PolicsLab: new data sources for a data informed policy making

Mariagrazia Zottoli, Contesti srl, zottoli@contesti.info
Carmelofrancesco Origlia, Contesti srl, origlia@contesti.info
Alfredo Fortunato, Contesti srl, fortunato@contesti.info
Sara Laurita, Università della Calabria, sara.laurita@unical.it

Abstract

Statistical agencies are encouraged to explore the use of new data sources, even in combination with existing ones, as input for official statistics thus, weighing these opportunities against the need to maintain a high quality of statistical output. Enrico Giovannini echoes this sentiment, stating: "Nowadays, NSIs are called to open the doors to new ways of constructing statistics in an era characterised by an infinity of new data sources", such as big data, open data, and crowdsourced data (i.e. collaborative platforms). The advent of the data revolution has generated an overwhelming increase in the amount of data available, giving rise to new opportunities for policy-makers to use a data-informed approach in defining policies. Official statistics data about economic activities is based on international classifications and, although very useful, they are not sufficient to keep at the same pace with sudden changes affecting local territories, as well as external factors that may influence a given territorial policy. Despite the possibilities, new data sources are not necessarily a panacea for the actors involved in the policy cycle who have to tackle numerous methodological and ethical challenges. This problem gave rise to the idea of offering users a concrete support tool for the definition of policies, by integrating various data sources that are useful for the adoption of an evidence-based approach. The paper presents the results of a case study in which we developed an "intelligent" platform for technological foresight called PolicsLab. The model is based on the valorisation of data available from traditional and non-traditional sources, as well as the construction of simple and intuitive features. The latter help to measure and predict territorial changes and transformations related to the innovation in Italian regions.

Keywords: PolicsLab, new data sources, data-informed, policy making support tool, innovation

1. Introduction

The paper presents the rationality of an information system able to support policy-makers in technological foresight. The model is based on the valorisation of data available from traditional and non-traditional sources as well as the construction of simple interpretative tools to measure and predict territorial changes and transformations related to the innovation system. Investments in research and innovation play a key role within the "Europe 2020- Strategy for smart, sustainable and inclusive growth". In this context, the Smart Specialisation Strategy (RIS3) is identified as a new paradigm to spread place-based policies at EU, national, and regional level. The Smart Specialisation concept describes the capability of an economic system to generate new specialisations through the discovery of new fields of opportunities based on the distinctive entrepreneurial strengths of each region (Foray, 2015).

Most of the structural changes generated by RIS3 imply the: (i) creation of the so-called related variety by recombining existing sectors or by developing new activities. (ii) Adoption of an entrepreneurial discovery process to identify current growth bottlenecks and knowledge-driven development opportunities (Hausmann and Rodrik, 2003).

2. Policy makers' information need

The Smart Specialisation approach requires national and regional policy-makers a non-trivial effort to develop evidence-based policies and innovation strategies based on a detailed analysis of socio-economic conditions with indicators reflecting the strengths and weaknesses of the innovation system (Kleibrink, 2016). Initial analysis carried out on approved RIS3, in fact, show a substantial continuity with previous policies and a choice of investment priorities weakly linked to the specific characteristics of local production peculiarities. For what concerns with the Italian situation, for example, it is not clear how regions have carried out the entrepreneurial discovery process, with reference to both new areas of specialisation and innovation leading actors (Caramis and Lucianetti, 2015). A critical issue is linked to the availability of data and information referring to: (i) the analysis of the regional innovation context; (ii) the identification of areas of specialisation and technological trajectories to invest in; (iii) the continuous verification of investment priorities defined in the light of

technological and international market scenarios and the effects of already implemented interventions.

2.1. The use of new data sources to fill the knowledge gap

The advent of data revolution has generated an overwhelming increase in the amount of data available, also outlining new opportunities for policy-makers to use a data-informed approach in defining policies and actions supporting innovation. In other words, the possibility of designing evidence-based frameworks for policy interventions is being envisaged. In this regard, there are two main types of data:

- open data (e.g. administrative data, demographic data and population statistics, economic indicators, etc.) currently used more intensively and linked together;
- data from social media, sensors and smartphones which are entirely new sources for policymaking, analysed with innovative methods such as sentiment analysis, location mapping or advanced social network analysis (Poel et al., 2015).

The use of data from official statistics based on international classifications is not sufficient to keep pace with sudden changes affecting local territories as well as external factors that may influence a given territorial policy. These datasets are limited in identifying innovative sectors, mapping innovation networks and characterising complex ecosystems (Crick et al., 2016).

