Key Factors for Obtaining Access to Big Data

Marc Debusschere, Ken Van Loon, Nico Waeyaert

Statistics Belgium

Abstract. In spite of years of testing and trying, the regular and recurrent use of big data and more particularly of privately owned big data for the production of official statistics is still almost non-existent in the ESS. This is not due, as was anticipated originally, to storage or processing difficulties or methodological issues, but almost exclusively to a lack of sustainable data access. Mobile phone data and similar types of big data are typically owned by private enterprises seeing them as valuable assets they are reluctant to share with statistical institutes, for various reasons. As a result, a realistic perspective of using them systematically and routinely for the production of official statistics is still lacking, although the promise they hold becomes ever more apparent.

Statistics Belgium (Statbel) has successfully implemented scanner data in its monthly calculations of the consumer price index (CPI), based on four key elements: a clear and detailed business case, high-level engagement, specific legislation and the fostering of trust by absolutely guaranteeing confidentiality and privacy. This approach may serve as a model for gaining access to mobile phone data and similar types of big data and setting up regular statistical production chains based on them.

1. Official Statistics and Big Data: State of the Art

The term ‘big data’ was coined at the end of the 1990’s for datasets which are too large, too fast-moving or too unstructured to be processed in a ‘normal’ way. The first documents exploring the potential of big data for official statistics, still in a largely theoretical way, date back only to the beginning of the present decade. For the European Statistical System (ESS) and even for official statistics globally the Scheveningen Memorandum on "Big Data and Official Statistics", adopted by the ESSC on 27 September 2013, is a landmark. In the memorandum, the DGINS acknowledge that big data represent new opportunities and challenges for official statistics.

As a result, Eurostat and many national statistical institutes established big data teams or taskforces, often in collaboration with academic researchers, to examine the potential of big data to complement or (partially) replace traditional survey and administrative data sources, with the explicit aim of making official statistics much more detailed and timely. The end of 2015 saw the launch of the EC-funded ESSnet Big Data, a 22-partner consortium conducting pilots on various types of big data: webscraping job vacancies and enterprise characteristics, ship position (AIS) data, smart meters, mobile phone data, …

The pilots conducted in the course of ESSnet Big Data and various national initiatives considerably expanded the knowledge about various types of big data and on how to handle them. This is particularly the case for mobile phone data, resulting in an impressively long list of publications. And yet, in spite of the time and money spent on numerous pilot studies at national and European level within the last three years, no instance is yet to be found of official statistics partially or wholly based on them. Even tested statistical use cases seem to be still lacking. At this moment only one-shot studies, based on limited datasets obtained from one operator in one country, have been conducted.

2. What is Blocking Us?

The first theoretical papers to discuss the use of big data for official statistics pointed out two potential barriers:

- The sheer volume, refresh rate and complexity of big data rendering it *next to impossible to store and process them with existing IT infrastructures and software*;
- The challenge to extract meaningful information from almost inconceivably large or unstructured datasets confronting us with *totally new methodological issues*.

Although these are indeed real and formidable problems, they don't look at present as insurmountable as they did before. Through work on real and synthetic datasets, the elaboration of theoretical use cases, progress in storage capacity and distributed computing, and the new approach of leaving the data where they are (also circumventing ownership, confidentiality and privacy issues) and instead moving the queries and processing algorithms about, it now looks more likely they can be solved. Consequently, it appears increasingly feasible to produce high-value high-quality official statistics based on mobile phone and similar big data, especially when combined with survey results, administrative data and other big data.

Data access, on the other hand, has proven to be much more problematic than anticipated. Different transfer models are possible, but for various reasons all have so far proven impracticable. Mobile phone data and many other types of big data are produced by private companies, as a side product or 'exhaust' of their operations. These data do not generate themselves, creating them from operational events and storing them in huge quantities in an exploitable database is optional with regard to the companies' core business, and a very costly affair indeed. Taking the example of mobile network operators, it is hardly conceivable to impose on them the obligation to lay out millions of euros to set up a database infrastructure solely catering for official statistics. Luckily,

most if not all operators already have this infrastructure in place for their own purposes of optimisation operations and commercialising their data.

If the data are available and exploitable, there are three possible ways in which they could be obtained as a regular source for official statistics: 1) buying; 2) demanding; 3) exchanging.

Statistical institutes are public institutions funded by taxpayers' money and they lack the means to pay huge sums of money for data. Moreover, it seems justifiable that data used for the public good should be made available free of charge by citizens and enterprises. Unfortunately, few companies owning big data in Europe seem to share this view and even if they are willing to provide data they expect to be paid very handsomely, at a level unaffordable for public institutions. In the course of ESSnet Big Data some limited mobile phone datasets have been bought for analysing and testing, which of course is distinct from buying data on a regular basis for statistical production.

The second option, demanding data for free from the owners, by way of a legal obligation, has major drawbacks. First of all, as already mentioned, setting up the required database structure is very costly and as a result hard to impose. Furthermore, the legal instruments at European or national level, either general or domain-specific, are still largely lacking or not specific enough; and taking into account the time required for adopting new legislation, it seems unlikely they will be soon forthcoming. And finally, forced cooperation seldom leads to optimal quality.

