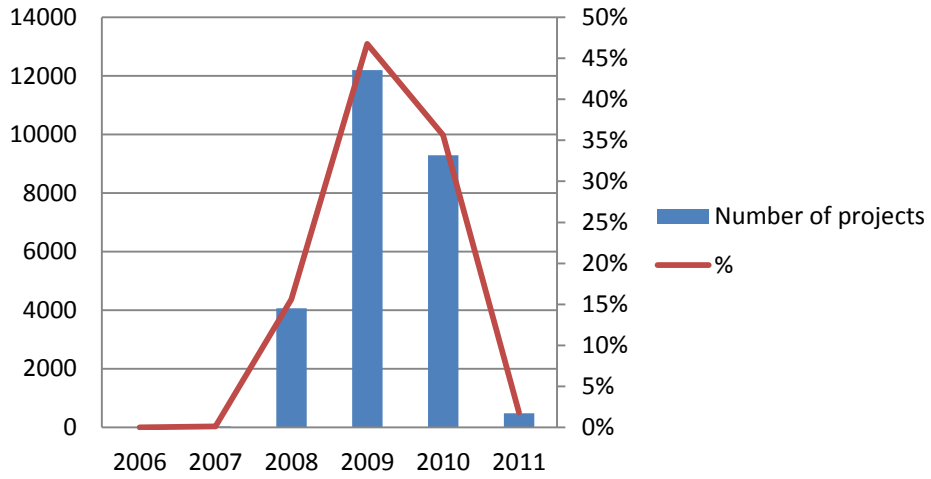
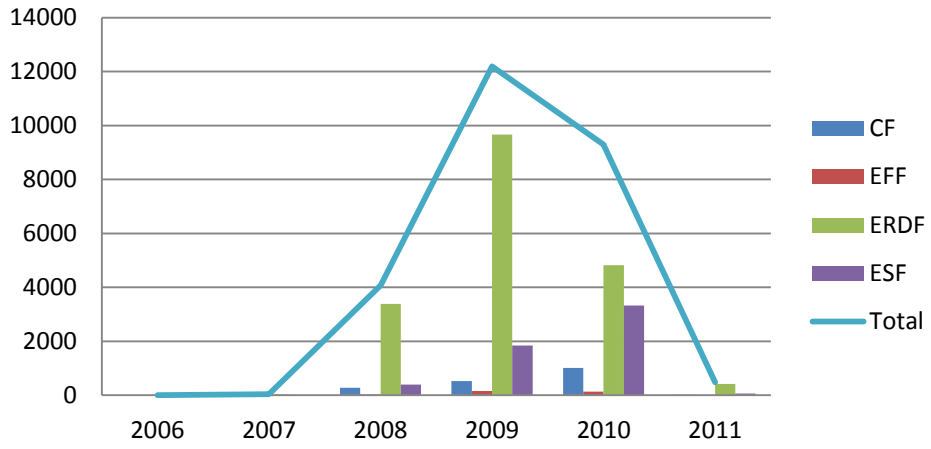


Table 2 presents the repartition of the granted project by funding type.

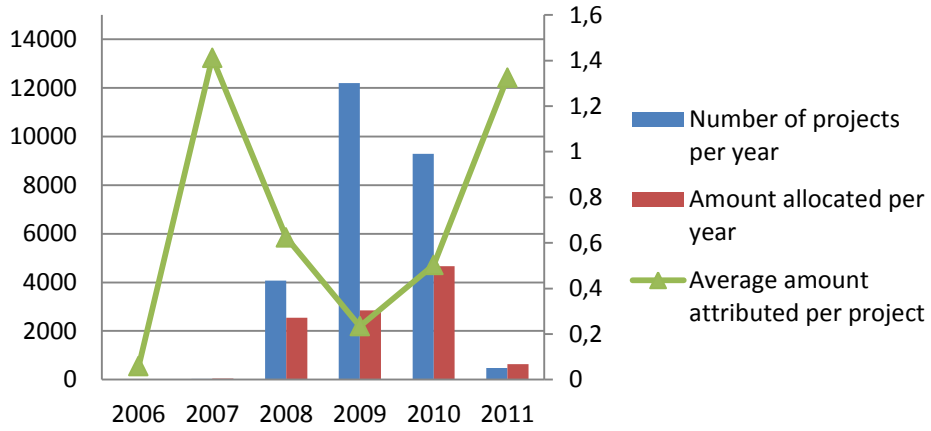
Table 2: Number of project granted per year by funding type



[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.