The main reasons that imply the use of new data sources are linked to the very nature of the concept of innovation. In particular: it involves novelties in terms of production factors, processes and results: new capacities, organisational forms and industries, by definition, not included in the classifications of existing economic activities which are not able to provide a detailed and precise picture of innovative activities in the business world and thus serve as a conceptual basis for effective analysis and evaluation on the part of public decision-makers (Crick et al., 2016). The unstructured data sources can offer an enrichment to traditional ones for calibrating policies and strategies by mean of a data-informed approach. High quality data, more granular and detailed would help to ask new types of questions and enable new research lines useful to assess the consequences of policy interventions (Einav et al., 2014).

New data sources are not, however, a panacea for policy-makers who have to tackle numerous methodological and ethical challenges (Einav et al., 2014), recognising how they integrate rather than replace traditional methods (Boyd et al., 2012). It is evident that in order to make the best use of the opportunities represented by these sources, it is necessary to fill, first of all, the gap in the competences of policy-makers towards an increase in data literacy.

Hence the intuition to offer them a concrete support in the definition of innovation policies, by means of a tool to integrate the various data sources toward an evidence-based approach in policy making.

3. PolicsLab, a smart platform for technological foresight

Nowadays, economists and scholars of regional sciences need to make a big effort in order to: (i) map new data sources and assessing their potential and limits; (ii) develop rigorous and reproducible scientific methodologies for gathering knowledge; (iii) verify on a continue basis the robustness of the produced evidence.

Starting from the analysis of S3 approach, an interdisciplinary working group has started developing an "intelligent" platform for supporting policy maker in performing technological foresight, by means of the integration of structured and non-structured sources. Known as PolicsLab, this platform offers an integrated overview of all relevant data and indicators in the domain of research and innovation policies in Italian regions.

The web platform consists of an integrated database containing information from traditional sources, represented by official statistics (such as Eurostat, Istat, etc.), available in open or linked open data format (e.g. indicators relating to R&D in the private sector, etc.). The tool also includes data retrieved from sources, previously rarely considered, such as RIS3 thematic platforms, web portals, blogs, patents, projects and scientific publications as well as data from social media. In particular, PolicsLab supports policy-makers in: (i) technological foresight, by defining medium/long term investment priorities on the basis of interconnections between research and business as well as facilitating the entrepreneurial discovery process; (ii) a horizon scanning, by means of information about innovative technological trajectories and emerging markets worldwide. PolicsLab provides users with easy-to-read dashboards containing interactive graphs about indicators useful for the analysis of

changes affecting regional innovation context, medatada and an interactive storytelling feature, for guiding them through the report interpretation.

The beta version of the platform, currently under development and testing stages, is set to provide a three-module structure. Each module will meet a specific information requirement.

The **Innovation System** Module allows the exploration of data and indicators available from official sources, useful to profile the innovation landscape at regional, national and international level (Table 1).

Table 1. Innovation System module

Policy maker's question	Indicator	Source
What are the characteristics of the regional innovation system?	R&D expenditure as a % of GDP	Istat
How is the region positioned respect to other European regions?	Regional Innovation Scoreboard	Eurostat

Source: Our analysis

The **Horizon Scanning** Module enables the consultation of structured forecast information on trends and market trends from main international sources (e.g. market size forecasts for specific technologies, VC investment trends by segment, etc.). The objective is to assess the consistency of technological trajectories identified by the regional context with recent and future developments on a global scale (Table 2).

Table 2. Horizon scanning module

Policy maker's question	Indicator	Source
Which are the trends in digital technologies?	Market size of technology drivers	Statista
What are the investment trends?	Investments and average number of private operator deals in innovative sectors	CB insights

Source: Our analysis

The **Foresight Module** allows the exploration of data and indicators, both from traditional as well as unstructured sources, enabling technological foresight exercise to define priority investment areas. Our research activity has been mainly focused on analysing research and innovation investments funded by cohesion policy (National Operational Programme on Research and Competitiveness 2007-2013) in order to gather useful insights for informing at a more granular level the entrepreneurial discovery process.

Table 2. Foresight module

Policy maker's question	Indicator	Source
In which research areas is the regional academic system more specialised?	University publications by Web of Science category and by S3 area	Thomson Reuters; Scopus
What is the degree of cooperation and the shape of networks in research and innovation projects? Which are the most relevant actors?	Research project's network	Open data NOP R&C; OpenCoesione

Source: Our analysis

4. The experimental use of new data sources

One of the main challenges the system aims to tackle is related to the priority economic activity concept, which differs from the economic sector notion and mainly refers to application domains of specific technologies and/or inventions. The traditional classification of economic activities loses of significance because, it does not allow the identification of priority application domains, even at the highest level of detail. For this purpose, the integration of sources, such as the textual corpus of patent databases or business websites, which so far have been majorly undervalued, is essential. To understand what companies are really specialised or investing in, we used the information kit on their portals, in order to enrich the analysis of sectoral crossovers emerging from collaborative projects funded by cohesion policy.

