Profiling Big Data sources to assess their selectivity

Piet Daas (pjh.daas@cbs.nl)[1], Joep Burger (j.burger@cbs.nl)[1]

Keywords: Big Data, profiling, social media, feature extraction, selectivity.

1. Introduction

In our modern world more and more data are being created and remain to be stored. These kinds of data, generally referred to as Big Data, are very interesting sources of information. They, for instance, may reflect traces of human or economic activity and could possibly be used for official statistics [1]. However, extracting information from Big Data for such purposes is challenging for a number of reasons. First, not all data are relevant for the research question at hand, which requires one to find the signal in the noise [2]. Second, most Big Data available are composed of events [3] and usually provide very little or no information on the unit that generated the data. Third, if information is available on the creator of the data it may not be easily linked to a specific person or company. Fourth, not all units in the target population that the researcher envisaged may be included in Big Data and the ones that are included are not a random sample from the target population. All in all, these issues make it challenging, to say the least, to use Big Data for the creation of official statistics. In this paper we will mainly focus on the fourth challenge, i.e. how to assess the selectivity of a Big Data source.

1.1. Social media as an example

Let's illustrate the research question at hand with an example. Many people in the Netherlands are active on social media: 70% of the population posted messages according to a recent European study [4]. Compared with a probability sample this is an extremely high coverage rate. In contrast to a probability sample, however, we do not know to what extent these social media accounts represent our target population. Also quite a number of social media accounts actually reflect the activity of companies (even though they are created by humans). These are different target populations: for persons the target population are the persons included in the population register of the Netherlands, while for companies these are the units in the Dutch statistical business register. Also, hypothetically, cyber savvy and extravert people are more likely to be active on social media than computer novices and introverts. In addition, not all activity is publicly available as some social media messages are private only. However, these are not uncommon issues as selective non-response in sample surveys also causes a deviation from representativeness. Without correction for selectivity, estimates will be biased.

2. Methods

A common method used to assess selectivity in sample surveys is by comparing the distribution of relevant background characteristics in the data source with their known distribution in the target population. In principle, the same approach could be applied to Big Data, although in practice this is not trivial [5]. In an ideal world, units are linked to a population register containing background characteristics. Our experiences on studying Big Data sources have revealed that many units hardly provide any information that could be used to deterministically link them to a population register. In a more realistic Big Data context, background characteristics will have to be derived from the Big Data source. The key question is how should this be done? We think that an approach called

[1] Statistics Netherlands, Heerlen, the Netherlands.

‘profiling’ is the best option. Here the term profiling refers to an approach from the field of information science. In this approach, large amounts of data are analysed with the aim of discovering patterns to discern groups of similar units [6]. In this abstract, we will discuss the topic of profiling units from a social media perspective, since social media are quite challenging in this respect and a concrete example helps to illustrate the challenges more clearly. Social media also have the advantage from an experimental point of view, since a lot of data is publically available and each unit in the population has a unique identifier: a user id. This in contrast to many other Big Data sources, of which the data may be owned by private companies or that may have computers or other electronic devices as units [1].

3. RESULTS

From an earlier study performed at Statistics Netherlands in cooperation with Erasmus University [7] we obtained a list of 380 thousand Twitter usernames which were—at that point in time and according to the location information on their user profile—all identified as Dutch Twitter users. This list is the starting point for the studies described below. Based on this list we will describe ways that could be used to profile Dutch Twitter users. Some of the methods have already been tested and some have not. For these studies only data available on public Twitter user accounts are used, meaning that anyone with a PC, a browser and an internet connection could access the data we studied. Since we respect the privacy of the users, no information on or examples of individual users will be provided.

3.1. Background characteristics

Many surveys use a similar set of background characteristics that correlate with target variables. In social statistics commonly used variables are: gender, age, income, education, origin, urbanicity, and household composition. For companies often used characteristics are: number of employees (size class), turnover, type of economic activity, and legal form. Because both persons and companies are active on social media, the first distinction that needs to be made is if the owner of an account represents a person or a company. Note that company is used here as a general term to describe both businesses and other public organizations. The distinction between persons and companies could be made by studying the username, user profile, link to the (company’s) website on the profile, the profile photo, the tweet frequency and the tweet content. As it is to be expected that companies will preferably be active on social media with a highly similar or identical name, a list of Dutch company names could be used to quickly identify them. It is to be expected that accounts of self-employed people and small-companies will be more difficult to distinguish from non-commercial personal accounts.

If an account is used by a private person, how could one determine its gender? This can be obtained from the username, the profile (bio-)information, the profile picture, an associated account (such as an account on LinkedIn) or the combined set of public tweets of that person. Combing the information will likely increase the chance of success. The same holds for age, where some first names occur more frequently for persons born in a particular period. The types of words and abbreviations used in tweets also provide clues on the age of the person writing them. Urbanicity could be derived from the location information in the user profile or from the location related content in the messages written. Level of education could be obtained from the job description (if available), the content of the messages related to this topic or from an associated LinkedIn account. The latter may also provide clues on income as does the location information in the user profile. Origin and household composition might additionally be derived by analysing the social network of a user.

For companies a lot of the required information could be obtained via the associated website. It is to be expected that companies provide such a URL in their official Twitter profile. On this website documents, such as company publications and links to annual reports, might be available from which more detailed information, such as turnover and number of employees, could be obtained. However, some companies might not be active on social media or someone could have created a 'fake' company account.

4. Conclusions

To check the viability of the above mentioned approaches, we start by manually profiling a sample of the Twitter database. This provides a dataset containing both input features (e.g. username, profile image, tweets) and outcome measurements (e.g. gender, age). The dataset is subsequently split (horizontally) into a training set and a test set. We can use these to train and test various approaches developed in the field of information sciences, such as artificial intelligence and machine learning methods. Here, the training set is used to train the algorithm the relation between input and output. The test set is used to compare predicted output with observed (manually profiled) output and to optimize the parameters of the algorithm. With these techniques one attempts to predict the background characteristics of the rest of the Twitter database using the input features. To enable the use of Big Data in official statistics it is important that successful profiling approaches are being developed.

References

[1] M. Glasson, J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, A. Khan, What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, (2013).

[2] N. Silver, The Signal and the Noise: Why So Many Predictions Fail—but Some Don't, Penguin Group, New York, USA, (2012).

[3] P.J.H. Daas, M.J.H. Puts, B. Buelens, P.A.M. van den Hurk, Big Data and Official Statistics. Journal of Official Statistics, NTTS special issue, (2014), accepted for publication.

[4] Eurostat, Internet access and use in 2012. Eurostat newsrelease, (2013), Located at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF

[5] B. Buelens, P. Daas, J. Burger, M. Puts, J. van den Brakel, Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands, (2014).

[6] M. Hildebrandt, S. Gutwirth, Profiling the European Citizen. Cross Disciplinary Perspectives. Springer, Dordrecht, (2013).

[7] P.J.H. Daas, M. Roos, M. van de Ven, J. Neroni, Twitter as a potential data source for statistics. Discussion paper 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands, (2012).