Finally, table 3 presents the number of projects and the amounts allocated per year (red and blue bars, relating to the left axis) as well as the average amount attributed per project (green curve, relating to the right axis). As can be seen by this graph, the average amount attributed per project is very volatile (ranking from 22 to 154,182,860 €). 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.

Table 3: Average amount allocated per year and per project



2.1.4 Data preparation for other 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, 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. While the present report only presents first econometric results on the regions of the Czech Republic, the data preparation is completed for the regions of both countries.

2.2 Task 2: Econometric Application: Preliminary analysis using data from the Czech Republic

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 research team envisages updating the Amadeus link using a more recent edition of the Amadeus database at a later stage of the project. This would also provide more recent data on possibly useful covariates such as firm size and so forth. For the initial analysis, this information is not needed, though.

The data on the ERDF 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.

[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.

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.

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 $\alpha^{DiD} = b_1 - b_2$, of 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. it 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).

In total that regression is performed with 12,887 firms and 9 year, resulting in 115,983 firm-year observations. The OLS 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. 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.

In the next step of the project we will apply a conditional difference-in-difference estimator. In our current estimation above, we do not use any other firm level information than the patent data and the treatment data. It could, however, be the case that the treated firms differ substantially from the control firms (e.g. in size and sector of activity) and could thus react differently to shocks. This would bias the estimation above. For instance, in the pre-treatment period, the treated firms have, on average, 345 employees whereas the non-treated firms have only 177 employees, on average.

Therefore we envisage to couple the DiD estimation with a matching estimator. A matching estimator would, for instance in the simplest form of "nearest neighbor" matching, search for a control observation within the control group that is the most similar firm to the treated firm. This is done on basis of observed characteristics, such as firm size and so forth, in the pre-treatment period. This matching exercise would yield matched pairs and subsequently the DiD regression is only run on these matched pairs. This conforms to the so-called "conditional DiD" estimator.

In a further analysis, we also take into account the amount of public funding instead of using simply a treatment indicator variable.

In similar vein, such an econometric exercise will also be conducted for the French data in the next phase of the project.

2.3 Task 3: Further analysis using additional data sources

During the inception meeting, it has been suggested that the research team chooses one region at which it looks more in detail because of the several pitfalls associated with the data that may confound the results of the econometric study. After giving the general overview of why this additional analysis might be useful, the research team suggests a region as well as an approach for the latter.

2.3.1 General overview

The goal of an adding an additional region to conduct more in-depth analysis would be to collect more detailed data in order to verify the results obtained with more narrow, publicly available data sources. Examples of pitfalls that may bias the estimation based on the purely publicly available data are:

- 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.
- The published grant dates may not correspond to the actual receipt of the project budget. We would thus expect an investment in year *t*, but in fact the firm only invested in *t*+1. One solution is to consider broader time windows in which the treatment most likely happened. However, this makes the measures that are used in the econometric analysis noisy and may lead again to an underestimation of the treatment effect.
- 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.
- With the publicly available data, we can only use patents as a proxy variable for innovation activity. As it is commonly known patents are a narrow measure for innovation. Thus it would be desirable to investigate a region where other, more comprehensive innovation data could be used. The Community Innovation Surveys could be a potential source. In order to use that data, however, the research team would need access to non-anonymized micro data as this would have to be supplemented with the Cohesion Policy recipient data.

If it would be possible to collaborate with regional authorities more closely and a more detailed data collection would become feasible, it could be checked to what extent the possible pitfalls mentioned above influence the results. This is not only useful to verify our initial analysis, but will also help to advise regions on future data reporting standards that should be met.

2.3.2 Suggestion of a region and an approach for the additional analysis

After careful consideration, the research group came to the conclusion that **Eastern Germany** would be a good candidate for more detailed analysis. The reason for this choice is threefold:

1. Since Eastern Germany is still considered a transition economy, it gets comparatively a high number of subsidies. Hence, the probability of having a sample large enough to do some more detailed analysis is higher than it might be for other regions.