The final option, creating a win-win situation by setting up partnerships with data owners to obtain mobile phone or other big data in exchange for non-monetary benefits, seems the most promising way forward with regard to quality and feasibility. In the case of mobile phone data, the benefits to the network operator can be intangibles such as a positive image by contributing to the common good, but also methodological assistance to exploit their own data and additional statistical data to increase their commercial value (see the article co-authored by Statbel, Eurostat and the Belgian mobile network operator Proximus [1] for an extensive discussion of this model). However, in practice this does not seem to work out, as not one single partnership between a statistical institute and a network operator for producing official statistics in a recurrent and systematic way has seen the light in any ESS country in the past three years ...

3. A Four-Way Approach to Data Access: the example of CPI and scanner data

An essential prerequisite for a more extensive use of big data as a source for official statistics, is solving the crucial data access issue. Because past experience shows there no easy solution exists, it might be worthwhile to review similar situations where official statistics receives data on a regular

basis from private partners, in order to identify the critical success factors. These could then be applied to mobile phone data and other privately owned big data.

Statistics Belgium has successfully implemented scanner data obtained from the major supermarket chains in its monthly CPI calculation. Four key factors were instrumental in obtaining this result [2]: a clear and detailed use case specifying costs versus benefits for all, high-level engagement and active support, specific legislation, and building trust to minimise fears and risk aversion.

3.1 Clear and Detailed Use Case

Approaching a private company owning data with a general and vague request, along the lines of 'Please give us whatever data you have, we may find some uses for them.', is sure to meet with a refusal. What is needed, instead, is a clear case stating which data exactly are needed for what statistical result serving what purpose, beneficial for whom and at what cost to the data provider, and why these data are indispensable to obtain the specified result in the most efficient and quality-oriented way.

In the particular case of the CPI, the high importance to society is obvious and self-evident. It is used for a wide variety of important objectives: indexation (wages, social benefits, rental and other contracts), the conduct of monetary policies, deflation of national accounts and conversion of nominal monetary values into real terms. These numerous high-profile applications have also turned it into a political top priority. Furthermore, it is clear to anyone that using scanner data is more efficient and delivers higher quality compared to noting down individual prices in selected shops. Also, the recurrent cost to the supermarket chain is negligible once an automatic data flow has been set up. And finally, an additional benefit to them is they can then broadcast their contribution to the common good by freely providing their data.

3.2 High-Level Engagement and Active Support

Precisely because of the high importance of the CPI for Belgian society and economy, supermarket chains were well aware that this data request was not just originating from the statistical office, for more or less academic reasons, but that it had the strong backing of employers' associations and unions, a wide range of social organisations and academics, and not in the least of the political authorities who have a strong interest in an uncontroversial CPI of the highest possible quality, detail and timeliness.

This high priority was systematically communicated by different stakeholders to the management of the supermarket chains, at a sufficiently high hierarchical level to ensure a clear decision, active engagement and concrete follow-up.

3.3 Building Trust

The main concern of private data owners is data security and respect for confidentiality towards their competitors also providing data to the statistical office. For this reason it is essential to have all contractual, physical and procedural safeguards in place to absolutely guarantee confidentiality (and respect for privacy, in the case of personal data). In practice this means:

- a fully operational, physically secure datawarehouse to receive and safely store the data;
- tested and documented procedures, including strictly managed and logged access by select and screened persons;
- a signed contract clearly stating the obligations and guarantees.

3.4 Legislation

Although not sufficient by itself to ensure full and willing cooperation, it definitely helps that the recently updated HICP regulation explicitly stipulates that scanner data must be made available to national statistical institutes for the calculation of HICP (which uses essentially the same data sources as the national CPI). Pointing out to data owners in preliminary discussions that compliance can be legally enforced probably increases the likelihood they will agree to a voluntary data transmission arrangement …

4. Getting Access to Privately Owned Big Data

The four-way approach described above, which has worked well to obtain scanner data for CPI, might be used as a general model for unblocking access to privately owned big data. The discussion below will focus on the more concrete example of mobile phone data, as they are an extremely important and promising big data source where data are available in directly usable formats, IT and methodological problems seem manageable and use cases have been tentatively identified. In principle, however, it should be possible to extend this approach to any other privately owned big data source.

Statbel has created a multi-disciplinary virtual big data team consisting of experts from various units which can be mobilised whenever data should become available and their specific competencies are needed for defining and refining use cases, negotiating conditions and procedures, drafting legal documents and contracts, and setting up or adapting statistical production lines.

4.1 Clear and Detailed Use Case

Big data sources such as mobile phone data are potentially useful for a large number of existing and novel statistics on population, migration, tourism, mobility and transport, time use, … This may seem fortunate to statisticians, but when approaching a data owner it may rather be a disadvantage because this could result in a data request which is general, vague and unspecific. Data owners are understandably reluctant to hand over data if it is unclear for what exact purpose. So, even if there is the idea to expand to multiple uses at a later stage, it is preferable to select and elaborate one particular and specific use case with which to approach a big data owner.

The use case must be detailed and precise, specifying:

1. what statistical product is to be the *output*, for a particular use and interest;

2. what selection of raw or transformed data is needed as *input* to arrive at the product;

3. what methodologically sound *statistical data processing* is to be used to that end.

Resolving the first issue, on the statistical output, is probably easier if an existing rather than a new statistical product is chosen, for several reasons: there is less need to argue the reasons and needs, modifying is easier than creating from scratch, and the focus is on the practical long-term modalities of changing an existing data flow.

An example of such a statistical product is the annual living place-workplace matrix which Statbel at present compiles annually at the level of municipalities on the basis of extractions from the population and social security administrative registers. A statistical use case was elaborated for this matrix using a selection of mobile phone positioning data consisting of approximately 121 million mobile phone positioning records (queried from a database containing several hundreds of billions of them). When tested and ready, the query and compilation process should result in a drastically more timely and geographically granular output, for hardly any cost or effort to the data owner and the statistical office.

4.2 High-Level Engagement and Active Support

Private companies owning big data are composed of different actors and units which do not necessarily share the same interests and perspective. Simplifying somewhat: the sales department wants to sell the data at the highest possible price, the business development people want to create new products based on their big data and are usually willing to share and collaborate to that end, the legal unit goes for zero risk (‘when in doubt, do nothing’), and the higher management would certainly like new products for profit, but is also very wary of the reputational risk with regard to the privacy concerns of the general public. The higher management ultimately decides, of course.

Experience has taught that good contacts with business development are necessary but not sufficient. As long as matters remain in the research/testing/pilot stage, collaboration is often quite good. But in order to move to regular statistical production, a decision at higher level, often the very highest level, is needed. This probably means that a breakthrough will only be possible when the management of private companies owning big data is approached by the highest possible level of the public authorities, and if there is political will to see things moving.

Another approach which has proven successful in the case of scanner data, is using the business sector federation and/or the official regulator as an intermediary to the individual enterprises, and in bringing the enterprises together to discuss issues collectively. The business federation often takes the somewhat longer-term and collective sector view and may be able to convince individual companies to do likewise. The official regulator has close contact with the enterprises and has a certain authority as a regulating and controlling body. Finally, assembling several private data owners in a meeting shows they are not the only one being asked and may allay fears of competitive disadvantage by providing data.

4.3 Building Trust

Private big data owners, and most particularly mobile network operators, are deeply concerned about their reputation with the press and general public which could ultimately even affect their stock market value and threaten the position of the CEO. Consequently, the strongest possible guarantees on data confidentiality and privacy have to be given by the statistical institute, in a concrete way, by:

- Intensive contacts and face to face meetings between the statistical institute and all concerned persons and units of the data owner (management, business development, legal, technical, IT), in order to discuss all issues and potential problems in detail and at the length which is required: legal guarantees on privacy and data confidentiality, data selection, transfer and storage, effort and investment required, costs, expected benefits from the mutual exchange of data and knowhow;

- Explicitly addressing and solving all concerns of the data owner, in every area: privacy, technical feasibility, procedures, cost, returns in terms of statistical and methodological assistance for product development;

- Resulting in a sufficiently detailed legal document (contract) to be signed by all parties.

4.4 Legislation

To support and if necessary enforce the request for data, legal provisions have to be in place which support and legitimate this request. Several possibilities, which can be combined, exist:

- General statistical legislation at the European and national levels has to be adapted to big data and specifically mention the right of national statistical institutes to obtain mobile phone data and similar privately owned big data; at this moment this possibility usually remains implicit, by formulas such as 'all data needed' not excluding big data;

- Legislation which might prevent data access (such as privacy, data protection or telecommunication laws) should allow the use for official statistics.

- Specific domain legislation at the European or national levels (such as the HICP Regulation) should explicitly mention access to specific types of data in order to produce specific statistics.

5. Conclusions

Privately owned big data, and more particularly mobile phone data, have the potential to revolutionise many domains of official statistics. The current state of research and testing suggests that storage, software and IT infrastructure, and methodology are sufficiently mature for this to happen, but that no data are yet forthcoming on a regular and stable basis, which of course makes it impossible to set up regular statistical production.

This problem will not be solved by adopting resolutions or drafting resolutions, but only by a concrete action plan which takes account of the existing bottlenecks and finds a way of removing them. The implementation of scanner data in the CPI may offer a model for such an action plan, consisting of four key elements: defining clear statistical use cases, obtaining high-level engagement and support, building trust and a mutually beneficial cooperation with data owners, and adopting appropriate legislation.

6. References

[1] Debusschere, M., Wirthmann, A., De Meersman, F. (2017), Official statistics and mobile network operators: a business model for partnerships (http://nt17.pg2.at/data/abstracts/abstract_125.html), NTTS Conference, 13-17 March 2017 (download page abstract and presentation)

[2] Van Loon, K., Roels, D. (2018), Integrating big data in the Belgian CPI (https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf), UNECE Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018 (PDF download)