Scuola di Dottorato
Università degli Studi di Milano-Bicocca

Department of

Economics, Management, and Statistics

Ph.D. program: **Economics and Statistics** Cycle: **XXXIV°**
Curriculum: **Statistics**

Essays on Inference for Non-probability Samples and Survey Data Integration

Surname: **Salvatore**
Name: **Camilla**
Registration number: **839357**

Supervisor: Prof. **Silvia Biffignandi**
Tutor: Prof. **Pietro Giorgio Lovaglio**
Coordinator: Prof. **Matteo Manera**

Academic Year: 2021-2022

Acknowledgments

I am grateful to my supervisor Silvia Biffignandi for her guidance and supervision throughout the course of my thesis, contributing to the conceptualization of the research and the discussion of the results. Since the beginning of my master's program, she has been a mentor and an inspiration, sparking my interest in statistics and giving me invaluable advice, suggestions, and opportunities that have contributed greatly to my growth as a researcher. I am truly grateful for her constant support and mentorship.

I would express my gratitude to Joseph Sakshaug, Arkadiusz Wiśniowski and Bella Struminskaya for their contribution to the supervision of the thesis, in particular for the conceptualization and the discussion of the paper in Chapter 3. It was a true pleasure to work with them, and they had a significant impact on my academic development. Thank you for all the time and effort you invested in me.

I greatly appreciate the support and expertise of Annamaria Bianchi in developing the paper included in Chapter 4 of my thesis. Throughout my research journey, she has been a constant source of support, providing me with valuable guidance from the start of my master studies. Collaborating with Annamaria and Silvia on the development of several papers was an enriching experience. Their support and expertise made the research process enjoyable and successful.

I am also grateful to the Utrecht University team, and in particular Bella, for making me feel welcomed and providing me with an excellent research environment. The warm and friendly environment made me truly enjoy my time in Utrecht.

I would also like to extend my gratitude to my family and friends who have supported me throughout my PhD journey. Their support and encouragement have been a constant source of motivation and strength for me, and I am deeply grateful for their love and support.

Contents

Chapter 1

Outline of the thesis

Probability sample (PS) surveys are considered the traditional gold-standard data source for studying socio-economic phenomena. As a function of the sampling design they allow for population inferences with measurable bounds of uncertainty (Neyman, 1934). They are structured with well-defined quality and methodological frameworks for dealing with sampling and non-sampling (e.g. coverage, measurement, non-response) errors (Biemer, 2010).

Although PS surveys are considered the gold standard for population inference, they are facing difficulties due to declining response rates and related increasing costs. Fielding large size probability samples can be cost prohibitive for many survey researchers and study sponsors. Thus, moving towards less expensive, but potentially biased, non-probability data is becoming a more common practice (Cornesse et al., 2020).

Non-probabilistic data sources include both traditional administrative data and more innovative data like volunteer web surveys and digital trace data (e.g. social media, google trends, data donation packages). While the use of administrative data is well-established in survey research, volunteer web surveys and especially digital trace data are gaining popularity due to their cost-effectiveness and ability to provide timely information (Couper, 2013; Nordbotten, 2010). These data sources can offer new perspectives and insights on various phenomena that cannot be studied using traditional data only, opening up new opportunities for making population inferences and supplementing traditional data (Japec et al., 2015).

While non-probabilistic data sources offer many advantages, they also come with limitations. For example, digital trace data are often unstructured (e.g. in the form of text or images) and require additional analysis to extract the information of interest. Moreover, drawing inference from non-probability samples is challenging because of the absence of a known sampling frame and random selection process. Additionally, there is no unique framework for evaluating their quality, and the lack of a benchmark measure can be a problem when studying new phenomena. Furthermore, it is important to evaluate the construct being measured, as it may be different from the one measured by traditional data sources (e.g. social media sentiment and sentiment measured on a Likert scale in a questionnaire).

Thus, from a statistical perspective, there are many challenges and research questions that need to be addressed, such as the possibility of doing inference with non-probabilistic data, the quality of these data, and whether these data sources can replace or supplement traditional PS surveys.

The focus of this thesis is on answering three research questions: 1) What is the evolution of the field and what new trends are emerging?, 2) Can probability and non-probability samples be combined in order to improve analytical inference and reduce survey costs?, and 3) How can traditional and digital trace data be combined to augment the information in traditional sources and better describe complex phenomena?

This thesis addresses the aforementioned three research questions and, in particular, contributes to the existing literature by a) providing a deeper understanding of the literature, and identifying current trends and research gaps for future investigations, b) developing an original Bayesian framework to combine probability and non-probability online surveys in a manner that improves analytic inference while also reducing survey costs, and c) developing a modular framework that allows for building composite smart indicators in order to augment the information available in traditional sources through digital trace data.

Each of the three research questions is addressed in one of the chapters.

Chapter 2 introduces the topic of inference for non-probability samples and survey data integration. The chapter provides an overview of the different types of non-probability samples and discusses their strengths and limitations. It also examines the challenges of integrating survey data with data from multiple sources and provides examples of how researchers are addressing these challenges. After providing a conceptual background, the chapter presents an original science mapping study using text mining and bibliometric tools, which makes this analysis different from other literature reviews. The study analyzes a collection of 1023 topic-related publications from the Scopus database published between the years 1937-2022. The originality of this paper lies in the fact that in addition to characterizing the field in terms of collaboration between authors and research trends, it also identifies research gaps and formulates a research agenda for future investigations. Hence, it addresses the first research question. This paper is currently under review. The paper was revised following the reviewers' comments, after receiving minor revisions.

Then, Chapter 3 and 4 address the problem of data integration and data augmentation. Each chapter focuses on a different type of data: structured and more *traditional* volunteer web surveys in Chapter 3 and unstructured textual data from social media (Twitter) in Chapter 4.

Chapter 3 is the result of a research project carried out with my supervisor and some international scholars who co-supervised me, including Joseph Sakshaug (Institute for Employment Research, Germany), Arkadiusz Wiśniowski (University of Manchester, UK), and Bella Struminskaya (Utrecht University, Netherlands). The focus of our work is on analytic inference, which is a topic rarely addressed in the literature, as it appears evident in the second Chapter. A similar framework has been developed for the analysis of continuous data (Sakshaug et al., 2019; Wiśniowski et al., 2020). We extend the methodology to account for binary outcome variables and we also provide a cost-analysis as an original contribution.

In order to address the second research question, the paper presents a novel Bayesian approach to integrate a small probability sample with a larger online non-probability sample (possibly affected by selection bias) to improve inferences about logistic regression coefficients and reduce survey costs. The approach can be applied in different contexts. We provide examples from socioeconomic contexts (volunteering, voting behavior, trust) as well as health contexts (smoking, health insurance coverage). Through the simulation and the real-life data analysis we show that the MSEs of regression coefficients are generally lower when implementing data integration with respect to the case of no data integration. Also, using assumed probability and non-probability sample costs, we show that potential cost savings are evident.

This work is accompanied by an interactive online application (Shiny App). In line with the reproducibility principle of open science, the app includes the replication code as well as additional insights into the research results. The app is designed to make it easier for researchers to apply or adapt the proposed framework to their research needs. A key feature of the app is its interactive cost-analysis tool. By entering probability and non-probability (per-unit) sample costs, researchers are able to compare different scenarios of costs. These results can be used as a reference for survey researchers interested in collecting and integrating a small probability sample with a larger non-probability one. This paper is currently under review. The paper was revised following the reviewers' comments, after receiving major revisions.

Chapter 4 is the result of the collaboration with my supervisor and Annamaria Bianchi (University of Bergamo). This study addresses the third research question, showing how digital trace data can be used to augment traditional data, thus feeding smart statistics. Our focus is on business statistics. The study begins with reviewing the main characteristics of traditional and novel business statistics sources. Then, an original general framework is developed to combine traditional and digital trace based indicators. It can be applied to new data sources and their integration with traditional ones. This framework is modular and it is composed of three layers, each describing the steps necessary for the technical construction of a smart indicator. The modularity of the framework is a key feature, as it allows for flexibility in its application. In fact, researchers can use the framework to explore different methodological variants within the same architecture, and potentially carry out improvements to specific modules or test for sensitivity of the results obtained at the different levels. In the second part of the paper, the methodology is illustrated through a practical exercise on the construction of a prototype indicator to measure the commitment of businesses to sustainability. This paper is currently under review. The paper was revised following the reviewers' comments, after receiving major revisions.

Finally, conclusions of this research are drawn in Chapter 5.

Bibliography

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3):145–156.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'neil, C., and Ushe, A. (2015). Big data in survey research: Aapor task force report. *The Public Opinion Quarterly*, 79(4):839–880.

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):558–625.

Nordbotten, S. (2010). The use of administrative data in official statistics-past, present and future: with special reference to the nordic countries. *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, page 205–223. Available at https://officialstatistics.wordpress.com/.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics*, 35(3):653–681.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1):120–147.

Chapter 2

Analysis of the Literature

Paper under review:
Salvatore C., Inference for Non-probability Samples and Survey Data Integration: Mapping the Literature through Text Mining Approaches. Revised After Minor Revision.

1 Introduction

The field of survey research has experienced a profound transformation since the end of the 1990s due to the opportunity to use new data sources to make population inferences or to be integrated with traditional surveys (Couper, 2013). Data integration is not new to survey researchers, who have already combined surveys based on probability-based samples (PS) with auxiliary data from censuses or administrative registers to enhance inference. However, as a result of technological progress and people's changing interaction with technologies, a variety of new data sources have become available, and their use for inferential purposes poses new challenges as well as opportunities.

Probabilistic surveys are designed to provide unbiased, accurate, and reliable population statistics. However, in practice, unbiasedness can be undermined by various factors, such as non-coverage, nonresponse, and other sources of error, as described by the Total Survey Error (TSE) framework (Biemer, 2010). Since the early 1980s, nonresponse, in particular, has increased significantly, primarily because of an increase in non-contacts and refusals (Luiten et al., 2020). Consequently, a rethinking of incentives strategy and increased fieldwork efforts have raised survey costs to the point that many organizations can no longer undertake large and prohibitively expensive PS surveys.

Starting from the 2000s, volunteer web surveys and (big) digital trace data (textual data from social media, Google searches and maps, sensor data, etc.) have become popular data sources that can potentially replace or be integrated with traditional PS surveys. In general, they provide a more convenient and timely source of information for understanding complex social phenomena (Japec et al., 2015). However, their non-probabilistic nature poses inferential and statistical challenges. The following paragraphs present three of these challenges, which will be discussed in more detail in Section 2.

The first challenge is selection bias that arises from the lack of a known selection mechanism and from the self-selection of individuals. Consequently, additional effort is required so that the estimates can be generalized. A second concern is the possibility that measurements of a particular construct may differ depending on the survey mode and characteristics of the auxiliary data sources. For instance, differences in measurement may arise when considering two surveys conducted in different modes (e.g., face-to-face vs. online) or one survey and a big data source (e.g., answers to a Likert scale vs. social media sentiment). As a third consideration, the quality of the data may also differ. Accordingly, ad-hoc quality and error frameworks need to be developed for each auxiliary source.

As a result of the above concerns, it is unlikely that data from non-probability samples (NPS) will replace traditional probabilistic surveys. However, supplementing a probabilistic survey with such auxiliary data is an appealing way to enhance inference while reducing the survey costs and respondent burden. The variety of these digital data requires more research on methodological aspects to address the statistical challenges mentioned above, as well as, applications to understand the potential benefits of building multi-source statistics. In particular, there are two main research streams (Rao, 2021). The first stream of research focuses on inference based on NPS (addressing quality issues and correcting selection bias using PS surveys). The second research stream aims to statistically integrate NPS with PS surveys. In both cases, a central assumption is a high-quality PS survey.

This study aims to provide an overview of the current state of research in survey data integration and inference for non-probability samples. For this purpose, we analyze a selection of publications related to that topic using text mining and bibliometric techniques. In terms of a bibliographic database, we consider Scopus. This database allows the collection of document metadata such as the title, year of publication, journal, authors, and abstract. As opposed to other literature reviews, the originality of this study lies in the use of bibliometric and text mining tools. These tools allow us to analyze a greater number of papers, identifying current research trends, and to suggest future research directions.

The paper is organized as follows. Section 2 provides the literature background and the context of this study. The objectives of the research and the data are presented in Section 3. Section 4 describes the methodology. A detailed discussion of the results can be found in Section 5. In conclusion, Section 6 outlines a research agenda and identifies remaining research gaps to be addressed.

2 Conceptual Background

This section focuses on two aspects. Firstly, it describes the context of this work which is essential in order to critically evaluate the results of our study which will provide further insights. Secondly, it reviews the methodological literature in light of the three statistical challenges described in Section 1.

It is becoming increasingly common for researchers and statistical institutes to integrate data and make inferences based on non-probabilistic samples. As a complement to survey data, administrative registers have often been used throughout history, and in recent decades they have played a key role in the production of official statistics (Nordbotten, 2010; Kreuter et al., 2010). However, the frontier of data integration and inference relates to three relatively new data sources: volunteer web surveys, big or digital trace data, and mobile data collection (Couper, 2013).

Volunteer web surveys and opt-in panels were developed during the second half of the 1990s but gained popularity only ten years later (Baker et al., 2010; Biffignandi and Bethlehem, 2021), especially for market research and public opinion studies. Even though hundreds (or thousands) of questionnaires can be filled out online in a relatively short time, concerns remain about the generalizability of the results to the general population due to the self-selection of individuals (Bethlehem, 2010). As a result, several methodologies have been developed to address coverage and selectivity issues.

Big or digital trace data are defined as digital data generated by human interaction and systems (e.g., sensor data, social media, google trends, transactions, etc.). They are not generated for statistical purposes (also known as organic data, see Groves 2011), but they can allow for measuring new phenomena (Stier et al., 2020). Since 2010, they have become increasingly popular in social science, mainly due to the diffusion of social media, which are particularly relevant to better understanding attitudes and behaviors (Ceron et al., 2016; Iacus and Porro, 2016). Also, statistical institutes are engaged in the production of experimental statistics based on big data (Daas et al., 2015). There are, however, selectivity and measurement issues that cannot be ignored, as demonstrated by the Google Flu experiment, which initially appeared promising but then failed to predict outbreaks (Lazer et al., 2014).

Finally, mobile data collection is directly linked to big data and developed in the last few years. Mobile surveys involve filling out surveys on, for example, tablets and smartphones, and collecting data using devices' sensors (e.g., photos, geolocation sensors, accelerometer, etc.). A benefit of sensor data is that it potentially provides objective data free from errors commonly associated with self-reports (Struminskaya et al., 2020). However, participation is voluntary and individuals decide whether and which data to share (Struminskaya et al., 2021).

Despite their differences, all three sources share the property of not being probabilistic. Nevertheless, given the variety of these data, the three statistical challenges (inference in presence of selection bias, measurement issues and quality aspects) described in Section 1 need to be addressed separately. Although the literature in this field is expanding rapidly, it is still limited. The following paragraphs present some of the studies addressing such issues.

Amaya et al. (2020) and Sen et al. (2021) explain how the Total Error Framework can be adapted to different big data sources. As for social media data, Salvatore et al. (2021) present a quality framework for Twitter data, while Amaya et al. (2021) address statistical issues related to Reddit data. An error framework for web-tracking data is presented by Bosch Jover and Revilla (2022). The opportunities and challenges associated with supplementing survey data with data from sensors and applications are discussed by Struminskaya et al. (2020).

Issues in representation and measurement when augmenting surveys with auxiliary data are addressed by Stier et al. (2020) and Braun and Kuljanin (2015). Einarsson et al. (2022) and Baker et al. (2010) also discuss measurement errors and mode effects in the context of online opt-in panels.

Despite the limited literature about data quality, error frameworks, and construct measurement, several studies focus on statistical inference in the presence of selection bias. Many traditional review articles have discussed the use of different inferential approaches to correct selection bias and integrate multi-source data. A comprehensive review of inference for non-probability samples has been published, for the first time, by Baker et al. (2013). In addition to reviewing the various non-probability sampling techniques, they also cover estimation and weight adjustment methods as well as considerations concerning the quality of the data.

Considering both missing-at-random (MAR) and missing-not-at-random (MNAR) selection mechanisms, Elliott and Valliant (2017) describe three methods of estimation from non-probability samples: quasi-randomization, superpopulation modeling, and doubly robust estimation. The authors provide a discussion of the respective advantages and disadvantages. The effectiveness of such approaches is then examined through the use of a simulation study in Valliant (2020).

Rao (2021) and Beaumont and Rao (2021) also review estimation methods, emphasizing data integration and demonstrating how big data can enhance small area estimation. Finally, Cornesse et al. (2020) review the empirical evidence of using NPS for inference, suggesting under which conditions it is possible to obtain the highest accuracy. More recently, review studies focused on machine learning and bayesian methods for data integration (Tsung et al., 2018; Breidt and Opsomer, 2017; Little, 2015).

The themes discussed above are expected to emerge from our analysis, as well as new insights regarding thematic evolution, potential applications and new research areas. The following section provides a detailed description of the research objectives.

3 Research Objectives and Data

3.1 Research Objectives

In contrast to the previous studies, this article offers an alternative and original perspective and situates itself within the discipline of science mapping. We consider a larger number of publications and, using bibliometric and text mining techniques, we are able to map the literature, providing an updated *big picture* of the field in terms of the research community and topics development. A comprehensive longitudinal analysis is conducted to identify research patterns and trends.

In particular, this study addresses the following research objectives (RO):

RO1. To understand the annual growth of the scientific production

RO2. To identify the most productive authors, the driving research groups, the leading outlets for publication, and in which topics authors are specialized (performance and social structure)

RO3. To explore the conceptual structure of the field

RO4. To understand the evolution of the conceptual structure over the years (thematic evolution)

Based on the results of our analysis, we identify the research gaps and the emerging topics. Thus, the ultimate goal of the study is the following:

RO5. To outline and provide practitioners with a research agenda for future investigations

3.2 Data

Bibliographic information can be retrieved from various databases, including Scopus, Web of Science (WoS), and Google Scholar. We consider the Scopus database. Compared to WoS, it has a more comprehensive list of publications. Further, it provides search and API tools for extracting data, resulting in higher quality data than Google Scholar, which is the most extensive database. Also, Google Scholar does not allow to define as specific and advanced search queries compared to Scopus and WoS.

A two-step retrieval strategy is used. First, a search query is formulated in order to retrieve publications about methodologies for data integration and statistical analysis of non-probability samples. The resulting list of documents is manually inspected in order to remove out-of-scope publications and keep and only topic-relevant documents. We refer to them as *seed* publications. Secondly, the dataset is expanded by selecting both cited and citing documents. This selection strategy aims to maximize the topic relevance and time coverage. In this way, the selected dataset's analysis should mirror the field's development.

The search query is based on the presence of keywords in the title and abstract, plus restrictions on language and subject area. Only publications (journal articles, conference proceedings, books, etc.) written in English and in the Mathematical field are considered. Appendix A discusses the keywords used to extract the publications in greater detail. Such keywords are identified based on the conceptual background outlined in Section 2.

The number of papers extracted by the query is 77, out of which 43 are considered as *seed* publications. With the inclusion of cited and citing publications, the full dataset accounts for 1675 items. However, we restrict our analysis to documents for which the title, abstract, year, outlet and author's identifiers are available. Thus, the final dataset contains 1023 publications. Figure 1 describes the data selection strategy and the cleaning progcess. Research papers are the prevalent document category (82%), followed by review papers (8%), books and book chapters (7%), and conference papers (3%).

In terms of authorship, 17% of documents are single-authored, 30% have two authors, 23% have three authors, and the remaining 30% have four or more authors. The publication years range from 1937 to the present.

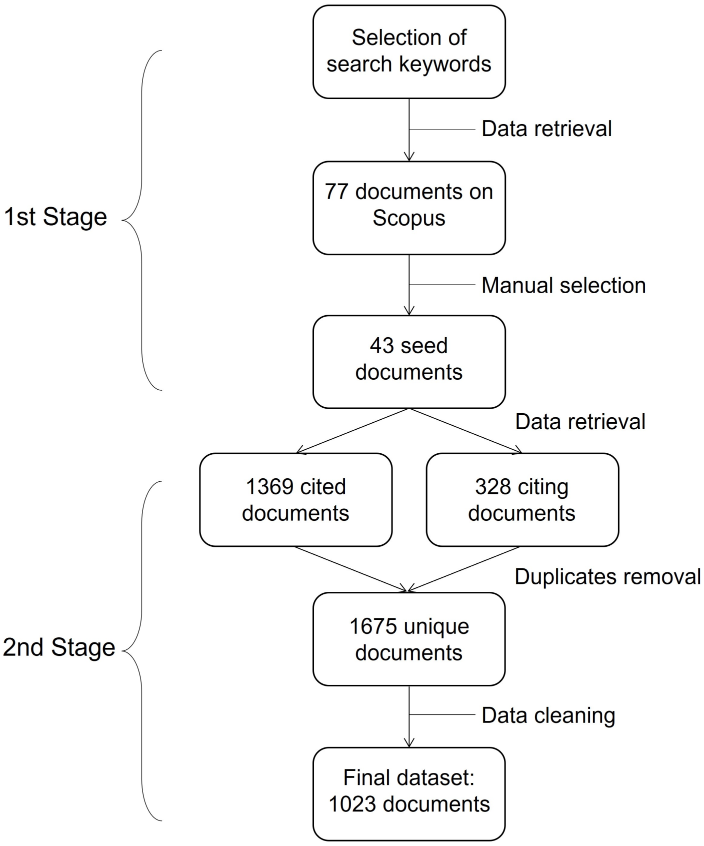


Figure 1: Data selection strategy.

4 Methods

Bibliometric analysis entails analyzing scientific publications and their metadata using statistics and text mining. Using such methodologies allows for the assessment of citations, field growth, conceptual structure, leading authors, trends, and scientific communities (Donthu et al., 2021). Bibliometrics has proven to be a valuable tool for providing a comprehensive overview of journals (Aria et al., 2020; Donthu et al., 2020) or research fields (Cuccurullo et al., 2016; Sánchez-Camacho et al., 2022; Belfiore et al., 2022).

A typical bibliometric study employs two main approaches. The first is performance analysis, which refers to the study of the authors' and journals' performance and co-citation analysis (Narin and Hamilton, 1996). The second is science mapping, which aims to identify the domain's structure in terms of topics and their evolution (Börner et al., 2003; Noyons and Van Raan, 1998). In both cases, statistical methods are used, including text mining, clustering, and, most importantly, network analysis. For an introduction to bibliometric analysis and methodologies please refer to Noyons et al. (1999) and Aria and Cuccurullo (2017).

Specifically, we use network tools to investigate both the social and conceptual structure (RO2-3-4). In the former case, collaboration networks among authors and countries are provided (Peters and Van Raan, 1991). In the latter, the co-words network is considered to identify clusters in topics and study their longitudinal evolution in the pre-defined subperiods (Callon et al., 1983). Themes are identified, in each subperiod, using a community detection algorithm named *walktrap* on the co-occurrence matrix of terms (Latapy and Pons, 2004). Then, the results can be plotted using the thematic diagram (Cobo et al., 2011). It is a Cartesian plane where Callon's centrality is on the x-axis, and Callon's density is on the y-axis (Callon et al., 1991).

The Centrality measures the interaction between networks (topics). Thus, it indicates the relative importance of a topic within the collection of documents. The density measures the strength of internal links among the terms describing the topic. Essentially, it is a measure of the topic's development. According to these definitions, each quadrant of the cartesian plane can be read as a different *theme typology*. In the upper-right quadrant, there are *motor-themes* which are both well developed and important in the field. On the upper-left side are the *niche-themes*, which are well developed but not strongly associated with other themes. *Emerging* or *disappearing themes* are in the lower-left quadrant (characterized by low centrality and low density). In the last quadrant, there are *transversal* and *basic themes*, which are well connected with most of the themes. In addition, a preliminary assessment of thematic evolution can be made by examining the word dynamics. It entails analyzing the popularity of terms (e.g., unigrams, bigrams etc.) in titles, abstracts or keywords list over the years.

Lastly, text mining tools are necessary to clean and prepare the data. It is especially important to clean abstracts since some of them include the journal's name and copyright symbols or follow a specific format divided into subsections (e.g., Introduction: [...], Motivation: [...], Results: [...]). Such structures are eliminated together with stopwords. Words are also singularized. We mainly consider the document's abstract for analyses, which provides a greater level of detail with respect to short titles. We also analyze keywords but only for preliminary analyses, which are only available for 783 documents.

To summarize the methods, Table 1 shows, for each of the research objectives, the methodology associated with it.

RO#	Objective	Methods/Approaches
RO1	Temporal evolution	- Time series plot
RO2	Performance and social structure	- Authors networks - Three-fields plot (Sankey diagram)
RO3	Conceptual Structure (CS)	- Co-words network analysis (Abstract and Keywords)
RO4	Thematic evolution (CS)	- Word dynamics (Keywords) - Thematic evolution map
RO5	Research Agenda	- Qualitative approach (global evaluation of research themes)

Table 1: Research objectives and relative methodology.

In order to perform the analysis, we use the "bibliometrix" R package (Aria and Cuccurullo, 2017). It allows to perform bibliometric analysis directly in R or using the accompanying interactive Shiny app.

5 Results and Discussion

5.1 RO1: Field development

Even though the field of survey data integration and inference for non-probability samples is still relatively new, our data retrieval strategy allows us to go back in time, providing a general perspective on the evolution of that field. As a matter of fact, the first paper in the dataset was published in 1937, and it is about the Straw election polls (Crossley, 1937).

Based on the 1023 documents published from 1937 to 2022, Figure 2 shows the year-wise distribution for the full and selected (clean) datasets following the procedures described in Section 3.2. Although 643 publications are excluded due to the absence of relevant information (Authors, Title, Abstract, Source, and Year), the two curves exhibit similar characteristics.

Prior to the 1990's, the number of publications is constant and low. Following the discussion about the conceptual background of this study, we expect this period to be characterized by fundamental papers dealing with general statistical methodologies, nonresponse, and polls.

Starting from the late 1990's, the number of publications increases, especially after 2005. Indeed, this is a very dynamic period characterized by the advent of big data and new data sources. We expect to have more insights through the thematic analysis. For this purpose, we consider five subperiods, which are shown in Figure 2 (1937-2005; 2006-2010, 2011-2015; 2016-2019; 2020-2022). The first subperiod, 1937-2005, covers the early developments in the field. For a more in-depth understanding of recent developments and to capture the dynamicity of the research field, the following subperiods cover approximately five years each. These partitions should allow to identify trends in research with a good level of detail. Indeed, considering the analysis of the conceptual background, we expect each period to be characterized by the rise of novel data sources, new statistical challenges and methodological advances. The period 2006-2010 should be characterized by an expansion of the web as a tool for data collection and an increased use of administrative data, as outlined in Section 2. After 2010, we expect the rise of new (digital trace) data sources as well as discussions regarding opportunities and challenges associated with the use of such data. A specific subperiod is assigned to the three years of the coronavirus pandemic(2020-2022).

Table 2 shows the number of documents for each subperiod in the full and selected (clean) data sets. As a result of the temporal division, each subperiod also has a similar number of documents.

Subperiod	**No. of publications (selected)**
1937-2005	366 (169)
2006-2010	276 (171)
2011-2015	369 (232)
2016-2019	386 (228)
2020-2022	254 (224)

Table 2: Number of publications by subperiod in the full and selected datasets.

Regarding *RO1*, it is evident that what was once a relatively young field has experienced rapid growth in recent years. Starting from 2010, the number of publications grew significantly. This growth can be explained and is aligned with the conceptual background (Section 2). From that year onward, web surveys became increasingly popular, and new data sources (e.g., big data and mobile data collection) became available.

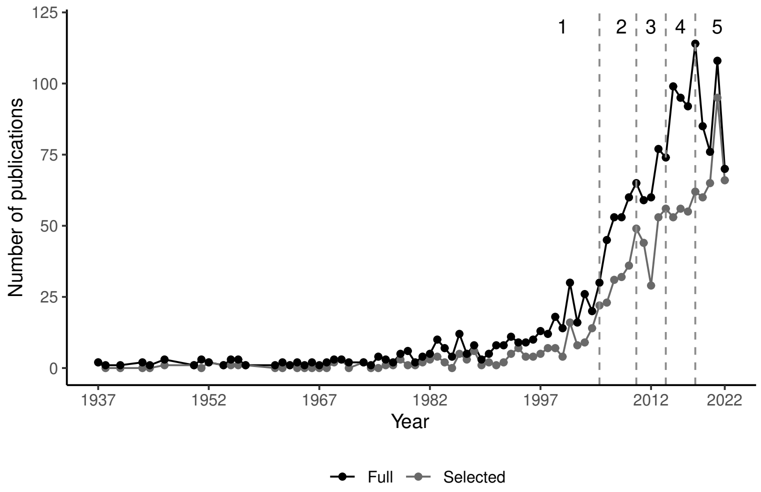


Figure 2: Year-wise distribution of publications in the full (black) and selected (grey) datasets. The five subperiods are indicated on top.

