Attributes for Big Data for Official Statistics – an Application to Scanner Data in Luxembourg

Ibtissam SAHIR (ibtissam.sahir@gopa.lu), Florabela Carausu, (florabela.carausu@gopa.lu),
Botir Radjabov (Botir.Radjabov@ext.statec.etat.lu)

Abstract:

Big Data is one of the key topics, around which the so-called data revolution has embarked. The increased usage of Big Data in the private sector has raised expectations from the public sector as well. With this in mind coupled with an increased pressure from users to digitize their production systems, statistical offices have also engaged in innovative projects to benefit from the opportunities which Big Data can offer. To balance users' expectations to the real possibilities of making good use of Big Data in official statistics, a set of attributes are proposed.

It should be noted that, the classic attributes of Big Data – the 4 Vs: volume, variety, velocity and veracity – are insufficient to explore their suitability for official statistics. As a first distinctive factor, the scope, for which official statistics needs or may use Big Data, is different from the same scope of the private sector. The former scope is decision-oriented analysis whereas the latter scope is an action-oriented analysis. Having the above-mentioned argument in mind, a case study in the field of official statistics so as an application of scanner data in Luxembourg is proposed and examined. The real life example from the study shows why Big Data suitability for official statistics using the classic attributes remains to be a challenge in itself.

Keywords: GOPA Luxembourg, Big Data, scanner data, CPI, classification, Price Index, STATEC, Price statistics

1. INTRODUCTION

"Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software."[1] Gartner introduced the attributes of Big Data in its definition, defining Big Data as "high-volume, high-velocity and/or high-variety information assets that demand cost-

[1] http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf

effective, innovative forms of information processing that enable enhanced insight, decision making and process automation."

It is generally accepted that Big Data can be explained according to the four Vs:

- Velocity – the speed at which the data is created, stored, analysed and visualised;
- Variety – unstructured data available from a variety of sources;
- Volume – huge amount of number of records and attributes.
- Veracity – (un)certainty of the data.

Some experts enhance the list of Big Data V attributes with others such as: Volatility, Variability, Value, etc. Nevertheless, the Vs attributes of Big Data do not capture[2]:

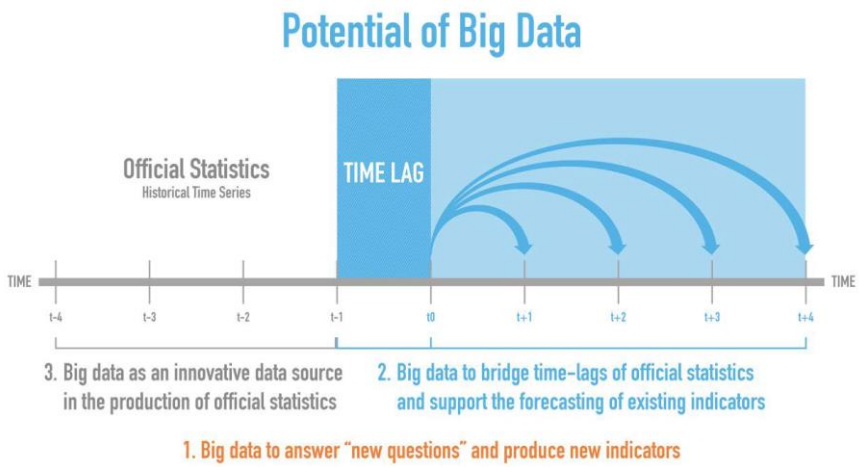
- the enlarged scope of the corresponding datasets;
- the extensive potential of making a use of this data;
- the lack of feasible and efficient means for database management and data processing applications.

The usage of new information, through Big Data, can maximise the potential of efficient policy making (monitoring and evaluation) if the associated risks are adequately prevented. Figure 1 is a graphical representation of the potentials of Big Data which can directly or indirectly benefit policy-making through:

- answering new questions and producing new indicators;
- bridging time lags in the availability of official statistics;
- supporting timelier forecasting of indicators; and
- providing an innovative data source in the production of official statistics.[3]

[2] P. Hackl (2015), Statistics Austria: 'Big Data: What Can Official Statistics Expect?'
[3] See IMF Staff Discussion Note: 'Big Data: Potential Challenges and Statistical Implications', SDN/17/06, September 2017

Figure 1: The Potential of Big Data



Source: IMF Staff Discussion Note: 'Big Data: Potential Challenges and Statistical Implications', SDN/17/06, September 2017

The usage of scanner data for price statistics can be potentially beneficial in bridging time lags in the availability of official statistics, supporting the timelier forecasting of indicators, and in providing an innovative data source in the production of official statistics. In addition, scanner data, being an innovative source in the context of official statistics production, presents as advantages, the reduction of response burden of enterprises, productivity gains for NSIs, and improved accuracy of price statistics.

