Improving Predictions of Response Propensities for Effective Adaptive Survey Design (ASD)

Shiya

Printed by: Ridderprint | https://www.ridderprint.nl/

Cover design by Shiya Wu

ISBN 978-94-6483-173-3

Support for the research in this dissertation was provided by the China Scholarship Council, under Grant number: 201606620038.

Improving Predictions of Response Propensities for Effective Adaptive Survey Design (ASD)

Verbetering van voorspellingen van responsneigingen voor een effectief adaptief enquêteontwerp

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 16 juni 2023 des middags te 12.15 uur

door

Shiya Wu

geboren op 1 november 1989

te Hubei, China

Promotor:

Prof. dr. J.G. Schouten

Copromotor:

Dr. M. Moerbeek

Beoordelingscommissie:

Prof. dr. J. van den Brakel (Maastricht University)

Prof. dr. I.G. Klugkist (Utrecht University)

Prof. dr. A.G.J. Van De Schoot (Utrecht University)

Dr. B. Struminskaya (Utrecht University)

Prof. dr. J. Wagner (University of Michigan)

Table of Contents

Chapter 1

General Introduction

Survey is a research method of gathering information (e.g., customer feedback) from a sample of individuals in a population. It typically involves asking a set of questions on a specific topic to the selected participants. Pioneered by sociologist Paul Lazarsfeld in the 1930s-40s, surveys were initially administered to study the phycological and cultural effects of radio. Today, they are commonly used to gain a fair representation of the socioeconomic status and well-being of individuals. Therefore, surveys become a vital tool for informed theory testing and decision making in social science.

Collecting *quality* survey data requires careful planning and investment of time and money. Over many decades, the method of data collection has evolved as the pressure to obtain quality data has grown. The survey practice faces several challenges, including difficulty answering research questions, difficulty recalling answers from long-term memory, increasing cost, declining participation possibly due to the lack of understanding of context (Krosnick, 1991; Groves et al., 2011). These challenges can limit the quality of survey data and make it harder to obtain the information needed to decisions making. Motivated by these factors, the data collection methods have evolved over time. From traditional and simple approaches (paper-and-pencil interviewing) to more sophisticated approaches (computed-assisted interviewing), survey practitioners have adapted their methods to overcome these challenges and maintain the quality. However, each method has its advantages and disadvantages. For example, web and multiple-device methods are low-cost and have become increasingly popular in recent years, whereas the use of multiple devices, such as tablets and smartphones, has rapidly taken over the web-only method (de Leeuw & Toepoel, 2018; Link et al., 2014; Mavletova, 2013; Zijlstra et al., 2018). This is because people without internet access are not represented in web survey samples. Additionally, breakoff—people start but incomplete the survey—is common in web surveys, and results in considerably low response rates (Manfreda et al., 2008; Musch & Reips, 2000; Peytchev, 2011). In this regard, a mixed survey method, which combines the strengths of diverse methods, can offer the greatest potential in terms of cost-quality balance for researchers and policy makers.

Finding a single mixed method that is ideal for all sampled units is unrealistic, as each method has a variety of features, and people have different preferences for in what condition, when, and how they like to be contacted and to participate. Adapting survey

features, such as the number of contact attempts, the data collection mode, and the type of incentives offered, to people's characteristics by intervening or tailoring in data collection can make the survey more attractive and accessible to different subgroups, improve response rates and representativeness, and the data quality.

Adaptive survey designs (ASD, Wagner, 2008; Schouten et al., 2017) have emerged as a data-driven recruitment method that optimize metrics of data collection progress, cost and quality (Schouten et al., 2009; Wagner, 2010). ASD provides a framework for such tailoring/ intervening. During the ongoing data collection process, metrics are monitored in an effort to inform fieldwork decisions (Kreuter, 2013). This dissertation focuses on developing a methodology to enhance the robustness of ASDs. The outline of this introduction is arranged as follows: Chapter 1.1 provides a brief context for ASD, Chapter 1.2 discusses the main objective of ASD, Chapter 1.3 describes the methodology behind effective ASD, and Chapter 1.4 explains the topic and contribution of this dissertation.

