United Nations

ECE/CES/2019/23

Economic and Social Council

Distr.: General
9 April 2019

Original: English

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Paris, 26-28 June 2019
Item 2 (b) of the provisional agenda
New data sources – accessibility and use
Session 2: Skills needed to use new data sources

New competencies needed to work with big data – lessons learnt so far

Note by the Statistical Office of Slovenia

Summary

This document describes lessons learned in the last six years by the Statistical Office of Slovenia regarding competencies needed to work with big data. It begins by listing the obtained competencies, the most important being those related to characteristics of new data sources, advanced methods for their processing, and communication with data holders. It continues with the ways the competencies were obtained, with learning by doing as the most important one, followed by exchanging experiences within the international statistical community. It concludes with a look into the future and mentions competencies retention as the most challenging task.

This document is presented to the 2019 Conference of European Statisticians seminar on “New data sources – accessibility and use”, session 2 “Skills needed to use new data sources” for discussion.

I. Introduction

1. Big data and its use have been one of the hottest topics in official statistics in recent years. The Statistical Office of Slovenia (SURS) has been dealing with obtaining, processing and analysing big data sources for more than six years. In this period, one project was concluded with the inclusion of a big data source, i.e. price scanner data, into regular statistical production. Other projects include analysing the potential of mobile phone location and transaction data, online job vacancy adverts, online prices, satellite images, financial transactions data and traffic sensors data. There are common issues to all of them, e.g. access, ethical considerations, methodology, technology and quality. There is, however, also one that seems the most horizontal, relevant for all projects and for all statistical phases, namely skills and knowledge needed.

2. This document elaborates on lessons learned by SURS regarding skills and knowledge when conducting big data projects. It begins by listing what new knowledge or skills were obtained, it continues with how they were obtained and concludes with the outlook for the future. In the paper, the term “competencies” is used to denote knowledge and skills.

II. What was learned

3. While implementing big data projects at SURS, new competencies were developed and existing improved or strengthened. Some competencies were new to some staff and some statistical areas but not to others; for example, some subject matter statisticians started to use more advanced IT tools during the projects, while, on the other hand, some IT experts are now more familiar with statistical methodological knowledge. The spread of already existing competencies within the office was an important effect of the work. Further on, a lot of competencies are not new to the “outside world” but are new to SURS; they have existed for several decades but have only recently started to get a more prominent role in Slovenian official statistics (e.g. data modelling). The most important new or improved competencies acquired by SURS with big data work include:

(a) More data sources are known. New data sources were recognized, and their characteristics studied (owners/holders, frequency, quality, content, accessibility). Examples include information generated by retailers (prices, quantities, codes of products sold), information generated by mobile phone usage (locations, transactions), systems for earth observations, information generated by traffic counters, etc.;

(b) There is more knowledge about legislation governing the existence of different types of data, mainly legislation in relation to the protection of personal data and legislation regarding electronic communications;

(c) More data collection modalities are used. For example, price scanner data are collected by accepting data files directly from enterprises. Such an approach was in the past used only for collecting data from public administration but not from enterprises. Further on, web scrapping techniques are now used to obtain data on prices in web stores and job adverts on web employment portals. For this to happen, it was necessary to acquire knowledge about the composition of internet sites, functioning of browsers, programming of specific and general web spiders and crawlers, and optimization of tags’ searching according to the HTML structure;

(d) New approaches for dealing with data providers were developed. Traditionally, enterprises were expected to provide data according to the prescribed form and content of data. Now, with price scanner data, each enterprise is providing data in a way and in a form which is adapted as much as possible to its possibilities and the way of operating. SURS thus learned how to be even more data supplier friendly;

(e) There is more knowledge about enterprises: what is their business model, what are they concerned about (revealing of the business secret, revealing of customers data, etc.), how do they see the public administration, their business and legislative environment, how do they function, what are they interested in;

