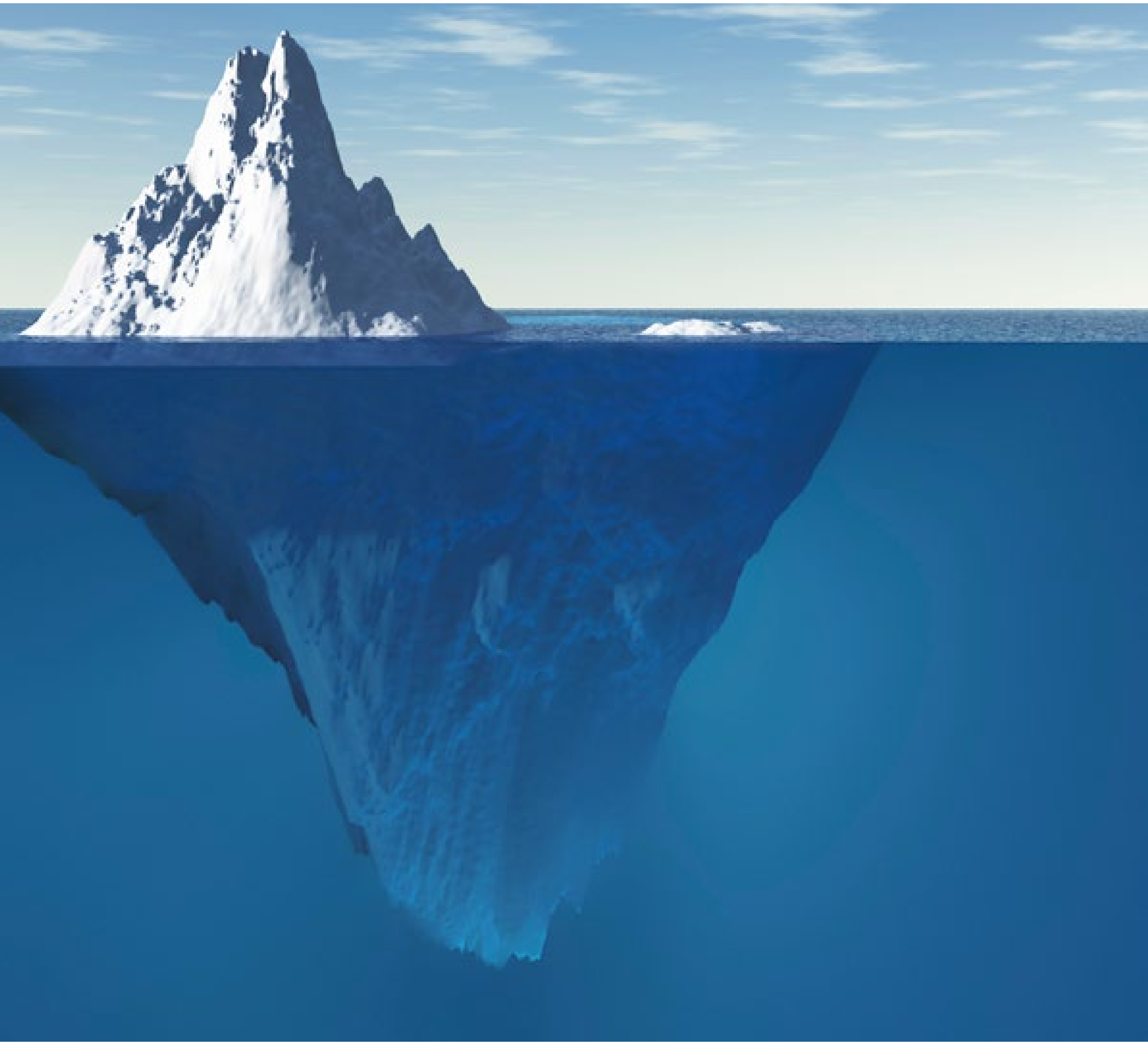
An overview of methods for treating selectivity in big data sources

MACIEJ BERĘSEWICZ, RISTO LEHTONEN, FERNANDO REIS, LOREDANA DI CONSIGLIO, MARTIN KARLBERG

2018 edition



STATISTICAL WORKING PAPERS

eurostat

An overview of methods for treating selectivity in big data sources

2018 edition

MACIEJ BERĘSEWICZ, RISTO LEHTONEN, FERNANDO REIS,
LOREDANA DI CONSIGLIO, MARTIN KARLBERG

Manuscript completed in July 2018

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2018

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Theme: General statistics
Collection: Statistical working papers

ISBN 978-92-79-88769-7 ISSN 2315-0807 doi: 10.2785/312232 KS-TC-18-004-EN-N

Contents

Abstract 6

Executive summary 7

Introduction to the study 7

What are big data? 7

Errors impairing inference from big data sources 8

Framework for focusing on selectivity 8

Selectivity issues for different data sources 9

Methods for correcting for selectivity 10

1 Introduction 13

1.1. Big data activities in official statistics at European level 13

1.2. A study on methods for treating selectivity in big data sources 14

1.3. Structure of the report 16

2 A statistical approach to big data 17

2.1. Defining big data 17

2.1.1. The three Vs of big data 17

2.1.2. Big data as non-probabilistic sample data 17

2.1.3. Designed data vs. organic data 18

2.1.4. Big data in the context of statistical data sources 18

2.1.5. Statistical features of big data 19

2.1.6. Enumeration of big data sources 20

2.1.7. A statistical definition of big data 22

2.2. Statistical characteristics of big data 22

2.2.1. Mediation by electronic platforms 22

2.2.2. Imperfect coverage 23

2.2.3. Missingness, selectivity and bias 24

2.2.4. Variables not observed directly 25

2.2.5. Unit identification problem 27

2.3. Addressing selectivity in big data 29

2.3.1. Big data as an opt-in panel survey 29

2.3.2. Big data survey design 31

2.3.3. The use of external sources to address selectivity 32

2.3.4. Measuring the self-selection process 33

3 Analysis of specific big data sources 35

3.1. Mobile network data 35

3.1.1. Description of the data source 35

3.1.2. Populations observed in mobile network data 36
3.1.3. Coverage of mobile network data 38
3.1.4. Selectivity in mobile network data 38
3.1.5. Selectivity of infrastructure — BTS and cells........ 39
3.1.6. Self-selection process on the mobile phone market — is it ignorable or not?........ 41
3.1.7. Limitations of mobile network data 42

3.2. Online social networks — Twitter........ 43

3.2.1. Description of the data source 43
3.2.2. Populations observed in Twitter data 44
3.2.3. Coverage of Twitter data 45
3.2.4. Selectivity in Twitter data 45
3.2.5. Self-selection process on Twitter — is it ignorable? 47
3.2.6. Limitations of Twitter data 48

3.3. Web activity data — Google Trends 49

3.3.1. Description of the data source 49
3.3.2. Populations observed in Google Trends 49
3.3.3. Coverage of Google Trends 50
3.3.4. Selectivity in Google Trends 51
3.3.5. Limitations of Google Trends data 52

3.4. Web activity data — Wikipedia usage 53

3.4.1. Description of the data source 53
3.4.2. Populations observed on Wikipedia 55
3.4.3. Limitations of Wikipedia data 55

4 Unit-level methods to correct selectivity 56

4.1. Pseudo-design approach — reweighting 56

4.1.1. Generalised weight share method 57
4.1.2. Model-free calibration and extensions 57
4.1.2.1. Model-free calibration 57
4.1.2.2. Extensions 58
4.1.3. Model-assisted calibration 60
4.1.4. Propensity weighting 62
4.1.5. Pseudo-empirical likelihood 63
4.1.6. Adjusted weights 64
4.1.7. Two-step weighting method 65
4.1.8. Software 67

4.2. Modelling approach 67

4.2.1. Heckman selection models 67
4.2.2. Small area estimation approach 68
4.2.3. Bayesian approach 68

4.2.3.1. Hierarchical Bayesian approach ... 69
4.2.3.2. Calibrated Bayes approach ... 70
4.2.3.3. Pattern mixture models — MNAR case ... 71
4.2.4. Machine-learning approach ... 72
4.2.4.1. K-nearest neighbours ... 72
4.2.4.2. Artificial Neural networks ... 72
4.2.4.3. Classification and regression trees ... 73
4.2.5. Software ... 74
4.3. Data linking approach ... 74
4.3.1. Record linkage ... 75
4.3.2. Sample matching ... 75
4.3.3. Software ... 76
5 Domain-level methods to correct selectivity ... 77
5.1. Pseudo-design methods — reweighting ... 77
5.2. Modelling approach ... 79
5.2.1. Direct estimation of bias ... 79
5.2.2. Blending of estimates ... 81
6 Conclusions ... 83
6.1. Sources of selectivity and correction methods in big data ... 84
6.1.1. Mobile network data ... 84
6.1.2. Twitter data ... 85
6.1.3. Google Trends data ... 86
6.1.4. Wikipedia data ... 87
6.1.5. Summary ... 87
6.2. Self-selection adjustments for big data sources ... 88
6.3. Recommended framework for adjusting selectivity in big data ... 89
Bibliography ... 91
Mathematical appendix ... 103
A.1 Notation ... 103
A.2 Quantifying selectivity ... 105
A.3 Bias quantification ... 107
A.4 Resampling with big data ... 110

Abstract

Official statistics is now considering seriously big data as a significant data source for producing statistics. It holds the potential for providing faster, cheaper, more detailed and completely new types of statistics. However, the use of big data brings also several challenges. One of them is the non-probabilistic character of most sources of big data, as very often they were not designed to produce statistics. The resulting selectivity bias is therefore a major concern when using big data. This paper presents a statistical approach to big data, searching for a definition meaningful from the statistical point of view and identifying its main statistical characteristics. It then argues that big data sources share many characteristics with Internet opt-in panel surveys and proposes this as a reference to address selectivity and coverage problems in big data. Coverage and the self-selection process are briefly discussed in mobile network data, Twitter, Google Trends and Wikipedia page views data. An overview of methods which can be used to address selectivity and eliminate, or mitigate, bias is then presented, covering both methods applied at individual level, i.e. at the level of the statistical unit, and at domain level, i.e. at the level of the produced statistics. Finally, the applicability of the methods to the several big data sources is briefly discussed and a framework for adjusting selectivity in big data is proposed.

Keywords: Big data, Selectivity.

Acknowledgements: This report is the result of a study to identify methods for treating selectivity in big data sources contracted[1] by the European Commission (Eurostat) to SOGETI. The study was conducted by Maciej Beręsewicz and Risto Lehtonen with terms of reference set up by Martin Karlberg, Loredana Di Consiglio and Fernando Reis. Maciej Beręsewicz was the main drafter of the report with contributions from the remaining authors. The report also benefited from the comments of the members of Eurostat task force on big data and of three external reviewers.

Authors:

Maciej Beręsewicz[2], Risto Lehtonen[3], Fernando Reis[4], Loredana Di Consiglio[5], Martin Karlberg[6]

[1] QTM9 of ESTAT/11111.2013.001-2015.084 under FWC ESTAT/11111.2013.001-2013.254

[2] Poznań University of Economics and Business, maciej.beresewicz@ue.poznan.pl

[3] University of Helsinki, risto.lehtonen@helsinki.fi

[4] European Commission (Eurostat), fernando.reis@ec.europa.eu

[5] European Commission (Eurostat) at the time of drafting the report. Can be contacted via diconsig@istat.it

[6] European Commission (Eurostat), martin.karlberg@ec.europa.eu

Executive summary

The European Statistical System (ESS) has committed itself to exploring the potential of big data for producing official statistics by adopting the Scheveningen Memorandum (*ESSC, 2013*) in 2013 and the **Big Data Action Plan and Roadmap** (*ESSC, 2014*) in 2014.

To implement the action plan, Eurostat launched several initiatives to explore the potential of big data and to address its challenges, such as the **ESSnet Big Data** project[7] and a **study covering ethics, communication, legal environment and skills**. Eurostat also launched a series of **in-house big data pilots** aimed at building internal technical expertise and inferring from its own experience the implications of big data at strategic level for official statistics in general, for the ESS and for Eurostat and the European Commission. One general challenge during these pilots was the selectivity of the big data sources used, since the fact that individuals self-select whether or not to use the technologies where big data are captured renders the samples biased.

Introduction to the study

The purpose of the methodological study which produced the results presented in this report was to help Eurostat address the selectivity of the big data sources used in its own pilots. The study also aimed to guide Eurostat in planning future development activities, internally and at the ESS level. It was much more specific than the methodologies analysed in the ESSnet Big Data project since it only addressed selectivity, making it possible to gain insights much more quickly.

The **main objective** of the study was to identify existing methods which could be used to address the selectivity in big data sources, in order to be able to make unbiased inference for populations of interest in official statistics (e.g. resident population between 15 and 65 years old).

The **approach** taken in the study was to address selectivity as a general term for selection errors resulting from:

- (self-selection) decisions of individuals (e.g. whether to tweet or use a particular mobile provider),
- decisions of the owners of the electronic platforms where data are captured (e.g. in terms of business concept, technical infrastructure), or
- the limitations of the technologies.

As a result, selectivity causes coverage, measurement and non-response (or missingness) errors, which introduce potential bias in estimates based on big data sources.

What are big data?

Previous attempts to identify or define big data highlighted the characteristics of this type of data. The level of detail is probably what characterises big data most. All other data types lack the level of detail we recognise in big data. This is certainly the result of these data being generated or captured in an automated way using IT systems or sensors. The automated generation or capture of highly detailed data results in massive datasets of very large volume. The ubiquity of IT systems, sensors and digitisation in our lives results not only in an extremely large overall volume of data in our society, but also in a huge variety of forms of data. Most of these data are organic and unstructured, and are rarely designed for statistical purposes. A special case is sensors. Sensors are sometimes set up specifically to collect data for statistical purposes, but have all the characteristics that distinguish big data from other types of data traditional in statistics.

A possible definition of big data which includes all these characteristics is that **big data are highly**

[7] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata.

detailed exhaust data automatically captured by sensors or generated during the use of IT systems.

Errors impairing inference from big data sources

From the statistical inference point of view, what really matters is the way in which big data are typically generated or captured. Specifically, the fact that big data are non-probabilistic, organic and unstructured means that they seldom allow us to directly measure our variables of interest, but rather enables us to measure proxies.

Big data are so highly detailed because of the automated data capture provided by **electronic platforms**. Just like questionnaires influence what is observed in traditional surveys, the electronic platforms providing the data capture in big data sources, together with the behaviour of individuals while interacting with those platforms, influence what is observed.

There are three consequences for big data of these electronic platforms mediating the data capture process. Firstly, the data captured often, but not always, do not correspond to the variables of interest for statistics. For example, the data captured in social networks in the form of natural language textual data are not suitable for use directly in statistical analysis. They need to be transformed into some variable of interest, e.g. the sentiment expressed in the communication, which is **unobserved** (i.e. not directly observed). This transformation involves **measurement errors** in the variables of interest.

Secondly, the most atomic object (e.g. a tweet) to which the data captured relates is typically not the statistical unit of primary interest (e.g. a person) and is sometimes only indirectly related to the statistical unit. For example, in mobile network data the most basic object is a communication event, about which we know the location of the antenna used for the communication and the identity of the mobile device used. The statistical unit of interest for statistics would be the individual using the mobile device. The statistical unit is certainly related to the observed object, because for each communication event there is a mobile device and each mobile device is used by one or more individuals. However, there are **several objects** observed in the data source and **the link with the statistical unit of interest involves uncertainty and errors, commonly known as unit error**.

Thirdly, access to and use of the electronic platform is a pre-condition for capturing the data. Those individuals (or more generically, objects) with no access to the platform or who do not use it will not be observed. If the data capturing platform is not deliberately set up for statistical purposes, the part of the population of interest observed will most probably be captured by non-random selection and the data will suffer from problems of **selectivity**.

Framework for focusing on selectivity

While all three types of error — measurement error, unit error and selectivity — have to be treated in order to conduct inference, this report focuses on selectivity. We propose a framework that separates selectivity treatment from the treatment of the other two types of errors. **The framework assumes that the target statistical unit, population and variables have already been defined based on user needs.** It includes an exploration phase that is needed before the big data source is used and which can run separately. It then includes a stage of pre-processing that addresses unit error and measurement errors. Although this stage does not address selectivity directly, it has implications for bias and successfully applying the adjustment methods. The following stage addresses the identification of the selectivity, by comparing distributions for background variables between the big data and additional sources and, if possible, estimates for the target variable. Finally, the framework helps identify the suitable method for dealing with selectivity error, giving priority to unit-level methods where possible.

In order to address selectivity, we can consider big data sources as a special case of an **internet opt-in panel survey**. In big data sources, we indirectly observe statistical units over time, just like in a panel survey. Common phenomena in panel surveys, such as attrition and births, are similar to phenomena in big data sources (e.g. changes to other mobile providers or decreasing interest in a particular social media, new customers coming to the provider, or new accounts being created). The self-selection mechanism in big data is similar to the opt-in mechanism in surveys. Finally, just like big data sources in general, internet surveys have frame issues because they work using an electronic platform (the internet in that specific case).

Approaching big data as a special case of an internet opt-in panel survey opens the door to methods that already exist in survey research literature and practice. In fact, many problems found in big data sources have previously been studied in the context of sample surveys (coverage, non-response, self-selection, measurement error).

Selectivity issues for different data sources

The several big data sources differ significantly, and therefore the selectivity mechanisms and the appropriate methods for treating that selectivity may be very different.

In the case of **mobile network data**, the data refer to communication events, base transceiver stations and SIM cards, rather than directly to the target population of individual persons. These objects are linked to individual people via SIM cards and corresponding mobile phone numbers. However, people can have more than one mobile number, and some SIM cards do not correspond to individual people. Eliminating or mitigating unit error, and its resulting bias, depends on the successfully identifying these cases. Selectivity (as well as measurement error) depends both on:

- the overall coverage of the target population by the mobile operators' samples to which we have access, and
- the overall coverage of the infrastructure of the mobile operators supplying the data.

The biggest challenge for correcting selectivity in this big data source is the difficulty in obtaining auxiliary information at unit level, which could be used to correct selectivity.

Twitter data suffer from both over-coverage and under-coverage. For example, some Twitter accounts are set up with the purpose of manipulating or influencing public opinion and would normally not be part of the target population. The development of algorithms to identify these accounts is challenging and still requires further research. 46 % of individuals living in the EU aged between 15 and 65 did not regularly use online social media in 2017 and were therefore not covered.([8]) The self-selection in Twitter data depends not only on the decision to be a Twitter user, but also on the particular circumstances in which individuals decide to tweet. There is evidence that the behaviour of individuals differs between different socioeconomic groups, so measuring these differences is fundamental to ensuring the correct level of selectivity. Information on many users is available or can be deducted from related information shared by the user, and this can be used to mitigate selectivity. Moreover, it is possible that non-representative samples of social media users can post content that potentially represents the larger population's opinions and experiences.

Google Trends is a web-based data source and therefore also suffers not only from a biased frame (internet users), but also the final data sample composition shifting significantly around major events. The biggest disadvantage of Google Trends is that individual data are not available at any level, which limits the applicability of selectivity correction methods. The direct use of Google Trends for statistics is therefore problematic, due to the lack of background information on the person that used Google's search engine. However, Google Trends is most commonly used as a proxy variable of a given phenomenon and as a covariate in statistical models used for nowcasting.

Wikipedia usage data are composed of several types of data. Besides the content of the articles themselves sourced by the users and several derivatives (e.g. Wikidata), there are data available on

([8]) Source: Eurostat (isoc_ci_ac_i).

the additions made to the articles and the number of times each article is consulted in every hour (page views). Just like Google Trends, there is no background information about which users view a page, but for the content and editions of the articles there is information at the level of the article and on some of the users.

Methods for correcting for selectivity

Methods to correct for selectivity can be applied at individual level, i.e. at the level of the statistical unit, or at domain level, i.e. at the level of the statistics produced.

Unit-level methods to correct selectivity include;

- pseudo-design methods, based on reweighting the individual records,
- model-based methods, and
- data linking approaches.

Reweighting is based on the assumption that population-level information of auxiliary variables comes from reliable sources (e.g. demographic characteristics from official statistics). If this assumption holds, then bias due to selectivity can often be adjusted and estimates made more efficient by introducing the available auxiliary data in the estimation procedure. Reweighting can be effective in reducing the selection bias of the target estimate if the auxiliary variables explain the selection mechanism and are related to the survey variables of interest. This assumption applies to most of the bias adjustment methods considered here.

Model-free calibration is the most common reweighting method in official statistics for improving the quality of sample estimates by using auxiliary information. Calibration is usually used to achieve coherence between the survey estimates and published official statistics, to improve the efficiency of estimates, and to account for the bias due to non-response. The method can also be used for correcting selectivity in big data sources when we can find auxiliary variables which are correlated with the selectivity mechanism and for which we know the population totals. Raking and post-stratification are two special cases of model-free calibration. Important extensions have been developed, including one-step and two-step calibration approaches for adjusting for ignorable and non-ignorable missingness.

Just as for reweighting, **model-assisted calibration** can be used to correct selectivity in big data sources when we have auxiliary variables which are correlated with the selectivity mechanism. However, contrary to model-free calibration where we use information on the population totals, in this case we use auxiliary variables at individual level for the entire population. This allows us to use flexible assisting models beyond the conventional linear model, e.g. generalised linear mixed models family members, and to obtain more precise estimates.

Weighs can also be adjusted directly to account for under-coverage and self-selection in big data sources. These **adjusted weights** would then reduce bias under certain conditions. An initial base weight can be adjusted for coverage and missingness and calibrated to auxiliary variables with correction factors which can be estimated independently from other data sources.

The **two-step weighting** method differs from the previous ones because instead of correcting initial weights, it finds the estimation weights by directly addressing the joint distribution between the variable of interest and auxiliary and paradata variables. It is based on an explicit behavioural model of access to the electronic platform capturing the big data and the usage decisions of individuals which generate the data. The method uses a supplementary, probability-based sample to correct these two sources of bias.

Propensity weighting is a commonly used method for reducing non-response bias in sample surveys. The method is based on directly modelling the probability (i.e. propensity) that each statistical unit is present in the big data sample. The propensities, when predicted with a model which accounts for the selectivity mechanism, can then be used to estimate statistics corrected for the selectivity in the big data source.

The **generalised weight share method** is useful when weights cannot be calculated directly for the statistical units but can be for the objects for which data are available from the big data source, as long as it is possible to link those objects to statistical units in an existing frame population. Basically this method aims to distribute the (possibly corrected) weights of the objects among the units of the target population.

The pseudo-empirical likelihood method provides yet another method for accounting for the selection bias, assuming the population means of the auxiliary variables are available.

The second group of methods, based on a **modelling approach**, relies on the use of the big data to estimate a model which will then be used with a representative sample to make estimations for the target population. In order for this approach to be effective, the model built on the big data sample should hold for the target population and should be as assumption-free as possible. Auxiliary data that contain information to account for self-selection should also be available and be included in the model specification. The modelling can be based, for example, on econometric selection models, a small area estimation approach, a Bayesian approach or a machine-learning approach.

Model-based **selection models** have been developed in econometrics to handle ignorable and non-ignorable missingness. In this approach, the selection mechanism is modelled in addition to the survey variable of interest by using the available auxiliary and covariate data.

In the **small area estimation approach**, the big data sample is combined with data from a representative sample (either from an existing survey or one launched for this purpose). Statistical models are used to link a target variable in the representative sample to correlated variables from the big data sample and to make estimations for small domains (i.e. small area). Combining the model prediction and the standard direct estimate for each area in a sensible way makes it possible to balance the bias in the big data and improve its ability to provide higher precision in the estimates for the target variables. Various approaches, including frequentist and Bayesian ones, and many commonly used model families are used in small area estimation.

In the **Bayesian approach**, several methods have been developed, for example hierarchical Bayesian methods and calibrated Bayes methods. For example, these Bayesian methods can be used to produce estimates for very fine strata from a big data source, and the results are then grossed up to the target population with information from the relative weight of each stratum taken from a frame or a representative sample. Another approach is the **pattern mixture model** which, instead of modelling the missingness mechanism directly, postulates the impact on the target variable of the missingness.

The last of the model-based approaches treated in this study is the **machine-learning approach**. In this approach a machine-learning algorithm which relates the variable of interest with some auxiliary variables which capture the selectivity process is trained on the big data. The variable of interest is then predicted in a frame population or for a representative sample which has the auxiliary variables. Machine learning is very promising because of its ability to model very complex and non-linear relationships between the variable of interest and auxiliary variables, and because of it usually has a very high level of predictive accuracy.

In a **data linking approach** to adjusting self-selection bias, data from a big data source are linked at individual level with data from a frame population or a representative sample. After the data are linked, unbiased estimations for the target population can be made either using a reweighting approach (such as post-stratification) or a modelling approach (such as prediction of the target variable). This normally uses auxiliary variables which account for the selectivity mechanism.

The **domain-level methods used to correct selectivity include:**

- pseudo-design methods, based on the reweighting of domain estimates (i.e. aggregated values for sub-populations — or even for the entire population), and
- model-based methods.

In the case of **reweighting** the domain-level estimates, the original estimate based on the big data is adjusted for the coverage of the electronic platform underlying the big data source, for the penetration rate of the electronic platform operator providing the data (if not all operators provide data), for the fraction of active users and an adjustment factor calibrating to known population totals. The adjustment factors may be known (from frame populations) or may need to be estimated independently.

The **modelling approach** at domain level assumes that an additional data source is available and provides ‘gold standard’ estimates of the target variable or a proxy of it. This additional data source can be a register or a survey based on a representative sample. One use of the gold standard for the target variable is to make a **direct estimation of the bias** present in the big data and to adjust the estimation based on big data. This estimation can be done by domain to account for possible bias levels depending on auxiliary variables. The gold standard can also be used to blend estimates from the big data and from the additional data to obtain an optimal combination from the unbiased (but more variable) estimate from a representative sample and the higher precision from a big data source.

1 Introduction

1.1. Big data activities in official statistics at European level

The European Statistical System (ESS) has committed itself to exploring the potential of big data for producing official statistics by adopting the Scheveningen Memorandum (*ESSC, 2013*) in 2013 and the **Big Data Action Plan and Roadmap** (*ESSC, 2014*) in 2014.

The Scheveningen Memorandum specifically acknowledges that the use of big data in the context of official statistics requires new developments in methodology and that the ESS should make a special effort to support these developments.

The overall purpose of the roadmap is to enable the ESS to gradually integrate big data sources into the production of European and national statistics and, in this way, contribute to the broader aims of the ESS Vision 2020. It is an immediate response to the demand for harnessing new data sources in statistical production. One of the topics identified within the ESS Big Data Roadmap is 'methods'; essentially what methods and methodologies will need to be employed to utilise big data sources within official statistics. The topic includes identifying and developing methods, and ultimately the production of guidelines and a toolbox.

Eurostat launched several initiatives to explore the potential of big data and to identify its challenges. The **ESSnet Big Data** project was launched in March 2016 for a maximum duration of 26 months. An ESSnet project consists of a network of several ESS organisations aiming to provide results that will be beneficial to the whole ESS. The ESSnet Big Data project was based on a set of pilots run by national statistical institutes. The purpose of the pilots was to explore the potential of using selected big data sources to produce official statistics and how the results could be applied to specific statistical domains. ESSnet aimed to generalise the findings of the pilots in terms of methodology, quality and IT infrastructure to identify how the selected big data sources from the pilots could be used in future within the European Statistical System.

In November 2015, Eurostat also launched a **study covering ethics, communication, the legal environment, skills and the organisation of a big data workshop**. The study aimed to share experience and was expected to last 23 months. The skills needs covered in this study are closely linked to the methodological challenges posed by the use of big data. One of the uses of the results of this study on selectivity treatment methods was to guide the identification of skills needs.

Most recently, Eurostat launched a series of **in-house big data pilots**. These aimed to build internal technical expertise and infer from Eurostat's own experience the strategic implications of big data for official statistics in general, for the ESS and for Eurostat and the European Commission. When running these pilots, Eurostat found one general challenge was the selectivity of the big data sources used.

1.2. A study on methods for treating selectivity in big data sources

The purpose of the study on methods for treating selectivity in big data, the results of which are presented in this report, was:

i. to help Eurostat address the selectivity of the big data sources used in its own pilots, and

ii. to guide Eurostat in planning future development activities, both internally and at ESS level.

This study was much more specific than the ESSnet Big Data project because it addressed only selectivity, allowing Eurostat to gain insights much more quickly.

The **main objective** of the study, which addresses both (i) and (ii), was to identify existing methods which could be used to address the selectivity in big data sources. Identifying these will help Eurostat make unbiased inference for populations of interest in official statistics (e.g. resident population between 15 and 65 years old).

The issue of selectivity, and the possibilities for addressing it, applies to all big data sources. However, the big data sources described in this report are taken as a starting point, in particular, mobile phone network data and social media which are data sources being explored by Eurostat.

The study lists the methods, as comprehensively as possible, describing their advantages and disadvantages and the level of maturity of each (i.e. whether the method is readily applicable or needs further development in terms of methodology or software tools.)

This study could then lead to further actions, such as developing methodologies, statistical tools or actions at European level.

To achieve these objectives, the study undertook the following **activities**:

- Activity 1 — Analysis of big data sources
  Based on the description of the particular big data sources and on a literature review of selectivity in big data sources, this activity identified the type of selectivity found in big data sources. The activity looked not just at selectivity issues but also at the characteristics of the sources since they may affect the applicability of methods to address selectivity (e.g. the existence or not of background characteristics).
- Activity 2 — Summary of literature on selectivity treatment methods

  This activity was composed of two sub-activities. The first one was a wide literature research on methods to deal with selectivity. This literature research spanned beyond the methods most commonly used in statistical offices and searched for potential methods in other statistical domains. The second sub-activity was a short summary of the relevant literature. The result of this activity is not included in this report and is made available online.[9]
- Activity 3 — Analysis and evaluation of methods

  Based on the description of the particular big data sources, including the specific causes of selectivity identified in activity 1 and on the literature summary conducted in activity 2, this activity consisted of selecting relevant methods and reviewing them in more depth. This took into

[9] https://ec.europa.eu/eurostat/cros/content/methods-treating-selectivity-big-data-sources_en

account not just selectivity issues but also the characteristics of the sources since they may affect the applicability of methods to address selectivity (e.g. the existence or not of background characteristics). The activity also involved classifying the literature review into categories. It also assessed whether there is anything ready to be deployed or, on the contrary, whether further developments are needed (e.g. software tools, further research).

The study started by defining *selectivity* as a general term for self-selection error resulting from decisions of individuals. These decisions could be unit-specific, for example whether to tweet or use a certain mobile provider), decisions of the owners of the technological platforms where data are captured (technology specific, for example in terms of business concept, technical infrastructure), or result from limitations of the technologies involved. The study therefore found that selectivity causes coverage, measurement and non-response error, which introduces potential bias in estimates based on big data sources.

Another concept that should be taken into account when treating selectivity in big data is **paradata**. This kind of information has been studied before in survey methodology research, and *Kreuter (2013, p. 3)* defines paradata as 'additional data that can be captured during the process of producing a survey statistics. Those data can be captured at all stages of the survey process and with very different granularities.' Moreover *Kreuter (2013, p. 4)* emphasises that paradata are not available prior to the data collection (as auxiliary variables) but are generated during collection, and that they can change the course of the data collection over time. *Kreuter (2013)* also points out that paradata are a key feature of the big data revolution for survey researchers and survey methodologists. In this study, we argue that big data are **paradata-designed,** while the data creation mechanism follows a self-selection process. While paradata vary between sources and can be data source specific, the paradata collected might be used to account for self-selection errors.

We therefore argue that methodological problems found when using big data are not new and have already been discussed in the literature on survey methodology. One main difference between traditional data sources used in statistics (including administrative registers) and big data is the *paradata* that are collected in big data on a scale never seen before in sample surveys. For example, Facebook collects data about its users even if they log out, and collects data on non-users via social plug-ins (*Acar et al., 2014*). A mobile device sends information regarding its characteristics and usage such as the operating system, hardware version, device settings, file and software names and types, battery and signal strength, and device identifiers, device locations, including specific geographic locations, such as through GPS, Bluetooth, or WiFi signals, etc. Moreover, there are several cookie-based technologies (such as canvas fingerprinting, evercookies) that make the tracking of movement on websites more persistent (*Acar et al., 2014*). The main reason for that is the growing use of automated and passive data collection (e.g. mobile devices, Internet of Things).

Moreover, we argue that, in terms of unit error, big data are similar to administrative records. Methods already available in the survey methodology can be applied to non-sampling errors such as over/under-coverage or non-response. We will also present terms used in the literature on big data and provide appropriate terms from the survey methodology (for example, profiling and imputation).

Further in the report we will introduce the concept of a **big data survey** to stress the similarities with existing methods in survey methodology, namely sample survey, register survey and internet survey. This opens up possible ways to deal with big data and points to appropriate methods from the survey methodology. However, some methods such as machine learning that before were not widely used in the survey methodology could be applied to big data surveys, in particular for imputation (see *Rey del Castillo 2012*, *Zapala 2015*).

We argue that big data in statistics is:

- another type of secondary data source (collected for other purposes),
- a new kind of internet survey (automated data collection made via the internet and Internet of Things),

- an opt-in panel (non-probability selection, longitudinal observation and data collection).

These statements will be justified further in the report based on the survey methodology literature.

1.3. Structure of the report

The report consists of seven chapters, a bibliography and an appendix. In the second chapter we try to define big data from a statistical point of view, in particular regarding statistical inference. We start in general terms, and then look into a selected group of big data sources in more detail. We address the methodological issues in measuring selectivity in big data. The second chapter starts by discussing problems in defining big data, in particular in official statistics.

We provide basic concepts and definitions taken from the survey methodology, in particular with reference to opt-in internet panels. We cover two issues, namely big data sources and unit-specific selectivity. This distinction is crucial because in the survey methodology, by definition, self-selection error is connected with the propensity to respond / take part in the survey. However, big data also suffers from coverage error resulting from the infrastructure or subset of the target population who uses the technology (e.g. internet or mobile phone) underlying the big data source. For clarification, we also provide an appendix which formalises the quantification of bias. Lastly, we briefly discuss the imputation of auxiliary and target variables which are not directly observed in most cases (in particular in big data sources). The issues presented in this part provide a general approach to identifying self-selection error and selecting appropriate methods for making estimations.

In chapter two we also explore the place of big data in data sources used for statistics. We then look at problems regarding the populations observed in big data sources and the problem of unit error and why it is important for the methods discussed in this report. We discuss the uncertainty connected with the transformation of big data objects into statistical units, the measurement error in auxiliary variables and linkage errors not considered by most of the current methods used in official statistics.

Chapter three describes the characteristics of selected big data sources. Mobile phones and social network data (Twitter, Google Trends, Wikipedia) will be presented in detail. The chapter looks in depth at issues such as describing the data source, how data are generated, the population that uses the platform from which the data sources originate and, finally, what kind of self-selection (informative or non-informative) might be observed in the data.

The fourth and fifth chapters discuss unit-level and domain-level methods that might be useful in adjusting for selectivity, and the suitability of these methods for selected data sources. We argue that access to object-level data might provide a more sophisticated approach to the problem.

The report ends with a summary of adjustments to big data sources, conclusions and recommendations for possible methods for adjusting big data sources using available data sources. We provide a general discussion of approaches for assessing and adjusting for selectivity. We recommend a general framework for adjusting for self-selection error in big data sources. This framework consists of three stages:

1. identifying variables and existing sources,
2. comparing estimates, and
3. selecting suitable methods.

The report finishes by discussing possible directions of future research in this field.

2 A statistical approach to big data

2.1. Defining big data

2.1.1. The three Vs of big data

The most well-known definition of big data, proposed by *Laney (2001)* and then discussed by *Beyer and Laney (2012)*, lists three characteristics of big data known as the 3 V's — (high) volume, (high) velocity and (high) variety. Volume refers to the amount of data in orders of magnitude of giga-, tera- or petabytes, which are difficult to analyse within the existing infrastructure. Velocity refers to how quickly these data are generated and their resolution in time. Variety refers to the multiple types of data available in big data, such as natural language textual data (e.g. social media posts), photos (e.g. posted on Instagram or Facebook), website logs, videos (e.g. camera surveillance), recordings or geocoded data.

However, this definition of big data is best suited to address the challenges brought by these new data sources in terms of IT infrastructure and information systems management. It doesn't highlight the most relevant statistical characteristics of big data.

