Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

A look into the future

Muhanad Sammar[1]

Disclaimer: The views and opinions expressed in this paper are those of the author and do not necessarily reflect the official policy or position of Statistics Sweden.

Abstract

The fact that the world is in continuous change and that new technologies are becoming widely available creates new opportunities and challenges for National Statistical Institutes (NSIs) worldwide. What if NSIs could access vast amounts of sophisticated data for free (or for a low cost) from enterprises? Could this facilitate the possibility for NSIs to disseminate more accurate indicators for the policy-makers and users, significantly reduce the response burden for companies, reduce costs for the NSIs and in the long run improve the living standards of the people in a country?

The time has now come for NSIs to find the best practice to align legislation, regulations and practices in relation to scanner data and big data. Without common ground, the prospect of reaching consensus is unlikely. The discussions need to start with how to define quality. If NSIs define and approach quality differently, this will lead to a highly undesirable situation, as NSIs will move further away from harmonisation.

Sweden was one of the leading countries that put these issues on the agenda for European cooperation; in 2012 Sweden implemented scanner data in the national Consumer Price Index after it was proven through research studies and statistical analyses that scanner data was significantly better than the manually collected data.

Key Words: Scanner data; Big data; Harmonization; Consumer Price Index.

1. Introduction

It is the responsibility of the producers of the statistics to create a sound foundation for a democratic society. An NSI that quickly can adapt to changes is compatible with sustainable development. On the other hand, lack of accuracy in the data from the producer of the statistics might have serious consequences to a country's economy. Decisions based on inaccurate or unrepresentative data may lead to unfortunate outcomes.

This paper introduces an initiative to modernise the price collection method using scanner data or big data for National Statistical Institutes (NSIs). Today we live and work in an IT era where information is available everywhere. Almost every successful company in the world stores large amount of information (big data) about their consumers as well as the goods and services they provide. Imagine if NSIs could access all this data for free (or for a low cost)? This might facilitate the possibility for NSIs to disseminate more accurate indicators for the policy-makers, significantly reduce the response burden for companies and at the same time reduce costs for the NSIs.

In 2012 Statistics Sweden implemented scanner data in the national Consumer Price Index and in the European Unions Harmonized Indices of Consumer Price. There was significant evidence that by replacing manually collected data with scanner data (for the daily necessities) the quality of the indices was improved. Scanner data has also been proven to be more cost effective.[2] Approximately 14 percent of all collected data used in the Swedish Consumer Price Index is from scanner data. Scanner data is mainly used in COICOP groups 01 (Food and non-alcoholic beverages, except for

[1] Muhanad Sammar, Statistics Sweden, Karlavägen 100, SE-104 51 STOCKHOLM, SUÈDE (muhanad.sammar@scb.se)

[2] Sammar, M, Norberg, A and Tongur, C (2013), "Issues on the Use of Scanner Data in the CPI." Available at http://www.dst.dk/da/Sites/ottawa-group/agenda.aspx

perishable fruits, vegetables and meat) and 02 (Alcoholic beverages and tobacco) but also in COICOP 05.5 (Tools and equipment for house and garden), 05.6 (Goods and services for routine household maintenance), 06.1 (Medical products, appliances and equipment), 09.3 (Other recreational items and equipment, garden and pets) and 12.1 (Personal care).

1.1 High non-response rate and inaccurate prices

In recent years, studies have shown that the collection of accurate and timely data for various surveys' has become relatively problematical. For NSIs to continue producing good quality data, there is no doubt that they will need to find new sources for collecting data. NSIs will also have to invest time in developing new methods that can be applicable to the use of new data source. In this section we will look at two studies that illustrate the extent of the problems with the traditional collection methods.

In 2010 a study was published by the Swedish Consumer Agency (Konsumentverket) with the scope of reviewing the price information in Swedish supermarkets.[3] A total of 13 500 product offers were examined in 291 stores. The research was conducted in the late summer of 2009 with the help of consumer advisors in 35 municipalities across the country. Two main findings of the study are:

- For 9 percent of the items in the survey, the prices were hard to find or could not be found at all. The lack of price information was larger in smaller shops.

- For 6 percent of the examined products, the prices on the shelves and packages were different from the purchase prices.

The second study of interest was about the difficulties with non-responses in telephone surveys. In July 2013, the Labour Force Survey (LFS) had a non-response rate of 32.5 percent while in the year 2000 the non-response rate for the same survey was down to 15 percent. Another survey that has developed similarly is the Household Budget Survey (HBS). In 2000 the non-response rate for the HBS was at 48 percent; in 2012 the non-response rate increased to 60 percent. The evidence is alarming and there are no signs that the trend will reverse.

The report concluded that the reason for the steady decline of the response rates over the last years were due to both external and internal factors[4] :

- The external factors identified
    - Difficulties finding relevant telephone numbers
    - Increasing reluctance towards answering among sampled units
    - Increased competition between different survey organisations and telemarketing companies

- The internal factors identified
    - Increasing workload at the Interview Unit
    - Poor interviewer training
    - Poor contact strategies
    - Inefficient work procedures

Both studies indicate that there is a high risk that both telephone surveys and manual price collection might become outdated in the near future. Besides the fact that both methods are costly, they may also be subject to measurement errors. For an NSI to continue with an old-fashioned approach cannot be seen as cost-effective in the long run. To

---

3 http://www.konsumentverket.se/Global/Konsumentverket.se/Best%C3%A4lla%20och%20ladda%20ner/rapporter/2010/2010_02_Prisinformation%20inom%20dagligvaruhandeln.pdf

4 Hörngren, Jan. Statistics Sweden's Overarching Project – Measures to Reduce Nonresponse in Individual and Household Surveys. 2011-05-30.

reverse the trend, more changes will be required, new policy changes must be adopted and a new mind-set will be needed.

In the next section, the author will present a vision of the future. Imagine a world where companies from all industries voluntarily submitted data to the NSIs. The NSIs would then function as spiders in a larger network by collecting, organising and processing the data. The company GS1 can be an important partner to achieve this vision.

1.2 GS1[5]

GS1 is an international not-for-profit association with member organisations in over 100 countries. Through cooperation, GS1 has developed sector-neutral standards and services which enable efficiency improvements in the flow of information and goods. GS1 is dedicated to the design and implementation of global standards and solutions to improve the efficiency and visibility of supply and demand chains globally and across sectors. The GS1 system of standards is the most widely used supply chain standards system in the world. GS1 provides a classification system that is structured logically with hierarchical levels called Global Product Classification (GPC).

1.3 GTIN code

Nearly all the trade items on the market are identified uniquely by a product code called the Global Trade Item Number (GTIN), provided by GS1 across the globe.[6] The code is often a 13 digit combination (12 positions for data and 1 position for a control digit). The first three digits usually identify the country where the manufacturer is registered. The country code is followed by 9 digits, of which the first part is a company prefix and the rest is the company numbering of their articles. A GTIN code for a product that is no longer available on the market can be reused for another product after a few years.

In November 2011 the European Commission issued a regulation on the provision of food information to consumers (EU Regulation 1169/2011). The regulation was set in motion in December 2014. In article 1, one can read:

"This Regulation provides the basis for the assurance of a high level of consumer protection in relation to food information, taking into account the differences in the perception of consumers and their information needs while ensuring the smooth functioning of the internal market."[7]

Furthermore it is written that the *"Regulation establishes the general principles, requirements and responsibilities governing food information, and in particular food labelling. It lays down the means to guarantee the right of consumers to information and procedures for the provision of food information, taking into account the need to provide sufficient flexibility to respond to future developments and new information requirements."*[8]

This means that there are strict rules that apply to GTIN codes according to regulations of the European Commission. For example, for a product where the declared net content has been changed, a new GTIN code will be required for the new item. A new GTIN code is also required when the wording of the existing product name, product brand or product description has been changed. Even when the change in packaging extends more than 20 percent in any axis, a new GTIN code will be required to the specific product. GTIN allocation rules specify what product changes require a new GTIN. Sometimes "the same" item has different GTINs in different markets. [9]

1.4 The missing link - GPC to COICOP mapping table

In addition, each GTIN is electronically linked to a unique GPC "brick code" by the manufacturers. The purpose with creating a GPC was to give manufacturers and retailers a common language for grouping products in the same way

---

[5] The source of the information in this section is www.gs1.org.

[6] GTIN should not be confused with the EAN (European Article Number) barcode. The EAN barcode is just a way to make the GTIN code machine-readable.

[7] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R1169&from=EN

[8] Ibid.

[9] http://www.gs1.org/1/gtinrules/index.php/p=overview

everywhere in the world. The brick level is the lowest classification level and consists of a group of narrowly defined products (e.g. perishable milk and milk products). In addition, master data about trade items is in many cases shared among trading partners by using an interconnected network of databases that are certified by GS1. This network is called the GDSN (Global Data Synchronization Network). GPC brick code is the chosen GS1 standard mandatory classification system for the GDSN. Furthermore, the GPC scheme is translated into many other languages and the GS1 system is certified by the International Organization for Standardization (ISO).

In Sweden, GS1 also provides a package of services (Validoo) that handles trade item information (such as ingredients, package size and packaging) and facilitates exchange of information between manufacturers and retailers. The Validoo service is connected to other GDSN data pools in other countries. In general, the code structure for attributes is the same within a country, but companies can enter different characteristics. It should be noted that the GS1 standards for item identification and classification have been developed primarily for supply chain and consumer applications. This means that in some cases it may not be perfectly optimised for statistics. Nevertheless, in these cases it is possible to submit work requests to GS1 to further develop the standards to meet the needs of the NSIs. In addition, many products are covered in the GS1 system, and the best coverage is currently for groceries and general merchandise. As information about the products attributes is not requested some producers chose not to fill in the information.

A smart solution for the NSIs is to create a classification key (mapping key) between GTIN and the Classification of Individual Consumption According to Purpose classification, COICOP. A mapping key will facilitate much of the work that today is done by staff at NSIs and in some cases reduce the duration of a specific task. E.g., NSIs will be more effective in the sampling process. A mapping key will also facilitate for NSIs to identify comparable products between countries. Constructing the mapping key will require a onetime effort.[10,11]

Steps needed to pilot

- Create GPC-to-COICOP mapping for a relevant item category
- Receive scanner data from retailer
- Download data from a GDSN data pool (same data pool as the retailer)
- Select GDSN data from relevant category (based on GPC)
- Replace GPC brick code with corresponding COICOP code
- Analyse scanner data

Please note it is not recommended to use the retailers own internal classification codes and should be avoided for mainly two reasons. Firstly, a GTIN code is harmonised over different store chains. If you have a specific product that is produced in a specific Swedish registered factory, the product will then have the same GTIN code in each and every store in the world where the product is sold. Secondly, the internal classification codes are unique only for each chain and not harmonised between different chains. If an NSI decides to go with the internal classification codes, then problems will occur when the retail chains change their codes (because then you have to remap the new codes either to the retailers' new classification or to the COICOP classification). In additional, by using the internal classification codes, you will not be able to compare price development with other countries (Purchasing Power Parity) nor will you be able to explain why GTIN codes disappear.

2. Challenges ahead

2.1 Challenges for NSIs

Over the past five years more and more countries have started to investigate the possibilities of using big data such as scanner data instead of traditional collection methods. The time has come for NSIs to align legislation, regulations and practices to be able to evaluate new sources of data such as big data and scanner data. Amendments must be made

[10] http://www.gs1.org/how-gpc-works

[11] http://www.gs1.org/1/productssolutions/gdsn/gpc/browser/

in the existing legislation and guidelines. The discussions have to start with defining the quality of the data. No action will only lead to further differences between NSIs and astray from the harmonization.

Another area worth exploring for NSIs is the data stored at companies' own databases. As most companies in Europe operate in a competitive market, one can strongly assume that many of them also obtain, collect and store large amounts of sophisticated data about their customers and their goods and services including variables such as prices, rebates, sold units, quantity etc. Some NSIs might argue that not all companies have good quality or sufficient data for NSIs to benefit from. The author agrees with the critics, but what prevents the NSIs from requesting the data they need? This leads to the next section: the challenges that NSIs are facing in the future. It is worth remembering that an NSI that engages other organisations and companies to cooperate in finding sustainable solutions and at the same time strives to be more cost-effective is consistent with what the public (and particularly the taxpayers) is expecting from an organisation. Arranging different types of workshops for various stakeholders, according to the author, would also signal to the public that the NSIs take their assignment of producing good quality indices very seriously.

2.2 Challenges for NSIs

For an NSI to be able to take advantage of a company's (scanner) data, it is necessary for them to have an open dialogue with retailers and their organisations (e.g. GS1s). It is also essential to make considerable efforts to create good relationships with people at various positions within the retail organisation and give the companies that provide data outstanding treatment. For example, the NSIs could offer the data providers guidance and support on how they can build their own indices and indicators for their own purposes and analyses. NSIs could also in return for receiving the data give the data provider's feedback. If NSIs would fail to secure data from companies, the alternative approach would be to encourage statistical offices and central banks (such as Eurostat and the European Central Bank) to agree to a text that obliges companies to provide data to NSIs and later present it to the European Commission. A legal text would allow for the NSIs to obtain big data and create good conditions for further research in the area. If that option also fails, a third alternative would be for NSIs to reach out to their own country's governments.

Once NSIs secure data from the companies, they should take the initiative to coordinate workshops with the data providers and their member organisations. The concept of these workshops could be to:

- Influence and inspire each other
- Share experiences
- Establish new contacts with industries and stakeholders on international as well as national levels
- Get insight of developments made by the companies (if a change was made by a company then the NSIs would thus be prepared for that change)
- Develop common policy concepts including goals, basic principles and priorities in the area of scanner data
- Institute new regulations that facilitate for NSIs to access scanner data from retailers
- Lay ground for a unified production system for scanner data processing (including data collection, data entry, data editing and data processing)

Suggested topics for discussion at the workshops might be:

- How to ensure better quality data
- Which variables might be useful for further analyses
- How prices are set on different products

NSIs should also organise at least once a year a more academically oriented workshop involving academics, researchers and experts within the subject area of big data/scanner data. The topics of discussion could be methodological issues (such as sample design, choice of index formula) and other issues connected to scanner data.

2.3 Challenges for statistical offices and central banks

Today the governments' awareness of the need for effective policies and programmes has improved around the globe; new regulations have been adopted at the national levels to prevent future financial crises occurring. The financial crises have also brought new demands on the NSIs from policy-makers, researchers, businesses and citizens to develop timelier, accurate, transparent and relevant indicators that are easily accessible to everyone.

The two questions that need to be raised soon are 1) what can Eurostat, other statistical offices and central banks in the world do to improve the statistics for some surveys/indicators produced at NSIs? This is especially important for surveys were the non-response rates are very high? And 2) how can NSIs access big data from companies and organisations without compromising confidence?

The author would like to see that statistical offices took much greater economic responsibility and at the same time provide solutions, financial and technical support to NSIs more than what they do today. To exemplify: instead of placing the responsibility on each country in the European Union to develop their own scanner data/big data production system, a more cost effective approach for Eurostat would be to develop its own system and share it with the member states.

When it comes to scanner data, the statistical offices need to come together and develop a manual as soon as possible identifying necessary steps for NSIs to take to ensure the quality of the scanner data from the companies. Discussions should also be about finding suitable methods for different surveys. By the time the manual on quality assurance has been finalised, a significant amount of countries might then have enough historical data to make empirical studies. Other topics that might be of interest for the NSIs are:

- A creation of a standard system for the use of scanner data
- Other technical aspects
- Customised quality assurance schemes for expanded use of scanner data
- How to manage data from multiple companies
- A legal text for data capturing (obligation to provide data to the NSI)

3. Concluding remarks

In this digital era NSIs are required to find new ways of thinking and acting to be able to continue to produce accurate indicators. New solutions need to be discovered and reviewed to meet the new demands. It is a fact that the consumer behaviour in Scandinavia over the last 10 years has significantly changed. From consumers physically going to shops to purchase commodities, more consumers have in recent years shifted their purchase to the internet and there are strong indications that these customers' shopping behaviour will continue.

The most favourable method for the NSIs must still be the method that best reflects the true reality. If a new data collection method is discovered to be as good as or better than the current method, then the new method should be more preferable than the old one. And if the new collection method is proven to also be less expensive than the current one, then that is one more significant argument that an NSI should consider changing the data collection method.

We all have sometime that necessity is the mother of invention but the question is do we have to wait till the need arises? The author welcomes a broad and in-depth discussion on the issue of scanner data and especially on data quality. Scanner data might today be the most suitable method to collect prices for some surveys at Statistic Sweden, but this does not necessarily mean that it will be the optimal solution for all time. As the author sees it, and based on his experience, this is a great time and an exceptional opportunity for NSIs to secure more big data.

References

Hörngren, Jan (2011), "Statistics Sweden's Overarching Project – Measures to Reduce Nonresponse in Individual and Household Surveys", unpublished report, Stockholm, Sweden: Statistics Sweden.

Sammar, M, Norberg, A and Tongur, C (2013), "Issues on the Use of Scanner Data in the CPI", paper presented at the Ottawa group meeting, Copenhagen, Denmark. Available at http://www.dst.dk/da/Sites/ottawa-group/agenda.aspx

REGULATION (EU) No 1169/2011, http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R1169&from=EN

Konsumentverket (2010). Available at http://www.konsumentverket.se/Global/Konsumentverket.se/Best%C3%A4lla%20och%20ladda%20ner/rapporter/2010/2010_02_Prisinformation%20inom%20dagligvaruhandeln.pdf

GS1 (2015), http://www.gs1.org

GS1 (2015), http://www.gs1.org/1/gtinrules/index.php/p=overview

GS1 (2015), http://www.gs1.org/how-gpc-works

GS1 (2015), http://www.gs1.org/1/productssolutions/gdsn/gpc/browser/