2. Since Eastern German firms get a lot of grants from different sources, it would be possible to compare the effect of each one of these sources and estimate what grant (or combination of grants) have the highest effect on the beneficiary firms (see here below for more details on this issue).

3. Since the research team has to do with German Authorities on a regular basis, getting access to the necessary data might be easier than for other regions. Furthermore, given the usual quality of German data the research team has worked with so far, it can be expected that the Eastern German data quality is such that a more detailed analysis could be realistically envisaged.

The **approach** suggested by the researchers lies in evaluated various cases of subsidy receipt, as outlined for example in the table below. This approach is inspired by a current research project by Czarnitzki and Lopes Bento and a previous study by Czarnitzki, Ebersberger and Fier (2007). The idea of this approach is to distinguish and simultaneously analyze the impact that several grants. Indeed, regions can benefit from public support from various sources (national funds, Cohesion fund, Framework contract etc.). These various funding sources all have different priorities, and hence different selection criteria and supposedly different impacts on the outcome variables. Basing our analysis on the matrix like the one in the Table below would thus allow seeing what the impact of a grant is by itself or in combination with others. In other words, we would compare several "actual" situations to their "counterfactual" situation. For example, case 1 in the Table below would compare firms that get only funding from Eastern German Authorities to their counterfactual situation of not getting subsidies for R&D and innovation. Case 4 would compare firms that get funds from the EU to a counterfactual situation where firms get public support from national Authorities.

These comparisons can be adapted. We could for instance compare national to EU grant without distinction of the EU funding source in a first step. Subsequently, the sample could be narrowed and consider only EU grant recipients. The latter could be divided according to the funding source, allowing to differentiate for instance between the Cohesion fund, the Framework contract, and funds stemming from other (national) sources. This analysis framework is very flexible and can be adapted according to the preferences of the Contracting Authority. The only requirement for the various cases to be feasible is to have a minimum of observations for a treatment and a control group. If this condition is fulfilled, the various econometric techniques explained in the previous section can be used in this framework, and the most adequate given the data will be retained.

The advantage of this approach is that it does not only allow to see whether getting grants is desirable compared to a counterfactual situation of not getting grants, but that it further allows to evaluate the impact between different grants (or the combination of the latter) and hence give much more precise policy conclusions. In addition to estimating whether we are facing crowding out effects of private investment by EU money, it allows to analyze the impact of the various grants (or the combination of the latter) on innovation input (R&D and innovation investment) as well as on innovation output (sales of new products, number of filed patents etc).

Subsidy effect on R&D intensity

	Dependent variable: R&D Intensity (R&D expenditures/sales * 100)			
		Actual status (m)		
Counterfactual (*l*)	No funding	Only national funding	Only EU funding	Funding from both sources
No funding		case 1	case 2	case 3
Only national funding	case 7		case 4	case 5
Only EU funding	case 8	case 9		case 6
Funding from both sources	case 10	case 11	case 12	

Note: The table reads from column to row. E.g., **case1**: "What would the output of firms that only getnational funding be, if they would not have been funded at all?"; **case 4**: "Would the output of firms that only get EU funding differ if they only got national funding?"; **case 10**: "Would a firm that gets no public support spend more on R&D and innovation if it would get funded from both, the EU and the national government?"

It has to be checked whether the study by Czarnitzki and Lopes Bento can be adapted to explore potential differences between cohesion policy and other policy instruments with the available data. In their original paper, Czarnitzki and Lopes Bento used data from the "Mannheim Innovation Panel" (MIP) which is an annual survey conducted by the Center for European Economic Research (ZEW), Mannheim. The data used in Czarnitzki and Lopes Bento (2011) cover a time period from 1992 to 2007. For the analysis of cohesion policy only a single cross-section of the Mannheim Innovation Panel could be used, as not every survey contains detailed questions on subsidies. The period of interest for the present study is 2007-2013. Thus, potentially the MIP surveys from 2008, 2009 and 2010 could be used to explore impacts of cohesion policy. However, only the 2009 wave of the MIP survey includes a question allowing to distinguish between different European Sources (i.e. Framework money vs money from other European sources). It would thus have to be checked whether the available data suffices to conduct a quantitative study as suggested here (with regard to the numbers of observations in the survey, and the overlap between the survey respondents and recipients of funds from cohesion policy programs).