5.2 RO2: Performance and social structure

To further characterize the scientific production, we consider authors, publications outlets and their link with main themes. Figure 3 shows the ten most popular authors and publication outlets. It has been necessary to conduct a match between the names and identifiers of the authors in order to compensate for different formats and misspellings. Journals have been abbreviated according to the ISO-4 standard. Figure 4 links them with the ten most popular bigrams in abstracts (i.e., two consecutive terms) by means of a Sankey diagram. It is a flow diagram and the width of the links corresponds to the flow rate. Authors are in the first column, bigrams in the second and publication outlets in the last one.

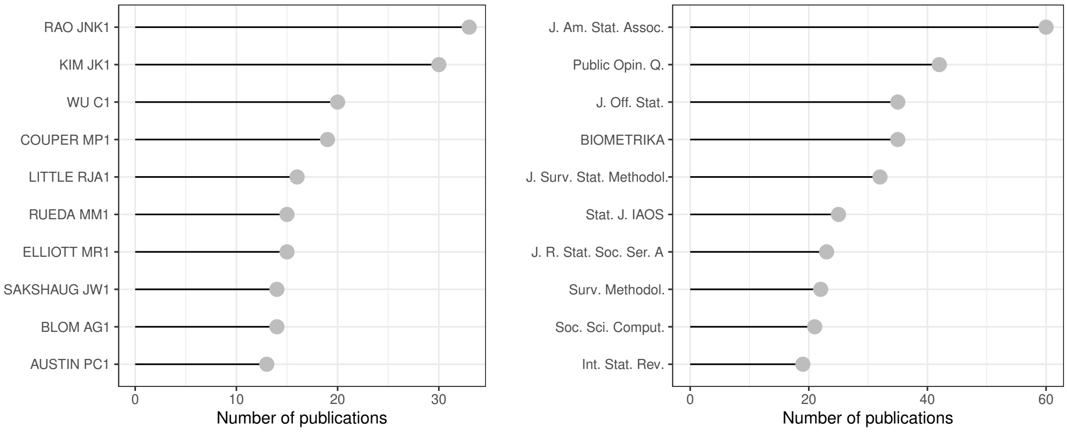


Figure 3: Top 10 authors and journals by number of publications.

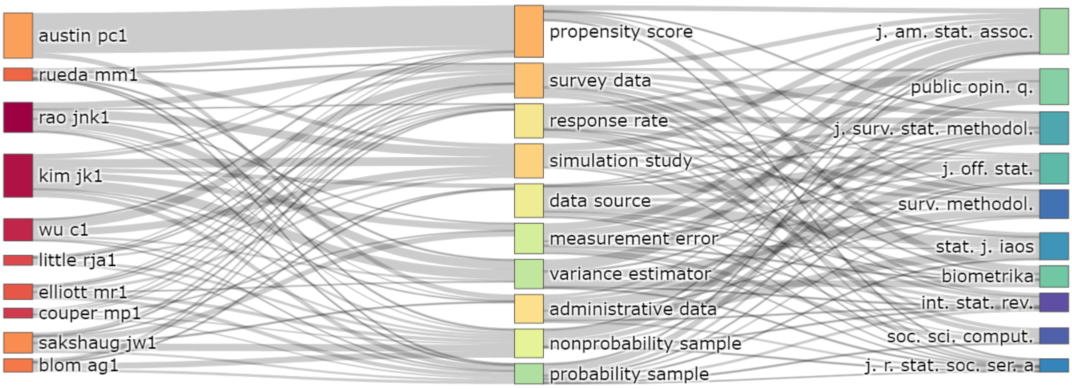


Figure 4: Three fields plot between authors, abstracts' bigrams and publication outlet.

In terms of research groups, Figure 5 shows the co-authorship network. In order to exclude one-off collaborations from the representation, the network analysis is based on the first 40 authors and restricted to those involved in at least two co-authored publications. Furthermore, the label size is proportional to the number of papers in the dataset, and the thickness of the edges, which indicate collaboration, is proportional to the number of co-authored papers. A total of nine driving research groups are identified. In order to gain a deeper understanding of the data, it is interesting to look at these three figures together.

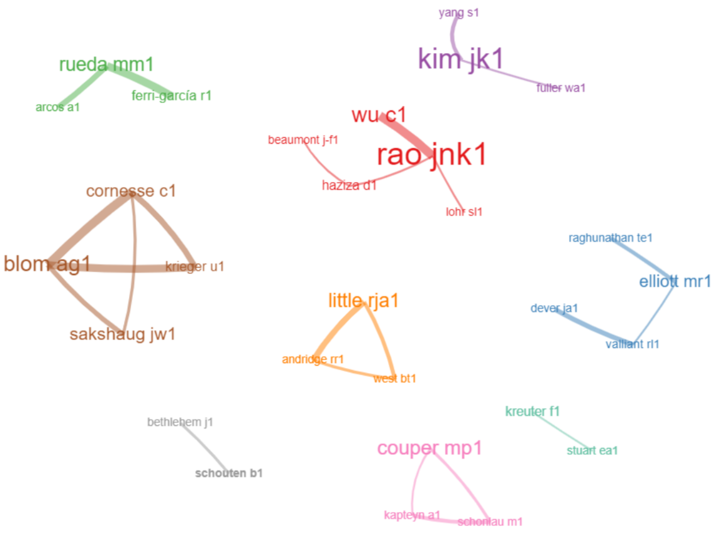


Figure 5: Author collaboration network.

Rao and Wu, the first and third top authors, are also part of the same cluster together with Haziza, Beaumont and Lohr. Their broad research topics mainly focus on survey weighting and the evaluation of inferential and data integration techniques using simulation studies. The second top author, Kim, collaborates with Yang and Fuller, considering a missing data perspective when analyzing NPS. Couper and his co-authors mainly address issues in web surveys and new data sources. The research group including Little, Andridge and West focuses primarily on selection bias and analytic inferences. Rueda and his co-authors focus on propensity score and calibration, while Elliot's group focuses mainly on model-based approaches. The collaboration among Sakshaug, Blom, Cornesse, and Krieger focuses on studies examining measurement error, administrative data, and online panels. The network does not include Austin, which has mainly one-off collaborations with many authors and is involved with medical statistics. Finally, two additional small groups are identified. The first one includes Kreuter and Stuart, which consider the perspective of causal inference when addressing selectivity. The second one is made up of Bethlehem and Schouten, which focuses on nonresponse and selection bias.

In terms of the most popular publication outlets, the Journal of the American Statistical Association takes the lead. Based on Figure 4, it is possible to identify bigrams (e.g., themes) that are distinctive to each journal, hence, identifying a polarity in themes discussed. For example, administrative data is primarily addressed by the Journal of Official Statistics and the Statistical Journal of the IAOS. Measurement error and response rates are specific to Public Opinion Quarterly and the Journal of Survey Statistics and Methodology. Studies about propensity scores or simulation studies are mainly published in the Journal of the American Statistical Association and Biometrika.

In terms of country production and collaboration, it is possible to look at Figure 6. The USA is the most productive country, followed by UK and Germany. The figure also shows the first ten collaboration edges, whose size is proportional to the number of co-authored documents. Major collaborations are evident between USA and other countries, primarily Canada, UK and Germany.

Figure 7 zooms in on European countries where the most productive and collaborative ones are UK, Germany, the Netherlands and Italy (with more than 150 publications each).

As for *RO2*, the analysis allowed us to determine which research groups are driving the research, which journals are the most influential, and how polarized the themes are within the field.

5.3 RO3: Conceptual structure

The conceptual structure of a field can be revealed through network analysis by mapping co-words. Indeed, each topic can be identified by a set of terms. Such

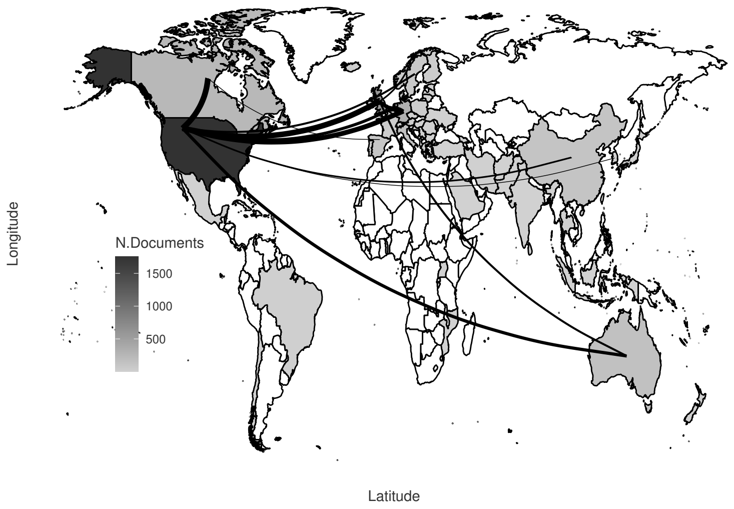


Figure 6: Country production by author affiliations and collaboration network (top 10).

terms are usually a set of keywords assigned by authors to their manuscripts or can be extracted from abstracts or titles. We consider bigrams extracted from abstracts which are more informative and descriptive than titles. Keywords are more distinctive of the document's topic, while abstracts' bigrams can help illustrate more details about studies. Therefore, we analyze both types of terms. The analysis of keywords is limited to 783 documents for which they are available. To have a static idea of the conceptual structure of the field, Figures 8 and 9 show the co-occurrence network considering keywords and abstracts' bigrams. The networks include the top 25 terms with at least two edges for both cases. Word clusters are characterized by different colors. The internal links between words within the same cluster have the same color. The gray color indicates external links between words that are assigned to different clusters but co-occur together in documents.

From the analysis of bigrams, it is possible to distinguish two main clusters. The first relates to different inferential methodologies (e.g., simulation study, propensity score, finite population, etc.). The second relates to substantive aspects such as the availability of new data sources which arose as a consequence of technological changes and their related issues (administrative data, PS and NPS, web survey/online panel, official statistics, measurement error, etc.). It is evident that there are many external (gray) links linking the two clusters, which indicates that they are highly interconnected. Keyword analysis also yields similar clusters related to methodological and practical aspects. In addition, there

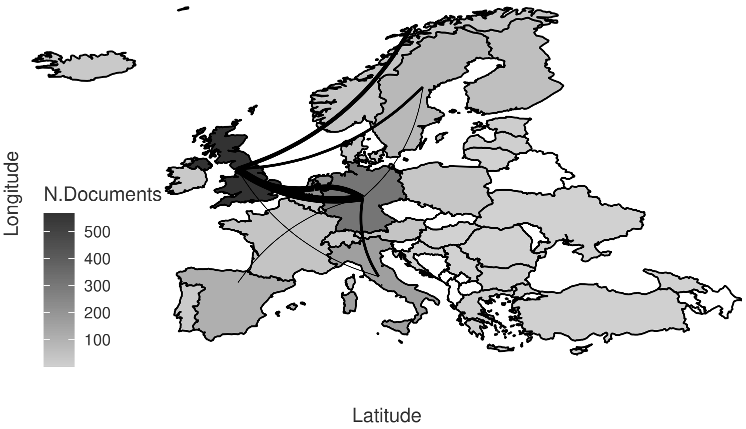


Figure 7: Country production by author affiliations and collaboration network (top 10 in Europe).

is also a society-related topic, the coronavirus pandemic. Indeed, online volunteer panels and social media-based surveys have been the subject of many social science studies concerning its impacts (Schaurer and Weiß, 2020).

This analysis, even though static, provides a general idea of the main topics in the fields, addressing *RO3*. The next step is the study of the conceptual structure over time. It concerns the evolution of themes through six subperiods, as discussed in Section 5.1. We consider the same categorization as for the themes emerged in this static analysis (methodological, substantive and applied/society-related).

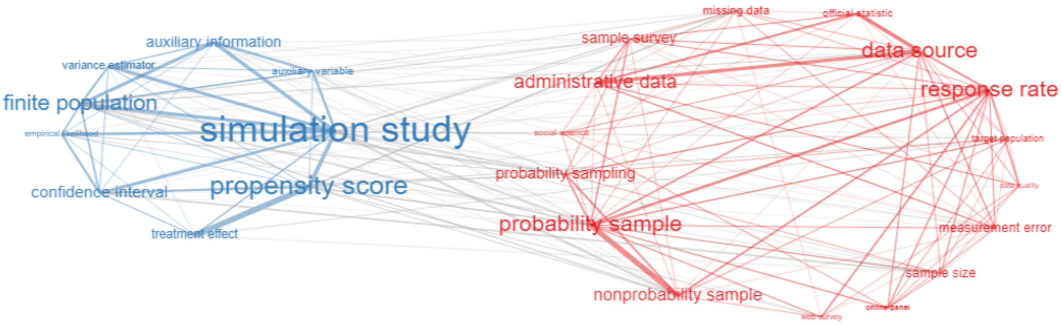


Figure 8: Abstracts' bigrams co-occurrence network.

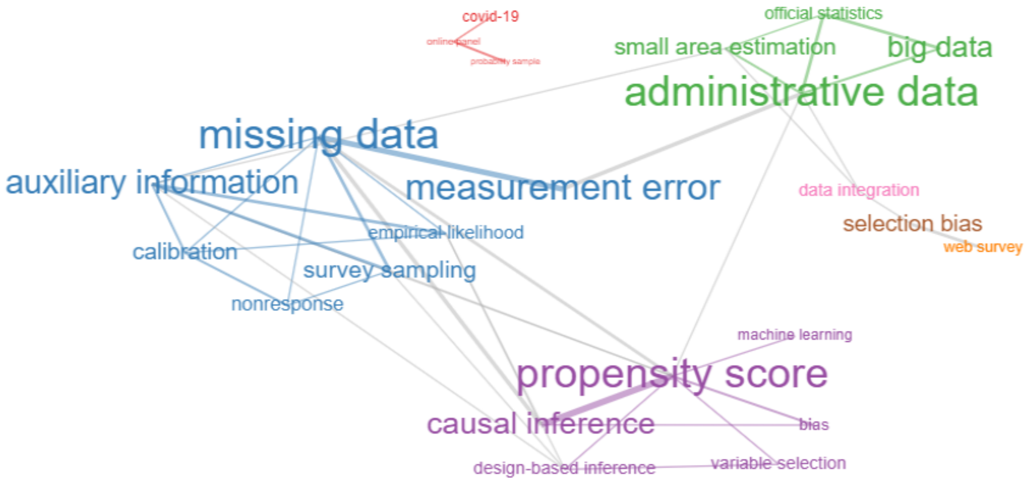


Figure 9: Keywords co-occurrence network.

5.4 RO4: Thematic evolution

This section examines the conceptual structure of the field through thematic evolution analysis. Using this method, we can identify the topics and their evolution during the five time slices under consideration (1937-2005; 2006-2010; 2011-2015; 2016-2019; 2020-2022). Essentially, it involves representing the terms that appear together in a document as a term co-occurrence network and implementing a community detection algorithm (walktrap) in order to identify themes (see Section 4 for more details). In regard to terms, we consider abstracts' bigrams that provide a good level of detail with respect to titles and keywords. In order to exclude infrequent bigrams, we restrict the analysis to those that appear in more than 3 documents, separately for each subperiod (which corresponds roughly to the 2% of documents). This is a common pre-processing step in text mining (Denny and Spirling, 2018).

However, before analyzing themes in greater detail, we focus our attention on keyword dynamics. Despite the fact that the analysis is limited to 783 documents, it does provide an overview of the most popular topics and their evolution over time. Figure 10 shows the cumulative frequency distribution for the top 10 keywords. Missing data is the first term appearing in the late 1970s. Indeed, the analysis of NPS can be approached as a missing data problem and the use of this keyword grew significantly from 2005 onward. Since the late 1990s, auxiliary information has been of interest to researchers. As an auxiliary data source to traditional surveys, administrative data (2005) and big data (2013) have emerged in recent years. On the other hand, classical statistical error issues (measurement error and selection bias) became more important and central in the methodological literature starting from 2010. Methods for data integration and inference using non-probability samples emerged as well, such as small area estimation, calibration, and propensity score (originally developed for causal inference). This dynamic is coherent with the conceptual background discussed in Section 2.

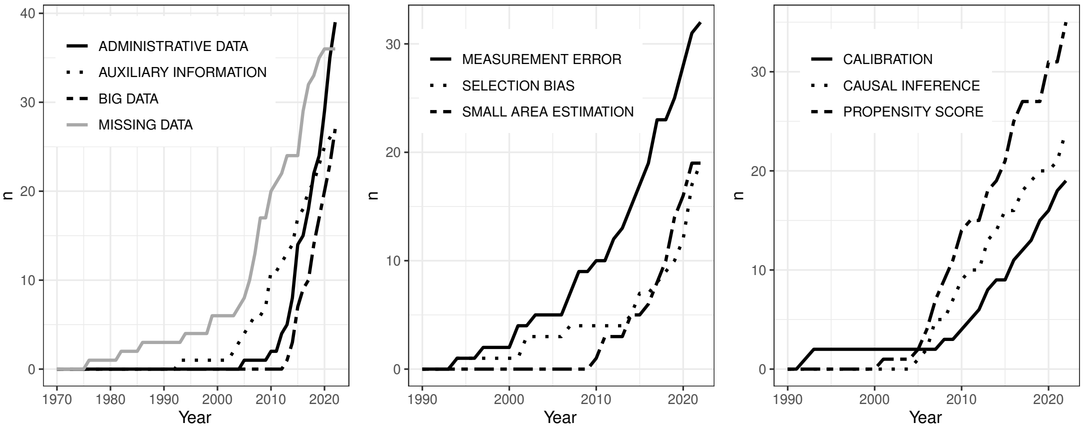


Figure 10: Top 10 Keywords dynamics (cumulative frequency distribution).

In order to gain further insight into themes in each subperiod, thematic maps can be constructed. The themes are sized in proportion to their importance in the collection of documents, and the most frequently occurring bigram is reported for each cluster (Fig. 11-12-13-14-15). When interpreting a cluster, we examine the documents most associated with it, along with other bigrams.

In this part, we adopt the same theme categorization as in the static conceptual structure analysis. Themes are classified in three categories of topics. The first one relates to methodological topics regarding inferential and data integration techniques (e.g. propensity score, variance estimation, regression analysis, etc.). The second one is about substantive topics that emerged as a consequence of technological innovation (e.g. register data, administrative data, online panel, social media, privacy paradox, linked data, etc.). The last class pertains to topics that reflect the research directions relevant to society and for which NPS data can be used (coronavirus pandemic, health care, educational attainment, etc.). The following subsections provide a detailed examination, organized according to the above-mentioned categories of topics, of each time slice. Detailed comments are provided only for the largest and most relevant clusters.

5.4.1 The first developments: 1937-2005

Prior to 2005 (Figure 11), it is possible to identify the *methodological theory* (e.g. variance estimation, measurement error, missing data, likelihood estimate) which is at the core of new inferential and data integration techniques. Measurement error and variance estimation are motor themes, which means they are highly interconnected to other topics, as well as highly developed within the field.

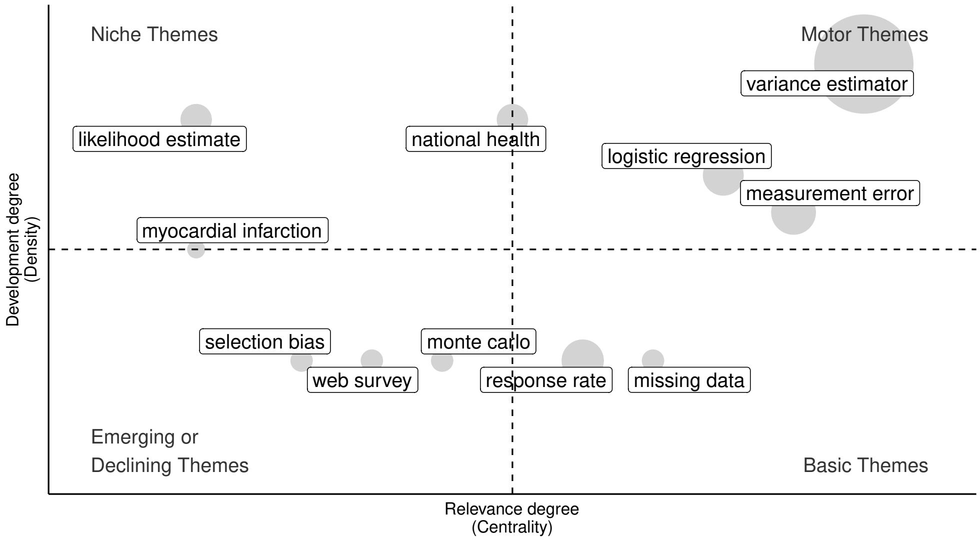


Figure 11: Thematic map 1937-2005.

Among *substantive topics* web surveys and selection bias emerges. The declining response rate is a basic theme, which means that it is generally studied in conjunction with other themes. For example, looking at associated documents, the relationship between selection bias, drop-out, and the response rate emerges, especially in relation to web surveys (Scharfstein et al., 1999; Bootsma-van der Wiel et al., 2002; Schonlau et al., 2004).

As part of *applied and society-related themes*, national health surveys and health registers are used to address migration and medical studies (myocardial infarction) (Scott and Kilbey, 1999; Austin et al., 2005).

5.4.2 Administrative data and web surveys: 2006-2010

In the second period, the biggest cluster is about *methods studies*, for which the most frequent bigram is simulation analysis (Fig. 12).

Looking at other bigrams and associated documents to that cluster, there are studies about propensity score models to address selection bias, variance estimation, sampling, and response rates. In the majority of these studies, such issues are addressed in relation to web surveys. For example, Bethlehem (2010) discusses self-selection and undercoverage in web surveys, and Schonlau et al. (2009) and Lee and Valliant (2009) address selection bias using the propensity score technique. Also, the statistical aspects of using administrative data in offi-

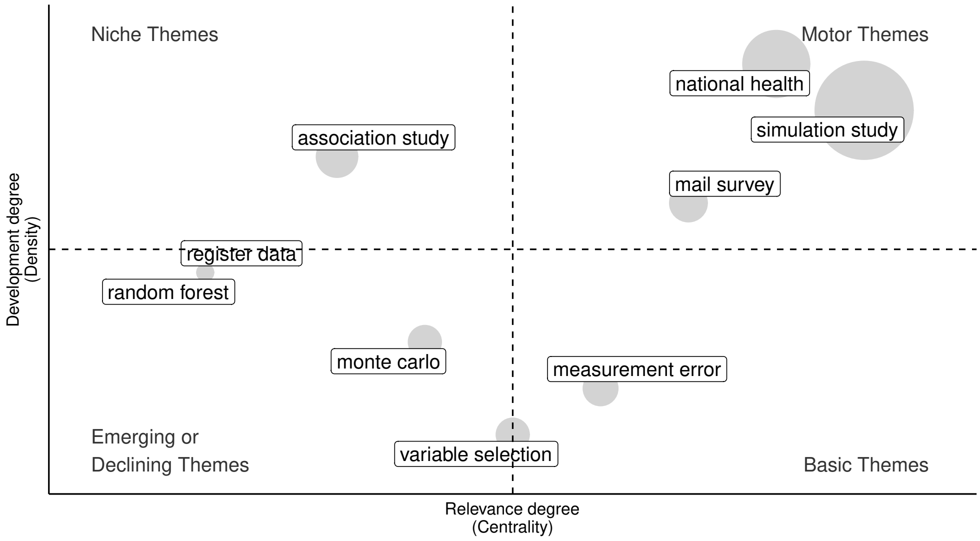


Figure 12: Thematic map 2006-2010.

cial statistics are discussed (Wallgren and Wallgren, 2007). Measurement error, which was a motor theme in the previous time slice, becomes less developed in the literature and moves to the category of basic themes.

As *substantive themes*, we find again mail surveys which is now a motor theme, indicating that it is well developed and strongly interconnected with other topics. This is also evident from the analysis of methods themes. Additionally, such studies also compare incentive effects between face-to-face and web surveys (Ryu et al., 2006).

Register data is an emerging topic that is connected to both methods and *applied* studies. For example, register data are used in the field of agriculture (Carfagna and Carfagna, 2010), demographic (Andersson and Scott, 2007) and health-related statistical studies (Raghunathan et al., 2007). A niche theme related to applied topics is genome-wide association studies.

5.4.3 New (big) data sources: 2011-2016

In line with the conceptual background, after 2010, web surveys and online panels became viable alternatives/supplements to traditional surveys, and new big data sources emerged (Fig. 13).

Indeed, as *substantive topics*, social media is an emerging theme, especially with reference to the analysis of Twitter data, while online panels and web surveys are basic and motor themes, respectively. In particular, the literature addresses the mode effect when considering mixed-mode surveys (Hox et al., 2015) or when comparing probability and non-probability (online) surveys (Erens et al., 2014).

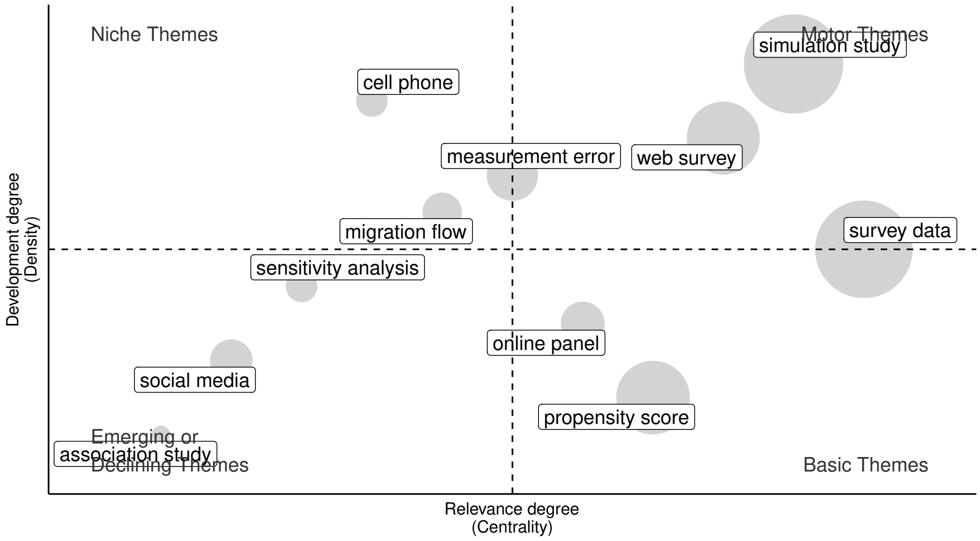


Figure 13: Thematic map 2011-2016.

A connected theme is the cluster of "survey data" which contains bigrams related to new data sources, administrative data, official statistics, survey mode, and data quality. Indeed, the opportunity and the challenges of using big data in survey research and official statistics are discussed in many studies with particular reference to the quality of the data (see for example, Struijs et al. (2014); Tam and Clarke (2015); Kitchin (2015)).

From a *methodological* point of view, the cluster related to simulation studies and methodologies for statistical inference is always a motor theme. The propensity score separates from this cluster and becomes a basic theme. In parallel, high dimensional propensity score methods emerge and applications are evaluated through sensitivity analysis (Rassen et al., 2011). The measurement error topic moves toward the direction of niche themes.

The main *applied topics* relate to genome-wide association studies (declining theme) and migration flows (niche theme). Besides these topics, also social media data are used to investigate various aspects, such as smoking behavior (Myslín et al., 2013) and communication about palliative medicine and physical activity (Nwosu et al., 2015; Zhang et al., 2013).

5.4.4 Mobile devices, data integration and the privacy paradox : 2016-2019

The fourth period is very dynamic in terms of themes (Fig. 14). As for the *methodological literature*, we can still see the presence of propensity score and missing data, plus new clusters about regression estimator (model and design based inference), machine learning methods (regression tree), adaptive lasso, nonresponse rate and survey error.

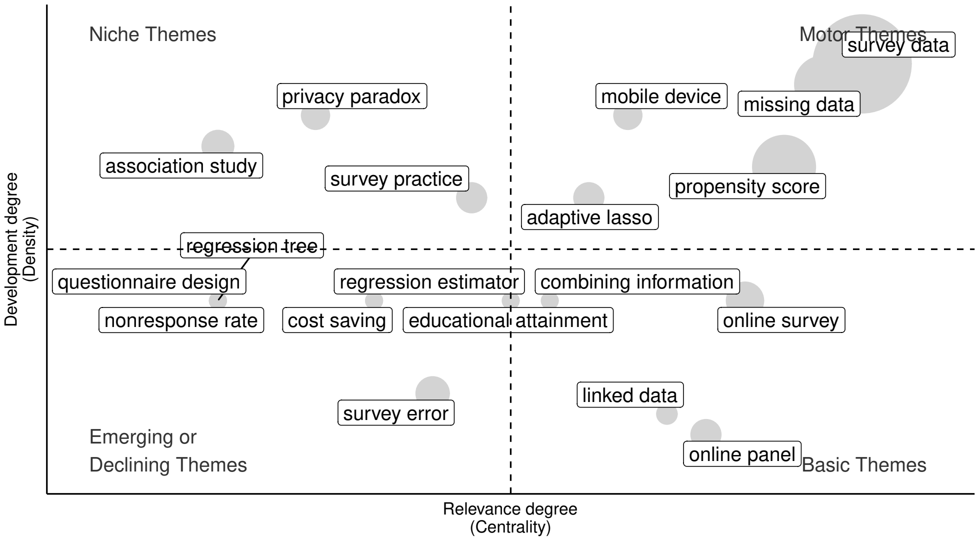


Figure 14: Thematic map 2016-2019.