2.1.2. Big data as non-probabilistic sample data

An early definition of big data which takes into account their statistical characteristics was proposed by *Horrigan (2013)* who regarded these data as 'non-sampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference'.

According to *Lohr (2009)*, a *sample* consists of any subset of the population and *probability sampling* is a method of sampling in which every subset of the population has a known probability of being included in the sample. On the contrary, in *non-probability sampling* the selection of units is not random (e.g. units are purposely selected or units decide themselves to take part in the survey). 'In a design-based (or randomization-theory) approach to sampling inference, the only difference between units sampled and units non-sampled is that the non-sampled ones could have been sampled during the sample selection process' *(Lohr, 2009)*. When non-probability sampling is used, it is not possible to apply probability (randomisation) theory to make proper inference about the target population of the panel or survey (*Bethlehem and Biffignandi, 2012, p. 445*).

The proposal in *Horrigan (2013)* included an aspect of big data that is fundamental for its use for statistical purposes, its **non-probabilistic character**.

Mercer, Kreuter, Keeter and Stuart (2017) discuss the theory and practice of non-probability surveys and seek to provide insight into the conditions under which non-probability surveys can be expected to provide estimates free of selection bias. *Baker et al. (2013)* presents a summary of results from work carried out by a task force set up by AAPOR (American Association for Public Opinion Research) whose aim was to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences about a larger population. A report published by the *National Academies of Sciences, Engineering, and Medicine (2017)* identifies and explores a set of key issues related to applying scientific inference to big data in biomedical applications:

a) inference about causal discoveries driven by large observational data,

b) inference about discoveries from data on large networks,

c) inference about discoveries based on integration of diverse data sets, and

d) inference when regularisation is used to simplify fitting of high-dimensional models.

Points b) and c) in particular are relevant for official statistics.

2.1.3. Designed data vs. organic data

Groves (2011a) and *Groves (2011b)* classify statistical data sources into two groups: *designed* and *organic*. *Designed* data refer to concepts and sources designed in official statistics (for instance surveys or censuses). In contrast, *organic* data denote 'data collectively assembled by society which reflects massive amounts of its behaviours and can be regarded as an ecosystem that is self-measuring in an increasingly broad scope'. The main difference between these two groups is that *designed* data are created by statisticians for statistical purposes (e.g. questions are pre-specified, the data collection process is controlled). Furthermore, the term *designed,* introduced by *Groves (2011a)*, is used where there are clear data collection times when questions are collected (e.g. on census day), while *organic* data are created in real time.

Organic data can sometimes be *designed* for business purposes. For example: Twitter allows users to write only 280 characters and offers a special character, # (the *hashtag*), to tag messages, which speeds up searches; Facebook classifies user information into groups that contain demographic, personal or activity data; Google corrects spelling errors in searches and also classifies search queries according to the categories used in Google Trends. One particular example of designed data is data captured by devices that are classified under the term Internet of Things (IoT). These devices are created by people and the data they collect were pre-specified by producers (e.g. logs). Another example is services that offer after-sales support and have special forms to input information and characteristics of products or services. Therefore, **organic** data are becoming more **designed** and structured (by technology) in order to collect as much information on products and users as possible.

Keller, Koonin and Shipp (2012) following *Groves (2011a)*, *Groves (2011b)* provided examples of organic data sources:

> 'location data (cell phone "externals", E-ZPass transponders, surveillance cameras), political preferences (voter registration records, voting in primaries, political party contributions), commercial information (credit card transactions, property sales, online searches, radio-frequency identification), health information (electronic medical records, hospital admittances, devices to monitor vital signs, pharmacy sales) and other organic data (optical, infra-red and spectral imagery, meteorological measurements, seismic and acoustic measurements, biological and chemical ionizing radiation).'

Two other concepts very similar to that of organic data are those of *exhaust data*, see e.g. *George et al. (2014)* and *Kreuter (2015)*, and *digital footprint*, see e.g. *Craig et al. (2011)* and *Fienberg et al. (2010)*. Just like organic data, both of these concepts refer to passively generated data.

2.1.4. Big data in the context of statistical data sources

Another way of understanding big data is to see it in the context of the several data sources used for statistics. *Citro (2014)* discussed big data and the internet in the context of multiple data sources. He proposed a classification of existing data sources into four groups:

- **'Surveys and censuses**, or a collection of data obtained from responses of individuals, who are queried on one or more topics as designed by the data collector (statistical agency, other government agency, and academic or commercial survey organization) according to principles of survey research with the goal of producing generalizable information for a defined population.'

- '**Administrative records** or a collection of data obtained from forms designed by an administrative body according to law, regulation, or policy for operating a program, such as paying benefits to eligible recipients or meeting payroll. Administrative records are usually ongoing and may be operated by government agencies, or non-governmental organizations.'
- '**Commercial transaction records**, or a collection of data obtained from electronic capture of purchases (e.g., groceries, real estate) initiated by a buyer but in a form determined by a seller (e.g., bar-coded product information and prices recorded by check-out scanners or records of product and price information for Web sales, such as through Amazon).'
- '**Interactions of individuals with the World Wide Web** by using commercially provided tools, such as a Web browser or social media site. This category covers a wide and ever-changing array of potential data sources for which there are no straightforward classifications. One defining characteristic is that individuals providing information, such as a Twitter post, act as autonomous agents: they are not asked to respond to a questionnaire or required to supply administrative information but, instead, are choosing to initiate an interaction.'

In comparison to traditional data sources, such as censuses or surveys, big data processing requires more effort, not only to clean the data but also to provide basic information about the people about whom the data are generated.

It must be noted that **big data can be gathered from administrative sources**, e.g. traffic sensor data, patient registers or aerial and satellite photos. However, in most cases, big data are associated with specific types of businesses and their customers — online services (e.g. Google, Facebook, Twitter), banking and finance (e.g. stock markets, global banking companies), supermarkets (e.g. Carrefour, Tesco), or telecommunications providers (e.g. Vodafone, Orange, Telefonica).

However, *Citro (2014)* did not discuss classifying data captured via sensors or cameras into the four groups specified above. Road sensor data could be classified as administrative records (as a result of certain administrative regulations) while data captured by apps in mobile devices could be classified either as a special transaction in commercial transaction records or as an interaction of individuals with the world wide web (or the internet in a broad sense). These issues are missing from *Citro's (2014)* analysis.

2.1.5. Statistical features of big data

Florescu et al. (2014) provides a comparison between traditional survey data, administrative data and big data based on 14 different features (F1 to F14). The non-probabilistic nature of big data is included as three features, 'representativeness and coverage difficult to assess (F7), bias unknown and possibly biased (F8) and the presence of both typical errors (e.g. missing data, reporting errors and outliers) although possibly less frequently occurring, and new types of errors' (F9). The organic, non-designed, nature of big data is also mentioned (F2), but it also includes the related feature of an ex-post design of statistical products (F1), which big data shares with administrative data. The IT characteristics of big data are included, considering it 'less persistent' (F10), with 'huge volume' (F11) and 'potentially much faster' (F12).

The comparison then includes other features that are also relevant for statistical production. It includes the 'higher potential for by-products' (F3), due to the higher detail provided by big data sources and the 'potentially greater comparability between countries' (F6), because many of the platforms, or at least the underlying technologies, capturing the data are shared worldwide. Finally, it includes the 'potentially inexpensive' (F13) nature of big data because data collection systems do not have to be set up on purpose (which does not mean that using big data sources does not imply other costs) and the fact that there is 'no incremental burden' (F14) for the data subjects.

Although these latter features are relevant from a statistical production point of view, they do not impact the validity of the inferences made based on big data, which is the focus of this report. An additional feature listed in *Florescu et al. (2014)* which does have an impact on the validity of statistical inference is the **unstructured** nature of big data, although it does recognise that there is a 'certain level of data structure, depending on the source of information' (F5). Unstructured data are data codified in a form that is not easy for statistical treatment or automated data processing. Typical unstructured data are natural language textual data, but these data also include pictures, video and sound recordings. This does not mean that unstructured data do not have any structure at all. Natural language has some structure, including orthographic rules and punctuation breaking text into sentences and paragraphs. The same is true for images which in order to be stored in the memory of a computer need to follow a pre-specified format. However, that format is not suitable for analytical use.

The features discussed above to some extent distinguish the data sources normally considered under the umbrella of big data (see more on this in the next section) from other more traditional sources. Importantly, *Florescu et al. (2014)* recognises that **not all data sources which would normally be considered big data share the same features**. For example, for the unstructured feature, it mentions that big data has a certain level of data structure, depending on the data source.

This highlights the difficulty of finding a comprehensive statistical definition of big data with a top-down approach, based on a detailed list of characteristics.

2.1.6. Enumeration of big data sources

Enumerating and categorising commonly recognised big data sources is another approach to understanding what big data are. The UNECE (United Nations Economic Commission for Europe) Task Team on Big Data, set up by the High Level Group for the Modernisation of Official Statistics, developed a classification of big data types which helps clarify what we have in mind when we talk about big data (*UNECE, 2013*).The classification includes three main types of big data sources. Firstly, it includes human generated data which is stored in digitised form. Nowadays this consists mostly of what people share on social networks, but it also includes other forms of human expression which were traditionally stored in non-digital forms, such as books and photographs printed on paper. A significant part of this data is unstructured but, some of it is structured, in particular the data that people register in a structured form on purpose, for example in the case of volunteered geographical information (e.g. open street map).

Secondly, the classification includes data that is generated as a by-product of IT systems which support particular business processes (process-mediated data). These data are still generated by people, when they interact with the IT systems, but people don't generate the data themselves. For this reason, this type of data is usually structured.

Thirdly, the classification includes machine-generated data, usually captured by some type of sensor. The sensors can be fixed sensors, which register either events or surroundings state at regular (and usually very frequent) points in time. They can also be mobile sensors, which register geolocation in addition to what fixed sensors register. Machine-generated data includes non-sensor data, in particular data captured and stored by IT systems about their own functioning (normally in log files).

An almost ubiquitous trait of all data sources in this classification is the level of detail of the data. With the exception of human-volunteered data, all data sources listed provide very high resolution in various dimensions (time, space, events, etc.).

This classification also facilitates the discussion about whether one of the features of big data is that it is designed. While for human-sourced data, it is clearly *not* designed (apart from volunteered data, in some cases) and process-mediated big data *is* designed (though not for statistical purposes), machine-generated/captured big data are often designed for statistical purposes (e.g. satellite images for remote sensing) or at least on purpose for data processing following requirements very similar to those of statistical use.[10]

Classification of types of big data (*UNECE, 2013*)

1. *Social Networks (human-sourced information)*
    - *1100. Social Networks: Facebook, Twitter, Tumblr etc.*
    - *1200. Blogs and comments*
    - *1300. Personal documents*
    - *1400. Pictures: Instagram, Flickr, Picasa etc.*
    - *1500. Videos: Youtube etc.*
    - *1600. Internet searches*
    - *1700. Mobile data content: text messages*
    - *1800. User-generated maps*
    - *1900. Email*
2. *Traditional Business systems (process-mediated data)*
    - *21. Data produced by Public Agencies*
        - *2110. Medical records*
    - *22. Data produced by businesses*
        - *2210. Commercial transactions*
        - *2220. Banking/stock records*
        - *2230. E-commerce*
        - *2240. Credit cards*
3. *Internet of Things (machine-generated data)*
    - *31. Data from sensors*
        - *311. Fixed sensors*
            - *3111. Home automation*
            - *3112. Weather/pollution sensors*
            - *3113. Traffic sensors/webcam*
            - *3114. Scientific sensors*
            - *3115. Security/surveillance videos/images*
        - *312. Mobile sensors (tracking)*
            - *3121. Mobile phone location*
            - *3122. Cars*
            - *3123. Satellite images*
    - *32. Data from computer systems*
        - *3210. Logs*
        - *3220. Web logs*

Source: UNECE (2013)

[10] Concerning internet searches, it should be noted that the data are composed of two elements, the search terms and the number of searches. The first element is human-sourced information, but the second one is derived from machine-generated data, in particular web logs.

2.1.7. A statistical definition of big data

All of the classifications above attempt to identify or define the relevant characteristics of big data. The level of detail is probably the most discriminant characteristic of big data. All other data types lack the level of detail seen in big data. This is certainly because big data are generated or captured in an automated way, via IT systems or sensors, which results in massive datasets of very large volume. The popularity of IT systems, sensors and digitisation of our lives means not only that there is an extremely large overall volume of data available in our society, but also that there is a big variety of forms of data. Most of these data are organic and unstructured, not designed for statistical purposes, but this is not always the case. Sensors are sometimes set up specifically to collect data for statistical purposes and even then the data have all the characteristics that distinguish big data from other types of data traditional in statistics.

A possible big data definition which includes all these characteristics is that **big data are highly detailed exhaust data automatically captured by sensors or generated through IT systems**. We define a **big data source** as a specific source of big data (e.g. Twitter or a specific mobile communications provider).

From the statistical inference point of view, what really matters is the way these data are generated, in particular their non-probabilistic character, and their organic and unstructured nature which means that often they do not consist of direct measurements of our variables of interest but only proxies.

2.2. Statistical characteristics of big data

2.2.1. Mediation by electronic platforms

Big data are so detailed because it is automatically captured by electronic platforms. Most big data sources have underlying electronic platforms which were not set up for statistical purposes. These platforms include telecommunications networks (e.g. mobile phone networks, the internet), electronic devices (e.g. those used to communicate or sensors) and the software applications providing services to users (e.g. social media, booking systems). Platforms may involve a single technology or a combination of several technologies.

The electronic platforms mediate statisticians' observation of reality, just like questionnaires do in traditional surveys. And just like questionnaires influence what is observed, the electronic platforms that capture big data, together with individuals' behaviour while interacting with these platforms, also influence what is observed. The characteristics of the specific electronic platforms influence the data captured at several levels: they influence which objects, individuals or events are observed and what can be measured, as well as frequency, timing, precision and reliability [11].

Let's take the internet as an example. The internet is a global and public system of interconnected computer networks that use the standardised internet protocol suite (TCP/IP). It is a network of networks consisting of millions of local networks and individual computers from all over the world. Email, world wide web (www/the web), FTP and other services are available to use via the internet computer network. The world wide web is an information system on the internet; it allows documents to be connected to other documents by hypertext links. To access the web, users rely mostly on web browsers, which are programs used to navigate the web by connecting to a web server, allowing the user to locate, access, and display hypertext documents or mobile applications that are computer programs designed to run on mobile devices, such as smartphones and tablets. Recently *the Internet*

[11] I.e. the probability of failure that a measurement is made at all (*IEEE Standards Association, 2010*).

of Things (IoT) has also gained recognition as a potential source of information. The Internet of Things is defined as the interconnection via the internet of computing devices embedded in everyday objects, enabling them to send and receive data (independently of the user).

The fact that electronic platforms mediate the data-capture process has three consequences for big data. Firstly, access to and use of the electronic platform is a pre-condition for capturing the data. Individuals or objects that have no access to the platform or that do not use it will not be observed. This means that in cases where the data capturing platform was not set on purpose for statistical use, part of the population is **not covered,** while the part of the population of interest observed will most probably be created by non-random selection. Consequently, the sample suffers from problems of **selectivity**.

Secondly, usually (though not always) the data captured does not correspond to the variables of interest for statistics. For example, the data captured in social networks in the form of natural language textual data are not suitable to be used directly in statistical analysis. They need to be transformed into some variable of interest, e.g. the sentiment expressed in the communication, which in this case is **not observed** directly.

Thirdly, very often the most basic unit or object (e.g. a tweet) to which the data captured relates is not the statistical unit of primary interest (e.g. a person) and sometimes only indirectly relates to the statistical unit of analysis most relevant for statistics. For example, in mobile network data the most basic unit is a communication event, where we know the location of the antenna used and can identify the mobile device used. A statistical unit of interest for statistics would probably be the individual using the device. The object and the statistical unit are certainly related, as for each communication event there is a mobile device and each mobile device is used by one or more individuals. However, there are **multiple analytical statistical units** which can be reconstructed and **the link with the statistical unit of interest involves uncertainty and errors**.

2.2.2. Imperfect coverage

The study of the coverage of statistical frames used in sample-based surveys is not new, and two cases are of particular relevance for big data sources: the internet (for internet-based surveys) and mobile phone numbers (e.g. for surveys based on random digit dialling).

The mobile phone penetration rate could be used as a proxy of the coverage of the target population of individuals. According to Eurostat,([12]) the mobile phone penetration rate varies across the EU, from 109 % to 231 %. This reflects potential over-coverage and duplication within each country, although we cannot assume that every single statistical unit is a mobile phone user.

From a statistical point of view, we can treat mobile phone numbers as a potential imperfect register or frame. However, as in the case of telephone surveys (*Lepkowski et al., 2007*), this is an indirect way to observe units from the target population. One should take into account six types of linkage in mobile phones (*Lepkowski et al., 2007, ch. 2.1.2*):

- one-to-one — linkage does not contribute to survey error, assuming no linkage error;
- one-to-none — there are units on the frame (i.e. one) that are not in the target population (i.e. none), frames must be screened to identify and remove certain numbers;
- none-to-one — linkage occurs in the telephone sampling when the target population is the entire or a subset of the target population, since statistical units without telephones are excluded;
- many-to-one — occurs when a single member in the target population can be selected via multiple frame entries (i.e. a person has several mobile phone numbers);
- one-to-many — there is one frame entry for multiple members of the population (people sharing a mobile phone number);

([12]) Source: Eurostat (isoc_tc_mcsupe).

- many-to-many — linkages occur when multiple people (the statistical unit) in a household are linked to multiple telephone numbers.

To provide more insight into the problem, basic definitions of frames are provided below.

According to the definition specified in *Lohr (2009)*, a **sampling frame** is a list, map, or other specification of sampling units in the population from which a sample may be selected.

According to *Wallgren and Wallgren (2014)*, a **register** aims to be a complete list of the objects in a specific group of objects or population. However, data on some objects can be missing due to quality deficiencies. The following data about an object's identity should be available so that the register can be updated and expanded with new variable values for each object: (1) complete listing and (2) known identities; these are thus the characteristics of a register. Catalogue, directory, list, register and registry are different terms for the same concept.

Recently, *Zhang (2015)* noted that using the term **list** is more natural than **register** in the context of register-based statistics, as well as in a number of situations outside official statistics, such as the sizing of wildlife, hard-to-reach or clandestine populations. Hence, Zhang proposed to use these two terms interchangeably.

We will use the generic term **frame** with regard to big data sources as the term **list** is a frame of a specific type (e.g. a list made on paper or several paper note cards).

Often, the quality of big data frames is unknown. Sometimes, frame holders (operators) are also unaware of frame quality because the frames were set up for other purposes. However, an increasing number of operators monetise the data captured by their electronic platforms and therefore are concerned with the quality of their frame.

Survey statisticians have dealt with frame problems before (e.g. in telephone surveys, cf. *Lepkowski et al. (2007)*). The problems related to lists in statistical literature are as follows:

- over-coverage — units that are not from the target population are included;
- under-coverage — units that are in the population are not observed;
- duplicates — units that are listed multiple times within a frame;
- multiple frames — there are several lists that partially measure the target population;
- lack of frame for the target population — there is no list that fully covers the target population.

Big data sources sometimes suffer from the problem of there not being an appropriate frame for the target population. For instance, if we would like to measure only the internet or mobile device population, we should have a list of all units for that population. Due to practical reasons, there is no such list. We would have to use other available lists, for instance phone numbers or social media portals. This problem was mentioned before in *Deville and Lavallée (2006)* and *Lavallée (2009)*.

We can now move on to the problem of multiple frames.

From the theoretical point of view, we can treat these operators (and their data) as an imperfect frame that may be directly or indirectly connected to the target population. Here, we start by denoting the population with the symbol $\mathbf{\Omega}$ and the population we are referring to using subscript. For instance, $\mathbf{\Omega}_{TP}$ means target population, $\mathbf{\Omega}_{GT}$ the population of Google Trends users. We denote frames using capital letters, and use subscripts to define data holders and population, for instance $A_{MT,1}$ denotes the list of mobile phones according to mobile provider 1 and $B_{BTS,2}$ denotes the list of base transceiver stations (BTS) under mobile provider 2. This notation will be used when the relation between these populations is further discussed in the context of selected big data sources.

2.2.3. Missingness, selectivity and bias

In the research literature, the term 'bias' is often used with different meanings, depending on the context. In design-based statistical inference for well-defined finite target populations, bias refers to a theoretical property of an estimator of a finite population parameter. An estimator is called 'design-unbiased' if its expectation over the entire universe of possible probability samples equals the true parameter. Many popular estimators of population (or sub-population) totals are ‘design-unbiased’ (the Horvitz-Thompson estimator is a good example).

As a reminder, the theoretically convenient property of design unbiasedness of design-based estimators can be jeopardised in the presence of non-probability sampling and informative unit non-response (if the response mechanism correlates with the variable of interest, cf. *Brick, 2013*) that will cause unknown selection bias. As a consequence, the originally design-unbiased estimator produces biased estimation for the target parameter. This also occurs for model-based estimators, whose design bias depends on the correctness of the model. In this case, we need to address both the design bias due to the estimator and the selection bias due to informative unit non-response. Thus, the selection bias is not a property of the original probability sampling procedure or the estimator itself, but it is still present in the estimates via the response/selection mechanism of the procedure that generates the observed data.

In big data sources, which are based on non-probability samples by definition, the selection bias becomes dominant. Therefore, we restrict the discussion to biases occurring due to the self-selection mechanism that is caused by individuals. We recognise that both design-based and model-based estimates from probability sampling and estimates from big data can suffer from an unknown selection bias. Therefore, similar statistical methods are applicable for both data source types. Further discussion on bias quantification is presented in Appendix A.3.

We define selectivity as a general term for self-selection error which results from individuals (unit-specific, e.g. whether to tweet or use a certain mobile phone network provider) and data holder decisions (data holder specific, e.g. in terms of business concept, technical infrastructure). This includes coverage, measurement or non-response (missingness) error, which further introduce potential bias in estimates based on big data sources only. Moreover, big data are **paradata-designed** while the data creation mechanism is following a self-selection process. Here, we summarise earlier findings and its implications on selectivity adjustment methods:

- **big data source specific error** — this is mainly related to coverage error that is not directly related to the decision of individuals but to either contingencies of the technology itself or decisions of the data source holder (or holder of the technology underlying the big data source). For instance, the infrastructure supporting mobile phone use might limit the possibility of identifying trips (e.g. to rural regions), not all units might have access to the internet, and social media restricts the population to units with internet access only (broadband/mobile).
- **unit-specific selectivity (self-selection)** — this is the result of any decision of individuals (or other entities) that affects its presence in the sample. In the case of big data, there are three types of decision to take into account: a) the decision to use the technological platform from which big data are being captured, e.g. the internet or a device connected to the mobile phone network; b) the decision to select a particular internet platform (e.g. one particular social media portal) or a particular mobile operator; and c) the decision to provide paradata that might be used to impute (pseudo-)responses or auxiliary variables. In the case of questionnaire surveys, we have this type of selectivity in the presence of non-response. Here we consider that selectivity might be related to the target variable (MNAR) or to auxiliary variables (MAR).

Based on the above, we distinguish two possible levels at which self-selection error can be approached — the unit level, which (formally) requires that each record refers to one unit, and the domain level, which requires unit-level data aggregated at a given domain level. However, in practice we should suspect that big data sources contain duplicated records referring to the same statistical unit or are affected by over-coverage.

2.2.4. Variables not observed directly

Very often, target and auxiliary variables are not directly measured in big data sources, in which case we can say that unobserved or latent variables are present.

The treatment of unobserved variables is not new. In some cases, variables are not directly observed because it might be expensive to measure or store them, or they may be removed on purpose to preserve privacy. In other cases, variables are unobservable because the variable is theoretical in nature (this is the case for intelligence, a hypothetical construct). More generally, we can also consider *true variables* that are measured with errors as unobserved variables.

Imputation provides one approach to obtaining estimates of unobserved variables. It is a procedure where missing values for one or more variables of interest are 'filled in' with substitutes. This procedure is very commonly applied to observed variables which are missing for some statistical units. The substitutes can be constructed according to a rule, or they can be observed values but for elements other than the statistical units for whom data are missing. In this sense, imputed values are artificial and can therefore contain errors, *Särndal and Lundström (2005)*. There are several techniques that are used to impute missing values, *Lohr (2009 ch. 8.6)*:

- deductive imputation,
- cell mean imputation,
- hot-deck imputation,
- regression imputation,
- cold-deck imputation,
- multiple imputation,
- machine-learning based imputation.

In addition to the methods presented above, literature in this area discusses **mass imputation**, particularly with reference to registers (*Houbiers, 2004*). Mass imputation is the procedure of imputing missing values for all units of the population, based on data from sample surveys and register sources, and possibly modelling methods as well. As a result, all units in the population have values for the target variables. The use of imputation with respect to big data can resemble mass imputation, where imputation of variables of interest is made for all units of the target population, based on values observed in big data sources. However, if the response mechanism underlying big data is missing not at random (MNAR), then imputation can lead to biased estimates. *Daalmans (2017)* discusses mass imputation for census estimation in the Dutch population census and also addresses the selectivity problem. For more information on imputation see *Rao (1996)*, *Fay (1996)*, *Shao (2003)*, *Särndal and Lundström (2005)*, *Rubin (1987)*, and *Rubin (1996)*.

Machine learning is commonly used with big data, where target and/or auxiliary variables are predicted from the captured data. It consists of a variety of methods, from more traditional linear statistical models to highly non-linear algorithmic methods, such as decision trees and artificial neural networks (*Hastie, Tibshirani and Friedman, 2008*). Although weaker in terms of interpretability and potentially in their capacity for out-of-sample generalisation ([13]), these algorithmic methods are very good at reproducing observed values of target or auxiliary variables in a training sample. Raw data in the form of natural language text, images, sound or sensor data that is not suitable to be used directly in statistical models and in most of these algorithms is pre-processed and transformed into **features** which can then be used. In the case of artificial neural networks, the raw data can even be used directly, in which case the features are derived automatically, or implicitly, by the algorithm itself.

Just like in the case of the more traditional imputation methods, the prediction from machine-learning algorithms is based on models and/or algorithms which introduce measurement errors, in terms of variance and possibly bias where the model/algorithm used is not valid.

This is particularly relevant when correcting selectivity in big data sources. The methods will often depend on auxiliary variables which, if there are measurement errors, will result in bias in the estimation of parameters of the models used and therefore potentially provoke errors in the

[13] The initial lack of generalisation capacity can usually be mitigated using existing methods, such as regularisation.

selectivity correction.

2.2.5. Unit identification problem

Some definitions state that the big data at hand equals the 'entire population'. However, from a statistical point of view, there is a variety of different possible populations and the target population of a statistical analysis should always be explicitly specified. Only then can we say if we have the entire population or not.

Wallgren and Wallgren (2014) provided the following definition of the process of defining a population: 'The population definition should clearly show which objects are included in that population. The object type should be clearly specified. In addition, a time reference and geographic delimitation should always be included. The geographic delimitation should also specify the relation that exists between the objects or statistical units and the geographical area.'

Moreover, *Wallgren and Wallgren (2014)* refers to four concepts related to populations:

- '**Population of interest** refers to the population in the theoretical question at hand.'
- '**Target population** refers to the operationalised population, the theoretical population of interest which has been translated into a concrete and examinable population, i.e. the population that is the **target** of the survey.'
- '**Frame population** refers to the object set that the **frame** actually gives rise to.'
- '**Register population** refers to the object set in the register that has been created for the survey in question, i.e. the population that is **actually** being surveyed.'

Table 1 lists several examples of target populations commonly used in official statistics.

Table 1: Examples of common target populations in official statistics

Statistical unit	Population delimitation
Individual (person)	Resident population in a country at a reference date
Individual (person)	Tourists who visited a country in previous 12 months
Household	Private households residing in a country

In big data sources, we observe several register populations that relate to these target populations. Tables 2 to 4 show examples of register populations and subsets of common target populations present in three big data sources: mobile network data, Twitter data and Google Trends data. Chapter 3 explore these in more detail.

Table 2: Examples of populations observed in mobile network operator data

Statistical unit / object type	Population delimitation
Register populations	
Base transceiver station (BTS)	BTS located in a country as of 1 January
Call detail record (CDR; communication event)	CDRs of mobile network operators operating in a country in the last month
SIM card (or mobile phone number)	SIM cards with at least one event in the last 12 months
Target populations	
Individual (person)	Mobile device users resident in a country

Table 3: Examples of populations observed in Twitter data (registration is required)

Statistical unit / object type	Population delimitation
Register populations	
Twitter post	Tweets posted in Dutch in the last 12 months
Twitter account	Accounts that published at least one tweet in the last 12 months
Target populations	
Individual (person)	Social media users with at least one post in the last 12 months

Table 4: Examples of populations observed in Google Trends data (registration is not required)

Statistical unit / object type	Population delimitation
Register populations	
Search query	Search queries made in the last 3 months
IP address	Unique IP addresses of search queries made in last 3 months
Web browser cookie[14]	Cookies identified in any website access in the last 3 months
Target populations	
Individual (person)	Google search users

Linking observed units to the units of the target population involves uncertainty and the introduction of unit identification errors, a.k.a. unit error (*Zhang, 2011*). Unit error occurs when base units from the register population are allocated to the wrong target units during the re-construction of the statistical units of the target population.

In order to draw valid statistical results from big data, it is necessary to exactly identify the observed register populations and specify the target population of interest. Identifying the relationship between these two populations might not be an easy task and can vary between sources (in Chapter 3 of this report we do this for a few selected big data sources). Moreover, the target population is rather stable over time while big data observed populations can vary significantly over time. For instance, the population of cookies is substantially bigger than the target population, as it is generated by each device and each browser. However, there are ways to track users across devices which make it possible to identify internet users (more details on cookies will be given in the section on Google Trends in Chapter 3).

Unit error has implications for the accuracy of statistics and also for the application of methods to correct selectivity in big data sources. Estimates based on big data objects or objects transformed into statistical units introduce bias in point estimates at population and domain levels. Due to duplicated records or over-coverage, the distribution of the target variable might be distorted, in particular leptokurtic ([15]). Moreover, unit errors will carry over to all statistics, such as income or population demographic statistics, which may or may not have severe consequences (*Zhang, 2011*). A unit-error theory proposed by *Zhang (2011)* should allow us to propagate the uncertainty to such induced statistics. *Zhang (2011)* proposed using an allocation matrix denoted by A^* to identify connections between base units and statistical units. In order to estimate the characteristics based on the allocation matrix, one needs an independent audit sample to identify A^*.

[14] A cookie is a way to identify a unique user.

[15] A leptokurtic distribution is a distribution with high kurtosis, by convention higher than 3. This situation occur when we have multiple objects that refer to the same unit (e.g. ads with the same characteristics refer to the same real estate).

Measurement errors in big data variables can be caused by unit identification errors and models that do not take this into account will provide **biased estimates of model parameters** *(Johnston, 1991)*. As a result, weighting procedures used to correct selectivity, such as propensity score adjustment or the model-based approach, might introduce additional bias to estimated characteristics such as means, proportions and totals.

Linkage errors can occur when probability-based methods are used to link records from two distinct data sets corresponding to the same target population. Relatively small linkage errors could lead to a substantial bias in estimating the relationship between variables from distinct datasets. See *Lahiri and Larsen (2005)*, *Chambers (2009)* and *Samart (2011)* for a comprehensive review of problems related to linkage errors.

In comparison to probability-based linkage, **statistical matching** methods assume that two records are matched based on a set of common variables and these records do not necessarily refer to the same unit. Such matching implies additional uncertainty of estimates (*Rivest, 2007*).

Most of the methods currently used for self-selection adjustment in official statistics do not take into account the problems above, simply because these problems did not exist in censuses or surveys (*Zhang, 2011*). Hence, one should identify and carefully study these errors when working with big data sources, and do so before using the methods discussed below.

2.3. Addressing selectivity in big data

2.3.1. Big data as an opt-in panel survey

According to Eurostat's definitions (*Eurostat, 2016*), **primary data** are data observed or collected directly from first-hand experience. Published data and data collected in the past or by other parties are called **secondary data**. In light of this definition, most big data sources should be treated as a secondary data source.

Bethlehem and Biffignandi (2012) provided the following definitions of surveys that should be taken into account when discussing the role of big data in statistics:

- **Internet survey** — a general term for various forms of data collection via the internet. It could be a web survey or an email survey.
- **Web survey** — a form of data collection via the internet, in which respondents complete a questionnaire on the world wide web. The questionnaire is accessible via a link on a web page.
- **Self-selection survey** — a survey for which the sample has been recruited by means of self-selection. Users can decide whether or not to participate.
- **Self-selection panel** (a.k.a. volunteer panel or opt-in panel) — a (web) panel for which people select themselves in response to a banner or pop-up window on the internet, or an advertisement in other media (radio, TV, and newspapers).

With this in mind, we argue that **big data sources** are:

- another type of secondary data source,
- a type of internet survey,
- a self-selection/opt-in panel.

Firstly, we treat big data sources as **secondary data sources,** mainly because they are not created by statisticians, or at least not for statistical purposes. The data from big data sources are delivered or obtained by statisticians in order to derive statistical information. This implies potential coverage issues and measurement errors. For instance, mobile providers' infrastructure might not reflect administrative borders or might not make it possible to identify the exact position in rural areas.

Secondly, following *Wallgren and Wallgren (2014, p. 10)*, the term **survey** is used generically to cover any activity that collects or acquires data for statistical purposes, including:

1. a census, which attempts to collect data from all members of a population,
2. a sample survey, in which data are collected from a (usually random) sample of population members,
3. a collection of data from administrative records, in which data are derived from records originally kept for non-statistical purposes,
4. a derived statistical activity, in which data are estimated, modelled, or otherwise derived from existing statistical data sources.

As the term 'survey' is used broadly and due to similarities with administrative sources (in particular that they are both secondary sources) and to the fact that we want to use these data for statistical purposes, we should use the term '**big data survey'** (cf. register survey or register-based survey).([16])

In addition, we propose to approach big data sources in the context of **internet surveys**. As *Bethlehem and Biffignandi (2012)* stated, this kind of survey is a general term for various forms of data collection via the internet. As argued in Section 2.2.1, big data are collected via electronic platforms, of which the internet is a representative example. Hence, when dealing with big data surveys, one can refer to the issues found in internet surveys and the methods developed to address them.

Finally, we argue that big data should be considered as an **opt-in panel**. *Lynn (2009)* defined a longitudinal survey, or panel, as one that collects data from the same sample elements on multiple occasions over time, and sample attrition (also referred to as panel attrition) as referring to the continued loss of respondents from the sample due to non-response at each wave of a longitudinal survey. *Smith, Lynn and Elliot (2016)* provided a classification of sample panel surveys.

Brüggen, van den Brakel and Krosnick (2016) discuss the role of auxiliary information in selection bias adjustment for opt-in panel data. The accuracy of results obtained from several opt-in online panels in the Netherlands was compared with the results of an independent probability sample. The available auxiliary information consisted of demographic and regional data from administrative data sources. In this study, the auxiliary data that was incorporated into weighting procedures appeared to be too weak to successfully explain the mechanism that generated the final panel data.

Buelens, Burger and van den Brakel (2015) present more positive results for selection bias adjustment, based on a simulation study using real data. They used data-generating mechanisms mimicking those often assumed in big data situations. The authors demonstrated the importance of strong auxiliary data that is capable of explaining the data-generating mechanism for a successful adjustment for the selection bias.

Bethlehem and Biffignandi (2012) provides a definition of a web (internet) panel. It is a survey based on a list of objects (e.g. companies, households or individuals) that are interviewed at different points in time (called panel waves). A web panel is expected to include a large number of objects, including certain known 'demographic' characteristics (in a wide sense). *Bethlehem and Biffignandi (2012)* provides the definition of an opt-in panel — panels based on self-selection recruitment. These panels are composed of respondents who voluntarily sign up (opt-in) for the panel. Here are some examples of how one can become a panel member:

- Participants go to the specific panel recruitment portal themselves. They could, for example, find out about the web panel via an advertisement in the media.

([16]) We do not recommend using the term 'big data-based survey' mainly because of the potential to confuse the term 'data-based' with the term 'database'.

- Participants are redirected through banners. Pop-ups are also often used for recruiting panel members.
- Some websites are designed to ‘sign up’ participants to several opt-in panels immediately upon entering the website.
- Participants are asked to register in a panel at the end of another survey — possibly offline. In this case, the panel is populated with a subset of the respondents of the initial survey.