1.1 Adaptive Survey Design (ASD)

1.1.1 Precedents

ASDs tailor different strategies, before or during data collection, for different sample units based on their characteristics. This concept can trace its history back to dynamic treatment regimes in the field of clinical trials (Murphy, 2003). A dynamic treatment regime consists of a sequence of decision rules, one for each stage of intervention. Instead of a "one-size-fits-all" intervention, the type, dosage amount and timing of the subsequent intervention are individualized to a patient, based on history and preceding treatments, with a particular focus on optimizing response. This approach has been used to develop adaptive treatments for, e.g., alcohol-dependent patients (Murphy et al., 2007). In this context, measures of alcohol abuse are tracked. A patient, who is observed with a second heavy drinking day within two months, is provided cognitive behavioral intervention to augment medication. Otherwise, low-cost medical management is provided at the end of the period. In the clinical trials research literature, adaptive designs, in general, allow for changing or modifying the characteristics of a trial as knowledge accrues, and therefore, play a role in the planning, design, and implementation stages (Coffey et al., 2012).

The mechanism behind adaptation in trials is the stimulus for adaptation in sample surveys, as identified by Wagner, 2008; Mercer et al., 2017. Some adaptations in trials, such as stopping rules, sequential multi-assignment randomized trials, preplanned protocols, and a data monitoring committee, can be adjusted to the survey context; see Rosenblum et al., (2019). Translating these processes into applications in surveys necessitates comprehension of what uncertainty in data collection gives rise to the use of ASDs.

1.1.2 Motivation

ASDs are a response to an increasingly difficult survey climate characterized by falling response rates, growing reluctance to participate, rising costs of collecting survey data, and etc. (De Leeuw & de Heer, 2002; Groves & Couper, 2012). The consequences are that information collected from individuals is less representative, and the objectives of a survey may backfire. This problem is severe when sticking to the "one path fits all sample members" approach to data collection (Axinn et al., 2011). This standard protocol is unable to fulfill the growing need for surveys for quality data collection. Statistics computed from data have many uses. They can reflect the attributes of a population, forecast changes over time, estimate population totals, help with decision making, and so on. However, the precision and accuracy of survey statistics are affected by several sources of survey error (Groves et al., 2011).

In each stage of a survey life cycle, errors can arise, resulting in final statistics of inferior quality, as well as misleading conclusions. The stages required to obtain a survey statistic are: (1) planning of a survey and outlining the steps to take when conducting it, (2) collecting and processing data, and (3) data analysis. Groves et al., (2011) differentiate the sources of errors from two dimensions: those related to who survey practitioners talk to (*representation*), and those related to what practitioners learn from those conversations (*measurement*). See a framework of sources of error in surveys in Figure 1.1, and a brief interpretation in Table 1.1. For example, a survey might be interested in the relationship between individuals and the labor market. Two kinds of inference are made on demand for such an understanding. On the basis of respondents' answers collected from sampled cases, a survey manager must first infer these estimates about the relationship, which is

the process of measurement. Next, in the process of representation, the manager infers the relationship in the population as a whole from those estimates.

