26th Meeting of the Wiesbaden Group on Business Registers
- Neuchâtel, 24 – 27 September 2018

Bertrand Loison[1]
Swiss Federal Statistical Office

Session 5 - New data sources, especially on enterprise groups

Swiss Federal Statistical Office - Data Innovation Strategy

Abstract

The development of big data and analytics, in particular data science, are set to be a significant disruptive innovation in the production of official statistics offering a range of opportunities, challenges and risks to the work of national and regional statistical institutions.

In November 2017, the Federal Statistical Office (FSO) published its "Data Innovation Strategy". This represents a first response by the FSO to the challenges brought about by the digital revolution. With this strategy, the FSO positions the importance of complementary analysis methods to increase and/or complete statistical production.

The paper describes the content of the strategy in particular the role that we attributed to administrative data and registers, provide a synoptic overview of the roadmap to implement the strategy, but also the new competencies and skills that we need to acquire.

Finally, we will present the five pilot projects which are currently trying to implement these complementary analytics methods (e.g., predictive analytics using approaches from advanced statistics, data science and/or machine learning) to existing (or traditional) and/or new (or non-traditional) data sources.

Keywords: Big Data, Data Science, machine learning, deep learning, administrative data

[1] Prof. Dr. Bertrand Loison is Board Member, Vice Director and Head of Division Registers at the Swiss Federal Statistical Office (SFSO) in Neuchâtel, Switzerland, nominated member of the national planning committee E-Government Switzerland and Swiss representative to the UN Global Working Group on Big Data for Official Statistics (UNECE). In addition, he is leading the "New Data Sources" working group at the SFSO that aims to implement the SFSO Data Innovation Strategy into the current statistical production. He is primarily interested in the change process that national statistical offices are facing in their use of new data sources. He is also Honorary Professor of Information systems at Haute école de gestion Arc - University of Applied Sciences and Arts Western Switzerland (HES-SO), School of Management.

I. Introduction

Even twenty years ago, information was difficult to access and had to be collected for particular purposes via surveys. The information obtained was therefore unique as there simply was no other alternative.

Over the past few decades, data collected by the public administration have become increasingly accessible for statistical purposes. The collection of statistical data using questionnaires was completed, and then replaced by administrative data sources when this became possible. In the context of all these developments, the information provided by the NSIs remained unique. In particular, the option of combining data from a number of sources has made official statistics even more precise.

The emergence of new data sources such as Big Data generates a potential benefit for the NSIs if we refer to work carried out by the Global Working Group on Big Data for Official Statistics [1] at a global level and in Europe by the ESSnet on Big Data [2]. It also, however, creates several substantial challenges.

II. Change of model

The main challenge faced by official statisticians in using Big Data is the truthfulness of data, something that is key to confidence in data. It includes the reliability, soundness, validity and quality of data as well as the transparency of the data production process.

Methodology represents another considerable challenge. Many Big Data type sources such as messages from social media are made up of observation data and are not deliberately designed for data analysis. Therefore, they do not have a target population, nor structure or quality. This is why it is difficult to apply **traditional statistical methods based on the sampling theory.**

The non-structured nature of these sources makes it difficult to extract information of statistical importance. For many Big Data sources, the interpretation of data and their relationship with social trends to be observed are far from evident. For example, public messages on a social network are, to a certain degree, a reflection of a general feeling in which the boundaries are far from clearly defined. Furthermore, if such data are to be used for statistical purposes, it is necessary to establish the representativeness of people who write public messages on a social network compared with the general population. The population of people using social media is also likely to change over time. Another major challenge is the volatility of this type of data, given that official statistics often take the form of time series analyses.

For the NSIs, the question is thus to know how the quality of official statistics can be guaranteed if they are all or partially produced from Big Data. The use of Big Data will incite changes to the model and greater use of **complementary analytics methods** (e.g. predictive analysis using advanced statistical techniques, data science and/or machine learning).

The FSO – a member of the Global Working Group on Big Data for Official Statistics since 2017 – identified these challenges and provided an initial response to these by publishing its strategy on data innovation in November 2017 [3]. The rest of this document presents this strategy and the pilot projects.

III. Primary and secondary Data

In the context of data innovation, the FSO divides data sources into primary and secondary data.

Primary ("made" or "designed") data have been collected – and designed – by the FSO for statistical purposes to explain and check the validity of specific existing ideas, i.e. through the operationalisation of theoretical

concepts. Learning from such primary data is known as primary (or top-down, i.e. explanatory and confirmatory) analytics. The corresponding analytics' paradigm is "deductive reasoning" that starts with an idea or theory ("idea first"). Examples of primary data are traditional data sources like censuses and surveys that have been collected by the FSO for statistical purposes.