Taking the above into account, we now summarise why we consider **big data as an opt-in panel**. The first argument is that with big data we indirectly observe statistical units over time, for instance by following mobile phone numbers or Twitter accounts assigned to one person. In big data we also observe attrition (e.g. churn to other mobile providers or decrease of interest in Twitter) and births (e.g. new customers sign up with the provider, new social media accounts are created). Because of the data-capture mechanism, big data should be treated as an opt-in panel. The self-selection mechanism means that certain people chose a given mobile provider or created an account on Twitter.

To sum up, big data share similar characteristics with

- secondary sources — turning non-statistical data sources into statistical information,
- internet surveys — because the data collection is mediated by an electronic platform (e.g. the internet),
- opt-in surveys — because the participation/selection mechanism is non-probabilistic,
- panel surveys — because we observe statistical units over time (with attrition).

This concept of treating big data as an opt-in internet panel is new. To the authors’ knowledge, the first references to panel surveys were made by *Diaz et al. (2016)* who argue that ‘online and social media activity function like an opt-in panel where different users engage to different degrees during different times’. Furthermore, *Diaz et al. (2016)* discussed the concept of ex-post panel creations from social media, which certainly has imperfections.

This discussion on the definition is important because considering these data sources opens an avenue to using methods that already exist in survey methodology. The problems found in big data sources were studied before in the context of sample surveys (coverage, non-response, self-selection, measurement error).

2.3.2. Big data survey design

Wallgren and Wallgren (2014) present register survey design as providing answers to the following questions:

- ‘How should the research objectives be defined? ’
- ‘What sources should be used? The possible sources should be analysed. Are they usable for the purposes, if used alone or in combinations with other sources? ’
- ‘Are special methods needed for creating the register population? Calendar year registers and longitudinal register (...) may require more advanced methods to be developed. ’
- ‘Are special methods needed for creating derived variables? Available variables can be used to classify statistical units. [For example] labour market activities can be used to classify both persons and enterprises into categories with different activity patterns. In such cases, advanced classification methods must be developed. ’

In the case of big data, we could also use the term ‘**big data survey design’** to disentangle its parts. Using this term allows us to separate elements of the process of using big data for official statistics. Moreover, at each stage of the big data survey design we could identify potential sources of selectivity. We argue that in the case of a big data survey design the following steps should be considered:

- big data sample — how given big data are created (business concepts underlying the big data source and self-selection mechanism),
- big data paradata — how given paradata are created at a given big data source (what data are collected);
- big data collection — how data are collected (obtained by or transmitted to the official statisticians),
- big data cleaning — how one should deal with the data cleaning process,
- big data auxiliary data — how to obtain auxiliary data for the estimation phase,
- big data estimation — how one should approach the estimation process.

2.3.3. The use of external sources to address selectivity

The selection of suitable methods for selectivity adjustment requires answering the following question: **do we have (independent) data sources that can be used to verify the quality of the big data source?** Such auxiliary data sources could be censuses, probabilistic sample surveys or administrative records (and derived statistical registers). One could also consider other big data sources. However, due to the non-statistical character of these data, they themselves may suffer from possibly unknown frame and selectivity issues. The importance of independent data sources that can be used to correct unit-error and/or estimate under-coverage or other errors has long been understood in sample surveys *(Zhang, 2011)*, e.g. in census under-coverage *(Wolter 1986)* or population size estimates (e.g. capture-recapture methods; *Fienberg 1972*, *Cormack 1989*, *Zhang, 2015*).

We see the following possible uses of external data sources, to:

- **Compare estimates based on big data and other available sources** — the main idea is to verify whether the big data source, without any correction such as weighting or modelling, can provide estimates close to the ones given by existing sources.
- **Use as auxiliary variables** —usage is limited to searching for possible auxiliary variables and totals for reweighting procedures.
- **Compare models built on big data and other available sources** — the main idea is to verify whether the model built on a big data source is close to the model built on a probability-based sample or other unbiased sources. The model should be the same for all data sources and the comparison of parameter estimates will provide information on whether using all available information removes self-selection error. This step will also provide information on whether the model built on the big data sample holds for the population and can be applied to predict target variables for population datasets (as for small area estimation). Here we make the assumption that the model built on the existing (statistical) data sources holds for the target population.
- **Link big data to available sources** — linkage might have several purposes, namely the identification of self-selection error (MAR, MNAR)([17]) and the correction of bias. For instance, *Lee (2006)* showed that an independent probability sample could reduce bias of estimates while using propensity weighting adjustments (see Section 4.1.4). If we use probabilistic or distance based methods to link records between the data sources, it would be possible to find out what types of unit are not observed in the big data source (have a high level of distance).

([17]) See annex A.2 for a definition of the several types of the missing data mechanism.

2.3.4. Measuring the self-selection process

Unit non-response has been recognised as a potential problem since the early days of probability sampling. Most surveys, especially in official statistics, employ large sample sizes and design-based estimators and reweighting methods with auxiliary information to adjust for non-response. While surveys employ methods to minimise non-response and its effects on estimates, in every survey there are some sampled units that do not respond *(Brick, 2013)*.

According to *Brick (2013)*, model assumptions and adjustments are made in an attempt to compensate for missing data. However, the survey estimates may be biased even after the model-based adjustments because the mechanisms that cause unit non-response are almost never adequately reflected in the model assumptions. Non-response also causes a loss in the precision of the survey estimates, primarily due to the reduced sample size and secondarily due to the result of the increased variation in survey weights. However, bias is the dominant component of the non-response-related error in the estimates, and non-response bias does not generally decrease as the sample size increases. Thus, bias is often the largest component of mean square errors of the estimates, even for subdomains where the sample size is large *(Brick, 2013, p. 330)*.

The focus of recent research on non-response can be summarised as follows *(Brick, 2013)*:

- studying causes of non-response mechanism — psychological and sociological factors,
- studying data collection processes to reduce non-response,
- studying statistical adjustments of the survey weights to adjust for non-response and keep the design-based mode of inference.

For big data sources, due to the non-probability character of the data creation process and the similarities to internet opt-in panels, the crucial problem is identifying self-selection errors. This concept is not new and was studied in research literature before, under the terms 'self-selection process', 'response mechanism' or 'missing data mechanism'.

Firstly, we focus on the problem of identifying missingness in the case of big data. In surveys, by non-response we understand either unit non-response (e.g. a person refuses to take part in the survey) or item non-response (e.g. a person takes part in the survey but refuses to provide an answer to a given question).

In the case of big data, it is common that the target variable (e.g. sentiment) is not directly measured. For such cases, *Little and Rubin (2014)* used the terms **factor analysis missing data pattern** or **latent-variable patterns with variables that are never observed**. It can be useful to view certain problems involving unobserved 'latent' variables as missing-data problems, where the latent variables are completely missing, and where ideas are applied from the missing-data theory to estimate the parameters.

Here, we assume that we have variables that manifest the latent variable. Moreover, measuring **latent** variables via **manifest** variables naturally introduces measurement errors and additional uncertainty.

To simplify, we assume that **response** in the context of big data can be understood as **either direct observation of the target variable or as a latent variable that can be imputed using a set of available paradata** (e.g. tweets, BTS locations). One should take into account that in the case of an unobserved variable we could introduce measurement errors caused by the methodology used to predict its value.

Another issue that should be considered for measuring self-selection is **attrition,** which was defined previously as continued loss of respondents from the sample due to non-response at each wave of a longitudinal survey. *Callegaro et al. (2014, ch. 1)* identified four kinds of attrition:

- 'Voluntary attrition — voluntary attrition is the proactive action of panel members to contact the company and ask to be removed from the panel. (…)'

- 'Passive attrition — more frequently, panel members simply stop answering surveys, or they change their email addresses without notifying the company. These members are also referred as 'sleepers', as they are not active, but some of them can be 'awakened' with specific initiatives. (…)'
- 'Mortality attrition — this occurs when a panel member dies or is no longer physically or mentally capable of answering surveys. (…)'
- 'Panel-induced attrition — lastly, the panel company can decide to 'retire' or force panel members out of the panel. (…)'

A similar concept of attrition, closer to the real situation in many big data sources, is that of **customer attrition,** which refers to the loss of clients or customers. In this case, two types of attrition are defined: **voluntary** when customers switch to another company or service provider and **involuntary** due to circumstances such as a customer's death, relocation etc.

The identification of, and adjustment for, the self-selection mechanism requires strong auxiliary variables. These variables should be:

- associated with the target variable;
- associated with the non-response mechanism.

Examples of such variables are:

- demographic characteristics that are the source of many post-stratification adjustments;
- common attitudinal questionnaire items known as webographic characteristics;
- observational and process data obtained during data collection, known as paradata.

For big data, we could consider the following variables, identified as paradata:

- use of the internet (e.g. webpages visited, 'likes');
- use of mobile phones (e.g. use of mobile services);
- type of device (e.g. smartphone, tablet).

However, one should take into account that using paradata to measure self-selection might not be an easy task *(Kreuter et al., 2010)*.

Tam and Clarke (2015) discuss important inferential issues when using big data sources in official statistics. They present a Bayesian framework to assess the conditions under which valid statistical inference can be drawn from big data and provide a Bayesian method for using big data sources to produce official statistics.

3 Analysis of specific big data sources

This chapter briefly discusses the characteristics of a few selected big data sources. Big data sources differ significantly, which means that the appropriate methods to treat selectivity may also be different. It covers mobile network positioning data, one case of social network data (Twitter) and two cases of web activity data (Google Trends and Wikipedia page views). These sources were chosen to cover a relatively broad spectrum of different characteristics found in big data.

3.1. Mobile network data

3.1.1. Description of the data source

Mobile network data means the data collected by mobile network operators (MNOs) as a by-product of the mobile telecommunications network.

This consists of two types of data. The first one is the communications network, i.e. who calls who, taken from calls and text messaging. The second one is positioning data. Whenever a device and the network communicate, the operator detects the approximate location of the device by identifying the antenna used for communication. Devices connect to antennas near to them, normally (although not always) the closest one. This allows individuals communicating via a certain antenna to be allocated to a spatial unit consisting of the points to which the antenna is the closest one. This section focuses on positioning data, but many of the conclusions also apply to other types of data taken from the mobile network.

These data are sometimes referred to simply as mobile phone data. However, it is important to distinguish them from the data collected by mobile phone devices themselves; this includes positioning data with different characteristics (e.g. greater precision) and data collected by sensors in the devices.

These data are mainly used to analyse the movements of people between countries, commuting or to measure economic activities. For instance, *Ahas et al. (2013)* assessed the feasibility of using mobile network positioning data for tourism statistics. *Ricciato et al. (2015)* researched the possibilities of estimating the population density distribution, and provided an overview of the types of data available from MNOs:

- Call detail records (CDRs) — ‘For each voice and data connection (or part of it) the network elements generate "tickets" that are sent to the billing system for charging purposes. The billing system stores these data in large databases, normally in the MNO warehouse. The term "Call Detail Records", and especially its acronym "CDR", commonly used nowadays to indicate all billing records, including those originated from data connections.’

- 'The Visitor Location Register and the Home Location Register — databases for subscriber data. The HLR stores the "permanent" subscriber parameters that are logically associated to the Subscriber Identity Module (SIM), like e.g. the IMSI[18].'
- Other systems:
  - 'Customer database. Every MNO maintains a data warehouse with private customer data. These are necessary e.g. for administrative, accounting and contractual purposes. (…).'
  - 'Lawful interception. Every MNO is obliged to maintain a lawful interception system and store certain data about the position and activity of its customers, to be made available to law enforcement staff upon order by a judge. (…).'
  - 'Location-based servers (LBS). Some operators deploy in their network commercial solutions to deliver so-called location-based services to part of their customers. These systems often involve one or more LBS servers connected to the network. These solutions are based on proprietary vendor technology, and their capabilities (in terms of share of population coverage and spatial accuracy) are highly dependent on the specific network configuration.'
  - 'Passive monitoring systems or signalling data. Some operators implement additional passive monitoring systems in support network operation and troubleshooting (e.g. *Ricciato, 2006, Ricciato et al., 2006*). These systems observe the whole signalling and traffic exchange between the network and the MSs [Mobile Stations, i.e. mobile devices] and can be used to infer the location of every MS with the highest possible spatial and temporal accuracy allowed by network-based data.'

In this section, we take CDR as the reference as this is the one most commonly used in research, although most conclusions apply to the others. CDRs normally include the following fields (*Ricciato et al., 2015*):

- 'IMSI (possibly encrypted),
- starting time and duration of the call or connection,
- type of call or connection (e.g. voice, text, data),
- Cell Global Identifier (CGI) of the starting cell where the call or connection was initiated. '

3.1.2. Populations observed in mobile network data

Problems affecting mobile network data are similar to those observed earlier in telephone surveys. *Lepkowski et al. (2007)* contains an overview of these problems. However, the main difference is that we do not sample or directly survey mobile phone users.

To assess how we can address selectivity in this big data source, we start by presenting the relationship between the several populations of objects and statistical units present in mobile network data (see Figure 1).

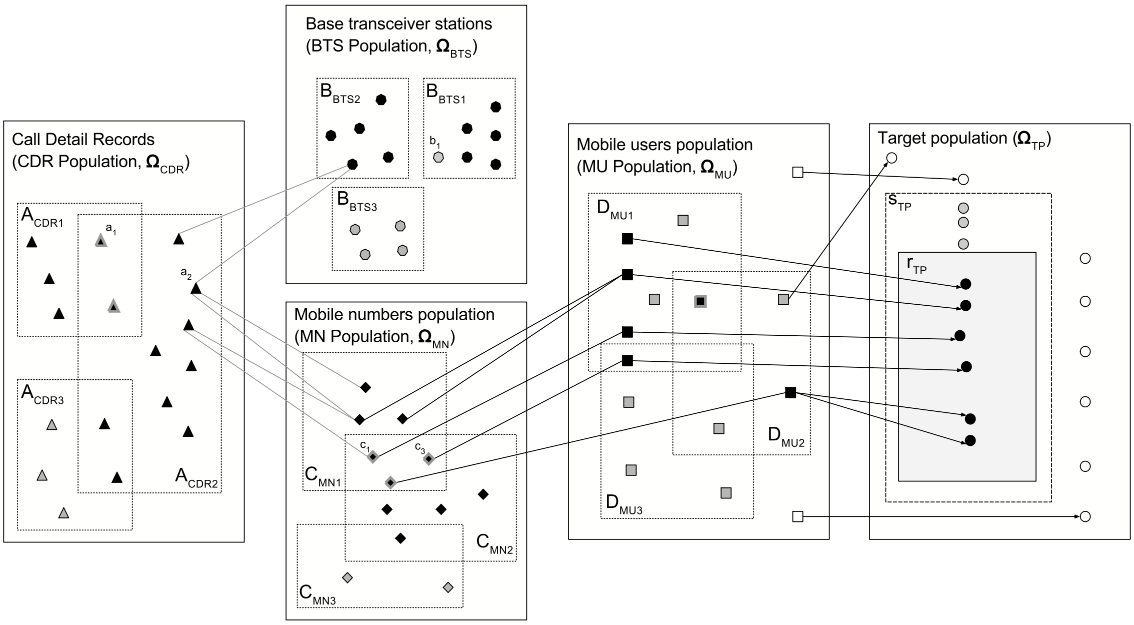
For simplicity's sake, we assume that there are three MNOs MNO$_1$, MNO$_2$, and MNO$_3$ and we obtain the data only from MNO$_1$ and MNO$_2$. Black shapes represent objects/units observed that are in frames; grey shapes refer to objects/units that are in frames but were not observed; and white shapes denote objects/units that are not covered by any frame (i.e. not captured by the data source). Solid grey arrows refer to links between objects, and solid black arrows refer to links involving statistical units.

The following populations of objects (solid black rectangles) and associated frames (dashed rectangles) can be identified:

[18] International Mobile Subscriber Identifier.

- CDR population denoted by Ω_{CDR}, with associated frames denoted by A_{CDR1}, $A_{\text{CDR2}}$and A_{CDR3}. A_{CDR1} and A_{CDR2}refer to the CDRs available respectively at MNO_1 and MNO_2 to which we have access, and A_{CDR3} refers to the frame of CDRs available at MNO_3 to which we did not have access. Naturally, the frames will overlap because each CDR is the record of a communication event between two mobile devices that can be clients of different MNOs. This would be the case for CDR a_1 in Figure 1, which would then be a case of double counting and over-coverage of the population of communication events.
- Mobile number population denoted by $\boldsymbol{\Omega}_{MN}$, with associated frames denoted by C_{MN1}, C_{MN2} and C_{MN3}. The overlap between the frames is presented due to possible transfers of mobile phone numbers between MNOs (e.g. c_1). Each CDR always involves two mobile numbers, and each mobile number will normally involve several CDRs.
- Base transceiver station (BTS) population denoted by $\boldsymbol{\Omega}_{BTS}$, with associated frames denoted by B_{BTS1}, B_{BTS2} and B_{BTS3}. The BTS frames do not overlap in most cases because each operator uses its own BTS infrastructure. However, in a few countries different operators share the same infrastructure; in that case, the frames overlap. Even for the BTS frame of a mobile operator that supplied data, not all BTSs might be present (e.g. b_1). This would be the case if no communication event involved the BTS for the time period for which data are available. Each CDR will have one BTS assigned, with each BTS usually linked to many different CDRs.

Figure 1 — Relationship between populations of different statistical units in mobile network data



The following populations of statistical units can be identified:

- Mobile user population denoted by $\boldsymbol{\Omega}_{MU}$, with associated frames denoted by D_{MU1}, D_{MU2} and D_{MU3}. White squares represent the mobile user population that is not observed in available frames (e.g. uses a small provider). Links between the mobile user population and the target population present cases of over-coverage and under-coverage.
- The target population is assumed to be the population of individuals denoted by $\boldsymbol{\Omega}_{TP}$, where we identify units without a mobile telephone number and units without a CDR either because they are from a different provider or because they do not have any call and 'responding' units r_{TP}.

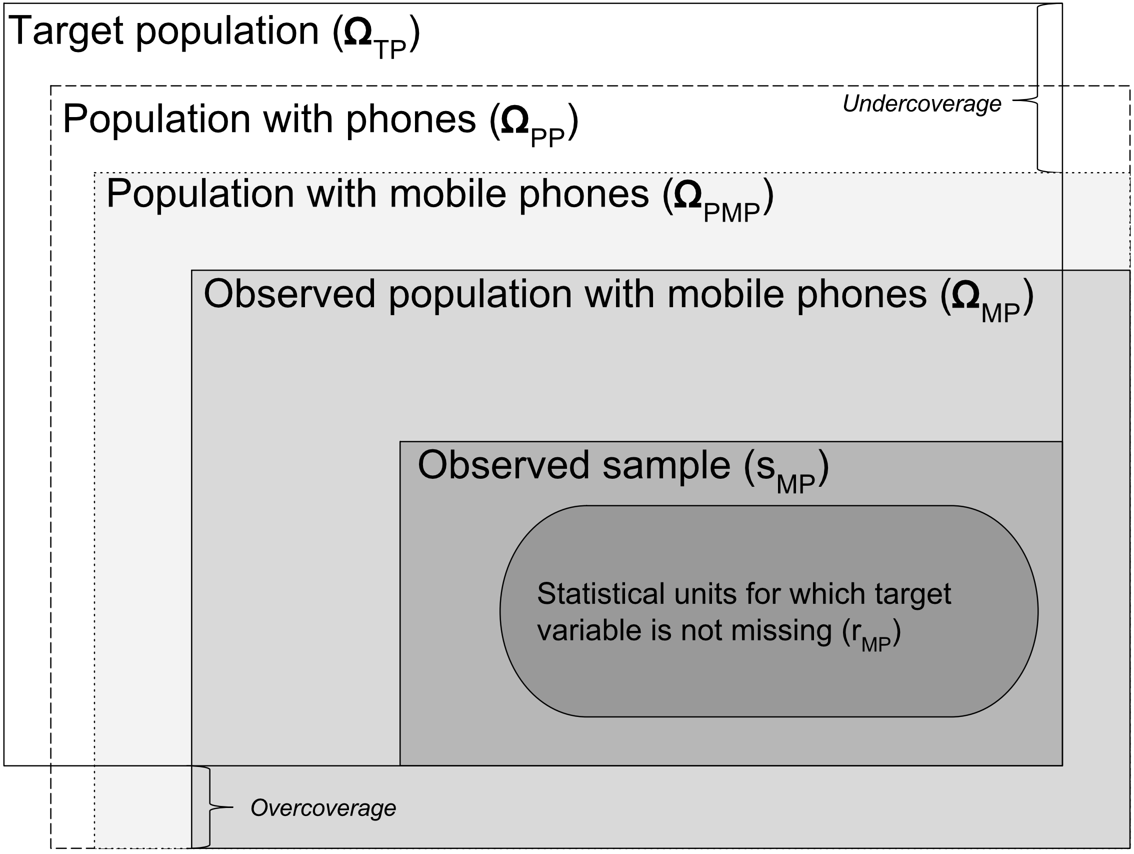
The relationship between populations provides some context on how these data are generated or captured and geolocated. Each row in the CDR has an assigned BTS. This relationship enables a certain location to be assigned to a mobile device at a certain moment in time.

An important issue from a statistical point of view is the fact that these relationships are complex and that multiple frames are observed.

3.1.3. Coverage of mobile network data

Figure 2 displays the relationship between the target population of individuals $\boldsymbol{\Omega}_{TP}$ and the final sample that can be obtained r_{MP}. We can distinguish between the population with phones (of any kind) $\boldsymbol{\Omega}_{PP}$ and the population with mobile phones $\boldsymbol{\Omega}_{PMP}$. The observed population $\boldsymbol{\Omega}_{MP}$ is the population that is observed within mobile frames, i.e. accessible from the mobile operators that provide the data, and s_{MP} is the sample that we have access to (e.g. limited to a certain period). We do not consider that $\boldsymbol{\Omega}_{MP} = s_{MP}$ because the sample observed is limited to members of the target population only, while $\boldsymbol{\Omega}_{MP}$ might suffer from over-coverage (e.g. objects that do not refer to any population member). Finally, we end with r_{MP} because we might observe missing data in the target variable for certain units. It might happen that, for all units from the observed sample, we do not have information on a target variable (e.g. location of residence) that needs to be imputed.

Figure 2 — From target to observed population in mobile network data


Target population ($\boldsymbol{\Omega}_{TP}$)
Population with phones ($\boldsymbol{\Omega}_{PP}$)
Undercoverage
Population with mobile phones ($\boldsymbol{\Omega}_{PMP}$)
Observed population with mobile phones ($\boldsymbol{\Omega}_{MP}$)
Observed sample (s_{MP})
Statistical units for which target variable is not missing (r_{MP})
Overcoverage

Finally, we would like to explain why self-selection (selectivity) is presented in terms of subsetting the target population. Regardless of the target population definition (e.g. trips, persons), some units will always be excluded. For instance, the mobile network infrastructure limits the exact identification of trips. However, this should be treated as a coverage error. Observed trips are a subset of the population of all trips (e.g. limit to one provider, only to those with mobile phones) and a trip can be identified ($R_i = 1$) that is also related to infrastructure coverage.

3.1.4. Selectivity in mobile network data

There are two sources of selectivity. One is connected with infrastructure (selectivity in coverage).

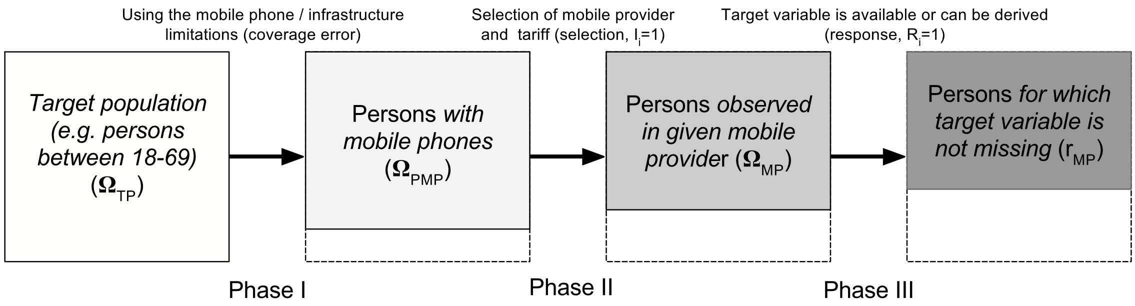
The second one is connected with mobile users (businesses and individuals, or contracts and prepaid cards). Mobile network data will be discussed in this context.

From the methodological point of view, we should treat mobile network data as an opt-in internet panel. Through mobile devices we can indirectly observe the target population, which can be persons or enterprises.

As mentioned, the positioning data from mobile networks refer to objects rather than the target population. Mobile operators are in possession of systems that provide background information on their customers (cf. customer relationship management systems). However, the analysis is normally limited to the CDR population because the relationship between the contractual client and the effective mobile device user is uncertain.

Figure 3 displays a hypothetical self-selection mechanism observed in mobile network data. We identify three phases. The first one refers to using mobile devices and infrastructure limitations, which corresponds to coverage error. The second phase refers to selecting a certain mobile provider and tariff. And finally, we refer to the sample of pseudo-respondents for which we have a response (observed directly or indirectly). We should point that if there is no data (i.e. no CDR because no calls were made) within a given time frame, data will also be missing.

Figure 3 — Self-selection mechanism in mobile network data


Using the mobile phone / infrastructure limitations (coverage error)
Selection of mobile provider and tariff (selection, I_i=1)
Target variable is available or can be derived (response, R_i=1)
Target population (e.g. persons between 18-69) (Ω_{TP})
Persons *with mobile phones* (Ω_{PMP})
Persons *observed in given mobile provider* (Ω_{MP})
Persons *for which target variable is not missing* (r_{MP})
Phase I
Phase II
Phase III

From a practical point of view, we would therefore like to address the following questions:

- What is the overall coverage of the target population in the mobile operators' frames to which we have access (population coverage)?
- What is the overall coverage of the infrastructure of the mobile operators supplying the data? (coverage by BTS)

The first question includes under-coverage of the target population. However, it should also be noted that the populations observed in mobile network operators refer to businesses and natural persons, which might overlap (e.g. one person with both a private and a business mobile phone). Over-coverage is therefore also an issue. The second question concerns the selectivity of the stations and at which level of aggregation these data can be used. The BTS is located according to the density of the population and is mainly presented in the literature using Voronoi diagrams (*Ricciato et al., 2015*).

3.1.5. Selectivity of infrastructure — BTS and cells

To assess the selectivity of the infrastructure, it is important to understand how it is built. The network infrastructure of telecommunications companies can, at the most basic level, be divided into:

- BTSs, which handle communication within a geographic area and can have multiple directional antennas linked to one BTS (each BTS has an exact location assigned to it);
- cellular geographic area covered by the same antenna, where each BTS is responsible for one or more cells.

BTSs have different technologies and might change over time (changes in the technologies from standard GSM to LTE). Modernisation of the infrastructure introduces additional coverage problems.

Telecommunications companies can only operate on limited radio frequencies. If all the frequencies are already taken by other users' calls or other services in the target area, then the next incoming calls will be handled by a more remote BTS. This process is called a handover. It will cause a measurement error (in the location of the device) and have an influence in case the statistical indicator is a function of the number of devices per BTS and may lead to bias. BTSs use a low power signal which, in conjunction with the use of multiple frequencies, allows radio frequencies to be reused without signal interferences between stations. As only certain frequencies are available to the operators, they need to adjust the placement of the BTSs so they do not interfere with each other. The areas covered by the stations are then divided by the usage of directional antennas placed with a certain angle.

In general, a strong mobile phone signal is more likely in urban areas, although these can also have some 'dead zones' with no reception. Cellular signals are designed to be resistant to multipath reception, which is most likely caused by a direct signal path being blocked by large buildings such as high-rise towers.

By contrast, many rural or sparsely inhabited areas lack any signal at all or have very weak fringe reception; many mobile network providers are trying to put up radio towers in areas most likely to be frequented by users, such as those along major roads. Even some national parks and other popular tourist destinations away from urban areas now have mobile phone reception. However, the location of radio towers within these areas is normally prohibited or strictly regulated, and is often difficult to arrange.

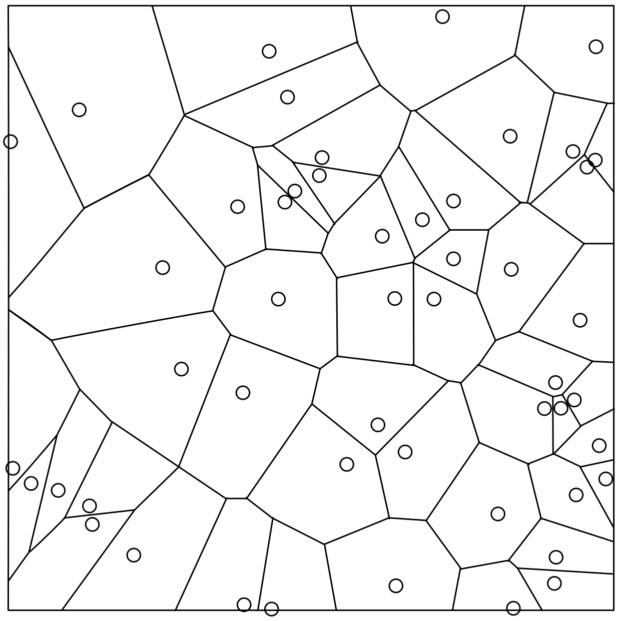
In areas where signal reception is usually strong, other factors can affect reception or may result in no reception at all (see RF interference). From inside a building with thick walls or made mostly of metal (or with thick reinforced concrete), signal attenuation may prevent a mobile phone from being used. Underground areas such as tunnels and metro stations will lack reception unless they are wired for cell signals.

There may also be gaps where the service contours of the individual base stations (cell towers) of the mobile provider (and/or its roaming partners) do not completely overlap.

Another key concept relates to the area covered by the BTS that is defined as a cell. The shape of the cells is not consistent with the administrative and statistical areas. In most cases, these cells are approximated using Voronoi diagrams, which represent the area covered by a single BTS. Figure 4 displays a Voronoi diagram with cells referring to BTSs. These cells represent the hypothetical coverage of a given area by 50 points. BTS coverage is correlated with the population density, which means that cities have high coverage of antennas compared to rural regions.

Because of the BTS infrastructure, it is not possible to provide the exact location of a mobile device.

Figure 4 — Example of Voronoi diagram based on BTS locations and resulting cells



Another issue is the existence of BTS frames in general. For example, according to the regulations in some countries, only a list of approvals is published. The data presented in such lists do not refer to BTSs that are currently in operation. The MNO has the final number of active BTSs.

Not all BTSs work all the time. For instance, there are several stations that are only enabled for special occasions (e.g. football matches, concerts). The main idea is therefore to verify the proportion of BTSs that are working continuously. This might enable us to verify the possibility of longitudinal observations.

To measure the coverage and selectivity of BTSs and cells, the following is needed:

- a list of all BTSs in a given country (if available),
- a list of all active BTSs of a given MNO (if available),
- a measurement of the coverage of cells (Voronoi diagrams) with regard to existing statistical and administrative regions,
- a list of BTSs working continuously or from time to time.

3.1.6. Self-selection process on the mobile phone market — is it ignorable or not?

As stated in the second part, we treat mobile network data in the context of opt-in panels. The reason is that each person decides whether to have a mobile phone, and also decides which mobile provider to select (including tariff, e.g. a prepaid card). There may well be situations in which the decision to select a provider is not independent (e.g. within the household). Finally, self-selection related to the target variable is mobile phone usage, which directly influences the possibility to measure mobility.

In general, potential outcomes of self-selection can be summarised as follows:

- the possession of a mobile device (yes or no, one or more),
- the selection of a mobile provider (one or more, which one),
- the selection of a tariff (contract, prepaid card, other) — do prepaid card users and contract users differ?
- the selection of device type (smartphone, basic mobile phone),

- the usage of a mobile phone.

The main issue in studying self-selection is to verify whether the mechanism is ignorable or not. This depends on what the target variable is and the situations where it is not recorded.

In the case of mobility, assuming full coverage of mobile phones and no attrition, the following cases are possible:

- people forget to take their mobile phones — this might be considered a random event, however the scale of such situations should be taken into account,
- people do not take their mobile phones on purpose — this cannot be treated as a random event. Another question that could be asked is the reason why such situations occur:
  - people are not very dependent on mobile phones (e.g. older people),
  - people do not want to be located (e.g. criminals).

However, bearing in mind how important mobile phones are for young people as well as for the working age population, these situations might be very unlikely.

Now, let us assume that we have access to the data from one mobile provider. We again consider the context of an opt-in panel. We start out with a general question — when is a missing data included in the target variable (e.g. mobility)? We could use panel attrition to explain this issue. In particular, we might consider the following cases:

- churn — a mobile user switches to another company/provider; this could be considered **voluntary attrition**,
- customer stops using a mobile phone — this could be due to 'churn' (migration to another provider but without having such information), not being very dependent on mobile phones (*passive attrition*) or death/relocation (*mortality attrition* — a given unit is outside the target population).

To sum up, we could consider churn, widely used in marketing, as *voluntary attrition*. However, it might not be connected to the target variable (e.g. mobility, economic activity), but to financial or private factors that do not influence the target variable.

Moreover, there are two main causes for mobile phone market selectivity: (i) infrastructure limitations, and (ii) individuals. The first cause makes it difficult to identify trips (if trips are regarded as the target population) where the infrastructure is sparse or in the case of overcrowded areas due to technical issues. The second cause refers to the decisions made by individuals — whether to use a certain mobile provider and mobile phone. As a result, we should carefully study these two sources to identify the missing at random (MAR) or missing not at random (MNAR) mechanism.

3.1.7. Limitations of mobile network data

First, the availability of background information on mobile device users might be related to the regulations of a given country. For instance, some countries recently approved 'anti-terrorist' laws that Ban the sale of prepaid cards without checking and recording the ID of the buyer.

Second, mobile phones are often registered to one person while other persons might be using them. For instance, individuals must have an ID card and be over 18 years old to have a contract. This problem also applies to prepaid SIM cards that might be bought by one person but used by another.

These problems refer to the concept of 'unit-error theory' that was introduced by *Zhang (2011)* in the context of register data. To sum up, we observe the following problems with inference based on mobile data:

- problems with defining and deriving populations observed in mobile network data,
- limited access to background information on mobile device users (place of residence, sex, age, employment status, marital status) — need for imputation (or profiling, as it is called in the literature),

- identification of over-coverage in frame(s),
- identification of statistical units,
- identification of exact locations,
- inclusion of uncertainty of imputation into estimates.

3.2. Online social networks — Twitter

3.2.1. Description of the data source

'Social media is the collection of websites and web-based systems that allow for mass interaction, conversation, and sharing among members of a network' (*Murphy, et al., 2013*). Online social media is not defined by a single type of platform or data. The list of popular platforms is long and can change rapidly. Widely recognised popular social media types include (*Murphy et al., 2014*):

- blogs (e.g. Blogger, WordPress, Tumblr),
- microblogs (e.g. Twitter),
- social networking services (e.g. Facebook),
- content sharing and discussion sites (e.g. YouTube, Reddit), and
- virtual worlds (e.g. Second Life).

Twitter is an online social networking service that enables users to post and read short 280-character messages called 'tweets'. Registered users can read and post tweets, while unregistered users can only read them. Users can access Twitter through the website interface, by text message or the mobile device app (*Twitter, 2016*).

Besides regular accounts, some accounts also are 'verified'. *Twitter (2016)* uses a blue verified badge to let people know that an account of public interest is authentic. This initially included accounts maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, business and other key interest areas. In July 2016, Twitter opened the possibility for any account of public interest to be verified (*The Verge, 2016*).

The company says it has some 190 000 verified accounts, although there are some 310 million monthly active users (the number of verified accounts can be checked at https://twitter.com/verified) (*The Verge, 2016*).

Twitter offers access to its data via an application programming interface (API). For instance, there is an R package 'twitteR' (*Gentry, 2015*) that can be used to obtain data from the Twitter Stream API. It requires registration on Twitter and the registration of an application to obtain API keys and tokens.