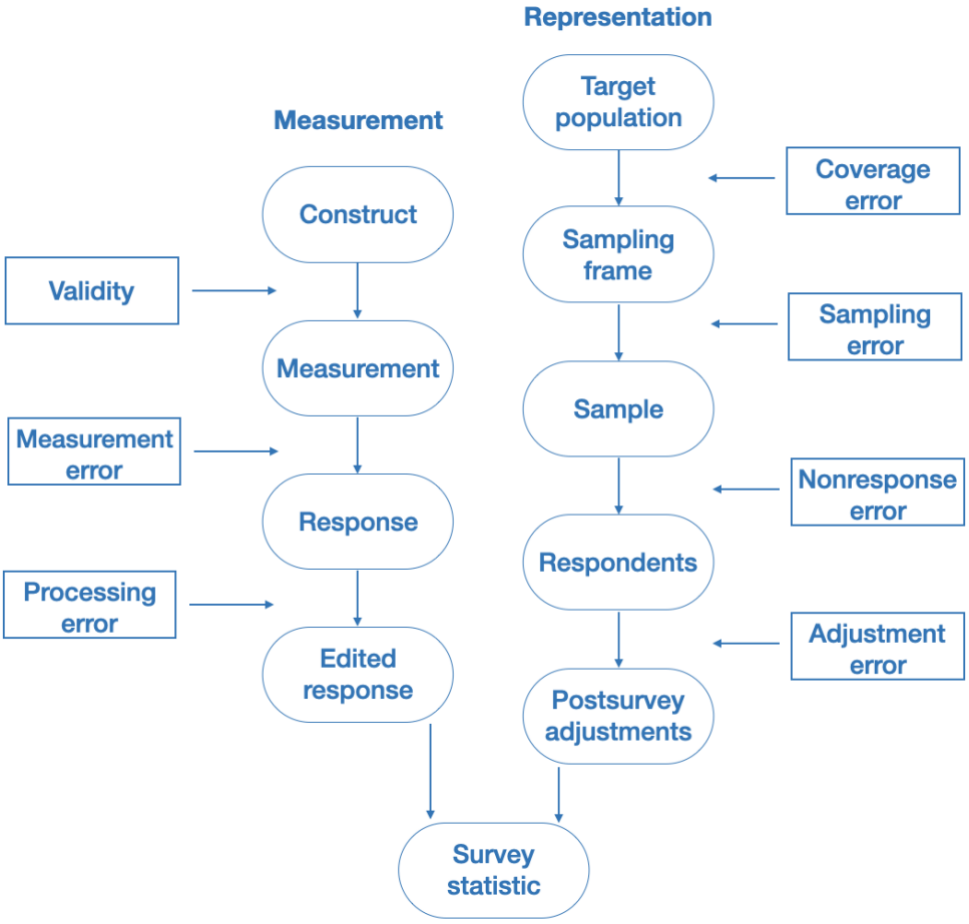

Figure 1.1 Survey life cycle from a quality perspective. Source: Groves et al., (2011); Raymer et al., (2015).

Each source of error affects the uncertainty of the survey estimates from two angles, namely, systematically and randomly. Systematic error and random variation lead to the statistical concepts of bias and variance, respectively. Bias, which causes a different survey estimate, refers to whether the estimate, on average, underestimates or overestimates the true value in the population of interest. A simple example of systematic measurement errors is the disparity in the number of sexual partners reported by men and women in face-to-face interviews (Tourangeau et al., 1997). Due to social desirability, women tend to underreport, while men tend to overreport. Another substantial concern, in addition to bias, is variability of the estimate. Random variation can generally reduce its precision, unlike systematic error that affects the point estimate. Precision describes the

extent to which survey estimates can be expected to vary, or "bounce around", if the survey was to be repeated many times. For instance, an estimate may be assumed to have a precision of ±10% at the 95% confidence interval. If the same survey was fielded 100 times, one is convinced that the results generated would fall within 10% of the true population value in 95 of the instances.

Taken together, those above-mentioned causes spur the development of ASDs. This technique aims to improve the sample representativeness and allocate cost-effective resources simultaneously.

Table 1.1 A brief description of survey error.

Error	**Description**
Coverage error	Arise from failure to encompass all components of the population being studied. Incomplete sampling frame often leads to these errors.
Sampling error	Occur when a selected sample does not represent the entire population.
Measurement error	Refer to the discrepancy between a measured quantity and its actual value.
Nonresponse error	Occur when the individuals that complete the interview differ systematically from those that were unable to be contacted and those that chose not to participate.
Processing error	Produced during data processing when the variable provided by the respondents differs from that in the estimation.
Adjustment error	Refer to the difference between a population parameter and its adjusted statistic.

1.1.3 Building blocks

Survey practitioners have noted that response behaviors vary depending on survey design features, such as the amount of incentives. Some sample units are more reluctant to

respond when higher incentives are assigned (Singer et al., 1999). This practical consideration forms the basis of tailoring.

The practices of tailoring are grounded in leverage-salience theory (Groves et al., 2000). The main idea behind this theory is sample *heterogeneity* in the attributes of a survey request relevant to the response decision. Those attributes include survey design features, such as the survey topic, survey mode, and incentive. Each survey design feature is commonly provided with leverage, salience, and valence, which can sway an indicvidual's participation decision.

Leverage refers to the intrinsic importance of a survey design feature to the decision. People assess a feature in a cognitive manner that places weight on how important that feature is. For any particular survey, individuals who value a feature (e.g., the topic), may volunteer to share their opinions. Others consider the topic to have negative leverage, yet may be lured into participating by incentives.

Salience is the emphasis placed on the feature by survey practitioners during the survey request. In the above example, those to whom the topic is most important are swayed to participate, when the survey request brings up the topic more saliently, in comparison to other survey design features. Correspondingly, the survey request should underline incentives more prominently for those motivated by incentives.

