Big Data approaches to estimating the impact of EU research funding on innovation development

Pukelis, L.; Stanciauskas, V.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Big Data approaches to estimating the impact of EU research funding on innovation development[1]

Dr Lukas Pukelis[*] and Vilius Stanciaukas[**]

[*]*Lukas.pukelis@ppmi.lt*
Institute of Public Policy and Management, Gedimino ave. 50, Vilnius, 01110, Lithuania.

[**] *vilius.stanciauskas@ppmi.lt*
Institute of Public Policy and Management, Gedimino ave. 50, Vilnius, 01110, Lithuania.

Introduction

The aim of this paper is to provide a brief overview is to discuss the limitations and difficulties associated with measuring the impact of research funding as well as to propose a novel approach how these limitations could be overcome. The proposed approach relies on sifting through and analyzing the vast amounts of data already posted on the web instead of engaging in efforts to collect new data. It also is highly automatable, allowing to collect and process unprecedently large amounts of data.

We are still in relatively early stages of this project and have completed only a few pilot tests of this approach, but the overall results appear to be of high quality and look very promising. Should our efforts be successful, using such approach would allow to gather more data at a lower cost and to significantly decrease the time-lag in the data sources. Furthermore, as such approach places no burden on the recipients of the research funding, it allows monitoring activities to be carried out even after the project ends, thus enabling the detection of the effects which take longer to manifest.

Context and Relevance

The EU spends significant amounts of money on research. Roughly €80 billion will be spent though EUs flagship "H2020" programme alone between 2014 and 2020, [2] with the amount of funding destined to increase in the programme's successor – "Horizon Europe".[3] The contribution of such sizable amounts of money naturally begs the question, what is the impact of the EU research investment? However, this question is notoriously difficult to answer for two major reasons: first – EU makes conscious effort to fund research with no immediate market applications; second – there hardly is any reliable data on the actual outputs and impacts of EU funded projects.

In order not to distort the market competition, EU funded research projects have to focus on early technology-readiness-levels (TLR 1-3). That means that even if technology developed

[1] This work was supported by the EU and draws insights from the "Data4Impact" project funded under H2020 "Transformations" topic aims to estimate the research outcomes, outputs, and impact of health-related EU-funded projects in FP7 and H2020. More information www.data4impact.eu;

[2] Horizon 2020 website < https://ec.europa.eu/programmes/horizon2020/en/horizon-2020-statistics>
[3] European Commission: Press Release <http://europa.eu/rapid/press-release_IP-18-4041_en.htm>

during the project is proven variable, its effects and market applications would take years to manifest. Naturally, by then tracing the particular market product to its source funding is extremely difficult.
Second, to ensure accountability and prevent fraud, EU already heavily monitors the funded projects and their participants. Project teams already have to fill-in various surveys during the project and after it finishes. Furthermore, in addition to the regular monitoring, the EU regularly commissions various studies to evaluate various aspects of its programmes and to gather perceptions from their participants. This results in what is called *survey fatigue* among the beneficiaries of EU funding.
Vast volume of surveys creates weariness and reduce the willingness to participate. Consequentially, these surveys tend to suffer from low response rates and low-quality responses.[4] These two factors together increasingly diminish the usefulness and validity of the surveys.[5] Finally, these surveys can only capture project outcomes, i.e. results delivered by the project, while it lasted. Capturing the effects of the project that manifested some time after its end-date raises significant survey administration challenges (e.g. identifying respondents whose address or place of employment might have changed) and is likely to suffer from even lower response rates, as the project participants are not obliged to respond after the project ends.
Therefore, measuring the impact of the EU-funded research projects is highly relevant, due to vast amount of resources committed for this purpose and highly arduous due to the challenges mentioned above. Any effort to measure the impact of EU research funding would require meeting two criteria: first, it has to address the issue of early TLR research funding. Second, it must do so in a non-invasive manner, without actively involving the beneficiaries in the measurement process.