Besides the tweets, it is possible to obtain information about users, including:

- name — name of the user
- screenName — screen name of the user
- id — ID value for this user
- lastStatus — last status update of the user
- description — user description
- statusesCount — number of status updates the user has had
- followersCount — number of followers for the user

- favoritesCount — number of favourites for the user
- friendsCount — number of followers for the user
- url — a URL associated with the user
- created — when the user was created
- protected — whether or not the user is protected
- verified — whether or not the user is verified
- location — location of the user
- listedCount — number of times the user appears in public lists
- followRequestSent — if authenticated via OAuth, will be TRUE if you have sent a friend request to the user
- profileImageUrl — URL of the user's profile image, if one exists.

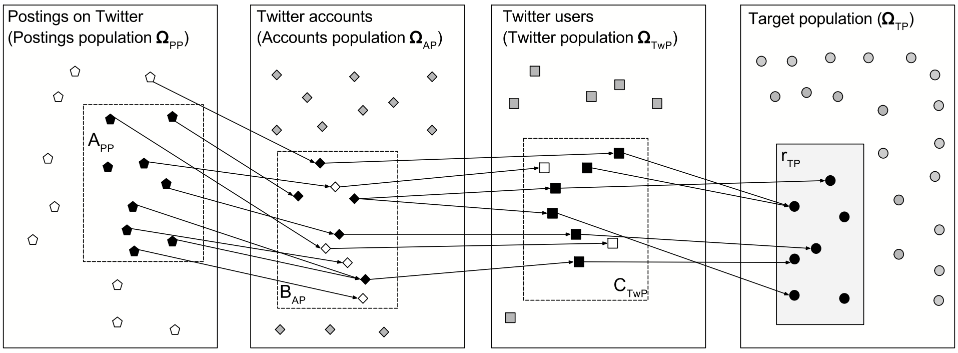
It is also possible to obtain followers (names and IDs) and friends (accounts that are followed with names and IDs).

3.2.2. Populations observed in Twitter data

Figure 5 displays the relationships between the several objects and statistical units observed in Twitter data, where we identify three populations. First, we identify the population of **tweets**, denoted by $\mathbf{\Omega}_{PP}$, which also includes retweets. Second, there is the population of **Twitter accounts**, denoted by $\mathbf{\Omega}_{AP}$, which includes 'fake accounts' (e.g. spammers), professional accounts or private accounts. Third, because the relationship between Twitter accounts and Twitter users is many-to-many, we identify the population of **Twitter users** (denoted $\mathbf{\Omega}_{TwP}$), which might significantly differ from the target population (denoted $\mathbf{\Omega}_{TP}$) but also from the population of social media users (denoted $\mathbf{\Omega}_{SMP}$). Dashed rectangles denote frames with regard to given populations that are created by data access. For instance, the size of the frame might vary if we are using a streaming or firehose API. Moreover, only public postings are available.

We start with a sample of tweets obtained (or purchased) from API that can be associated with Twitter accounts. Black diamonds refer to Twitter accounts that can be associated with real users, while white diamonds refer to fake, bot or institutional accounts. Grey diamonds refer to accounts that are not observed in the sample. Based on the Twitter account population, we also identify connections with Twitter users that are a subset of the target population. White squares denote Twitter users that (1) cannot be associated with the target population, (2) do not belong to the target population, or (3) we cannot derive a target variable ($R_i \neq 0$).

Figure 5 — Relationship between populations of different statistical units in Twitter data


Postings on Twitter
(Postings population $\mathbf{\Omega}_{PP}$)
A_{PP}
Twitter accounts
(Accounts population $\mathbf{\Omega}_{AP}$)
B_{AP}
Twitter users
(Twitter population $\mathbf{\Omega}_{TwP}$)
C_{TwP}
Target population ($\mathbf{\Omega}_{TP}$)
r_{TP}

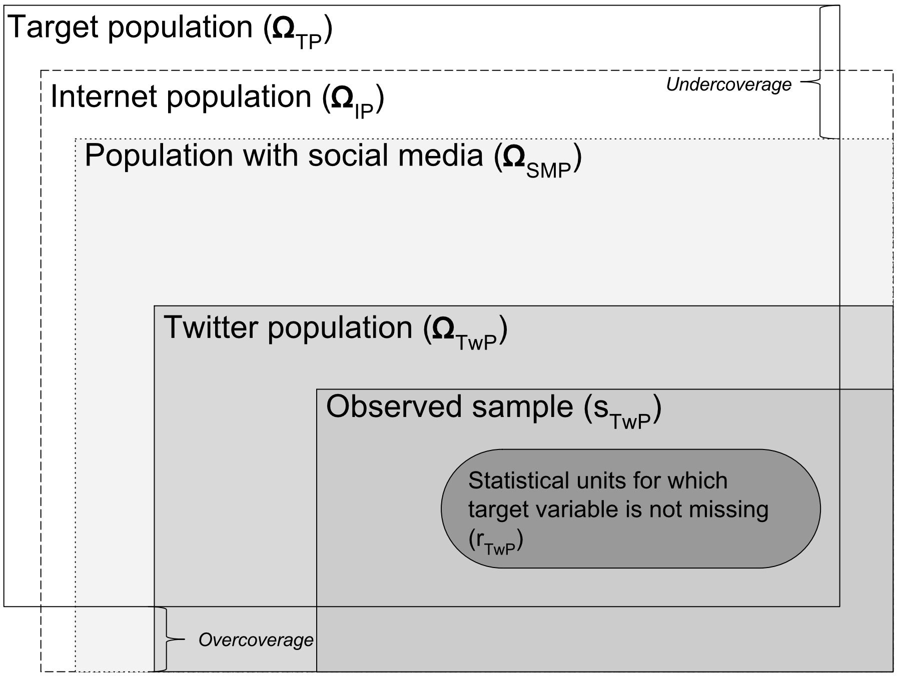
3.2.3. Coverage of Twitter data

Figure 6 shows the relationship between the target population ($\boldsymbol{\Omega}_{TP}$) and the populations observed on Twitter. We distinguish between the internet population ($\boldsymbol{\Omega}_{IP}$), the population with social media ($\boldsymbol{\Omega}_{SMP}$) and the Twitter population ($\boldsymbol{\Omega}_{TwP}$). Finally, depending on the data collection method we get an observed sample (s_{TwP}). The crucial part is to determine the target variable (which is often an unobserved variable) and we then get r_{TwP}.

The Twitter account population suffers from both over-coverage and under-coverage. The problem of over-coverage affects accounts that are associated with enterprises or are used to generate spam. It has been reported that some 7 % of Twitter accounts are used to produce spam or to automatically generate tweets (*Our Social Times, 2016*). *Chu et al. (2012)* proposed the following criteria to indicate whether an account may be a bot: (i) ‘periodic and regular timing’ of tweets, (ii) whether the tweet content contains known spam, (iii) the ratio of tweets from mobile versus desktop, compared to an average human Twitter user.

Moreover, there are Twitter accounts associated with private companies, newspapers, politicians or other entities (e.g. NGOs) that should not be associated with the target population. However, in the case of politicians, this might not be straightforward as some of them might use Twitter for personal opinions rather than official political ones.

Figure 6 — From target to observed population in Twitter data


Target population ($\boldsymbol{\Omega}_{TP}$)
Internet population ($\boldsymbol{\Omega}_{IP}$)
Undercoverage
Population with social media ($\boldsymbol{\Omega}_{SMP}$)
Twitter population ($\boldsymbol{\Omega}_{TwP}$)
Observed sample (s_{TwP})
Statistical units for which target variable is not missing (r_{TwP})
Overcoverage

3.2.4. Selectivity in Twitter data

We can identify three phases in the self-selection process present in Twitter data (see Figure 7). The first one is connected to the coverage error, in particular the propensity to have access to the internet. According to the Community survey on ICT usage by individuals and households, 15 % of households in the EU did not have access to the internet in 2016.

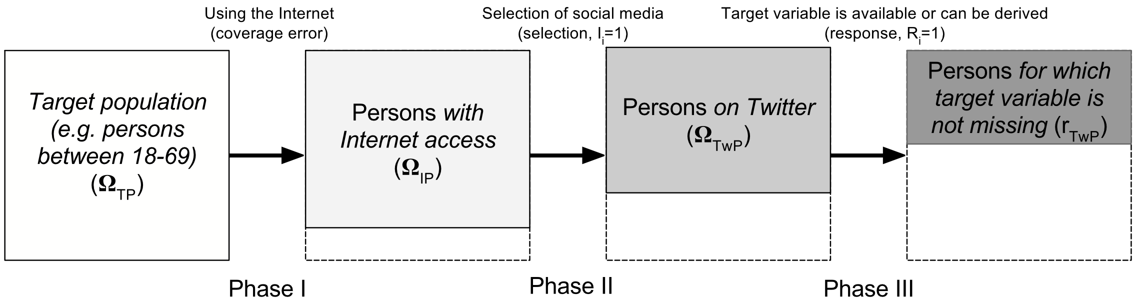
The second one is associated with the use of online social media in general and, in particular, with having a Twitter account. According to the survey, 46 % of EU residents aged between 15 and 65 did not regularly use online social media, Twitter or any other, in 2017.([19]) The use of online social media is dependent on the socioeconomic background of individuals, making the sample of users of online social media, and probably also those of Twitter, unbalanced. For example, while 85 % of individuals aged between 16 and 19 used online social media, only 45 % of those aged between 45 and 54 did so. While 63 % of individuals with higher education used online social media, only 39 % of those without secondary education did so.

The third phase relates to the activity of posting tweets as data are captured only when individuals tweet. If the statistics are produced over a period of time when some individuals never tweeted, then they will not be covered. The propensity to tweet might depend on the period of time (e.g. during New Year's Eve), space (e.g. during holidays) and also on the particular subject, which will often be related to the target variable.

To account for the self-selection error in Twitter data, the following should be taken into account:

- internet coverage — internet access, including mobile access,
- frame error — propensity to use social media and have a Twitter account,
- missingness — propensity to tweet;

Figure 7 — Self-selection mechanism in Twitter


Using the Internet (coverage error)
Selection of social media (selection, I_i=1)
Target variable is available or can be derived (response, R_i=1)
Target population (e.g. persons between 18-69) (Ω_{TP})
Persons *with Internet access* (Ω_{IP})
Persons *on Twitter* (Ω_{TwP})
Persons *for which target variable is not missing* (r_{TwP})
Phase I
Phase II
Phase III

One issue to be taken into account is the treatment of retweets. We consider tweets to be a manifestation of opinions on certain topics. Retweets can be considered as sharing someone else's opinion on a certain topic, although this might not necessarily be the case. Moreover, retweeting might be accompanied or followed with comments reflecting either agreement or disagreement, which might further introduce errors. To assess self-selection on Twitter, it is therefore crucial to determine which kinds of tweets should be associated with the target variable.

We stated previously that big data are paradata-designed and that the usage of these paradata is crucial to determine the self-selection mechanism. Most Twitter accounts do not have specified background information on the user and this must be derived. Once again, we should treat these variables as latent variables that can be imputed using paradata. There is a growing interest in research on determining the characteristics of users based on their activity on Twitter.

Bearing in mind the above, the following paradata might be taken into account when measuring selectivity:

- background information:
  - type of account (private, professional, institutional),
  - verification of profile,
  - name, surname or username,
  - photo and/or background photo,
  - description (might include occupation),

[19] Source: Eurostat (isoc_ci_ac_i).

- activity information:
  - number of tweets,
  - number of retweets,
  - number of users followed,
  - number of users following certain users,
  - hashtags used,
  - number of conversations and retweets,
  - types of users followed (e.g. politicians),
  - type of device used (web client, mobile phone, TweetDeck[20]), other),
  - activity over time.

3.2.5. Self-selection process on Twitter — is it ignorable?

A key question on the estimation based on big data sources is to verify whether the self-selection process is ignorable or not.

‘The problem of coverage in social media research is a nuanced one that goes deeper than the problem of Internet access alone. Once people are online, they behave in a variety of ways. This has been referred to as "differential use". The nature of social media data is such that the production of information online is almost never distributed equally across individuals. Some people post far more information than the average user and other individuals tend to lurk in the background, rarely, if ever, generating their own content (*Gruzd and Haythornthwaite, 2013*). Individuals who never post information may be invisible to certain sampling techniques. Those who rarely post information might be systematically undersampled. This process can bias the results of data collection toward the heaviest users. Differences between frequent and infrequent posters can be addressed by weighting individuals by the inverse of the frequency with which they post, yet ‘lurkers’ may be systematically different from active individuals in terms of their privacy preferences, opinions, and behaviors. Differences between posters and never-posters may be impossible to establish without alternative methods of data collection.’ *(Murphy et al., 2014, p. 19-20)*

‘Another factor affecting representativeness concerns access to the Internet and proficiency in using it (*Stern, Bilgen and Dillman, 2014*). There are clearly differences between Internet access disparities and the variation found in social media usage. For example, research has demonstrated that race, education, rurality, and socio-economic status play a role in social media usage and proficiency (*Stern, Adams and Elasser, 2009; Witte and Mannon, 2010*). Therefore, there are many nuances related to the questions of who uses social media and how that we have yet to fully understand. Concentrated study in this area will be important as social media use in research continues to evolve. ’ *(Murphy et al., 2014, p. 20)*

In order to understand whether certain posts are non-ignorable, we should take into account proxies to remove bias. Such a proxy might be the activity within a certain account, the kind of information that is shared or who a certain person follows. For example, on Facebook, the default privacy setting is that only articulated connections (friends) can see a user’s posts; on Twitter, the default privacy setting gives the public access to a user’s posts. Twitter allows users to have pseudonyms, whereas Facebook encourages ‘real name’ use, which alters how people create accounts (individuals are more likely to have multiple accounts on Twitter) and how organisations use these sites (including uncivil or manipulative *trolling*). Facebook users also have to agree to become friends (symmetric linking), whereas Twitter users can follow a user without reciprocity (asymmetric linking); this

[20] TweetDeck is a social media dashboard application for management of Twitter accounts. See https://tweetdeck.twitter.com/.

seemingly subtle difference has been associated with major differences in social practices of posting on those sites *(Schober et al., 2016, p. 203)*.

3.2.6. Limitations of Twitter data

Morstatter et al. (2013) recently compared two APIs that are available for Twitter. The first is a public streaming API that consists of 1 % of all tweets in a given period. According to the analysis, the sample will return at most 1 % of all the tweets created on Twitter at a given time. Once the number of tweets matching the given parameters reaches 1 % of all tweets, Twitter will begin to sample the data returned to the user. The methods that Twitter employs to sample this data are currently unknown. The streaming API takes three parameters: keywords (words, phrases or hashtags), geographical boundary boxes, and user ID. One way to overcome the 1 % limitation is to use the Twitter Firehose — a feed provided by Twitter that allows access to 100 % of all public tweets. *Morstatter et al. (2013)* concluded that:

- ‘Topical analysis is most accurate when we get more data from the Streaming API.
- The Streaming API almost returns the complete set of the geotagged tweets despite the sampling. (…) Although the number of geotagged tweets is still very small in general (1 %), researchers using this information can be confident that they work with an almost complete sample of Twitter data when geographic boundary boxes are used for data collection.
- In the case of top hashtag analysis, the Streaming API sometimes reveals negative correlation in the top hashtags, while the randomly sampled data exhibits very high positive correlation with the Firehose data.’

The overall finding of *Morstatter et al. (2013)* is that the suitability of the streaming API heavily depends on the coverage and the type of analysis that the researcher wishes to perform.

Diaz et al. (2016) also recently used Firehose to capture all of the English language tweets made between 1 August 2012 and 6 November 2012 that mentioned the words *Obama* and/or *Romney*. ‘For each tweet, we captured the text of the tweet, shared URLs, the date and time when the message was written, and profile information about the author, including their name and self-identified location.’

In general, the following limitations exist regarding the use of Twitter for statistics:

- limited access to background data (sex, age),
- problems with latent variables— background and target variables are observed indirectly,
- research indicates that Twitter is used for political discussions,
- bots and non-human accounts,
- localisation problems,
- ‘pointless babble’ messages,
- is a retweet an endorsement?

Schober et al. (2016) provided an interesting discussion about social media. They argued that for social media analyses, topic coverage can in principle be achieved without population coverage. In other words, other mechanisms of information propagation that are particular to the dynamics of social media may lead a corpus of social media posts to reflect the broader population’s collective opinion and experience through a range of (not yet fully understood) possible mechanisms. *Schober et al. (2016)* identify three mechanisms:

- ‘**audience design**: social media postings may reflect users’ judgments of what their audience (friends and followers and unknown others) is interested in hearing about right now, and the collection of audience interests across the many networks within a massive social media site may reflect — even represent — the broader public’s interests;

- **propagation**: ideas that resonate more broadly (both within and beyond a user's social media network) may be more likely to survive and flourish or "cascade" — be "liked", retweeted, replied to, and lead to new followers or friends;
- **media reflection**: social media postings reflect the issues disseminated by multiple media outlets and thereby reflect the current public agenda.'

Schober et al. (2016) propose that non-representative samples of social media users can post content that potentially represents the larger population's opinions and experiences.

3.3. Web activity data — Google Trends

3.3.1. Description of the data source

Google Trends[21] is an index computed from the individual search queries performed by users on the Google search engine *(Google Trends, 2016a)*. The indexes are created based on a sample of individual search queries that changes every day.

Google Trends adjusts the index so it is comparable between search terms and regions. Otherwise, places with the most search volume would always be ranked the highest. To do this, each data point is divided by the total number of searches of the geography and time range to which it refers to compare relative popularity. The resulting numbers are then scaled to a range of 0 to 100 *(Google Trends, 2016b)*. For example, users in Fiji and Canada may have the same search index value if they are equally likely to search for a hotel. However, they may not have the same number of total searches for this term.

Search queries are classified into 25 different categories based on natural language processing methods applied to the search terms *(Choi and Varian, 2012)*. The number of subcategories varies from 9 (for the category *Property*) to 157 (for the category *Business & Industrial*). For instance, in the *Jobs & Education* category there are two main subcategories (*Jobs* and *Education*), with *Jobs* including four subcategories: *Resumes & Portfolios*, *Job Listings*, *Career Resources* and *Planning and Developer Jobs*.

3.3.2. Populations observed in Google Trends

Google does not provide access to the data on the individual search queries. However, it is still useful to clarify the different objects of the index.

The most basic object is the **search query**, with the query population denoted by $\boldsymbol{\Omega}_{SQ}$ (see Figure 8). Some search queries are excluded *(Google Trends, 2016a)*:

- searches made by very few people: Google Trends (GT) only analyses data for popular terms, so search terms with low volumes appear as 0,
- special characters: GT filters out queries with apostrophes and other special characters (e.g. *King's dog* will be converted to *Kings dog*),
- duplicate searches: GT eliminates repeated searches from the same person over a short period of time.

Duplicate search queries are eliminated with the help of cookies,[22] which are used to distinguish users. The cookie created by the Google search engine includes a **user ID** that identifies the user

[21] https://trends.google.com/.

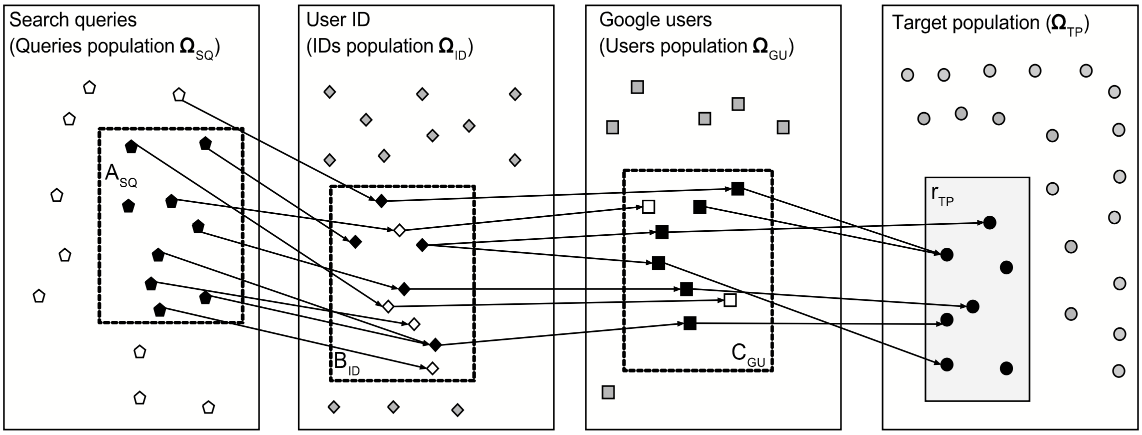
[22] A cookie is a file stored by the web browser containing data interchanged with websites over different browsing sessions.

and the device used. This is the second object, with the ID population denoted by $\mathbf{\Omega}_{ID}$. The user ID is a Universal Analytics feature that Google uses to associate multiple sessions (and any activity within those sessions) with a unique ID.

With it, Google offers a more accurate count of different **Google users**, with the user population denoted by $\mathbf{\Omega}_{GU}$. For more details, see *Google Analytics (2016a)*, *Google Analytics (2016b)* and *Google Analytics (2016c)*.

However, each user ID is counted as a unique user, so the same user making the same search query on different devices is counted more than once. Different web browsers on the same device, as well as instances of private browsing, are also counted as unique devices. If a user signs in on two different browsers on the same laptop, two unique devices are attributed to that one user ID.

Figure 8 — Relationship between populations of different statistical units in Google Trends

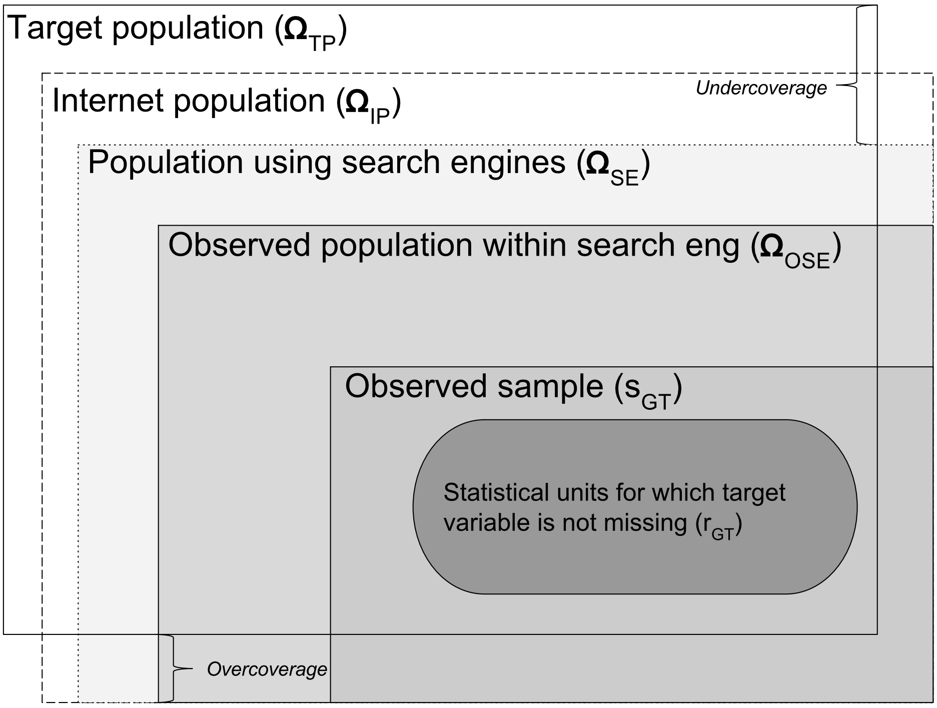

Search queries
(Queries population $\mathbf{\Omega}_{SQ}$)
A_{SQ}
User ID
(IDs population $\mathbf{\Omega}_{ID}$)
B_{ID}
Google users
(Users population $\mathbf{\Omega}_{GU}$)
C_{GU}
Target population ($\mathbf{\Omega}_{TP}$)
r_{TP}

3.3.3. Coverage of Google Trends

The coverage by Google Trends depends on the internet access (broadband, mobile) and on the internet population. As a natural consequence, the Google user population is a subset of the internet population, in particular units that use the Google search engine.

Figure 9 displays the relationship between the Google Trends population and the target population. In addition to the populations specified previously, we identify populations of search engine users and a sample of users that are using the Google search engine. Finally, we identify a sample of pseudo-respondents that send search queries via Google.

Figure 9 — Target population on Google Trends


Target population (Ω_{TP})
Internet population (Ω_{IP})
Undercoverage
Population using search engines (Ω_{SE})
Observed population within search eng (Ω_{OSE})
Observed sample (s_{GT})
Statistical units for which target variable is not missing (r_{GT})
Overcoverage

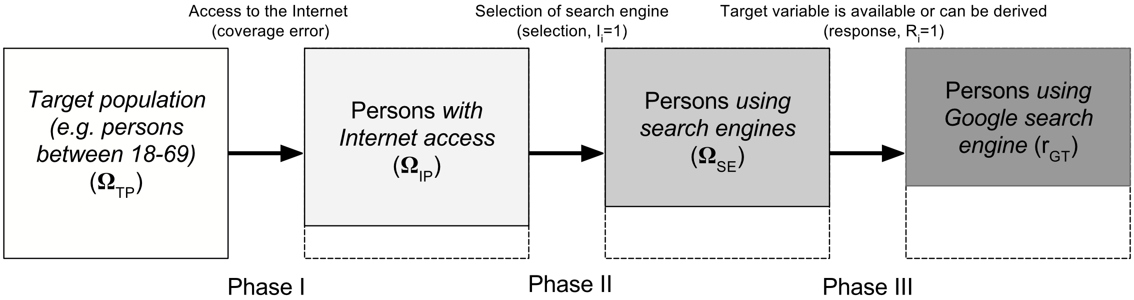
Similar to other big data sources, we do not directly observe the target variable, e.g. looking for a job, but rather a manifestation of it by means of search queries. We associate the usage of certain keywords as a willingness to change or look for a job. For instance, there are several papers that predict the unemployment rate based on Google Trends data *(Fondeur and Karamé, 2013; Vicente et al., 2015)*.

Nonetheless, Google Trends does not distinguish the populations' indices presented by Google. We might assume that it is correlated with the number of daily or monthly internet users. Moreover, Google users might also be considered as an opt-in panel. However, *Diaz et al. (2016)* show that: (i) online user demographics are not only biased, but change dramatically depending on major events, (ii) individual level engagement varies greatly depending on major events, and (iii) the nature of a user's activity in online and social media discussions also changes depending on major events.

3.3.4. Selectivity in Google Trends

The self-selection mechanism in Google Trends can only be approximated. Due to the lack of access to unit-level data or queries that are stratified by sex or age groups, it is not possible to reweight it based on the estimated internet population. However, discrepancies between Google Trends published on a given spatial level with target variables estimated from register or survey data might provide an approximation of the spatial self-selection process. Figure 10 shows the hypothetical self-selection mechanism in Google Trends.

Figure 10 — Self-selection mechanism in Google Trends


Access to the Internet
(coverage error)
Selection of search engine
(selection, I_i=1)
Target variable is available or can be derived
(response, R_i=1)
Target population (e.g. persons between 18-69) (Ω_{TP})
Persons with Internet access (Ω_{IP})
Persons using search engines (Ω_{SE})
Persons using Google search engine (r_{GT})
Phase I
Phase II
Phase III

For the sake of clarity, the following self-selection process sources can be observed in Google Trends:

- internet coverage — propensity for having access to the internet (e.g. broadband, mobile),
- internet usage (e.g. in last 3 months) — propensity for being active on the internet,
- search engine usage — propensity for using the Google search engine.

Assessing Google Trends self-selection is possible by using internet panels of representative random samples that follow the behaviour of internet users, e.g. by logging individuals' web activities, including the use of search engines (special software is installed on users' devices). This data collection would of course require the consent of users and need to be done in an unobtrusive way.

3.3.5. Limitations of Google Trends data

Google Trends has most commonly been used as a proxy variable of a given phenomenon or a covariate in nowcasting (the prediction of the present, the very near future and the very recent past). The direct use of Google Trends for statistics is more problematic due to the lack of background information on the person that used the Google search engine. For instance, a given computer can be used by several users, and many people can use it to look for the information online. However, there are examples where typing can be used in discrimination analysis to verify or identify persons.

Another limitation affecting Google Trends is the spatial availability of data. These data are currently only available at country or regional level. However, it is possible to limit the query on search terms to cities (e.g. by writing "work in Warsaw").

The next problem is the algorithms that Google uses to classify terms into categories. The description provides just a brief description of the algorithm, and the classification error is unknown. However, we might assume that it is low due to the high number of queries supplemented by the intersection with query results.

Another issue is the time frequency provided by Google Trends. By default, these data are available on a weekly basis, which sometimes makes the calculation of monthly data difficult. However, it is possible to obtain daily data by selecting the option 'last 90 days' on the Google Trends webpage or by using packages to download the data, e.g. gtrendsR (*Massicotte and Eddelbuettel, 2016*). These data are scaled from 0-100 for a 90 day period, which requires a transformation to be comparable with other periods. One possible solution is to take overlapping periods of 90 days and rescale them using the ratio between the numbers of searches for the overlapping days.

The following limitations should be considered when using Google Trends:

- limited access to background information about users,
- constant changes to the internet population limits the possibilities of using weighting schemes,
- the spatial aggregation of Google Trends is limited and varies between countries,
- measurement error — the target variable is not directly observed in Google Trends, it is only related to search queries.

3.4. Web activity data — Wikipedia usage

3.4.1. Description of the data source

Wikipedia was founded in 2001 by Jimmy Wales and Larry Sanger. The objective was to create a free online encyclopedia that anyone can edit. In the last 15 years, it has grown to 38 million articles in 246 languages. It is widely used, with 21 million page views an hour. According to official EU statistics, 44 % of individuals aged 16 to 74 living in the EU *consulted wikis to acquire knowledge (e.g. Wikipedia)* in 2013. The figure was 69 % for those aged between 16 and 24 *(Reis et al. 2016)*.

Giles (2005) presented a report where it is argued that Wikipedia comes close to Britannica in terms of the accuracy of its science entries. *Hilles (2014)* summarised that 'we should realise that Wikipedia is indeed "good enough" when it comes to basic facts and information'.

Brown (2011) found that while Wikipedia's political data appear to be accurate, they contain serious errors of omission. Wikipedia's omissions follow a predictable pattern: coverage is best on topics that are more recent or prominent. *Brown (2011)* argued that even if Wikipedia is perfectly accurate, it is not suitable as a sole source for students, who ought to be consulting better resources than encyclopedias.

Wikipedia page views provide the number of page views for each article per hour. The following factors may influence the popularity of a page:

- general popularity — a subject well known to most people will probably get more views than one that is naturally more obscure,
- current events — the subject of (or related to) a current event will likely get many more views when it is receiving media coverage than when it is little discussed by the public,
- current, unrelated events — auburn (the colour) received a spike in views immediately after an important American football game involving Auburn University's football team, even though the two articles have no direct link at all (other than the hair colour being prominently located on top of the disambiguation page Auburn),
- incoming links from other Wikipedia pages — a page is more likely to get viewed when other Wikipedia pages link to it.

Wikistats([23]) provides the following services to obtain data about:

- General
  - database reports — an index of automatically generated reports about English Wikipedia (GitHub),
  - size of Wikipedia — by the count of articles and gigabytes of data,
  - size of Wikipedia in volumes — to see the size of Wikipedia in print encyclopedia volumes,
  - size comparisons — comparisons against other encyclopedias and information collections,
  - modelling Wikipedia's growth — analysis of the count of articles, attempting to fit mathematical growth models,
  - number of words, edits, articles (by month), e.g. for English,

([23]) https://stats.wikimedia.org

  - time between edits — length of time (measured in days) between each block of 10 million edits made in Wikipedia,
  - article traffic jumps — a place to document unusual jumps in article traffic.
- Articles and edits
  - list of Wikipedians by number of edits — 10 000 editors with the highest edit counts, updated daily,
  - list of Wikipedians by number of recent edits — updated based on data as of 16 January 2011,
  - list of Wikipedians by article count — 5 000 editors with the highest article counts updated weekly.
- Images
- Links
  - external links ranking — top 1 000 most linked domains from external links as of April 2011 (all namespaces),
  - Linkypedia.
- Page views
  - Pageviews Analysis — article traffic statistics, also allowing comparisons between articles (documentation),
  - Wikitop — top 30 most popular articles by categories, with user comments on traffic jumps,
  - Wikirank — most viewed or most important articles by Wikidata categories,
  - Wikipedia’s reach, traffic and ranking compared to other websites — graphs and comparison statistics provided by Alexa Internet,
  - Wikipedia compared with other sites — graph using the above to compare Wikipedia against various top 10 sites,
  - lists of popular pages by WikiProject — these projects use pageview data to focus article improvement efforts on popular but poor quality articles,
  - Top 25 Report — top 25 most popular articles in a weekly chart,
  - TreeViews — monthly view statistics for category trees,
  - Wikitrends — top 10 today/this week/this month, and trends (also links for Wikipedia in other languages),
  - popular pages — current weekly 5 000 most popular articles based on raw data,
  - multiyear ranking of most viewed pages — top 100 list of most viewed pages for 2007-2016,
  - Wikipedia is more popular than... — a list of Alexa traffic comparisons,
  - traffic stats calculation — a PHP program that calculates monthly stats for a list of articles, using stats data provided by the Pageview API.
- Deletion and vandalism statistics
  - statistics about deletion and other administrator actions can be found at the WP:Adminstats page,
  - a list of recently deleted files can be found at https://en.wikipedia.org/wiki/Special:Log/delete,

- revision deletion statistics can be found here.

Moreover, it is possible to obtain dump files of Wikipedia page views([24]) as well as for the content of the articles. There are also external tools such as the wikipediatrend R package (*Meissner and Team, 2016*).

3.4.2. Populations observed on Wikipedia

The main question on using Wikipedia for official statistics is the population that we are interested in. For instance, *Reis et al. (2016)* used the number of page views, not counting access via mobile devices and access identified as done by non-humans (i.e. bots) and the content of the articles.

We can distinguish between the following populations of statistical units observed in Wikipedia data sources:

- population of Wikipedia users — users that visit Wikipedia to acquire knowledge,
- population of Wikipedia editors — active users that create, edit or delete articles,
- population of Wikipedia articles — knowledge that is acquired from Wikipedia,
- population of Wikipedia activities — observed in log files that provide information on user activities such as page views.

The usage of Wikipedia might also be associated with topic analysis similar to Twitter data or Google Trends. For instance, we would be interested in the popularity of certain articles or categories of articles that might reflect a topic.

3.4.3. Limitations of Wikipedia data

The limitations of Wikipedia are similar to those observed in Google Trends, in particular:

- limited access to background information about users (some for content and editions and none for page views),
- constant changes to the Wikipedia users' population (as well as the internet population), which limits the possibilities to use weighting schemes,
- measurement error — target variable is not directly observed, it is simply related to the access or editing of articles.

However, unlike Google Trends, Wikipedia provides current and historic object-level data consisting of logs and user activities. These data are publicly available on a daily basis with timestamps and for each Wikipedia language version.

[24] https://dumps.wikimedia.org/other/pagecounts-ez/.

4 Unit-level methods to correct selectivity

4.1. Pseudo-design approach — reweighting

Reweighting means adjusting the weighting system in sample surveys or opt-in panels to account for non-response or selectivity and to improve accuracy. The method presupposes having access to auxiliary variables from reliable sources (e.g. demographic characteristics from official statistics or independent probability samples) that correlate with the study variables and/or the selection mechanism. Weighting is associated with sample surveys, where the initial design weights (inverses of inclusion probabilities) can be used if there are no non-sampling errors. However, initial design weights are inappropriate for non-sampling errors such as selective non-response. In these cases, adjusted weights are often constructed by using a reweighting method such as post-stratification or calibration. Reweighting methods are routinely used in official statistics and have much potential in accounting for selection bias in big data sources.

A particular property of big data sources is that we need to deal with an unknown probability of inclusion. The question is then how we can apply reweighting methods when inclusion probabilities are unknown. The unknown inclusion probabilities are often substituted by pseudo-inclusion probabilities that are estimated from the available data. The situation is similar in other types of non-probability samples. *Elliott (2009)* discusses the ‘pseudo weights’ for combining data from probability and non-probability samples. *Elliott and Valliant (2017)* discuss methods where the unknown inclusion probabilities are estimated under a pseudo design-based approach. In this method, the pseudo-inclusion probabilities are estimated using the available covariate or auxiliary variables that are known both for the sampled and non-sampled units. The authors also discuss model-based methods for this purpose. If there is a lack of covariate information, an assumption of equal inclusion probabilities may be postulated. One of the methods explained further on in this report can then be used to adjust the initial weights.