2. THE USE OF SCANNER DATA BY STATEC

In the Framework Agreement on assistance to the implementation and monitoring of the Community Statistical Program 2015-2017 and 2018-2021, a project in the field of "Prices" is implemented by GOPA Luxembourg, aiming essentially to support the improvements and the modernization of price statistics of the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (STATEC).

Scanner data is being increasingly used by National Statistical Institutes (NSI) for the calculation of price indices. Several NSIs started to use scanner data for the compilation of the Consumer Price Index (CPI) instead of physically collecting prices data. In Luxembourg, the collaboration was put in place with several retailers who agreed to transmit their data to STATEC every month. Several tasks have been accomplished to guarantee that the process goes smoothly and all eventual problems are anticipated and previously taken into account.

It should be noted that scanner data are big files of transaction data, which differs from retailer to retailer. The files datasets typically contain complete coverage of items sold by a retailer at all their locations as well as items quantities sold and sales received by the retailer for these items. In general, the files mainly contain information such as: EAN, Item Label, Number of Units Sold, Sales, Month and etc.

3. METHODS

This paper presents the methodological approach adopted by STATEC and GOPA Luxembourg consultant to use the scanner data for integration in the formers official price statistics. Datasets received contain different product classifications which need to be mapped to a single CPI classification.

The first step in processing scanner data is classifying the individual items into Classification of Individual Consumption according to Purpose (COICOP) groups, which is the standard classification used for compiling a CPI.

A cascading infrastructure was put in place for the automatic classification procedure, namely:

1. Search for a common attribute
2. Dynamic mapping table
3. Text mining
4. Manual classification

The different solutions chosen are executed from the most discriminating to the most probabilistic. A learning set was firstly built by manual classification. In general, around 80% of the items available in the scanner datasets can be classified with this cascading approach.

The implementation of new data sources and methods in any statistical series requires careful consideration of the statistical impacts (outliers, missing data, etc.). From those items which have been classified, not all items will enter the index compilation. In line with the recommendation of Eurostat within the use of the dynamic method where baskets are resampled more frequently and the process is automatized to a large extent, an outlier filter which flags prices that drop below or rise above given thresholds is put

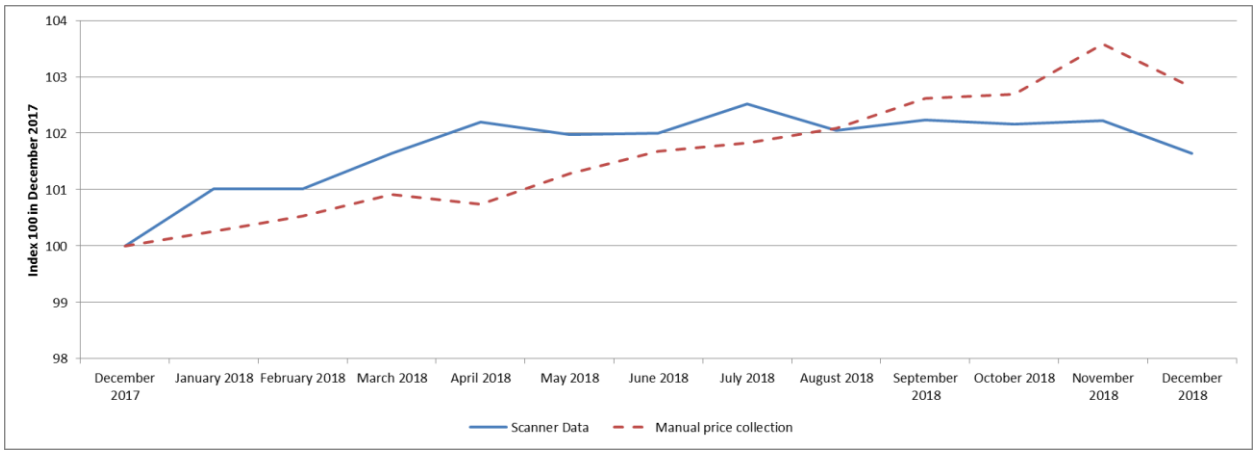
in place. The method applied is the thresholds proposed by Van der Grient & de Haan (2010).

