United Nations
Economic Commission for Europe
Statistical Division

CONFERENCE OF EUROPEAN STATISTICIANS WP.13
Workshop on Statistical Data Collection: Riding the Data Deluge 20 April 2015
29 April – 1 May, Washington D.C., United States of America

Usage of new data sources at SURS

Boro Nikić, Tomaž Špeh and Zvone Klun (Statistical Office Republic of Slovenia)

Abstract and Paper

New data sources, ranging from diverse business transactions to social media, high-resolution sensors, and the Internet of Things, are creating a digital tidal wave of big data. The statistical community has recognized the potential of new data sources which have also been described as a transformative tool for official statistics provided that these data are collected, processed and integrated accordingly with careful attention to protection of privacy. Integration of these data can play an important role in improving the accuracy, timeliness, and relevance of statistics at a lower cost than expanding existing data collections. Thus innovative and automated approaches have to be implemented in order to successfully integrate new data sources into existing solutions. SURS has recently started several projects in order to discover the usage of new data sources. New approaches for collecting price data started to be introduced (e.g. scanned data, scrapped data and data collected using mobile devices). Mobile positioning data started to be used to enhance statistics about population (active, retired, etc.), distribution regarding time and location, and population mobility. Scraped job vacancies data started to be used as a quality measure for traditionally collected data. This paper provides an overview of ongoing projects and related technological developments in this field at SURS.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection: Riding the Wave of the Data Deluge
Washington, 29 April – 1 May 2015

Session 2: New Tools and Methods

USAGE OF NEW DATA SOURCES AT SURS

Prepared by Boro Nikić, Tomaž Špeh, Zvone Klun, Statistical Office of the Republic of Slovenia

I. Summary

1. New data sources, ranging from diverse business transactions to social media, high-resolution sensors, and the Internet of Things, are creating a digital tidal wave of big data. The statistical community has recognized the potential of new data sources, which have also been described as a transformative tool for official statistics provided that these data are collected, processed and integrated accordingly with careful attention to privacy protection.

2. Integration of these data can play an important role in improving the accuracy, timeliness, and relevance of statistics at a lower cost than expanding existing data collections. Thus innovative and automated approaches have to be implemented to successfully integrate new data sources into existing solutions.

3. SURS has recently started several projects to discover the usage of new data sources. New approaches for collecting price data started to be introduced (e.g. scanned data, scrapped data and data collected using mobile devices). Mobile positioning data started to be used to enhance statistics about population (active, retired, etc.), distribution regarding time and location, and population mobility. Scraped job vacancies data started to be used as a quality measure for traditionally collected data.

4. This paper provides an overview of ongoing projects and related technological developments in this field at SURS.

II. Background

5. Statistical authorities are under pressure to produce data faster and at lower cost, to become more responsive to users´ demands, while at the same time providing high quality output. One way to fulfil this is to make more use of already available data sources. The Statistical Office of the Republic of Slovenia (SURS) has been moving towards an increased use of administrative data sources for statistical purposes as a substitution and/or as a complement to information previously collected by sample surveys. Administrative data sources are similar in structure, but are not the result of a sample survey. They are typically collected in support of some administrative process. The variables are not chosen or defined by the NSI, as opposed to the variables in sample surveys. Alternative data sources are becoming increasingly important. They are not comprised of a set of records directly corresponding to units in a target population. These kinds of data sources often register events. Such

data can be generated as a by-product of some process unrelated to statistics or administration. Since these data files are often much larger in size than sample data or administrative registers, the term ‘big data’ is sometimes used in these cases.

6. Within SURS there is a growing need to modernize statistical data collection and processing. Key words are decreasing costs and administrative burden, and increasing efficiency and flexibility. We are investigating ways to disclose all kinds of new data sources that become available through the global use of modern technologies such as the Internet, mobile phones, automated scanning techniques, etc. At the same time, new technologies offer broad new possibilities to modernize statistical data processing.

7. SURS collects data in different ways and in different formats. To facilitate the data collection in business surveys, web collection system ESTAT has been developed. For some administrative data sources with frequent data exchange, the direct database replication method is used; the majority of others come into the office through the Government Secure Data Exchange Hub or SFTP server. Recently, the need for supporting efficient acquisition of alternative data sources (for example scraped data, mobile data, scanned data) has emerged. The new technologies bring with them new challenges in such areas as efficient processing, integration into a multi-source environment, privacy and security issues, and cooperation with partners outside official statistics.

8. Statistical data processing has always been a demanding, time consuming and consequently quite expensive task. To overcome or at least reduce the gap between the above mentioned demands, in recent years a lot of effort has been put into the rationalization of statistical data processing. As a consequence, SURS is developing a modernized statistical data processing system consisting of a few modules which aim at “covering” the different parts of the statistical process (e.g. data validation, data correction and imputation, aggregation and standard error estimation, tabulation).

