*Survey Research in Times of Big Data**

Investigación con encuestas en los tiempos del big data

Pablo Cabrera-Álvarez

Institute for Social and Economic Research de la Universidad de Essex
Universidad de Salamanca
pablocal@usal.es (ESPAÑA)
ORCID: https://orcid.org/0000-0001-8105-5908

Recibido: 14.09.2020
Aceptado: 19.12.2021

ABSTRACT

Although surveys still dominate the research landscape in social sciences, alternative data sources such as social media posts or GPS data open a whole range of opportunities for researchers. In this scenario, some voices advocate for a progressive substitution of survey data. They anticipate that big data, which is cheaper and faster than surveys, will be enough to answer relevant research questions. However, this optimism contrasts with all the quality and accessibility issues associated with big data such as the lack of coverage or data ownership and restricted accessibility. The aim of this paper is to explore how, nowadays, the combination of big data and surveys results in significant improvements in data quality and survey costs.

KEY WORDS

Survey methodology, big data, administrative data, social media data, data linkage.

RESUMEN

La encuesta es la técnica de investigación predominante en la investigación en Ciencias Sociales. Sin embargo, la aparición de otras fuentes de datos como

* El proyecto que ha generado estos resultados ha contado con el apoyo de una beca de la Fundación Bancaria "la Caixa" (ID 100010434), cuyo código es LCF/BQ/ES16/11570005

las publicaciones en redes sociales o los datos generados por GPS suponen nuevas oportunidades para la investigación. En este escenario, algunas voces han defendido la idea de que, debido a su menor coste y la velocidad a la que se generan, los big data irán sustituyendo progresivamente a los datos de encuesta. Sin embargo, este optimismo contrasta con los problemas de calidad y accesibilidad que presentan los big data como la fata de cobertura de algunos grupos de la población o el acceso restringido a alguna de estas fuentes. Este artículo, a partir de una revisión profunda de la literatura de los últimos años, explora como la cooperación entre los big data y las encuestas resulta en mejoras significativas de la calidad de los datos y una reducción de los costes.

PALABRAS CLAVE

Metodología de encuestas, big data, datos administrativos, datos de redes sociales, combinación de datos.

1. INTRODUCTION

We are in the era of big data. Every minute Twitter users post 511,000 tweets, people send more than 188 million of emails, and Google processes more than four million searches (DOMO 2019). All these actions leave a digital trace behind; it may be a log of metadata, the content of the publication itself or the reactions to it. These traces are stored and constitute a potential data source for research. This flood of granular and cheap data is possible thanks to the technological developments that allow the storage and processing of the data. The rise of big data contrasts with the complex and costly process of survey data collection, the dominant paradigm in the social sciences since Neyman (1934) published his work about inference from probability samples. Parallel to the rise of big data, it is the appearance of substantial challenges in the field of survey research. Two of these challenges are the general drop of response rates in the last decades and the expansion of internet data collection methodologies.

Nowadays, while surveys still dominate the research landscape in social sciences, alternative data sources such as social media posts or GPS data open a whole range of opportunities for researchers. In this scenario, some voices advocate for a progressive substitution of survey data. They anticipate that big data, which is cheaper and faster than surveys, will be enough to answer relevant research questions. However, this optimism contrasts with all the quality and accessibility issues associated with big data such as the lack of coverage or data ownership and restricted accessibility.

The aim of this paper is not to discuss the future of survey research or the potential of big data. Instead, it explores how, nowadays, the combination of big data and surveys results in significant improvements in data quality and survey costs. In other words, the central postulate of this paper is that surveys and big

data together can attain what neither of them could get on their own. This paper, which does not pretend to be exhaustive, presents a selection of the latest advancements in the field of survey research that involves the use of big data to highlight the benefits of combining both data sources.

The first section of the paper presents a definition and a typology of big data focused on survey research. The second section discusses whether big data can substitute surveys and the benefits of combining both types of data, while the third explores the caveats of linking surveys and big data. The next section outlines some of the cutting-edge advancements in the field of survey research that involve surveys and big data. Finally, the paper concludes with some reflections about the present and future of survey research.

2. THE CONCEPT OF BIG DATA

Big data is an ample term used in different contexts such as academia, business, or media. Although the notion of a large volume of data is common to most definitions, the scholars have not reached a consensus about the ground characteristics of big data (Ward and Barker 2013). One of the first and most extended definitions focuses on three features: velocity, volume, and variety (Laney 2001). Velocity because big data production occurs at a high rate; volume refers to the necessary large size of the data that generally cannot be processed by a single machine; and variety is related to the unstructured format of the data that is raw and typically requires a substantial effort to clean and format. Another early, however less known, definition of big data covers all interactions among individuals, institutions, and things that are recorded and stored digitally (Negroponte et al. 1997). This definition highlights two main characteristics, the origin of big data, interactions, and the critical role of technology development and digital capacity.

A typology of big data focused on survey research helps to establish the scope of this review (Callegaro and Yang 2018; Japec et al. 2015). This typology aligns with the one presented by Baker (2017) that splits big datasets into three groups based on the source of the data. These groups are transaction data, data derived from the internet of things, and social media data.

Transaction data refers to the records produced in the context of an interaction. These interactions can involve individuals and organisations, including the public administration. For instance, a person who applies for a benefits scheme generates a record that is stored in a governmental database. Also, these records contain an ID, like the social insurance number or the passport number, that enables to link it with other data sources. This system of linking several transaction datasets allows building high dimensional data that is of maximum interest for social research. In the context of survey research, there are three subcategories of transaction data that are especially valuable: administrative data, commercial data, and paradata.

Every interaction between a governmental agency and an individual leaves a trace that is recorded and stored, and can be used for research (Playford et al. 2016). However, answering research questions is not the primary objective of this information, that generally pursues to smooth the management and evaluation of the programs (Baker 2017; Woollard 2014). This second life of administrative data can benefit social research by supplementing surveys with direct measurements.

Using administrative records has several advantages in contrast to survey. First, the use of administrative records prevents the impact of measurement error (Connelly et al. 2016; Künn 2015). Administrative data tend to be more accurate than surveys, especially if the questions inquire about the respondent's past or socially desirable behaviours. Second, the fact that administrative records are produced and stored over time enables the use of a longitudinal perspective in research (Connelly et al. 2016). A unique identifier allows gathering the records for the same individual over time, which contrasts with data collected using cross-sectional surveys, that frequently suffer from recall errors.