Imputations are made for items which are either temporarily missing or, as explained before, are considered "outliers" or "dumped prices". The principle is that a price is imputed if it was included in the sample in a previous period. The idea is that if an item was well sold in a previous month it is likely that, once it is again available, it will be well sold again and it is desired to take into account this possible price change in price index calculations.

4. Results

Currently, scanner data covers roughly 5% of the 2018 Harmonised Index of Consumer Prices (HICP) of Luxembourg basket. The current coverage of scanner data is limited to COICOP division 01 (food products and non-alcoholic beverages) with exclusion of strongly seasonal items. The result of the compilation of the index for with the "automated scanner data methodology", which has been described above, can be seen in **Figure 2** together with the same result of the "manual price collection methodology". The base period for both sub-indices is December 2017.

Figure 2: Comparison between the CPI manual price collection and scanner data for COICOP division 01



Source: STATEC

The usage of scanner data for official statistics, alongside with the advantages featured, also presents a set of issues that have to be tackled in order to maximize its benefits, such as:

- Methodological issues;
- Identification and treatment of relaunched products;
- Products homogeneity;
- Representability;
- Treatment of rebates;
- Investment costs and partnerships with data providers.

As mentioned above, the scope for which official statistics needs or may use Big Data is different from usage of Big Data by the private sector. For decision-oriented analysis (i.e. policy making analysis), the continuity of information from Big Data is critical. This is especially applicable to the presented case study of scanner data in Luxembourg. Continuous scanner data is required by statistical office of Luxembourg to produce price indexes every month.

It is argued that it is impossible to efficiently explore Big Data suitability for official statistics using classic attributes of Big Data. In this sense, it should be noted that scanner data, i.e. a Big Data source, should offer a certain level of structure, always containing such attributes as month of purchase, item label, item price, item turnover and etc. whereas Big Data is usually characterised by being unstructured. Moreover, whether velocity of scanner data is of the same nature as of Big Data remains debatable. This means that it might be true that the speed by which scanner data is created and stored is fast enough, but it is certainly slower when scanner data needs to be analysed and visualised as it has been carefully described above.

The application of Big Data attributes to scanner data – as Big Data for official statistics – could guide towards refining the list of attributes in the context of Big Data, considering, as essential attributes, the benefits the usage of Big Data can bring balanced with the limitations of already existing data sources, such as:

- Timeliness vs continuity;
- Response burden vs. production costs.

Nevertheless, such attributes shall be assessed for each (type of) dataset exclusively, having in mind the potential indicators derived or the statistical domain worked in, and not on a general basis.

5. Conclusions

STATEC is aiming to include more retailers and more COICOP sub-groups such as seasonal food products, cleaning and maintenance products as well as personal hygiene products into compilation of CPI/HICP using scanner data.

Currently, the National Statistical Institute of Luxembourg (STATEC) is using bilateral dynamic index compilation approach for scanner data on «Food» and «Non-Alcoholic Beverages» Classification of Individual Consumption by Purpose (COICOP) groups calculation. This approach implies that prices of goods of two consecutive months are taken into account with basket of goods resampled every month. STATEC is utilizing machine learning algorithms in items classification process to make classification process faster and almost automatic. Future goals regarding scanner data integration in CPI/HICP estimation are to include COICOP groups of «Fresh Fruits», «Fresh Vegetables» and «Alcoholic Beverages» into STATEC production system as well as to integrate multilateral index complication approach for scanner data for the above mentioned COICOP groups.

A proper methodology is under development with the support of GOPA Luxembourg consultant to address the issue of seasonality and seasonal products using scanner data. The compilation methodology should also be improved with respect to issues of replacements, quality adjustments as well as product linking. In particular, the use of multilateral methods such as Gini-Éltetö-Köves-Szulc (GEKS) and weighted time product dummy (WTPD) for scanner data is under investigation and more research in the area of scanner data is needed to adequately use scanner data in the future. Preliminary results of the applicability of the above mentioned multilateral methods can be seen in **Figure 3 and Figure 4,** analysing groups of fresh fruits and fresh vegetables respectively.