9. One of the challenges in this process of change is the integration of sources and collection modes and following to that the standardization of collection methods and technologies. Besides this, a second and probably even bigger challenge is the integration of the collected data into the statistical production: How to make optimal use of all available data sources (existing and new)?

10. SURS has recently started several projects to discover the usage of new data sources. New approaches for collecting price data started to be introduced (e.g. scanned data, scrapped data and data collected using mobile devices). Mobile positioning data started to be used to enhance statistics about population (active, retired, etc.), distribution regarding time and location, and population mobility. Scraped job vacancies data started to be used as a quality measure for traditionally collected data.

III. Description of experimental projects

A. Modernisation of consumer price collection and compilation

11. In the last few years, one of the challenges and initiatives to insure the quality of data and harmonisation of price statistics is to explore new technical and methodological solutions for data collection and data compilation by using different methods and data sources. These activities are oriented in the direction of modernising of consumer price collection and production process by using different tools and data sources and by analysing and exploring different new ways of automated data collection.

12. SURS started with the activities in this field already within the projects “Food price monitoring tool methodological and practical improvements” (to be finalised by the end of August 2015) and “Multipurpose price statistics” (finished by the end of March 2014). Experiences gained have shown that there is still room for the methodological and technical improvement.

13. Within the project the following activities will be carried out:

14. Introduction of scanner data as a new approach to simplifying data collection and compilation of price statistics, aimed at the reduction of the burden of price collectors and retailers and the creation of a data warehouse for price statistics. This approach should generate more high-quality data which should also meet the European Union’s quality requirements. As already proved, the use of scanner data appears to be a promising source for the collection of prices for temporal and spatial price comparisons, in particular for food and beverage products. SURS has already started working on scanner data but, nevertheless, additional analysis and development of new and improved methodological and technical solutions in data collection and production process are required.

15. SURS is carrying out different analyses with regard to scanner data based on findings and recommendations already gained in this field with the view to find the most appropriate methodological solutions to integrate them into the existing CPI/HICP production system to produce high-quality statistical results.

16. The following activities are ongoing:

- To develop a system for secure scanner data acquisition,

- To continue with the inclusion of other retailers in the pricing sample and integration of scanner data in the production system (software updates),

- To continue with further work on the development of methods for integration, storing and processing scanner data (technical aspect); e.g. procedures for optimization of processes due to different data sources,

- To develop a common structured database due to different data sources (scanner data, traditional collected prices) for CPI/HICP, DAP and PPP purposes,

- To examine and develop methodological solutions for the compilation of price data; data cleaning, quality adjustments, calculation of average prices, distribution and calculation of weights, calculation of average prices and price indices,

- To assess the sampling frame (calculation of the quality indicators),

- To carry out analysis with regard to the quality aspect due to the implementation of the new data sources; parallel calculation of price indices based on two different data sources,

- To further investigate and study best solutions.

17. If the proposed actions of processing and storing scanner data produce results comparable with the existing quality of data, SURS will take into consideration the inclusion of the proposed solution in the regular data production process.

18. The use of electronic devices for price collection. Based on the past experiences and recommendations (MSs and Eurostat), SURS's staff will study the best way to develop efficiency in the price collection as well as in the compilation of prices by using electronic devices. SURS has considered different requirements to find the optimal solution for the selection of the electronic device for the purpose of price collection.

19. The following activities have been implemented:

- To review, examine and analyse the existing practice of MSs already using electronic devices for price collection (CPI/HICP and PPP purposes),

- To study requirements for the most appropriate device from the view of SURS's price collectors (size of the screen, the weight of the device, battery capacity, offline/online connection, etc.) and from the view of the IT developer (data storage, backup possibilities, security aspects, etc.),

- To develop software for management of the back-office processes; main functionalities should be as follows: management of interviewers, stores, products characteristics and price data,

- To develop software/application for data entry as the existing application is not appropriate for electronic devices; main functionalities of the data entry application should be as follows: editing, basic logical controls, visibility of historical data, insertion of comments, transmission of data, export of data (in agreed format), inclusion of new products,

- To introduce methodological solutions for price collection,

- To test and evaluate the application/solutions,

- To develop methods for integration of collected data into the existing production system.

20. The use and analysis of prices collected on the Internet. The process of automated collection of information from the web using a computer software technique is known as web scrapping. It is a process based on a computer software program which is used to browse the web in an automated, orderly and methodical manner with the aim to provide up-to-date information of all web pages visited and to convert contents (e.g. product descriptions) that are not well structured into well-formatted contents. Web scrapers are also known as web spiders, automatic indexers or web robots.