Overview of the previous research and data

To the authors best knowledge, there have not been any systematic attempts to comprehensively evaluate the impact of the EU funding. By this we mean attempts to evaluate the impact of EU funding which would comprehensively cover different aspects/ types of impacts and provide data which would be detailed representative at a low level (participating entity) and thus could be aggregated to higher levels (regional/ sectoral/ national).
Rather there have been many attempts to look into certain aspects of the EU funding in isolation. Notable examples include: study on the impact of EU funding on researchers' careers,[6] attempts to estimate the overall added value of EU funding,[7] or various attempts to estimate the impact of the EU funding on the various regions in Europe.[8]
These studies utilized three main methodologies to generate inferences about the impact of the European funding: surveys, case studies, or econometric modelling. The concerns over the quality of such studies were already briefly discussed above. Meanwhile, the case studies, though provide rich and valuable data on fine-grained impacts of EU funding cannot be used

[4] Based on professional experience conducting vast amount of surveys which target the recipients of EU-funding.
[5] Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. *Improving survey response: Lessons learned from the European Social Survey*. John Wiley & Sons, 2010.
[6] "Research Careers in Europe" – study commissioned by DG <https://publications.europa.eu/en/publication-detail/-/publication/c97be578-9aa5-11e6-868c-01aa75ed71a1>
[7] EC "Assessment of the Union Added Value and the Economic Impact of the EU Framework Programmes" <>
[8] E. g.: Becker, S.O., Egger, P.H. and Von Ehrlich, M., 2010. Going NUTS: The effect of EU Structural Funds on regional performance. *Journal of Public Economics, 94*(9-10), pp.578-590.
Or
Wanzenböck, I. and Piribauer, P., 2018. R&D networks and regional knowledge production in Europe: Evidence from a space-time model. *Papers in Regional Science, 97*, pp.S1-S24.

to draw general conclusions and at best can be used only as anecdotal evidence. The econometric modelling, on the other hand paints a very broad picture and aims to evaluate the impacts on very high (country or regional) level. Yet, even these results have to be taken with a grain of salt, as they tend to rely on broad and simplistic assumptions.[9]

In other words, previous attempts to estimate the impact of EU funding faced two main problems: their scope was narrow and, therefore, they were able to provide insights only to a specific aspect of the impact of the EU funding and their methods were limiting in a sense that they were able to provide only either very broad and undetailed account of the impact (econometric studies), unrepresentative set of stories of how this impact unfolded (case studies) or somewhat unrepresented picture of participants' opinions (surveys).

The main reason why the previous studies were limited in their thematic scope or their level of analysis was the fact that they heavily relied on manual labor to gather and process the data. Since data collection through interviews and case studies place a significant labor demands of both data providers (interviewees/ survey respondents) and data collectors (scholars/ policy analysts), these efforts are naturally limited in scope. Econometric studies, meanwhile, are limited by the collected, cleaned and processed data that is available to the researchers. Such data are usually provided by the national statistical offices and processed by such entities as Eurostat, which rely on extensive and inflexible bureaucracies and, therefore, to introduce new indicators and to acquire data for them is a gargantuan task, which requires extensive coordination and takes years to implement.

Having discussed the limitations of the existing efforts and approaches, naturally, raises the question whether a viable alternative exists to overcome these limitations. We propose that this could be done by utilizing Big Data approaches. Since its inception in 1991 the amount of data on the 'World Wide Web' (the visible part of the internet) has been increasing steadily, while the data in the rest of internet (data repositories, databases, databanks, etc.) has been increasing exponentially.[10] Researchers have been developing tools and methodologies how this data could be mined and analyzed for the last two decades,[11] and recent increases in the computational power and increased availability of cloud computing services made the tools to carry out these tasks available to ordinary researchers outside major scientific centers and computer labs.

We propose that the vast amounts of data already stored on the Web could be mined, processed and analyzed to generate insights into policy-relevant questions, such as the impact of EU funding. Mining data from the web can yield data on the level of individual participants (persons or institutions), covering vast range of possible outputs, outcomes and impacts. These processes also have a large automation potential, meaning that data collection can be carried out on an unprecedented scale. Finally, such data structure can be easily aggregated to higher levels, yielding insights into the impacts at a national or regional level.