In the case of web surveys, individuals sometimes ‘respond’ more than once. Moreover, such respondents may do so deliberately in order to influence the results, and without disclosing their multiple responses. This is a challenging problem because those ‘respondents’ may use different computers, or even the same computer with a different IP address. One possible way to deal with this problem is to model the number of times an individual ‘responds’, for example using a log-linear model that incorporates the auxiliary covariate variables used for modelling. Similar modelling methods can be applied to big data sources.

4.1.1. Generalised weight share method

The generalised weight share method (GWSM) may be used for correcting selectivity in big data sources when the statistical units in the big data source can be linked to statistical units in an existing frame population.

The GWSM was initially proposed to deal with the weighting procedure in longitudinal surveys. The problem that the GWSM aims to solve is specifying weights when statistical units change with time. For instance, when a household is sampled but splits or merges with another household in the next survey cycle.

The GWSM is also used in indirect sampling to address practical problems regarding the lack of sampling frames for a given population. For instance, when there is no sampling frame for households and a frame of dwellings has to be used instead. Basically, the GWSM aims to share sampling (or calibrated) weights to units of the target population.

We will follow the notation for the GWSM based on *Deville and Lavallée (2006)* and *Lavallée (2009)*. Let's assume that we have two populations $\mathbf{\Omega}_A$ and $\mathbf{\Omega}_B$. For the first population we have a sampling frame and we have selected a sample. However, our target population is $\mathbf{\Omega}_B$ for which unfortunately we do not have such a frame. However, we assume that units in $\mathbf{\Omega}_A$ are linked to units in $\mathbf{\Omega}_B$. Hence, we would like to take this into account and we would like to use the structure of linkage between units to provide weights for units selected from population $\mathbf{\Omega}_B$. Note that the size of sample s^B will be a random variable because we do not control the sample size obtained through indirect sampling.

Equation (1) presents the vector $\mathbf{W}$ of GWSM weights $w_i = \sum_{j=1}^{N_A} t_j^A \, \tilde{\theta}_{ji}^{AB} / \pi_j^A$:

$$\mathbf{W} = \widetilde{\mathbf{\Theta}}_{AB}' \, \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A, \tag{1}$$

where $\mathbf{\Pi}_A = diag(\boldsymbol{\pi}^A)$ of size $N^A \times N^A$ and $\boldsymbol{\pi}^A = \{\pi_1^A, \dots, \pi_{N^A}^A\}'$ denotes probabilities of units selected for the sample s^A. Then, let $\mathbf{t}^A = \{t_1^A, \dots, t_{N^A}^A\}'$ where $t_j^A = 1$ if $j \in s^A$, and 0 otherwise. Let $\mathbf{T}_A = diag(\mathbf{t}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the indicator variables t_j^A. Let $\mathbf{1}_A$ be a column vector of 1's of size N^A. Finally, $\widetilde{\mathbf{\Theta}}_{AB}$ denotes the standardised link matrix given by

$$\widetilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB} \left[diag(\mathbf{1}' \mathbf{\Theta}_{AB}) \right]^{-1}, \tag{2}$$

where $\mathbf{\Theta}_{AB} = \left[\theta_{ji}^{AB}\right]$ is a link matrix between $\mathbf{\Omega}_A$ and $\mathbf{\Omega}_B$ of size $N^A \times N^B$, where $\theta_{ji}^{AB} > 0$ indicates that two units are linked, otherwise units are not related to each other. For efficiency reasons, matrices used to derive these weights should be sparse (see the Armadillo C++ library which might be useful for that purpose http://arma.sourceforge.net).

Lavallée (2007) shows that it is possible to associate the calibration of *Deville and Särndal (1992)* to the GWSM method in order to adjust the GWSM weights so that they reproduce the known population totals of the auxiliary variables. *Maia (2009)* proposes using the GWSM in connection with multiple sampling frames in order to account for coverage bias.

4.1.2. Model-free calibration and extensions

4.1.2.1. MODEL-FREE CALIBRATION

Model-free calibration is the most common reweighting method used in official statistics to improve the quality of sample estimates using auxiliary information. This method is called 'model-free' (*Särndal 2007 p. 107*) because it is not necessary to explicitly specify a model to derive the calibration equations. The method is also called multi-purpose calibration, because the same weighting scheme can be applied for all study variables in a survey. This property is often considered important in official statistics.

Calibration is usually used for three purposes:

i. to achieve coherence between the survey estimates and published official statistics,

ii. to improve efficiency of estimates, and

iii. to account for the bias due to non-response.

Calibration can also be used to correct selectivity in big data sources when we can find auxiliary variables that are correlated with the selection mechanism and for which we know the population totals.

The formal background for calibration was proposed by *Deville and Särndal (1992)* and is summarised below. Special cases of calibration are known as ‘raking’ and ‘post-stratification’.

Let $\mathbf{d} = (d_1, \ldots, d_n)'$ be a vector of design weights ($d_i = 1/\pi_i$), $\mathbf{w} = (w_1, \ldots, w_n)'$ a vector of final weights, x_j the auxiliary variable for which totals are known $\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}$, where N is the population size and n is a sample size. Therefore, in order to get $\mathbf{w}$ we need to resolve the following set of equations, see e.g. *Vanderhoeft (2001)*:

$$\begin{aligned} D(\mathbf{w}, \mathbf{d}) &= \sum_{i=1}^{n} d_i \, G(\frac{w_i}{d_i}) \rightarrow min, \\ \sum_{i=1}^{n} w_i \, x_{ij} &= \mathbf{X}_j, \mathrm{j} \ = \ 1, \ldots, \mathrm{k}. \end{aligned} \tag{3}$$

Equation (3) is known as a calibration equation, which should be fulfilled in order to ensure that the sample totals and the known population totals of the auxiliary x-variables are equal. A set of calibration equations can further restrict the weights given by the following equation:

$$L \leq \frac{w_i}{d_i} \leq U \tag{4}$$

where $0 \leq L \leq 1 \leq U$ and $i = 1, \ldots, n$. Equations (3) - (4) minimise distance function G between known weights $\mathbf{d}$ and unknown weights $\mathbf{w}$. *Deville and Särndal (1992)* listed five different distance functions with somewhat varying properties but sharing the property of producing asymptotically equivalent calibration estimators, see also *Deville, Särndal and Sautory (1993)* who also discuss computational methods. Commonly used distance measures in calibration practice include the ‘linear method’ with distance function $G(x) = 1/2(x-1)^2$, where $x = w_i/d_i$. This distance function corresponds to the minimisation of a chi-square distance. Another is the multiplicative (or raking ratio) method with distance function $G(x) = x \log x - x + 1, \ x > 0$. A property of linear calibration is that calibration weights w_i are obtained analytically whereas the raking method requires iterative computation.

The calibration estimator for population total $Y = \sum_{i \in U} \mathrm{y}_i$ can be expressed as:

$$\hat{Y}_{cal} = \sum_{i \in s} w_i \, \mathrm{y}_i, \tag{5}$$

where the weights w_i are obtained by using one of the distance functions. The linear method produces weights given by $w_i = d_i(1 + \boldsymbol{\lambda}'\boldsymbol{x}_i)$ with $\boldsymbol{\lambda}' = (\sum_{i \in U} \boldsymbol{x}_i - \sum_{i \in s} d_{\boldsymbol{i}} \boldsymbol{x}_i)'(\sum_{i \in s} d_{\boldsymbol{i}} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}$, where $\boldsymbol{\lambda}$ is a Lagrange multiplier vector. *Deville, Särndal and Sautory (1993)* developed a generalised raking method as an extension to the classical raking ratio of *Deming and Stephan (1940)* and introduced the CALMAR software for making generalised raking estimates.

4.1.2.2. EXTENSIONS

Initially, calibration did not take into account non-response. *Deville (2000)* introduced a generalised calibration approach for weighting for non-response in surveys. *Lundström and Särndal (1999)* and *Särndal and Lundström (2005)* extended calibration to non-response to an item or unit. The main idea behind this approach is to get a set of weights that account for both non-response and auxiliary variables. The success of the calibration approach to reducing bias caused by non-response depends on the strength of the auxiliary variables and on the assumed non-response mechanism. Typically, calibration can reduce bias when auxiliary x-variables are related to an assumed non-response mechanism of missing at random (MAR) type. However, if a set of x-variables does not account for the MAR mechanism, calibration will not necessarily reduce the bias.

Recent extensions in calibration methodology offer the potential to adjust selection bias in big data sources. *Särndal and Lundström (2005)* introduce a one-step approach for non-response adjustment with simultaneous calibration to auxiliary variable totals and calibrated non-response adjustment. They also developed a two-step approach with non-response adjustment followed by a separate calibration phase. *Chang and Kott (2008)* discuss calibration to adjust for unit non-response when the response model and covariates in calibration may differ. *Kott and Liao (2015)* propose calibration weighting for non-response adjustment when performed in two steps: from the respondent sample to the full sample to remove the response bias and then from the full sample to the population to decrease variance. *Haziza and Lesage (2016)* also discuss weighting procedures for unit non-response in surveys in two cases:

i. non-response calibration weighting without explicitly estimating the response probabilities, also called the one-step approach, and

ii. a two-step approach with non-response propensity weighting followed by calibration.

As a summary, two basic approaches to adjusting for selection bias are presented in the calibration framework: a one-step approach and a two-step approach. In the one-step approach, calibration and reweighting are carried out in a single step, without explicitly modelling the response mechanism. The two-step approach consists of two separate steps. Firstly, the response mechanism is modelled by using the auxiliary variables available both for respondents and non-respondents in order to obtain the estimated response probabilities (propensities). Then, non-response-adjusted weights for the respondents are calculated by multiplying the initial design weights by a non-response adjustment factor, which is defined as the inverse of the estimated propensities. In the second step, the auxiliary variables available for the respondents and the known population totals of the auxiliary variables are used to calibrate the non-response-adjusted weights on the population totals of the auxiliaries. Both approaches require access to strong auxiliary variables that explain the missingness mechanism if they are to successfully adjust for selection bias (*Särndal and Lundström 2005*; *Haziza and Lesage 2016*).

Calibration approaches have been developed for the missing not at random (MNAR) situation signalling a non-ignorable selectivity pattern. Some examples or recent contributions for treating non-ignorable non-response are the following. *Kott and Chang (2010)* discuss adjusting non-ignorable non-response using calibration weighting in the prediction and quasi-randomisation approaches. *Matei and Ranalli (2015)* propose a latent modelling approach for treating non-ignorable missingness. *Kott and Liao (2017)* present a calibration weighting method for removing bias when, as with MNAR, unit non-response is a function of one or more survey variables.

Ranalli et al. (2016) present a calibration estimation approach for dual-frame surveys. *Guandalini and Tillé (2017)* propose design-based estimators that are calibrated on estimated totals from multiple surveys instead of calibrating on known population totals. *Yeager et al. (2011)* examined the accuracy of telephone and internet surveys based on probability sampling, and of internet surveys based on non-probability sampling. The accuracy of the non-probability internet surveys varied much more than the probability sample surveys. Post-stratification improved the overall accuracy of some of the non-probability sample surveys but decreased the overall accuracy of others, in the cases considered. *Lenau and Münnich (2017)* discuss approaches to compensating the selectivity of non-probability sampling, including calibration and propensity-scoring methods.

Calibration is also widely used in registers. *Wallgren and Wallgren (2014 ch. 11)*, for example, provide an overview of how reweighting is applied in administrative sources, and this could be also applied to big data sources. In that case an artificial vector of initial weights $\mathbf{w} = \mathbf{1}$ is created and calibrated to match known population totals. Another approach is repeated weighting where all register-based census tables are calibrated to have the same population totals (*Houbiers et. al. 2003*).

4.1.3. Model-assisted calibration

Just as in the case of model-free calibration, model calibration (design-based model-assisted calibration) can be used to correct selectivity in big data sources when we can find auxiliary variables which are correlated with the selectivity mechanism. However, contrary to model-free calibration where we only need population totals, in this case we need auxiliary x-variables at individual level for the entire population (or at least to a higher detail).

Model calibration was proposed by *Wu and Sitter (2001)* and is a method of searching for calibrated weights by minimising a distance measure between design weights and new weights, which satisfy certain calibration constraints on the **predicted values** of the target variable y from a model. In this approach we assume that the relationship between the target variable y and some auxiliary variables x can be described using a proper statistical model. The predicted values for all population elements can then be computed by using the estimated model and the unit-level auxiliary x-data. A benefit of model calibration is the possibility to use flexible assisting models in addition to the conventional linear model, e.g. members of the generalised linear mixed models family. Good examples are logistic models for binary and polytomous survey variables and Poisson models for count variables. Mixed models can also be used.

Consequently, when the new weights are applied to the predicted values of y variable in the sample, they reproduce the population sum of the predicted values of y variable.

It is also important that the new weights calibrated under the selected distance measure should be as close as possible to the sampling weights . In model-assisted calibration, *Wu and Sitter (2001)* assume that the relationship between y and $\mathbf{x}$ can be described by a superpopulation model through the first and second moments:

$$\begin{cases} E_U(y_i) = f(\mathbf{x}_i, \boldsymbol{\beta}) \\ D_U^2(y_i) = v_i^2\sigma^2 \end{cases}, \tag{6}$$

where $f(\mathbf{x}_i, \boldsymbol{\beta})$ is a known function of x and $\boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)'$ and σ^2 are unknown superpopulation parameters which have to be estimated, and v_i is a known function of $\mathbf{x}_i$. E_U and D_U^2 denote the expectation and variance with respect to the superpopulation model. In this approach we also assume that $(y_1, \mathbf{x}_1), \dots, (y_k, \mathbf{x}_k)$ are mutually independent.

The model described by *Wu and Sitter (2001)* is very general and includes both linear and non-linear regression models. We assume that the main goal is to estimate the total value of the variable y given by $Y = \sum_{i=1}^{N} y_i$. Moreover, we assume that auxiliary variables $x_1, \dots, x_k$ exist and that $f(\mathbf{x}_i, \boldsymbol{\beta})$ is the linking model between y and the auxiliary variables. Using data from sample s and all auxiliary variables $x_1, \dots, x_k$ we find predicted values

$$\hat{y}_i = f\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right), i = 1, \dots, N \tag{7}$$

where *N* is the number of population elements and the weighted least squares estimator of parameter β is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\boldsymbol{\Pi}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Pi}^{-1}\mathbf{y}_s, \tag{8}$$

where matrix $\boldsymbol{X_s}$ contains sample values of the covariates and $\boldsymbol{\Pi}$ is a diagonal matrix consisting of first-order inclusion probabilities π_i:

$$\boldsymbol{\Pi} = diag(\pi_1, \dots, \pi_n) = \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \pi_n \end{bmatrix}. \tag{9}$$

Taking into account the chi-square distance function, the problem of finding calibration weights (w_i) given initial weights $(d_i = 1/\pi_i)$ in the model-assisted approach can be formulated as follows:

$$\begin{cases} D(\mathbf{w}, \mathbf{d}) = \dfrac{1}{2} \sum\limits_{i \in s} \dfrac{(w_i - d_i)^2}{d_i} \rightarrow min \\ \sum\limits_{i \in s} w_i \hat{y}_i = \sum\limits_{i \in U} \hat{y}_{\mathrm{i}} = \sum\limits_{i \in U} f\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right) \\ \sum\limits_{i \in s} w_i = N \end{cases} \tag{10}$$

The problem of finding calibration weights in the model-assisted approach can be formulated equivalently as follows:

$$\begin{cases} D(\mathbf{w}, \mathbf{d}) = \dfrac{1}{2} \sum\limits_{i \in s} \dfrac{(w_i - d_i)^2}{d_i} \rightarrow min \\ \sum\limits_{i \in s} \mathrm{w_i} \mathbf{z}_i = \sum\limits_{i \in U} \mathbf{z}_i = \mathbf{z}_U \end{cases} , \tag{11}$$

where

$$\mathbf{z}_i = \left(\hat{y}_i, \mathbf{1}\right)' \tag{12}$$

and

$$\mathbf{z}_U = \left(\sum_{i \in U} \hat{y}_i, N\right)'. \tag{13}$$

The model-assisted calibration estimator for totals takes the following form:

$$\hat{Y}_{mcal} = \textstyle\sum_{i \in s} w_i \, y_i \, , \tag{14}$$

where the vector of calibration weights $\mathbf{w} = (w_1, w_2, \dots, w_n)'$ is obtained as the following minimisation problem:

$$\mathbf{w} = argmin_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}) \, , \tag{15}$$

under the constraint:

$$\textstyle\sum_{i \in s} w_i \, \mathbf{z}_i = \mathbf{z}_U \, . \tag{16}$$

The solution to the minimisation problem using the linear calibration method (see Section 4.1.3) with the chi-square distance is a vector of calibration weights $\mathbf{w} = (w_1, w_2, \dots, w_n)'$, for which

$$w_i = d_i + d_i(\mathbf{z}_U - \hat{\mathbf{z}})' \left(\sum_{i \in s} d_i \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \mathbf{z}_i \ , \tag{17}$$

where:

$$\hat{\mathbf{z}} = \left(\sum_{i \in s} d_i \, \hat{y}_i \, , \sum_{i \in s} d_i\right)' \ . \tag{18}$$

The *Wu and Sitter (2001)* approach was further studied by other researchers. For example, *Montanari and Ranalli (2005)* present a non-parametric model calibration method in survey sampling. *Park and Kim (2014)* discuss instrumental-variable calibration estimation in survey sampling. *Lehtonen and Veijanen (2012, 2016)* applied model calibration for small area estimation.

In the context of big data, *Chatterjee et al. (2016)* discuss constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. They develop regression models based on individual-level data from an 'internal' study while utilising summary-level information, such as information on parameters for reduced models, from an 'external' big data source and identify a set of constraints that link internal and external models.

The applicability of calibration (either model-free or model-assisted) for selection bias adjustments depends on the availability of auxiliary variables **x** (from register data sources, census, probability sample, etc.) that are correlated with the target variable y (from the big data source) and with the selection mechanism.

4.1.4. Propensity weighting

Propensity weighting is a method commonly used to reduce non-response bias in sample surveys. The method is based on modelling of the probability (i.e. propensity) ρ_i that a population element is included in the big data source. The propensities, when predicted with a model which accounts for the selectivity mechanism, can then be used to estimate statistics corrected for the selectivity in the big data source. To be computable, covariate information both on respondents and non-respondents is needed; this can be obtained from the big data source and/or from auxiliary data sources.

The approach whereby response propensity is accounted for by an auxiliary sample (s_{TP}) is used for self-selection web panels (*Lee, 2006*) or in propensity score matching in observational studies (*Rosenbaum and Rubin 1983*).

The propensity can be written as $\rho_i = \mathrm{P}(R_\mathrm{i} = 1 | \mathbf{X} = \boldsymbol{x}_i)$, where $R_\mathrm{i} = 1$ if unit *i* responds and zero otherwise, and $\mathbf{X}$ is the set of variables known for both respondents and non-respondents (*Rosenbaum and Rubin 1983*). To estimate the unknown propensity we thus need information on the full sample. There are several models that can be used to estimate ρ_i (e.g. *Bethlehem and Biffignandi, 2012*):

$$log(\frac{\rho_i(\mathbf{x}_i)}{1-\rho_i(\mathbf{x}_i)}) = \boldsymbol{\beta}'\mathbf{x}_i \ , \tag{19}$$

where equation (19) is a logit model, $\boldsymbol{\beta}$ is a vector of regression parameters.

Another model is a probit model

$$\Phi^{-1}(\rho_i(\mathbf{x}_i)) = \boldsymbol{\beta}'\mathbf{x}_i \ , \tag{20}$$

or in more generic terms, a generalised linear model

$$E\left(g\big(\rho_i(\mathbf{x}_i)\big)\right) = \boldsymbol{\beta}'\mathbf{x}_i \ , \tag{21}$$

where g is a link function that should be specified beforehand. Classification and regression trees (CART), given by (22) (*Breiman et al. 1984*) can also be used:

$$g(\rho_i(\mathbf{x}_i)) = f(\mathbf{x}_i) \ , \tag{22}$$

These models can also be extended by using additional variables derived from paradata, which can influence the propensity score. These variables can refer to brokers' or owners' characteristics, as well as paradata taken from big data sources. Models (19) - (22) can also be extended by using the mixed model approach.

Once the propensities have been estimated, a bias-adjusted weight for observed data can be obtained as $w_i = 1/(\pi_i\hat{\rho}_i)$, where π_i denotes the (pseudo) inclusion probability and $\hat{\rho}_i$ is the estimated propensity for element *i*. This weight can be further calibrated on additional auxiliary variables, for example by the two-step calibration approach presented in Section 4.1.2. It is important to note that successful selection bias reduction presupposes a response model with good capability to explain the response mechanism.

Lee (2006) presents an overview of the propensity score method for selection bias adjustment in volunteer panel web surveys and notes that the method may decrease bias but increase variance. The author notes that it is critical to include covariates that are highly related to the study outcomes and further, that the role of non-demographic variables did not seem critical to improving propensity score adjustment in the cases considered. *Lee and Valliant (2009)* derived a combination of propensity weighting and calibration weighting to account for the bias in volunteer panel web surveys. In the resulting two-step method, the design weights are first adjusted by propensity scores to correct for selection bias due to non-probability sampling, and the adjusted weights are then calibrated to auxiliary variable totals for the target population in order to adjust for coverage bias. *Peress (2010)* introduced the adjustment for non-ignorable non-response by using paradata as auxiliary information in a propensity weighting procedure. *Valliant and Dever (2011)* discuss the estimation of propensity adjustments for volunteer web surveys by using a randomly selected reference sample and estimating the probabilities of being a web volunteer via propensity modelling. The authors identify several options for using the estimated propensities when estimating population quantities and call for a careful analysis to justify these methods. *Franks, Airoldi and Rubin (2016)* present models for non-ignorable missing data with an approach where the joint distribution of observed data and missing data is specified through non-standard conditional distributions.

Similarly, as in the case of calibration, selection bias adjustment requires the availability of strong auxiliary variables (from register data sources, census, probability sample, etc.) which are correlated with the target variable y (from the big data source) and moreover, are correlated with the selection mechanism.

4.1.5. Pseudo-empirical likelihood

The pseudo-empirical likelihood method (*Owen 2001*) has been presented as a non-parametric data-driven method for finite population problems. Empirical likelihood is a likelihood function derived by assuming that the distribution has support only on the observed sample points.

Following notation in *Wu (2006, p. 240)*, the pseudo empirical maximum likelihood (PEL) estimator of the population mean $\overline{y} = N^{-1}\sum_{i=1}^{N} y_i$ is computed as $\hat{\overline{y}}_{PEL} = \sum_{i=1}^{n} \hat{p}_i \, y_i$ where the weights $\hat{p}_i$ are obtained by maximising in p_i the pseudo empirical log-likelihood function

$$l_{ns}(\mathbf{p}) = n^* \sum_{i \in s} d_i^* \, log(p_i) \tag{23}$$

subject to the set of constraints

$$0 < p_i < 1 \ ,$$
$$\sum_{i \in s} p_i = 1 \ ,$$
$$\sum_{i \in s} p_i \, \mathbf{x}_i = \overline{\mathbf{X}}$$

where $d_i^* = d_i / \sum_{i \in s} d_i$ are the normalised design weights, $d_i = 1/\pi_i$, s refers to the sample, n^* is the effective sample size and $\overline{\mathbf{X}}$ is the vector of population level means. The second order inclusion probabilities or their approximations are needed in the computation of n^* which can make the method untractable for practical purposes (unless the selections are independent, when $\pi_{ij} = \pi_i \pi_j$). By solving the optimisation problem using a standard Lagrange multiplier argument, it can be shown that

$$\hat{p}_i = \frac{d_i^*}{1+\boldsymbol{\lambda}'(\mathbf{x}_i - \overline{\mathbf{X}})}, i \in s \ , \tag{24}$$

where the vector-valued Lagrange multiplier is

$$g_1(\boldsymbol{\lambda}) = \sum_{i \in s} \frac{d_i^*(\mathbf{x}_i - \overline{\mathbf{X}})}{1+\boldsymbol{\lambda}'(\mathbf{x}_i - \overline{\mathbf{X}})} = 0 \ . \tag{25}$$

The major computational task here is to find the solution to $g_1(\boldsymbol{\lambda}) = 0$. For more information on extensions and review papers on the likelihood approach see *Owen (2013)*, *Rao and Wu (2010)*, *Feder and Pfeffermann (2015)* or *Wu and Lu (2016)*.

A practical problem in the pseudo empirical likelihood method is the requirement for joint inclusion probabilities and design effects, which makes the method computationally intractable. *Berger and De La Torres (2016)* present a computationally simpler empirical likelihood approach as an extension of the empirical likelihood approach to complex survey data. This method does not need the joint inclusion probabilities and design effects but requires known first-order inclusion probabilities (or their approximations, e.g. propensities), which in the case of big data sources are unknown and should be estimated. *Berger (2017)* presents an overview of empirical likelihood methods under unit non-response using auxiliary data on the population. *Luo and Pang (2017)* propose a quantile empirical-likelihood-based method for dealing with non-response bias when a missing at random (MAR) or ignorable non-response mechanism is assumed; see additional references therein. *Liu (2017)* discusses empirical likelihood for the response mean of generalised linear models with MAR responses. However, even if promising, the application of empirical likelihood methods to selection bias adjustment in big data analytics is still limited.

4.1.6. Adjusted weights

Following *Lepkowski et al. (2007, ch2)*, weights can also be adjusted to account for under-coverage and self-selection in big data sources. Estimation based on these weights would then allow unbiased estimates. This problem was considered earlier in sample surveys, in particular in telephone surveys. *Lepkowski et al. (2007 ch. 2)* provided an overall method to weight in order to account for errors in telephone surveys:

$$w_i = d_i a_i^{nr} a_i^{cov} a_i^{trim} a_i^{cal} \ , \tag{26}$$

where d_i is a base weight (probability-based weight), a_i^{nr} is an adjustment for non-response, a_i^{cov} is an adjustment for incomplete mobile/internet coverage, a_i^{trim} is an adjustment to control variation among weights and a_i^{cal} is an adjustment to calibrate the weights.

In general,

$$d_i = \pi_i^{-1}\ , \tag{27}$$

but for big data sources it might be approximated to using, for instance, a log-linear model approach (see propensity weighting). It might be also possible to create pseudo-inclusion probabilities (*Elliott 2009; Elliott and Valliant 2017*) or start with $\pi_i = 1$, as in the case of registers.

Adjustment factors are based on the estimation of propensities and calibration. For instance

$$a_i^{nr} = \hat{\rho}_i^{-1}\ , \tag{28}$$

is the probability to respond (estimated via modelling or **weighting class approach**; also applies to aggregates, see equations (25) — (28)). Following the same logic of using estimated response propensities to adjust for non-response, an estimate of the propensity for telephone coverage is needed to compute an adjustment for under-coverage in telephone samples (*Lepkowski et al. 2007, p. 48*). The adjustment to the existing weight for the i-th respondent with temporary discontinuation in the h-th cell is then computed as:

$$a_{hi}^{cov} = \hat{c}_h^{-1}\ , \tag{29}$$

where $\hat{c}_h$ is the estimated telephone coverage rate for people in households with service discontinuation. *Lepkowski et al. (2007 ch. 2)* say that this adjustment is only practical for relatively large samples, because:

- sub-national sources of data to estimate the totals of households without a telephone service or the totals of households with service interruption that currently do not have a telephone service are very limited,
- the coverage adjustment affects only households with service interruption in the last year, a small percentage of the 2.3 % of adults with an interruption of 1 day or more.

Finally, a_i^{trim} corrects for too extreme weights and a_i^{cal} adjusts weights for known population totals.

Similarly, the application of methods to big data sources requires strong auxiliary variables that are associated with the self-selection mechanism. These variables should be related to the internet, social media or mobile phone usage. Applicability might be limited by available auxiliary variables from surveys or registers and might vary between countries.

4.1.7. Two-step weighting method

Börsch-Supan et al. (2004) proposed a two-step weighting method for weighting data from online surveys that is based on an explicit behavioural model of internet access and survey participation decisions. Such models were also studied by econometricians, in particular *Heckman (1976)* and *McFadden et al. (1977)*.

This method can be used to correct selectivity in big data sources because, just like in internet surveys, in big data sources there is an electronic platform mediating the target population and the data-capture process. This method differs from the previous ones in the sense that instead of correcting initial design weights, it addresses directly the joint distribution of the variable of interest and auxiliary and paradata variables.

The procedure has the following two steps:

1. correction for internet bias (propensity to use the internet);
2. correction for participation bias (using an offline survey).

The method requires a supplementary, probability-based sample for correcting these two sources of bias. The aim of the weighting procedure is to make the online data representative — i.e. to recover from the online data a density for *y* which is proportional to the density of *y* found in the population as a whole. To this end, it is useful to describe the response behaviour in an online survey in terms of a data-generating process, that is, as conditional distributions of the various sets of variables of interest. This data-generating process represents a behavioural model of online survey participation (*Börsch-Supan et al. 2004*).

The joint distribution of the variables of interest in the population, denoted $f(y, v, x)$ can be represented by the following data-generating process that factorises $f(y, v, x)$ in three components:

$$f(x, y, v) = f_1(x) f_2(v\,|x) f_3(y\,|v, x)\ , \tag{30}$$

where x, y, v are defined as previously for propensity score weighting. The distribution of the response variables of interest, $f_3(y \mid v, x)$, is conditional on the socio-demographic characteristics of a person and of their attitudes, v and x. The distribution of attitudes is, in turn, conditional on socio-demographic characteristics — this is captured by $f_2(v\,|x)$. Finally, the distribution of the pre-determined characteristics, $f_1(x)$, is unconditional.

This is the formal reason why weights need to be based on a representative offline survey — while $f_1(x)$ is known from official sources such as census data, and while $f_2(v|x)$ might be recovered from other surveys, $f_3(y \mid v, x)$ is generally not available from other sources.

The data-generating process of the online sample diverges from the density in the population along two dimensions: (i) internet access and (ii) the participation decision. These dimensions are reflected by two additional densities $f_5(y, v, x)$ and $f_6(y, v, x)$, respectively.

These densities represent the probabilities of participation at each stage of the participation process, stratified by the *y, v,* and *x* variables. Accordingly, the conditional probability distribution of online responses given internet access and participation is

$$f(x, y, v \mid online) = \frac{f_1(x) f_2(v\,|x) f_3(y\,|v,x) f_5(y,v,x) f_6(y,v,x)}{\sum_x \sum_v \sum_y f_1(x) f_2(v\,|x) f_3(y\,|v,x) f_5(y,v,x) f_6(y,v,x)}\ . \tag{31}$$

If the distribution f_5 and f_6 can be obtained, then weighting can be used to correct the data and determine the distribution of interest.

The two-step participation model therefore results in a two-step procedure to determine the weights. However, note that the second step of the weighting process that corrects for the participation decision is less straightforward to implement. In fact, in contrast to the first step where both individuals with and without internet access could be observed in the offline sample, the second-stage counterfactual (i.e. self-selected non-participants) are neither observed in the online nor in the offline samples. Therefore, one cannot construct a convenient indicator binary participation variable and estimate the corresponding participation probabilities directly. Rather, weights need to be constructed indirectly based on the observed values of the (*y, z, x*) variables in the online and offline samples. For the weighting to be successful the online sample must have reached all strata of the population being investigated. (*Börsch-Supan et al. 2004*).

Börsch-Supan et al. (2004) applied this approach to Perspektive Deutschland's online survey and a small traditional CAPI survey that used a random sample of the adult population. In the first stage they used a probit model, calibrated on the offline sample, to predict the *ex ante* probability of using the internet for every observation in the online sample. The explanatory variables used were socio-demographic variables that are known to influence internet access, as well as psychographic variables (i.e. *x* and *v* variables, respectively).

In the second stage, *Börsch-Supan et al. (2004)* applied iterative proportional fitting (raking) to adjust marginal response distributions in the online survey to those in the offline survey. The core set of raking variables contains basic socio-demographics and psychographic variables. The social-political focus of the questionnaire meant it focused on psychographic variables that reflect social responsibility, the willingness to engage oneself, and risk attitude. Variables were derived from the

CAPI survey and from the German microcensus. When the raking process was completed, the online survey was weighted using the combination of first-stage internet access weights and second-stage participation weights.

4.1.8. Software

Weighting procedures are available in several statistical packages. Calibration can be computed using many SAS, SPSS and R programs developed in statistical agencies and elsewhere, for example Statistics Canada, INSEE, and ISTAT. Examples are SAS macro GES of Statistics Canada, Calmar SAS Macro developed by INSEE, SPSS Module g-CALIB-S of Statistics Belgium and `ReGenesees` developed in R by ISTAT. Other examples of R packages are `survey` or `laeken`.

Empirical likelihood methods could be implemented in any programming language. Several R packages are already available, namely, `emplik` or `emplik2` that contain empirical likelihood ratio tests for means/quantiles/hazards from possibly censored and/or truncated data. The R package `glmc` fits generalised linear models where the parameters are subject to linear constraints. There is also a paper by *Wu (2005)* who provided R/S-Plus scripts for the pseudo-empirical likelihood method in survey sampling.

Propensity score weighting can be calculated using any software because it relies on modelling binomial distribution. Currently, the GWSM is not implemented, but it is not demanding and can be implemented in any software, namely SAS `PROC IML`, the `Matrix` package for sparse matrices and `Armadillo` a C++ library for linear algebra (see also `RcppArmadillo` in R). In case of very large datasets, one can consider using `Spark` and its `MLlib` package (http://spark.apache.org/docs/latest/mllib-guide.html).

4.2. Modelling approach

The modelling approach to account for self-selection in non-probability surveys has been widely discussed in the literature on sample surveys. In order to use this approach to correct selectivity the following steps should be taken:

- create careful model specifications — here we assume that models built on a big data sample hold for population. This model should be as assumption-free as possible (e.g. algorithmic approach);
- compile powerful auxiliary data that contain information to account for the self-selection error — without such variables it might not be possible to reduce bias. We assume that the variables used in the model will be able to account for self-selection.

Further in the section we will present, though without being exhaustive, possible approaches that are currently studied in the literature. Both ignorable and non-ignorable missingness are discussed.

4.2.1. Heckman selection models

Much has recently been written on handling ignorable (MAR) and non-ignorable (MNAR) missingness in selection models. This builds on research by *Heckman (1979)*, *McFadden et al. (1977)* and *Diggle and Kenward (1994)*, and many others. Methodology developments and applications are often in observational and longitudinal surveys in economics and medical sciences. This section contains a brief literature review on a selection of proposed methods potentially relevant to selection bias assessment and adjustment.

In econometric literature, a strict distinction is drawn between ignorable and non-ignorable missingness. For example, *Verbeek and Nijman (1996)* present econometric methods for the treatment of selection bias in longitudinal and panel surveys. In addressing ignorable and non-ignorable missingness, the authors state that for the non-ignorable response mechanism both imputation and weighting strategies require a model-based approach in which the selection rule is specified and estimated. Therefore they concentrate on model-based approaches in which both the observed and the missing data are modelled.

Gad (2011) discusses a selection model for longitudinal data with non-ignorable missing values. The paper presents a model for continuous longitudinal data with non-ignorable non-monotone missing values and assumes two separate models: a multivariate normal model for the study variable of interest and a binomial model for the missingness mechanism. Parameters in the adopted model are estimated using the stochastic expectation-maximisation algorithm. The approach is illustrated with an empirical study. In the same way as many other selection model approaches, Gad's approach resembles the two-step weighting methods for adjusting for non-ignorable non-response discussed in Section 4.1.2.

Wang, Bartlett and Ryan (2017) proposed a Bayesian selection model for correcting for non-ignorable non-response bias in a logistic modelling. The authors present a strategy for modelling non-ignorable missingness where the probability of non-response depends on the outcome. The bias in regression estimates is quantified for a logistic regression model. The non-identifiability of the observed likelihood is shown using a non-ignorable missing data mechanism. A Bayesian framework for model estimation is proposed as a flexible approach for incorporating different missing data assumptions and conducting a sensitivity analysis.

Many of these and other methodological developments in the area have potential to help treat selection bias in big data sources.

4.2.2. Small area estimation approach

The small area estimation (SAE) methodology has been developed to produce reliable estimates of different characteristics of interest, such as means, count, quantiles or ratios for domains for which only small samples are available, see *Rao and Molina (2015)*.