Furthermore, administrative records reduce the level of respondents' burden by shortening the questionnaire. The interviewer can skip a substantial number of questions if the interviewee agrees to match, for instance, their school or tax records (Connelly et al. 2016). Concerning the coverage of the population, administrative records are exhaustive since all citizens need to be in the registers (Hand 2018). This advantage is especially relevant when the study aims to cover small subgroups of the population, such as the patients affected by a rare disease.

Commercial datasets combine records from different sources such as administrative registers, survey data, and transaction databases owned by companies. This information is employed in the context of marketing analysis to perform market segmentation and drive campaigns (Peytchev and Raghunathan 2013). An example of this is the Experian database in the United States, which is made up of more than 3,500 public and proprietary sources (Pasek et al. 2014).

Finally, paradata are by-products generated during the survey interaction (Kreuter 2013). These by-products include calling records, interviewers' observations, questionnaire timestamps, or navigation logs of web surveys. This type of data, which is specific to the field of survey research, is useful to monitor and refine the survey process.

Internet of Things (IoT) covers the digital traces captured by sensors and other interconnected devices (Gerschenfeld, Krikorian and Cohen 2004). Ten years ago, some scholars suggested that the appearance of smartphones would have a decisive impact on data collection (Lazer et al. 2009; Raento, Oulasvirta and Eagle 2009). Technology has developed and spread in the population. In Europe (UE-27), Eurostat (2016) estimates that 77% of the residents are smartphone users, while in some countries such as Sweden, this percentage reaches 92%. This transformation has enabled the collection of passive data (Stier et al. 2019), an opportunity to collect granular and rich information while reducing respondent burden.

Social media is another source of data that has gained prominence in the last few years. Social media are a set of digital platforms that allow interacting by storing and delivering information (Murphy, Hill and Dean 2013). Every day, the majority of the population log on their social media accounts, share information, and interact with other users. Eurostat data shows that 56% of the EU-27 aged 16-74 participated in social media in 2018. Also, the emergence of social media in the last 20 years has democratised content production by giving people the ability to publish and generated the concept of online community (Scott and Jacka 2012). In the field of social research, this data allows observing social interaction in an unintrusive manner.

3. CAN BIG DATA REPLACE SURVEYS?

Some scholars have depicted a dark future for surveys anticipating that other data sources will substitute them (Savage and Burrows 2007). The deficits of surveys reinforce this idea. Surveys are facing a changing environment where the expansion of the internet has opened the door to new methods of data collection, while the response rates are declining (De Leeuw, Hox and Luiten 2018). Certainly, the emergence of the internet is the opportunity to use web surveys to speed up and lower the costs of data collection. Yet a substantive part of the population, normally older people and form disadvantaged households, does not have internet access, which complicates the selection of probability samples and the inference process for general population surveys (Elliott and Valliant 2017).

The challenges faced by surveys run in parallel to the optimism generated by the rise of big data. Big data might be adequate to answer some questions in the area of social research, but this is not the case on most of the occasions. Even when the use of big data shows results comparable to surveys, there is a counterpart. For instance, Tusmajan and his colleagues (2011) made an accurate prediction of the 2009 general election in Germany based on the frequencies of party mentions on Twitter. Nonetheless, those who tried to replicate this method in a different context obtained inaccurate results (Gayo-Avello 2012).

One of the main issues affecting social media and other sources of digital trace data is selection bias (Hsieh and Murphy 2017; Schober et al. 2016). Selection bias occurs when a part of the target population is not present in the data. This lack of coverage is not an issue if the target population is restricted to those using a social media platform or owning smartphones. However, most of the studies in social research do not focus on these groups, and, on many occasions, aim to cover the general population. The case of Twitter users in Great Britain serves to illustrate this. In Great Britain, the people with a Twitter account are younger and had higher qualifications than the average (Sloan 2017). Similar deviations are found when comparing the general population to smartphone users. The use of smartphones correlates with some sociodemographic characteristics such as age and education (Keusch et al. 2020; Jäckle et al. 2019; Wenz, Jäckle

and Couper 2019). In contrast, selection bias is less of an issue for administrative records, which tend to cover most of the population.

Big data also present measurement issues. Sometimes, the definition of the concepts in social research exceeds the formulations used in big data sources (Hsieh and Murphy 2017). Hand (2018) uses a simple example to illustrate this issue. For some time, the trends from the British Crime Survey and the police records evolved in opposite directions due to the use of different definitions. Furthermore, apart from using other concepts, big data is not exempt from measurement error. In a recent publication, Bähr and his colleagues (2020) show that geolocation sensor data suffer from different sources of error such as the manufacturer and operating system settings, research design, third-party apps, and the participants' behaviour.

Big datasets tend to have a high number of cases, but few covariates (Couper 2013). This scarcity of covariates is not an issue if the objective is to estimate a single figure. However, most of the time, social research is about exploring relationships between variables, and relevant covariates are required. Linked to this issue is the lack of attitudinal or sociodemographic measures in most big data sources (Salganik 2017). However, an essential part of social research focuses on attitudinal data. Certainly, social media data can be used to derive attitudinal measures, or administrative records tend to capture demographics. Still, sometimes, the fact that the research team cannot control big data production imposes severe limitations. Moreover, digital trace and social media data can suffer from a lack of stability (Schober et al. 2016). Social media platforms tend to change over time and even disappear. Other issues are the access and privacy policies. Most of the time, big data sources are proprietary and access, therefore, is restricted (Couper 2013).

The flaws of big data make implausible to think of a near future without surveys. Likewise, the use of surveys is not exempt from issues and challenges. The need for accurate statistics imposes the collaboration of surveys and big data. They can cooperate to overcome their imperfections by building an enhanced data environment. Some scholars have openly advocated for this combination as a form of refining survey data quality (Forsyth and Boucher 2015; Miller 2017; Kalton 2019). Others have shown that surveys can help to improve the quality of big data (Rafei, Flannagan and Elliott 2020; Kim and Tam 2020). Either way, both approaches support the idea that surveys and big data can achieve together what they cannot accomplish on their own. In the last few years there have been efforts form industry and academia to test different combinations of big and survey data and new spaces have emerged to enhance these new research streams such as the BigSur18 and BigSur20 conferences organised by the European Survey Research Association (Hill et al. 2019).