Figure 3: Comparison of GEKS and WTPD for fresh fruits

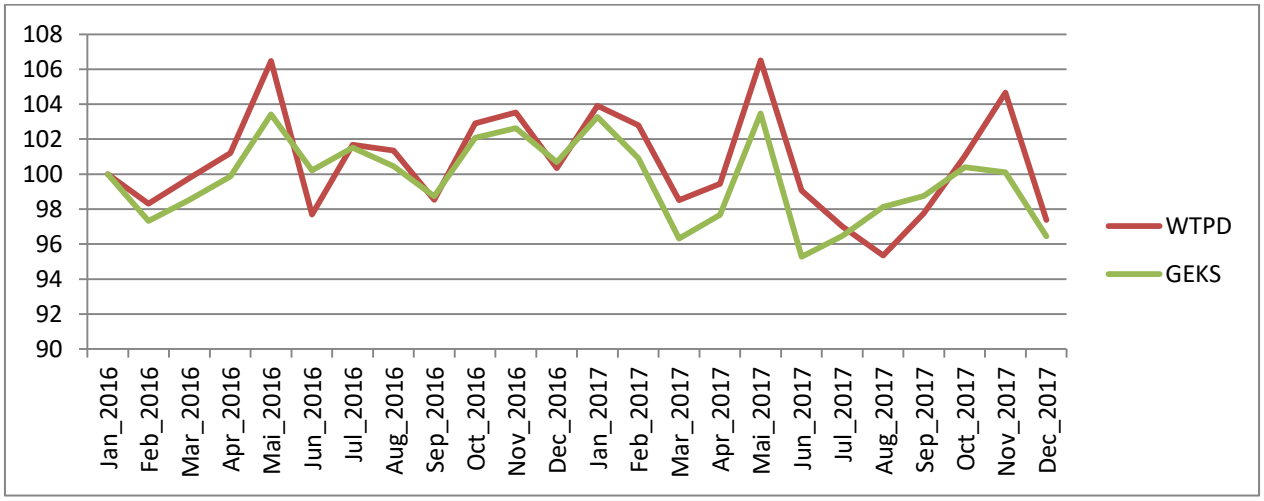
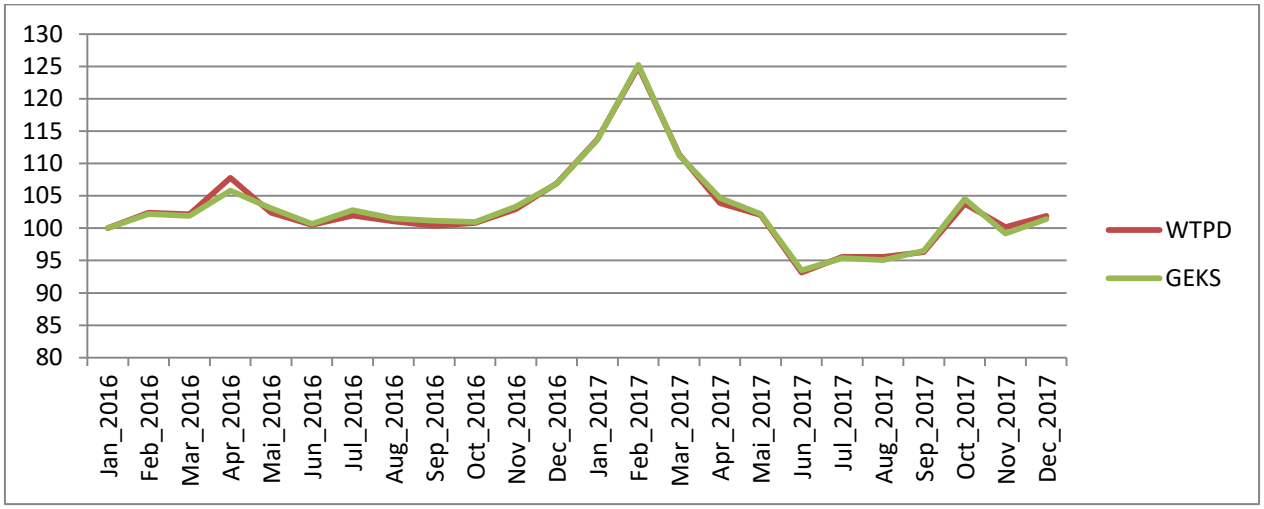


Figure 4: Comparison of GEKS and WTPD for fresh vegetables



The application of Big Data attributes to scanner data – as Big Data for official statistics – could guide towards refining the list of attributes in the context of Big Data, considering, as essential attributes, the benefits the usage of Big Data can bring balanced with the limitations of already existing data sources.

Nevertheless, such attributes shall be assessed for each (type of) dataset exclusively, having in mind the potential indicators derived or the statistical domain worked in, and not on a general basis.