The clusters of simulation studies, measurement error, and other methodologies merge with the cluster of survey data (which included administrative data, new data sources and official statistics). This new cluster reflects the temporal dynamics of topics. Although these *methods and substantive themes* have taken different paths in the past (emerging, niche, or basic themes), they are now very well integrated within each other and well developed in the literature. As a result, a mixed cluster is formed.

Within the *substantive themes*, online survey and panel take the position of basic themes, while mobile device and technology is one of the leading topics in the research (motor theme). Some studies discuss the opportunity of administering a questionnaire on smartphones or other mobile devices, and the differences in measurement and response rate between devices/modes (Revilla et al., 2016; Lugtig and Toepoel, 2016; Elevelt et al., 2019). An important related concept is the willingness of respondents to use mobile apps for surveys and sharing data (Wenz et al., 2019; Jäckle et al., 2019; Keusch et al., 2019). As we move into the digital age, privacy concerns related to the donation of personal data are becoming more relevant. It is still a niche theme, and few authors discuss the privacy paradox, which refers to the discrepancy between what respondents claim and their actual behavior with regard to online behavior and personal data protection (Barth and De Jong, 2017).

From a data integration perspective, the "combining information" cluster is a basic theme (Kim et al., 2018; Park et al., 2017). Similarly, also the topic of linked data is a basic theme. The purpose of the technique is to combine information from different sources in order to develop a new, richer dataset (Davern et al., 2019). In the literature, the cost-saving argument emerges as a rationale for integrating survey data and using new data sources. In fact, the objective of many studies is to develop methodologies that allow for inferences to be drawn, potentially resulting in cost savings (Sakshaug et al., 2019).

The genome-wide association studies are still being studied as a part of *application themes*, but they have become a niche topic over the years. Due to the wide range of topics, no other specific application clusters emerge from this analysis.

5.4.5 Recent developments and the coronavirus pandemic: 2020-2022

In the last three years, the coronavirus pandemic has shaped the research, not only in terms of *applied research* (health and socio-economic impacts of the pandemic), but also in terms of data collection (*methods and substantive topics*).

Indeed, researchers were forced to change the method of collecting data from face-to-face surveys to either online data collection or telephone surveys (Fig. 15). An example from the "online panel" cluster is the transition from the German Internet Panel to the Mannheim Corona study. The objective was to adapt the infrastructure to collect daily data in order to provide practitioners with updated information to study the socio-economic effects of the pandemic (Blom et al., 2020; Cornesse et al., 2021). In this context, social media might also be relevant for administering surveys (Lehdonvirta et al., 2021; Bradley et al., 2021). The "coronavirus pandemic" is part of the survey data cluster, which is a motor theme. Similarly, also machine learning is a motor theme, which means that both topics are well developed and highly interconnected with other themes.

Considering the current scenario, in which several data sources are available and methodologies are being developed to address inferential aspects, the theme of error sources emerges Dever (2020).

The coronavirus pandemic made it clear the role of technology in survey research and the need to develop inferential frameworks and data integration techniques in order to make use of auxiliary data (digital trace, web surveys, passive data collection, and administrative data). It implies the study of different aspects, including measurement error, selection bias, different error sources, and new sampling strategies.

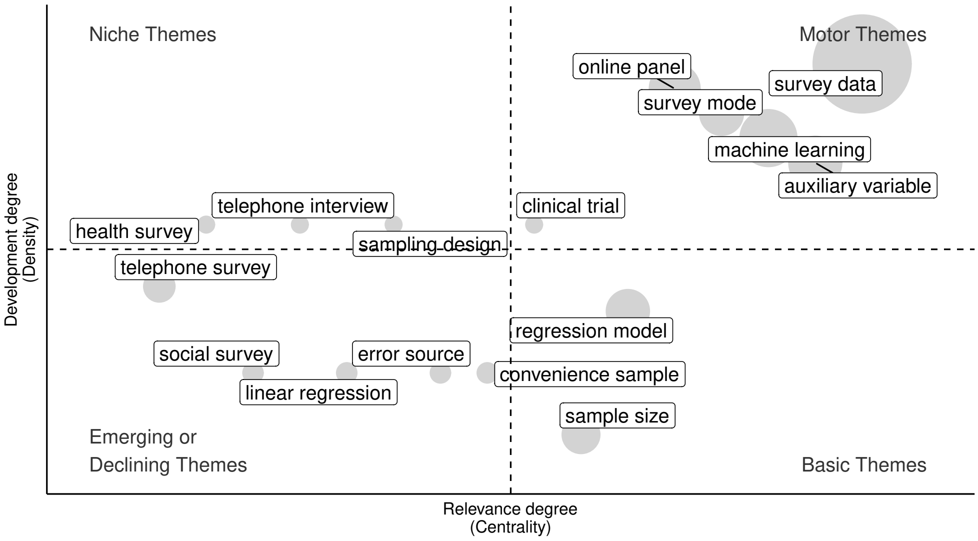


Figure 15: Thematic map 2020-2022.

In order to gain a better understanding of how themes have evolved over time, addressing *RO4*, the thematic evolution analysis was performed taking into account the three categories of topics identified in the static conceptual structure. With respect to substantive and methodological research, a cyclical pattern has emerged. Many of the themes shifted between the four dimensions considered (emerging, niche, motor, and basic). It is important to note that substantive and methodological themes are also closely interconnected. As soon as a new data source is discovered and new opportunities are investigated, new methods are developed to address inferential aspects.

In terms of applied research, the themes revealed by our analysis are mainly related to health and medical studies. One possible reason is that large amounts of health registers and claims data are readily available, making methodological studies through simulation analysis easier. Besides educational attainment and migration flow studies, other massive socioeconomic topics do not emerge. It may also be due to the wide variety of aspects that do not constitute a singular topic. As a matter of fact, when reviewing documents, we find applications related to agriculture, demographics, psychology, and social statistics.

6 Concluding remarks

6.1 Main Findings

A deep transformation is occurring in survey research with regard to the use and integration of new data sources for inference. The literature has been reviewed in many papers in light of methodological advancements, but a comprehensive study about the evolution of the field is lacking. In order to address this gap, we map the literature by providing a link between methodological, substantive, and applied themes. We employ an original approach that combines tools for bibliometric analysis and text mining in order to achieve this goal. In contrast to previous literature reviews, this study analyzes a greater number of papers in order to gain a deeper understanding of how research has evolved in response to changes in data sources and technology diffusion. This is crucial for identifying emerging trends for future research.

In particular, this paper provides an original contribution to the literature in two ways. Firstly, it characterizes the field of inference for NPS and survey data integration in terms of bibliometric performance and social structure (*RO1-3*). The leading research groups and the most productive authors are identified. Several collaborations between countries have emerged, primarily between the United States and Germany, and with reference to European countries, between United kingdom and Germany. There is also evidence of a polarity in the topics covered by journals.

Secondly, our study outlines the evolution of the field in terms of conceptual structure (*RO4*). The results of this analysis indicate that advances in survey research and technology are closely related topics. As a matter of fact, technology is both a tool and a driver of innovation. In our digital era, the research is becoming increasingly data-driven, so the need for a methodologically sound framework for inference is crucial. There is evidence of a cyclical pattern in the topic evolution across the four dimensions (emerging/declining, niche, motor, and basic) and in terms of topic typology. Indeed, new methodological aspects are investigated as soon as a new data source becomes available.

However, this study entails some limitations. Firstly, only one source (Scopus) is considered. Although it is one of the largest bibliographic databases and provides high quality data, some results may be missing. However, in the scientometric literature, different sources have been compared and there is evidence of a high level of overlap between them.(Falagas et al., 2008; Harzing and Alakangas, 2016). Secondly, the formulation of the query may affect the results (selectivity). To understand the extent of this issue, we performed a sensitivity analysis using different keywords and identified the query described in Appendix A. Thirdly, we do not consider publications that lack adequate information, as described in Section 3. As a result, there are fewer documents in the final collection. While we are aware of the concerns outlined above, we believe that the study is valuable in explaining the main themes and their evolution. Indeed, our bibliometric analysis is consistent with the conceptual background described in Section 2. The results are coherent and allow a better understanding of the social and conceptual structure of the field.

As a conclusion to this paper, we address the last objective of the research. Thus, we identify gaps in the literature based on our analyses and we outline a research agenda for future investigations (*RO5*).

6.2 A research agenda for future investigations

The thematic analysis of the field of survey data integration and inference for non-probability samples reveals that it has undergone significant changes in response to the rise of new data sources and the challenges they present. In general, we observe a shift from the early period of research, when most focus was placed on aspects related to traditional (interview-based) probability sample surveys, to new areas of research. This shift has been accelerated by the pandemic which has emphasized the need to innovate in survey research, making use of different survey modes, new data sources, and of non-traditional methods in survey methodology, like machine learning.

The transition from traditional interview surveys to telephone and web surveys is a long-standing trend in the field. Through the thematic analysis, we have observed an evolution in online surveys, starting with web and mail surveys and progressing to online (opt-in) panels and web surveys administered on mobile devices (e.g. smartphones, tablets). This transition has led to new considerations for questionnaire design, and further research is needed to understand how to optimally design and integrate surveys that are administered using different modes and devices.

The pandemic has also increased the need for timely statistics for real-time monitoring and understanding emerging social aspects. This has led to a greater use of volunteer-web surveys and alternative data sources, such as social media, which in turn has brought increased attention to inferential and data quality aspects. An emerging topic that requires further investigation is the classification of error sources in novel data sources. As data integration advances, it is also necessary to develop quality frameworks for evaluating combined products, and to understand how errors arise, accumulate, and interact throughout the entire process of inference and data integration.

With the use of digital trace data as an alternative or supplement to surveys, new privacy concerns have been raised. The ability to easily collect this data online or through donations from individuals has raised questions about the treatment of personal information and individuals' willingness to share it. Similar to consent in surveys, individuals' willingness to share their digital data (passive data collection) should be further investigated. The analysis of the literature reveals a contradiction between privacy concerns and actual online behavior (*privacy paradox*), which needs to be clarified.

Volunteer web-surveys and digital trace data share the same non-probabilistic nature. Thus, from a methodological perspective, the study of selectivity and the variables associated with it (selection or auxiliary variables) has been highlighted in the literature in recent years (Figure 15). An open problem relates to the scenario where the selection mechanism is "missing not at random" (i.e., participation directly depends on the outcome variable of interest), which requires further research.

So far, statistical frameworks have primarily focused on the estimation of finite populations quantities. However, even analytic estimates (such as regression and correlation coefficients) are susceptible to selection bias. This direction has been rarely explored in the literature, and further developments are needed. As non-traditional methods in survey search, machine learning, in particular, is a topic that has gained significant attention in recent years (2016-2022), especially during the pandemic. It encompasses not only to the analysis of unstructured data, but also to the application of such algorithms to address classic survey methodology issues, including survey weighting, data integration and variable selection.

On the basis of our analysis, non-probabilistic data sources should not be viewed as substitutes for probability sample surveys, but rather as supplements to them. PS surveys are still the gold standard in research, and new technologies and data can help to address some practical issues (for example, nonresponse) and augment the information to gain a better understanding of the phenomena. This is coherent with other literature review studies (Cornesse et al., 2020; Brick, 2011). From our analysis, it appears clear that research in this field is moving towards the use of new data sources and survey modes. One key driver of this trend is cost savings (Figure 14). Traditional PS surveys are facing challenges due to rising non-response rates and costs, making non-probability data a more cost-effective alternative. However, it is important to note that new inferential and data quality considerations must be taken into account when using non-probability data.

In conclusion, addressing the challenges and opportunities presented by non-probability data requires not only the development of methodological approaches, but also qualitative evaluations. For that reason, the collaboration between researchers from different research areas will be a key aspect for the development of the field.

Appendix

A Search Query

The search query has been selected after a sensitivity analysis considering different keywords. The objective is to select methodological papers about inferential-related topics and data integration with non-probability samples. The symbol "*" has the role of wildcards. For example "sampl*" returns both sample/samples and sampling. Plurals are considered internally by the search function. For more information about formulating search queries in Scopus, please refer to the Scopus Search Guide[1]. The search query is made by four elements linked with the AND operator:

1. **TITLE**: "data integration" OR inference OR estimat* OR integrat* OR combin* OR compar* OR "selection bias" OR "self selection" OR selectivity OR representativ* OR "non probabili* sampl*" OR "nonprobabili* sampl*" OR "nonprobabili* survey*" OR "non probabili* survey*" OR "online panel*" OR "volunteer web survey*" OR "volunteer online survey*" OR "volunteer data" OR "nonprobabili* data" OR "non probabili* data" OR "smartphone survey*" OR "digital trace data" OR "administrative data" OR "mobile data" OR "self administ*"

2. **ABSTRACT**: "data integration" OR inference OR integrat* OR combin* OR "selection bias" OR "self selection" OR selectivity) AND ("non probabili* sampl*" OR "nonprobabili* sampl*" OR "nonprobabili* survey*" OR "non probabili* survey*" OR "online panel*" OR "volunteer web survey" OR "volunteer online survey" OR "volunteer data" OR "nonprobabili* data" OR "non probabili* data" OR "smartphone survey*" OR "digital trace data" OR "administrative data" OR "self administ*"

3. **SUBJECT**: "MATH"

4. **LANGUAGE**: "English"

[1] http://schema.elsevier.com/dtds/document/bkapi/search/SCOPUSSearchTips.htm

Bibliography

Amaya, A., Bach, R., Keusch, F., and Kreuter, F. (2021). New data sources in social science research: Things to know before working with reddit data. *Social science computer review*, 39(5):943–960.

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: Adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, 8(1):89–119.

Andersson, G. and Scott, K. (2007). Childbearing dynamics of couples in a universalistic welfare state: The role of labor-market status, country of origin, and gender. *Demographic research*, 17:897–938.

Aria, M. and Cuccurullo, C. (2017). bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of informetrics*, 11(4):959–975.

Aria, M., Misuraca, M., and Spano, M. (2020). Mapping the evolution of social research and data science on 30 years of social indicators research. *Social indicators research*, 149(3):803–831.

Austin, P. C., Mamdani, M. M., Stukel, T. A., Anderson, G. M., and Tu, J. V. (2005). The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in medicine*, 24(10):1563–1578.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., et al. (2010). Research synthesis: Aapor report on online panels. *Public Opinion Quarterly*, 74(4):711–781.

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(2):90–143.

Barth, S. and De Jong, M. D. (2017). The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, 34(7):1038–1058.

Beaumont, J.-F. and Rao, J. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? surv. *The Survey Statistician*, 83:11–22.

Belfiore, A., Cuccurullo, C., and Aria, M. (2022). Iot in healthcare: A scientometric analysis. *Technological Forecasting and Social Change*, 184:122001.

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

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, 74(5):817–848.

Biffignandi, S. and Bethlehem, J. (2021). *Handbook of web surveys.* John Wiley & Sons.

Blom, A. G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., et al. (2020). High frequency and high quality survey data collection. *Survey Research Methods*, 14(2):171–178.

Bootsma-van der Wiel, A. v., Van Exel, E., De Craen, A., Gussekloo, J., Lagaay, A., Knook, D., and Westendorp, R. (2002). A high response is not essential to prevent selection bias: results from the leiden 85-plus study. *Journal of clinical epidemiology*, 55(11):1119–1125.

Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255.

Bosch Jover, O. and Revilla, M. (2022). When survey science met web tracking: presenting an error framework for metered data. *Journal of the Royal Statistical Society. Series A: Statistics in Society.*

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700.

Braun, M. T. and Kuljanin, G. (2015). Big data and the challenge of construct validity. *Industrial and Organizational Psychology*, 8(4):521–527.

Breidt, F. J. and Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science*, 32(2):190 – 205.

Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75(5):872–888.

Callon, M., Courtial, J.-P., and Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205.

Callon, M., Courtial, J.-P., Turner, W. A., and Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social science information*, 22(2):191–235.

Carfagna, E. and Carfagna, A. (2010). Alternative sampling frames and administrative data. what is the best data source for agricultural statistics? *Agricultural survey methods*, pages 45–61.

Ceron, A., Curini, L., and Iacus, S. M. (2016). *Politics and big data: Nowcasting and forecasting elections with social media.* Routledge.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., and Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of informetrics*, 5(1):146–166.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Cornesse, C., Krieger, U., Sohnius, M.-L., Fikel, M., Friedel, S., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., et al. (2021). From german internet panel to mannheim corona study: Adaptable probability-based online panel infrastructures during the pandemic. *Journal of the Royal Statistical Society: Series A (Statistics in Society).*

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3):145–156.

Crossley, A. M. (1937). Straw polls in 1936. *Public Opinion Quarterly*, 1(1):24–35.

Cuccurullo, C., Aria, M., and Sarto, F. (2016). Foundations and trends in performance management. a twenty-five years bibliometric analysis in business and public administration domains. *Scientometrics*, 108(2):595–611.

Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262.

Davern, M. E., Meyer, B. D., and Mittag, N. K. (2019). Creating improved survey data products using linked administrative-survey data. *Journal of Survey Statistics and Methodology*, 7(3):440–463.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Dever, J. A. (2020). Discussion of "how errors cumulate: Two examples" by roger tourangeau. *Journal of Survey Statistics and Methodology*, 8(3):433–441.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., and Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296.

Donthu, N., Kumar, S., and Pattnaik, D. (2020). Forty-five years of journal of business research: A bibliometric analysis. *Journal of Business Research*, 109:1–14.

Einarsson, H., Sakshaug, J. W., Cernat, A., Cornesse, C., and Blom, A. G. (2022). Measurement equivalence in probability and nonprobability online panels. *International Journal of Market Research*, 64(4):484–505.

Elevelt, A., Lugtig, P., and Toepoel, V. (2019). Doing a time use survey on smartphones only: What factors predict nonresponse at different stages of the survey process? *Survey Research Methods*, 13(2):195–213.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Erens, B., Burkill, S., Couper, M. P., Conrad, F., Clifton, S., Tanton, C., Phelps, A., Datta, J., Mercer, C. H., Sonnenberg, P., et al. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. *Journal of medical Internet research*, 16(12):e3382.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., and Pappas, G. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342.

Groves, R. M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly*, 75(5):861–871.

Harzing, A.-W. and Alakangas, S. (2016). Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804.

Hox, J. J., De Leeuw, E. D., and Zijlmans, E. A. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in psychology*, 6:87.

Iacus, S. M. and Porro, G. (2016). *Subjective Well-Being and Social Media*. Routledge.

Jäckle, A., Burton, J., Couper, M. P., and Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Coverage and participation rates and biases. *Survey Research Methods*, 13(1):23–44.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'neil, C., and Ushe, A. (2015). Big data in survey research: Aapor task force report. *The Public Opinion Quarterly*, 79(4):839–880.

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., and Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public opinion quarterly*, 83(S1):210–235.

Kim, J. K., Wang, Z., Zhu, Z., and Cruze, N. B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics*, 23(2):175–189.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, 31(3):471–481.

Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly*, 74(5):880–906.

Latapy, M. and Pons, P. (2004). Computing communities in large networks using random walks. *arXiv preprint cond-mat/0412368*.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343.

Lehdonvirta, V., Oksanen, A., Räsänen, P., and Blank, G. (2021). Social media, web, and panel surveys: using non-probability samples in social and policy research. *Policy & internet*, 13(1):134–155.

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.

Lugtig, P. and Toepoel, V. (2016). The use of pcs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, 34(1):78–94.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3):469–487.

Myslín, M., Zhu, S.-H., Chapman, W., Conway, M., et al. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8):e2534.

Narin, F. and Hamilton, K. (1996). Bibliometric performance measures. *Scientometrics*, 36(3):293–310.

Nordbotten, S. (2010). The use of administrative data in official statistics-past, present and future: with special reference to the nordic countries. *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, page 205–223. Available at https://officialstatistics.wordpress.com/.

Noyons, E., Moed, H., and Van Raan, A. (1999). Integrating research performance analysis and science mapping. *Scientometrics*, 46(3):591–604.

Noyons, E. and Van Raan, A. (1998). Advanced mapping of science and technology. *Scientometrics*, 41(1-2):61–67.

Nwosu, A. C., Debattista, M., Rooney, C., and Mason, S. (2015). Social media and palliative medicine: a retrospective 2-year analysis of global twitter data to evaluate the use of technology to communicate about issues at the end of life. *BMJ supportive & palliative care*, 5(2):207–212.

Park, S., Kim, J. K., and Stukel, D. (2017). A measurement error model approach to survey data integration: combining information from two surveys. *Metron*, 75(3):345–357.

Peters, H. and Van Raan, A. (1991). Structuring scientific activities by co-author analysis: An expercise on a university faculty level. *Scientometrics*, 20(1):235–255.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., and Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102(478):474–486.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1):242–272.

Rassen, J. A., Glynn, R. J., Brookhart, M. A., and Schneeweiss, S. (2011). Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American journal of epidemiology*, 173(12):1404–1413.

Revilla, M., Toninelli, D., Ochoa, C., and Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*.

Ryu, E., Couper, M. P., and Marans, R. W. (2006). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18(1):89–106.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics*, 35(3):653–681.

Salvatore, C., Biffignandi, S., and Bianchi, A. (2021). Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156(2):601–630.

Sánchez-Camacho, C., Carranza, R., Martín-Consuegra, D., and Díaz, E. (2022). Evolution, trends and future research lines in corporate social responsibility and tourism: A bibliometric analysis and science mapping. *Sustainable Development*, 30(3):462–476.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Schaurer, I. and Weiß, B. (2020). Investigating selection bias of online surveys on coronavirus-related behavioral outcomes. *Survey Research Methods*, 14(2):103–108.

Schonlau, M., Van Soest, A., Kapteyn, A., and Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, 37(3):291–318.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. H. (2004). A comparison between responses from a propensity-weighted web survey and an identical rdd survey. *Social science computer review*, 22(1):128–138.

Scott, A. and Kilbey, T. (1999). Can patient registers give an improved measure of internal migration in england and wales? *Population Trends*, 96:44–55.

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*, 85(S1):399–422.

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5):503–516.

Struijs, P., Braaksma, B., and Daas, P. J. (2014). Official statistics and big data. *Big Data & Society*, 1(1):2053951714538417.

Struminskaya, B., Lugtig, P., Keusch, F., and Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social Science Computer Review*, 0(0):0894439320979951.

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., and Dolmans, R. (2021). Sharing Data Collected with Smartphone Sensors: Willingness, Participation, and Nonparticipation Bias. *Public Opinion Quarterly*, 85(S1):423–462.

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.

Tsung, C., Kuang, J., Valliant, R. L., and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive lasso. *Survey Methodology*, 44(1):117–145.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*, volume 553. John Wiley & Sons.

Wenz, A., Jackle, A., and Couper, M. P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, 13(1):1–22.

Zhang, N., Campo, S., Janz, K. F., Eckler, P., Yang, J., Snetselaar, L. G., Signorini, A., et al. (2013). Electronic word of mouth on twitter about physical activity in the united states: exploratory infodemiology study. *Journal of medical Internet research*, 15(11):e2870.

Chapter 3

Bayesian integration of probability and non-probability samples for logistic regression

Paper under review:
Salvatore C., Biffignandi S., Sakshaug J., Wiśniowski A., Struminskaya B. Bayesian integration of probability and non-probability samples for logistic regression. Revised After Major Revision.

1 Introduction

Probability sampling has long been considered the gold standard method for designing large-scale, population-based surveys. It provides a scientifically sound framework for making population-based inference as a function of the sample design with measurable bounds of uncertainty (Neyman, 1934). All population units have a known (or knowable) non-zero chance of selection and through inverse probability weighting it is possible to construct design-unbiased estimators (Kish, 1965). However, unbiasedness is threatened by the practical realities of surveys, including non-coverage, non-response, and other error sources defined in the Total Survey Error (TSE) framework (Biemer, 2010). Non-response in particular has significantly increased over the years, mainly driven by increasing non-contacts and refusals. Strategies to cope with non-response, such as through higher incentives or more intensive fieldwork efforts have failed to stymie this trend, and survey costs have increased to a point where fielding large probability sample (PS) surveys has become cost-prohibitive for many survey researchers and study sponsors (Luiten et al., 2020).

For these reasons, among others, such as convenience, many researchers have embraced the more affordable and timely alternative of non-probability sample (NPS) surveys. Although non-probability samples are used in a variety of *big data* contexts (e.g., social media, sensors, etc.), this article focuses on NPS surveys, specifically those conducted via online access (or volunteer opt-in) panels. Due to their relatively low cost, the use of online NPS surveys has increased in recent years (Biffignandi and Bethlehem, 2021). However, because they rely on individuals self-selecting themselves into the panel and agreeing to take part in periodic surveys, they are strongly susceptible to selection bias, often more so than carefully-designed PS surveys (Cornesse et al., 2020). Consequently, one of the current statistical challenges is developing approaches for reducing selection bias in NPS surveys or integrating NPS surveys with PS surveys that are assumed to be of higher quality. The present study focuses on the latter approach.

Specifically, we propose a Bayesian approach to supplement a small PS survey with information from a parallel NPS survey to improve inference about logistic regression coefficients for a binary outcome. Several strongly-informative priors constructed from the NPS information are evaluated in terms of the mean-squared error (MSE) of the posterior estimates through a simulation study and real data application. We show that supplementing PS surveys with these priors produces coefficient estimates with lower MSEs compared to not using any NPS information and relying solely on the PS data for inference. An R Shiny web app[1] is provided that displays the algorithm and full results, and includes an interactive cost analysis tool to estimate potential cost savings of the method.

The remainder of the article is organized as follows. Section 2 provides background on the topic and reviews the relevant literature. Section 3 outlines the research aims. Section 4 introduces the methodological framework. Section 5 presents the simulation results and Section 6 the results of the real data application. The article concludes in Section 7 with a general discussion of the findings, recommendations, and potential research extensions.

2 Background

Participants in NPS surveys are typically recruited from online access panels or directly from visited web sites, including search engines and social media sites (Baker et al., 2010). Thousands of internet users opt-in to online access panels and crowd-sourcing platforms, often in exchange for incentives or rewards to complete

[1] https://bayesdataintegration.shinyapps.io/shiny_bayes_data_integration/

periodic web surveys. Thus, NPS web surveys can be completed by a large number of respondents rapidly and at low cost, enabling timely statistics (Astley et al., 2021; Kreuter et al., 2020) and reaching rare or hard-to-interview populations (Berzofsky et al., 2018). However, the drawbacks of online access panels are that there is no explicit sampling frame of the general internet population, the data generating process is typically outside the researcher's control, and the lack of a known random selection mechanism renders the classical design-based approach to inference inappropriate. Moreover, parts of the general population are not covered, such as people without internet access and those who were not exposed to, or targeted by, the access panel's advertising efforts (Bethlehem, 2010). For these reasons, serious concerns remain about the generalizability of NPS survey estimates.

Although NPS surveys are convenient and cost-efficient, the empirical evidence suggests that the accuracy of the resulting estimates is usually lower than those obtained from PS surveys (Cornesse et al., 2020; Yeager et al., 2011). Moreover, there is no unified inferential framework for NPS surveys and error frameworks for such data, akin to the TSE framework for PS surveys, have only recently begun to emerge (Amaya et al., 2020). Inference usually relies on modelling and statistical adjustments based on benchmark data, including high-quality PS surveys or official statistics (Baker et al., 2013; Elliott and Valliant, 2017; Valliant, 2020; Dever and Shook-Sa, 2015; Dutwin and Buskirk, 2017). In this setting, where PS surveys are known to have higher data quality but are expensive and NPS surveys are convenient and more affordable but can suffer from large selection biases, a natural avenue of research is the integration of both PS and NPS surveys to exploit their respective advantages in a way that overcomes their respective disadvantages and minimizes overall survey costs (Couper, 2013; Miller, 2017; Beaumont, 2020; Rao, 2021).

Classic data integration approaches include the construction of pseudo-weights and/or calibrated weights based on auxiliary variables or population totals (Elliot, 2009; DiSogra et al., 2011; Robbins et al., 2020; Raghunathan et al., 2021). An alternative approach is doubly-robust inference where the quasi-randomization approach for the construction of pseudo-weights and the modelling of the outcome variable of interest are combined (Yang et al., 2020). A key aspect of this approach is that the estimator is approximately unbiased if either one of the models is correctly specified. Another approach is mass imputation which comes from the missing data literature (Kim et al., 2021). Small area estimation (SAE) methods are also applied for integrating multiple data sources (Ganesh et al., 2017;

Beaumont and Rao, 2021). Moreover, the availability of new data sources and unstructured *big data* offers new methodological possibilities (Stier et al., 2020). Kim and Tam (2021) address the problem of finite population inference when integrating *big data* sources and a PS survey accounting for both selection bias and measurement error without making missing at random (MAR) assumptions. Another option is to integrate both data sources under a Bayesian framework using latent class or hierarchical models (Alexander et al., 2020; Hsiao et al., 2020; Sakshaug et al., 2019; Wiśniowski et al., 2020).