(f) Change in the fundamental paradigm of statistics production was recognized and is gradually being accepted. Traditionally, a statistician’s work starts with the output. Once users’ needs are known and output defined, considerations about the data sources and best ways for data collection and processing start (from output to input). With big data, there is usually the opposite approach. First, input data sources are available, and a statistician only then starts to analyse whether there is any statistical value in them (from input to output). With this new way of approaching statistics production, there are more competencies in SURS for the selection of the most “talkative” data, for data integration and for deciding on research direction;

(g) Thinking in a broad perspective is more common. With big data sources, SURS is learning to be open-minded when considering possible uses. For example, the question about possible uses of satellite images in statistics production was discussed extensively within SURS with broad participation of staff who were invited to present their proposals;

(h) New methodologies were mastered. With the price scanner data, the methodology for the compilation of the consumer price index for food and beverages based on representative products was abandoned and the full coverage was introduced. With this, the compilation of new types of indices was also learned;

(i) Knowledge of data integration practices is upgraded. Integration of data from different sources is not new to SURS as administrative records have already been used for several decades. The novelty is the integration of directly collected and big data inputs;

(j) Approaches for optimizing programmes for processing large amounts of data and performing processors’ intensive operations are mastered. For example, a lot of effort was put into understanding and preparing programmes in a multithreading way which enables faster data collection, analysis, processing and aggregation. Multithreaded methods are used for analysing, amending, changing and aggregating traffic counters data; due to their amount (tens of gigabytes) standard loops would be much more time-consuming. Multithreaded programmes are used also for scrapping and storing of websites (hundreds of thousands of them);

(k) Knowledge on text mining and machine learning, used for the structuring of unstructured data and for classification of data with different algorithms, is being acquired. The methods are tested for classifying consumer products according to the COICOP classification, for matching business units and for structuring web published job adverts;

(l) Data modelling together with machine learning for forecasting and compiling statistics is much more familiar. This includes knowledge for understanding and developing of models, and comprehending limitations and implications of combining models. For example, forecasting economic aggregates with linear regression using traffic counters data and turnover was introduced. The method for compiling job vacancies from unstructured data using logistic regression, linear regression, nearest neighbour method and the AdaBoost ensemble method was developed;

(m) New programming tools and libraries are in use, use of existing ones is more common; for example, technologies (.NET, ASP.NET, ML.NET frameworks), tools (MS Visual Studio, SQL Server Data Tools - SQL Server Integration Services), languages (C#, R, Python), databases (MS SQL Server, Oracle), Python libraries (TensorFlow, Scikit-Learn, Keras), etc.;

(n) Due to high level of instability in tools and business models of their providers (e.g. some tools are for some time available as an open source, then suddenly as fully commercial ones; tools were available yesterday but no longer today), SURS learned how to be more flexible and adaptive to such changes so that they do not cause major disruptions in the work;

(o) There is more knowledge on coordination and connection of all above-mentioned competencies, and recognition that it is the most useful for staff to have several different competencies (for example coding, mathematics, statistical methodology).

III. How it was learned

4. From an organisational point of view, SURS approached big data in the same way as other development work. Development work is considered to be an integral part of each statistician's work. There is no special unit that would deal only with innovation, progress, improvements, etc. SURS believes that separating development and regular work leads to segregation between them, to the creation of stove-pipes, to the alienation of "development statisticians" from practical issues and "regular statisticians" from progress, and to additional problems in introducing changes. The same holds true for working with big data. This work is now organized within formally established projects with the participation of employees from different organizational units.

5. More specifically, competencies needed to work with big data were acquired (and are still being acquired) in different ways, namely:

(a) Learning by doing. Conducting projects and practical work is by far the most important way to gain competencies. Staff from different units (statistical methodologists, IT experts, mathematicians) are included in project teams and their good cooperation is crucial for efficient exchange of ideas, knowledge, etc. In addition, some statisticians participate in several projects, which is beneficial for transferring solutions from one project to another;

(b) Participation in international activities: UNECE Big Data Task Force, ESS (European Statistical System) Task Force on Big Data, ESSNet BigData1 (web scraping job vacancies, early estimates, methodology), ESSnet BigData2 (job vacancies, financial transactions), ESS Expert Group on Scanner Data;