In contrast, secondary (observational or "found" or "organic") data have been collected – and designed – for other reasons, often without FSO supervision, and could be used to create new ideas or theories. Learning from such secondary data is known as secondary (or bottom-up, i.e. exploratory and predictive) analytics. The corresponding analytics' paradigm corresponds to "inductive reasoning" that starts with data ("data first"). Examples of secondary data are non-traditional data sources such as FSO internal and external register data, administrative data, and other digital data from devices, machines, sensors, satellites, drones and social media. As secondary data sources were not designed to be used directly in official statistical production systems, they need to be made fit for purpose for statistical inference, i.e. so that conclusions can be drawn from them for the purpose of official statistics by deductive reasoning.

To do so, secondary data can be further classified into identifiable and non-identifiable data. Identifiable data can be meaningfully associated with a single unit at a given place and time, such as an individual, institution, product or geographical location (e.g. register data, administrative data, satellite imagery, geospatial information and product barcodes). Non-identifiable data cannot be made identifiable at any such level (e.g. Google trends data, Twitter feeds and other forms of social media). Identifiable secondary data could be made fit for purpose for statistical inference if their veracity has been successfully assessed (as is the case with the FSO's current use of its internal register data), whereas non-identifiable secondary data are of limited use for statistical inference because it is not possible to assess their veracity.

IV. The inductive-deductive reasoning cycle

It is important to note that the two approaches of analytics (i.e. inductive and deductive reasoning) are complementary and should proceed iteratively (Image 1) and side by side in order to enable continuous improvement and data-informed decision and policy making. This implies that the analytics methods currently used at the FSO will still be needed together with complementary analytics methods.

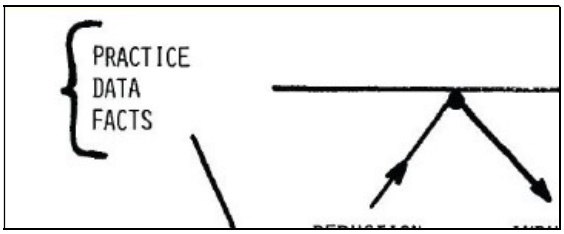

Image 1: Inductive-deductive reasoning cycle[2]

For example, the information (ideas) generated by inductive reasoning cannot explain if and why these discoveries are useful and to what extent they are valid. The confirmatory tools of deductive reasoning will be needed to confirm the discoveries and evaluate the quality of decisions based on those discoveries.

V. Strategic objectives

The first strategic objective of the data innovation strategy is directed towards creating awareness that data innovation is a main strategic issue.

Moreover, the first strategic objective will be progressively revised and updated depending on how the steps defined below are fulfilled.

[2] Box, G.E.P. (1976). Science and statistics. Journal of the American Statistical Association, 71, 791-799.

Strategic objective 1: Develop data innovation guidelines and investigate the feasibility of the application of complementary analytics methods to existing (or traditional) and/or new (or non-traditional) data sources, along with the goal of augmenting and/or complementing any existing basic statistical production for which data innovation makes sense.

Data innovation requires a paradigm shift by combining deductive and inductive reasoning that will necessarily lead to a change in the production and communication of official statistics.

It is important to note that no common generic methodological approach exists for the procedures outlined above for using data innovation in official statistical production. The challenges and opportunities are specific to each data innovation application and type of data. As such, current FSO process frameworks and process models may need to be adapted and extended to enable data innovation.

This means that communication on these issues is key. This leads us to the second strategic objective.

Strategic objective 2: Develop and implement FSO internal and external communication measures to increase awareness of the added value of data innovation in official statistics and the related paradigm shift.

VI. Roadmap

The first step is the application through FSO internal pilot projects of complementary analytical methods to existing (or traditional) FSO internal primary data sources and already matched identifiable secondary data sources (if such exist).

It is about augmenting and/ or complementing existing learning from data, i.e. to use such data sources in new ways to gain practical experience and make an inventory of existing challenges, resources, skills and technologies in order to perform data innovation and investigate "quick-wins".

If feasible, a subsequent second step could then be to complement and/or augment existing statistical production at the FSO with data innovation generated from the application of complementary analytics methods to additional secondary data already in use at the FSO.

A subsequent third step could be to apply complementary analytics methods to only new – hitherto unused at the FSO – secondary data to investigate and produce new statistical information and statistics in particular statistical domains.

VII. Preferred data source sequence

The preferred data source sequence for the FSO's data innovation strategy is describe below.

