On Big Data based statistical inference

Piet J.H. Daas, Marco J.H. Puts & Robbert Renssen

Statistics Netherlands, Center for Big Data Statistics

INTRODUCTION

Big data is a hot topic and many are studying its potential application. This is usually not done from a generic viewpoint. For official statistics, the latter is essential and it is becoming more and more apparent that -to produce reliable big data based statistics- a new way of working is required [1]. Henceforth, there is an urge to have a general, fundamental statistical framework to describe phenomena hidden in big data.

Suppose a researcher wants to determine the mean of the unknown variable μ of a target population.

Administrative source (covering whole population): determine y for i=1,…,N_{pop}; with y_i for all N units in the population (1).

Big data source (covering unknown part of population): determine z for the big data population for j=1,…,M_{BD} (2).

In both cases it is assumed that y and z are approximations of μ. Hence the mean of μ is approximated by:

$$\bar{\mu} \cong \hat{\bar{y}} = \frac{1}{N}\sum_{i=1}^{N_{pop}} y_i \quad (1), \qquad \bar{\mu} \cong \hat{\bar{z}} = \frac{1}{M_{BD}}\sum_{j=1}^{M_{BD}} z_j \quad (2)$$

In (2) it is assumed that the big data source covers a huge part of the target population, i.e. the units missing can be ignored.

The more z is related to y and the more the big data population M is equivalent to the target population N, the more likely the series of z and y correlate and cointegrate. The equations indicate two important issues:

i) Variables y and z are assumed to be approximations of the original concept μ. Because of differences a bias may be introduced.

ii) Differences between the units in the target population and the population generating the data in the big data source may result in biased estimates.

VARIABLES

Different variables can measure comparable phenomena. An interesting example of this are the time series of consumer confidence (survey based) and social media sentiment (big data based) in the Netherlands. Both series show a remarkable high correlation; r = 0.9 [2, 3]. Another example is the high correlation between the quarterly GDP of the Netherlands and the traffic intensity in the country [4]. To assure the quality of the data in big data sources specific checking and correction methods have been developed [5].

Center for Big Data Statistics

POPULATIONS

Information on the composition of the units in a big data sources is important for reliable estimates. However, this is not an easy task as many of the commonly used identifiers for statistics, so-called background characteristics, are absent in many big data sources. Extracting features is a way to deal with that challenge. For Twitter, it has been demonstrated that gender can be reliably derived from a combination of features [6]. The approach shown in the table below resulted in an estimate for gender with an accuracy of 96.5%.

Unassigned before (%)	*Assigned (% of total)*	*Feature used to assign gender*
100	18	Short bio
82	64	First name
18	14.6	Tweet content
3.4	1	Picture (faces)
2.4	2.4	Assign male

MIXED POPULATIONS

Another issue is that social media messages are created by mixed populations of units. Persons, companies and other types of organizations are active on it. These accounts might be Dutch or not.

Determining if an account is Dutch (or not) is essential when studying the population relevant for official statistics. The population producing Twitter messages in 2015 contained 9.1 million unique Twitter user IDs. A very computational efficient process was developed to identify Dutch accounts with an overall recall of 93% (Figure 1). In the end, nearly 1.4 million Twitter user IDs were identified as Dutch.

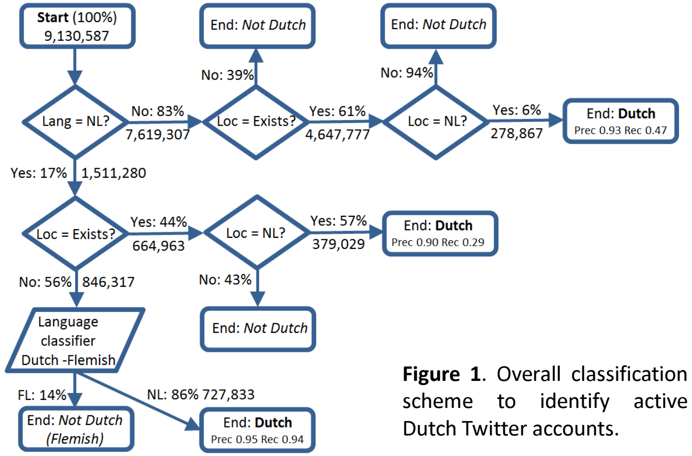

Figure 1. Overall classification scheme to identify active Dutch Twitter accounts.