21. More and more prices of products and services can be found on the different Internet pages (for CPI/HICP and PPP purposes). Manual price collection from the Internet sites is very time consuming due to different reasons (large amount of requested and needed information for selected product or service), which could be solved by new solutions, e.g. use of web scrapping.

The following activities are ongoing:

- To review, examine and analyse the existing documents on web scrapping (available documentation from MSs, Eurostat, other available data sources),

- To identify the target product groups (focused on product groups where a large amount of data has to be downloaded, e.g. rental market, cars, etc., for CPI/HICP and PPP purposes),

- To identify relevant Internet sites (possibility of using web scrapping),

- To verify technical and legal aspects of the data source (where necessary set up contacts),

- To study specific IT software tools for web scrapping (possible acquisition and installation of specific IT software tools),

- To study specific IT software tools for parsing unstructured scraped data (possible acquisition and installation of specific IT software tools),

- To evaluate possible methodological issues related to Internet collections,

- To assess possible technical solutions for regular production,

- To examine advantages and disadvantages of the approach.

Based on the study, SURS will decide if it is appropriate to change the existing method of price collection with the new approach – web scrapping.

B. Scraped job vacancies data

22. Until 2014 SURS had produced the Job Vacancies (JV) statistics completely based on the administrative sources. The enterprises were obliged to report every job vacancy to the Employment Service of Slovenia. SURS produced quarterly statistic based on the data from the Employment Service of Slovenia. The statistics which have been published were totals of job vacancies on the domains defined by activity of units and domains defined by size of units.

23. In the framework of activities to reduce the administrative burden of enterprises, the Slovenian Government has adopted regulations that liberate private businesses from reporting job vacancies to the Employment Service of Slovenia. Consequently, SURS started to conduct the quarterly survey on the population of units from the private sector to ensure the reliability of statistics. The survey is costly and (again) presents the burden for the reporting units; therefore, SURS started to investigate alternative sources of JV data.

24. In 2014 the main Slovenian websites advertising job vacancies were identified. At the same time the various tools for web scraping of data were identified and tested. The main disadvantage of “web scraping tools” is that they could collect data only from very structured websites and the main disadvantage of the “scraped” data is that the data often do not contain the information about the location of the JV and activity of the enterprises. To collect this information, SURS tested the procedures of record linkage of scraped data with the Business Register. The aim is to assign the ID to units from the Business Register and consequently the information about the size and the activity of units.

25. Because the survey based on a sample started in 2015, the quality of the "scraped data" will be tested this year. On the other hand, the "scraped data" will be used to benchmark the quality of survey data and statistics. In the next step the various models based on "scraped" and survey data will be tested to combine all three sources (administrative, survey and web JV data) in the production of JV statistics.

C. Mobile positioning data

26. In the form of EU GRANT "Merging statistics and geospatial information in Member States" SURS started the project the main objective of which is to establish an IT environment (IT infrastructure and human resources) to test the possibility of acquiring and processing the data primarily produced by the mobile operators. The action should reveal the potential deficiencies that have to be considered when the integrated solution is to be established in the period after this action. Particular focus of the action is on acquiring additional experience and skills regarding the handling of such data regarding its nature, size and applicability to standard GIS processes.

27. The specific objectives of the action are:

- Improving the integration of geo-information and geo-reference into the statistical production process including the official registers

  The result should be the production of datasets in a format manageable with the existing GIS equipment thus improving the temporal dimension of the already available geo-statistical data and creating the possibility of geo-referencing beyond the "static" Register of Spatial Units, e.g. day/night time population distribution. Consequently, the geo-referencing potential of the registers (e.g. Central Population Register) will be automatically improved regarding the space-time accuracy.

- Establishing internal and external processes required for a continuous update of the geo-reference of official registers and administrative data

  Lessons learnt and systems established will reveal the nature of possible cooperation between SURS and the mobile operators after this action. The action should clarify the processes needed to be executed over a longer time by either of the involved parties. The update is feasible within the restrains of the possibility of managing the data from the mobile operators regarding the time-space accuracy.