4. COMBINING BIG DATA AND SURVEYS

There are several approaches to combine big data and surveys. The method to be used depends on the characteristics of the datasets. The first relevant factor is whether the records in the databases belong to the same entity. The second requirement is the existence of a set of variables that uniquely identify the entities in the datasets.

Meeting the two conditions allow performing a one-to-one linkage using deterministic or probabilistic methods. However, if the datasets contain information from different elements or the identification is not practicable, a model-based approach can be employed. This model-based strategy, which is called statistical matching or data fusion, aims to match records based on a set of common characteristics. Alternatively, a statistical model can translate the information from one data source to the other. This is the case of techniques like imputation, small area estimation, and hierarchical models (Lohr and Raghunathan 2017). Finally, this paper also considers a case of combination that consists of applying big data related methods such as machine learning and artificial intelligence to survey research.

The most extended form of data matching is the deterministic linkage. This method requires a set of unique identifiers for each case in all the datasets. Typical examples of unique identifiers are the national insurance number, the passport ID, or the employee code. This type of matching is not exempt from errors since the registers might be outdated, or the identifiers of some records can contain errors. In these cases, the use of a probability approach can help to increase the number of matches if some merging variables contain errors (Calderwood and Lessof 2009). Sometimes, even though the data sources cover the same elements, the variables do not uniquely identify them or contain errors. Then, the use of a probabilistic approach can enable the data matching. The probabilistic matching relies on an algorithm and a set of quasi-identifiers which are variables that can identify pairs of cases with some probability such as surname, date of birth, or address.

However, beyond the technical details, there are legal and ethical barriers to perform a one-to-one deterministic or probability linkage. First, the research team needs to have access to the databases. The previous section addressed the issues that imposes the proprietary nature of most of the big data sources. The most obvious implication for research is that the data collected by companies is not normally usable while the governments have the obligation to preserve the privacy of individuals and organizations. Only some countries have systems in place to perform data matching in a secure environment. Second, survey respondents must give their informed consent to the data linkage. The data linkage process is also critical for data quality since the differences between those accepting and refusing to share their information might bias the estimates. This issue has attracted the attention of some researchers. For instance, some studies showed that agreeing to link administrative is related to the respondents' cognitive skills, trust in the survey organisation, or privacy concerns in the data linkage request

(Jäckle et al. 2018; Sala, Burton and Knies 2013; Sakshaug et al. 2012). Also, in relation to smartphone and social media data, some studies have detected a significant resistance to share personal data (Baghal et al. 2019; Revilla, Couper, and Ochoa 2019).

The use of model-based approaches allows combining surveys and big datasets in situations where the records belong to different entities. This approach includes statistical matching, imputation, small area estimation, and the use of hierarchical methods (Lohr and Raghunathan 2017). In contrast to the deterministic and probability record linkage methods, statistical matching is used to merge records that belong to different entities based on a set of characteristics present in both datasets (Moriarity and Scheuren 2001). Another method of combining big data and surveys is imputation. In this approach, a statistical model is built to predict a target variable using a set of covariates shared by all datasets. Then, the model is employed to predict the values in the datasets where this measure is missing (Carpenter and Kenward 2012). In small area estimation, administrative data and surveys join forces to produce statistical estimates for small areas such as census tracks or population subgroups. This method helps to estimate summary statistics where survey estimates would be imprecise due to the small sample size by combining the prediction from a model of the statistic for the subgroup and the estimate from the survey data (Rao and Molina 2015; Fay and Herriot 1979). Similarly, hierarchical models are also used to synthesise summary statistics or individual records. These models allow combining estimates from different studies or the individual records nested in the studies (Cooper, Hedges and Valentine 2019).

Finally, another approach consists of using machine learning and artificial intelligence to treat surveys. The use of machine learning is being extended to some areas of survey research, such as the calculation of response propensities for the computation of non-response weights (Buskirk 2018; Kern, Klausch and Kreuter 2019). These tools help to solve classical problems of survey research more efficiently. Likewise, the use of artificial intelligence is also helping to improve the efficiency of tasks such as the generation of sample frames using satellite images and gridded population data (Chew et al. 2018).

5. DEVELOPMENTS USING BIG DATA AND SURVEYS

The synergy between surveys and big data can have a variety of purposes. This section presents some developments in which the combination of both sources leads to survey enhancement, measurement improvements, solve issues related to representativeness, or facilitate fieldwork management.

5.1. Big data to enhance surveys

The most recurrent case of synergy between surveys and big data consists of supplementing the survey with covariates from other sources. This merge generates a joint dataset that broadens the scope of the survey or improves the quality of the measures. This approach has been used in the area of official statistics for a long time. This is the case of the Census Longitudinal Study in England and Wales. This study, which started in 1971, links census records and administrative data about vital events for a sample of 500,000 individuals. In recent years, the number of studies that use a form of big data to enhance surveys is growing (e.g. Biddle et al. 2019; Dissing et al. 2021; Möller et al. 2019).

Eady and his colleagues (2019) researched social media consumption to establish whether people tend to live in online bubbles where they only receive insights from ideologically aligned users. To answer this question, they used a dataset ensembled by YouGov, which contained survey and Twitter data. The representative sample of Twitter users was linked to the content of the accounts they followed. The final dataset, which comprised 1,496 survey respondents and 1,2 billion of tweets from 642,345 accounts, allowed them to replicate respondents' timelines.

Cornwell and Cagney (2017) used smartphones to research the mobility of older adults. They wanted to assess whether the elderlies spend their time in their neighbourhood or have a more extensive area of movement. For this research, they selected a convenience sample of 60 elderlies in New York City and equipped them with smartphones. The devices were programmed to send the GPS location every five minutes for four days. Besides, the GPS measures were supplemented by an initial questionnaire and a set of ecological momentary assessments, which are short questionnaires about the location, experiences, and activities. This information allowed them to track the movements of the sample through granular and accurate information collected using GPS.