The feasibility of such an analysis will be checked after further discussions with DG Regio on the selection of beneficiary regions for this study.

3 Timeline

Table: Time schedule and deliverables

Inception report	25 February 2011	√
Progress reports	By the 25th of each month (or next working day) in months where no other deliverable	√
Presentation of methodological findings at meeting in Hungary	May	√
Interim report – i.e. task 1 plus an initial draft for task 2	18 July 2011	√
Interim meeting	28 & 29 July 2011	
Draft final report	18 November 2011	
Presentation of findings at a Polish presidency event in Warsaw	12 December 2011	
Final report	24 December 2011	

Detailed information on the progress made on a monthly basis can be found in Appendix D.

4 Operational management plan

4.1 Time schedule

The project started in January 2011 and will end in December 2011. The table below shows the different steps that have already been executed so far and that further will be executed during the foreseen time period.

Table: Time schedule and deliverables

	Quarter 1			Quarter 2			Quarter 3			Quarter 4		
	1	2	3	4	5	6	7	8	9	10	11	12
	Task 1			**Task 1**			**Task 2**			**Task 2 & 3**		
Deliverable 1: Inception report		√										
Task 1: Data preparation						√						
Download of data of 27 member states												
Data quality assessment and choice of the regions												
Data conversion into database compatible format												
Creation of complete database tables												
Merging of data of retained regions with other datasets												
Deliverable 2: Interim report						√						
Task 2: Economic and counterfactual work and analysis												
Literature review and desk research												
Evaluation and choice of various micro econometric methods												
Micro econometric estimation and testing												
Interpretation of the results												
Task 3: Conclusions and headline figures												
Assessment of re-applying the used methods elsewhere for ERDF support												
Deliverable 3: Draft final report											●	
Deliverable 4: Final report												●
Deliverable 5: Presentation												●
	Completed						Ongoing					

4.2 Work plan

The following table outlines the repartition of the tasks and responsibilities between the team members.

Table: Division of tasks and responsibilities

	D. Czarnitzki	T. Doherr	C. Lopes Bento
Overall coordination	●		
Task 1: Data preparation	●	●	●
Data conversion and creation of complete database tables		●	
Merging of various datasets		●	
Manual check of hits between merged datasets	●	●	●
Task 2: Economic and counterfactual work and analysis	●		●
Literature review, desk research & evaluation of micro econometric tools	●		●
Micro econometric estimations and testing	●		●
Interpretation of the results	●		●
Task 3: Conclusions and headline figures	●		
Assessment of re-applying the used methods elsewhere for ERDF support	●		
Presentation of the findings	●		
	● main actor	●	cooperation

Furthermore, it has been agreed at the kick-off meeting that regular coordination and updates on respective progress between work package 1 and work package 2 will be assured. WP1 has been awarded to ASVAPP. The first meeting between the consultants of WP1 and WP2 will take place at the inception meeting in Brussels on 15. March 2011. Thereafter, meetings will be held as judged necessary.

5 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)	Contains activity codes.
Molise	Pdf	* Beneficiary institution * Project name * Year * Amount (detail)	Contains activity codes.
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.
	Friuli Venezia Giulia				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	Search engine	

	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 be available**.
	Severen tsentralen	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 be available**.
	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 be available.**
	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 be available.**
	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 be available.**
	Yuzhun tsentralen	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 be available.**
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 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.	**Czech Republic**	Praha	excel			
		Stredni Cechy	pdf	* Beneficiary * Project * Amount	institution name	**Detailed information available only in Czech.** The titles are also available in English.
		Jihozapad	pdf	* Beneficiary * Project * Amount	institution name	**Detailed information available only in Czech.** The titles are also available in English.
		Severozapad	Excel	* Beneficiary * Project * Amount	institution name	**Detailed information available only in Czech.** The titles are also available in English.
		Severovychod			ERROR ON THE PAGE	
		Jihovychod	Excel Pdf	* Beneficiary * Project * Amount	institution name	**Detailed information available only in Czech**. The titles are also available in English.
		Stredni Morava	word	* Beneficiary * Project * Amount	institution name	**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 * Project * Amount	institution name	Classified per activity.
		Mazowieckie	RAR Excel	* Beneficiary * Project * Amount	institution name	Classified per activity.
		Malopolskie	RAR Excel	same for all regions		
23.	**Romania**	Nord-Vest	search engine			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.