Many studies on combining probability and non-probability samples focus on finite population inference. Nevertheless, other types of inference can also be of interest, such as the study of associations and model parameters. While descriptive estimates tend to have larger discrepancies between the two sample types compared to correlations and regression coefficients (Pasek, 2016), the literature is mixed with some studies reporting strong correspondence between PS and NPS surveys for regression coefficients and other studies reporting larger discrepancies (Malhotra and Krosnick, 2007; Callegaro et al., 2014; Thompson and Pickett, 2020). The presence of selection bias in NPS surveys for regression coefficients has recently been studied by West et al. (2021) who propose indices of non-ignorable selection bias in linear and probit regression models based on a pattern-mixture model and on the availability of aggregate auxiliary data.

3 Research Aims

The present study focuses on integrating PS and NPS survey data for improving analytic inference about coefficients for logistic regression models and potentially reducing survey costs, which is still an emerging topic in the literature. Our contribution focuses on supplementing a small PS survey with information from a parallel NPS survey with overlapping variables. In order to combine the information coming from the two samples, we consider a Bayesian framework where inference is based on the PS survey and available information from the NPS survey is supplied through a strongly-informative prior. Sakshaug et al. (2019) and Wiśniowski et al. (2020) proposed a similar framework for the analysis of continuous data using linear regression. However, categorical data analysis, and particularly the modeling of binary outcomes is of key interest in the social and health sciences, where the objective is to study the classification of behaviors, attitudes, and characteristics (e.g. healthcare coverage, unemployment, voting, illness, among others). Thus, we extend the previous work by developing an approach for modeling binary outcomes with covariates using logistic regression and leave the analysis of other categorical data types to future work.

To evaluate the proposed method, we conduct a simulation study to compare the performance of several strongly-informative priors in terms of mean-squared error (MSE) of the posterior estimates according to different selection mechanisms, selection probabilities, and sample sizes. The comparisons are made in reference to a weakly-informative ("baseline") prior in which no NPS information is supplied and inference is based on the PS survey data alone. In contrast to previous studies, which generally assume a MAR selection mechanism for the NPS data, we do not make such an assumption and also evaluate the framework in the missing not at random (MNAR) context, where the NPS selection mechanism depends on the outcome variable of interest. Under this framework, incorporating biased NPS data through a strongly-informative prior is likely to result in posterior estimates that have more bias, but possibly less variance compared to using a naïve prior that does not utilize any NPS information. Thus, our main interest lies in investigating under which conditions this reduction in variance offsets increases in bias, thus leading to lower posterior MSEs relative to a probability-only sample. We expect that any MSE reductions will be most evident when considering small PS sizes, where the strongly-informative priors will have the most influence on reducing the posterior variance, but that this result will be moderated by the underlying selection mechanism and level of bias in the NPS data.

In addition to the simulation study, we evaluate the strongly-informative priors through a real data application involving a nationally representative, probability-based web survey and several overlapping non-probability web surveys with potentially different selection mechanisms. A cost analysis is performed to study the extent to which the strongly-informative priors produce posterior estimates at a lower cost for the same MSE as would be obtained from a (potentially more expensive) probability-only sample. Here, we expect that the largest potential cost savings will occur when PS sizes are small and the MSE reduction is notable.

4 Methodology

4.1 The Bayesian inferential framework

The Bayesian framework offers a unified approach for integrating multiple data sources of different sizes and quality in a natural way, that is, through the prior structure. In the proposed methodology, inference is based on the PS survey and additional information from the NPS survey is incorporated through a strongly-informative prior. The aim is to improve inference about logistic regression model parameters for a small probability sample by integrating information from a larger parallel non-probability sample. We assume that the PS survey is of high quality (i.e., unbiased) despite its potentially small sample size and that the NPS survey might be subject to large selection biases and thus has lower quality.

We consider logistic regression to model a binary outcome with covariates. Let us denote with Y_{PS} the binary response vector of size $n_{PS} \times 1$ and X_{PS} the $n_{PS} \times k$ design matrix from a PS survey. The PS data are denoted by $D_{PS} = (n_{PS}, Y_{PS}, X_{PS})$. Similarly, the data from a parallel NPS survey are denoted by $D_{NPS} = (n_{NPS}, Y_{NPS}, X_{NPS})$. The likelihoods of the NPS and PS data are denoted by $L(\boldsymbol{\beta}|D_{NPS})$ and $L(\boldsymbol{\beta}|D_{PS})$, respectively.

The logistic model is presented in Equation 1, where $\theta_i = \frac{\exp(\boldsymbol{X_i'\beta})}{1+\exp(\boldsymbol{X_i'\beta})}$ are the success probabilities:

$$\begin{aligned} &Y_{PSi} \sim Ber(\theta_i), \\ &logit(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_{0PS} + \sum_{j=1}^{k} \beta_{jPS} X_{PSij} \quad \text{for} \quad i = 1, \dots, n_{PS} \end{aligned} \tag{1}$$

Inference is expressed through the posterior distribution, $\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})$, which is based on the likelihood function for the PS data $L(\boldsymbol{\beta}|D_{PS})$ and the prior distribution $\pi(\boldsymbol{\beta})$ through Bayes theorem.

4.2 Construction of the prior distributions

Eliciting the prior is a key step in Bayesian analysis. A strongly-informative prior distribution incorporates previous information or beliefs about the parameters before the data are observed. Such information may come from the literature, historical data, or expert opinions. If there is no previous information or beliefs to incorporate, then vague or weakly-informative priors are used to reflect this lack of knowledge.

4.2.1 Strongly-informative priors

We derive several strongly-informative prior specifications that incorporate information from a NPS survey about a model parameter. We first consider multiple variations of a normally distributed prior, which is a common choice for constructing priors. The general idea behind this class of proposed priors is to set the location parameter equal to the Maximum-Likelihood estimate (MLE) of the target parameter from the substantive model of interest obtained from the NPS data and scale this information through the scale parameter. We propose different formulations for the scale parameter taking into account the distance between the MLEs from the separate PS and NPS data sources and the NPS size. We refer to this class of priors as distance priors.

The first prior from this class of distance priors is simply referred to as the **Distance prior**, which was originally proposed by (Sakshaug et al., 2019) for continuous outcomes:

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{jNPS}, |\hat{\beta}_{jPS} - \hat{\beta}_{jNPS}|\right). \tag{2}$$

The scale parameter is set equal to the absolute difference between the two MLEs from the PS and NPS data, denoted by $\hat{\beta}_{PS}$ and $\hat{\beta}_{NPS}$, respectively. These MLEs can be estimated using standard logistic regression functions in statistical software packages. The larger the absolute difference (an indication of larger selection bias in the NPS data), the smaller the influence of the prior on the posterior.

A shortcoming of such a formulation is that the scale parameter can be very small or equal to zero in extremely unlikely cases (Sakshaug et al., 2019). Thus, as an alternative we propose to take the maximum value of the squared difference between the ML estimates and the variance of the MLE based on the NPS data. Then, we use the NPS size to shrink the prior around $\hat{\beta}_{jNPS}$. To do that, we use an inverse logarithmic scaling factor $\frac{1}{\log(n_{NPS})}$, where n_{NPS} is the length of the vector of the NPS data (Wiśniowski et al., 2020). This may lead to potentially more bias but lower posterior variance. We refer to this prior as the **Distance-log prior**:

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{jNPS}, \sqrt{\frac{1}{\log(n_{NPS})} \cdot \max\left((\hat{\beta}_{jPS} - \hat{\beta}_{jNPS})^2, \hat{\sigma}^2_{\beta_j NPS}\right)}\right) \tag{3}$$

While these two prior formulations are not new, the literature suggests that they've never been applied in a logistic regression setting.

We propose a slightly modified prior specification using the common logarithm instead of the natural logarithm, which will result in a slightly wider distribution. This prior is referred to as the **Distance-log10 prior**:

$$\beta_j \sim \mathcal{N}\left(\hat{\beta}_{jNPS}, \sqrt{\frac{1}{\log_{10}(n_{NPS})} \cdot \max\left((\hat{\beta}_{jPS} - \hat{\beta}_{jNPS})^2, \hat{\sigma}^2_{\beta_j NPS}\right)}\right) \tag{4}$$

If participation in the NPS survey depends directly on the outcome variable of interest, then also the intercept will be biased. Thus, we consider a mixed formulation for the distance prior specifications in Eqs. 2, 3, and 4, where the prior for the intercept is replaced by a weakly-informative Student-t prior distribution with three degrees of freedom, t_3. We refer to this set of priors as the **Mixed-distance priors** (Mixed-Distance, Mixed-Distance-log, and Mixed-Distance-log10, respectively). In all of these prior formulations, the issue of using the PS data twice arises. Indeed, inference is based on the PS data which are also used as a reference to construct the scale (variance) parameter of the priors. The use of PS data on the second-order prior component essentially serves as protection against a prior informed from a severely biased NPS data source from dominating the posterior inference.

Lastly, we propose the **Power prior** to integrate the two sample types. Chen et al. (1999; 2000) and Ibrahim et al. (2000) introduced this new class of strongly-informative prior distributions based on the availability of historical data. The prior's properties have been discussed in different contexts, e.g. in regression models (Ibrahim et al., 2000), variable selection, logistic regression (Chen et al., 1999), and generalized linear models (GLM) (Chen et al., 2000). The term *historical data* refers to both data from previous studies or similar parallel studies that are used to inform the model parameters. The power prior is mainly used in clinical trials and health applications (De Santis, 2006; Ibrahim et al., 2012). To the best of our knowledge, this is the first time that the power prior has been used to integrate probability and non-probability sample surveys.

In our context, the NPS survey can be viewed as the *historical data*. The degree of influence of the NPS data on the posterior inference is determined by the power parameter $0 \leq a \leq 1$. The case of $a = 0$ corresponds to no borrowing of information while $a = 1$ reflects the case of full borrowing. Although it is possible to specify a prior for the power parameter a, we consider a fixed value for this parameter. The initial prior for $\boldsymbol{\beta}$ is denoted by $\pi_0(\boldsymbol{\beta})$. Equation 5 shows the prior specification for fixed a:

$$\pi(\boldsymbol{\beta}, a|D_{NPS}) \propto L(\boldsymbol{\beta}|D_{NPS})^a \pi_0(\boldsymbol{\beta}). \tag{5}$$

Thus, the resulting posterior in Equation 6 is also proportional to the NPS data:

$$\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS}, a) \propto L(\boldsymbol{\beta}|D_{PS})L(\boldsymbol{\beta}|D_{NPS})^{a}\pi_0(\boldsymbol{\beta}). \tag{6}$$

We set the prior $\pi_0(\boldsymbol{\beta})$ to be weakly informative as in Eq. 7. We choose not to set a to an arbitrary fixed value because it may lead to high MSE values in the presence of high selection bias. Instead, we select a in an automated fashion according to the similarity between the MLEs obtained from the PS and NPS data. Specifically, we set a equal to the p-value resulting from the Hotelling's T^2 test for the difference between the two vectors, $\hat{\boldsymbol{\beta}}_{PS}$ and $\hat{\boldsymbol{\beta}}_{NPS}$. P-values close to 1 indicate strong evidence for the null hypothesis and, in such cases, it is suggested to *borrow* more heavily from the NPS data. For p-values close to 0 (indicating significant differences between the two vectors), the amount of information borrowed is smaller or, in the worst-case, is nil. This method allows for automatic rescaling of $L(\boldsymbol{\beta}|D_{NPS})$ based on the difference between the MLEs.

4.2.2 Weakly-informative (baseline) prior

To form the basis for evaluating the informative priors, we evaluated several vague priors to serve as baseline priors, including uniform and some normally-distributed priors centered around zero with large scale parameters. However, the results were not satisfying, especially for small sample sizes. Thus, we discarded these priors and focused our attention on weakly-informative priors.

We refer to Gelman et al. (2008) for a discussion of suitable weakly-informative priors for logistic regression. The authors do not recommend the use of vague normal priors, and give preference to the location-scale family of the Student's t-distribution. In particular, they suggest a t-density function with 7 degrees of freedom and scale equal to 2.5, which is close to the likelihood of a single binomial trial. A more conservative choice is also proposed, which is a Cauchy prior with scale parameter equal to 2.5. However, Ghosh et al. (2018) show that in such cases sampling from the posterior is challenging in the presence of separation and, thus, it is recommended to use a t-distribution with degrees of freedom ν between 3 and 7. We use $\nu = 3$. The resulting prior is formalized in Equation 7 and referred to as the **Baseline prior**:

$$\beta_j \sim \mathit{Student}\left(\nu = 3, \mu = 0, s = 2.5\right). \tag{7}$$

4.3 Posterior estimation

For the simulation and real data application, the posterior distributions based on both strongly-informative and weakly-informative priors are numerically approximated. We use the No-U-Turn sampler, implemented in R (R Core Team, 2020) and Stan (Stan Development Team, 2019), which is a variant of the Hamiltonian Monte Carlo algorithm. Specifically, we used the R packages rstan (Stan Development Team, 2021) and rstanarm (Goodrich et al., 2020). The posterior distributions were obtained using four MCMC chains with samples of 7,000 each and 3,500 burn-in samples which ensured convergence of all chains.

5 Simulation study

5.1 The simulation framework

The simulation study is designed to evaluate the proposed strongly-informative priors under a variety of real-world settings. All settings involve the analysis of binary outcome variables, which is a common application in the social and health sciences. For example, a researcher might be interested in analyzing the propensity to commit a crime, become divorced, drop out of school, or become inflicted with an illness. In some cases, the binary outcome is unbalanced, i.e., there is a greater proportion of zeros than ones (or vice versa). For this reason, we consider both balanced and unbalanced outcomes in the simulation study (and also in the application in Section 6). To reflect further practical scenarios, we also consider different PS and NPS sizes when setting up the simulation. Because PS surveys are the gold-standard for inference, but large sample sizes can be prohibitively expensive, we focus our attention on a range of PS sizes, from very small (50-100 cases) to modestly large (up to 1,000 cases). On the contrary, we consider larger NPS sizes which we reasonably assume to be more affordable than similarly-sized PS surveys.

Based on these practical considerations, we assume that the outcome variable is generated from the logistic model in Equation 1 with two binary predictors: $X_{i1} \sim Ber(0.5)$ and $X_{i2} \sim Ber(0.5)$. In order to test the stability of the results, we consider three specifications for the population regression coefficients, $\beta = (\beta_0,$

β_1, β_2), namely:

$$\beta_{NEG} \in (0.5, -1.3, -0.9)$$
$$\beta_{MIX} \in (0.5, -1.3, 0.9)$$
$$\beta_{POS} \in (0.5, 1.3, 0.9)$$

These specifications consider both balanced and unbalanced scenarios for the outcome and different cell proportions when combining the three variables. The proportions of Y are 0.37, 0.57, and 0.81, respectively. Table 1 shows the cross-tabulation of the outcomes with the covariates under these three scenarios.

Table 1: Cross-tabulation of variables used in the three simulated populations.

Y	X_1	X_2	NEG	MIX	POS
0	0	0	0.09	0.09	0.09
1	0	0	0.16	0.16	0.16
0	1	0	0.17	0.17	0.04
1	1	0	0.08	0.08	0.21
0	0	1	0.15	0.05	0.05
1	0	1	0.10	0.20	0.20
0	1	1	0.21	0.12	0.02
1	1	1	0.04	0.13	0.23

Under this model, we simulate a population of size $N = 1,000,000$. The PS is then drawn from this population with simple random sampling without replacement (*srswor*). We consider different probability sample sizes, including very small and larger sizes, $n_{PS} \in \{50, 100, 150, 200, 300, 500, 750, 1000\}$.

For generating the NPS we first simulate a self-selected panel of individuals who declared their willingness to complete online surveys. From this panel we extract two simple random samples without replacement of different sizes, $n_{NPS} \in \{1000, 5000\}$. We assume that the panel population reflects an online access panel and is thus affected by self-selection. Indeed, the real-life process for joining the panel and eventually completing the survey includes different stages of selection (Valliant and Dever, 2011). First, individuals must have an internet connection and visit the recruitment website. Then they decide to join the panel completing all the required steps. For a specific survey, a sample of individuals is selected from the panel and they can decide whether to participate or not. For simplicity, we assume that all selected units will participate and fully complete the questionnaire (i.e., no unit nonresponse, item nonresponse, or break-offs) without measurement error, just as we assume for the PS survey. Thus, we assign to each population unit a positive probability of participation denoted by p. In general, we set p to be low in order to account for the selection process described above. However, due to the rise of *big data* sources and the potential of conducting surveys through social media, it may be possible to reach a very large, but very specific part of the population. Thus, we also consider higher values of p.

When the probability of participation depends directly on Y, we are in the case of non-ignorable selection bias (or missing not at random; MNAR). Otherwise, if p depends only on observed covariates then we have a missing at random (MAR) selection mechanism; that is, after controlling for X_1 and X_2 in the model, the coefficients will be unbiased. In the latter case, we account for all variables that explain the selection mechanism. However, the MAR assumption is strong and may not hold in practice. Thus, it is important to consider different selection mechanisms when evaluating the proposed data integration method.

We consider five selection mechanisms: (1) p depends on Y only (MNAR); (2) p depends on Y and X_1 (MNAR); (3) p depends on Y and X_2 (MNAR); (4) p depends on X_1 and X_2 (MAR); and (5) p depends on Y, X_1, and X_2 (MNAR). If the probabilities of participation are equal for all units, then there is no selection bias. To introduce bias we consider four scenarios of varying probabilities of participation p for specific subgroups defined by the value of the selection variables:

$$p = \begin{cases} \{0.10, 0.20, 0.50, 0.90\} & \textit{if the value of the selection variable(s) is 1} \\ 0.10 & \textit{otherwise} \end{cases}$$

where $p = 0.1$ reflects the case of no selection bias and $p = 0.9$ high selection bias.

Then the probability of participation p is used to generate the participation indicator $P_i \sim Ber(p_i)$ for $i \in \{1, ..., N\}$ for each individual in the population. It follows that the size of the panel N_{Panel} is random. Both PS and NPS are constructed cumulatively and thus, all cases in the smaller samples are always included in the larger ones. We consider standardized covariates for comparability and also because this can reduce auto-correlation in MCMC chains.

The simulation is repeated 100 times (the results were consistent with more repetitions). In order to compare the performance of the strongly-informative priors against the weakly-informative baseline prior, we consider the MSE of the posterior estimates. Given the true value of the generic coefficient β, namely β^*, the MSE is defined as follows:

$$\begin{aligned} MSE\left(\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})\right) &= Bias^2\left(\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})\right) + Var\left(\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})\right) \\ &= \bar{\pi}(\boldsymbol{\beta}|D_{PS}, D_{NPS}) - \beta^* + Var\left(\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})\right), \end{aligned} \tag{8}$$

where $\bar{\pi}(\boldsymbol{\beta}|D_{PS}, D_{NPS})$ is the mean of the posterior distribution for a given coefficient and $Var(\pi(\boldsymbol{\beta}|D_{PS}, D_{NPS})$ is the posterior variance. When describing the results, we always refer to the median value of MSE estimates obtained from the 100 iterations.

When working with small samples sizes and categorical variables, the problem of quasi-complete separation and the Hauck-Donner effect may arise (Yee, 2021). In such cases, even if the algorithm converges without any evidence of predicted probabilities numerically equal to 0 or 1, coefficients and standard errors can assume very large and implausible values. We observed the presence of such issues for a few small samples (mainly of size 50). For this reason, we use the median instead of the mean across all simulations to limit the influence of a small number of outliers. The mean values are also available in the Shiny web app.

5.2 Simulation results

For ease of visualization, Figure 1 shows the median MSEs of the regression coefficients for three selected informative priors: Power, Distance, and Mixed-Distance. The results are shown for the case of no selection bias ($p = 0.1$) and high selection bias ($p = 0.9$) with sample sizes restricted to $n_{PS} \leq 500$ and for $n_{NPS} = 5000$. For $n_{PS} \geq 500$, the MSEs of the informative priors are indistinguishable from those of the baseline prior, and between $n_{NPS} = 1000$ and $n_{NPS} = 5000$ the differences in MSEs are minuscule.

The first three rows of each panel in Figure 1 show that for the MAR and MNAR selection scenarios with no bias ($p = 0.1$), all strongly-informative priors produce remarkably lower MSEs than the baseline (weakly-informative) prior, especially when $n_{PS} < 200$. The only exception is for the intercept, where the Mixed-Distance prior, by design, yields MSEs close to the baseline prior. In the presence of large selection bias ($p = 0.9$; the bottom three rows of each panel in Figure 1), the strongly-informative priors also reduce the MSEs relative to the baseline prior, but to a lesser extent than the no bias case. The MSE reduc-

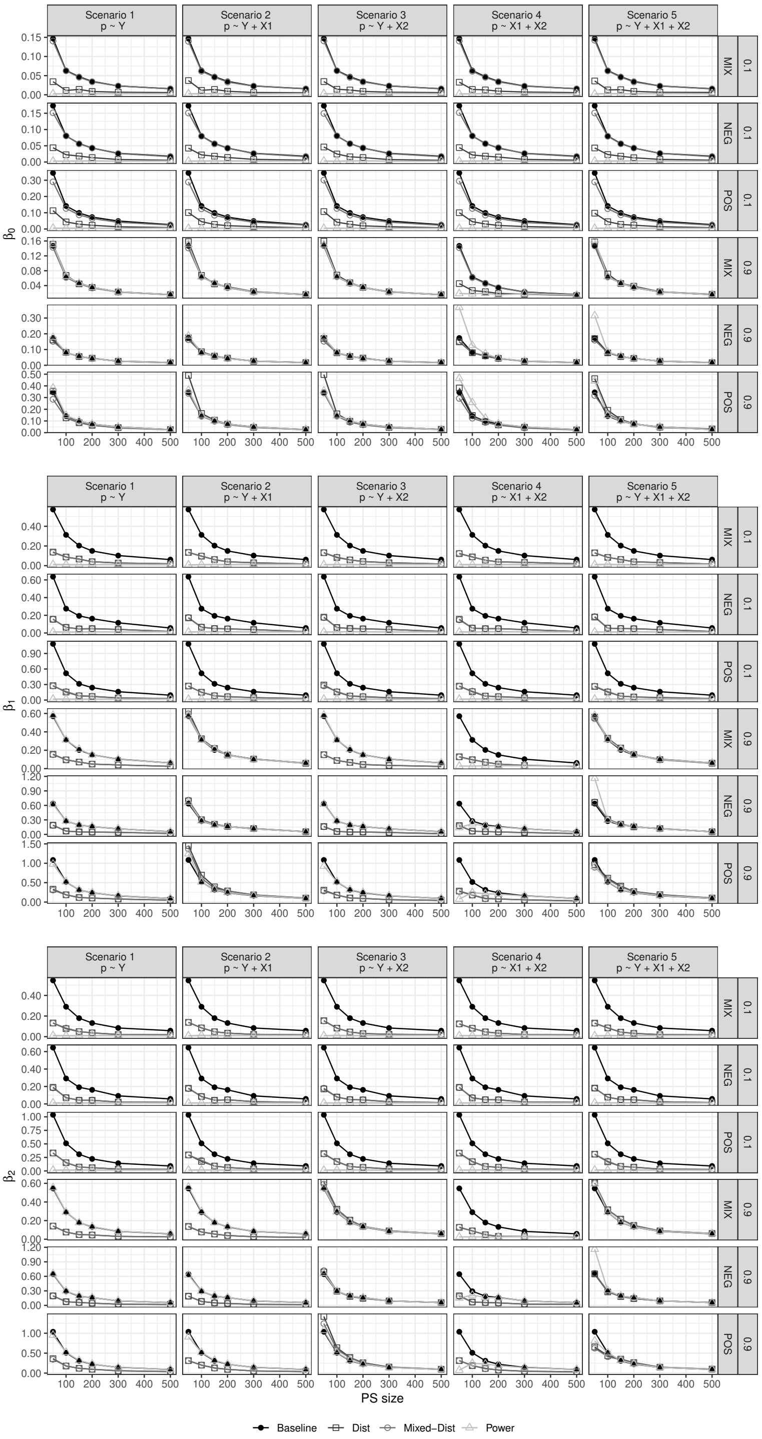


Figure 1: Median MSEs for regression coefficients averaged over 100 simulations and for $n_{NPS} = 5000$ by probability sample (PS) size.
Note: Four priors are considered: Distance (Dist), Mixed-Distance (Mixed-Dist), Power, and Baseline. Each panel shows the combination of the five selection scenarios, the three population models: MIX - $\beta_{MIX} \in (0.5, -1.3, 0.9)$, NEG - $\beta_{NEG} \in (0.5, -1.3, -0.9)$, and POS - $\beta_{POS} \in (0.5, 1.3, 0.9)$, and the case with no selection bias ($p = 0.1$) and high selection bias ($p = 0.9$).

tions are hardly affected by whether the outcome is balanced (MIX) or unbalanced (POS and NEG), with the exception that the Power prior tends to perform slightly worse than the baseline prior for very small unbalanced samples. This exception notwithstanding, we may conclude from the figure that the strongly-informative priors produce MSEs that are generally smaller or, in the case of high selection bias, similar to those of the baseline prior for all coefficients and selection scenarios.

The full results (including the bias and variance components of the MSE) for all priors and sample sizes are available in the Shiny web app under the tab *Simulation/Results.* The results clearly show a bias-variance trade-off. In general, the strongly-informative priors lead to the posterior estimates being more biased than the estimates based on the baseline prior (which uses PS data only). However, the results also demonstrate that a careful scaling of the NPS information in the prior reduces the variance and, thus, improves the MSE relative to the baseline prior.

The bias-variance trade-off is especially critical for the scenario with the highest level of selection bias ($p = 0.9$), where the Distance prior yields the smallest MSE relative to the other prior formulations. In the worst case, the Distance prior performs similarly to the baseline prior, while for the Distance-log10 and Distance-log priors the MSEs can become slightly higher than those of the baseline prior. The Power prior performs especially well for small PS sizes (50-150 observations) and in the MAR selection scenario.

To summarize the full results, we look at the overall performance of each prior across all simulation settings and coefficients, keeping the NPS size equal to 5000. The first column of Table 2 shows the percentage of instances where the MSEs obtained using a strongly-informative prior is lower than the MSE obtained using the baseline prior. With this measure, the Mixeded-Distance and Mixed-Distance-log10 priors perform best, each yielding 82% of the MSEs smaller than the baseline prior. Further, to better understand how large the differences are when the MSEs of the strongly-informative priors are *worse* (i.e. larger) than those of the corresponding baseline prior, we calculate the relative difference (RD) between those MSEs defined as:

$$RD = \frac{MSE_{INF} - MSE_{BASE}}{MSE_{BASE}} \quad \text{if } MSE_{INF} > MSE_{BASE}. \tag{9}$$

The other columns of Table 2 show the percentage of instances where the relative difference is lower than 5%, 10%, 20%, and 30%, respectively, when $MSE_{INF} > MSE_{BASE}$. In general, a relative difference up to 5% may be considered small, moderate up to $10 - 20\%$, and large otherwise, though we acknowledge such judgments are subjective.

We observe that the Distance and Mixed-Distance priors perform best, i.e. almost always with relative differences smaller than 30% (96% and 100%, respectively), followed by the Power, Mixed-Distance-log10, Distance-log10, Mixed-Distance-log10, and Distance-log priors.

Strongly-Inf. Priors	$MSE_{INF} \leq MSE_{BASE}$	$MSE_{INF} > MSE_{BASE}$			
		$\leq 5\%$ RD	$\leq 10\%$ RD	$\leq 20\%$ RD	$\leq 30\%$ RD
Dist	78	45	73	91	96
Mixed-Dist	82	62	82	98	100
Dist-log	65	5	11	21	29
Mixed-Dist-log	78	17	27	38	49
Dist-log10	72	17	34	60	74
Mixed-Dist-log10	82	33	45	72	85
Power	64	74	80	86	89

Table 2: The percentage of instances where the MSE obtained using the informative prior is lower than the MSE obtained using the corresponding baseline prior ($MSE_{INF} \leq MSE_{BASE}$), and the percentage of instances where the relative difference (RD) is lower than a prespecified threshold <5%, <10%, <20%, and <30% for the instances where $MSE_{INF} > MSE_{BASE}$.

Note: The priors are: Distance (Dist), Mixed-Distance (Mixed-Dist), Distance-log (Dist-log), Mixed-Distance-log (Mixed-Dist-log), Distance-log10 (Dist-log10), Mixed-Distance-log10 (Mixed-Dist-log10), and Power. A detailed breakdown of results by sample sizes, selection and bias scenarios, and balanced/unbalanced outcomes are available in the Shiny web app under the menu *Simulation/Summary.*

The simulation study demonstrated that the use of strongly-informative priors is beneficial to improve the MSEs of coefficient estimates from logistic regression models, especially when the PS size is smaller than 200 observations. However, the amount of such improvements depends on the level of selection bias in the NPS data and the selection mechanism. In the worst-case selection bias scenario ($p = 0.9$), there is evidence that the MSEs from the strongly-informative priors are similar to those of the baseline prior. In the rare instances where $MSE_{INF} > MSE_{BASE}$, the differences are usually relatively small. The MSE reductions are mainly driven by a reduction in the posterior variance which offsets the increase in bias. Overall, the Mixed-Distance prior performs best, with 82% of the MSEs being lower than those of the baseline prior and the remaining MSEs never exceeding a 30% relative difference (Table 2).