Methods for data acquisition and analysis

Estimating the full range of the outputs, outcomes, and impacts of the EU funding is mammoth task and largely beyond the scope of this paper. Instead we suggest to focus on a particular aspect of EU funding impact – boosting innovation as an example of how this could

[9] Salter, A.J. and Martin, B.R., 2001. The economic benefits of publicly funded basic research: a critical review. *Research policy*, *30*(3), pp.509-532.

[10] Jeff Schultz "How Much Data is Created on the Internet Each Day?" <https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/>

[11] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N., 2000. Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), pp.12-23.

be accomplished. This is an example illustrating how the Big Data approaches could serve to help estimate the impact of the EU funding within the confines of this paper. However, it is only a part of a larger conceptual framework[12] employed to this end in one of the projects, "Data4Impact"[13] which we are currently carrying out together with our consortium partners.

Research Design

We intend to use Counterfactual Impact Evaluation (CIE)[14] design to answer this question. CIE is an adaptation of the randomized controlled trail (RCT) study design used in medical and pharmaceutical sciences used to measure and evaluate the effects of policy intervention. It uses the comparison of the subjects who benefited from policy intervention (treatment group) and randomly chosen similar group of non-beneficiaries (control group).

We intend to use all the participants of the projects funded under FP7 as the treatment group, while the control group will be composed of randomly selected European enterprises. Our sample will consist of roughly 30, 000 entities will be constructed of 28 sub-samples for each member state. In each national sub-sample there will be around 1 100 enterprises chosen using randomized quota sampling. The sub samples will be representative at a national level and will reflect the economic structure, in terms of the turnover share per NACE sector group from the overall national enterprise turnover. We distinguish between NACE sector groups: B-E – manufacturing; F – construction; and G-N (excl. K) – services.

NB! At the moment, we are still in the process of gathering and processing data on both the treatment and control groups. Therefore, the preliminary results presented in the paper will focus on the pilot/POC run we performed on the participants of "FP7 Health" projects.

Method for data acquisition

We acquired data on the participant companies from their websites. We developed specialized web-scrapers/crawlers, which entered a specific web-domain and collected all textual data where-in. The crawlers were constructed to obey 'Robots.txt' instructions and to avoid pages which were likely to contain sensitive or personal data, such as contact information, emails, etc.

There were overall, 1492 distinct enterprise participants in FP7-Health. Out of them, we collected data on 1301 companies. However, since some companies have gone out of business or abandoned their web-presence for other reasons, out of these 1301 domains, 1154 were functional at the time of data collection. This means that we achieved the data collection rate of roughly 80% which is good, keeping in mind that some of the FP7 Health projects ended a decade ago.

Method for data analysis

The main question that we seek to answer with the data is whether a particular enterprise is innovative or in other words, whether it has produced at least one innovation output during a given period. Additionally, we seek to know what kind of innovation it was and whether it is market-ready.

Given the extremely large quantity of data and the need to parse through such quantity of data on a regular basis, the only feasible way to answer this question is to employ an automated

[12] Data4Impact: "Conceptual Framework" <http://www.data4impact.eu/wp-content/uploads/2018/05/D2.1_Conceptual-Framework_M6.pdf>

[13] Data4Impact Project website < http://www.data4impact.eu/>

[14] EC Science Hub: Counterfactual Impact Evaluation < https://ec.europa.eu/jrc/en/research-topic/counterfactual-impact-evaluation >

solution. To this end, we have developed a specialized algorithm[15] which was trained on a sub-set of manually labelled data. The algorithm recognizes innovation mentions in text and performs binary classification: 0 – no mention of innovation; 1 – mention of innovation output.
Naturally, as any automated solution, the algorithm does not work perfectly and has a small share of false positives and false negatives. Yet the overall performance of the model is very satisfactory with the F1 score[16] around 0.88.