SAE uses statistical models to link survey outcome or response variables to a set of auxiliary variables known for small areas to predict small area-level target variables. This combines the model prediction and the standard direct estimate for each area in a sensible way, **balancing bias and precision**.

The simpler models are based on linear regression but more sophisticated (parametric and non-parametric) models have been proposed to deal with non-linearity and complex sampling designs. See *Rao and Molina (2015)* for an overview.

The SAE methodology is used by different national statistical institutes in different areas, in particular to estimate quantities related to the labour market, agriculture or business statistics. It is also useful in mapping poverty. For instance, the World Bank has used the SAE methodology to prepare poverty maps for more than 60 countries all over the world.

Marchetti et al. (2015) propose three ways to use big data sources together with SAE techniques, and show how big data has the potential to mirror aspects of well-being and other socioeconomic phenomena. *Marchetti, Giusti and Pratesi (2015)* demonstrate the use of big data from Twitter for small area estimation of households' share of food consumption expenditure in Italy.

4.2.3. Bayesian approach

The Bayesian approach in survey inference (*Ericson 1969, 1988*; *Basu 1971*; *Scott 1977*; *Binder 1982*; *Rubin 1983, 1987*; *Ghosh and Meeden 1997*; *Sedransk 2008*; *Little 2003, 2004*; *Fienberg, 2011*) requires the specification of a prior distribution for the population units. A complete Bayesian approach would also require modelling the observation/inclusion mechanism, often under an ignorable (MAR) missingness pattern. Extensions to non-ignorable missingness also have been proposed.

Tam and Clarke (2015) present a Bayesian framework to assess the conditions under which valid statistical inference can be drawn from big data, and provide a Bayesian method for using big data sources to produce official statistics. The proposed Bayesian framework for making big data inferences is based on conceptualised transformation, sampling and censoring processes that are applied to the measurements from the big data sources. Proper inference will require modelling all three processes. The authors note that the required modelling exercise can be complex. However, if certain sampling and censoring ignorability conditions are fulfilled, inference can be made on the big data measurements as if they were generated from a random sample.

4.2.3.1. HIERARCHICAL BAYESIAN APPROACH

The inclusion of a selection mechanism in the Bayesian approach has been always challenging, as modelling this mechanism is highly complicated. *Gelman (2007)* presents a model-based Bayesian approach to model the target variable in a first phase and suggests reweighting estimates using post-stratification in a second phase. *Gelman (2007)* proposes that models should contain all variables that were used for sample selection (that were used to create weights) and other variables that may influence the modelled outcome, representing a standard model-based practice to account for the sampling complexities. *Brick (2013)* maintains that powerful auxiliary variables are required in order to reduce bias in sample surveys that suffer from non-response. Further, *Brick (2013)* argues that weighting procedures provide consistent estimates and are easier to apply than model-based approaches.

Wang et al. (2015) applied a Bayesian hierarchical model with post-stratification to a non-representative poll of Xbox users. We will focus on a detailed description of this paper to provide insights into how such models could be built for big data sources.

During the 2012 US presidential campaign, Wang et al. conducted 750 148 interviews with 345 858 unique respondents on the Xbox gaming platform during the 45 days preceding the election. Xbox Live subscribers were asked to provide baseline information about themselves in a registration survey, including demographics, party identification, and ideological self-placement. Each day, a new survey was offered and respondents could choose whether they wished to complete it. The final sample size was equal to 83 283 users who responded at least once prior to the first presidential debate on 3 October. In total, these respondents completed 336 805 interviews, or an average of about four interviews per respondent.

Wang et al. (2015) considered all possible combinations of sex (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3 categories) and 2008 vote (3 categories) in the model. They considered the following models:

- The first models predict whether a respondent supports a major-party candidate

$$\mathrm{P}(y_i \in \{Obama, Romney\}) = logit^{-1}\big(\alpha_0 + \alpha_1(state\ last\ vote\ share) + a_{j[i]}^{state} + a_{j[i]}^{edu} + a_{j[i]}^{sex} + a_{j[i]}^{age} + a_{j[i]}^{race} + a_{j[i]}^{partyID} + a_{j[i]}^{ideology} + a_{j[i]}^{lastVote}\big), \tag{32}$$

where α_0 is the fixed baseline intercept and α_1 is the fixed slope for Obama's fraction of the two-party vote share in the respondent's state in the last presidential election. The terms $a_{j[i]}^{state}$, $a_{j[i]}^{edu}$,$a_{j[i]}^{sex}$ and so on correspond to the varying coefficients associated with each categorical variable. Here, the subscript $j[i]$ indicates the cell to which the i-th respondent belongs. The varying coefficients are given by independent prior distributions

$$a_{j[i]}^{var} \sim N(0, \sigma_{var}^2), \tag{33}$$

where $var \in \{\text{state}, \text{edu}, \dots, \text{party ID}\}$ is used for simplicity. To complete the full Bayesian specification, the variance parameters are assigned a hyperprior distribution $\sigma_{var}^2 \sim inv\text{-}\chi^2(\nu, \sigma_0^2)$ with a weak prior specification for the remaining parameters, ν and σ_0^2.

- The second model predicts support for Obama, given that the respondent supports a major-party candidate

$$\begin{aligned}P(\,y_i = Obama \mid y_i \in \{Obama, Romney\}\,) \\ = logit^{-1}\Big(\beta_0 + \beta_1(state\ last\ vote\ share) + b_{j[i]}^{state} + b_{j[i]}^{edu} + b_{j[i]}^{sex} \\ + b_{j[i]}^{age} + b_{j[i]}^{race} + b_{j[i]}^{partyID} + b_{j[i]}^{ideology} + b_{j[i]}^{lastVote}\Big)\end{aligned} \tag{34}$$

and

$$\begin{aligned}&b_{j[i]}^{var} \sim N(0, \eta_{var}^2),\\ &\eta_{var}^2 \sim inv\text{-}\chi^2(\mu, \eta_0^2).\end{aligned} \tag{35}$$

Finally, after the modelling process they applied a post-stratification using exit poll data from the 2008 presidential election instead, due to the lack of auxiliary variables on party ID or ideology. In total, 101 638 respondents were surveyed in state and national exit polls. The estimator is given by:

$$\hat{y} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}, \tag{36}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population.

Their results indicate small differences in the share of Obama voters estimated by the proposed approach based on Xbox data on the day before the election and estimates obtained from the 2012 national exit poll. With insufficient demographic information on respondents, inadequate population-level statistics, or a lack of historical election poll results, it would have been difficult to generate accurate forecasts from non-representative data.

Reilly, Gelman and Katz (2001) studied the situation when population totals are unknown, but a proxy variable under a dynamic model was used. They noted that one of the greatest practical limitations to the use of post-stratification is the need to know the proportion of the population in each stratum. We have population-level information only for certain variables, so it appears that post-stratification is useful only if our quantity of interest is related to one of a handful of characteristics for which we have population-level information. *Reilly et al. (2001)* proposed a dynamic (space-state) model for the variable by which we post-stratify, thereby estimating the strata weights from the sample.

4.2.3.2. CALIBRATED BAYES APPROACH

Little (2012) proposed an alternative philosophy for survey inference under the name calibrated Bayes (CB), where inferences for a particular data set are Bayesian, but models are chosen to yield inferences that have good design-based properties.

Little (2015) stressed the ‘status quo’ in current official statistics, which might be termed the ‘design/model compromise’. This favours design-based inference for descriptive statistics such as means and totals based on large probability samples, and model-based inference for questions that are not well addressed by the design-based approach, such as small area estimation, survey non-response and response errors.

Little (2012) argued that CB resolves the design/model compromise conflict and capitalises on the strengths of both frequentist and Bayesian approaches. Features of the CB approach to surveys include incorporating survey design information into the model, and models with weak prior distributions that avoid strong parametric assumptions.

Features of CB models for surveys are that:

a. relatively weak prior distributions should be favoured so that the evidence in the data dominates the evidence in the prior; and

b. model checks become an important feature of the analysis *(Little, 2012)*.

Little (2012) argues that the best approach is to take the strengths of both paradigms, that is:

- Bayesian statistics is strong for inference under an assumed model, but is relatively weak for developing and assessing models.
- Frequentist statistics provide a useful tool for developing and assessing models, but are a weak tool for inference under an assumed model.

Hierarchical Bayes models yield estimates close to 'direct' estimates when sample sizes are large, and as the sample size decreases, move seamlessly towards predictions from a fixed-effect model. Thus, for a typical hierarchical Bayes model, the posterior mean of the population $\bar{y}_d$ in domain d, given covariate information $\mathbf{X}$ has the form

$$E(\bar{y}_d|data) = \gamma_d\bar{y}_d + (1-\gamma_d)(\bar{y}_d + (\bar{\mathbf{x}}_d - \bar{\mathbf{X}})'\hat{\boldsymbol{\beta}}), \tag{37}$$

where $(\bar{y}_d + (\bar{\mathbf{x}}_d - \bar{\mathbf{X}})'\hat{\boldsymbol{\beta}})$ is the regression prediction for the mean of y aggregated over all areas, and $\gamma_d = n_d\sigma^2/(n_d\sigma^2 + \tau^2)$, where $\bar{y}_d, \bar{\mathbf{x}}_d, n_d$, are the sample means of y and $\mathbf{x}$ and sample size in area d,assigns most of the weight to the sample mean when n_d is large, and most of the weight to the regression prediction over all areas when $n_d$is small.

Little (2015) discussed using CB in the context of big data; 'With non-probability samples, issues of selection bias are not solved by the CB paradigm, but CB provides a useful theoretical framework for addressing them. Thus one strategy is to incorporate auxiliary population information into the model, using post-stratification or other approaches to reduce the selection bias explained by these factors'.

Little (2011) discusses propensity modelling in the context of the CB approach for the treatment of missingness.

4.2.3.3. PATTERN MIXTURE MODELS — MNAR CASE

The pattern mixture model (PMM) was introduced by *Little (1993)* in order to deal with MNAR case (or non-ignorable non-response). This is different to the selection mechanism model. The selection mechanism model accounts for selectivity by factorising the unobserved full data between the observed data and the missingness mechanism,

$$P(Y, R; \psi, \theta) = P(Y; \theta)P(R|Y; \psi), \tag{38}$$

where Y is the variable of interest for the full data, R is the response (or missingness) variable, θ is a parameter of the probability density of $Y$and ψ is a parameter of the probability density of R.

In contrast, the PMM factorises the unobserved full data between the density of the full data for each missingness pattern and the probability (i.e. weight) of that missingness pattern,

$$P(Y, R; \psi, \theta) = P(R; \phi)P(Y|R; \theta). \tag{39}$$

The PMM, including auxiliary variables, can be written as follows,

$$P(Y, R, X; \psi, \theta) = P(R, X; \phi)P(Y|R, X; \theta) \tag{40}$$

where the parameters ψ and θ are a priori independent. $P(Y|X, R = 1, \theta_1)$ is the distribution of observed data, $P(Y|X, R = 0, \theta)$ is the distribution of the missing data and $P(R, \phi)$ the marginal response probability.

The general procedure for using the PMM is as follows:

- Draw θ_1^* from its posterior distribution using $P(Y|X, R = 1, \theta_1)$;
- Specify the posterior $P(\theta|\theta_1)$ a priori (for instance, $\theta_0 = \theta_1 + k$ where k is a fixed constant);

- Draw $\theta_0^* \ P(\theta_0|\theta_1)$;
- Impute missing data from $P(Y|X, R = 0, \theta_0^*)$.

The PMM requires making untestable assumptions to help identify the model. The model can reduce or eliminate bias only under the condition that the additional assumptions are valid. On the other hand, when the assumptions are not valid the PMM can produce bias even for MAR data. The PMM is an ideal tool when sensitivity analysis is required (*Enders, 2010, p. 300*).

'(…) The pattern mixture model's reliance on untestable assumptions is its weakness. Although the model does not rely explicitly on distributional assumptions, it does require the user to specify values for the inestimable parameters. To the extent that these assumed values are correct, the model can reduce or eliminate bias. However, specifying the wrong values can produce substantial bias, even when the data are MAR (*Demirtas and Schafer, 2003*). (…) However, some methodologists argue that this is actually an advantage of the model because it forces researchers to make their assumptions explicit. (…) However, the possibility of using different approaches to generate values for the inestimable parameters makes the pattern mixture model an ideal tool for sensitivity analyses because you can examine the stability of the resulting estimates across a variety of scenarios.' (*Enders, 2010, p. 300*).

4.2.4. Machine-learning approach

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. These algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions (*Friedman, Hastie and Tibshirani, 2001*). A comprehensive overview of machine-learning techniques can be found in *Breiman et al. (1984)*, *Breiman (2001)* or *Hastie, Tibshirani and Friedman (2016)*.

What distinguishes many machine-learning models from the traditional statistical models is that they are based on algorithms. This makes it possible to specify highly non-linear relationships between the variable of interest Y and auxiliary variables X which can capture the selectivity mechanism.

Buelens et al. (2015) perform a simulation study to compare methods in removing bias. They conclude that 'the algorithmic approaches are more flexible than the pseudo-design-based and traditional model-based methods and should be considered in real world settings, in particular when the relations between auxiliary and target variables are complex and non-linear.'

Here we present briefly some of the machine learning algorithms that can be used to adjust for self-selection bias.

4.2.4.1. K-NEAREST NEIGHBOURS

The k-nearest neighbour method (*Hastie et al. 2016*) is a non-parametric method that predicts the unknown value of the variable of interest Y of the unobserved units by averaging the k nearest observed values. Typically, the distance measure taken is the (weighted) Euclidean distance for a set of auxiliary variables $\mathbf{X}$. The method requires the choice of and appropriate distance function, a set of covariates $\mathbf{X}$ and the number of neighbours k (*Buelens, et al., 2015*).

4.2.4.2. ARTIFICIAL NEURAL NETWORKS

An artificial neural network is an algorithmic method mimicking the architecture of the human brain in a simplified manner. The method creates predictions by propagating inputs, i.e. predictors, throughout a network of artificial neurons (nodes) laid out in layers, where multiple hidden layers are possible. Each node in the network applies a multiplicative weight to all its inputs and an 'activation function' to the weighted sum. The output of a node in one layer serves as the input to the nodes in the next layer. This is known as a feed-forward neural network (*Buelens et al., 2015*). Learning or training (fitting a neural network) consists of determining the weights of the neurons which minimise the error between the predicted values and the true values in the training data.

The algorithm itself also has tuning parameters (i.e. hyper-parameters), namely the number of hidden layers, their size and the choice of activation function.

4.2.4.3. CLASSIFICATION AND REGRESSION TREES

Classification and regression trees (CART), discussed in a summary paper by *Loh (2014)*, do not assume a linear relationship between the target variable and predictors. The main idea is to create an **optimal** (according to some criterion) split of the $\mathbf{x}$ space into subsets. A binary tree splits the data into two subsets at each node, to maximise the variance between the two groups. Starting from a single root node, a data set is split into two branches of the tree which in turn split the data again. Some stopping criterion is applied to decide whether data at a particular node get split — if not, the node is said to be a leaf node.

Here we report some specific extensions of trees based on *Loh (2014)*:

- **Classification rule with unbiased interaction selection and estimation (CRUISE)** — splits each node into multiple children nodes, with their number depending on the number of distinct Y values.
- **Bagging** — uses an ensemble of unpruned CART trees constructed from bootstrap samples of the data.
- **Random forest** — weakens the dependence among the CART trees by using a random subset of X variables for split selection at each node of a tree.
- **Boosting** — sequentially constructs the classifiers in the ensemble by putting more weight on the observations misclassified in the previous step.
- **Bayesian model averaging** — prior distributions are placed on the set of tree models and stochastic search is used to find the good ones.
- **Model-based recursive partitioning** — fits least-squares, logistic and other models using the score functions of M-estimators. Special cases include standard maximum and pseudo-likelihood models. It achieves unbiased variable selection by choosing split variables on the basis of structural break tests for the score function. The algorithm is not unbiased if some X variables are used for both fitting and splitting.

CART and clustered data

Hajjem, Bellavance and Larocque (2011) propose a regression tree that takes into account the clustering of units. The basic idea behind the proposed mixed effects regression tree is to dissociate the fixed effects from the random ones. *Hajjem et al. (2011)* used a standard regression tree to model the fixed effects and a node-invariant linear structure to model the random effects. The method is implemented using a standard tree algorithm within the framework of the expectation-maximisation (EM) algorithm.

The proposed model is as follows:

$$\begin{aligned} &y_i = f(x_i) + Z_i u_i + \epsilon_i \\ &u_i \sim N(0, D), \\ &\epsilon_i \sim N(0, R_i). \end{aligned} \tag{41}$$

The MERT (Minimum Error Rate Training) algorithm is the ML-based EM algorithm in which we replace the linear structure used to estimate the fixed part of the model by a standard tree structure. The algorithm is as follows:

Step 0: Set $r = 0$, let $\hat{u}_{i(0)} = 0$, $\hat{\sigma}_{(0)} = 1$ and $D_{(0)} = I_q$

Step 1: Set $r = r + 1$, update $y^*_{i(r)}, \hat{f}\,(x_i)_{(r)}$ and $\hat{u}_{i(r)}$

$$y^*_{i(r)} = y_i - Z_i\,\hat{u}_{i(r-1)}\,, i = 1, \dots, n,$$

Let $\hat{f}\,(x_i)_{(r)}$ be an estimate of $f(x_i)$ obtained from a standard tree algorithm with $y^*_{i(r)}$ as responses and $X_i, i = 1, \dots, n$, as covariates. Note that the tree is built as usual using all N individual observations as inputs along with their covariate vectors,

$\hat{u}_{i(r)} = \hat{D}_{(r-1)}\, Z_i'(y_i - \hat{f}\,(x_i)_{(r)}), i = 1, \dots, n$, where $\hat{V}_{i(r-1)} = Z_i\, \hat{D}_{(r-1)}\, Z_i' + \sigma^2_{(r-1)} I_{n_i}$

Step 2: update $\hat{\sigma}^2_{(\mathrm{r})}$ and $\hat{D}_{(\mathrm{r})}$ using:

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \left\{ \hat{\epsilon}'_{i(r)}\, \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)}\, [n_i - \hat{\sigma}^2_{(r-1)}\, \mathrm{trace}(\hat{V}_{i(r-1)})] \right\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \left\{ \hat{u}_{i(r)}\, \hat{u}'_{i(r)} + \hat{D}_{(r-1)}\, [\hat{D}_{(r-1)} - \hat{D}_{(r-1)}\, Z_i'\, \hat{V}^{-1}_{i(r-1)}\, Z_i\, \hat{D}_{(r-1)})] \right\}$$

where $\hat{\epsilon}_{i(r)} = y_i - \hat{f}\,(x_i)_{(r)} - Z_i\, \hat{u}_{i(r)}$.

Step 3: Repeat steps 1 and 2 until convergence, monitored by computing, at each iteration, the following generalised log-likelihood (GLL) criterion:

$$\mathrm{GLL}(f, u_i \mid y) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \{(y_i - f(x_i) - Z_i u_i)' R_i^{-1} (y_i - f(x_i) - Z_i u_i) + u_i' D^{-1} u_i + \log|D| + \log|R_i|\},$$

where $R_i = \sigma^2 I_{n_i}$.

Another approach was introduced in *Sela and Simonoff (2012)*. This approach considered the case of longitudinal and clustered data. The algorithm is similar to MERT, but assumes a more complex covariance structure for random effects. Most recently, *Hajjem, Bellavance and Larocque (2014)* extended their earlier proposal by replacing CART by Random Forest algorithm.

Wang et al. (2012) studied measurement errors in the online social networks classification, which might be useful if we want to detect clusters accounting for correlated errors within clusters. Other methods that are used in machine learning can be found in *Lazer et. al. (2009)*.

4.2.5. Software

Currently neither the M-quantile nor the model-based calibration are available in the most common statistical software environments. The model-based calibration could be easily implemented (for example in R) while the M-quantile requires more attention. There are some R functions for M-quantile models that include Gaussian, Binomial and Poisson distributions.

For the Bayesian modelling approach there are several software options, including Bugs, WinBugs, `Stan` (http://mc-stan.org), `R-INLA` (a powerful tool using integrated nested Laplace approximations, http://www.r-inla.org) or several R packages, namely `MCMCglmm` or `arm`. SAS also enables MCMC estimation via PROC MCMC. These models are computationally demanding and both require a high number of CPUs and RAM, but are fairly easy to parallelise.

Machine learning methods are widely available in several languages such as R and `Python` (e.g. `mlr`, `caret`, `sci-kit`). Software such as `scala`, `Spark` or `h2o.io` has also been developed to implement machine learning dedicated to big data. Nonetheless, some of the CART are only available as compiled programs that require some additional API to make them more user-friendly (e.g. GUIDE — http://pages.stat.wisc.edu/~loh/guide.html).

4.3. Data linking approach

In a data linking approach to self-selection bias adjustment, data from a big data source are linked at individual level with data from a frame population or a representative sample. After the data have been linked, unbiased estimations for the target population can be made, either via a reweighting approach such as post-stratification, or a modelling approach such as prediction of the target variable. This normally involves using auxiliary variables which account for the selectivity mechanism.

Basically, there are two possible data linking methods: probabilistic record linkage, and statistical matching. The first method assumes that we link two or more records that refer to the same statistical unit (for example person, enterprise, entity) while the latter assumes that we link records that are similar, given a distance function, but that don't necessarily refer to the same statistical unit.

4.3.1. Record linkage

In record linkage, the records in the two files are compared with one another using available information, which typically does not include unique, error-free personal codes. Individuals can be compared on surname, first name, age or date of birth, and other variables. Some of these matching variables carry a lot of information for identifying individuals, whereas others (e.g. age or sex) contain very little. However, some comparisons are useful for discriminating between certain people, such as individuals living in the same household. Information can be missing or recorded with typographical or spelling errors (*Lahiri and Larsen, 2005*).

Record linkage might be a required step to complement target units with auxiliary variables to be used in models or for weighting. *Daas et al. (2016)* examine different ways of extracting and linking auxiliary information with big data sources and apply a selected method to Dutch Twitter data.

4.3.2. Sample matching

Recently, *Lavallée (2015)* discuss the use of generalised weight share method in the context of big data and opt-in surveys. *Lavallée (2015)* discusses the sample matching proposed by *Rivers (2007)* which consists of selecting a probabilistic sample from a sampling frame, and linking this sample to a web panel of respondents using statistical matching. *Rivers (2007)* notes that if the pool of respondents on the web panel is sufficiently large and diverse, the matched sample is guaranteed to have approximately the same joint distribution of the auxiliary variables as a probability sample s. In addition to web panels, sample matching can also be applied to a dataset from a big data source.

Sample matching is similar to nearest neighbour imputation and is highly dependent on the modelling assumptions underlying the matching of sample s to units on the dataset from the big data source.

The basic setting for sample matching is as follows:

- We want to estimate the total $y = \sum_{i=1}^{N} y_i$ of a population U, based on Horvitz-Thompson estimator $\hat{y} = \sum_{i=1}^{n} y_i w_i$.
- However, the variable *y* is not measured in a probability survey.
- Therefore, we use the variable y_k for unit k observed in the big data source that is 'closest' to unit i in U.
- Unit k (the 'closest' to unit i) is determined by distance function based on $\mathbf{x}$. In the event of a tie, a value can always be selected at random from the identified indices k.
- Then, the matching estimator is given by:

$$\tilde{y} = \sum_{i=1}^{n} \frac{y_i^*}{\pi_i}, \qquad (42)$$

where n is the sample size of the probability sample and y_i^* refers to y variable observed in the big data source.

The main technical conditions required for sample matching to work are (*Rivers, 2007*):

- the observations $(y_i, \mathbf{x}_i, R_i)$ are i.i.d, where y_i is the target variable,

- the missingness variable R is independent of the vector $(y, \mathbf{x})$, which corresponds to the missing completely at random (MCAR) mechanism.

The MCAR condition is extremely strong, but may be replaced with the following weaker conditions:

- the selection of the panel is ignorable: the indicator variable R is independent of y, given the auxiliary variables $\mathbf{x}$;
- the big data set covers all the relevant portions of the population U.

The condition of having relevant proportions of the target population is critical for sample matching to yield results which are comparable to a conventional sample.

4.3.3. Software

Record linkage could be considered as a special classification problem. Hence, machine learning techniques could be applied. There are several software devoted to record linkage, for instance `RecordLinkage` or `Relais` package in R which handles fairly large data sets. Statistical matching or sample matching requires both calculating distance functions, and imputation (cf. `StatMatch` package).

5 Domain-level methods to correct selectivity

The domain-level approach refers to models that assume data aggregated at a certain level. The term ‘domain’ does not only mean spatial aggregation (e.g. LAU 1), but also cross-classifications (e.g. cross-classifications by sex and labour status).

This level might be the result of either the transformation of object-level /unit-level data or be created by the data holder. In the second case, statisticians do not have control over the data cleaning process, which propagates errors (in the form of over-coverage and duplicated records). However, NSIs should establish collaboration with the data holder to agree in advance on the appropriate data aggregation.

Similar to the case of unit-level data, we distinguish two approaches: (i) reweighting; and (ii) modelling. These can also be combined. For the modelling part we might assume two cases: (i) unbiased estimates at domain level are available from existing data sources; and (ii) such data are not available. Instead a marginal distribution of certain characteristics is present.

5.1. Pseudo-design methods — reweighting

In general, the reweighting at domain level approach depends on the data available. Here we distinguish two cases:

1. Estimates of θ_d based on a big data source, denoted by $\breve{\theta}_d$, for domain $d = 1, \ldots, D$ are available without any information on cross-classifications;
2. Estimates of θ_{cd} based on big data source, denoted by $\breve{\theta}_{cd}$, for domain $\mathrm{d} = 1, \ldots, \mathrm{D}$ with additional cross-classifications denoted by c, which could be sex, age or other characteristics that could be derived from a big data source. This also includes the case where data are available in time, in which case we add the subscript t, that is $\breve{\theta}_{cdt}$.

In the first case, we can consider the following correction for $\breve{\theta}_d$, which can be written as a linear estimator:

$$\breve{\theta}_d^{adj} = \breve{\theta}_d \times a^{cover} \times a^{active} \times a^{share}, \tag{43}$$

Where:

a^{cover} refers to an adjustment for the coverage of the big data source,

a^{active} refers to an adjustment for the fraction of the active users’ population (e.g. people who are using the electronic platform underlying the big data source),

and, finally, a third adjustment denoted by a^{share} referring to the penetration rate of the particular electronic platform operator supplying the data (in case where not all operators provide data).

These adjustments could also be done per domain d with a_d^{share}. This is because using an overall adjustment based on the whole target population might be highly biased. For instance, such a situation occurs for simple synthetic small area estimators such as broad area ratio estimator (BARE) or composite estimators (*Rao and Molina, 2015*). If domain-level estimates are available, the linear estimator is given by:

$$\breve{\theta}_d^{adj} = \breve{\theta}_d \times a_d^{cover} \times a_d^{active} \times a_d^{share}. \tag{44}$$

For the second case, when cross-classification data such as sex or age groups are available, the linear estimator is extended by an additional adjustment accounting for known population totals (a_{cd}^{cal}),

$$\breve{\theta}_{cd}^{adj} = \breve{\theta}_{cd} \times a_d^{cover} \times a_d^{active} \times a_d^{share} \times a_{cd}^{cal}. \tag{45}$$

Often, $a_d^{cover}, a_d^{active}, a_d^{share}$ are unknown and should be estimated. We assume that a_{cd}^{cal} is known from registers, but it could also need to be estimated based on sample survey. The linear estimator then has the following form:

$$\breve{\theta}_d^{adj} = \breve{\theta}_{cd} \times \hat{a}_d^{cover} \times \hat{a}_d^{active} \times \hat{a}_d^{share} \times a_{cd}^{cal}. \tag{46}$$

Finally, we could compare the distribution of $\breve{\theta}_{cd}^{adj}$, $\breve{\theta}_{cd}$ and true values (if available) from a register, denoted by $\tilde{\theta}_{cd}$ or a representative sample, denoted by $\hat{\theta}_{cd}$, to see the impact of the adjustments on estimates.

If bias is still present, this might provide information about two issues:

1. An important auxiliary variable that could adjust for bias is missing.
2. The linear form of adjustment is wrong.
3. The missing data mechanism could be missing not at random (MNAR) and the correction using adjustment factors might not remove bias.

To investigate the MNAR case for binary data we could use the approach discussed by *Zhang (1999)* and *Zhang et al. (2013)*.

Reweighting at domain level has the following drawbacks, in particular for small areas:

- coverage of the electronic platform underlying the big data source for each domain might not be available and would need to be estimated;
- direct estimates based on sample surveys might provide unbiased estimates, but with high variance — in this case small area estimation models (e.g. synthetic estimates) could be considered;
- the penetration rate of the electronic platform operator providing the data (when not all operators are considered) might be unknown and a naive approach such as national average could be used.

5.2. Modelling approach

Modelling domain-level data that suffers from selectivity bias is challenging. A simple modelling process that only takes into account the big data will not remove or reduce this bias. The use of additional data is unavoidable, in particular estimates of the target variable or of a proxy of it.

Therefore, in this report we present methods that could be used to remove bias but assume that an external data source that provides 'gold standard' estimates is available. For the modelling process we consider that either direct estimates or adjusted estimates based on the big data are used. We recommend the prior use of reweighting methods to remove bias that is connected with coverage error before the modelling process.

5.2.1. Direct estimation of bias

Following *Fosen and Zhang (2011)*, let us assume that we are interested in estimating bias of characteristic $\breve{\theta}$ that was estimated based on big data. Two cases should be considered. The first one is given by equation (47) and refers to the case when the true value θ is estimated based on a sample survey denoted by $\hat{\theta}$:

$$Bias(\breve{\theta}) = E(\breve{\theta}) - E(\hat{\theta}), \tag{47}$$

and for this case $MSE(\breve{\theta})$ is given by:

$$MSE(\breve{\theta}) = V(\breve{\theta}) + \left(E(\breve{\theta}) - E(\hat{\theta})\right)^2. \tag{48}$$

The second case is given by equation (49) and refers to the case when the true value θ is estimated using register data $\tilde{\theta}$ and this estimator is assumed to be unbiased or adjusted for bias:

$$Bias(\breve{\theta}) = E(\breve{\theta}) - E(\tilde{\theta}), \tag{49}$$

and for this case $MSE(\breve{\theta})$ is given by:

$$MSE(\breve{\theta}) = V(\breve{\theta}) + \left(E(\breve{\theta}) - E(\tilde{\theta})\right)^2. \tag{50}$$

In the cases above we assume that estimates of θ based on register data ($\tilde{\theta}$) or a sample survey ($\hat{\theta}$) are unbiased, while big data-based estimates ($\breve{\theta}$) are biased. In this setting we assume that the expectation of $\breve{\theta}$ will be given by:

$$E(\breve{\theta}) = \theta + b, \tag{51}$$

where θ is the true value and b is a constant that refers to bias and $\hat{\theta}$ will be given by:

$$E(\hat{\theta}) = \theta + e, e \sim N(0, \psi), \tag{52}$$

where ψ is known sampling variance.

For register-based estimates of θ we can assume two cases:

$$E(\tilde{\theta}) = \theta, \tag{53}$$

or

$$Bias(\tilde{\theta}) = \theta + \tilde{e}\,, \tilde{e} \sim N(0,\xi). \tag{54}$$

For the first case we assume that register-based estimates are unbiased and, owing to their character, do not contain an additional error component. In the second case it is assumed that $E(\tilde{\theta})$ was initially biased, but thanks to bias-adjustment methods it was reduced and, as a result, an extra error component $\tilde{e} \sim N(0,\xi)$ has been introduced, where ξ is a known (sampling) variance resulting from the bias-adjustment procedure. Therefore, the estimator of $Bias(\breve{\theta})$ based on a representative sample, given by equation (47), has the form:

$$\widehat{\text{Bias}}(\breve{\theta}) = E(\breve{\theta}) - E(\hat{\theta}) = \theta + b - (\theta + e) = b + e, e \sim N(0,\psi), \tag{55}$$

and for register-based estimates, given by equation (49), the estimator of the bias for the two cases listed above is given by:

$$\begin{aligned} \widehat{\text{Bias}}(\breve{\theta}) &= E(\breve{\theta}) - E(\tilde{\theta}) = \theta + b - \theta = b, \\ \widehat{\text{Bias}}(\breve{\theta}) &= E(\breve{\theta}) - E(\tilde{\theta}) = \theta + b - (\theta + e) = b + e, e \sim N(0,\xi)^{\cdot} \end{aligned} \tag{56}$$

Now let us assume that we are interested in estimating bias at domain level. Therefore, equation (47) will be given by the following equation:

$$Bias(\breve{\theta}_d) = E(\breve{\theta}_d) - \hat{\theta}_d, \tag{57}$$

where d refers to domain and $\hat{\theta}_d$ refers to the unbiased estimate of θ_d

The approaches presented above, without considering estimates at domain level, are more appropriate when the estimation of $Bias(\breve{\theta})$ is made at a level for which the variance of $\hat{\theta}$ or bias-adjusted $\tilde{\theta}$ is low, that is, estimates of θ are characterised by *high* precision. However, when the estimation is made for domains that were not planned before conducting the survey, the estimation of θ_d based on a sample survey is characterised by *low* precision (high coefficient of variation). Therefore, a direct estimation of $Bias(\breve{\theta}_d)$ is inappropriate and other approaches should be considered. One solution could be to apply a model-based approach normally applied in small area estimation.

Fosen and Zhang (2011) and *Zhang (2012)* proposed a model-based approach to estimate bias in a register-based survey at domain level. This approach takes into account the fact that θ_d, based on a sample survey, is estimated with low precision, and that a direct estimation of bias may be unreliable. The approach proposed by *Fosen and Zhang (2011)* and *Zhang (2012)* is based on the following assumptions:

- two data sources are available: a register which covers the target population and a separate sample survey of the same population;
- y is the target variable of interest (in *Zhang (2012)* it was unemployment rate);
- θ denotes a target statistic of interest (for example the population mean of y);
- $\tilde{\theta}$ denotes a register-based estimator of θ, which is assumed to be biased;
- $\hat{\theta}$ denotes a survey-based estimator of θ, which is treated as a *gold standard* (reference point), hence it is assumed to be unbiased.

Fosen and Zhang (2011) and *Zhang (2012)* applied small area estimation techniques to estimate the bias of $\tilde{\theta}$, which is defined by the following equation:

$$Bias(\tilde{\theta}_d) = \tilde{\theta}_d - \hat{\theta}_d, \qquad (58)$$

where $\tilde{\theta}_d$ is an estimator of θ_d based on the register

and

$\hat{\theta}_d$ is an unbiased estimator of θ_d based on sample data in domain d.

Assuming that $\tilde{\theta}_d = \theta_d + b_d$ is a biased register-based estimate of θ for domain d, $\hat{\theta}_d = \theta_d + e_i$ is an unbiased survey estimate of θ for domain d (being b_d the bias of $\tilde{\theta}_d$ and e_d the sampling error of $\hat{\theta}_d$) we get:

$$Bias(\tilde{\theta}_d) = \tilde{\theta}_d - \hat{\theta}_d = (\theta_d + b_d) - (\theta_d + e_i) = b_d + \epsilon_d, \qquad (59)$$

where $\epsilon_d = -e_d$ and $e_d \sim N(0, \psi_d)$.

Fosen and Zhang (2011) introduced a random-effects model of bias:

$$b_d = \beta + v_d, \qquad (60)$$

where $E(v_d) = 0$ and $V(v_d) = \sigma_v^2$, which yields the following linear mixed model given by:

$$Bias(\tilde{\theta}_d) = \beta + v_d + \epsilon_d. \qquad (61)$$

Fitting the model gives us:

$$Bi\hat{a}s\,(\tilde{\theta}_d) = \hat{b}_d = \hat{\beta} + \hat{v}_d, \qquad (62)$$

where $\hat{v}_d = \hat{\gamma}_d\,(Bias(\tilde{\theta}_d) - \hat{\beta}), \hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \tilde{\psi}_d)$ and $\tilde{\psi}_d$ is a smoothing function of ψ_d.

As a result we should analyse $Bi\hat{a}s\,(\tilde{\theta}_d)$ to verify the level of bias. In particular, it might be useful to model bias over time to verify whether the level of bias is constant over time. Furthermore, estimated bias might be used to correct big data estimates for self-selection errors. This approach was further developed by *Beręsewicz (2016)* who extended this model to cover several data sources as well as the autocorrelation of bias.