6 Application: American Trends Panel

6.1 The data

In order to evaluate the method in a practical setting, a real data application is presented with an actual PS survey, the American Trends Panel (ATP; Keeter, 2019), and nine parallel NPS web surveys carried out by different vendors, which reflect real-world selection scenarios. The ATP is the Pew Research Center's probability-based online panel used for conducting public opinion research. It is representative of the general population aged 18 years and older in the U.S and covers both the online and the offline population – before 2016 offline individuals were provided with paper-questionnaires or were interviewed by telephone, while in the subsequent years panelists were supplied with the necessary technological tools.

Panel members were originally recruited in 2014 from the Political Polarization and Typology survey (Dimock et al., 2014), a national RDD survey. Additional panelists have been recruited via random-digit-dial telephone surveys in 2015, 2017, and 2018. Panelists are invited to complete at least one survey in each monthly wave. Survey duration is 15 minutes and a system of financial incentives is implemented. The data we analyze were collected in waves 5 ([dataset] Pew Research Center, 2014a), 7 ([dataset] Pew Research Center, 2014b), and 10 ([dataset] Pew Research Center, 2015a) in 2014 and 2015. We note that Wave 5 was part of a mode experiment in which panel members who use the internet were randomly assigned to either web or telephone mode. We analyze the full sample independently of the mode, though a sensitivity check yielded the same conclusions when excluding the telephone cases.

During the same period, Pew sponsored the parallel collection of nine NPS web surveys from different vendors (Kennedy et al., 2016; [dataset] Pew Research Center, 2015b). The same questionnaire was administrated to all respondents with the questions overlapping with those in the ATP, but in different waves. The required sample size was about 1,000 respondents. All vendors implemented quota sampling based on different variables, including age, gender, education, and also other non-demographic variables. More survey details, including target population, response rate, and sample sizes are available in Kennedy et al. (2016) and in the Shiny web app under the menu *Real Data Analysis/Data*.

Six categorical outcome variables are considered. In the case of non-binary classification, variables were re-coded in a binary fashion. All question wordings are available in the Shiny web app under the menu *Real Data Analysis/Data/Variable*

Coding. The questions relate to smoking at least 100 cigarettes in one's entire life (SMOKING; 1 = yes, 0= no), volunteering in the last 12 months (VOLUNTEERING; 1 = yes, 0 = no), health insurance coverage (HEALTHCARE COVERAGE; 1 = yes, 0 = no), frequency of voting in local elections (ALWAYS VOTE; 1 = always, 0 = otherwise), how many people they trust in their neighborhood (NEIGHBORHOOD TRUST; 1 = all people, 0 = otherwise), and how safe they feel when walking in their neighborhood at night (NEIGHBORHOOD SAFETY; 1 = very safe, 0 = otherwise).

Covariates include binary age (AGE; 1 = 50+, 0 = otherwise), gender (GENDER; 1 = male, 0 = female), education (EDU; 1 = college graduate or higher, 0 = otherwise), and the continuous survey weight variable (SVY WEIGHT; log-transformed). Only some vendors provided weights but the Pew team constructed ATP-style weights for all NPS surveys with the aim to reduce selection bias through a raking adjustment to population benchmarks. We use these weights in the analysis by including them as covariates in the regression models. Before analyzing the data, we drop all observations with missing values and standardize all covariates. More details on the percentage of missing data and the final sample size for each variable are available in the Shiny web app (*Real Data Analysis/Data/Description*).

Figure 2 shows the estimated proportions with 95% confidence intervals for a selection of outcome variables (smoking, always vote, and neighborhood trust) across all samples. While there is no evidence of significant differences between the ATP and NPS estimates for the smoking and neighborhood trust variables, differences are evident for the always vote variable in NPS surveys C, D, F, H, and I. The proportions of the other outcome variables are presented in the Shiny web app (*Real Data Analysis/Data/Additional Plots*).

Logistic regression coefficient estimates, based on maximum-likelihood estimation, are also provided separately for the PS and NPS survey data. As an example, Figure 3 shows the estimates for the smoking variable. The figures for always vote and neighborhood trust are in Appendix A and those for the remaining variables are available in the Shiny web app (*Real Data Analysis/Data/Additional Plots*). Detailed results about the parameter estimates, standard errors, and goodness-of-fit statistics are available in the Online Appendix. Figure 3 shows that the NPS regression coefficients for smoking differ only slightly from the ATP estimates, with the exception of the education variable in sample NP-D and the gender variable in sample NP-I. For the neighborhood trust variable (Figure 7), the NPS and ATP coefficients are very similar, except for age in samples NP-A, NP-B, and

NP-D. For always vote (Figure 6), there are notable differences for the education variable where coefficients have opposite signs for all NPS surveys. The same is true for gender in NP-I.

From these comparative analyses we conclude that the descriptive estimates are more dissimilar between the ATP and NPS surveys compared to the regression estimates, which is consistent with the literature (Pasek, 2016). Contrary to the simulation study, which included scenarios with high selection bias, it appears that in the considered real data the regression coefficients are not heavily affected by a high level of bias.

To apply the proposed methodology, we simulate a situation in which only a small PS size survey is conducted along with a parallel NPS survey. We consider PS sizes $n_{PS} \in \{50, 100, 150, 200, 300, 500\}$. The PS data are drawn with *srswor* from the full ATP data and are assumed to be unbiased. The samples are constructed cumulatively, such that respondents selected for the smaller samples are included in the larger samples. For the NPS surveys, the original sample sizes (approximately 1,000 each) are used.

To compute the posterior bias, the true values are defined by the vector of ML estimates obtained using the full ATP sample (where the sample size ranges between 3,106-3,331 respondents depending on the outcome variable of interest), namely $\boldsymbol{\beta}^*$, which provides an unbiased result by assumption. The entire procedure is repeated 100 times and, as in the simulation study, the median MSE values are reported across all repetitions. Only for the healthcare coverage variable, which is highly unbalanced, the model was not estimable for some iterations which is likely due to lack of variation for the smallest PS sizes. As an *ad hoc* remedy, additional iterations were performed for this outcome variable until 100 estimable results were obtained.

6.2 Results

For brevity, we discuss the results for only a selection of outcome variables: smoking, always vote, and neighborhood trust, and only one NPS data source (NP-A). The results for all other outcome variables and NPS data sources are shown in the Shiny web app under the menu *Real Data Analysis/Results and Summary.* Figure 4 shows the median posterior bias, variance, and MSE for the smoking outcome across the 100 repetitions and for the selected priors as in the simulation study. The figures for always vote and neighborhood trust are shown in Appendix A (Figs. 8 and 9).

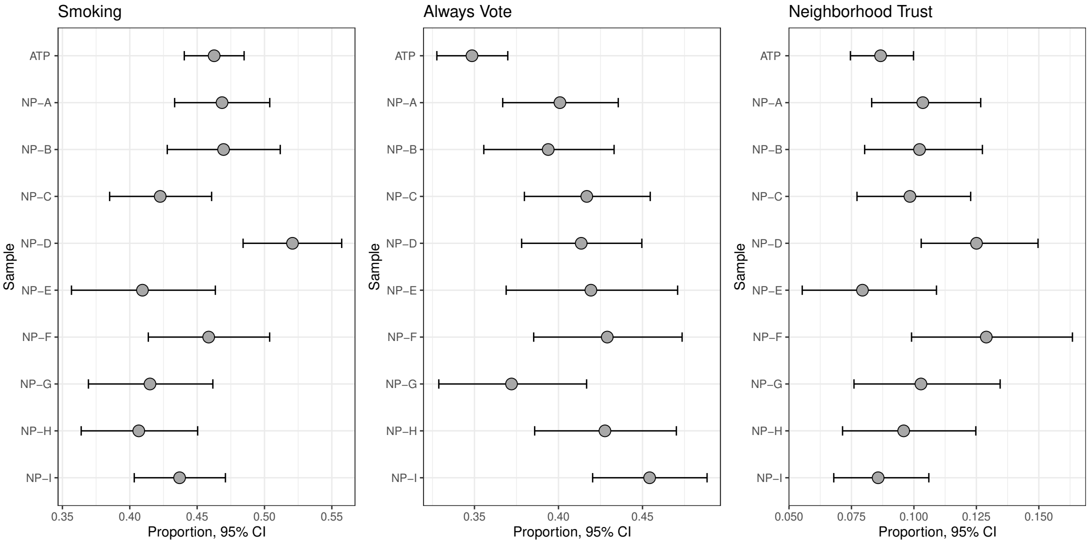


Figure 2: Estimated sample proportions (weighted) with 95% confidence intervals for a selection of outcome variables for each survey.

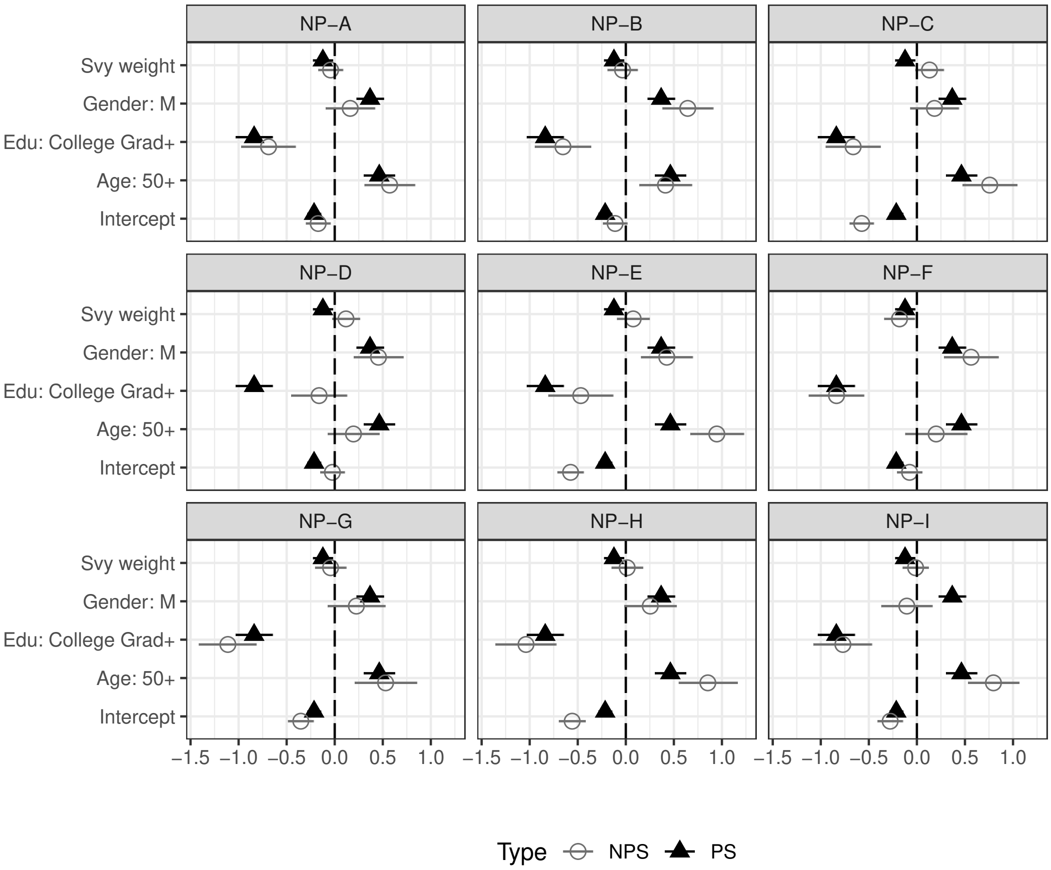


Figure 3: Maximum likelihood estimates of logistic regression coefficients and 95% confidence intervals for the smoking outcome variable in the ATP (triangles) and nine non-probability surveys (circles).

For the smoking and neighborhood trust outcomes, which are affected by a low level of selection bias (see Figures 3 and 7), all strongly-informative priors produce lower MSEs than the baseline prior for all coefficients particularly when $n_{PS} < 200$. The reduction in posterior MSEs is mainly driven by a reduction in the posterior variance. The Power prior leads to the largest reductions in MSE relative to the baseline prior, especially for the smallest PS sizes ($n_{PS} \in \{50, 100\}$). However, the more dissimilar are the maximum likelihood (ML) coefficient estimates between the PS and NPS surveys, the more similar the MSEs of the estimates based on strongly-informative priors become to those of the baseline prior. Indeed, for the neighborhood trust variable in sample NP-A, the maximum-likelihood estimate for the age coefficient is significantly different from the ATP estimate (see Figure 7), which yields MSE curves for the strongly-informative priors that are similar to those of the baseline prior (Figure 9).

For the always vote variable, the strongly-informative priors reduce the MSEs with respect to the baseline prior for almost all coefficients, or in the worst case, the MSE curves are similar to the baseline. In particular, as expected from the analysis of the maximum-likelihood regression coefficients (see Fig. 8), there is no significant reduction in the MSEs using the strongly-informative priors for the education coefficient. Indeed, the Distance and the Mixed-Distance priors produce MSEs that are equal to or slightly lower than the baseline prior. In contrast, the Power prior produces larger MSEs than the baseline prior for PS sizes between 100 and 200. Thus, while the distance priors protect against excessive selection bias in the education coefficient, this is not true of the Power prior, which decreases the posterior variance but not enough to offset the large education bias. In general, the largest reduction in MSEs relative to the baseline prior is achieved through the Distance-log prior, which is driven by reductions in the posterior variance.

As for the simulation study (Section 5), we summarize the results of the application in Table 3 by showing the percentage of instances where the MSE of a coefficient using the strongly-informative priors is lower than the MSE obtained using the baseline prior, and, for the instances where the performance of the strongly-informative priors is worse (i.e. $MSE_{INF} > MSE_{BASE}$), the percentage of instances where the relative difference (Eq. 9) is lower than 5% and 30%, representing the two extremes, across all PS sizes and NPS surveys used to construct the strongly-informative priors. In the Shiny web app under the menu *Real Data Analysis/Summary*, such a table is available for each combination of n_{PS} and NPS survey. As inferred from Section 6.1, most regression coefficients from the NPS data are not affected by a high level of selection bias, thus, the strongly-informative priors lead to lower MSEs in most cases, as expected.

The neighborhood trust and healthcare coverage variables are two particularly notable cases in terms of MSE reduction using the strongly-informative priors. In more than 99% of instances, the distance priors yield lower MSEs than the baseline prior, indicating improvements in the MSEs regardless of which NPS survey is used to supply the prior information. The Power prior performs similarly and such percentages are about 98% and 97% for these two outcomes, respectively. Similar results are achieved when considering the smoking and neighborhood safety outcomes, where the best priors are the Mixed-Distance-log10 (99%) and the Distance (97.4%) for the two outcomes, respectively. For always vote, the best prior is the Distance-log10 (93.3%) and for volunteering it is the Mixed-Distance-log10 together with the Distance prior (97.8%). The performance of the Power prior is generally worse compared to the distance priors. The worst result is for the always vote and neighborhood safety outcomes where in only 67% of cases does the Power prior produce lower MSEs than the baseline prior. Nevertheless, for all strongly-informative priors the relative difference usually does not exceed 30% when $MSE_{INF} > MSE_{BASE}$.

In summary, there is evidence that the strongly-informative priors reduce MSEs for logistic regression coefficients relative to a weakly-informative baseline prior in real-world settings. The smaller MSEs are driven by a reduction in variability and the largest reductions occur for small PS sizes (50-200 observations). These results are also consistent with the simulation study. Moreover, the results slightly vary according to which NPS survey is used. When selection bias is low (as for neighborhood trust, healthcare coverage), all priors perform similarly well, although the largest reductions in MSEs are achieved with the Power prior for very small PS sizes (50-100 observations) and with the Distance-log prior (and its mixed version) for sample sizes up to 200 observations. However, as the selection bias increases, the Distance or Distance-log10 priors (and their mixed versions) tend to be superior. The mixed priors usually perform better than their non-mixed counterparts and the results are generally consistent across balanced and unbalanced outcome variables.

	Smoking			Always vote			Volunteering		
		$MSE_{INF} > MSE_{BASE}$			$MSE_{INF} > MSE_{BASE}$			$MSE_{INF} > MSE_{BASE}$	
Strongly-Inf. Priors	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD
Dist	94.8	14.3	100	92.6	65	100	91.5	65.2	100
Mixed-Dist	97.0	100	100	90.4	69.2	100	97.8	83.3	100
Dist-log	92.2	14.3	95.2	79.6	40	94.6	87.4	35.3	97.1
Mixed-Dist-log	96.3	60	100	82.2	35.4	95.8	96.7	44.4	100
Dist-log10	94.8	14.3	100	93.3	55.6	100	91.1	66.7	100
Mixed-Dist-log10	99.6	100	100	89.3	65.5	100	97.8	83.3	100
Power	90.7	56	76	67.8	39.1	80.5	73.3	65.3	76.4

	Neighborhood Trust			Neighborhood Safety			Healthcare Coverage		
		$MSE_{INF} > MSE_{BASE}$			$MSE_{INF} > MSE_{BASE}$			$MSE_{INF} > MSE_{BASE}$	
Strongly-Inf. Priors	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD	$MSE_{INF} \leq MSE_{BASE}$	$\leq$ 5% RD	$\leq$ 30% RD
Dist	99.6	0	100	97.4	85.7	100	97.4	50	100
Mixed-Dist	99.3	50	100	95.6	91.7	100	95.6	50	100
Dist-log	99.6	0	100	93.3	27.8	100	93.3	75	100
Mixed-Dist-log	99.6	0	100	94.1	31.2	100	94.1	50	100
Dist-log10	99.6	100	100	97.4	57.1	100	97.4	0	100
Mixed-Dist-log10	99.3	50	100	96.7	77.8	100	96.7	50	100
Power	98.2	60	100	66.3	69.2	90.1	66.3	62.5	100

Table 3: The percentage of instances where the MSE obtained using the strongly-informative prior is lower than the MSE obtained using the corresponding baseline prior ($MSE_{INF} \leq MSE_{BASE}$), and the percentage of instances where the relative difference (RD) is lower than a pre-specified threshold of <5% and <30% for the instances where $MSE_{INF} > MSE_{BASE}$.

Note: The priors are: Distance (Dist), Mixed-Distance (Mixed-Dist), Distance-log (Dist-log), Mixed-Distance-log(Mixed-Dist-log), Distance-log10 (Dist-log10), Mixed-Distance-log10 (Mixed-Dist-log10), and Power. A detailed breakdown of results by PS sizes and NPS surveys are available in the Shiny web app under the menu *Real Data Analysis/Summary*.

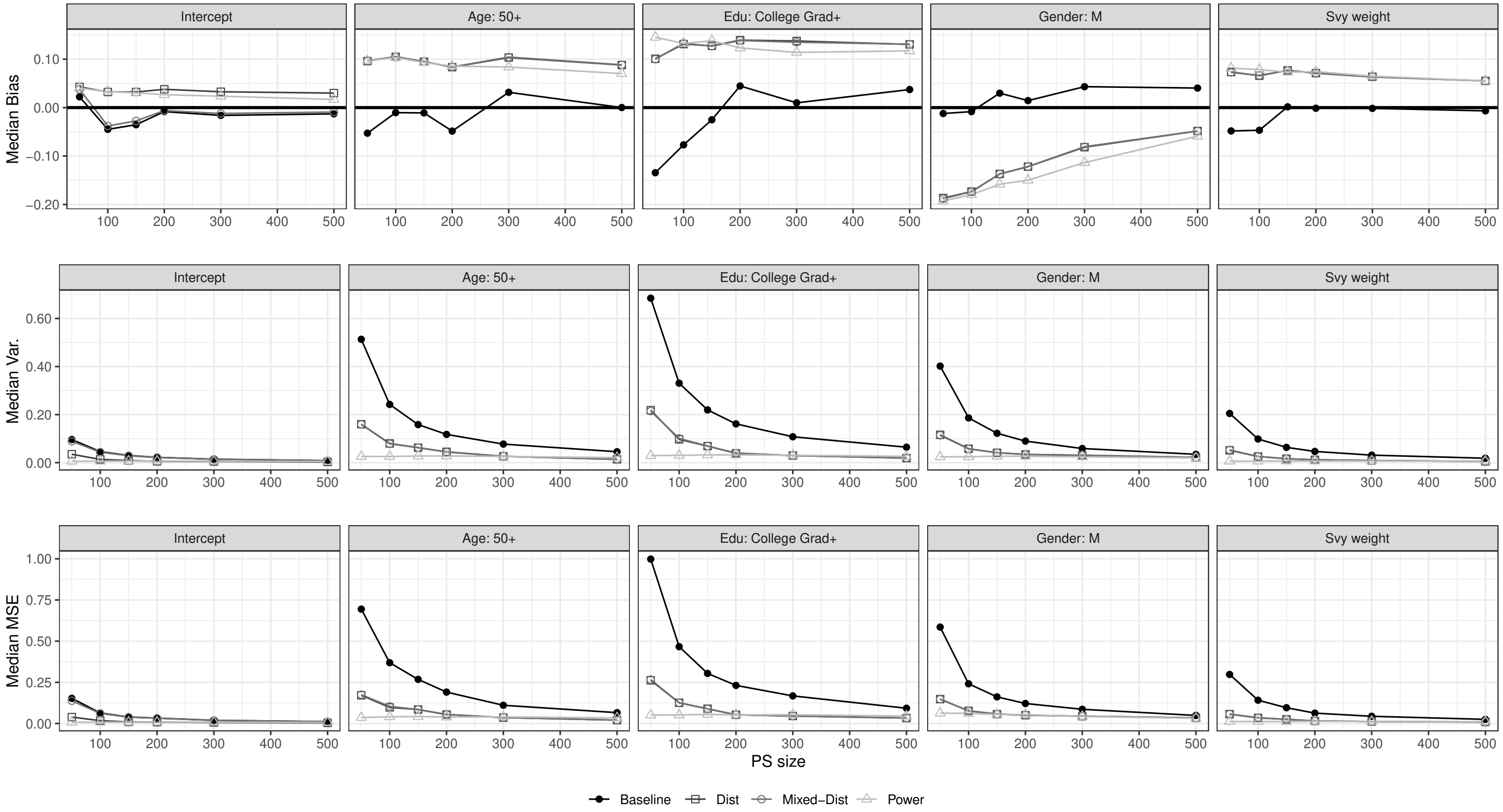


Figure 4: Median Bias, Variance, and MSE of coefficient estimates for the smoking outcome using the NP-A survey to supply the prior information.

Note: Four priors are considered: Distance (Dist), Mixed-Distance (Mixed-Dist), Power, and Baseline.

6.3 Cost analysis

The simulation and real-data application showed that supplementing a small PS survey (100-200 cases) with prior information from a parallel NPS survey results in lower MSEs of logistic regression coefficients compared to not supplementing. To the extent that NPS surveys are less expensive than PS surveys, these results suggest that the same MSE values might be achieved at a lower cost with an integrated sample compared to a larger and potentially more expensive standalone PS survey. Cost savings are an important justification for integrating PS and NPS data and should be considered when applying the proposed data method. To explore the potential for cost savings, we implement a cost analysis which takes into account hypothetical, yet realistic, costs for the ATP and NPS data sources.

To demonstrate the extent to which the strongly-informative priors lead to cost savings (or losses), we first estimate the expected cost of fielding a PS-only survey with baseline prior that would achieve the same MSE as fielding parallel PS and NPS surveys with strongly-informative priors, and then compare it to the cost of fielding the parallel surveys. The cost analysis can be performed interactively within the Shiny web app under the menu *Real Data Analysis/Cost Analysis*, where users can specify different per-respondent costs for the PS and NPS surveys. For illustration, here we assume the cost per respondent to be \$5 in the NPS survey and \$30 in the PS survey. The cost of the PS survey equates to roughly \$2 per interview minute for a 15-minute interview, which is consistent with the cost of similar PS surveys[2].

The cost analysis follows a three-step approach. First, we fit a model to learn the cost-MSE structure of the PS data. To this end, we run a linear regression model of hypothetical PS survey costs on 100 repetitions of MSEs obtained using the baseline prior from the real-data application (Section 6). Both the outcome (cost) and the covariates (the MSEs for each coefficient) are log-transformed. Second, we use this model to predict the expected PS cost given the median MSEs obtained using the strongly-informative priors and each NPS survey. Third, we calculate the total costs for the PS-only survey, as well as the blended PS and NPS surveys by multiplying the per-respondent cost of the PS and NPS surveys by their respective sample sizes. To compute the expected cost savings (or losses), we compare the expected PS-only survey cost with the cost of the blended PS and NPS surveys.

[2] https://openpanelalliance.org/pricing.php

The results reveal a mixed picture as the presence and amount of cost savings depends on the prior structure and the NPS survey used. Figure 5 shows the distribution of percent savings(+)/losses(-) for the smoking outcome variable across the nine NPS surveys for different PS sizes (50, 100, 150) and for the Distance, Distance-log, Distance log-10, and Power priors. Figures 10 and 11 show the same plots for the always vote and neighborhood trust outcomes, respectively. An interactive visualization is available for all outcome variables, PS sizes, and priors in the Shiny web app under (*Real Data Analysis/Cost Analysis/Savings Distribution*). Looking at the three figures, it is evident that for a PS size of 50, the Power prior nearly always leads to higher cost savings compared to the other strongly-informative priors. The Distance-log prior also leads to cost savings in most cases. The Distance and the Distance-log10 priors always lead to losses. However, as the PS size increases ($n_{PS} \geq 100$), the pattern reverses and the Power prior starts to generate losses rather than savings and the performance of the distance priors improves, especially the Distance-log prior which generates the largest cost savings. In general, the median level of cost savings for the mixed-distance priors is lower than for the non-mixed formulations, except for the healthcare and neighborhood trust outcomes, where the median savings are similar for both prior types or slightly higher for the mixed priors.

Table 4 summarizes, for each PS size, the prior formulation (and corresponding NPS survey) that leads to the largest cost savings (in percent) for the selected outcomes: smoking, neighborhood trust, and always vote. Results for the additional outcomes are available in the Shiny web app under (*Real Data Analysis/Cost Analysis/ Max. Savings*). For small PS sizes (50-100), the Power prior yields the largest percent cost savings across the three outcomes (range: 37%-68%). In the case of larger PS sizes (150-500), the distance priors, specifically the Distance-log (range: 42%-58%) and Mixed-Distance-log (56%-61%) priors produce the largest cost savings for the three outcomes. Thus, there is indications of a potential cost savings (up to 68%) by integrating the PS and NPS surveys using informative priors under assumed per-respondent costs. However, whether savings occur and the amount of those savings varies depending on which NPS survey is used to construct the prior.

7 Discussion

Integrating probability sample (PS) surveys with non-probability sample (NPS) data has received a lot of attention recently, mainly due to the convenience,

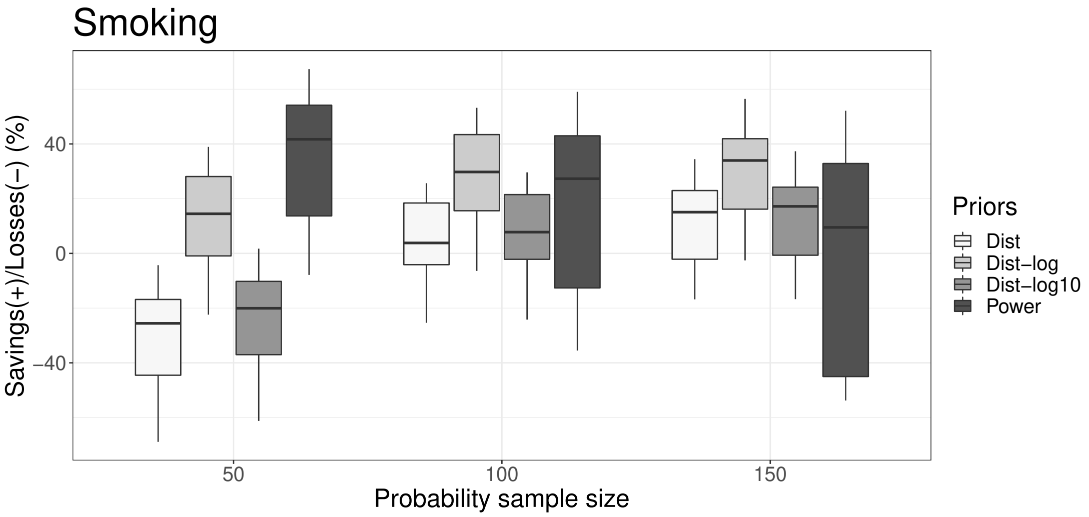


Figure 5: Percentage cost savings(+)/losses(-) for the smoking outcome.

timeliness, and cost-effectiveness of online access panels and the availability of *big data* sources. However, the presence of selection bias in NPS data is why researchers often view them as a supplement, rather than a replacement, for PS survey data. While previous data integration approaches have focused on finite population inference, approaches for analytic inference about model parameters are still emerging. This article contributes to this growing research area by building on previous Bayesian data integration approaches to improve analytic inference about parameters of logistic regression models, which is of great importance in the social sciences for studying attitudes, behaviors, and characteristics of populations. We proposed several novel strongly-informative priors that exploit auxiliary information from a parallel NPS survey data to improve coefficient estimates from small PS size surveys.