Preliminary results

To date, we have only performed binary classification, i.e. determined whether an enterprise is innovative. Currently, we are developing another algorithm to determine the innovation type, stage and assign it to a particular NACE code.
Results indicate that around 42% of companies which participated in FP7 Health have recently introduced an innovation output and hence are innovative companies. However, by itself this number tells us fairly little – in order to draw conclusions whether this number is high or low, we need to place it in a context. Unfortunately, we were unable to compute the same metric for the control group by the paper revision deadline and therefore, have to rely on other data sources for to make this comparison. One of the most notable attempts to measure the prevalence of innovation among European companies is "Community Innovation Survey". Naturally, the results of CIS and our work are not directly comparable, as they use different methodologies and different definition of innovation (ours is a bit more conservative). However, a crude juxta-positioning is still possible. For the 2014 (most recent year available) the CIS indicator "Share of SMEs which introduced a new product or process over the last year" was 30.6% (see Table 1).

Table 1. Prevalence of innovations according to different measurements

Innovation prevalence in company websites (machine-coded)	Innovation prevalence in company websites (hand-coded)	Community Innovation Survey Data
42%	50.4%	30.6%

It is worth noting that when we manually labelled a sub-set of data to train the model (300 company domains with around 120 000 individual web-pages), the share of innovative companies in the sub-sample was 50.4%, while the share of innovative companies in the remainder of the population which was labelled by the model was 42%.
Currently, we are analyzing the model output to determine whether this difference was due to sampling errors or due to some quirks in the model performance. Additionally, we are labelling a larger sample of companies to further train the model.

Challenges ahead

The results we obtained look very promising and are in-line with the expectation that participants of EU-funded projects should be more innovative than the population mean. Nonetheless, as we still are in very early stages of this effort many significant challenges remain. The biggest among them is having very little control over what data can be gathered

[15] We used a Multi-Layered Perceptron, written in Python programming language, utilizing the "TensorFlow" library.
[16] Yang, Y. and Liu, X., 1999, August. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-49). ACM.

from via Big Data approaches. Using traditional surveys offers the advantage of being able to ask the target group precisely the right questions. Meanwhile, data mining can only yield data that is already on the web, i.e. one can only gather the information that has already been posted by the target group and is accessible to the public. In the pilot case presented above, this has not been an issue, as the companies generally are willing to advertise the new innovations they have developed. However, it is highly likely that with the other indicators this will not be the case and instead we will have to develop an approach to work around these limitations.

Another challenge ahead is to optimize the performance of the predictive model. Though the preliminary results look promising, our experience indicates that the biggest efforts are required not for the initial development, but rather fine-tuning of the model.

References

"Research Careers in Europe" – study commissioned by DG RTD <https://publications.europa.eu/en/publication-detail/-/publication/c97be578-9aa5-11e6-868c-01aa75ed71a1>

Becker, S.O., Egger, P.H. and Von Ehrlich, M., 2010. Going NUTS: The effect of EU Structural Funds on regional performance. *Journal of Public Economics*, *94*(9-10), pp.578-590.

Data4Impact Project website < http://www.data4impact.eu/>

Data4Impact: "Conceptual Framework" <http://www.data4impact.eu/wp-content/uploads/2018/05/D2.1_Conceptual-Framework_M6.pdf>

EC "Assessment of the Union Added Value and the Economic Impact of the EU Framework Programmes" <>

EC Science Hub: Counterfactual Impact Evaluation < https://ec.europa.eu/jrc/en/research-topic/counterfactual-impact-evaluation >

EC: Press Release <http://europa.eu/rapid/press-release_IP-18-4041_en.htm>

Horizon 2020 website < https://ec.europa.eu/programmes/horizon2020/en/horizon-2020-statistics>

Schultz, J. "How Much Data is Created on the Internet Each Day?" <https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/>

Salter, A.J. and Martin, B.R., 2001. The economic benefits of publicly funded basic research: a critical review. *Research policy*, *30*(3), pp.509-532.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N., 2000. Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), pp.12-23.

Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. *Improving survey response: Lessons learned from the European Social Survey*. John Wiley & Sons, 2010.

Wanzenböck, I. and Piribauer, P., 2018. R&D networks and regional knowledge production in Europe: Evidence from a space-time model. *Papers in Regional Science*, *97*, pp.S1-S24.

Yang, Y. and Liu, X., 1999, August. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-49). ACM.