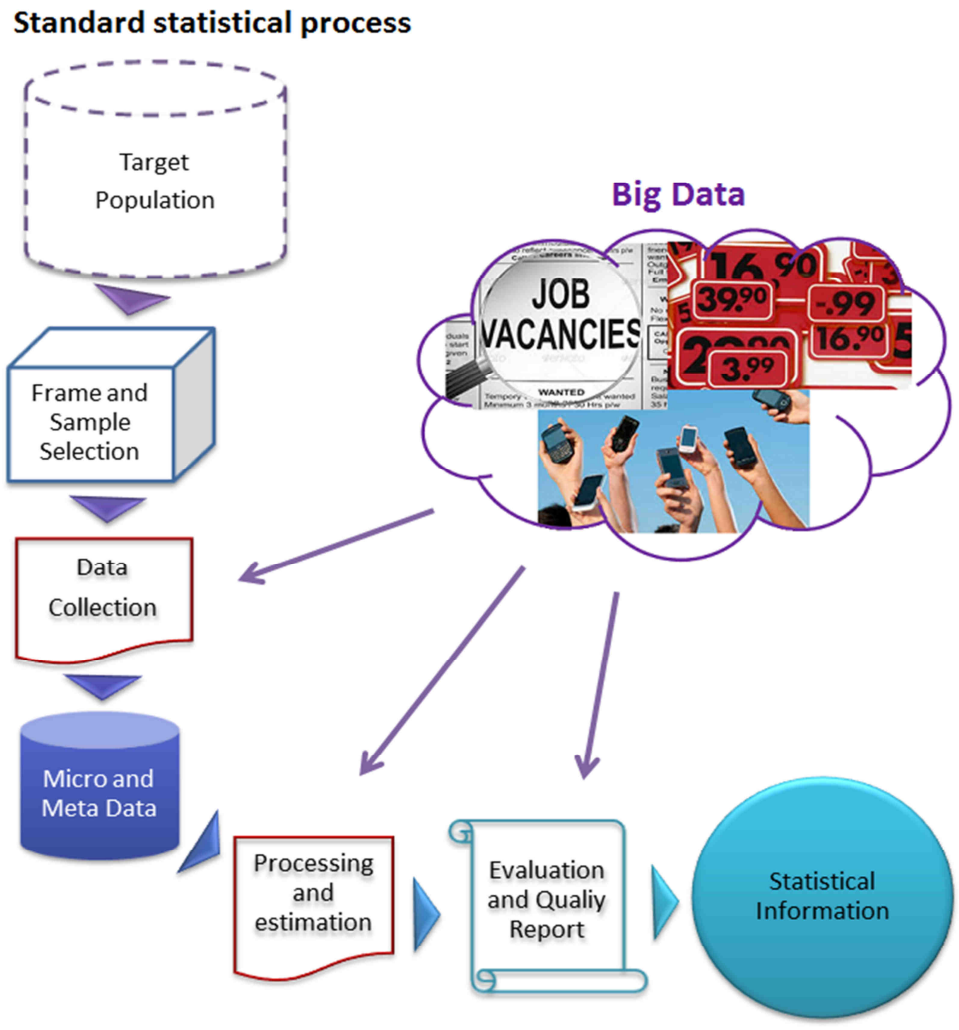
- Illustrating how linking geo- and statistical information and corresponding metadata provides additional value and creates new information

  Additional value of thus produced geo-statistics will be presented through various case studies where the potential of the data acquired from the mobile operators will be compared to the potential of the time-space dimension supported by the existing register-based data. SURS will strive to produce at least some datasets which will be applicable also to KASPeR or its successor StaGe (http://gis.stat.si/) to provide interactive visualisation of these data. Although the current development of the StaGe does not include the functionalities that would enable the visualisation potential of "mobile" geo-statistical data, it is positively estimated that the data adequately adapted could be included and presented effectively.

28. Under those objectives SURS started to negotiate with the largest mobile operators in Slovenia to acquire the data from them. At the end of the 2014 SURS received a set of mobile data from April 2014 to November 2014. The data are anonymised in the sense that they are transmitted in the form of three variables; anonymised IMEI of the phone, the time of the outgoing event (call, SMS or

connection in web using mobile card) and coordinates of base stations (antennas). The data from mobile operators have great potential in the official statistics also for the other government institutions (e.g. emergency services). The usage of the data could cover many statistical areas from tourism statistics to social statistics (active-inactive population, segmentation of populations by activity, migration of population, etc.).

29. SURS's long term strategy is to establish a completely secure IT environment to store and manipulate the sensitive data. SURS will also try to establish a long-term partnership with the private data owners with the aim to secure the continuous access to their data or even to process the data at owner's location to prepare (non-sensitive) aggregates which could be transmitted to SURS. SURS is also aware of the importance of positive public perception of the usage of sensitive data in official statistics, so activities will also be implemented to prepare the communication strategy towards the public.



Picture 1: Standard statistical process and new sources

IV. Conclusion

30. The new sources of data will play an important role in the future in the official statistics. With the rapidly developing ICT sector and the mass use of new sources, not only the world but also the whole universe has become small for a human being. Over the last decade as much information has been created as during the entire period of civilization. The possible usage of new sources (mainly Big Data) cover:

- Production of completely new type of statistics
- Production of existing statistics
- Production of flash statistics
- Benchmarking of results from exiting sources and validation of data
- New type (mode) of collection of micro data

But we should be careful with the replacement of the existing sources with new ones and statistics derived from them due to the problem of data sustainability. Before employing new data in the official statistical production, NSIs should have longer series of statistical results and a solution in the case of unexpected interruption in data access.