The strongly-informative priors were evaluated through a simulation study which showed that they achieve smaller MSEs for coefficient estimates compared to those achieved exclusively using PS survey data with a weakly-informative baseline prior. This was particularly true for the case of no (or low) selection bias in the NPS data and for PS sizes less than 200, where the Distance-log and the Power priors produced the smallest MSEs, effectively driving down the posterior variance. In the case of large selection bias, the strongly-informative priors performed mostly similarly to the baseline prior with the Distance prior being superior to the other strongly-informative prior formulations for small PS sizes. We then evaluated the approach in a real-data application by modeling six binary outcomes from an actual PS survey and several parallel NPS surveys

PS size	Results	Smoking	Neighborhood Trust	Always Vote
50	**Strongly-Inf. Prior**	Power	Power	Power
	NPS survey	NP-A	NP-I	NP-H
	Exp. Cost (Base. prior)	$20,250	$20,433	$10,359
	Blended Cost (Inf. prior)	$6,610	$6,500	$6,535
	Savings %	67.36	68.19	36.91
100	**Strongly-Inf. Prior**	Power	Power	Dist-log
	NPS survey	NP-A	NP-I	NP-H
	Exp. Cost (Base. prior)	$19,807	$19,746	$11,516
	Exp. Cost (Best prior)	$8,110	$8,000	$8,035
	Savings %	59.05	59.48	30.23
150	**Strongly-Inf. Prior**	Dist-log	Mixed-Dist-log	Dist-log
	NPS survey	NP-A	NP-H	NP-H
	Exp. Cost (Base. prior)	$22,075	$23,998	$18,234
	Exp. Cost (Best prior)	$9,610	$9,535	$9,535
	Savings %	56.47	60.27	47.71
200	**Strongly-Inf. Prior**	Dist-log	Mixed-Dist-log	Dist-log
	NPS survey	NP-A	NP-H	NP-H
	Exp. Cost (Base. prior)	$26,710	$28,809	$23,173
	Exp. Cost (Best prior)	$11,110	$11,035	$11,035
	Savings %	58.4	61.27	52.38
300	**Strongly-Inf. Prior**	Dist-log	Mixed-Dist-log	Dist-log
	NPS survey	NP-A	NP-H	NP-H
	Exp. Cost (Base. prior)	$30,677	$34,932	$27,357
	Exp. Cost (Best prior)	$14,110	$14,035	$14,035
	Savings %	54.00	59.82	48.70
500	**Strongly-Inf. Prior**	Dist-log	Mixed-Dist-log	Dist-log
	NPS survey	NP-A	NP-I	NP-H
	Exp. Cost (Base. prior)	$36,352	$45,870	$34,732
	Exp. Cost (Best prior)	$20,110	$20,000	$20,035
	Savings %	44.68	56.4	42.32

Table 4: Best performing strongly-informative priors in terms of percent cost savings for a selection of outcome variables and different PS sizes.

reflecting different selection scenarios one might face in practice. The strongly-informative priors again yielded significant reductions in MSEs, compared to the baseline prior, especially for the smaller PS sample sizes. The Power prior was superior to the other strongly-informative priors in terms of MSE reduction for very small sample sizes (50-100), whereas the Distance-log prior and its mixed version performed better for slightly larger sample sizes (150-200). For PS sizes larger than 200, all of the strongly-informative priors performed similarly to the baseline prior with respect to MSEs.

An important novelty of the method lies in its ability to achieve the same MSE values as would a larger (and likely more expensive) PS-only survey at a potentially lower cost. Using assumed but realistic cost data for the parallel PS and NPS surveys, we showed indications of potential cost savings for the informative priors for different PS sizes. In general, for a PS survey of 50-100 respondents, the Power prior showed high potential cost savings (up to 68%), while the Distance-log and its mixed version were among the best performers for larger PS sizes (achieving potential cost savings up to about 60%). Thus, researchers with low-to-moderate budgets may benefit from using the proposed data integration strategy to minimize both costs and errors.

As a general recommendation for practitioners, the Power prior appears to be the most appropriate choice for small PS sizes (up to 100 observations) and the Distance-log and Mixed-Distance-log priors for larger PS sizes. These recommendations hold regardless of the outcome variable and the NPS survey considered. Nevertheless, we recommend performing a sensitivity analysis and comparing estimates obtained using different priors.

The present study entails some limitations. First, the method rests on the assumption that the PS survey is unbiased or less biased than the parallel NPS survey. This assumption may not always hold in practice. In addition, the method does not account for measurement errors, which may differ between PS and NPS survey data (Einarsson et al., 2022). We leave these topics for future work. Moreover, it would be worthwhile to extend the current framework to other types of categorical variables (e.g. multinomial, ordinal) and account for complex sample design features (e.g. stratification). The approach may also be extended in a multivariate setting by taking into account the distributions of several outcomes simultaneously. Exploring alternative methods for selecting the Power parameter is another topic for future development.

In conclusion, while many researchers are moving away from large and expensive PS surveys and shifting towards more convenient and less expensive NPS surveys, integrating both sample types in a way that overcomes their respective weaknesses is an attractive approach. The proposed data integration method can be easily implemented in any statistical software which supports Bayesian computation. To assist researchers, an R Shiny web app has been developed which provides the replication code for applying the methodology and allows readers to browse the full results of the simulation and application in more detail (see Appendix B). In addition, a key feature of the Shiny app is the possibility to dynamically implement the cost analysis with user-entered PS and NPS cost data. This may be useful for practitioners interested in collecting and integrating parallel PS and NPS survey data and wish to compare different cost scenarios.

Appendices

A Additional materials for the application

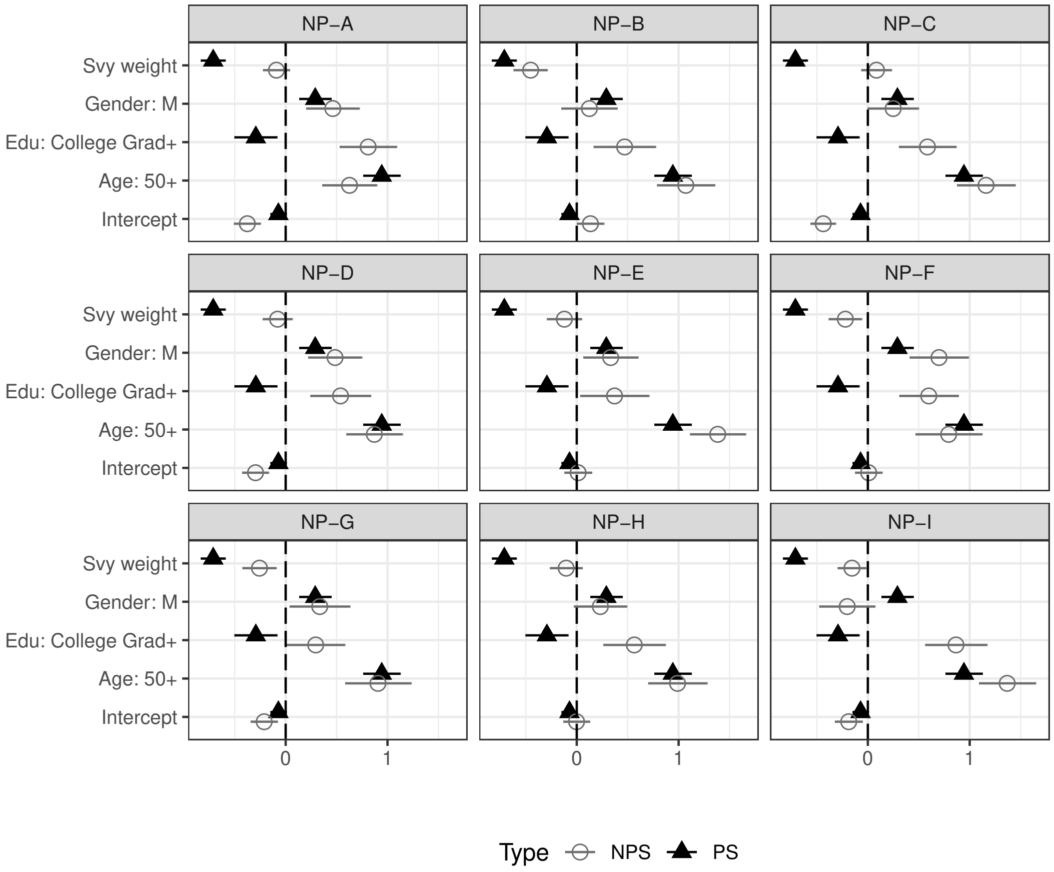


Figure 6: Maximum likelihood estimates of logistic regression coefficients and 95% confidence intervals for the always vote outcome variable in the ATP (triangles) and nine non-probability surveys (circles).

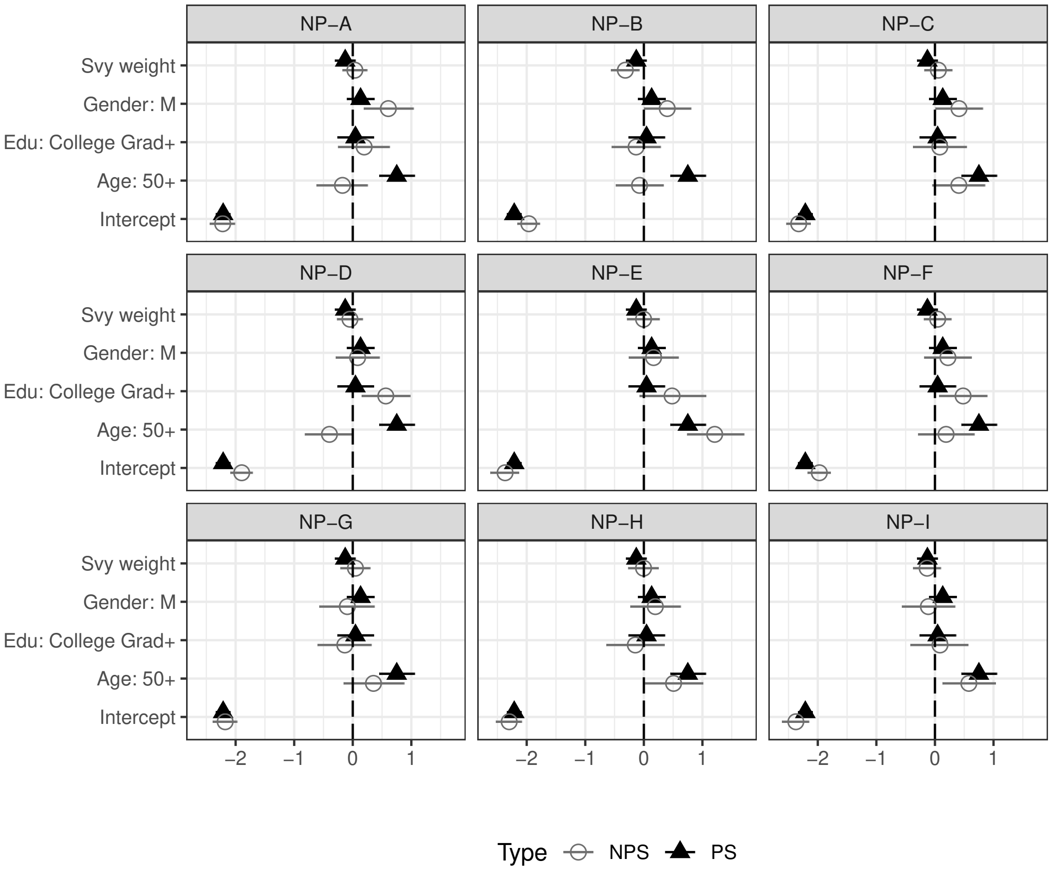


Figure 7: Maximum likelihood estimates of logistic regression coefficients and 95% confidence intervals for the neighborhood trust outcome variable in the ATP (triangles) and nine non-probability surveys (circles).

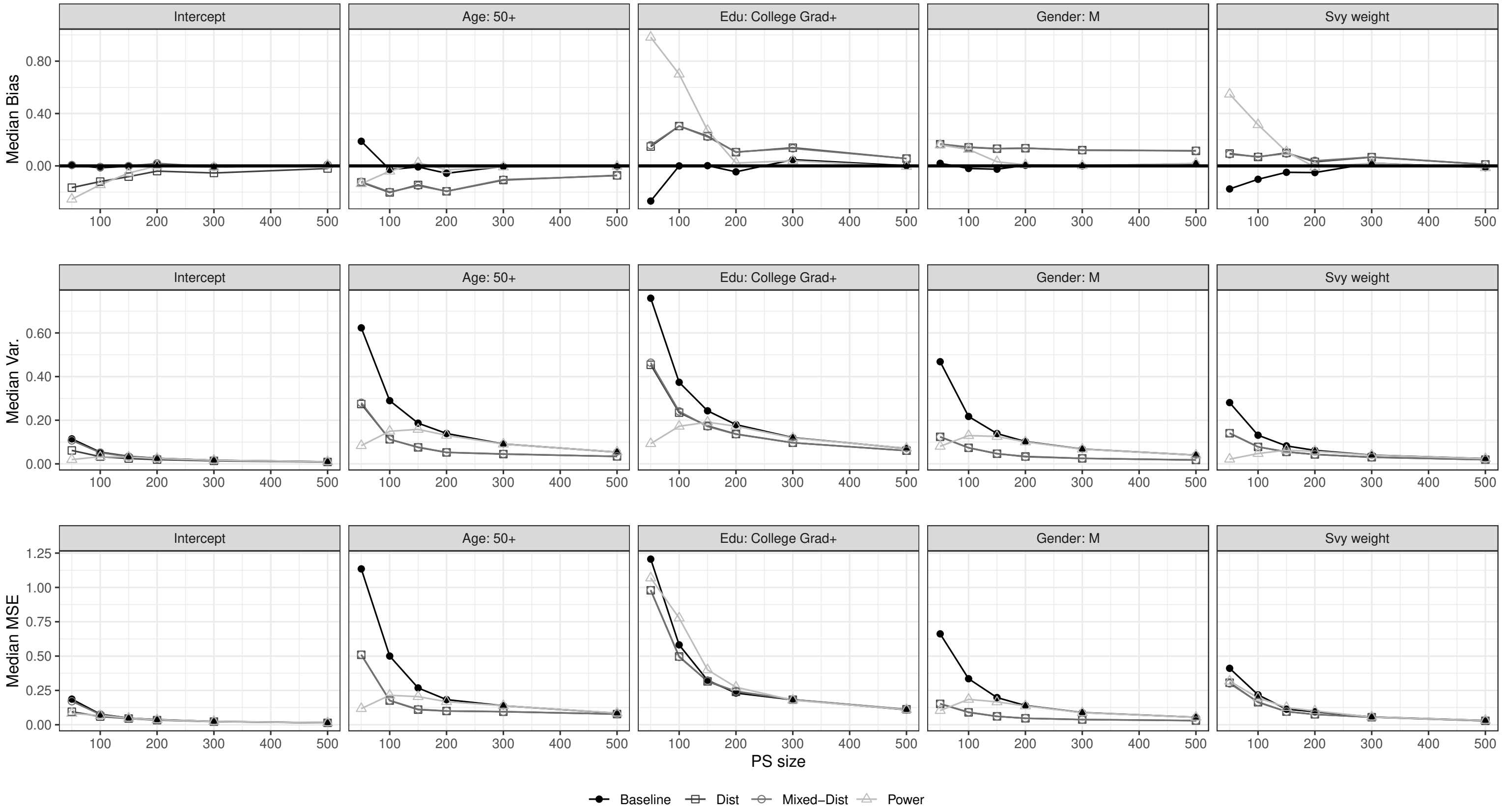


Figure 8: Median Bias, Variance, and MSE of coefficient estimates for the always vote outcome using NP-A survey as prior information.

Note: Four priors are considered: Distance (Dist), Mixed-Distance (Mixed-Dist), Power, and Baseline.

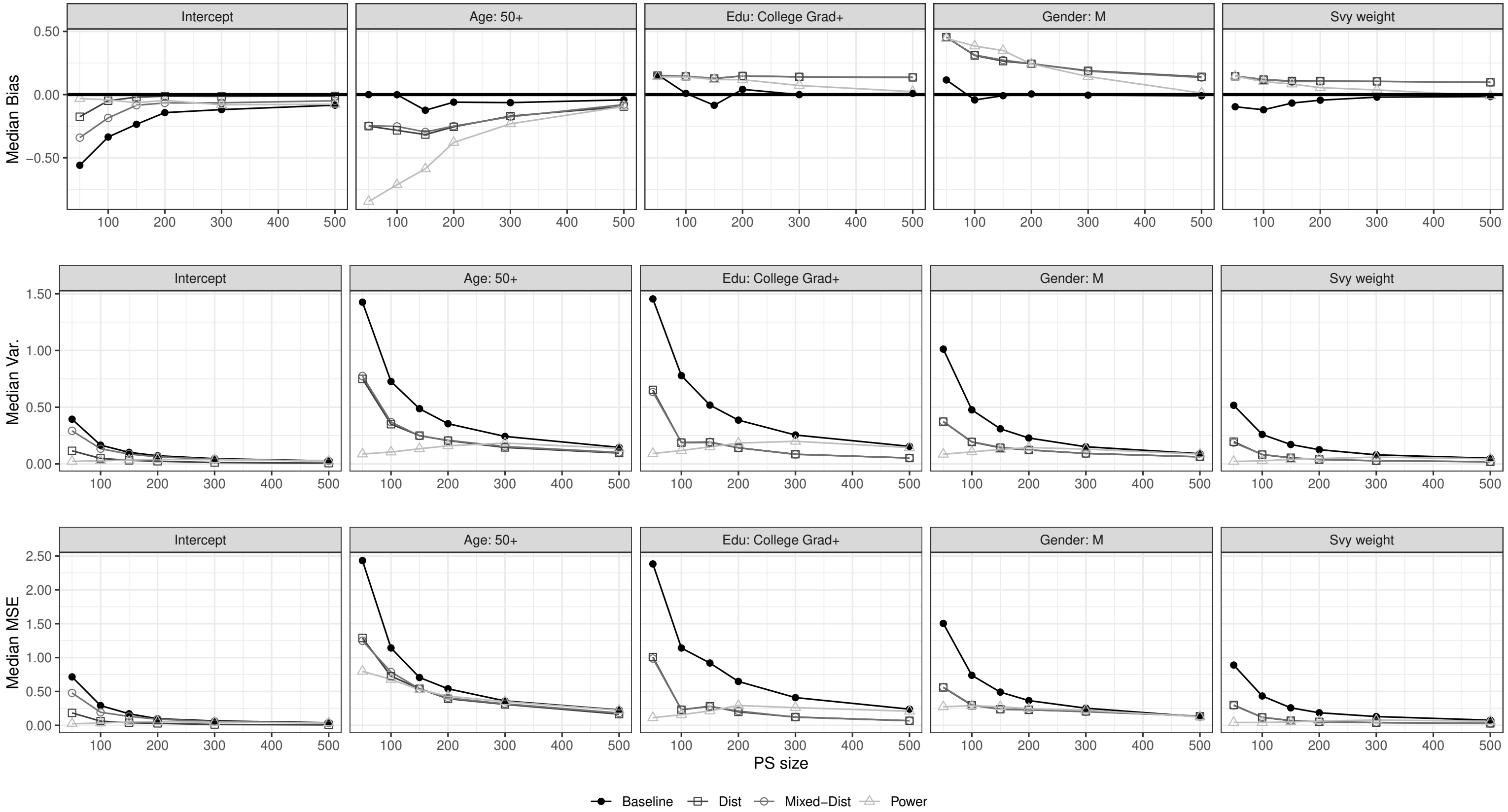


Figure 9: Median Bias, Variance, and MSE of coefficient estimates for the neighborhood trust outcome using NP-A survey as prior information.

Note: Four priors are considered: Distance (Dist), Mixed-Distance (Mixed-Dist), Power, and Baseline.

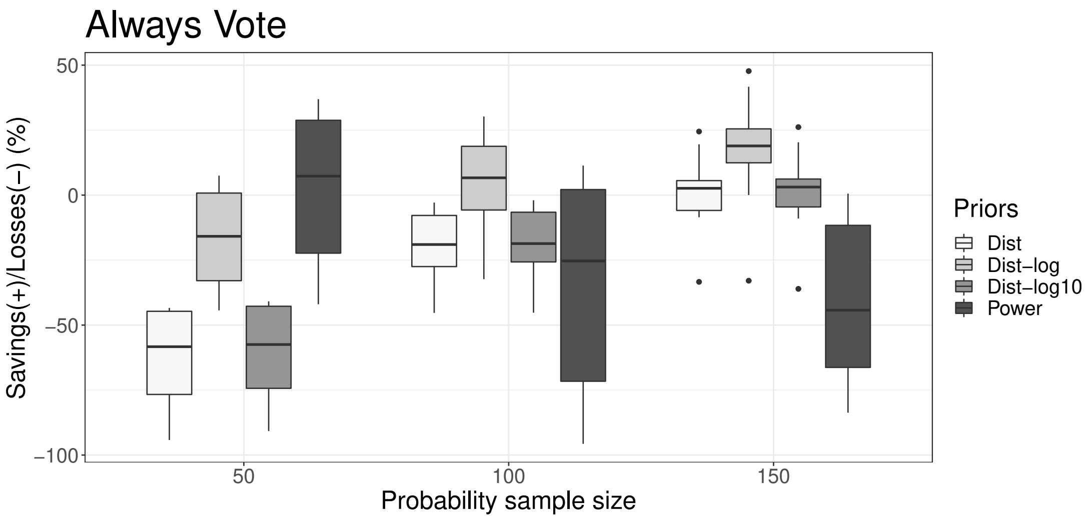


Figure 10: Percentage cost savings(+)/losses(-) for the always vote outcome. Four priors are considered: Distance (Dist), Distance-log (Dist-log), Distance-log10 (Dist-log10), and Power.

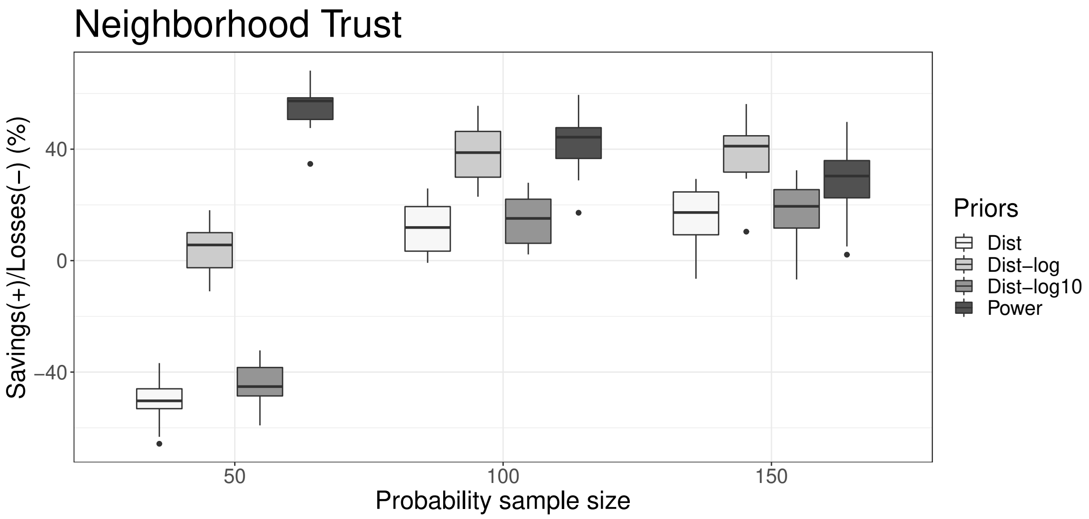
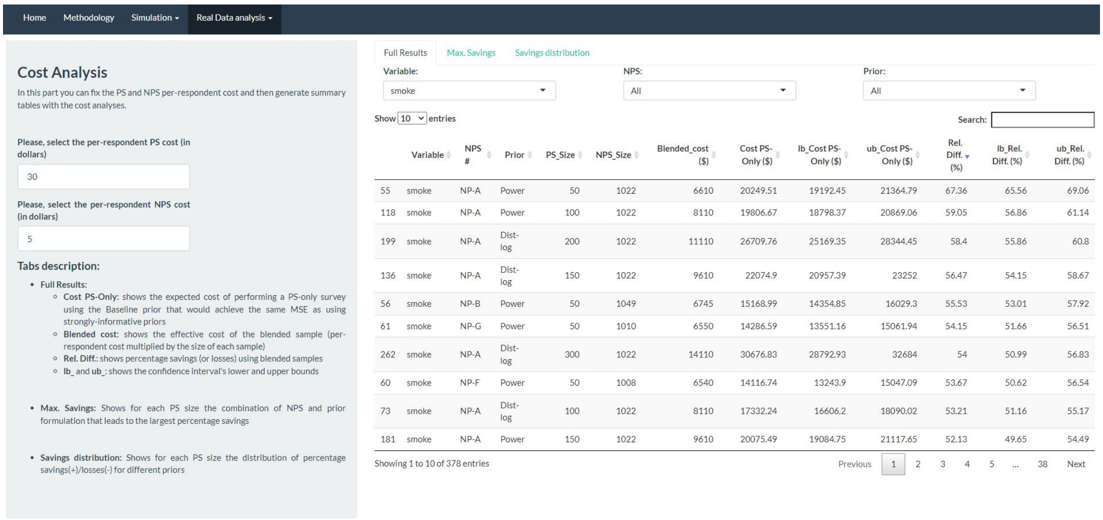


Figure 11: Percentage cost savings(+)/losses(-) for the Neighborhood Trust outcome. Four priors are considered: Distance (Dist), Distance-log (Dist-log), Distance-log10 (Dist-log10), and Power.

B The Shiny App

The Shiny web application was developed in R using the Shiny package (RStudio, Inc, 2022) and can be accessed at: `https://bayesdataintegration.shinyapps.io/shiny_bayes_data_integration/`. The app is divided into three sections. In the Methodology section the Bayesian framework is introduced along with the relevant code used to estimate the model parameters using different priors. In the



Cost Analysis

In this part you can fix the PS and NPS per-respondent cost and then generate summary tables with the cost analyses.

Please, select the per-respondent PS cost (in dollars)

30

Please, select the per-respondent NPS cost (in dollars)

5

Tabs description:

- **Full Results:**
  - **Cost PS-Only:** shows the expected cost of performing a PS-only survey using the Baseline prior that would achieve the same MSE as using strongly-informative priors
  - **Blended cost:** shows the effective cost of the blended sample (per-respondent cost multiplied by the size of each sample)
  - **Rel. Diff.:** shows percentage savings (or losses) using blended samples
  - **lb_** and **ub_:** shows the confidence interval's lower and upper bounds
- **Max. Savings:** Shows for each PS size the combination of NPS and prior formulation that leads to the largest percentage savings
- **Savings distribution:** Shows for each PS size the distribution of percentage savings(+)/losses(-) for different priors

Variable: smoke

NPS: All

Prior: All

Show 10 entries

Search:

	Variable	NPS #	Prior	PS_Size	NPS_Size	Blended_cost ($)	Cost PS-Only ($)	lb_Cost PS-Only ($)	ub_Cost PS-Only ($)	Rel. Diff. (%)	lb_Rel. Diff. (%)	ub_Rel. Diff. (%)
55	smoke	NP-A	Power	50	1022	6610	20249.51	19192.45	21364.79	67.36	65.56	69.06
118	smoke	NP-A	Power	100	1022	8110	19806.67	18798.37	20869.06	59.05	56.86	61.14
199	smoke	NP-A	Dist-log	200	1022	11110	26709.76	25169.35	28344.45	58.4	55.86	60.8
136	smoke	NP-A	Dist-log	150	1022	9610	22074.9	20957.39	23252	56.47	54.15	58.67
56	smoke	NP-B	Power	50	1049	6745	15168.99	14354.85	16029.3	55.53	53.01	57.92
61	smoke	NP-G	Power	50	1010	6550	14286.59	13551.16	15061.94	54.15	51.66	56.51
262	smoke	NP-A	Dist-log	300	1022	14110	30676.83	28792.93	32684	54	50.99	56.83
60	smoke	NP-F	Power	50	1008	6540	14116.74	13243.9	15047.09	53.67	50.62	56.54
73	smoke	NP-A	Dist-log	100	1022	8110	17332.24	16606.2	18090.02	53.21	51.16	55.17
181	smoke	NP-A	Power	150	1022	9610	20075.49	19084.75	21117.65	52.13	49.65	54.49

Showing 1 to 10 of 378 entries

Figure 12: The Shiny App: Cost Analysis

Simulation section the code for generating the population and the two samples (PS and NPS) is provided. Finally, all the steps are combined and the full simulation code is presented. Additional results including plots and tables for Section 5 are also available. In the Real Data Analysis section the data sources are presented in more detail and additional plots and tables for Section 6 are available. The Cost Analysis menu, under *Real Data Analysis/Cost Analysis*, contains the interactive application for replicating the cost analysis for user-entered PS and NPS costs (Fig. 12).

Bibliography

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the united states. *Population Research and Policy Review*, pages 1–28.