5.2.2. Blending of estimates

The approach we describe in this section can also be associated with measurement error, when we assume that the target variable is measured at domain level with bias. The model was proposed by *Ybarra and Lohr (2008)* and further extended by *Lohr and Brick (2012)* assuming two sources, one of which is biased.

Lohr and Brick (2012) consider three cases:

1. Both data sources are measured without error

$$\lambda \hat{y}_{1d} + (1 - \lambda)\,\hat{y}_{2d}, \qquad (63)$$

where $\hat{y}_{1d}$is the target variable measured in the first source and $\hat{y}_{2d}$ the one measured in the second source.

2. The second data source is biased, and bias (denoted by b_d) is additive

$$\lambda \hat{y}_{1d} + (1 - \lambda)\left(\hat{y}_{2d} - \hat{b}_d\right). \qquad (64)$$

3. The second data source is biased, and bias is multiplicative

$$\lambda \hat{y}_{1d} + (1-\lambda)\, \hat{b}_d\, \hat{y}_{2d}, \tag{65}$$

where $\hat{y}_{1d}$ is the estimate of the target variable based on the first source,, and $\hat{y}_{2d}$ is the estimate of the same variable based on the second source (e. g. big data); the estimator $\hat{y}_{1d}$ is assumed in this case to be unbiased whereas the second estimator $\hat{y}_{2d}$ is potentially biased.

For additive bias *Lohr and Brick (2012)* started with simple model:

$$\begin{pmatrix} \hat{y}_{1d} \\ \hat{y}_{2d} \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_d \\ \theta_d + \eta_d \end{pmatrix}, \sigma^2 \begin{pmatrix} 1/n_{1yd} & 0 \\ 0 & 1/n_{2yd} \end{pmatrix}\right], \tag{66}$$

where n_{1yd} is the sample size in the first source and n_{2yd} is the sample size in the second (biased) source for domain d.

Manzi et al. (2011) discuss modelling bias in combining small area prevalence estimates from multiple surveys. In this approach, bias modelling is crucial, and they propose a series of Bayesian hierarchical models which allow for additive biases.

The main model that *Manzi et al. (2011)* propose is as follows:

$$y_{ds} \sim N(\theta_d + \mu_s, \sigma_{ds}^2 + \tau_s^2), \tag{67}$$

where μ_s and τ_s^2 represent respectively the mean bias and the bias variance for data source s. Following *Manzi et al. (2011, p. 36)*, the model suffers from an identifiability issue that becomes obvious because, for example, all θ_i could be increased by an arbitrary constant and all μ_s decreased by the same constant without changing the fit of the model. Therefore, *Manzi et al. (2011)* added a constraint assuming that the overall estimate of $\overline{\theta}$ is known:

$$\theta_d = \overline{\theta} \times D - \textstyle\sum_{d=1}^{D-1} \theta_d, \tag{68}$$

where d refers to domain $d = 1, \dots, D$.

From the distribution above we have

$$y_{ds} - \mu_s \sim N(\theta_d, \sigma_{ds}^2 + \tau_j^2), \tag{69}$$

and then, for known μ_s and τ_j^2. The analysis is essentially a fixed-effect meta-analysis with $\hat{\theta}_d = \sum_d (y_{ds} - \mu_s)\, w_{ds}$ where

$$w_{ds} = \tfrac{1}{\sigma^{ds} + \tau_s^2} / \textstyle\sum^s \tfrac{1}{\sigma^{ds} + \tau_s^2}, \tag{70}$$

and

$$\overline{w}_s = \textstyle\sum^s w_{ds} / D. \tag{71}$$

Manzi et al. (2011) further extended the model to take into account the correlation between data sources and the time trend.

6 Conclusions

Big data are certainly a new and challenging data source for statistics. Self-selection and the resulting non-probability nature of these data introduce potential bias into estimates based only on these data. However, some of the problems linked to big data usage have been studied earlier in the case of sample or opt-in/internet surveys.

The main issues concerning selectivity in big data sources can be summarised as follows:

- **Big data are paradata-designed** — the paradata collected in big data are enormous and might be used to account for self-selection errors. For instance, psychological or network demographics could be used as covariates in a model of self-selectivity. Understanding the big data source and the paradata associated with it is therefore a key aspect.
- **Big data specific selectivity** — big data sources are heterogeneous, and the type of self-selection error varies between sources. For instance, we might expect self-selection to be informative on Twitter data, whereas for mobile network data it is not.
- **Unit error** — this is one of the main obstacles present in big data. In the majority of cases, we observe the target population indirectly, and the identification of statistical units might be problematic. Probabilistic record linkage methods might help identify the correct statistical units.
- **Unobserved (latent) variables** — the measurement of demographic variables as well as the target variable itself can often be indirect measures. Imputation might be considered as a special case of filling missing data in a latent variable (e.g. machine learning or measurement error models could be applied).
- **Measurement errors** — due to the lack of background information and the need for imputation of variables, measurement errors at an individual level can be significant. However, in such cases we might still have statistical equivalence (the same estimated statistics despite errors in units), and the results may be accurate. *Zhang and Fosen (2012)* made some remarks on this topic.
- **Topic representativeness** — *Schober et al. (2016)* make a distinction between population and topic coverage, and argued that social media might be topic representative. 'For social media analyses, topic coverage can, in principle, be achieved without population coverage. That is, other mechanisms of information propagation that are particular to the dynamics of social media may lead a corpus of social media posts to reflect the broader population's collective opinion and experience, through a range of (not yet fully understood) possible mechanisms. A collection of posts may accurately distill larger conversations in the full population despite the lack of population coverage among posters, perhaps because those who post may have particular access to—or are opinion-formers or elite communicators.'

- **Discrepancies in detail between big data and official statistics** — the population in big data is often on a level that is not available with sufficient accuracy from the designed-based estimation in sample surveys. We could use small area estimation to estimate the internet coverage at LAU 1 or LAU 2 level.

6.1. Sources of selectivity and correction methods in big data

In this section, we provide a summary of sources of selectivity and an overview on the suitability of applying the above-mentioned methods and their limitations for the big data sources addressed in this report.

6.1.1. Mobile network data

The following sources of selectivity can affect mobile network data:

- global level — infrastructure:
  - coverage of provider's BTSs,
  - coverage and signal strength,
  - mobile phone penetration rate in the population,
  - market share of mobile network operators.
- individual level:
  - demographics,
  - usage of mobile phone data (correlated with demographics),
  - type of contract (prepaid, postpaid),
  - labour market.

In some cases, mobile network operators currently provide national statistical institutes with data only aggregated at domain level. Nevertheless, data might be available at object level, hopefully at an increasing rate, even if object-level data, when made available, often lack data on the demographic background of the actual user of the phone number. This information is sometimes imputed based on the analysis of activities of mobile phone users (*Blondel, Decuyper and Krings, 2015*).

The following methods could be used depending on the target population:

- The target units are *trips* (without reference to persons) to study mobility:
  - The use of the generalised weight share method (GWSM) could be explored to correct coverage errors in a mobile operator's infrastructure. GWSM would assign weights to BTSs with regard to the population in related administrative areas.
    - When we restrict data to commuters, available data on this population of trips might be used for reweighting.
  - The selection of reference data is crucial — there is a need for sources that provide information on commuting. Possible sources could be the Urban Audit or analysis of commuters based on administrative registers or the Census of Population and Housing. However, the drawback is that such data may cover only trips between home and work,

which are only part of the trips available in mobile network data; access to such data might be limited and vary between countries.

- The target units are *persons*:
  - Reweighting might be a suitable method that takes into account coverage errors as well as demographic auxiliary variables. Modelling might be problematic in particular when the target variable is an unobserved variable. It might be questionable whether a model built on data from certain mobile users holds for the target population. This limits the possibilities of the application of small area, unit-level models.
  - Moreover, there is limited access to demographic variables, or access that is limited only to sex and age. Auxiliary variables should then be imputed on the basis of activities and BTS logging.

Even if a unit-level approach is recommended whenever possible, if only aggregated data are available, the GWSM method could be used to reweight BTS data to estimate the target variable in the administrative areas (e.g. LAU 1).

We suggest a two-step correction to adjust for infrastructure (coverage error) and self-selection errors (persons). Existing survey estimates of mobile device users could be extended with the information about the contract or mobile operator.

In using the correction methods:

- models should be adjusted for imputation uncertainty (i.e. measurement error),
- models should be adjusted for uncertainty in transforming objects to units (i.e. unit error).

It might be difficult to remove selectivity bias using auxiliary variables from official statistics. For instance, the information available in official statistics might refer to different populations, or only proxy variables are available.

6.1.2. Twitter data

The following sources of selectivity can affect Twitter data:

- global level — infrastructure:
  - internet (including mobile) coverage,
  - social media usage,
  - ‘hot topics’ discussed in the news/TV/internet.
- individual level:
  - demographics,
  - activities on Twitter,
  - interest in topics on Twitter (e.g. following accounts related to politics),
  - tweets themselves (willing to share opinions).

It might be possible to remove self-selection errors from Twitter data. However, it requires several steps:

- Removal of accounts that do not belong to the target population (such as bots, organisations). There are several research questions that should be answered, e.g. whether accounts set up to promote specific political views should be included in the final dataset (they might belong to the target population).
- The acquisition of paradata associated with a Twitter account might be useful for the initial weighting of an account as well as the modelling process (e.g. usage of Twitter).
- The imputation of basic demographic variables as well as the target variable is needed (e.g. via machine learning techniques).

Unit-level methods could be applied to Twitter data. To apply them, we could use two settings:

- Assume one-to-one connections between accounts and persons and then apply the unit-level approach, which might consist of reweighting as well as modelling the approach. If possible, build the same model on external data (e.g. sample survey) and compare estimates.
- Assume many-to-many connections between accounts and persons, then transform objects into statistical units and use the inverse of number of links between accounts and persons as weight for estimation. Follow the steps above for one-to-one cases with and without these weights.

Weighting and modelling could be based on:

- demographic variables — sex, age, nationality, etc.,
- region — based on description or geolocation of tweets,
- variables derived from paradata — number of tweets and their intensity, number of followers and following accounts.

In using the correction methods:

- models should be adjusted for imputation uncertainty (i.e. measurement error),
- models should be adjusted for uncertainty in transforming objects to units.

In the case of Twitter, the data unit-level approach is probably recommended. However, if aggregated, data are present without any additional information (e.g. demographic); a bias modelling approach could then be used. However, this approach assumes that there is a second data source that provides unbiased estimates of the target variable characteristics.

ICT surveys might be extended to provide more information about the use of social media, in particular on using Twitter, Facebook and other portals. In addition, another piece of useful information might relate to the reason for using certain media. For instance, we might assume that Twitter is used mainly for news and political discussions.

6.1.3. Google Trends data

The following sources of selectivity can affect Google Trends data:

- global level:
  - internet (including mobile) coverage
  - usage of the Google search engine,
  - keywords among the most popular ones.
- individual level:
  - topics searched for on the internet.

It is not possible to apply unit level methods to Google Trends in its present form due to the lack of object-level . However, the domain-level approach could be applied to Google Trends. Google Trends is available at national and regional level and for time frequencies much higher than normally available in official statistics (e.g. daily, hourly). One possible approach might be the following:

1. collect Google Trends data at the lowest possible regional level,
2. data should be aggregated on a monthly or quarterly basis,
3. reweight these data according to the average use of the internet at a certain domain level (regions).

There are three possibilities to reweight Google Trends data:

- internet coverage error — there are still people who do not use the internet, and this might vary between regions,

- daily/monthly internet usage — what is the share of people using the internet on a daily basis; we might assume that these people are the population of Google search users,
- Google users — share of users that use the Google search engine.

Another approach might be to adjust for self-selection bias using estimates of the target variable or a proxy (see Section 5.2). These data might also be used as an auxiliary variable or, as current research indicates, for nowcasting of short-term statistics.

Extending or conducting new surveys devoted to the Google search engine for reweighting purposes does not seem to be required for Google Trends in its present form. However, information about changes in the use of the Google search engine may be useful for identifying structural breaks in the nowcasting models and possibly correcting them.

Finally, the bias observed in Google Trends data might also be connected with measurement errors because the concepts observed in Google Trends might be very different from those in official statistics.

6.1.4. Wikipedia data

The following sources of selectivity can affect Wikipedia data:

- global level:
  - internet (including mobile) coverage,
  - usage of online encyclopedias,
  - ‘hot topics’ discussed in the news/TV/internet.
- individual level:
  - demographics,
  - activities on Wikipedia (e.g. editing),
  - interest in topics on Wikipedia.

If the target population is composed of persons, in particular Wikipedia users, the self-selection error might not be removed due to the lack of auxiliary variables. If the target population is the list of topics and its popularity, unit models might be applied. In fact, unit-level models can be used to explain the popularity of Wikipedia articles given their description (e.g. what is a topic).

In its present form, it might not be possible to completely remove self-selection from Wikipedia. However, these data might be used as an auxiliary variable or, as current research indicates, for nowcasting of short-term statistics.

Extending or conducting new surveys to collect information on people’s use of Wikipedia for reweighting purposes does not seem to be required. However, information about changes in its usage may be useful for identifying structural breaks in the models for nowcasting and possibly correcting them.

As is the case for Google Trends data, the bias observed in Wikipedia data might also be connected with measurement errors because the concepts might be very different from those in official statistics.

6.1.5. Summary

The tables below present an overview of the characteristics of the above-mentioned big data sources.

Table5 – An overview of topics related to self-selection in selected big data sources

Characteristic	Mobile	Twitter	Google Trends	Wikipedia

Unit-level corrections	possible	possible	impossible	limited
Domain-level corrections	possible	possible	very limited	limited
Existing sources	available but limited	unavailable (only for social media)	unavailable	available
Background on population	available but limited	limited	unavailable	limited
Paradata	available	available	unavailable	available
MNAR	unlikely	very likely	likely	likely
Measurement error in target variable	fairly likely	very likely	very likely	very likely

Table 6 — Big data sources and possibility of obtaining variables to correct self-selection errors

Characteristic	Mobile	Twitter	Google Trends	Wikipedia
Sex	possible	limited but possible to infer	very limited	very limited
Age	possible	limited but possible to infer	very limited	very limited
Residence	precise	limited but possible to infer	very limited	very limited
Occupation	limited	limited but possible to infer	very limited	very limited
Marital status	limited	limited	very limited	very limited
Employment status	limited	limited	very limited	very limited
Spatial aggregation	municipality	regional/cities	regional/cities	regional/cities

6.2. Self-selection adjustments for big data sources

In this report, we focused on methods that might be useful for bias adjustments. We presented and discussed the feasibility of using available methods, namely weighting procedures, model-based approaches and a combination of them. There are a vast number of methods that could be used for these purposes, but in the end they are limited by (i) available auxiliary information, and (ii) existing data sources.

The research on self-selection methods for big data should be presented together with a careful description of these sources and the imputation of target and auxiliary variables. Only powerful information imputed from the big data paradata can provide information on whether bias can be reduced. To underline this problem, in line with *Zhang et al. (2013)* we present a table with the effects of reweighting on estimates, which can also be applied to the model-based approach.

Table 7 — Relationship between bias and variance, its association with target

variable (y), auxiliary variables $(\mathbf{x})$ and response indicator (R)

	Low association (x, y)	**High association (x, y)**
Low association $(x,\ R)$	Little effect on bias	Little effect on bias
	Little effect on variance	Variance reduction
High association (x,R)	Little effect on bias	Bias reduction
	Variance inflation	Variance reduction

The table shows that we can only reduce bias using big data sources if there is a high association of x, R and y. Otherwise, external sources such as surveys or administrative data should be used.

Another general issue is that technical and data-related possibilities to assess, and account for, the selective property of big data sources vary a lot between EU countries depending on the maturity of the statistical infrastructure in a given country. Big data sources are often available on a multi-national and worldwide scale, while access to the relevant data for correcting self-selection is limited and might be impossible to apply in certain countries. For example, 'survey countries' (where surveys are the basic data source) in comparison to 'register countries' (where registers are extensively used) means limitations to providing the relevant auxiliary data on a low level of aggregation.

6.3. Recommended framework for adjusting selectivity in big data

We propose the following framework to deal with selectivity in big data sources. It assumes that the target statistical unit, population and variables have already been defined based on the identified needs of users. It includes an exploration phase that needs to run sequentially before the big data source is used, although it can take place separately. It then includes a phase of big data pre-processing. Even if not addressing selectivity, it has direct implications for bias and the successful application of the adjustment methods.

Exploration phase of big data sources

1. Define the big data population and objects.
2. Study the big data source to verify whether it is possible to derive the target variable and auxiliary variables.
3. Study the paradata available in the big data source, which might be used for the imputation of: (i) background information (e.g. sex, age, departure-destination location) that can be used to correct self-selection errors, (ii) the target variable (or its proxy).

Stage 0: Pre-processing big data

1. If possible, search for the appropriate method to impute variables from point 3 of the exploration phase and report on its uncertainty (accuracy, root mean square erros,).
2. If possible, transform objects into statistical units (unit identification).

Stage I: Identify selectivity

1. Select suitable existing auxiliary data sources among censuses, sample surveys and statistical registers. If not available, consider conducting new surveys, adding questions to existing surveys or look for new administrative sources.

2. If possible, compare the distributions of auxiliary variables.
3. Compute statistics of interest for the target variable derived from big data sources and compare them with existing data sources if possible.
4. If possible, compare estimates broken down by auxiliary variables at domain level with existing data sources.
5. Detect discrepancies between the distribution of auxiliary and target variables.

Stage II: Select suitable methods for dealing with selectivity errors

1. Verify whether unit-level methods can be applied or if the domain-level approach is the only possible one
2. If the unit-level approach can be used:
   - If possible, apply basic weighting procedures such as calibration (e.g. raking, post-stratification) to correct for coverage errors and adjust for known totals.
   - If results are not satisfactory, apply a propensity score with weighting procedures.
   - If results are not satisfactory, consider a model-based approach with weighting adjustments.
3. If the unit-level approach is not possible or not working, consider the aggregation of data, and:
   - If possible, apply basic weighting procedures such as calibration (e.g. raking, post-stratification) to correct for coverage errors and adjust for known totals.
   - If results are not satisfactory, adjust for bias using domain-level estimates.

Bibliography

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., and Diaz, C. (2014). *The web never forgets: Persistent tracking mechanisms in the wild*. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (pp. 674–689). ACM.

Acar, G., Van Alsenoy, B., Piessens, F., Diaz, C., and Preneel, B. (2015). *Facebook tracking through social plug-ins*. Retrieved from https://securehomes.esat.kuleuven.be/~gacar/fb_tracking/fb_plugins.pdf

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L., Nurmi, O., Potier, F., Schmücker, D., Sonntag, U., and Tiru, M. (2013). *Feasibility study on the use of mobile positioning data for tourism statistics*. Retrieved from http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies

Baker R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., and Tourangeau, R. (2013). *Summary Report of the AAPOR Task Force on Non-probability Sampling*. Journal of Survey Statistics and Methodology 1, 90–143. Retrieved from https://www.researchgate.net/publication/273561892

Beręsewicz, M. (2016). *Internet data sources for real estate market analysis*. PhD Dissertation, available from the author.

Berger, Y. (2017). *Empirical Likelihood approach for unit non-response*. In Articus, C., et al. (2017) Future needs in statistics. InGRID Deliverable 23.2. Trier: InGRID project. https://inclusivegrowth.be/downloads/output/d23-2-final.pdf

Berger, Y.G., and De La Riva Torres, O. (2016). *An empirical likelihood approach for inference under complex sampling design*. Journal of the Royal Statistical Society, Series B, 78(2), 319-341. URL http://dx.doi.org/10.1111/rssb.12115.

Bethlehem, J. (2010). *Selection bias in web surveys*. International Statistical Review, 78(2), 161–188. Wiley Online Library.

Bethlehem, J., and Biffignandi, S. (2012). *Handbook of web surveys*. John Wiley & Sons.

Beyer, M., and Laney, D. (2012). *The importance of 'big data': A definition*. Retrieved from https://www.gartner.com/doc/2057415/importance-big-data-definition

Biemer, P. (2014), *Total Survey Error: Adapting the Paradigm for Big Data*, Retrieved from http://www.niss.org/sites/default/files/biemer_ITSEW2014_Presentation.pdf

Blondel, V. D., Decuyper, A., and Krings, G. (2015). *A survey of results on mobile phone datasets analysis. EPJ Data Science*, 4(1), 1. Springer Berlin Heidelberg.

Börsch-Supan, A., Elsner, D., Fassbender, H., Kiefer, R., McFadden, D., and Winter, J. (2004). *How to make internet surveys representative: A case study of a two-step weighting procedure*.

Breiman, L. (2001). *Statistical modelling: The two cultures* (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231. Institute of Mathematical Statistics.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Brick, J. M. (2013). *Unit Nonresponse and Weighting Adjustments: A Critical Review*. Journal of Official statistics, 29(3), 329–353.

Brown, A. R. (2011). *Wikipedia as a data source for political scientists: Accuracy and completeness of coverage*. PS: Political Science & Politics, 44(02), 339–343. Cambridge Univ Press.

Buelens, B., Burger, J., and Brakel, J. van den. (2015). *Predictive inference for non-probability samples: a simulation study.* Retrieved from https://www.researchgate.net/publication/283441457_Predictive_inference_for_non-probability_samples_a_simulation_study

Brüggen, E., van den Brakel, J., and Krosnick, J. (2016). *Establishing the accuracy of online panels for survey research*. CBS, The Netherlands, Discussion papers 2016|04.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (Eds.). (2014). *Online panel research: A data quality perspective*. John Wiley & Sons.

Chambers, R. (2009). *Regression Analysis of Probability-Linked Data.* Official statistics research series. Wellington: Statistics New Zeland. Retrieved from http://www.statisphere.govt.nz/official-statistics-research/series/default.htm

Chambers, R., and Chandra, H. (2013). *A random effect block bootstrap for clustered data.* Journal of Computational and Graphical Statistics, 22(2), 452–470. Taylor & Francis.

Chambers, R., and Clark, R. (2012). *An introduction to model-based survey sampling with applications* (Vol. 37). OUP Oxford.

Chambers, R., and Tzavidis, N. (2006). *M -quantile models for small area estimation*. Biometrica, 93(2), 255–268.

Chang, T., and Kott, P.S. (2008). *Using calibration weighting to adjust for nonresponse under a plausible model*. Biometrika 95 (555-571).

Chatterjee, N., Chen, Y.H., Maas, P., and Carroll, R.J. (2016). *Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources*. J Am Stat Assoc. 111(513), 107-117. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4994914/

Choi, H., and Varian, H. (2012). *Predicting the present with Google Trends*. Economic Record, 88(1), 2–9. Wiley Online Library.

Chu, Z., Gianvecchio, S., Wang, H., and Jojodia, S. (2012). *Detecting automation of Twitter accunts: Are you a human, bot or cyborg?* IEEE Transactions on Dependable and Secure Computing, 9(6), 811–824. IEEE.

Citro, C. F. (2014). *From multiple modes for surveys to multiple data sources for estimates*. Survey methodology, 40(2), 137–161.

Cormack, R. M. (1989). *Log-linear models for capture-recapture*. Biometrics, 395–413.

Craig, T., and Ludloff, M. E. (2011). *Privacy and big data*. ‘O'Reilly Media, Inc.’

Daalmans, J. (2017). *Mass imputation for census estimation*. CBS, The Netherlands, Discussion Papers 2017|04.

Daas, P., Puts, M., Buelens, B., and Van den Hurk, P. (2015). *Big data as a source for official statistics*. Journal of Official Statistics, 2(31).

Demidenko, E. (2004). *Mixed Models. Theory and Applications*. New York: Wiley.

Deming, W.E., and Stephan, F.F. (1940). *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*. Annals of Mathematical Statistics, 11, 427-444.

Dever, J. A., and Valliant, R., (2006), *A Comparison of Model-Based and Model-Assisted Estimators under Ignorable and Non-Ignorable Nonresponse.* Proceedings of the Section on Survey Research Methods, Washington DC: American Statistical Association, 2938–2945.

Deville, J.-C. (2000) *Generalized calibration and application to weighting for non-response*. In: Bethlehem J.G. and van der Heijden, P.G.M. (eds) COMPSTAT. Physica, Heidelberg.

Deville, J.-C., and Lavallée, P. (2006). *Indirect sampling: The foundations of the generalized weight share method.* Survey Methodology, 32(2), 165.

Deville, J.-C., and Särndal, C.-E. (1992). *Calibration estimators in survey sampling*. Journal of the American statistical Association, 87(418), 376–382. Taylor & Francis Group.

Deville, J.-C., Särndal C.-E., and Sautory, O. (1993). *Generalized Raking Procedures in Survey Sampling*. Journal of the American Statistical Association, 88(423), 1013-1020.

Di Consiglio L., Tuoto T. (2017) *Small Area Estimation in the Presence of Linkage Errors*. In: Ferraro M. et al. (eds) Soft Methods for Data Science. Advances in Intelligent Systems and Computing, vol 456. Springer, Cham

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., and Rothschild, D. (2016). *Online and social media data as an imperfect continuous panel survey*. PloS one, 11(1), e0145406. Public Library of Science.

Diggle, P.J., and Kenward, M.G. (1994). *Informative dropout in longitudinal data analysis (with discussion)*. Journal of Royal Statistical Society B 43 (49-93).

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Efron, B., and Tibshirani, R. (1986). *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy.* Statistical science, 54–75. JSTOR.

Elliott, M.R. (2009). *Combining data from probability and non- probability samples using pseudo-weights*. Survey Practice, August 2009.

Elliott, M.R., and Valliant, R. (2017). *Inference for nonprobability samples*. Statist. Sci. 32(2), 249-264.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

European Statistical System Committee (2013). *Scheveningen Memorandum on 'Big Data and Official Statistics'*. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/SCHEVENINGEN_MEMORANDUM Final version.pdf

European Statistical System Committee (2014). *Big Data Action Plan and Roadmap*. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/ESSC doc 22_8_2014_EN_Final with ESSC opinion.pdf

Eurostat (2016). Glossary – *primary data*. Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Primary_data

Fabrizi, E., Salvati, N., Pratesi, M., and Tzavidis, N. (2014). *Outlier robust model- assisted small area estimation*. Biometrical Journal, 56(1), 157–175.

Fay, R. E. (1996). *Alternative paradigms for the analysis of imputed survey data*. Journal of the American Statistical Association, 91(434), 490–498. Taylor & Francis Group.

Feder, M., and Pfeffermann, D. (2015). *Statistical inference under non-ignorable sampling and non-response*. University of Southampton. Retrieved from http://eprints.soton.ac.uk/378245/

Fienberg, S. E. (1972). *The multiple recapture census for closed populations and incomplete 2k contingency tables*. Biometrika, *59*(3), 591–603.

Fondeur, Y., and Karamé, F. (2013). *Can Google data help predict French youth unemployment?* Economic Modelling, 30, 117–125. Elsevier B.V. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0264999312002490

Fosen, J., and Zhang, L.-C. (2011). *The approach to quality evaluation of the micro-integrated employment statistics*.

Franks, A.M., Airoldi, E.M., and Rubin, D.B. (2016). *Non-standard conditionally specified models for non-ignorable missing data*. Retrieved from https://arxiv.org/pdf/1603.06045.pdf

Gad, A.M. (2011). *A selection model for longitudinal data with non-ignorable non-monotone missing values*. Journal of Data Science 9 (171-180).

Gelman, A. (2007). *Struggles with Survey Weighting and Regression Modelling*. Statistical Science, 22(2), 153–164. Retrieved from http://projecteuclid.org/euclid.ss/1190905511

Gelman A., and Hill, J. (2009). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gentry, J. (2015). *TwitteR: R based twitter client*. Retrieved from https://CRAN.R-project.org/package=twitteR

George, G., Haas, M. R., and Pentland, A. (2014). *Big data and management*. Academy of Management Journal, 57(2), 321-326.

Giles, J. (2005). *Internet encyclopaedias go head to head*. Nature, 438(7070), 900–901. Nature Publishing Group.

Google Analytics. (2016a). *Cookies and user identification*. Retrieved from https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id

Google Analytics. (2016b). *Google analytics cookie usage on website*. Retrieved from https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage

Google Analytics. (2016c). *User iD and cross device reports*. Retrieved from https://support.google.com/analytics/topic/6009743?hl=en&ref_topic=1007027

Google Analytics. (2016d). *Limits of user-ID views & Cross Device reports*. Retrieved from https://support.google.com/analytics/answer/3223194

Google Trends. (2016a). *Where trends data comes from*. Retrieved from https://support.google.com/trends/answer/4355213?hl=en&ref_topic=4365599

Google Trends. (2016b). *How trends data is adjusted*. Retrieved from https://support.google.com/trends/answer/4365533?hl=en&ref_topic=4365599

Groves, R. M. (2011a). *'Designed data' and 'organic data'*. Retrieved from http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/

Groves, R. M. (2011b). *Three Eras of Survey Research*. Public Opinion Quarterly, 75(5), 861–871. Retrieved from http://poq.oxfordjournals.org/cgi/doi/10.1093/poq/nfr057

Guandalini, A., and Tillé, Y. (2017). *Design-based estimators calibrated on estimated totals from multiple surveys*. International Statistical Review 85, 250–269.

Hajjem, A., Bellavance, F., and Larocque, D. (2011). *Mixed effects regression trees for clustered data*. Statistics and Probability Letters, 81(4), 451–459. Elsevier B.V. Retrieved from http://dx.doi.org/10.1016/j.spl.2010.12.003

Hajjem, A., Bellavance, F., and Larocque, D. (2014). *Mixed-effects random forest for clustered data*. Journal of Statistical Computation and Simulation, 84(6), 1313–1328.

Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning*, 2nd edition. New York, NY, USA: Springer New York Inc.

Haziza, H., and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics 32, 129–145.

Heckman, J. J. (1976): *The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models*. Annals of Economic and Social Measurement, 5, 475–492.

Hilles, S. (2014). *To use or not to use? The credibility of Wikipedia*. Public Services Quarterly, 10(3), 245–251. Taylor & Francis.

Horrigan, M. W. (2013). *Big data: A perspective from the BLS*. Retrieved from http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/

Houbiers, M. (2004). *Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands*. Journal of official statistics, 20(1), 55–75.

Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H., and Snijders, V. (2003). *Estimating consistent table sets: position paper on repeated weighting*. Statistics Netherlands, Discussion paper, 3005, 2003.

IEEE Standards Association. (2010). *Systems and software engineering - Vocabulary* ISO/IEC/IEEE 24765: 2010. Iso/Iec/Ieee, 24765, 1-418.

Janecek A., Valerio D., Hummel K. A., Ricciato F., and Hlavacs, H. (2015) *The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring*. IEEE Transaction on Intelligent Transportation Systems.

Johnston, J. (1991). *Econometric methods*. Third [International] Edition) McGraw-Hill Companies, Inc.

Keller, S. A., Koonin, S. E., and Shipp, S. (2012). *Big data and city living–what can it do for us?* Significance, 9(4), 4–7. Wiley Online Library.

Kim, G., and Chambers, R. (2012). *Regression analysis under incomplete linkage*. Statistica Neerlandica, 56(9), 2756–2770. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167947312001089

Kott, P.S., and Chang, T. (2010). *Using calibration weighting to adjust for nonignorable unit nonresponse*. JASA, 105(491), 1265-1275.

Kott, P.S., and Liao, D. (2015). *One step or two? Calibration weighting from a complete list frame with nonresponse*. Survey Methodology 41(1), 165-181.

Kott, P.S. and Liao, D. (2017). *Calibration weighting for nonresponse that is Not Missing at Random: Allowing more calibration than response-model variables*. Journal of Survey Statistics and Methodology, 5(2) (159–174).
Retrieved from https://doi.org/10.1093/jssam/smx003

Kreuter, F. (Ed.). (2013). *Improving surveys with paradata: Analytic uses of process information* (Vol. 581). John Wiley & Sons.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., et al. (2010). *Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(2), 389–407. Wiley Online Library.

Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neill, C., and Usher, A. (2015). *AAPOR Report on Big Data* (No. 4eb9b798fd5b42a8b53a9249c7661dd8). Mathematica Policy Research.

Lahiri, P., and Larsen, M. D. (2005). *Regression Analysis with Linked Data*. Journal of the American Statistical Association, 100(469), 222–230.

Laney, D. (2001). *3D Data Management: Controlling data volume, velocity and variety*. META Group [now Gartner] Research Note. Retrieved from http://goo.gl/Bo3GS

Lavallée, P. (2009). *Indirect sampling* (Vol. 7397). Springer Science & Business Media.

Lavallée, P. (2007). Indirect Sampling. Springer Series in Statistics. Springer, New York, NY. Chapter: GWSM and Calibration, 121-150. Retrieved from https://link.springer.com/chapter/10.1007/978-0-387-70782-2_7

Lavallée, P. (2015). *Sample Matching: Toward a probabilistic approach for Web surveys and Big Data?*. Plenary Session. ITACOSM 2015 4th ITAlianConference on Survey Methodology Rome, June 24-26, 2015.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., and Jebara, T. (2009) *Computational social science*. Science, 323, 721.

Lee, S. (2006). *Propensity score adjustment as a weighting scheme for volunteer panel web surveys*. Journal of Official Statistics, 22(2), 329.

Lee, S., and Valliant, R. L. (2009). *Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment*. Sociological Methods & Research, 37(3), 319-343.

Lehtonen, R., and Veijanen, A. (2012). *Small area poverty estimation by model calibration.* Journal of the Indian Society of Agricultural Statistics, 66(1), 125–133.

Lehtonen, R. and Veijanen, A. (2016). *Model-assisted methods for small area estimation of poverty indicators*. In: Pratesi M. (ed.) *Analysis of Poverty Data by Small Area Estimation*. Chichester: Wiley, 109–127.

Lenau, S., and Münnich, R. (2017). *The use of New Data and non-probability sampling*. In Articus, C. et al. (2017). Future needs in statistics. InGRID Deliverable 23.2. Trier: InGRID project. Retrieved from https://inclusivegrowth.be/downloads/output/d23-2-final.pdf

Lepkowski, J. M., Tucker, C., Brick, J. M., De Leeuw, E. D., Japec, L., Lavrakas, P. J., Link, M. W., et al. (Eds.). (2007). *Advances in telephone survey methodology* (Vol. 538). John Wiley & Sons.

Little, R. J. (1993). *Pattern-mixture models for multivariate incomplete data*. Journal of the American Statistical Association, 88(421), 125-134.

Little, R. (2011). *Calibrated Bayes, for statistics in general, and missing data in particular*. Statistical Science 26(2), 162–174.

Little, R. J. (2012). *Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder)*. Journal of Official Statistics, 28(3), 309–372.

Little, R. J., and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J. (2015). *Calibrated Bayes, an inferential paradigm for official statistics in the era of big data*. Statistical Journal of the IAOS, 31(4), 555–563. IOS Press.

Liu, X. (2017). *Empirical likelihood for the response mean of generalized linear models with missing at random responses*. Communications in Statistics - Simulation and Computation 46, 164-173.

Loh, W.-Y. (2014). *Fifty years of classification and regression trees*. International Statistical Review, 82(3), 329–348. Wiley Online Library.

Lohr, S. (2009). *Sampling: Design and analysis*. Cengage Learning.

Lohr, S., and Brick, J. (2012). *Blending domain estimates from two victimization surveys with possible bias.* Canadian Journal of Statistics, 40(4), 679–696. Retrieved from http://onlinelibrary.wiley.com/wol1/doi/10.1002/cjs.11153/full

Lundström, S., and Särndal, C.-E. (1999). *Calibration as a standard method for treatment of nonresponse*. Journal of Official Statistics, 15(2), 305–328. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=152305

Luo, S., and Pang, S. (2017). *Empirical likelihood for quantile regression models with response data missing at random*. Open Math. 15 (317–330). DOI 10.1515/math-2017-0028

Lynn, P. (Ed.). (2009). *Methodology of longitudinal surveys*. John Wiley & Sons.

Maia, M. (2009). *Indirect sampling in context of multiple frames*. Proceedings of Section on Survey Research Methods, 2009. Retrieved from http://ww2.amstat.org/sections/srms/Proceedings/y2009/Files/303803.pdf

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., and Thompson, S. G. (2011). *Modelling bias in combining small area prevalence estimates from multiple surveys.* Journal of the Royal Statistical Society. Series A (Statistics in Society), 174(1), 31–50.