(c) Study visits at statistical offices having more experience using big data sources, studying their documents and practices. These were especially used in the area of using scanner data;

(d) Participation in international conferences, workshops, courses, etc., either physical or virtual (webinars, web courses, web workshops). For example: Big data: effective processing and analysis of very large and unstructured data for official statistics; The use of R in official statistics; Can a statistician become a data scientist?; Introduction to big data and its tools; Big data sources – web, social media and text analytics; Advanced big data sources: mobile phone and other sensors; Automated collection of online prices; Macroeconomic forecasting; Data science: linear regression; Symbolic data analysis; Data analytics workshop for official statistics; Machine learning and econometrics; Big data UN global conference, etc.;

(e) Self-study using resources available on the internet (e.g. manuals, handbooks, videos, software developer platforms, forums);

(f) Attending lectures at the university. SURS's staff attended several lectures within full-time study at the Faculty of Computer and Information Science relevant for processing big data;

(g) Cooperation with institutes. SURS cooperates mainly with the "Jožef Stefan" Institute, predominantly in the area of advanced algorithms for language-based text classification;

(h) Knowledge is also obtained during studies, i.e. before a person joins SURS. Currently, the most relevant are studies of computer science, information science, mathematics, physics, social sciences and statistics. Given that a person has a basic knowledge and understanding of programming, object programming and data structures, knowledge on machine learning, data mining and text mining can be obtained relatively quickly, at least in a basic form.

6. One of the more important lessons learned about training is that there is no point in stock-piling knowledge, just in case someone will need it. Earlier, many members of the staff attended courses on big data but then either did not use it or left the office before being included in the big data projects. Now there is a possibility to gain knowledge in any of the above-mentioned ways when a person is assigned to the work where such knowledge is needed.

IV. What next

7. For the future of competencies, three issues are worth mentioning.

8. First, missing competencies. The most pressing need is to upgrade competencies in the areas of more complex text mining, understanding machine learning, data modelling, and integrating big data. In addition, there is a new need to gain competencies for data processing using distributed computing with clustering, neural networks, decision trees and random forest. The competencies will be acquired with all the ways listed above.

9. Second, the distribution of competencies. So far, only a limited number of staff were included in the big data work and gaining competencies for it. For the future it will be necessary to assure that competencies are distributed among other staff as well. This will become easier with each successful integration of a new data source into the regular statistical production, i.e. with each successful project the following steps will be easier. Thus, for managers one of the biggest challenges will be how to motivate their staff for embracing the changes that digitalization is bringing to official statistics.

10. Regarding this issue, very rapid advancement of these sciences needs to be mentioned as well. For an individual, it is already difficult to be knowledgeable in all branches of machine learning and text mining. In the private sector, a process of specialization of these professions has already started. Nevertheless, for an expert it is still necessary to be familiar with at least basics of all elements and to follow new developments. It is thus to be expected that also in official statistics there will be more specialization of data scientists. This will result in even more experts needed to cover all elements of big data work.

11. Third, the retention of competencies. Due to the high demand for data scientists and subsequent high remunerations in the private sector, it is quite difficult to retain staff with relevant competencies. At least in the medium term, staff retention will be one of the main, if not the main, human resource challenges at SURS.

V. Conclusion

12. Willingness to deal with new topics and researching new ideas, not being afraid of failures, being curious, having a positive attitude to constant changes and enthusiasm for working with data are characteristics that are optimal to work with big data. Above these, competencies need to be built to be fully equipped for achieving results. The two most general lessons about competencies learned by SURS in conducting big data projects are the following. First, with big data not that much is different from any other performed activity; there are some competencies that are slightly more specific to big data sources, but a large majority of them are the same. Second, acquiring competencies is not a serious problem. It requires intellectual effort, time and money but otherwise there are ample opportunities for gaining knowledge and skills. The really serious issue is also the most classical one: how to identify relevant new and existing employees, how to motivate them to gain new competencies and assume new responsibilities, and how to retain them in order to keep competencies within the office.