Preferred data source sequence

1. FSO internal primary data sources and already matched identifiable secondary data sources (if they are already used in FSO's current statistical production);

2. additional secondary data sources already in use at the FSO;

3. new – until now unused at the FSO – secondary data sources.

The aim of this sequence is to raise efficiency, reduce costs and minimise the administrative burden on businesses and individuals. Moreover, this sequence should allow for a better understanding of methods, technologies and tools, without adding unnecessary complexities as would be the case by starting with the third step, which might raise legal, technological (IT) and related capacity problems at the same time. Furthermore, greater experience of complementary analytics

methods will sharpen the skills needed to find new data sources adapted to the specific aims of official statistical production.

VIII. Pilot projects

As mentioned above, the FSO has started five pilot projects that are currently being implemented.

Pilot project 1: One of the FSO's key tasks is the correct coding of the economic activities of enterprises. This project strives to automate the coding of the economic activities of enterprises using machine learning methods applied to data already available within the FSO (data from surveys, descriptions in the commercial register, key words, explanatory notes for classifications etc.) with a view to supporting the production units.

Pilot project 2: The FSO's land use statistics form an invaluable tool for the long-term observation of the territory. This project involves learning and mastering the use of artificial intelligence (IA) technologies to eventually automate (even partially) the visual interpretation of aerial images in order to detect and classify changes.

Pilot project 3: The project evaluates the potential of the small area estimation method for the Job Statistics. The aim is to produce reliable estimates of the total number of jobs and FTEs for cantons, major towns and NOGA[3] levels that were not anticipated in the sample plan.

Most economic statistics are based on sample surveys. The survey plan is generally established at the two digits of the classification of economic activities (NOGA2) and the major regions (NUTS 2). There is also considerable demand for estimates of results at detailed levels (cantons, communes, NOGA3, NOGA4, etc., MIGS (Eurostat's main industrial groupings), employers' associations and professional organisations. In these cases, the FSO's usual response is to suggest an increase in sample numbers in co-financing. This practice is costly for clients but also goes against aims to reduce the burden on enterprises. The aim of the present project is to evaluate the small area estimation method's potential to respond to these needs without increasing sample sizes.

Pilot project 4: Statistical offices carry out plausibility tests to check the quality and reliability of administrative data and survey data. Data that are either clearly incorrect or seem at least questionable are sent back to data suppliers with a correction request or comment. Until now, such plausibility tests have mainly been carried out at two different levels: either through manual checks or automated processes using threshold values and logic tests.

This process of two-way plausibility checks involves a great deal of work. In some cases, staff are required to manually check the data again, in other cases rules are implemented with further checks sometimes called for. This rule-based approach has developed from previous experience but is not necessarily exhaustive nor always precise. Machine learning could help to ensure faster and more accurate checks.

This approach would rely on an algorithm using historical data at first. Based on a previous data analysis, a target variable can be defined that should be able to be predicted by the algorithm. Only then can the algorithm be used for the prediction. As the final stage, the predicted and actual values of the target variables are compared and the predictive accuracy can be evaluated. Finally, a feedback mechanism is also used to send an automatic explanation to data suppliers.

[3] General Classification of Economic Activities (NOGA) takes into account both the framework conditions set by the Statistical classification of economic activities in the European Community (NACE, rev. 2) and the needs of various interest groups in Switzerland.

Pilot project 5: The grouping of typical prospective trajectory patterns concerning the receipt of benefits in the social security system and employment and the estimation of group affiliation through the use of individual variables and retrospective trajectory data applying a machine learning approach.

IX. Conclusion

The new data sources (Big Data) is expected to have a considerable impact on organisations for which data production and analysis are key. The National Statistical Institutes (NSI) are no exception to this.

Official statistics are often taken for granted. However, where public confidence is missing, society lacks an important foundation for pragmatic discussion and the creation of public policies based on convincing evidence. Professional standards and norms thus play a key role in ensuring trust in official statistics [4], [5], [6].

Big Data should therefore only be considered in the production of official statistics while respecting the scientific code of ethics.

Notes

[1] https://unstats.un.org/bigdata/

[2] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

[3] https://www.bfs.admin.ch/bfs/fr/home/actualites/quoi-de-neuf.gnpdetail.2017-0673.html

[4] https://www.eda.admin.ch/dea/en/home/bilaterale-abkommen/ueberblick/bilaterale-abkommen-2/statistik.html

[5] https://www.bfs.admin.ch/bfs/fr/home/ofs/engagement-qualite.html

[6] https://www.bfs.admin.ch/bfs/fr/home/services/appariement-donnees/generalites.html