Marchetti, S, Giusti, C., and Pratesi, M. (2015). *The use of Big Data from Twitter for small area estimation of households' share of food consumption expenditure in Italy*. In Articus C. et al. (2017) Future needs in statistics. InGRID Deliverable 23.2. Trier: InGRID project.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L. (2015). *Small area model-based estimators using big data sources*. Journal of Official Statistics, 31(2), 263–281. Retrieved from http://doi.org/http://dx.doi.org/10.1515/JOS-2015-0017

Massicotte, P., and Eddelbuettel, D. (2016). *GtrendsR: R functions to perform and display google trends queries*. Retrieved from https://CRAN.R-project.org/package=gtrendsR

Matei, A. and Ranalli, M.G. (2015). *Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach*. Survey Methodology, 41(1), 145-164.

McFadden, D., S. Cosslett, G. Duguay, and W. S. Jung (1977). *Demographic data for policy analysis*. Urban Travel Demand Forecasting Project, Final Report, Volume VIII, Institute of Transportation Studies, University of California, Berkeley.

Meissner, P., and Team, R. C. (2016). *Wikipediatrend: Public subject attention via wikipedia page view statistics*. Retrieved from https://CRAN.R-project.org/package=wikipediatrend

Mercer, A.W., Kreuter, F., Keeter, S., and Stuart, E.A. (2017). *Theory and practice in nonprobability surveys. Parallels between causal inference and survey inference*. Public Opinion Quarterly, 81, Special Issue, 250–279.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). *Every missingness not at random model has a missingness at random counterpart with equal fit*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(2), 371–388. Wiley Online Library.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). *Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose*. arXiv preprint arXiv:1306.5204.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., and Pasek, J., et al. (2014). *Social Media in Public Opinion Research* Executive Summary of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. Public Opinion Quarterly, 78(4), 788–794. AAPOR. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Social_Media_Report_FNL.pdf

National Academies of Sciences, Engineering, and Medicine. (2017). *Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/24654.

Nordman, D. J., and Lahiri, S. N. (2004). *On optimal spatial subsample size for variance estimation*. Annals of statistics, 1981–2027. JSTOR.

Our Social Times. (2016). *7% of Twitter users are not human*. Retrieved from http://oursocialtimes.com/7-of-twitter-users-are-not-human/.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Owen, A. B. (2013). *Self-concordance for empirical likelihood*. Canadian Journal of Statistics, 41(3), 387–397.

Park, S. and Kim, J.K. (2014). *Instrumental-variable calibration estimation in survey sampling*. Statistica Sinica 24, 1001-1015. Retrieved from http://dx.doi.org/10.5705/ss.2013.038

Peress, M. (2010). *Correcting for Survey Nonresponse Using Variable Response Propensity*. Journal of the American Statistical Association 105(492), 1418-1430.

Pfeffermann, D. (2011). *Modelling of complex survey data: Why model? Why is it a problem? How can we approach it*. Survey Methodology, 37(2), 115–136.

Pfeffermann, D., and Sverchkov, M. (2003). *Fitting generalized linear models under informative sampling*. Analysis of survey Data, 175–195. Wiley, New York, USA.

Pfeffermann, D., and Sverchkov, M. (2007). *Small-Area Estimation under Informative Probability Sampling of Areas and Within the Selected Areas*. Journal of the American Statistical Association, 102(480), 1427–1439. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1198/016214507000001094

Pratesi, M., (Ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. Wiley.

Ranalli, M.G., Arcos, A., Rueda, M.M., et al. (2016). *Calibration estimation in dual-frame surveys*. Stat Methods Appl 25 (321-349). Retrieved from https://doi.org/10.1007/s10260-015-0336-5

Rao, J. (1996). *On variance estimation with imputed survey data*. Journal of the American Statistical Association, 91(434), 499–06. Taylor & Francis.

Rao, J. N. K., and Molina, I. (2015) *Small area estimation*, 2nd ed. Wiley

Rao, J. N. K., and Wu, C. (2010). *Pseudo–Empirical Likelihood Inference for Multiple Frame Surveys*. Journal of the American Statistical Association, 105(492), 1494–1503. Retrieved from http://www.tandfonline.com/doi/abs/10.1198/jasa.2010.tm09534

Reilly, C., Gelman, A., and Katz, J. (2001). *Poststratification without Population Level Information on the Poststratifying Variable with Application to Political Polling*. Journal of the American Statistical Association, 96(453), 1–11.

Reis, F., Di Consiglio, L., Kovachev, B., Wirthmann, A., and Skaliotis, M. (2016). *Comparative assessment of three quality frameworks for statistics derived from big data: the cases of Wikipedia page views and Automatic Identification Systems*. In European Conference on Quality in Official Statistics (Q2016), Madrid (Vol. 31).

Rey del Castillo, P. (2012). *Use of machine learning methods to impute categorical data*. In Conference on European Statistics.

Ricciato, F. (2006) *Traffic monitoring and analysis for the optimization of a 3G network* IEEE Wireless Communications — Special Issue on 3G/4G/WLAN/WMAN Planning, 13(6).

Ricciato, F., Svoboda, P., Motz, J., Fleischer, W., Sedlak, M., Karner, M., Pilz, R., Romirer-Maierhofer, P., Hasenleithner, E., Jäger, W., Krüger, P., Vacirca, F., Rupp, M. (2006). *Traffic monitoring and analysis in 3G networks: lessons learned from the METAWIN project*. Elektrotechnik & Informationstechnik.

Ricciato, F., Widhalm, P., Craglia, M., and Pantisano, F. (2015). *Extraction of population density distribution from network-based mobile phone data*. Retrieved from https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Final-%20jrc-AIT-MNO-study-compressed_1.pdf

Riddles, M. K. (2013). *Propensity score adjusted method for missing data* (PhD Thesis). Iowa State University. Retrieved from http://lib.dr.iastate.edu/etd/13287

Rivers, D. (2007). *Sampling for web surveys*. In Joint statistical meeting.

Roger, S., Bivand, R. S., and Pebesma, E. J. (2013). *Applied spatial data analysis with R*, John Wiley & Sons.

Rosenbaum, P. R., and Rubin, D. B. (1983). *The central role of the propensity score in observational studies for causal effects*. Biometrika, 70(1), 41–55. Biometrika Trust.

Rubin, D. B. (1976). *Inference and missing data*. Biometrika, 63(3), 581–592. Biometrika Trust.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Rubin, D. B. (1996). *Multiple imputation after 18+ years*. Journal of the American statistical Association, 91(434), 473–489. Taylor & Francis.

Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. (2012). *Small area estimation via M-quantile geographically weighted regression*. Test, 21(1), 1–28.

Samart, K. (2011). *Analysis of probabilistically linked data* (PhD thesis). Doctor of Philosophy thesis, School of Mathematics; Applied Statistics, University of Wollongong. Retrieved from http://ro.uow.edu.au/theses/3513/

Samart, K., and Chambers, R. (2014). *Linear Regression with Nested Errors Using Probability-Linked Data*. Australian & New Zealand Journal of Statistics, 56(1), 27–46. Retrieved from http://doi.wiley.com/10.1111/anzs.12052

Särndal, C.-E. (2007). *The calibration approach in survey theory and practice*. Survey Methodology, 33(2), 99–119.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., and Conrad, F. G. (2016). *Research Synthesis. Social Media Analyses for Social Measurement*. Public Opinion Quarterly, 80(1), 180–211.

Sela, R. J., and Simonoff, J. S. (2012). *RE-EM trees: a data mining approach for longitudinal and clustered data.* Machine Learning, 86, 169–207.

Shao, J. (2003). *Impact of the bootstrap on sample surveys*. Statistical Science, 18(2), 191–198. Institute of Mathematical Statistics.

Shirley, K. E., and Gelman, A. (2015). *Hierarchical models for estimating state and demographic trends in US death penalty public opinion*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(1), 1-28.

Smith, P., Lynn, P., and Elliot, D. (2016). *Sample design for longitudinal surveys*. In P. Lynn (Ed.), Methodology of longitudinal surveys (pp. 21–34). John Wiley & Sons.

Stern, M., Adams, A., and Elasser, S. (2009). *Digital Inequality and Place: The Effects of Technological Diffusion on Internet Proficiency and Usage across Rural, Suburban, and Urban Counties*. Sociological Inquiry, 79(4), 391–417.

Sutradhar, Rao and Pandit (2010) *Inferences in longitudinal mixed models for survey data.* Journal of the Indian Society of Agricultural Statistics, a special issue in Memory of Dr. G. R. Seth, 64, 177–189.

Tam S.-M., and Clarke, F. (2015). *Big Data, official statistics and some initiatives by the Australian Bureau of Statistics*. International Statistical Review, 83(3), 436–448.

The Verge. (2016). *Twitter now let anyone request a verified account*. Retrieved from http://www.theverge.com/2016/7/19/12227490/twitter-opening-verified-account-user-form

Tuoto, T., Fusco, D., and Di Consiglio, L. (2016). *Exploring solutions for linking Big Data in Official Statistics*. SIS 2016. Conference proceedings ISBN: 9788861970618.

Twitter. (2016). *About verified accounts*. Retrieved from https://support.twitter.com/groups/31-twitter-basics/topics/111-features/articles/119135-about-verified-accounts

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., and Chambers, R. (2015). *Robust small area prediction for counts.* Statistical methods in medical research, 24(3), 373–395.

Valliant, R., and Dever J.A. (2011). *Estimating propensity adjustments for volunteer web surveys*, Sociological Methods & Research, 40(1), pp. 105-137.

Vanderhoeft, C. (2001). *Generalised Calibration at Statistics Belgium*. SPSS Module g-CALIB-S and Current Practises. Statistics Belgium.

Verbeek, M., and Nijman, T. (1996). *Incomplete panels and selection bias*. In Mátyás, L. and Patrick S. (Eds.) (1996). The Econometrics of Panel Data: A Handbook of the Theory with Applications. Kluwer Academic Publishers (449-490).

Vicente, M. R., López-Menéndez, A. J., and Pérez, R. (2015). *Technological Forecasting & Social Change Forecasting unemployment with internet search data: Does it help to improve predictions* *when job destruction is skyrocketing?* Technological Forecasting & Social Change, 92, 132–139. Elsevier Inc. Retrieved from http://dx.doi.org/10.1016/j.techfore.2014.12.005

Vonesh, E.F. (2012). *Generalized Linear and Nonlinear Models for Correlated Data*. Theory and Applications Using SAS. SAS Institute.

Wallgren, A., and Wallgren, B. (2014). *Register-based Statistics*. Wiley series in survey methodology (Second.). John Wiley & Sons, Inc.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). *Measurement error in network data: A re-classification.* Social Networks, 34(4), 396-409.

Wang, W., Rothschildb, D., Goelb, S., and Gelman, A. (2015). *Forecasting elections with non-representative polls.* International Journal of Forecasting, 21(3), 980–991.

Wang, J.J.J., Bartlett, M. and Ryan, L. (2017). *Non-ignorable missingness in logistic regression*. Statistics in Medicine 36 (3005–3021). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28574592

Wolter, K. M. (1986). *Some coverage error models for census data*. Journal of the American Statistical Association, 81(394), 337–346.

Wu, C. (2005). *Algorithms and R codes for the pseudo empirical likelihood method in survey sampling*. Survey Methodology, 31(2), 239–243.

Wu, C., and Lu, W. W. (2016). *Calibration Weighting Methods for Complex Surveys*. International Statistical Review, 84(1), 79–98.

Wu, C., and Sitter, R. R. (2001). *A model-calibration approach to using complete auxiliary information from survey data*. Journal of the American Statistical Association, 96(453), 185–193. Taylor & Francis.

Ybarra, L. M. R., and Lohr, S. L. (2008). *Small area estimation when auxiliary information is measured with error*. Biometrika, 95(4), 919–931. Retrieved from http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/asn048

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., & Wang, R. (2011). *Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples*. Public Opinion Quarterly, 75(4), 709–747.

Zabala, F. (2015). *Let the data speak: Machine learning methods for data editing and imputation*. Conference on European Statistics.

Zhang, L.-C. (1999). *A note on post-stratification when analyzing binary survey data subject to nonresponse*. Journal of Official Statistics, 15(2), 329–334.

Zhang, L.-C. (2011). *A Unit-Error Theory for Register-Based Household Statistics.* Journal of Official Statistics, 27(3), 415–432.

Zhang, L. C. (2012). *On the accuracy of register-based census employment statistics*. In European Conference on Quality in Official Statistics (Q2012), May.

Zhang, L.-C. (2015). *On modelling register coverage errors.* Journal of Official Statistics, 31(3), 381–396.

Zhang, L.-C. (2016). *Multisource statistics: Quality and statistical methods*. Training course provided in European Conference on Quality in Official Statistics (Q2016), May.

Zhang, L.-C., & Fosen, J. (2012). *A Modelling Approach for Uncertainty Assessment of Register-based Small Area Statistics*. Journal of the Indian Society of Agricultural Statistics, 66(1), 91–104.

Zhang, L.-C., Thomsen, I., & Kleven, Ø. (2013). *On the Use of Auxiliary and Paradata for Dealing with Non-sampling Errors in Household Surveys*. International Statistical Review, 81(2), 270–288.

Mathematical appendix

A.1 Notation

a^{active}	Adjustment for a fraction of the active users' population (e.g. people who use the internet on a daily or monthly basis)
a_i^{cal}	Adjustment to calibrate the weights
a_i^{cov}	Adjustment for incomplete mobile/internet coverage
a_i^{nr}	Adjustment for nonresponse
a^{share}	Market share of a given company
a_i^{trim}	Adjustment to control variation among weights
$cor_{\rho,y}$	Correlation coefficient between the target variable y and the response behaviour ρ
$\boldsymbol{d}$	Vector of design weights
$d_i = 1/\pi_i$	Design weights that are the inverse of the probability π_i that a given unit i will be sampled
$D(\boldsymbol{w}, \boldsymbol{d})$	Distance function between calibrated weights $\boldsymbol{w}$ and design weights $\boldsymbol{d}$
$E_{M^*}(\theta)$	Expectation with respect to the working model M^*
$E_{\varsigma}(\hat{\theta})$	Expectation with respect to the sampling distribution (ς)
$f(\boldsymbol{x})$	Function of auxiliary variables
$f(y\|\boldsymbol{x})$	Conditional function of y given x
I_i	Inclusion indicator — if statistical unit i is included in sample then $I_i = 1$, otherwise $I_i = 0$
$inv\text{-}\chi^2(\nu, \sigma_0^2)$	Inverse χ^2 distribution
$N(\mu, \sigma^2)$	Normal distribution
N_{NI}/N	Proportion of people without internet access
$P(\bullet)$	Probability of $\bullet$
$\hat{p}_i$	Weight of statistical unit i from pseudo empirical log-likelihood maximisation
$q_{ij} = n_{ij}/n$	Gross sample proportion
r_{ds}	Sample of respondents for data source ds
R_i	Response indicator — if statistical unit i responded to survey then $R_i = 1$, otherwise $R_i = 0$
s_{MP}	Sample from $\boldsymbol{\Omega}_{MP}$ to which we have access
s_y	Standard deviation of the target variable
s_ρ	Standard deviation of the response probabilities
$\tilde{\boldsymbol{u}}$	Predicted random effect
$\boldsymbol{v}$	Vector of paradata variables
$\boldsymbol{w}$	Vector of calibrated weights
$\boldsymbol{x}$	Vector of auxiliary variables

$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}$	Known total of auxiliary variable j
y	Target variable
$\overline{y}$	Mean of the target variable y
$\tilde{y}$	Matching estimator of target variable y
$\overline{y}_I$	Mean for the internet population
$\overline{y}_{NI}$	Mean for the non-internet population
$\overline{y}_{nr}$	Mean in the non-respondent stratum
$\overline{y}_{pst}$	Post-stratified estimate of the mean for the respondents
$\overline{y}_r$	Estimate of the mean for the respondents
θ	Target characteristic of target variable (e.g. mean, total)
θ_d	Target characteristic of target variable in domain $d = 1, \ldots, D$ (e.g. mean, total)
$\hat{\theta}$	Direct estimator of target characteristic
$\check{\theta}_{cd}$	Direct estimator based on big data source of θ_{cd} for domain $d = 1, \ldots, D$ with additional cross-classifications denoted by c, which could be sex, age or other characteristics that could be derived from big data sources
$\check{\theta}_d$	Direct estimator based on big data source of θ_d for domain $d = 1, \ldots, D$ available without any information on cross-classifications
ρ	Response probability (response propensity, propensity score)
ϱ_q	Asymmetric loss function
ψ_q	Asymmetric influence function
$\mathbf{\Omega}_{AP}$	Twitter ‘accounts’ population
$\mathbf{\Omega}_{BD}$	Big data population
$\mathbf{\Omega}_{IP}$	Internet population
$\mathbf{\Omega}_{MN}$	Population of mobile numbers
$\mathbf{\Omega}_{MP}$	Population that is observed within mobile frames
$\mathbf{\Omega}_{MU}$	Population of mobile phone users
$\mathbf{\Omega}_{PP}$	Twitter ‘posting’ population
$\mathbf{\Omega}_{RP}$	Register population
$\mathbf{\Omega}_{SE}$	Users of search engines population
$\mathbf{\Omega}_{SMP}$	Population with social media
$\mathbf{\Omega}_{TP}$	Target (statistical) population
$\mathbf{\Omega}_{TwP}$	Twitter population

A.2 Quantifying selectivity

In general, we can represent the missingness (or non-response) and coverage components of selectivity in big data sources as follows.[25] By s we denote sample and by r the set of respondents (or pseudo-respondents in the case of administrative and big data sources), and in both cases sample size n_s and n_r are of random size. We define a response indicator variable R_i as

$$R_i = \begin{cases} 1 & if\ i \in r\ (element\ i\ responds, i.e. is\ not\ missing) \\ 0 & if\ i \notin r\ (element\ i\ does\ not\ respond, i.e. is\ missing) \end{cases} \tag{72}$$

The probability that a given unit will respond, i.e. will not be missing (response propensity, propensity score) is given by

$$\rho_i = E(R_i = 1|\mathbf{x}, y) = \ P(R_i = 1|\mathbf{x}, y), \tag{73}$$

where x refers to auxiliary variables (e.g. demographics), v stands for paradata variables that reflect an individual's attitudes (e.g. number of Twitter followers) and y is the target variable. $I_i = 1$ refers to inclusion in the big data sample (which can be modelled in a similar manner as R_i when s is a non-probabilistic sample).

$$I_i = \begin{cases} 1 & if\ i \in s \\ 0 & if\ i \notin s \end{cases} \tag{74}$$

Rubin (1976) discusses three missing data mechanisms:

1. missing completely at random (MCAR);
2. missing at random (MAR);
3. missing not at random (MNAR);

The **MCAR** mechanism occurs when missingness is due to random events such as system failures and interruptions in the data collection process, which are not associated with either x, v or y. Under MCAR, the response probability (or non-missing probability) does not depend on $\mathbf{x}, \mathbf{v}$ and y:

$$P(R_i = 1|\mathbf{x}, y) = P(R_i = 1). \tag{75}$$

The **MAR** mechanism is connected with $\mathbf{x}$ and/or $\mathbf{v}$, but not the target variable y. MAR can also be the result of the sample selection process (conditional upon $I_i = 1$). Under MAR, the response probability is the expectation of the response indicator variable conditional on auxiliary variables, but not on the target variable itself:

$$P(R_i = 1|\mathbf{x}, y) = P(R_i = 1|\mathbf{x}). \tag{76}$$

In the **MNAR** mechanism, missingness is related not only to auxiliary variables, but also to the target variable y. Ignoring missingness in the presence of the MNAR mechanism may result in large biases and erroneous inferences (*Rubin, 1976, Pfeffermann, 2011*). The response probability under MNAR is the expectation of the response indicator (response propensity) variable conditional upon auxiliary variables and the target variable itself:

$$P(R_i = 1|\mathbf{x}, y) \neq P(R_i = 1|\mathbf{x}). \tag{77}$$

[25] The measurement error component of selectivity is not represented.

In practice, identification of all units that are members of the internet population denoted by $\boldsymbol{\Omega}_{IP}$ or the big data population denoted by $\boldsymbol{\Omega}_{BD}$ is unlikely. It is more realistic to assume that information is available on all units of the target population denoted by $\boldsymbol{\Omega}_{TP}$ and on the set of respondents denoted by r. When only $\boldsymbol{\Omega}_{TP}$ and r are available, expectations of R_i for the case are given by equations (78) and (79):

$$Pr(R_i = 1 \cap I_i = 1 | \mathbf{x}, \mathbf{v}, y) = Pr(R_i = 1 \cap I_i = 1 | \mathbf{x}, \mathbf{v}), \tag{78}$$

and

$$Pr(R_i = 1 \cap I_i = 1 | \mathbf{x}, \mathbf{v}, y) \neq Pr(R_i = 1 \cap I_i = 1 | \mathbf{x}, \mathbf{v}). \tag{79}$$

Another case is when a probabilistic sample denoted by s_{TP} taken from $\boldsymbol{\Omega}_{TP}$ or a population related to $\boldsymbol{\Omega}_{TP}$ and r is available. In the first case, we reduce the number of conditions in (1) and (2) by eliminating I_i, which results in the expectation of R_i, which for $\forall_{i\in r} R_i = 1$ and $\forall_{i\notin r} R_i = 0$.

Response propensities are expressed in a similar way as in (3) and (4) only when r and s_{TP} or the register population denoted by $\boldsymbol{\Omega}_{RP}$ are available. However, the definition of R_i is different and is given by the following equation:

$$R_i = \begin{cases} 1 & \text{if } i \in r, \\ 0 & \text{if } i \in s_{TP} \backslash r \text{ or } i \in \boldsymbol{\Omega}_{RP} \backslash r. \end{cases} \tag{80}$$

In equation (80), we realistically assume that $r \cap s_{TP} \neq \emptyset$ and $r \cap \boldsymbol{\Omega}_{RP} \neq \emptyset$. However, based on the literature, *Brick (2013)* stated that the MNAR model cannot be distinguished from its MAR counterpart based on the observed data and pointed to two papers on MNAR, stating that ‘Every missingness not at random model has a missingness at random counterpart with equal fit’ (*Molenberghs, Beunckens, Sotto, & Kenward, 2008*).

A.3 Bias quantification

To quantify errors connected with the two sources of selectivity, we introduce basic notation regarding bias and estimation. The basic definition of bias of estimator $\hat{\theta}$ is the difference between its expectation $E(\hat{\theta})$ and the true value of the characteristic target variable y given by θ (e.g. average income, number of trips). This relationship is expressed by the formula:

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right) \tag{81}$$

which reduces to

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \tag{82}$$

when θ is taken as a fixed (non-random) quantity. Mean square error (MSE) of $\hat{\theta}$ is the sum of variance of $\hat{\theta}$ given by $V(\hat{\theta})$ and the square of the bias of $\hat{\theta}$. MSE of $\hat{\theta}$ is defined by:

$$MSE(\hat{\theta}) = V(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2. \tag{83}$$

The coefficient of variation of $\hat{\theta}$ is the ratio of the square root of $MSE(\hat{\theta})$ with estimator $\hat{\theta}$:

$$CV(\hat{\theta}) = \frac{\sqrt{MSE(\hat{\theta})}}{\hat{\theta}}. \tag{84}$$

To properly assess bias, we should take into account that two main groups of estimators can be distinguished in the survey methodology – design-based and model-based. The first consists of estimators that are unbiased by definition as they use the randomisation principle; this states that the random sampling distribution is the only means by which valid inferences are made. As a result, bias of such an estimator is taken with respect to the sampling distribution (ς) instead of a model. Bias of a designed-based or model-assisted estimator of θ, assuming 100 % response, is then given by:

$$E_{\varsigma}(\hat{\theta}) - \theta, \tag{85}$$

where $\hat{\theta}$ can be estimated using either Horvitz-Thompson or a model-assisted estimator such as generalised regression (GREG) estimators. A model-assisted technique such as generalised regression estimation is one of many estimation methods used by design-based practitioners. GREG relies on a set of auxiliary variables to produce efficient survey estimates.

The characteristics of model-based estimators (e.g. BLUP, *best linear unbiased prediction* estimator) are derived with respect to a specified model. For example, the prediction bias of $\hat{\theta}$, which is evaluated based on the working model M^*, assuming 100 % response, is

$$E_{M^*}\left(\hat{\theta}_{BLUP} - \theta\right) = E_{M^*}\left(\hat{\theta}_{BLUP}\right) - E_{M^*}(\theta), \tag{86}$$

This type of inference is heavily dependent on M^* closely matching the true superpopulation model M. If the underlying model is correct, then this bias is equal to 0. However, if some other model, for example $\widetilde{\mathrm{M}}$, is correct (or at least a better description of y than M^*), then $\hat{\theta}_{BLUP}$ is biased.

BLUP by definition has the minimum model variance among the set of unbiased estimators — a desirable property for producing efficient estimates. Assuming correct model specification, BLUP will therefore be more efficient than a design-based estimator. However, if the underlying model is misspecified and the sample is large, bias might dominate the MSE of $\hat{\theta}_{BLUP}$. In that case, a design-based method tends to provide more efficient estimation.

Dever & Valliant (2006) studied model-assisted estimators — GREG, modified GREG using only respondents, and response-adjusted regression — and model-based (BLUP) estimators with ignorable and non-ignorable non-response. Their simulation study showed that the model-based estimation is characterised by the lowest relative bias even with non-ignorable non-response. However, as *Dever & Valliant (2006)* noted, results are limited to their simulation study where the target population was generated from a simple polynomial model and variables responsible for response were included in the modelling process. For more details on the model-based approach in survey sampling, see *Chambers & Clark (2012)*.

A 100 % response rate is often assumed in theory, but is rare in practice. Surveys and censuses face problems regarding coverage or non-response. Let us assume that we are interested in estimating an average denoted by $\bar{y}$ (e.g. average income).The Horvitz-Thompson estimator (which is unbiased by definition) of $\bar{y}$ then produces a biased estimation due to a non-sampling error, or non-response in this case. The bias is given by:

$$Bias(\bar{y}) \approx (1-p)(\bar{y}_r - \bar{y}_{nr}), \tag{87}$$

where p is the proportion of units in the respondent stratum, $\bar{y}_r$ is the mean in the respondent stratum, and $\bar{y}_{nr}$ is the mean in the non-respondent stratum. *Bethlehem (2010)* presented several cases and potential bias (for estimating an average) with non-sampling errors:

Bias due to coverage (in particular internet coverage)

$$Bias(\bar{y}) = \frac{N_{NI}}{N}(\bar{y}_I - \bar{y}_{NI}), \tag{88}$$

N_{NI}/N stands for the proportion of people without internet access. The bias is higher when a larger proportion of the population does not have access to the internet. $\bar{y}_I - \bar{y}_{NI}$ stands for the contrast $\bar{y}_I$ between the internet population and the non-internet population. The more the mean of the target variable differs for these two sub-populations, the larger the bias will be.

- Due to non-response

$$Bias(\bar{y}) = \frac{cor_{\rho,y} s_\rho s_y}{\bar{\rho}}, \tag{89}$$

where $\bar{\rho}$ is the mean of the response probability over all units in the sample. Furthermore, $cor_{\rho,y}$ is the correlation coefficient between the target variable and the response behaviour, s_y is the standard deviation of the target variable and s_ρ is the standard deviation of the response probabilities. *Bethlehem (2010)* stated that bias vanishes when:

- All response propensities are equal. Again, this is the case in which the self-selection process can be compared with a simple random sample.
- All values of the target variable are equal. This situation is very unlikely to occur in practice. No survey would be necessary. One observation would be sufficient.
- There is no relationship between the target variable and the response behaviour. It means that the participation does not depend on the value of the target variable.

In addition, *Zhang, Thomsen & Kleven (2013)* presented an approach to evaluate bias under the MNAR assumption for binary data. Denote by $q_{ij} = n_{ij}/n$ the gross sample proportion of $(x, y) = (i, j)$ for $i, j = 0.1$, and r_{ij} the corresponding within-group non-response rate (R=1, if statistical unit responded to the survey). The expected gross sample distribution is given below:

	Y = 1		Y = 0	
	R = 1	R = 0	R = 1	R = 0
X = 1	$n_r(1,1) = q_{11}(1 - r_{11})$	$q_{11}r_{11}$	$n_r(1,0) = q_{10}(1 - r_{10})$	$q_{10}r_{10}$
X = 0	$n_r(0,1) = q_{01}(1 - r_{01})$	$q_{01}r_{01}$	$n_r(0,0) = q_{00}(1 - r_{00})$	$q_{00}r_{00}$

Following *Zhang et al. (2013)*, the gross sample mean of the target variable Y is $p = q_{11} + q_{01}$, and the mean of the known auxiliary variable X is $q = q_{11} + q_{10}$. With non-response, the observed sample mean is given by $\bar{y}_r = [n_r(1,1) + n_r(0,1)]/n_r$, where $n_r(i,j)$ is the size of the corresponding respondent subsample $(X,Y) = (i,j)$ and n_r is the total number of respondents. Reweighting w.r.t X based on the MAR assumption $(f(r,y|x) = f(r|y)f(y|x))$ yields the gross sample mean $\bar{y}_{pst} = qn_r(1,1)/n_{r,1} + (1-q)n_r(0,1)/n_{r,0}$, where $n_{r,i} = n_r(i,1) + n_r(i,0)$ is the number of respondents where $X = i$.

The MNAR assumption implies that $r_{i0} = E(r_0)$ and $r_{i1} = E(r_1)$ for $i = 0{,}1$, with $r_0 \neq r_1$, and the bias of $\bar{y}_r$ (estimate for respondents) and $\bar{y}_{pst}$ (post-stratified estimate) can be evaluated. Although the biases themselves depend on the unknown quantities r_0 and r_1, it turns out that the ratio between them can be given as:

$$\frac{Bias(\bar{y}_r)}{Bias(\bar{y}_{pst})} = \frac{\dfrac{(E[n_r(0,1) + E[n_r(1,1)])(E[n_r(0,0)] + E[n_r(1,0)])}{E[n_r]}}{\dfrac{E[n_r(1,1)]E[n_r(1,0)]}{E[n_r,1]} + \dfrac{E[n_r(0,1)]E[n_r(0,0)]}{E[n_r,0]}} \tag{90}$$

which can be estimated from the observed sample alone. Combined with the observed $\bar{y}_r$ and $\bar{y}_{pst}$, we can then derive an estimate of the bias of $\bar{y}_{pst}$ for the MNAR assumption.

Zhang et al. (2013) also provided an approach for the K-category case, i.e. $X,Y = 1.2,\ldots,K$. For the MNAR assumption, we have $f(x|y) = f(x|y,r=1) = f(x|y,r=0)$, such that:

$$f(X=k) = \sum_{j=1}^{K} f\,(Y=j)f(X=k,y=j) = \sum_{j=1}^{K} f\,(Y=j)f(X=k|y=j,r=1). \tag{91}$$

Defining vectors $\mathbf{p} = (f(Y=1), f(Y=2), \ldots, f(Y=K))^T$ and $\mathbf{q} = (f(X=1), f(X=2), \ldots, f(X=K))^T$ and matrix $B_{k,k}$, whose (k,j)-th element is given by $P(X=k|y=j,r=1)$. We then have:

$$\mathbf{q} = \mathbf{Bp}. \tag{92}$$

Now that $\mathbf{q}$ is known and $\mathbf{B}$ can be estimated from the response subsample, we can obtain a method-of-moments estimator of $\mathbf{p}$ for the MNAR assumption and with it the bias of any estimator based on the alternative MCAR or MAR assumption.

A.4 Resampling with big data

This report considers two cases where resampling can be used with big data. First, resampling methods for estimation and testing that consist of methods such as jackknife and bootstrap (parametric, semiparametric, nonparametric, block). Second, resampling methods for model assessment and selection, in particular leave-one-out cross-validation and k-fold cross-validation. Spatial sub-sampling methods, which take into account the structure of the sampling region, are not taken into consideration here (*Nordman & Lahiri, 2004*).

Block bootstrap, a resampling method for estimation and testing with big data

Efron & Tibshirani (1986) developed the bootstrap technique for parametric inference for independent and identically distributed data. Originally, bootstrap did not take into account the fact that observations are correlated/clustered, and we know that big data sources contain paradata collected for the same units over time. As a result, we should include methods that account for correlated data.

The block bootstrap (*Chambers & Chandra, 2013*) extends this to take into account the hierarchical dependence structure of the clustered and multilevel data that are characteristic of random-effects models. The block bootstrap replications rely on level-specific empirical residuals to construct bootstrap samples. The method is simple to implement, free of both the distribution and dependence assumptions of the parametric bootstrap, and is consistent when the mixed model assumptions are valid. Results based on Monte Carlo simulation show that the method seems robust to failure of the dependence assumptions of the assumed mixed model:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \qquad (93)$$

The random effect block (REB) bootstrap algorithm (type = 0) is as follows:

- Calculate the nonparametric residual quantities for the fitted model:
  - marginal residuals: $\mathbf{e} = \mathbf{y} - \widehat{\boldsymbol{\beta}}\mathbf{X}$
  - predicted random effects: $\tilde{\boldsymbol{u}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e}$
  - error terms: $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{e} - \mathbf{Z}\,\tilde{\boldsymbol{u}}$
- Take a simple random sample with replacement of the groups and extract the corresponding elements $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\epsilon}}$
- Generate bootstrap samples via the fitted model equation $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\boldsymbol{u}} + \tilde{\boldsymbol{\epsilon}}$
- Refit the model and extract the statistic(s) of interest
- Repeat steps 2-4 B times.

Variation 1 (type = 1): The first variation of the REB bootstrap zero-centres and rescales the residual quantities prior to resampling.

Variation 2 (type = 2): The second variation of the REB bootstrap scales the estimates and centres the bootstrap distributions (i.e. adjusts for bias) after REB bootstrapping.

The main argument for bootstrap methods for mixed models, in particular a block bootstrap approach, is because of the correlation between observations in big data sources. Due to automatic data collection, we might have many data points for one unit and these units will also be correlated within domains (e.g. LAU 1). Taking into account such characteristics is crucial for correctly estimating standard errors when using models in selectivity correction methods.

Resampling for model validation

Resampling for model validation is common in machine learning, where it is used to assess how the results of a statistical analysis generalise to an independent dataset. In these procedures, we partition data into two datasets: (i) training data, used to estimate the model; and (ii) testing data, not used in the estimation of the model parameters and which is used to compare predictions based on the model with observed values. The selection of these datasets should be random and might involve stratification.

The most frequently used methods for sub-sampling (*Friedman, Hastie, & Tibshirani, 2001*) include:

- k**-fold cross-validation** — the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation.
- **Leave-p-out cross-validation** — involves using p observations as the validation set and the remaining observations as the training set. This is repeated on all possible ways of cutting the original sample on a validation set of p observations and a training set of the remaining observations.
- **Leave-one-out cross-validation** — is a particular case of leave-p-out cross-validation with $p = 1$. The process looks similar to jackknife. However, with cross-validation you compute a statistic using the left-out sample(s), whereas with jackknifing you compute a statistic using the kept samples only.

The application of sub-sampling used in machine learning could be applied when building models. This approach might provide insight into which variables should be used, what is the prediction error is as well as point to appropriate methods either for self-selection adjustments or imputation.

KS-TC--18-004-EN-N

An overview of methods for treating selectivity in big data sources

The official statistics community is now seriously considering big data as a significant data source for producing statistics. It holds the potential for providing faster, cheaper, more detailed and completely new types of statistics. However, the use of big data also brings several challenges. One of them is the non-probabilistic character of most sources of big data, as very often, they were not designed to produce statistics. The resulting selectivity bias is therefore a major concern when using big data. This paper presents a statistical approach to big data, searching for a definition meaningful from the statistical point of view and identifying their main statistical characteristics. It then argues that big data sources share many characteristics with Internet opt-in panel surveys and proposes this as a reference to address selectivity and coverage problems in big data. Coverage and the self-selection process are briefly discussed in mobile network data, Twitter, Google Trends and Wikipedia page views data. An overview of methods which can be used to address selectivity and eliminate, or mitigate, bias is then presented, covering both methods applied at individual level, i.e. at the level of the statistical unit, and at domain level, i.e. at the level of the produced statistics. Finally, the applicability of the methods to the several big data sources is briefly discussed and a framework for adjusting selectivity in big data is proposed.

For more information
http://ec.europa.eu/eurostat/

Publications Office

ISBN 978-92-79-88769-7