Meyer and Mittag (2019) provide an example of how linking a survey with administrative data affects the quality of economic related measures. They merged the sample of the Current Population Survey (2008-2013) from New York with administrative records from benefits programs including information about the amounts received. First, they compared the administrative and the survey reports data in order to assess the impact of measurement error. Then, they analysed the effect of government transfers on the level of deprivation using administrative data instead of survey reports. They found out that respondents on low income tend to misreport the amount of money received from government transfers. This also affected the assessment of the program, which had a more significant impact than what survey data analysis showed.

5.2. Big data and surveys together to tackle measurement error

Measurement error occurs when the response in the questionnaire differs from the actual characteristic of the sample unit (Groves et al. 2013). This phenomenon has several causes, such as recall mistakes or response modifications due to judgement. The latter is especially striking in measures of attitudes and behaviours affected by social desirability. Alterations in the question position, wording, or response categories may help to reduce the impact of measurement error. However, a benchmark is needed to evaluate the level of bias in the responses. Big data sources are adequate to perform a measure validation. Indeed, the combination of survey data and administrative records for this purpose is not a new idea (Ferber et al. 1969; Parry and Crossley 1950). Also, related to measurement, machine learning emerges as an alternative to ease the coding of open-ended questions.

The appearance of big data offers a genuine opportunity to assess whether a survey accurately measures the population characteristics. One illustrative case is the use of administrative records to research on the causes of electoral turnout overestimation in surveys. In some countries, the public administration keeps a record of those who voted in the elections which can be linked to survey responses. The scholars have examined two hypotheses that could explain the turnout overestimation. The first is related to a deliberate misreporting in which some respondents hide their intention of not participating in the election. The second, which covers the effects of sample selection, states that those with lower levels of interest in the elections are less likely to take part in the survey. The use of administrative records offers an opportunity to test these hypotheses by linking the individual records from the voting files with the survey (Ansolabehere and Hersh 2012; Selb and Munzert 2013; Enamorado and Imai 2019).

Sometimes surveys are assumed to be the benchmark to validate big data measures. Hersh (2015) performed such an exercise to validate the variable race contained in the Catalist database, one of the commercial databases used in American politics to organise electoral campaigns and target voters. This database is composed of several sources being the voter records the most important. However, the electoral legislation, which is different for each State, shapes the availability of data at the individual level. Some states do not collect information about the race at the registration stage. Therefore, to fill-in that variable, Catalist employs an imputation algorithm using other variables. This research validated the race variable in the Catalist database by linking the American National Election Study.

The use of big data for validation is not restricted to transaction data. In the last years, scholars are using data collected from smartphones, sensors, and social media to evaluate the accuracy of survey self-reports (e.g. Boase and Ling 2013; Scharkow 2016). Vraga and Tully (2018) compared the self-reports of news consumption with behavioural data tracked using a web analytics software. Haenschen (2018) combined Facebook and survey data to assess how the self-reports about social media usage departs from reality. Henderson and his collea-

gues (2019) replicated this exercise using Twitter. They collected data from a subsample of adults in the US who had Twitter and asked them for permission to link their responses to their Twitter data. They investigated how engagement with the social platform affects the accuracy of self-reports.

However, using big data to validate survey measures present some drawbacks. Jürgens, Stark and Magnin (2019) identified three types of biases that affect this type of validation analysis, sample selection, tracking device selection, and data generation errors. The first, sample selection, refers to the composition of the sample and the possible deviations with respect to the target population. The tracking device selection refers to the fact that those accepting to cooperate with the data collection, which generally involves downloading and installing and application, may differ from those not taking part. Finally, during the data generation, technical issues may arise, or the individuals' behaviour may change due to the awareness about the tracking device. In this study, for instance, they show that overestimation of survey self-reports is more likely to happen when the tracker application is on the smartphone rather than on the desktops or laptops.

Coding open-ended questions is another field of survey research where the emergence of big data is having an impact. In this case, the innovations do not come from the use of new data sources but the development of machine learning techniques (Gweon et al. 2017). Schonlau and Couper (2016) combined human coders, the use of text mining techniques, and multinomial boosting —a type of machine learning model— to classify the responses to open-ended questions. In this experiment, human coders classified a random sample of responses to train the machine learning model before applying the algorithm to the rest of the sample. The authors, who applied the methodology to two surveys, found that half of the responses can be classified automatically with an accuracy of 80%. However, the performance of the algorithms also depends on how accurately the human coders classify the training set of responses (He and Schonlau 2019). Also, some experiments have used unsupervised topic modelling in which the algorithm joins the responses based exclusively on their content, and no human coding is needed. Pietsch and Lessman (2018) tested different machine learning models, including Latent Feature Latent Dirichlet Allocation, Biterm Topic Model, and Word Network Topic Model. They concluded that research could benefit from these techniques for topic exploration in some instances.

5.3. Survey representativeness and inference for big data

The possibility of having access to massive volumes of data also entails an opportunity to study the issues related to representativeness in survey research. Sampling can use GPS and satellite images to outperform the quality of traditional sampling frames in certain contexts. Administrative records can be useful to research the effects of non-response by describing those not taking part in the survey. Additionally, big data can benefit from the techniques employed to infer

from nonprobability surveys, and surveys can rely on big data sources to adjust survey estimates.

The dominant framework in survey research establishes that a random selection of elements is necessary to infer the characteristics from the sample to the population. Drawing a probability sample requires a sampling frame – a full list of the population elements, such as the census or other administrative registers. However, sometimes administrative records are not up to date, do not exist, or are unreliable, as it happens in some developing countries. In such cases, gridded population data works as a valid alternative to more traditional sampling frames. The generation of a gridded population dataset consists of splitting up the territory where the target population lives in little squares and calculating a population count for each. The population counts are computed using models that combine administrative data, spatial covariates, and satellite images (Thomson et al. 2017). In some cases, when the administrative data is unreliable, other data such as mobile phone connection logs can be used. Besides, machine learning and deep learning techniques are being used to classify satellite images and generate the population counts of each grid cell (Stevens et al. 2015; Chew et al. 2018).

The study of the effects of non-response can also benefit from the use of big data sources. The problem of investigating the effects of non-response has to do with the lack of information about those not taking part in the survey. The use of big data allows observing nonrespondents' characteristics and assessing whether the survey estimates are biased. For this purpose the use of administrative records is especially useful (McMinn et al. 2019; Sakshaug and Eckman 2017). Other researches have experimented linking administrative aggregate data based on geographical identifiers (Biemer and Peytchev 2012).