Discerning between accounts of persons and companies is also essential for official statistics. For Twitter accounts, a total of 99 unique features were extracted (Figure 2). Next, Machine Learning based approaches were trained and their performance to discern between persons and non-persons and between companies and other non-persons were evaluated. Persons could be discerned with an accuracy of 94%. From the non-persons part, companies could be identified with an accuracy of 80%.

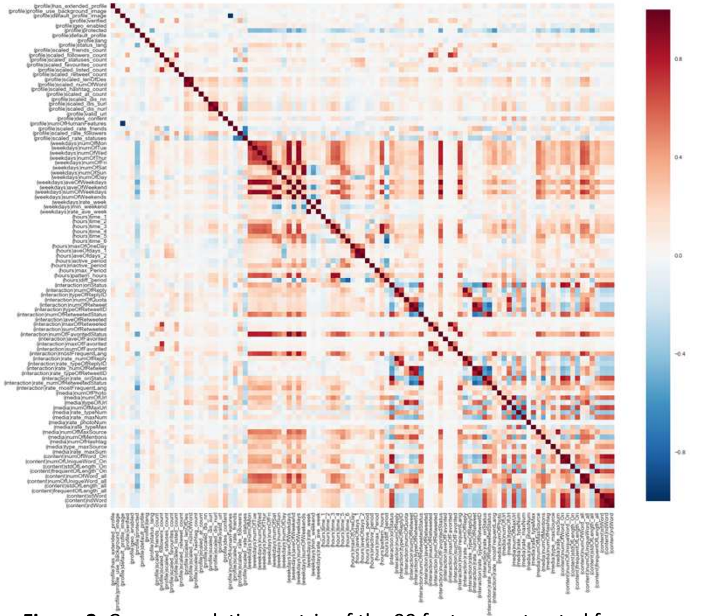

Figure 2. Cross-correlation matrix of the 99 features extracted from Twitter account data and their associated tweets.

To conclude. It is important that statisticians realize that a concept can be measured, i.e. operationalized, in various ways. This is not only important in the context of Big Data, but also for any other data source used; e.g. admin and survey data. Big question is: What do we want to measure?

The authors thank Ali Hürriyetoglu, Agata Troost, Joep Burger, Chang Sun and Quinty Daas for their assistance in preparing this poster.

REFERENCES

[1] WP8 (2017) Results of Work Package 8, on Methodology, Quality and IT, of the ESSnet Big Data.

[2] Daas, P.J.H., Puts, M.J.H. (2014) Social Media Sentiment and Consumer Confidence. *European Central Bank Statistics Paper Series No. 5*, Frankfurt, Germany.

[3] Van den Brakel, J., Söhler, E., Daas, P., Buelens, B. (2017) Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, in press.

[4] Daas, P.J.H., Braaksma, B., Aly, R., Engelhardt, Y., Hiemstra, D., Zurita Milla, R. (2016) Big Data Masterclass and DataCamp 2015. *Discussion paper 201615*, Statistics Netherlands,.

[5] Puts, M., Daas, P., de Waal, T. (2017) Finding Errors in Big Data. In: *The Best Writing on Mathematics 2016*, Princeton, USA. (Pitici, M., ed), pp. 291-299, Princeton University Press, USA.

[6] Daas, P.J.H., Burger, J., Quan, L., ten Bosch, O., Puts, M. (2016) Profiling of Twitter Users: a big data selectivity study. *Discussion paper 201606*, Statistics Netherlands.