6 Appendix B: Example of a data reporting structure

Format: excel

Reporting language: English

Name of beneficiary, complete address and legal form	ID	Operation	Operational Programme	Fund EU		European Union funding							
					Duration of the projects in months (including exact start and end date)	Date of allocation	Amounts allocated (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.**

7 Appendix C: 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.[5] As treatment effects models usually consider discrete treatments we start with such methodologies, and briefly mention possible extensions for multiple or continuous treatments afterwards.

7.1 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:[6]

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

where Y^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

$$E\left(Y^C \middle| S=1\right) \neq E\left(Y^C \middle| S=0\right) \qquad (2)$$

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

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

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.

Suppose the outcome equation has following form

$$\begin{aligned} Y &= X\beta + S\alpha + U \quad \text{if} \quad S = 1 \\ Y &= X\beta + U \quad \text{if} \quad S = 0 \end{aligned} \tag{3}$$

where X denotes a set of exogenous variables, β their 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

$$S^* = Z\gamma + V, \tag{4}$$

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:

$$S = \begin{cases} 1 & \text{if } S^* > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

In the following we refer to this as selection equation.

7.1.1 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:

$$\begin{aligned} E(Y \mid S = 1) &= X\beta + \alpha + \rho\phi\left(\frac{Z\gamma}{\sigma_V}\right)\Phi\left(\frac{Z\gamma}{\sigma_V}\right)^{-1} \\ E(Y \mid S = 0) &= X\beta - \rho\phi\left(\frac{Z\gamma}{\sigma_V}\right)\left[1 - \Phi\left(\frac{Z\gamma}{\sigma_V}\right)\right]^{-1} \end{aligned} \tag{6}$$

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. (6) and estimate the two equations separately with least squares. In order to obtain TT, we calculate

$$\alpha = X\left(\beta_1 - \beta_0\right) + \left(\rho_1 - \rho_0\right)\phi\left(\frac{Z\gamma}{\sigma_V}\right)\Phi\left(\frac{Z\gamma}{\sigma_V}\right)^{-1} \qquad (7)$$

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.

7.1.2 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. [7]

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.

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

7.1.3 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. (2). This condition means that the receipt of subsidies and potential outcome are independent for firms with the same set of exogenous characteristics

$$Y^T, Y^C \perp S \mid X = x. \tag{8}$$

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

$$E\left(Y^C \mid S=1, X=x\right) = E\left(Y^C \mid S=0, X=x\right) \tag{9}$$

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

$$\alpha = E\left(Y^T \mid S=1, X=x\right) - E\left(Y^C \mid S=0, X=x\right) \tag{10}$$

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

$$\hat{\alpha} = \sum_{i \in T} \left(Y_i - \sum_{j \in C} w_{ij} Y_j \right). \qquad (11)$$

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.

7.1.4 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:

$$\alpha_{TT}^{DiD} = \left(E\left(Y_{t_1} \mid S=1\right) - E\left(Y_{t_0} \mid S=1\right) \right) - \left(E\left(Y_{t_1} \mid S=0\right) - E\left(Y_{t_0} \mid S=0\right) \right) \qquad (12)$$

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.

7.2 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.

8 Appendix D: Monthly progress

MONTHLY PROGRESS REPORT	Page 1/1

Prof. Dr. Dirk Czarnitzki K.U.Leuven	Naamsestraat 69, 3000 Leuven, Belgium. Phone: +32 16 326 906; Fax: +32 16 326 732.

Contract details		Report details	
Contract nbre:	**Project title**	**Month ending:**	**Report Nbre :**
2010.CE.16.0.	Counterfactual impact evaluation of cohesion policy, work package 2: Examples From	February	1

Tasks accomplished this month:

- Data of all the regions of the 27 Member states downloaded;
- Quality of the available data assessed;
- According to their quality and the decisions taken during the kick-off meeting, several regions have been selected for further analysis, namely:
  - France (all regions),
  - Wales and London,
  - Czech Republic (all regions),
  - Slovenia (all regions)
  - Slovakia (all regions)
  - Poland (all regions)
  - Spain (data yet to be received in excel of word format)
- For the selected regions, started data conversion into a database compatible format;
- Separate regional spreadsheets per country converted into complete database tables;
- French beneficiaries matched to French companies in the Amadeus database, by country, then including regional information. Matching completed;
- Conversion of Czech Republic data into database compatible format and converted into one database table;
- Matching of Czech Republic beneficiaries with the Amadeus database ongoing;
- Inception report drafted.