Alliance", O. P. (2022). Pricing.

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: Adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, 8(1):89–119.

Astley, C. M., Tuli, G., Mc Cord, K. A., Cohn, E. L., Rader, B., Varrelman, T. J., Chiu, S. L., Deng, X., Stewart, K., Farag, T. H., Barkume, K. M., LaRocca, S., Morris, K. A., Kreuter, F., and Brownstein, J. S. (2021). Global monitoring of the impact of the covid-19 pandemic through online surveys sampled from the facebook user base. *Proceedings of the National Academy of Sciences*, 118(51).

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., et al. (2010). Research synthesis: Aapor report on online panels. *Public Opinion Quarterly*, 74(4):711–781.

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(2):90–143.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46(1):1–29.

Beaumont, J.-F. and Rao, J. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? surv. *The Survey Statistician*, 83:11–22.

Berzofsky, M. E., McKay, T., Hsieh, Y. P., and Smith, A. (2018). Probability-based samples on twitter: Methodology and application. *Survey Practice*, 11(2):4936.

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

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848.

Biffignandi, S. and Bethlehem, J. (2021). *Handbook of Web Surveys*. John Wiley & Sons.

Callegaro, M., Villar, A., Yeager, D. S., and Krosnick, J. A. (2014). *A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples*, pages 23–53. Wiley.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84(1-2):121–137.

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999). Prior elicitation, variable selection and bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):223–242.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3):145–156.

[dataset] Pew Research Center (2014a). American trends panel wave 5. [Data set]. https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-5/.

[dataset] Pew Research Center (2014b). American trends panel wave 7. [Data set]. https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-7/.

[dataset] Pew Research Center (2015a). American trends panel wave 10. [Data set]. https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-10/.

[dataset] Pew Research Center (2015b). Online nonprobability landscape study. [Data set]. https://www.pewresearch.org/methods/dataset/online-nonprobability-landscape-study/.

De Santis, F. (2006). Power priors and their use in clinical trials. *The American Statistician*, 60(2):122–129.

Dever, J. and Shook-Sa, B. (2015). The utility of weighting methods for reducing errors in opt-in web studies. In *International Total Survey Error Conference, Baltimore, MD*.

Dimock, M., Doherty, C., Kiley, J., and Oates, R. (2014). Political polarization in the american public. Technical report, Pew Research Center.

DiSogra, C., Cobb, C., Chan, E., and Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM)*.

Dutwin, D. and Buskirk, T. D. (2017). Apples to Oranges or Gala versus Golden Delicious?: Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples. *Public Opinion Quarterly*, 81(S1):213–239.

Einarsson, H., Sakshaug, J. W., Cernat, A., Cornesse, C., and Blom, A. G. (2022). Measurement equivalence in probability and nonprobability online panels. *International Journal of Market Research*.

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Ganesh, N., Pineau, V., Chakraborty, A., and Dennis, J. M. (2017). Combining probability and non-probability samples using small area estimation. In *JSM Proceedings, Survey Research Methods Section*, pages 1657–1667.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Ghosh, J., Li, Y., and Mitra, R. (2018). On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383.

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1.

Hanson, T. E., Branscum, A. J., and Johnson, W. O. (2014). Informative g-priors for logistic regression. *Bayesian Analysis*, 9(3):597–612.

Hsiao, Y., Fiorio, L., Wakefield, J., Zagheni, E., et al. (2020). Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends. Technical report, Max Planck Institute for Demographic Research, Rostock, Germany.

Ibrahim, J. G., Chen, M.-H., et al. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Ibrahim, J. G., Chen, M.-H., Xia, H. A., and Liu, T. (2012). Bayesian meta-experimental design: evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics*, 68(2):578–586.

Keeter, S. (2019). Growing and improving pew research center's american trends panel. Technical report, Pew Research Center.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016). Evaluating online nonprobability surveys. Technical report, Pew Research Center.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):941–963.

Kim, J.-K. and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2):382–401.

Kish, L. (1965). *Survey Sampling.* John Wiley and Sons, Inc., New York.

Kreuter, F., Barkay, N., Bilinski, A., Bradford, A., Chiu, S., Eliat, R., Fan, J., Galili, T., Haimovich, D., Kim, B., et al. (2020). Partnering with facebook on a university-based rapid turn-around global survey. *Survey Research Methods: SRM*, 14(2):159–163.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2):293–312.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3):469–487.

Malhotra, N. and Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 anes to internet surveys with nonprobability samples. *Political Analysis*, 15(3):286–323.

Miller, P. V. (2017). Is There a Future for Surveys? *Public Opinion Quarterly*, 81(S1):205–212.

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):558–625.

Pasek, J. (2016). When will nonprobability surveys mirror probability surveys? considering types of inference and weighting strategies as criteria for correspondence. *International Journal of Public Opinion Research*, 28(2):269–291.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rafei, A., Flannagan, C. A., and 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.

Raghunathan, T., Ghosh, K., Rosen, A., Imbriano, P., Stewart, S., Bondarenko, I., Messer, K., Berglund, P., Shaffer, J., and Cutler, D. (2021). Combining information from multiple data sources to assess population health. *Journal of Survey Statistics and Methodology*, 9(3):598–625.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1):242–272.

Robbins, M. W., Ghosh-Dastidar, B., and Ramchand, R. (2020). Blending Probability and Nonprobability Samples with Applications to a Survey of Military Caregivers. *Journal of Survey Statistics and Methodology*, 9(5):1114–1145.

RStudio, Inc (2022). shiny: Easy web applications in r. URL: http://shiny.rstudio.com.

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics*, 35(3):653–681.

Stan Development Team (2019). Stan modeling language users guide and reference manual, version 2.29.

Stan Development Team (2021). RStan: the R interface to Stan. R package version 2.21.3.

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5):503–516.

Thompson, A. J. and Pickett, J. T. (2020). Are relational inferences from crowdsourced and opt-in samples generalizable? comparing criminal justice attitudes in the gss and five online samples. *Journal of Quantitative Criminology*, 36(4):907–932.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.

Valliant, R., Dever, J. A., and Kreuter, F. (2018a). Basic steps in weighting. In *Practical Tools for Designing and Weighting Survey Samples*, pages 321–367. Springer.

Valliant, R., Dever, J. A., and Kreuter, F. (2018b). Nonprobability sampling. In *Practical Tools for Designing and Weighting Survey Samples*, pages 565–603. Springer.

West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., and Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15(3):1556–1581.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1):120–147.

Yang, S., Kim, J. K., and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):445–465.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., and 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.

Yee, T. W. (2021). On the hauck–donner effect in wald tests: Detection, tipping points, and parameter space characterization. *Journal of the American Statistical Association*, 0(0):1–12.

Chapter 4

Augmenting Business Statistics Information by Combining Traditional and Textual Data

Paper under review:
Salvatore C., Biffignandi S., Bianchi A. Augmenting Business Statistics Information by Combining Traditional Data with Textual Data: A Composite Indicator Approach. Revised After Major Revision.

1 Introduction

The availability of new data has led to an expansion of data collection methods, moving beyond traditional primary data collection to the extraction of statistics from non-traditional sources. These sources, referred to as big, or digital trace/behavioral data, include, among others, social media posts, Google trends and mobile phone data (i.e., location, photos, and other sensor data), and are produced by human online/digital behaviors and interactions (Howison et al., 2011).

Digital trace data are not generated for statistical purposes but can serve as a convenient and timely source of information for understanding and measuring (new) complex socio-economic phenomena (Japec et al., 2015). These new data sources provide a basis for the multi-purpose extraction of different statistical indicators, which complement the traditionally available statistical information and feed smart statistics (Trappmann et al., 2022; Stier et al., 2020; Struminskaya et al., 2020)

The cost of collecting and processing high quality traditional data, such as surveys, is increasing, and the process of deriving statistical products from this data is demanding and time-consuming (Luiten et al., 2020). To address these issues, the integration of traditional and digital trace data for producing innovative statistics and indicators is a promising approach. This can enhance the timeliness, providing a finer spatial and temporal resolution, a higher level of detail, new perspectives, and new insights on phenomena, while also reducing the production cost of (official) statistics (Ricciato et al., 2020).

Research on indicators constructed from non-traditional sources, particularly textual data from social media, is prevalent in the literature. Social media are commonly used to better understanding attitudes and behaviors with reference to social phenomena (Ceron et al., 2016; Luhmann, 2017; Iacus et al., 2020; Rill et al., 2014). Further, a number of experimental statistics have been developed by National Statistical Institutes (NSIs) using such textual data to study social tensions[1] and consumers' confidence in the economy (see, for example, Daas and Puts (2014) and the Istat's Social Mood on Economy Index[2]). However, studies combining traditional and digital trace data-based indicators are scarce.

As it appears evident from the literature, the study of social aspects using unstructured data is prevalent. Nevertheless, the production of business statistics can particularly benefit from the use of new data sources. These can be used in a variety of ways, including enhancing the information for a given unit (Bender and Sakshaug, 2021). For example, Statistics Canada used sensor data to augment administrative data and produce more efficient small area estimates for business statistics (Thomassin, 2018). Similarly, Statistics Netherlands (CBS) is committed in enhancing business statistics, using web-scraped data from companies' website in order to detect innovative companies and improve the quality of the appointed NACE codes (Daas and van der Doef, 2021; Roelands et al., 2018). The Italian National Statistical Institute (ISTAT) is also committed in developing experimental statistics based on businesses' websites in order to identify their activities or to augment the information collected through the traditional survey on Information and Communication Technologies (Barcaroli et al., 2015, 2016; De Fausti et al., 2019).

In this paper, we focus on business statistics and propose a general methodological framework for the construction of composite indicators that are generated by combining traditional and innovative (e.g. social media or web-based) indica-

[1] https://www.cbs.nl/en-gb/about-us/innovation/project/social-tensions-indicator-gauging-society

[2] https://www.istat.it/en/experimental-statistics/experiments-on-big-data

tors. The framework is developed following a modular approach for entailing the use of digital and unstructured data (and relative metadata) in measuring new phenomena in business statistics. Another original aspect is that we propose to process metadata[3] in order to build innovative indicators. Processing metadata is an emerging aspect in the analysis of digital trace data and existing experiences rely mainly on checking and improving the quality of the metadata, whereas the computation of indicators based on metadata is a novel contribution.

To the purpose of providing an example to practitioners, we develop an illustrative exercise to demonstrate how to implement the proposed method. It serves as a prototype application which shows the steps to be undertaken to build up new, innovative, indicators based both on unstructured and structured data. In our exercise, we consider a commercial database as traditional source for structured data and Twitter as new data source for unstructured data. We focus on the case where data about the same units are available in both sources. However, a similar approach can be adopted at a more aggregate level, namely in the case such individual information is not available.

The reminder of the article is the following. Section 2 discusses the challenges of constructing smart business statistics. Section 3 presents a modular architecture for the construction of such statistics and the framework to build composite indicators. Section 4 illustrates the practical exercise on the construction of a prototype indicator. Section 5 discusses the results and conclusions are drawn in Section 6.

2 Challenges of augmenting business statistics with unstructured data

Traditionally, business statistics are derived from survey data, like the European Company Surveys[4], the Business and consumer surveys (BCS)[5] and other surveys carried out by NSIs. In these cases, the data are structured, the data-generating process is under the researchers' control, and errors are allocated along the whole survey process according to the Total Survey Error (TSE) framework (Biemer,

[3] In order to avoid confusion, we clarify the use of the term metadata in the context of digital unstructured data. It differs from the definition used in statistics, i.e., the information that is required in order to interpret and use statistics. In this context, metadata refers to additional information about the main data of interest. In Twitter, for example, the tweet represents the main data and the date of publication, likes, links, and images are metadata.

[4] https://www.eurofound.europa.eu/surveys/european-company-surveys

[5] https://ec.europa.eu/eurostat/web/euro-indicators/business-and-consumer-surveys

2010). Consequently, surveys are considered as a high-quality data source for business statistics.

Alongside surveys, other popular sources for business statistics are administrative or commercial business data (Costanzo, 2011). These are still structured data, quality is checked and improved when necessary. These data are not primarily collected for statistical or research purposes. For that reason, they are usually referred to as secondary data. Business registers, documents from local authorities (e.g., tax authority), and law-mandatory reporting are all example of administrative data. Commercial business data are provided by private companies, for example, Bureau van Dijk[6], Bloomberg[7], and Refinitiv[8].

More recently, the digital transformation has resulted in the emergence of new sources for business and economic statistics (Bernal and Sejersen, 2021). For example, social media posts, annual reports, businesses websites and newspaper articles can be used to study new aspects or gain additional information about companies. In this respect, the production of statistics using traditional data enhanced with new data available from digital sources are referred to as smart statistics. One of the advantages of smart statistics is the ability to augment the information, thereby providing richer insights into the topic of interest. However, there are also several challenges to be considered. In the following discussion, we focus our attention on unstructured textual data.

To begin with, it is necessary to extract the data of interest using, for example, web-scraping or Application Programming Interfaces (APIs). Online data are not static. Hence, during data extraction, researchers must be aware of issues pertaining to the changes in data over time, coverage, reliability, and validity of the data, among others. Social media posts, for instance, can be modified or deleted over time, and related metadata can also change (e.g., likes, replies, and shares). Therefore, the results may differ based on the timing of retrieval. Similarly, different formulation of the search query in terms of the keyword specified, such as when extracting social media posts or newspaper articles based on firm names or products, can result in the delivery of different data.

Another issue that might arise when one wants to obtain unit level observations, for example, studying the external communication of businesses on social media, is the problem of identifying the right accounts. For instance, not all businesses are present on social media, or they may have multiple accounts related to specific types of communication (e.g., general communication, promotion

[6] https://www.bvdinfo.com/en-gb/

[7] https://www.bloomberg.com/

[8] https://www.refinitiv.com/en/financial-data

and advertisement, business news, clients assistance, recruiting and topic-specific accounts for communicating their socially-responsible behavior). This leads to selection and coverage issues that might affect the quality of the data.

Secondly, unstructured textual data must be transformed into structured data. This can be accomplished in different ways according to the purpose of the analysis. For example, sentiment analysis, topic modeling, and other classification or clustering algorithms can be applied. Moreover, the results might be influenced by the various data cleaning and pre-processing choices (Denny and Spirling, 2018; Symeonidis et al., 2018).

Like survey data, also the analysis of unstructured textual data is susceptible to errors. In this direction, there are efforts being made to adapt the TSE framework to such data, but currently, there is not a general framework in order to account, measure and evaluate errors and data quality (Salvatore et al., 2021; Amaya et al., 2020; Sen et al., 2021). Data sources have different characteristics, which require different quality frameworks. The importance of these aspects becomes especially evident when integrating data from different sources, where it is crucial to understand how errors arise, accumulate, and interact during the entire integration process (De Waal et al., 2019). These are all emerging topics in the literature.

While all these factors should be considered when combining data, our focus here is on proposing a procedure to develop composite indicators based on the integration of different types of data, structured and unstructured, derived from traditional and non-traditional sources.

3 Methodology

3.1 A modular framework for the construction of smart business statistics

To produce smart business statistics using unstructured textual data, we develop a modular methodological approach in three layers. This is an adaption of the modular organization into three layers introduced by Ricciato et al. (2020).

In the first layer, the data are collected and transformed into structured data. Such data and their relative metadata need then to be interpreted by statisticians and serve as input for the second layer. The processing of metadata to complement the analysis of unstructured digital data has been examined in a limited number of studies. Indeed, it is an emerging topic and applications relate user/account profiling (Perez et al., 2018; Daas et al., 2016), and geo-spatial applications (Da Mota and Pickering, 2021; Rosales Sánchez et al., 2017). As original contribution, we propose to use social media metadata for the construction of composite indicators as shown in the prototype application (Section 4).

In the second block, innovative statistical information is extracted, and indicators are computed. The first and the second layer are augmenting statistical information through the creation of new indicators generated using textual unstructured data.

In the third layer innovative statistics and indicators are used to augment the already available traditional data. Depending on the specific use-case, this can be achieved through methods such as linkage, statistical integration, or by combining indicators. As a result, Smart Business Statistics are produced. Figure 1 summarizes the framework described above.

Based on the modular architecture in three layers, we show how to generate smart business composite indicators combining structured and unstructured data (e.g. textual data from social media and websites, or other innovative data sources).

The modular approach is useful when dealing with new and complex data sources and their integration with traditional ones. Modularity also allows other researchers and practitioners to explore other methodological variants (instances) within the same methodological architecture, and possibly propose improvements to specific modules or test sensitivity of the obtained results. Moreover, if a researcher wants to apply the proposed procedure for the construction of a composite indicator in its own research context, it is possible to proceed across the whole set of three layers or to compute the composite indicator only going through the second and third layer if the elementary indicators have already been computed.

3.2 The Composite Indicator Approach

3.2.1 Background concepts

Before describing the proposed methodology, we shortly remind that when constructing composite indicators, it is necessary to consider and take decisions on different aspects (Mazziotta and Pareto, 2013).

First of all, the theoretical framework of the substantial research topic has to be defined. This is crucial for the choice of the data and the variables' definition. It is also important to guide the researcher in the construction process of the composite indicator with respect to methodological decisions related to

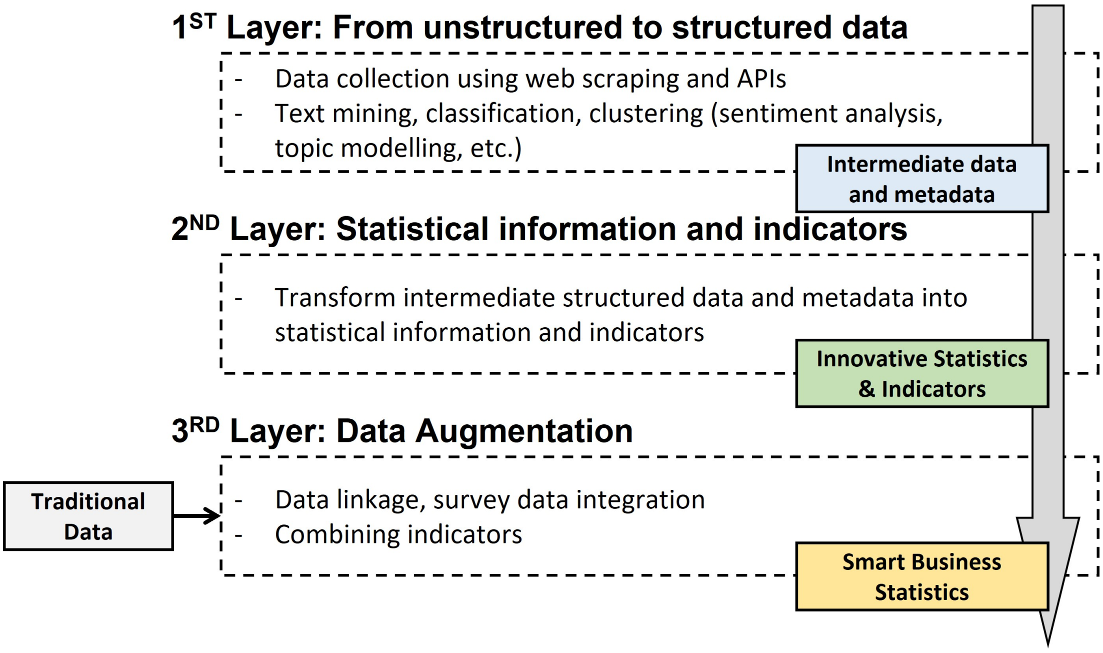

1ST Layer: From unstructured to structured data
- Data collection using web scraping and APIs
- Text mining, classification, clustering (sentiment analysis, topic modelling, etc.)
Intermediate data and metadata
2ND Layer: Statistical information and indicators
- Transform intermediate structured data and metadata into statistical information and indicators
Innovative Statistics & Indicators
3RD Layer: Data Augmentation
Traditional Data
- Data linkage, survey data integration
- Combining indicators
Smart Business Statistics

Figure 1: Modular methodological framework for producing smart business statistics

the normalization of the indicators and the aggregation strategy. Normalization is performed in order to ensure comparability. Based on the variable type (e.g., continuous, categorical, or ordinal) and the aggregation strategy, this can be accomplished in a variety of ways. Common methods are the standardization (z-score), min-max transformation (or re-scaling) or the transformation to index numbers (Mazziotta and Pareto, 2020).

Aggregation refers to the combination of the individual indicators in order to create a composite indicator. This phase entails considerations on the polarity and the importance of each elementary indicator and the identification of the technique to synthesize the elementary indicators. To properly insert the original indicators into the aggregation procedure polarity of indicators should be carefully considered. The polarity of an indicator refers to the direction of the relationship between the indicator and the phenomenon to be measured. The polarity is positive (negative) if the dimension is positively (negatively) associated to the phenomenon.

The selection of the aggregation technique depends on the level of compensability of the individual indicators, which refers to the possibility of balancing a disadvantage on some indicators with a sufficiently large advantage on others. This should be based on theoretical evaluations. In this respect, there are three types of aggregation approaches depending on the degree of compensability: compensatory, partial compensatory, and non-compensatory. For example, full-compensatory aggregation is obtained with the arithmetic mean. In the case of individual indicators from unstructured data, this can be the case of the topic proportion resulting from a topic model.

Partial-compensatory approaches relate, for example, to the computation of geometric, harmonic, quadratic means, or specific methods like the Mazziotta-Pareto procedure (De Muro et al., 2011). For example, one could consider the social media dimension related to communication aspects of a certain phenomenon to be partially replaceable with traditional measurements of the same phenomenon. Non-compensatory aggregation is usually performed following multi-criteria approaches.

Aggregation also involves the identification of weights associated to the individual indicators. Weights reflect the relative importance of the indicators to be combined. When no weights are specified, all indicators are implicitly weighed equally. Alternatively, weights can be determined according to subjective and expert evaluations, or statistical methods, such as Principal Component Analysis. However, weights should only be specified when there is a strong theoretical basis for doing so, otherwise a no-weighting strategy should be adopted (Booysen, 2002; Mazziotta and Pareto, 2022). Attention should be paid to implicit importance associated to the original elementary indicators in the case of subsequent aggregations. For a complete overview of composite indicators construction, please refer to Mazziotta and Pareto (2020), OECD (2008) and Booysen (2002).

When developing composite indicators, it is important to evaluate the quality of the results taking into consideration the impact of the different methodological decisions that have been made. This includes normalization methods, weighting approaches, and the evaluation of uncertainty in the weights of sub-indicators. In the literature, various possible procedures for evaluating quality are suggested, mainly uncertainty analysis (UA) and sensitivity analysis (SA). UA focuses on how uncertainty in the input factors propagates through the structure of the composite indicators and influence its value. SA studies how much each individual source of uncertainty contributes to the output variance. For a general discussion of the procedures, please refer to Saisana et al. (2005).

In addition to these traditional quality aspects, when working with unstructured data or non-traditional data sources, new quality considerations arise. For example, results may be affected by data extraction techniques (e.g. selection of social media accounts of webpages), pre-processing (e.g. data cleaning) and analytical choices (e.g. machine learning methods to extract the information).

While these topics are important and currently being discussed in the literature, they are beyond the scope of this paper, which focuses on presenting a general framework for data augmentation.

3.2.2 Procedures of the approach

As regards our original contribution, we present a methodology for constructing a) simple and composite indexes that measure new aspect of phenomena using new data sources and b) a composite indicator that integrates traditional and non-traditional indexes. To do that, we follow our adaption of the modular approach originally proposed by Ricciato et al. (2020, see Figure 1). At first, we focus on the construction of the innovative index (INN-INDEX). We assume that the traditional index is already available (TRAD-IDNEX), and we compute the augmented index (SMART-INDEX).

It is important to note that the theoretical framework of the phenomenon being measured plays a crucial role in the construction of the index. All decisions that should be taken at the various step of the three layers and of the composite indicator construction must align with this framework.

In the first layer, elementary indicators are identified. The second layer includes the aggregation of sub-indicators which are, then, combined in the smart indicator in the third layer. Our proposed modular layer approach is illustrated in Figure 2. Starting from a set of individual traditional and innovative indicators, at the first level, individual indicators are aggregated to describe the traditional and innovative dimensions of interest, respectively.

By way of example, assume that, according to the theoretical framework, there are two relevant dimensions that can be measured by the innovative data source, namely D_1 and D_2 and let $I_{D_1,1}, \dots, I_{D_1,i}, \dots, I_{D_1,n}$ be the n individual indicators related to dimension D_1 and $I_{D_2,1}, \dots, I_{D_2,j}, \dots, I_{D_2,m}$ be m individual indicators related to dimension D_2. Such indicators and dimensions must be identified based on theoretical, empirical, pragmatic, or intuitive considerations (Booysen, 2002). The elementary indicators are combined in order to generate two composite indicators measuring each dimensions of interest, CI_{D_1} and CI_{D_2} respectively. The approach may be extended to more dimensions depending on the characteristics of the phenomenon and the innovative source being studied.

These indicators are then further aggregated to create the INN-INDEX. This is the second level of aggregation. The same methodology can be applied to obtain a traditional indicator if one does not already exist. Moving to the third layer, the third level of aggregation relates the construction of the innovative smart composite indicator. In the second and third levels, attention should be paid to avoid double normalization. As stated before, all choices made along the three levels of aggregation are determined by the phenomena under examination.

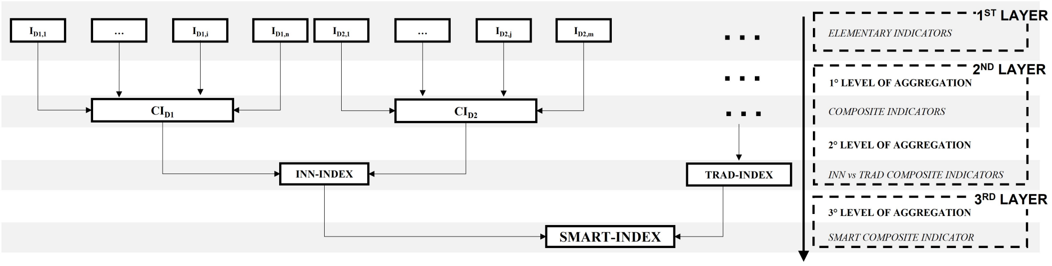


Figure 2: Composite Indicator construction strategy on three layers.

We illustrate how to apply the proposed methodology through a practical exercise that shows how to construct a prototype composite indicator for measuring Corporate Social Responsibility in the next section.

4 Construction of a prototype

4.1 Context and theoretical framework

This application focuses on the construction of a composite indicator in the field of business statistics and sustainability. Socially responsible behaviors of businesses are linked to the concept of Corporate Social Responsibility (CSR)[9]. Given its multi-faceted nature, measuring CSR activities is naturally related to the use of composite indicators, which allow us to summarize complex or multi-dimensional phenomena (Dahlsrud, 2008).

The aim of this practical exercise is to measure CSR commitment based on a comprehensive view, including both effective commitment (as traditionally considered) and online communication of CSR-related activities. Despite its importance, the study of online business communication with respect to sustainability is still a relatively under-examined and emerging topic (Araujo and Kollat, 2018; Chae and Park, 2018).

[9] CSR refers to the implementation of activities aiming at the improvement of firms' reputation and at positively impacting the society (Carroll et al., 1991). A related aspect, that is becoming more and more important nowadays, is the online communication of CSR activities, which can be investigated thanks to the availability of social media data. Indeed, listening to the online communication is useful to researchers and policy makers in order to monitor the behavior of the business with reference to the implementation of sustainable development and with respect to the Agenda 2030.

Our contribution is to demonstrate how the modular framework can be applied in practice. We show the various steps that should be undertaken for the technical construction of a smart indicator to measure CSR. By providing a step-by-step guide for the technical construction of the indicator, we aim to show how to effectively use social media data from Twitter in conjunction with (already available) traditional data to create a comprehensive indicator that accounts for various aspects of CSR (augmenting information). Thus, in this context, the INN-INDEX is based on social media and renamed SM-INDEX.

It is beyond the scope of this paper to provide a comprehensive examination of the CSR theoretical framework or to fully evaluate the meaning of the computed indicators. Further details regarding the construction and quality evaluation of these indicators will be explored in an ongoing study.

4.2 The application of the modular framework

For the sake of illustration, we consider the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies as of the composition in August 2020. We retrieved the full list of firms, jointly with the corresponding activity sector from Bloomberg. With respect to sectors classification, Bloomberg adopts the Global Industry Classification Standard (GICS) developed by MSCI and S&P Dow Jones.

For the traditional indicator, we consider the Environmental, Social and Corporate Governance (ESG) database provided by Refinitiv, one of the world's largest providers of financial markets data and infrastructure (commercial data). Data for listed companies refer to their sustainability performance considering various aspects, including emission reductions, social programs, and economic performance. The database collects publicly reported data, checked for quality, and provides a CSR-Strategy Score. This reflects a company's practices to integrate economic (financial), social and environmental dimensions into its day-to-day decision-making process and it ranges between 0 and 100. The CSR-Strategy Score is the traditional indicator we consider (TRAD-INDEX). It should be noted that it is not available for all firms. For the purpose of our example, we only consider the firms for which information are present in both the traditional and digital data source.