We assumed that companies share information about their specialisation and investment areas in the *About* section of their portal. So we scraped the web sites of all the beneficiaries of the National Operational Programme on Research and Competitiveness 2007-2013 located in Calabria..

During our analysis aimed to define an appropriate corpus for applying text analysis techniques, we come up with three important findings: (i) The universities *About* section was too generic and not very informative for our analysis purpose, so we have excluded them from the analysis. (ii) Many beneficiaries, in general small companies, presented a very short and *About* with respect to other with a very detailed list of specialisations and activities. (iii) For several Ateco[1] subcategories the number of beneficiaries was not sufficient for granting an adequate significance level of the results

As a consequence, we performed traditional cleaning and processing analysis on a corpus composed by the beneficiaries belonging to the J-61.00.00 – Computer programming activities and M-72.19.09- Other research and experimental development on natural sciences and engineering for a total of 32 documents.(i.e. lowering cases, removing stopwords, etc.). Since we were interested in analysing not only single words but also more complex verbal expression, we tokenized the text in unigram (single words) and bigram (pairs of words). Consequently we realised a Document Term Matrix and performed a topic modelling algorithm, in order to find more detailed specialisation in the companies' activities, which led us to associate to predefined Ateco subcategories the following priority domains (Table 3).

[1] Ateco is the standard classification of economic activities used in Italy by national statistical agencies which derives directly from ISIC and NACE.

Table 3. Priority domains per Ateco subcategory

Ateco subcategory	Priority domains
J-62.01.00	Big data Internet of Things eCommerce Data analytics Business intelligence Smart home
M-72.19.09	Manufacturing Billing systems Healthcare Biomedical field Agrifood Cybersecurity Fashion ePayments Security management System integration

Source: Our analysis on web scraped data

Starting from the results of our analysis we managed to detail the network of cooperation in R&D projects among the beneficiaries pertaining to the selected Ateco subcategories. This exercise aimed to define cross sectoral priority domains in order to highlight the regional specificities concerning with the innovation ecosystem. Possible emerging models of recombination and diversification refer to:

- Manufacturing and smart home
- Biomedical field/health care and big data/IoT
- Agrifood and data analytics/big data/e-commerce
- e-commerce and e-payments/billing systems/cyber security

5. Conclusions and future developments

This paper aimed to represent the first research results for defining an informative system to support policy makers in the definition and revision of Smart Specialisation Strategy. Regional policy makers are called to identify some priority domains which could determine important structural changes in the whole business ecosystem. The first analysis on some of the regional strategies shows a substantial continuity with previous policies, with a lack of metrics, indicators and appropriate data and classifications, policy makers are not able to assess the innovation progress and evaluate transformations.

PolicsLab represents a first attempt of answering to policy makers' needs, integrating traditional and non-traditional sources and implementing interactive features useful for measuring and forecasting regional developments related to innovation.

In the coming months, with regards to the experimentation on new data sources, the working group's efforts will focus on improving current research output through the analysis of research projects abstracts in order to enrich the information derived from the companies' *About* sections, validating the results on a wider sample of companies, and defining a classification algorithm in order to associate each beneficiary to a specific priority applications domain.

5. References

Boyd D., Crawford K. (2012), Critical questions for Big Data, in Information, Communication & Society. 1-18.

Caramis A., Lucianetti L. F. (2015); Le Strategie di Smart Specialisation delle Regioni Italiane. Paper presented at XXXV Italian Conference of Regional Sciences, Arcavacata di Rende (CS), 14-16 September.

Crick T., Mateos-Garcia J., Bakhshi H., Westlake S. (2016), Innovation Policy-Making in the Big Data Era, in Data for policy: book of abstract.

Einav L., Levin J. (2014), The data revolution and economic analysis (downloaded from 093.145.166.179 on July 19, 2017)

Foray D. (2015), Smart Specialisation: Opportunities and Challenges for Regional

Hausmann R., Rodrik D. (2003), Economic Development as Self-Discovery, Journal of Development Economics, 72, pp. 603–633.

Kleibrink A. (2016), Data-driven innovation policies in Europe: mapping methods and sources, in Data for policy: book of abstract

Poel M., Schroeder R., Treperman J., Rubinstein M., Meyer E., Mahieu B., Scholten C., Svetachova M. (2015), Data for policy: A study of big data and other innovative data-driven approaches for evidence-informed policy making.