Another area of synergy is the application of the evidence accumulated in survey research to infer from nonprobability samples. There is some parallelism between nonprobability samples and some forms of big data. In the field of survey methodology, the inference from nonprobability surveys is a topic that has gained relevance in the last decades due to the rise of internet data collection methods (Baker et al. 2013). There are two main strategies that, based on statistical models, are used to infer from nonprobability samples (Valliant 2019). The first is quasi-randomisation in which a statistical model is used to calculate the pseudo probabilities of selection for the elements in the nonprobability sample. The second strategy is based on superpopulation models. Furthermore, in the last years, the expansion of Bayesian models has helped to develop new methods to combine survey and big data to boost model-based inference (Gelman 2007; Mercer 2018). For example, Wang and colleagues (2015) used data from Xbox users to forecast the 2012 US presidential elections. The original data was heavily skewed towards males and young people, however, using a multilevel regression with poststratification model, it was possible to rebalance the sample and make an accurate forecast of the vote.

Some examples show how surveys and big data can be used together to produce estimates. The Directorate-General for Regional and Urban Policy of the European Commission (2019) carried out a feasibility study to assess whether

some economic activity indicators for urban areas in Germany can be estimated using aggregate mobile data and the Labour Force Survey (LFS). For this, they employed small area estimation. The model was built using the LFS data while the aggregate mobile data were the covariates used to generate the area estimates. Klingwort and his colleagues (2019) combined the Dutch Road Freight Transport Survey, the Dutch vehicle and enterprise registers, and weigh-in-motion road sensor data to correct the bias of surveys estimates about the number of transportations and the total weight.

Machine learning techniques are being used to improve the models that adjust surveys after data collection. The computation of survey adjustments is another field where machine learning can be used to improve the results by substituting the traditional parametric models (Buelens, Burger and van den Brakel 2018). Chen and his colleagues (2018) used an adaptative LASSO model to compute calibration weights that performed slightly better than the traditional linear calibration model. Likewise, Ferri-García and Rueda (2020) compared the performance of logistic regression and some machine learning techniques such as Random Forests, GBM, k-Nearest Neighbours, and Naïve Bayes to compute the pseudo probabilities of selection. They showed through a set of simulations that using machine learning techniques outperform the traditional logistic regression model.

5.4. Fieldwork applications using paradata

Paradata, the by-products generated during the survey process, such as the call records or the questionnaire completion times, are of great interest in tracking and adjusting the fieldwork process. The use of paradata can help to improve contact rates, monitor representativeness during the fieldwork, generate data quality indicators, and study non-response (Kreuter 2013). Recently, other sources like GPS are also contributing to improving the data collection process by adding meaningful information to fieldwork monitoring.

The emergence of paradata has fostered the generation of quality indicators to monitor survey data collection. These indicators are the base for survey responsive designs. A responsive design monitors a set of process and quality indicators and alters survey design features during the data collection to improve survey cost efficiency and the quality of the estimates (Groves and Heeringa 2006). The National Survey of Family Growth in the United States, for instance, achieved a significant increase in the number of completed interviews, from 12,500 (2002-2003) to 22,500 (2006-2010), partially by using paradata to inform fieldwork decisions (Kirgis and Lepkowski 2013). The management team used paradata like interviewer observations and call records to build a response propensity model. The predicted response propensities were used to select the cases more likely to respond, which were reissued during the second phase of fieldwork. Focusing the interviewers' efforts on these cases caused a rise of response rates while limiting the costs.

Some recent applications of paradata also include the prediction and adjustment of panel attrition in a web panel survey (Roßmann and Gummer 2015). This research assessed whether some paradata like past participation history or response times are helpful to predict response propensities and adjust the sample. The analysis concluded that some paradata such as response times or history call help to predict panel attrition. Similar findings found Durrant and her colleagues (2017) using data from the United Kingdom Household Longitudinal Survey, a face-to-face longitudinal survey. Other authors used paradata to study interviewer effects and how these effects influence interviewing quality (Sharma 2019). Similarly, another study used GPS data to track interviewers travel behaviour in a face-to-face survey in order to establish the potential of this information to improve fieldwork management (Olson and Wagner 2015).

6. FINAL THOUGHTS

Along these pages I have discussed the qualities and issues associated with big data, some of the challenges faced by survey research, and the potentials arising from the cooperation between these two worlds. Here are some reflections emerging from this review.

Big data has been around for a while and, for example, the use of administrative records to validate survey measures goes back to the fifties of the past century. Big data is part of a broader technological change. The capacity to store and process data has increased exponentially in the last decades and so has done the data available for research purposes. Some of these sources, like administrative data, are not new to survey research; however, others, like satellite images or social media posts, offer excellent opportunities to refine data quality and enhance social research.

There is no expectation that big data will be able to substitute surveys in the near future. This paper outlined some of the limitations associated with big data sources —the inability to cover the whole population, the instability of some data sources, or the measures definitions—. These barriers are not minor issues given that the main characteristic of surveys is a double inference process: the extrapolation of sample characteristics to the population and the inference from individuals' responses to respondents' characteristics. However, there are some cases in which big data sources could substitute surveys in the area of social research. An example is the use of administrative records to reduce the burden on respondents and the survey costs. But, even in these cases, surveys are still necessary if the objective is to explore the relationships between variables.

Other barriers that prevent big data from substituting surveys are the lack of access to most of the databases and the legal and ethical requirements to perform the data linkage. Big data does not mean open data since companies and governments are in control of the data sources. In this scenario, researchers must generate synergies with the database owners to access data or await that the company implements a data-access policy. Either way, the scientific community

needs to identify the relevant datasets for research and advocate for an open-data framework.

However, data access is not only a matter of property. There are legal and ethical requirements in place to preserve the rights of the citizens and organizations who origin the data. The process of data linkage needs to ensure that citizens anonymity is guaranteed. Given the increasing importance of big data, authorities need to develop flexible systems that allow the use of data while protecting citizens' rights. Some countries, which have a long tradition using data for policy research, such as the United Kingdom, have developed institutions that control data quality, release, and access in the area of social sciences. This is another area of work to foster the use of big data in the context of survey research.

Also, when using big data, we need to acknowledge that they are prone to error. This paper presents some researches devoted to assessing the quality of some big data sources. For example, Bähr and his colleagues (2020) have developed a framework to identify the possible sources of error when dealing with sensor data. This research is utterly necessary before we adopt any source of big data. Besides, it is the responsibility of the research team to think about the definition of the big data concepts and the data generation process to anticipate possible issues at the analysis stage.

Despite all these warnings, the emergence of big data is already an opportunity for survey research. Survey researchers are working to integrate the opportunities of big data into the field. One of the purposes of this review was to illustrate this process. However, to utilise the potential of this cooperation, we should go beyond combining data and techniques from both fields. Data scientists and programmers need to gain presence in survey teams, and the survey methodologists need to understand what these roles can bring to the field. Overall, this exciting time of change is opening new opportunities to improve data quality and reduce costs. Big data has come to foster the future of survey research.

Nevertheless, this cooperation should not be a one-way transaction from data science to survey research. The developments from the field of survey methodology are also valuable in many data science projects. Survey researchers have thought for a long time about issues such as the inference from nonprobability samples or the best way to measure complex constructs. Data science can benefit from all this knowledge. It is not a coincidence, for instance, that companies that mostly work with big data, such as Facebook, incorporate survey methodologists in their teams. Again, the cooperation between these two worlds brings new opportunities to shape the future of data collection methodologies.

7. BIBLIOGRAPHY

AL BAGHAL, T., SLOAN, L., JESSOP, C., WILLIAMS, M. L., BURNAP, P. (2019): "Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies", Social Science Computer Review.

ANSOLABEHERE, S., HERSH, E. (2012): "Validation: What big data reveal about survey misreporting and the real electorate", Political Analysis, 20, 4, 437–459.
BÄHR, S., HAAS, G.-C., KEUSCH, F., KREUTER, F., TRAPPMANN, M. (2020): "Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data", Social Science Computer Review.
BAKER, R. (2017): Big Data. In: Total Survey Error in Practice. John Wiley & Sons, Inc., Hoboken, NJ, USA, 47–69.
BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. J., TOURANGEAU, R. (2013): "Summary report of the aapor task force on non-probability sampling", Journal of Survey Statistics and Methodology, 1, 2, 90–105.
BIDDLE, N., BREUNIG, R., MARKHAM, F., WOKKER, C. (2019): "Introducing the Longitudinal Multi-Agency Data Integration Project and Its Role in Understanding Income Dynamics in Australia", Australian Economic Review, 52, 4, 476–495.
BIEMER, P. P., PEYTCHEV, A. (2012): "Census geocoding for nonresponse bias evaluation in telephone surveys", Public Opinion Quarterly, 76, 3, 432–452.
BOASE, J., LING, R. (2013): "Measuring Mobile Phone Use: Self-Report Versus Log Data", Journal of Computer-Mediated Communication, 18, 4, 508–519.
BUELENS, B., BURGER, J., VAN DEN BRAKEL, J. A. (2018): "Comparing Inference Methods for Non-probability Samples", International Statistical Review, 2, 86, 322–343.
BUSKIRK, T. D. (2018): "Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research", Survey Practice, 11, 1, 1–13.
CALDERWOOD, L., LESSOF, C. (2009): Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn, P. (ed.): Methodology of Longitudinal Surveys. John Wiley & Sons, Ltd, Chichester, UK, 55–72.
CALLEGARO, M., YANG, Y. (2018): The Role of Surveys in the Era of "Big Data." In: The Palgrave Handbook of Survey Research. Springer International Publishing, Cham, 175–192.
CARPENTER, J., KENWARD, M. (2012): Multiple Imputation and its Application.
CHEN, J. K., VALLIANT, R. L., ELLIOTT, M. R. (2018): "Model-assisted calibration of non-probability sample survey data using adaptive LASSO", Survey Methodology, 44, 1, 117–145.
CHEW, R. F., AMER, S., JONES, K., UNANGST, J., CAJKA, J., ALLPRESS, J., BRUHN, M. (2018): "Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery", International Journal of Health Geographics, 17, 1, 1–17.
CONNELLY, R., PLAYFORD, C. J., GAYLE, V., DIBBEN, C. (2016): "The role of administrative data in the big data revolution in social science research", Social Science Research, 59, 1–12.
COOPER, H., HEDGES, L. V., VALENTINE, J. C. (2019): The Handbook of Research Synthesis and Meta-Analysis. Russell Sage Foundation.
CORNWELL, E. Y., CAGNEY, K. A. (2017): "Aging in activity space: Results from smartphone-based GPS-tracking of urban seniors", Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 72, 5, 864–875.
COUPER, M. P. (2013): "Is the sky falling? New technology, changing media, and the future of surveys", Survey Research Methods, 7, 3, 145–156.

DE LEEUW, E. D., HOX, J. J., LUITEN, A. (2018): "International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data", Survey Methods: Insights from the Field, 1–11.

DISSING, A. S., ROD, N. H., GERDS, T. A., LUND, R. (2021): "Smartphone interactions and mental well-being in young adults : A longitudinal study based on objective high-resolution smartphone data", Scandinavian Journal of Public Health, 49, 3, 325–332.

DOMO (2019): Data never sleeps, https://www.domo.com/learn/data-never-sleeps-6.

DURRANT, G. B., MASLOVSKAYA, O., SMITH, P. W. F. (2017): "Using prior wave information and paradata: Can they help to predict response outcomes and call sequence length in a longitudinal study?", Journal of Official Statistics, 33, 3, 801–833.

EADY, G., NAGLER, J., GUESS, A., ZILINSKY, J., TUCKER, J. A. (2019): "How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data", SAGE Open, 1, 9.

ELLIOTT, M. R., VALLIANT, R. (2017): "Inference for Nonprobability Samples", Statistical Science, 32, 2, 249–264.

ENAMORADO, T., IMAI, K. (2019): "Validating Self-Reported Turnout by Linking Public Opinion Surveys with Administrative Records", Public Opinion Quarterly, 83, 4, 723–748.

EUROPEAN COMMISSION (2019): City data from LFS and Big Data.

EUROSTAT (2016): Internet use by individuals. https://ec.europa.eu/eurostat/documents/2995521/7771139/9-20122016-BP-EN.pdf/f023d81a-dce2-4959-93e3-8cc7082b6edd

FAY, R. E., HERRIOT, R. A. (1979): "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", Journal of the American Statistical Association, 366a, 74, 269–277.