For the construction of the social-media based index (SM-INDEX), to be integrated with the traditional one, we follow the modular methodologies proposed in Section 3.1. First, elementary indicators are identified and are used as input

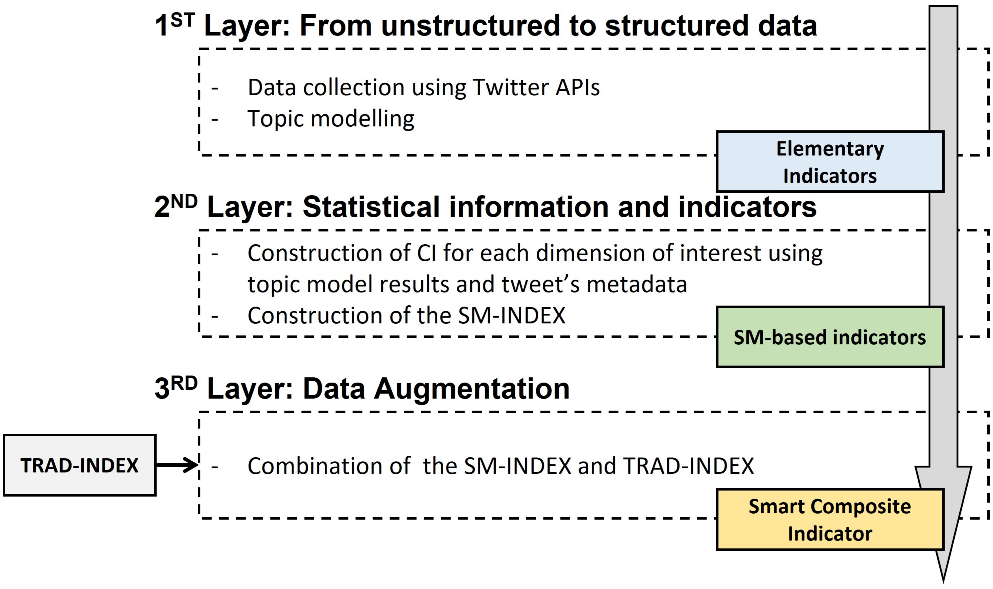

1ST Layer: From unstructured to structured data
- Data collection using Twitter APIs
- Topic modelling
Elementary Indicators
2ND Layer: Statistical information and indicators
- Construction of CI for each dimension of interest using topic model results and tweet's metadata
- Construction of the SM-INDEX
SM-based indicators
3RD Layer: Data Augmentation
TRAD-INDEX
- Combination of the SM-INDEX and TRAD-INDEX
Smart Composite Indicator

Figure 3: Modular methodological framework applied to the specific empirical exercise

for the construction of the innovative social media-based index (second layer). Then, as part of the third layer we combine the two indicators to produce a smart business composite indicator. Figure 3 summarizes the process described above. The following sections discuss in greater detail the proposed layers.

4.2.1 The first layer: elementary indicators

Following the tasks in the first layer, we identified and retrieved the data form the official Twitter accounts of the companies. Given that companies may have several Twitter accounts, we focused primarily on CSR accounts and, in case these are not available, on the news or multipurpose ones. The objective is to reduce the noise (no-CSR tweets) in the data. Two companies, namely Apple and Walgreens Boots Alliance, turned out not to have a Twitter account, thus, leading to the inclusion of 28 firms and 42 different accounts (18 CSR, 5 news-type and 19 multipurpose) in the analysis. We use the same data retrieved by Salvatore et al. (2022). They refer to the 2019 year and the total number of messages retrieved is 25,148. We then apply Structural Topic Model (STM) which allows to discover and link the topics to the CSR dimensions, namely economic, social, environmental and general (or mixed). A short description of the STM and topic model output can be found in Appendix A and B. Results can be found in greater detail in Salvatore et al.(2022) .

Generally, social media communication differs in content (the topic discussed) and modality (the way it is conducted). Thus, we consider two dimensions to build the SM-based indicator. The first one refers to the communication content in tweets, i.e., to the text which refers to the communication of CSR activities in one of its dimensions, economic, social, environmental and general (or mixed). The second one refers to communication modality (media richness from tweets metadata). This is an important aspect for the communication to be effective and to engage with customers and stakeholders. We expect that the higher the media richness, the more effective the communication will be (Araujo and Kollat, 2018).

Topic model results (proportion of text about CSR dimensions) represent the elementary indicators with respect to the content dimensions. For the modality dimension we consider tweets' metadata. In this respect, each tweet can contain hashtags (defining the topic of posts and allowing users to associate the tweet with all other tweets using the same identifying hashtags), mentions (engaging with other users), media (e.g., photos), and links (to external web pages). These elementary indicators represent the output of the first layer, which is the base for the construction of intermediate composite indicators in the second layer.

4.2.2 The second layer: Development of the social media-based indicator

The composite indicator for the content dimension is constructed by considering the output of the topic model as its elementary indicators. Specifically, the proportion of text devoted to each CSR dimension for each tweet is used. We assume that these proportions are substitutes (compensatory aggregation) with the same importance (no weight). To obtain the composite indicator, we take the sum of these proportions at the tweet level and then aggregate them at the firm level by taking the arithmetic mean (first innovative indicator).

The composite indicator for the modality dimension is based on tweets' metadata. Similar to the content dimension, we consider elementary indicators to be substitutes (compensatory aggregation) with the same importance (no weight). The elementary indicators used are the presence of hashtags, mentions, media, and links (binary variables). For each tweet, we sum these individual indicators, obtaining a score between 0 and 4. We then aggregate these scores at the firm level by computing the arithmetic mean (second innovative indicator).

Once the modality and the content indexes are constructed, it is necessary to combine them to obtain the SM-INDEX. In this case we propose to apply

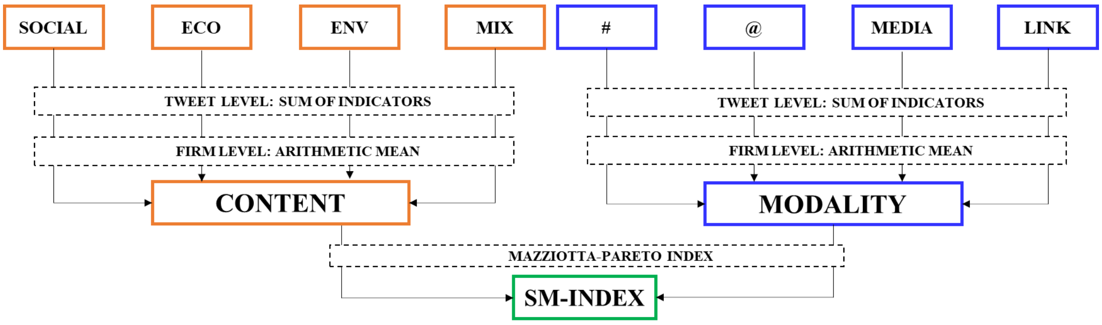

SOCIAL
ECO
ENV
MIX
#
@
MEDIA
LINK
TWEET LEVEL: SUM OF INDICATORS
TWEET LEVEL: SUM OF INDICATORS
FIRM LEVEL: ARITHMETIC MEAN
FIRM LEVEL: ARITHMETIC MEAN
CONTENT
MODALITY
MAZZIOTTA-PARETO INDEX
SM-INDEX

Figure 4: Composite Indicator aggregation strategy.

the Mazziotta-Pareto index (MPI) which is partially compensatory recognizing that the two dimensions are equally important but partially substitute to gain efficiency in CSR communication. Indeed, a deficiency in the content can be partially compensated by effective communication (and vice versa). It is based on a non-linear function that, starting from the arithmetic mean of the normalized indicators, introduces a penalty for units with unbalanced indicators (De Muro et al., 2011). To compute the MPI, given the data matrix $X = \{x_{ij}\}$, we proceed with standardization as follows

$$z_{ij} = 100 + \frac{((x_{ij} - M_{xj}))}{S_{xj}} \cdot 10 \tag{1}$$

where i refers to the unit and j to the indicator (content and modality respectively), and M and S refer to the mean and standard deviation of the content and modality indexes. Next, given the positive polarity of the indicators, we compute the MPI

$$MPI_i = M_{z_i} - S_{z_i} \cdot cv_{z_i} \tag{2}$$

where z refers to the standardized data as in (1) and M_{z_i}, S_{z_i}, cv_{z_i} denote the mean, standard deviation, and coefficient of variation of the normalized values for unit i, respectively. Figure 4 summarizes the aggregation approach described above.

4.2.3 The third layer: Development of an augmented information composite indicator

Considering the SM-INDEX and the TRAD-INDEX, it is possible to build a combined innovative smart indicator (SMART-INDEX). The TRAD-INDEX is standardized before the combination, while the SM-INDEX is not, being the aggregation output of previously standardized indicators. For the aggregation of SM-INDEX and TRAD-INDEX, we propose to apply the MPI, considering the

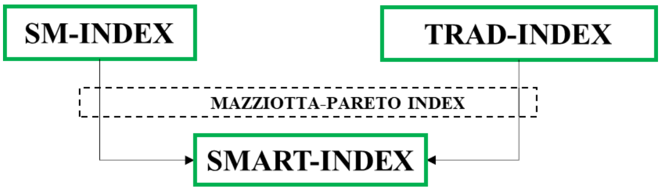


Figure 5: Composite Indicator aggregation strategy.

positive polarity of the indicators (Figure 5). Indeed, we assume that the two dimensions are partially compensatory, i.e., efficient communication might compensate low effective commitment and high effective commitment might compensate scarce communication.

The SMART-INDEX measures the commitment in a more comprehensive way, considering not only the effective commitment (traditional indicator) but also the effort in online CSR communication (social media indicator). Figure 5 summarizes the methodology to combine the two indicators.

5 Results and Discussion

Figure 5 shows the values of the social media-based, traditional and combined indicators for each company. The table with detailed result is available in Appendix C. For the TRAD-INDEX the standardized values according to (1) used as input for the Mazziotta-Pareto index are reported.

The TRAD-INDEX is very similar across all companies, except for Boeing, Chevron, Honeywell, and McDonalds for which it is particularly low and below 100 indicating a low level of effective commitment. A rational behind this similarity is that the index is constructed considering mainly compliance to laws and regulation with respect to CSR reporting that, nowadays, is a common practice for most companies. The SM-INDEX allows to discriminate better the communication about CSR commitment among firms.

The combination of the two indicators provides an innovative measure of CSR commitment and communication effectiveness, giving additional insights to researchers. Table 2 in Appendix C provides the ranking of firms based on the SM-INDEX, TRAD-INDEX, and the SMART-INDEX, respectively. Generally, firms that rank highly on the SM-INDEX place low on the TRAD-INDEX (and vice versa). Companies in the services sector (e.g., Tech and HC) have a higher position on the SM-INDEX and a lower position on the TRAD-INDEX. A possible explanation could be that firms in the services sector have a high need for communication via their websites, whereas firms in other sectors do not. This

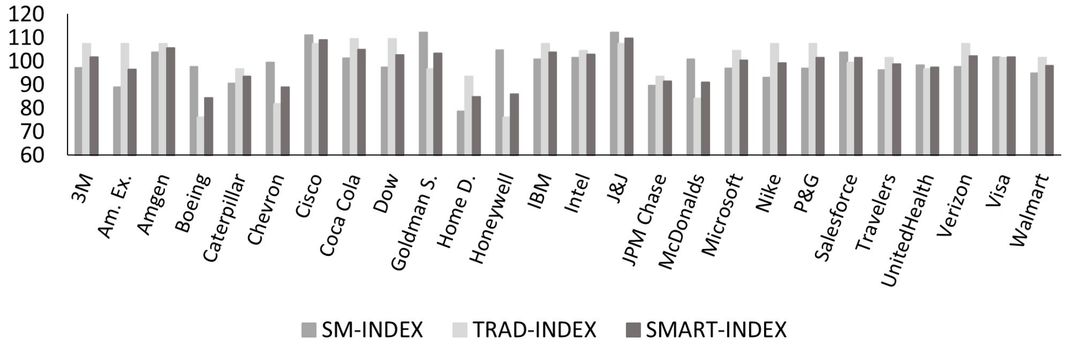


Figure 6: Social media-based (SM-INDEX), traditional (TRAD-INDEX) and smart indexes.

may be because other methods of communicating sustainability are possible when offering a consumer product (such as information on the package).

Due to their equal weighting, the SMART-INDEX provides a middle ground between the two. Nevertheless, researchers may decide to use a different weighting strategy according to their practical and theoretical evaluations (Mazziotta and Pareto, 2022).

The quality of the resulting innovative composite indicators (SM-INDEX and SMART-INDEX), can be difficult to asses as there is no benchmark to compare them to. Further analyses, such as uncertainty and sensitivity analyses, can help understand how methodological choices in the construction of the indices affect the results (Saisana et al., 2005). However, such approaches should be enlarged in order to take into account emerging aspects form novel data sources (such as selection of social media accounts, data pre-processing and analytical methods to transform unstructured data to structured one) and the multi-source nature of the process (Rocci et al., 2022). These issues are being addressed in an ongoing study and are out of scope of the present paper.

6 Conclusions

The availability of new sources of data, such as social media, provides an excellent opportunity for augmenting business statistics and examining new aspects of phenomena of interest. In spite of this, statistical challenges and errors exist throughout the entire analysis process, from identification of the units of interest in the digital data source to data collection, pre-processing, analysis, and data augmentation. In the first part of the paper, these challenges are briefly discussed. As a means of augmenting the data, we propose a modular methodological framework organized in three layers that defines the tasks and the outputs of each block. In this study, we focus on the case of composite indicators as the basis for augmentation. We demonstrate how the combination of traditional and digital textual data can be used to derive smart composite business indicators.

The second part of the paper demonstrates, using a prototype application, how the proposed methodology can be applied to real-life data. The specific empirical exercise of measuring CSR proved that traditional and social media-based indicators measure different aspects of the phenomenon, and enriched information is derived through data augmentation. The resulting smart index provides an innovative measure of CSR commitment and communication effectiveness.

This application can serve as a prototype for the construction of socio-economic indicators, contributing to the advancement of methodological knowledge for the construction of socio-economic indicators based on traditional data augmented with textual data. A similar modular approach and composite indicator methodological framework can be applied to other contexts. As an innovative aspect, we also use Twitter metadata to enhance the information and construct the SM-INDEX. As metadata usage in the data processing is an emerging topic, more research is required to understand the opportunities and statistical challenges resulting from its use. We expect more research to be conducted in this area.

A single digital data source was considered to augment traditional data in this paper. The proposed framework, however, allows the consideration of multiple data sources. For example, researchers may supplement traditional data with website information, social media posts, and newspaper articles. Further research will be conducted in this area in the future.

It is worth noticing that the proposed approach relies on the possibility of identifying the units under investigation on the smart source of data. This is some way a specific advantage for business surveys and very difficult in the case units are individuals. In such cases, a similar methodology can nevertheless be developed by considering preliminary aggregation. This direction of research would require specific attention and could be the topic for further investigations.

Finally, evaluating the quality of innovative indicators is an important area for future research. In fact, in addition to traditional quality dimensions and techniques, it is necessary to identify specific quality dimensions that are relevant to the data source and use case.

Appendices

A The Structural Topic Model (STM)

In order to identify the content of unstructured textual data, a common approach is to implement topic modeling (TM). It is an unsupervised learning technique which allows to study the underlying properties of a text in order to discover the topics discussed and get signals from the data. Among the different algorithms to implement TM, we select the STM which was originally designed to analyze open-ended survey questions, and which is becoming increasingly popular due to the possibility of estimating models including document-level metadata and, thus, characterizing the relationship between topics and metadata.

In the following, we briefly introduce the STM algorithm. For more details, please refer to Roberts et al. (2016). Figure 7 represents the model in plate notation. A topic is defined as a mixture over words and a document as a mixture over topics. In STM, document-metadata influences two components of the model, the topical prevalence that is defined as the proportion of the document that is associated to a topic, and the topical content that refers to the usage rate of word in a topic.

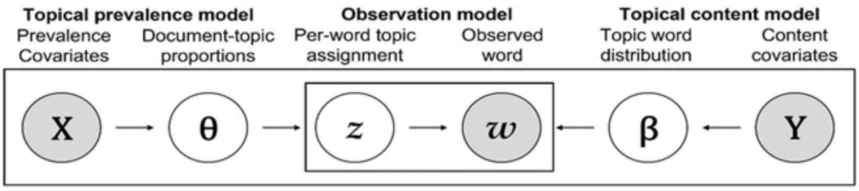


Figure 7: Structural Topic Model. Source: Amended from Roberts et al. (2016)

For the case study, we consider a previous work where topical prevalence covariates were included and the effect of time and sector on the discussion proportion of topics as part of a larger application-oriented study. As output, the STM model provides the per-word and per-document topic probabilities. We focus on the latter, i.e., we consider the probability of a document to be generated from a specific topic (also referred as to the proportion of text generated from a topic) as the input to build social media-based indexes. For our analyses, we use R and, in particular, the stm package (Roberts et al., 2019) to estimate the model and the quanteda package (Benoit et al., 2018) to clean and prepare the data.

B Details about topic modeling results

We identified 47 topics, 36 of which related to CSR activities. Table 1 shows an example of the topics for each CSR dimension. More details are available in Salvatore et. al (2022).

CSR Dimension	Description of Topics
Economic	- CEO talks about leadership - Economic impact of the business - Announcement of partnerships
Social	- Social impacts of innovation and digitalization - Accessibility and inclusiveness (disability) - Creating a better world for everyone - Fighting discriminations - Preserving the culture of communities - Sustaining small businesses - Workplace well-being
Environment	- Reducing emissions and pollution - Clean water - Marine Conservation
Mixed-General CSR	- Sponsorship of events

Table 1: Summary of topic modeling results.

C Details about composite indicators

Firm	SM-INDEX	TRAD-INDEX	SMART-INDEX	Rank. SM-INDEX	Rank. TRAD-INDEX	Rank. SMART-INDEX
J&J	112.20	107.30	109.64	2	2 (=)	1
Cisco	111.00	107.30	109.09	3	2 (=)	2
Amgen	103.85	107.30	105.52	5	2 (=)	3
Coca Cola	101.19	109.42	104.98	9	1 (=)	4
IBM	100.74	107.30	103.82	10	2 (=)	5
Goldman Sachs	112.30	96.65	103.30	1	6 (=)	6
Intel	101.49	104.44	102.92	8	3 (=)	7
Dow	97.36	109.42	102.69	16	1 (=)	8
Verizon	97.69	107.30	102.05	14	2 (=)	9
3M	97.09	107.30	101.69	17	2 (=)	10
Visa	101.67	101.59	101.63	7	4 (=)	11
Procter & Gamble	96.81	107.30	101.52	19	2 (=)	12
Salesforce	103.66	99.36	101.42	6	5	13
Microsoft	96.81	104.44	100.34	18	3 (=)	14
Nike	93.13	107.30	99.22	22	2 (=)	15
Travelers	96.32	101.59	98.81	20	4 (=)	16
Walmart	94.82	101.59	97.97	21	4 (=)	17
UnitedHealth	98.30	96.65	97.46	13	6 (=)	18
American Express	88.89	107.30	96.37	25	2 (=)	19
Caterpillar	90.61	96.65	93.43	23	6 (=)	20
JPMorgan Chase	89.70	93.52	91.53	24	7 (=)	21
McDonalds	100.70	84.09	90.90	11	8	22
Chevron	99.53	81.76	88.90	12	9	23
Honeywell	104.66	76.14	85.90	4	10 (=)	24
Home Depot	78.63	93.52	84.79	26	7 (=)	25
Boeing	97.63	76.14	84.23	15	10 (=)	26

Table 2: Social media-based (SM-INDEX), traditional (TRAD-INDEX) and smart indexes values with ranking.

Bibliography

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, 8(1):89–119.

Araujo, T. and Kollat, J. (2018). Communicating effectively about csr on twitter: The power of engaging strategies and storytelling elements. *Internet Research*.

Barcaroli, G., Nurra, A., Salamone, S., Scannapieco, M., Scarnò, M., and Summa, D. (2015). Internet as data source in the istat survey on ict in enterprises. *Austrian Journal of Statistics*, 44(2):31–43.

Barcaroli, G., Scannapieco, M., and Summa, D. (2016). On the use of internet as a data source for official statistics: a strategy for identifying enterprises on the web. *Rivista italiana di economia, demografia e statistica*, 70(4):20–41.

Bender, S. and Sakshaug, J. (2021). Data sources for business statistics: What has changed? *The Survey Statistician*.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Bernal, I. and Sejersen, T. (2021). Big data for economic statistics. Technical report, Stats Brief, Issue 28, United Nations.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, 74(5):817–848.

Booysen, F. (2002). An overview and evaluation of composite indices of development. *Social indicators research*, 59(2):115–151.

Carroll, A. B. et al. (1991). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business horizons*, 34(4):39–48.

Ceron, A., Curini, L., and Iacus, S. M. (2016). *Politics and big data: Nowcasting and forecasting elections with social media*. Routledge.

Chae, B. and Park, E. (2018). Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling. *Sustainability*, 10(7):2231.

Costanzo, L. (2011). Use of administrative data and use of estimation methods for business statistics in europe: an overview. In *Admin Data ESSnet Workshop "Using Admin Data-Estimation approaches"(Vilnius*.

Da Mota, V. T. and Pickering, C. (2021). Assessing the popularity of urban beaches using metadata from social media images as a rapid tool for coastal management. *Ocean & Coastal Management*, 203:105519.

Daas, P. J., Burger, J., Le, Q., ten Bosch, O., and Puts, M. (2016). Profiling of twitter users: a big data selectivity study. Technical report, CBS discussion paper.

Daas, P. J. and Puts, M. J. (2014). Social media sentiment and consumer confidence. Technical report, ECB Statistics Paper.

Daas, P. J. and van der Doef, S. (2021). Using website texts to detect innovative companies. Technical report, CBS Working paper no.: 01-21.

Dahlsrud, A. (2008). How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate social responsibility and environmental management*, 15(1):1–13.

De Fausti, F., Pugliese, F., and Zardetto, D. (2019). Towards automated website classification by deep learning. *arXiv preprint arXiv:1910.09991*.

De Muro, P., Mazziotta, M., and Pareto, A. (2011). Composite indices of development and poverty: An application to mdgs. *Social indicators research*, 104(1):1–18.

De Waal, T., van Delden, A., and Scholtus, S. (2019). Quality measures for multisource statistics. *Statistical Journal of the IAOS*, 35(2):179–192.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Howison, J., Wiggins, A., and Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12(12):2.

Iacus, S. M., Porro, G., Salini, S., and Siletti, E. (2020). An italian composite subjective well-being index: The voice of twitter users from 2012 to 2017. *Social Indicators Research*, pages 1–19.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). Big data in survey research: Aapor task force report. *Public Opinion Quarterly*, 79(4):839–880.

Luhmann, M. (2017). Using big data to study subjective well-being. *Current Opinion in Behavioral Sciences*, 18:28–33.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3):469–487.

Mazziotta, M. and Pareto, A. (2013). Methods for constructing composite indices: One for all or all for one. *Rivista Italiana di Economia Demografia e Statistica*, 67(2):67–80.

Mazziotta, M. and Pareto, A. (2020). *A Brief History of Time: From the Big Bang to Black Holes*. G. Giappichelli Editore, Torino.

Mazziotta, M. and Pareto, A. (2022). Weighting in composite indices construction: the case of the mazziotta-pareto index. *Rivista Italiana di Economia Demografia e Statistica*, Forthcoming.

OECD (2008). *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.

Perez, B., Musolesi, M., and Stringhini, G. (2018). You are your metadata: Identification and obfuscation of social media users using metadata information. In *Twelfth International AAAI Conference on Web and Social Media*.

Ricciato, F., Wirthmann, A., and Hahn, M. (2020). Trusted smart statistics: How new data will change official statistics. *Data & Policy*, 2.

Rill, S., Reinel, D., Scheidt, J., and Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69:24–33.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software*, 91:1–40.

Rocci, F., Varriale, R., and Luzi, O. (2022). Total process error: An approach for assessing and monitoring the quality of multisource processes. *Journal of Official Statistics*, 38(2):533–556.

Roelands, M., van Delden, A., and Windmeijer, D. (2018). Classifying businesses by economic activity using web-based text mining. Technical report, CBS discussion paper.

Rosales Sánchez, C., Craglia, M., and Bregt, A. K. (2017). New data sources for social indicators: the case study of contacting politicians by twitter. *International journal of digital earth*, 10(8):829–845.

Saisana, M., Saltelli, A., and Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):307–323.

Salvatore, C., Biffignandi, S., and Bianchi, A. (2021). Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156(2):601–630.

Salvatore, C., Biffignandi, S., and Bianchi, A. (2022). Corporate social responsibility activities through twitter: From topic model analysis to indexes measuring communication characteristics. *Social Indicators Research*, 164(3):1217–1248.

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1):399–422.

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5):503–516.

Struminskaya, B., Lugtig, P., Keusch, F., and Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social Science Computer Review*, 0(0):0894439320979951.

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110:298–310.

Thomassin, M. (2018). The migration of the canadian census of agriculture to an integrated business program without contact with respondents. In *Fifth International Workshop on Business Data Collection Methodology, Lisbon.*

Trappmann, M., Haas, G.-C., Malich, S., Keusch, F., Bähr, S., Kreuter, F., and Schwarz, S. (2022). Augmenting survey data with digital trace data: Is there a threat to panel retention? *Journal of Survey Statistics and Methodology.*

Chapter 5

Conclusions and final remarks

This research work addressed one of the emerging issues in survey research: the use of novel (non-probabilistic) data sources for inference and their integration with traditional data to augment the available information. It employed a diverse range of methodologies, including bibliometrics, text mining, Bayesian inference, and composite indicators, to address three research questions.

Through an original literature review analysis which uses text mining and bibliometric tools, we are able to answer to the first research question, thus providing insights into the evolution of the field in response to the rise of new data sources in order to exploit their advantages and address their challenges. It showed a shift from traditional in-person interviews to telephone and web-based surveys, including volunteer and opt-in panels. With the increasing use of mobile devices for online surveys, new considerations have emerged regarding questionnaire design and methods for combining different survey modes.

The pandemic emphasized the need for real-time data and the opportunities deriving from digital trace data to measure new phenomena. This has brought increased focus on the inferential and data quality aspects that need to be further explored in future studies. Digital trace data encompasses various sources, such as social media, Google trends, and data donation packages, which are often unstructured and possess distinct characteristics. As a result, different quality and inferential aspects should be considered separately for each source, as emerging from the literature. Additionally, understanding individuals' willingness to share their digital data (similar to consent in surveys) has emerged as an important aspect to be studied.

As a final point, the literature analysis suggested that while probability sample surveys are still central in survey research, their integration with non-probabilistic data, both structured (e.g. volunteer web surveys) and unstructured (e.g. digital trace data), has the potential to augment the available information, improving inference and reducing the production costs of statistics. From an inferential perspective, there are additional areas that require further exploration, such as the study of the selection mechanisms, in particular when it is missing-not-at-random, the use of different statistical approaches (e.g. machine learning) and the development of new methodologies to integrate data.

Given this overview, the thesis addresses two more research questions: how to improve analytic inference combining probability and non-probability samples in a way that also reduces costs and how to produce smart statistics combining traditional and digital trace data.

The paper in Chapter 3 presented a novel Bayesian data integration approach to improve analytic inference about parameters of logistic regression. Through a simulation and a case study, we demonstrated the effectiveness of the proposed approach, which results in more efficient regression estimates and lower survey costs. The Shiny app is one of the key contribution of this paper as it not only allows for the replication of the study, but also enables researchers who are interested in applying our approach to perform an interactive cost analysis. While our study only considers the presence of selection bias in the non-probability samples, a future research direction could be the development of a similar framework to address measurement error. This is an area we are currently exploring in ongoing research. Further, the current framework could be extended to other types of categorical variables (e.g. multinomial, ordinal) and to incorporate complex sample design features (e.g. stratification).

The paper in Chapter 4 provides researchers with a modular framework in order to construct business smart indicators. It consists of three layers, each of which defines specific tasks and outputs. A key advantage of the approach is its modularity, which enables researchers to tailor the framework to their specific needs and to explore other methodological variants (instances) within the same methodological architecture. We illustrate the proposed approach through a practical exercise that demonstrates how traditional and innovative indicators measure different aspects of the phenomenon, and how data augmentation leads to enriched information. However, the study only examines two data sources (one traditional and one innovative). Future research could expand the framework to include multiple data sources, such as social media, newspapers, and websites. An important area of research that needs further exploration and is currently being addressed in an ongoing study is the assessment of the indicators' quality. It is essential to identify specific quality dimensions that account for the multi-source nature of the integration problem, in addition to traditional quality dimensions and techniques. These dimensions should be specific to the data source and case study.

In conclusion, research in the field of survey data integration and inference for non-probability samples is expanding and becoming increasingly dynamic. Combining different data sources, especially traditional and innovative ones, is a powerful way to gain a comprehensive understanding of a topic, exploring new perspectives, and can result in new and valuable insights.

This thesis contributes to the current debate in the literature by providing original methodological results and considering a broad perspective in terms of analytical tools (text mining, Bayesian inference and composite indicators) and data sources (volunteer web surveys and textual data from social media).