Planned activities for next month:

- Conversion of remaining regions into complete and compatible database tables;
- Matching the of the various databases to the Amadeus data;
- Assessing the quality of the matched data of the various regions and decide on 5 regions/macro-regions for further matching with patent data;
- Presentation of the inception report, the problems encountered and the lessons learned at the inception meeting;
- Presentation by the IT expert on data specific problems and on ways to improve the data collection to facilitate future evaluations at the inception meeting.

General comments:

None at this stage.

MONTHLY PROGRESS REPORT

Prof. Dr. Dirk Czarnitzki
K.U.Leuven

Naamsestraat 69, 3000 Leuven, Belgium.
Phone: +32 16 326 906; Fax: +32 16 326 732.

Contract details		Report details	
Contract nbre:	**Project title**	**Month ending:**	**Report Nbre :**
2010.CE.16.0.	Counterfactual impact evaluation of cohesion policy, work package 2: Examples From	April	2

Tasks accomplished this month:

- Check data matched with Amadeus database and potential hits have been manually checked.
- Control group for Czech recipients has been drawn.
- Czech data is now being linked to patent data.
- Beneficiary data of Wales, London, Flanders data have been matched to Amadeus and have been manually checked.
- French data has been matched to Amadeus and manual checks of potential hits are ongoing.

Planned activities for next month:

- Polish and Slovakian beneficiary data match to Amadeus has to be manually checked.
- Completion of manual checks for France.
- Control groups will be drawn for Flanders, London, Wales and France from Amadeus database.
- Control groups have to be drawn for Poland and Slovakia from Amadeus database.
- Manual quality check of the match between Czech firm-level data and the patent database.

General comments:

None at this stage.

MONTHLY PROGRESS REPORT	Page 1/1

Prof. Dr. Dirk Czarnitzki K.U.Leuven	Naamsestraat 69, 3000 Leuven, Belgium. Phone: +32 16 326 906; Fax: +32 16 326 732.

Contract details		Report details	
Contract nbre:	**Project title**	**Month ending:**	**Report Nbre :**
2010.CE.16.0.	Counterfactual impact evaluation of cohesion policy, work package 2: Examples From	May	3

Tasks accomplished this month:

- Czech data is linked to patent data. Quality checks ongoing
- Beneficiary data of Wales, London, Flanders are being quality checked and control group is being drawn.
- French data has been matched to Amadeus and manual checks of potential hits are ongoing.
- Polish and Slovakian beneficiary data match to Amadeus is being manually checked.
- Ongoing:
    - Econometric literature review
    - Economic literature review

Planned activities for next month:

- Manual checks for France is still ongoing. Completion is planned for June.
- Control groups will be drawn for Flanders, London, Wales and France from Amadeus database.
- Control groups have to be drawn for Poland and Slovakia from Amadeus database.
- First descriptive statistics for Czech Republic
- First econometric analysis on Czech data

General comments:

None at this stage.

MONTHLY PROGRESS REPORT	Page 1/1

Prof. Dr. Dirk Czarnitzki	Naamsestraat 69, 3000 Leuven, Belgium.
K.U.Leuven	Phone: +32 16 326 906: Fax: +32 16 326 732.

Contract details		Report details	
Contract nbre:	**Project title**	**Month ending:**	**Report Nbre :**
2010.CE.16.0.	Counterfactual impact evaluation of cohesion policy, work package 2: Examples From	June	4

Tasks accomplished this month:

- French data has been matched to Amadeus and manual checks of potential hits have been completed.
- Polish and Slovakian beneficiary data match to Amadeus is being manually checked.
- Ongoing:
  - Econometric literature review
  - Economic literature review
- Drafting interim report

Planned activities for next month:

- Linking French cleaned French recipient data to Amadeus and Patent data.
- First descriptive and econometric analyzes on French data
- Finishing descriptive and initial econometric analyzes for the Czech data.
- Finishing the interim report
- Presentation and discussion of interim report.

General comments:

None at this stage.

8 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, 497-517.

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: