Towards Big Data methodology: a generic Big Data based statistical process

**Keywords:** Big Data, methodology, statistics production, statistical process.

# Introduction

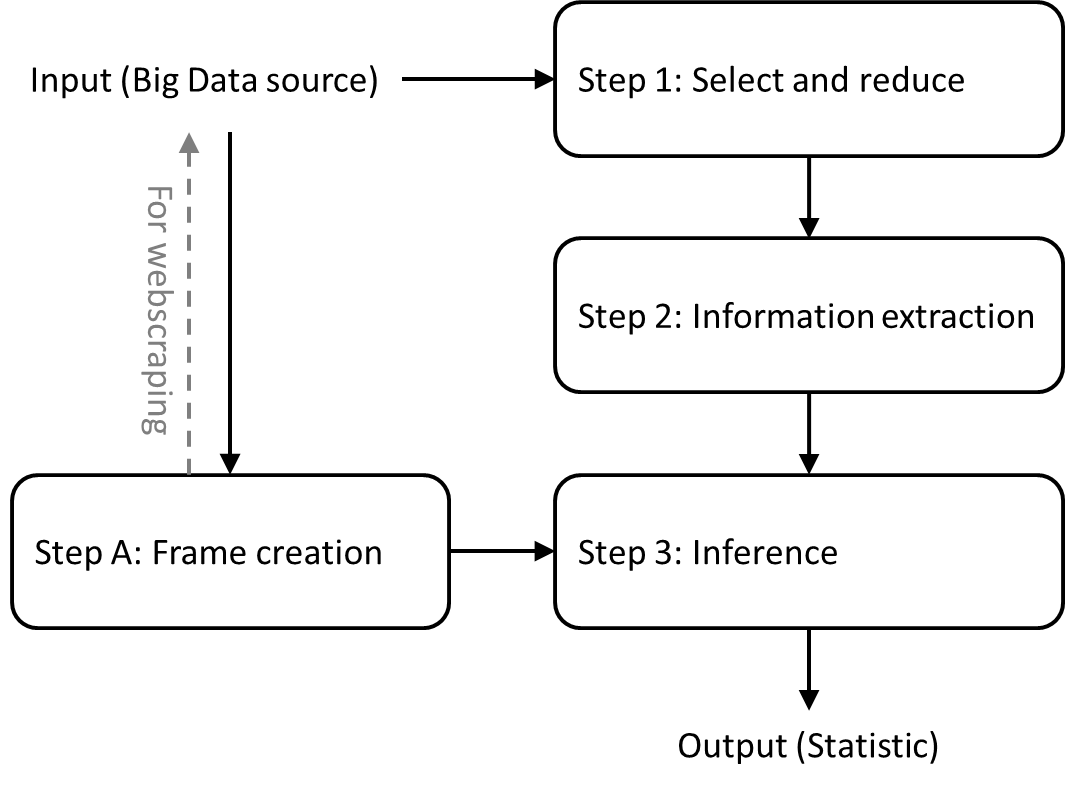
More and more National Statistical Institutes are investigating the potential of using Big Data. This has resulted in a number of statistics created that use Big Data at various stages of production. An overview of the 13 examples found [1] is shown in Table 1. Not all processes have already completely matured and only two of them are in production. These numbers will undoubtable increase in the near future.

Table 1. Overview of Big Data based official statistics and their status.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Statistics** | **Data sources** | **Countries** | **Status** | **Ref.** |
| 1 | Consumer Price Index | Scanner data, web scraped prices | Various countries | In production | [2] |
| 2 | Traffic intensities | Road sensors | NL | In production | [3] |
| 3 | Online job vacancies | Scraped job vacancies | ESSnet | Towards implementation | [4] |
| 4 | Enterprise characteristics | Scraped websites | ESSnet | Towards implementation | [4] |
| 5 | Electricity/energy consumption | Smart meter data | ESSnet | Towards implementation | [4] |
| 6 | Maritime and Inland waterway statistics | Automatic Identification System data | ESSnet | Towards implementation | [4] |
| 7 | Financial transaction based statistic | Bank transaction data | ESSnet | Exploratory | [5] |
| 8 | Earth observation derived statistics | Satellite and Arial pictures | ESSnet | Towards implementation | [4,5] |
| 9 | Mobile network derived statistics | Mobile Network Operations data | ESSnet | Towards implementation | [5] |
| 10 | Innovative tourism statistics | Various sources (e.g. webpages) | ESSnet | Exploratory | [5] |
| 11 | Innovative companies (small) | Scraped websites | NL | Towards implementation | [6] |
| 12 | Social mood on economy index | Social media messages | IT | Experimental | [7] |
| 13 | Mobile phone derived outbound tourism | Mobile Network Operations data | FI, EE | Experimental | [8,9] |

In this paper the processes developed for the examples listed combined with the personal experiences of the authors are used to derive a generic process by which Big Data can used in official statistics production. Be aware that the focus of this process is on using Big Data as the main data source and *not* when used as an additional data source; see [10] for more info on this very important difference.

The basis for the workflow is the one described for the production of Traffic Intensity statistics based on road sensors [3,11]. In this process, first a select and reduce step is applied to the raw data (Step 1). This seriously reduces the amount of relevant data remaining; e.g. lots of unneeded data, such as variables and units, are removed. Next information is extracted from the data remaining (Step 2). This, combined with the frame (Step A), is used in the last inference step (Step 3) to produce a statistic. In Figure 1 an overview is given of this generic process.



**Figure 1. Overview of the 4 steps required in Big Data based statistics production.**

# Results

While studying the list of Big Data based statistics included in Table 1, it became clear to the authors that the process shown in Figure 1 -in principle- contained the (general) building blocks observed in many of the other Big Data based statistical processes. These steps are described in more detail below.

## Step 1: Select and reduce

Usually, the first step in the process involves a selection at the unit and variable level. As a result the amount of data is often seriously reduced [3]. Goal of this step is either predominantly focussed on the removal of unwanted data, the selection of relevant records or an a combination of both. It may require the need for specific quality indicators used as selection criteria. Examples of this step are: i) the selection of products from scanner data [2], ii) the selection of road sensor data to only include data of high quality of a specific set of variables needed [3], iii) the selection of AIS-messages that contain info on the position of ship [4], iv) the splitting and selection of areas on satellite pictures for the particular region or country studied [5], and v) the selection of social media messages containing specific words [7].

## Step 2: Information extraction

In the next step, the data remaining is processed in such a way that the information required for the specific statistical need is obtained. This may simply involve extracting a value from a dataset, may require applying a filter or a transformation of the data, may involve a model-based approach or a combination of these steps. In the model-based approach, usually an underlying, not-directly measurable variable is estimated. Examples of information extraction are: i) determine the elementary aggregate of similar products sold [2,12], ii) correct for the missing data of road sensors [3], iii) remove outliers in the location data of ships [4], iv) determine the type of crop grown an a patch of land from a picture [4], v) determine if a company is innovative or not from web site texts [6], vi) determine the sentiment of social media messages [7], and vii) derive if a particular social media message indicates a person with a particular disease [13]. From these examples it is clear that very divers methods can be applied in this step.

## Step A: Frame creation

In addition to the steps listed above, somewhere in the process a link has to be made between the population included in the Big Data source and the target population of the statistics. Here, two situations may occur. The first is webscraping. Here, usually, the Business Register is used to select business websites (indicated by the dashed line in Fig. 1). Establishing the relation between a (statistical) business unit and a website is important here. For this, various approaches have been developed, such as URL-finding [14] or use the URL provided by the Chamber of Commerce [6]. By comparing some of the data on the website, such as the name, address and Chamber of Commerce number, with those in the business register one is able to check if the correct link has been made.

The second situation requires getting information from the units included in the Big Data source. This can be challenging and may sometimes seem impossible. A very challenging example is the population active in social media [7]. Here, accounts of persons and companies are included. Many of these accounts do not provide enough information, such as background characteristics, to determine the exact person or company to which the profile belongs. When that occurs, one can try to derive any information relevant for that (indirectly) from the data that is available. An example of this is deriving the gender (and age) from the profile of a Twitter account [15].

Examples of studies in which the target population frame has been successfully obtained from the Big Data source alone are: i) traffic intensity statistics based on road sensor data [3] and iii) AIS-based maritime travel data [4]. In all these cases, the population units are (nearly) completely included in the data used. To obtain the target population, it was essential to link the sensors to the roads location data [3] and remove the ghost ships and non-maritime ships from the AIS-data [4], respectively. In cases where the population is not fully included, one may need to use non-probability based approaches, apply capture-recapture methods or decide to use the Big Data extracted information as a proxy. Examples of the latter are online job vacancies as a proxy for vacancy statistics [4] or social media sentiment as a proxy for economic mood [7].

## Step 3: Inference

In the end of the process the frame and the information extracted meet. In this step, nearly always a model-based approach is used in which one aims to infer statistics for the entire population [10] and corrects for any biases [6]. In a number of cases this is the second step of the process where models are used; e.g. no. 2 and 11 in Table 1. Examples of step 3 are i) creating a CPI-index based on the products sold [2], ii) estimating the traffic intensity for road segments [3], iii) estimating the number of small and large innovative companies in a country [6], and iv) estimating the number of susceptible with COVID19-symptoms on social media [13].

# Conclusions

In this paper, we propose a general process for using Big Data as the main data source in official statistics production. This process is composed of two data processing steps, one inference step and one population oriented step. We have meticulously compared this process with those documented for the 13 Big Data based statistics examples listed in Table 1. It was found that all examples at least described the first two data processing steps. For three cases all steps have been implemented and estimates for the target population have been produced; i.e. no. 1, 2 and 11 in Table 1. For some cases the results are used as a proxy for the target population; i.e. no 3-5, 12 and 13 in Table 1. Here, linking the Big Data units to the target population and correcting for the missing part of the population remain the biggest challenges. Some sources enable an expansion of regular statistics, e.g. no. 6, 8 and 10 in Table 1. All others have not reached these stages yet, but are well on their way to produce Big Data based statistics.

# References

1. Daas, P., et al. First draft of the Methodological report. Deliverable K5, ESSnet Big Data II (2020).
2. Eurostat. Practical Guide for Processing Supermarket Scanner Data. Working paper, (2017) September.
3. Puts, M.J.H., Daas, P.J.H., Tennekes, M., de Blois, C. Using huge amounts of road sensor data for official statistics. AIMS Mathematics 4, (2019), 12-25.
4. ESSnet Big Data I, Final deliverables of Workpackages 2-7, (2018).
5. ESSnet Big Data II, Final deliverables of Workpackages B-J, (2020).
6. Daas, P.J.H., van der Doef, S. Detecting Innovative Companies via their Website. Statistical Journal of IAOS (2020), *in press*.
7. ISTAT. Social Mood on Economy Index. Methodological note, (2020), Webpage.
8. Nurmi, O., Piela, P. The Use of Mobile Phone Data in Tourism Statistics. Paper for the 62nd ISI World Statistics Congress (2019), Kuala Lumpur, Malaysia.
9. Ahas, R., Tiru, M., Saluveer, E., Demunter, C. Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. Paper for the NTTS. (2011), Brussels, Belgium.
10. Consten, A. et al. Report describing the methodology of using Big Data for official statistics and the most important questions for future studies. Workpackage 8, (2018), ESSnet Big Data I
11. Puts, M., Advanced big data sources - Mobile phone and other sensors, ESTP training course, (2016), slide 5.
12. Willenborg, L. Transitivity of price indices, CBS discussion paper, (2018), May.
13. Puts, M., Daas, P. Applications of big data to official statistics: Social Media. Webinar 3 of ESTP Big Data training course, (2020), slides 27-33.
14. Van Delden, A., Windmeijer, D., ten Bosch, O. Searching for business websites, CBS Discussion paper, (2019).
15. Daas, P.J.H., Burger, J., Quan, L., ten Bosch, O., Puts, M. Profiling of Twitter Users: a big data selectivity study. Discussion paper 201606, (2016), Statistics Netherlands, The Hague/Heerlen, The Netherlands.