FERBER, R., FORSYTHE, J., GUTHRIE, H. W., MAYNES, E. S. (1969): "Validation of a National Survey of Consumer Financial Characteristics: Savings Accounts", The Review of Economics and Statistics, 436–444.

FERRI-GARCÍA, R., DEL MAR RUEDA, M. (2020): "Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys", PLoS ONE, 15, 4, 1–19.

FORSYTH, J., BOUCHER, L. (2015): "Why Big Data Is Not Enough", Research World, 50, 2015, 26–27.

GAYO-AVELLO, D. (2012): ""I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data", CoRR.

GELMAN, A. (2007): "Struggles with survey weighting and regression modeling", Statistical Science, 22, 2, 153–164.

GERSCHENFELD, N., KRIKORIAN, R., COHEN, D. (2004): "The Sevenfold Way", Scientific American, 291, 4, 76–81.

GROVES, R. M., FOWLER, F. J., JR., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E., TOURANGEAU, R. (2013): Survey Methodology, John Wiley & Sons.

GROVES, R. M., HEERINGA, S. G. (2006): "Responsive design for household surveys: tools for actively controlling survey errors and costs", Journal of the Royal Statistical Society: Series A (Statistics in Society), 169, 3, 439–457.

GWEON, H., SCHONLAU, M., KACZMIREK, L., BLOHM, M., STEINER, S. (2017): "Three methods for occupation coding based on statistical learning", Journal of Official Statistics, 33, 1, 101–122.

HAENSCHEN, K. (2018): "Self-Reported Versus Digitally Recorded: Measuring Political Activity on Facebook", Social Science Computer Review.

HAND, D. J. (2018): "Statistical challenges of administrative and transaction data", Journal of the Royal Statistical Society. Series A: Statistics in Society, 181, 3, 555–605.

HE, Z., SCHONLAU, M. (2019): "Automatic Coding of Text Answers to Open-Ended Questions: Should You Double Code the Training Data?", Social Science Computer Review, 1–12.

HENDERSON, M., JIANG, K., JOHNSON, M., PORTER, L. (2019): "Measuring Twitter Use: Validating Survey-Based Measures", Social Science Computer Review, 1–21.

HERSH, E. D. (2015): Hacking the electorate: How campaigns perceive voters, Cambridge University Press.

HILL, C. A., BIEMER, P. P., BUSKIRK, T. D., CALLEGARO, M., CORDOVA CAZAR, A. L., ECK, A., JAPEC L., KIRCHNER, A., KOLENIKOV, S., LYBERG, L.E., STURGIS, P. (2019): "Exploring new statistical frontiers at the intersection of survey science and Big Data: Convergence at 'Bigsurv18.'", Survey Research Methods, 13, 1.

HSIEH, Y. P., MURPHY, J. (2017): "Total Twitter Error", en Total Survey Error in Practice, Wiley & Sons, 23–46.

JÄCKLE, A., BENINGER, K., BURTON, J., COUPER, M. P. (2018): "Understanding data linkage consent in longitudinal surveys", Understanding Society Working Paper Series, University of Essex.

JÄCKLE, A., GAIA, A., LESSOF, C., COUPER, M. P. (2019): "A review of new technologies and data sources for measuring household finances: Implications for total survey error", Understanding Society Working paper Series, University of Essex.

JAPEC, T. F. M. I. L., KREUTER, F., BERG, M., BIEMER, P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C., USHER, A. (2015): "AAPOR Report on Big Data", American Association for Public Opinion Research.

JÜRGENS, P., STARK, B., MAGIN, M. (2019): "Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data", Social Science Computer Review, 1–16.

KALTON, G. (2019): "Developments in Survey Research over the Past 60 Years: A Personal Perspective", International Statistical Review, 87, S1, S10–S30.

KERN, C., KLAUSCH, T., KREUTER, F. (2019): "Tree-based machine learning methods for survey research", Survey Research Methods, 13, 1, 73–93.

KEUSCH, F., BÄHR, S., HAAS, G. C., KREUTER, F., TRAPPMANN, M. (2020): "Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey", Sociological Methods and Research.

KIM, J., TAM, S.-M. (2020): "Data Integration by combining big data and survey sample data for finite population inference", International Statistical Review, 1–30.

KIRGIS, N. G., LEPKOWSKI, J. M. (2013): Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006-2010 National Survey of Family Growth. In: Improving Surveys with Paradata. John Wiley & Sons, Inc., Hoboken, New Jersey, 121–144.

KLINGWORT, J., BUELENS, B., SCHNELL, R. (2019): "Capture–Recapture Techniques for Transport Survey Estimate Adjustment Using Permanently Installed Highway-Sensors", Social Science Computer Review.
KREUTER, F. (2013): Improving Surveys with Paradata: Analytic Uses of Process Information. John Wiley & Sons.
KÜNN, S. (2015): "The challenges of linking survey and administrative data", IZA World of Labor.
LANEY, D. (2001): "META Delta", Application Delivery Strategies.
LAZER, D., BREWER, D., CHRISTAKIS, N., FOWLER, J., KING, G. (2009): "Life in the network: the coming age of computational social science", Science, 5915, 323, 721–723.
LOHR, S. L., RAGHUNATHAN, T. E. (2017): "Combining Survey Data with Other Data Sources", Statistical Science, 32, 2, 293–312.
MCMINN, M. A., MARTIKAINEN, P., GORMAN, E., RISSANEN, H., HÄRKÄNEN, T., TOLONEN, H., LEYLAND, A. H., GRAY, L. (2019): "Validation of non-participation bias methodology based on record-linked Finnish register-based health survey data: A protocol paper", BMJ Open, 9, 4, 1–6.
MERCER, A. W. (2018): Selection Bias in Nonprobability surveys: a causal inference approach, Doctoral dissertation, University of Maryland, College Park.
MEYER, B. D., MITTAG, N. (2019): "Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net", American Economic Journal: Applied Economics, 11, 2, 176–204.
MILLER, P. V. (2017): "Is There a Future for Surveys?", Public Opinion Quarterly, 81, 205–212.
MÖLLER, J., VAN DE VELDE, R. N., MERTEN, L., PUSCHMANN, C. (2019): "Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit?", Social Science Computer Review.
MORIARITY, C., SCHEUREN, F. (2001): "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure", Journal of Official Statistics, 3, 17, 407.
MURPHY, J., HILL, C. A., DEAN, E. (2013): Social Media, Sociality, and Survey Research. In: Social Media, Sociality, and Survey Research. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1–33.
NEGROPONTE, N. HARRINGTON, R., MCKAY, S. R., CHRISTIAN, W. (1997): "Being digital", Computers in Physics, 11, 3, 261–262.
NEYMAN, J. (1934): "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", Journal of the Royal Statistical Society, 97, 4, 558.
OLSON, K., WAGNER, J. (2015): "A feasibility test of using smartphones to collect GPS information in face-to-face surveys", Survey Research Methods, 9, 1, 1–13.
PARRY, H. J., CROSSLEY, H. M. (1950): "Validity of responses to survey questions", Public Opinion Quarterly, 14, 1, 61–80.
PASEK, J., JANG, S. M., COBB, C. L., DENNIS, J. M., DISOGRA, C. (2014): "Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data", Public Opinion Quarterly, 78, 4, 889–916.
PEYTCHEV, A., RAGHUNATHAN, T. (2013): "Evaluation and Use of Commercial Data for Nonresponse Bias Adjustment", American Association for Public opinion Research annual conference.
PIETSCH, A.-S., LESSMANN, S. (2018): "Topic modeling for analyzing open-ended survey responses", Journal of Business Analytics, 2, 1, 93–116.

PLAYFORD, C. J., GAYLE, V., CONNELLY, R., GRAY, A. J. J. G. (2016): "Administrative social science data: The challenge of reproducible research", Big Data and Society, 3, 2, 1–13.
POLIDORO, F., GIANNINI, R., CONTE, R. Lo, MOSCA, S., ROSSETTI, F. (2015): "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation", Statistical Journal of the IAOS, 31, 2, 165–176.
RAENTO, M., OULASVIRTA, A., EAGLE, N. (2009): "Smartphones: An emerging tool for social scientists", Sociological Methods and Research, 37, 3, 426–454.
RAFEI, A., FLANNAGAN, C. A. C., ELLIOTT, M. R. (2020): "Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using bayesian additive regression trees", Journal of Survey Statistics and Methodology, 8, 1, 148–180.
RAO, J. N. K., MOLINA, I. (2015): Small Area Estimation: Second Edition. John Wiley & Sons, Inc, Hoboken, NJ, USA.
REVILLA, M., COUPER, M. P., OCHOA, C. (2019): "Willingness of online panelists to perform additional tasks", Methods, Data, Analyses, 13, 2, 223–251.
ROSSMANN, J., GUMMER, T. (2015): "Using Paradata to Predict and Correct for Panel Attrition", Social Science Computer Review, 34, 3, 312–332.
SAKSHAUG, J. W., COUPER, M. P., OFSTEDAL, M. B., WEIR, D. R. (2012): "Linking Survey and Administrative Records", Sociological Methods & Research, 41, 4, 535–569.
SAKSHAUG, J. W., ECKMAN, S. (2017): "Are survey nonrespondents willing to provide consent to use administrative records? Evidence from a nonresponse follow-up survey in Germany", Public Opinion Quarterly, 81, 2, 495–522.
SALA, E., BURTON, J., KNIES, G. (2013): "Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics", Sociological Methods & Research, 41, 3, 414–439.
SALGANIK, M. J. (2017): Bit by Bit: Social Research in the Digital Age. Princeton University Press.
SAVAGE, M., BURROWS, R. (2007): "The Coming Crisis of Empirical Sociology", Sociology, 41, 5, 885–899.
SCHARKOW, M. (2016): "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data", Communication Methods and Measures, 10, 1, 13–27.
SCHOBER, M. F., PASEK, J., GUGGENHEIM, L., LAMPE, C., CONRAD, F. G. (2016): "Social Media Analyses for Social Measurement", Public Opinion Quarterly, 80, 1, 180–211.
SCHONLAU, M., COUPER, M. P. (2016): "Semi-automated categorization of open-ended questions", Survey Research Methods, 10, 2, 143–152.
SCOTT, P. R., JACKA, M. (2012): Auditing Social Media. John Wiley & Sons, Inc., Hoboken, NJ, USA.
SELB, P., MUNZERT, S. (2013): "Voter overrepresentation, vote misreporting, and turnout bias in postelection surveys", Electoral Studies, 32, 1, 186–196.
SHARMA, S. N. (2019): "Paradata , Interviewing Quality , and Interviewer Effects", Doctoral Dissertation.
SLOAN, L. (2017): "Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015", Social Media + Society, 3, 1.

STEVENS, F. R., GAUGHAN, A. E., LINARD, C., TATEM, A. J. (2015): "Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data", PLoS ONE, 10, 2, 1–22.

STIER, S., BREUER, J., SIEGERS, P., THORSON, K. (2019): "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field", Social Science Computer Review.

THOMSON, D. R., STEVENS, F. R., RUKTANONCHAI, N. W., TATEM, A. J., CASTRO, M. C. (2017): "GridSample: An R package to generate household survey primary sampling units (PSUs) from gridded population data", International Journal of Health Geographics, 16, 1, 1–19.

VALLIANT, R. (2019): "Comparing Alternatives for Estimation from Nonprobability Samples", Journal of Survey Statistics and Methodology, 1–33.

VRAGA, E. K., TULLY, M. (2018): "Who Is Exposed to News? It Depends on How You Measure: Examining Self-Reported Versus Behavioral News Exposure Measures", Social Science Computer Review.

WANG, W., ROTHSCHILD, D., GOEL, S., GELMAN, A. (2015): "Forecasting elections with non-representative polls", International Journal of Forecasting, 31, 3, 980–991.

WARD, J. S., BARKER, A. (2013): "Undefined By Data: A Survey of Big Data Definitions", arXiv preprint arXiv:1309.5821.

WENZ, A., JÄCKLE, A., COUPER, M. P. (2019): "Willingness to use mobile technologies for data collection in a probability household panel", Survey Research Methods, 13, 1, 1–22.

WOOLLARD, M. (2014): Administrative Data: Problems and Benefits: A perspective from the United Kingdom. SCIVERO, Berlin