Valence pertains to being a part of leverage, and is positively or negatively related to the participation decision. Any feature with a positive valence weight can encourage individuals to consent to participate in an interview. Conversely, any negative valence feature encourages refusal.

This theory stresses that survey practitioners play an important role in making survey design features salient to people. Additionally, the theory places great emphasis on converting refusals, that is, optimizing perceived survey costs, in an effort to reduce nonresponse (e.g., Groves & McGonagle, 2001). The perceived cost is the cost that undertaking any particular survey incurs, and relates closely to the risk of survey errors (Groves, 2005). An attempt to convert refusals in fact affects the people's perception of the benefits and costs of the decision-making process. In a particular survey, the

likelihood that a sampled person who will become a respondent is called the *response propensity* (Lavrakas, 2008). Due to the heterogeneity discussed above, sampled persons differ in the likelihood of becoming a respondent. By implication, the propensity to respond is a function of the characteristics of an individual and the design features to assign.

Therefore, heterogeneity is the crux of ASDs. Applying a homogenous survey design protocols to all individuals is evidently obsolete. As inspired by leverage-salience theory, protocols should instead be tailored and adapted to each respondent. Survey practitioners can use collected data to identify the most effective protocol—an explicit set of the sequence, dosage, and timing—that works best for different individuals at the onset of or during the course of data collection. The data used come from information available on the sampling frame and/or register, commercial auxiliary data, the growing ubiquity of data generated in the data collection process, and real-time data collection systems (Couper, 1998).

1.1.4 Variations

ASD vs. Responsive Survey Design (RSD). A recent research topic considered is tailored survey designs to counteract survey errors, i.e., reduce nonresponse, which is the greatest focus in survey methodology literature. There are two terms that are rather similar in concept but different in application. Groves & Heeringa (2006) first introduced the term "responsive", which was used afterward by, e.g., Couper & Wagner (2011); Wagner et al., (2020). This was followed by the term "adaptive", employed by Wagner, (2008); Schouten et al., (2013). The main idea behind both ASDs and RSDs is to tailor design features to optimize the quality of the survey estimates and survey cost. Occasionally, RSDs are regarded as a special case of ASDs; see Bethlehem et al., (2011). However, the difference should be noted. RSDs reply on data collected during the course of fieldwork, called paradata, and response data. Measures are identified in the first phase, and based on these measures, changes are made to tailor design features to specific sample units. In contrast, ASDs are based on prior information available and specify potentially influential design features before the start of data collection.

Dynamic vs. Static ASD. ASDs can be classified as static or dynamic designs (Schouten et al., 2013). Both allocate strategies (a set of design features) based on prior information on the interaction between sample units and the available treatments. Designs that employ auxiliary data available before the onset of fieldwork are termed static, while those that additionally consider paradata are termed dynamic. For static designs, the allocation of strategies is set before the start of fieldwork. For dynamic designs, the strategy set is known beforehand, but the allocation itself can be made only when paradata is available.

ASD vs. post-survey adjustment. The well-established technique of post-survey adjustment is applied to adjust survey results to bring them into line with known population characteristics (Kalton & Flores-Cervantes, 2003). These adjustments attempt to reduce sampling variation, and to compensate for the effect of missing data due to an incomplete sampling frame or nonresponse. Weight is allocated to survey respondents who are the selected representatives of missed groups of individuals. The objectives of these adjustments overlap with those of ASDs. This implies that utilizing the given frame and administrative data in these adjustments is as effective as ASDs. ASDs have additional benefits of improving the quality of survey estimates (bias and variance) beyond post-survey adjustment methods. See Särndal & Lundquist, (2014); Schouten et al., (2016); Zhang & Wagner, (2022) for theoretical and empirical evidence. ASD is a less expensive approach and is worth the effort. Nevertheless, if the objective is to obtain smaller biases and variances, ASD would be combined with post-survey adjustment, e.g., Särndal & Lundquist, (2019).

1.2 The Main Objective: Improve Balanced or Representative

Wagner, (2008) pioneered the use of adaptive designs in survey methodology as newly developed data collection methods to address the nonresponse problem. Tailoring strategies that vary in individuals can promote the propensity to respond. These approaches (ASDs) can be used simultaneously to increase the response rates, improve sample representativeness, and control data collection costs.

As described in Table 1.1, the nonresponse suffered by most surveys results from sampled individuals not providing the required information. This lack of response has serious impacts on the quality of survey statistics, and contributes to bias in estimates. In

particular, nonresponse bias is a major concern in surveys with low response rates. Due to the potential high nonresponse rate, sources of bias in estimates must be identified in a nonresponse analysis. For this reason, the survey methodology literature mostly restricts quality-cost trade-off assessments to this issue. Some studies have additionally extended the objectives of ASD to measurement error (Calinescu et al., 2012; Calinescu & Schouten, 2016).

As Chapter 1.1.2 noted, bias is the systematic difference between an estimate and the true value. *Nonresponse bias* of a response mean is defined as the product of the amount of nonresponse and of the difference between respondents and nonrespondents. This can be represented analogously as the ratio of the covariance between the survey variable and the response propensity, and the mean response propensity (Bethlehem, 1988).

Nonresponse bias is unspecified for survey variables, as the true value of the population is unknown. In recent years, survey methodologists have proposed surrogates of nonresponse bias as an indirect way to analyze bias. These surrogates, called indicators, can ensure the quality of the data available to survey users and data collection managers. Moreover, such indicators serve as quality criteria to help make decisions and optimize the allocation of limited resources in ASD.

The most broadly used quality indicator is the response rate (Biemer & Lyberg, 2003). This indicator acts at best as the upper limit of nonresponse bias. For a high response rate, it is not necessarily appropriate to jump to the conclusion that the estimates are accurate enough. Nonresponse bias can thereby be worsened in a setting in which one increases the response rate, while the difference between nonrespondents and respondents increases concurrently. This built-in risk would occur in the situation that a survey pursues a higher response rate. Focusing only on this indicator, which rarely addresses the nonresponse bias, is misleading. On its own, the response rate may be a poor indicator of nonresponse bias, as exemplified in the survey methodology literature (e.g., Groves & Peytcheva, 2008; Schouten, 2004).

To manage the nonresponse bias risk well, alternatives to the response rates are in demand (Wagner, 2012). Nonresponse indicators for data collection are split into those that provide overall insight into the consequence of the risk, and those that provide greater

detail. One often-studied group of overall indicators are representative indicators: R-indicators and the coefficient of variation of response propensities (CV for short). Schouten et al., (2009) defined that a response mechanism is called *strongly representative* with the respect to the sample in case the response probabilities are constant. A response mechanism is *weakly representative* with the respect to the sample for auxiliary variables if the corresponding response propensities are identical. Each is defined as a function of response propensity. The R-indicator measures the distance of response propensities to the response rate, estimated by a statistical model based on observed auxiliary variables. The statistical concept behind this transforms the standard deviation of subgroup response propensities to a 0-1 interval. This form serves as a measure of variation in subgroup response rates. Schouten et al., (2016) empirically show that low variability indicates a low risk of nonresponse bias and representativeness. One may call into question the performance of the R-indicator when monitoring and controlling the survey process. One potential limitation is that at low response rates, representativeness is probably overestimated in early data collection. Thus, the decision of when to modify methods or to end data collection becomes problematic. In contrast, CV in such scenarios is more attractive than R-indicator (Moore et al., 2018 & 2021). CV is the variation in subgroup response rates standardized by dividing by the mean response propensity; a low CV indicates representativeness. The magnitude of CV specifies the maximum standardized nonresponse bias of the survey variables (Schouten et al., 2011).

CV and its counterpart, R-indicator, are informative regarding actual bias. These overall indicators, however, cannot identify subgroups to improve the quality. To be optimally informative, partial indicators, which are partial decompositions from overall indicators, measure propensity variation associated with auxiliary variables (De Heij et al., 2015; Schouten & Shlomo, 2017). They enable a fine-grained examination of which variables and/or which categories within each cause an absence of representativeness, by how much, whether the bias will persist after adjustment on variables, and what subgroups are targeted when modifying methods; for example, Nishimura et al., (2016); Sakshaug & Antoni, (2019). Additionally, partial indicators can be classified as unconditional or conditional forms. Unconditional indicators describe the between variance and measure deviation from representativeness for each auxiliary variable, while conditional indicators

correspond to the within variance and represent deviation from conditional representativeness. See Shlomo & Schouten, (2013) for the theoretical details.

Overall indicators, together with partial decompositions, summarize bias and underpin the planning of, monitoring, and management of the data collection progress. Besides, they have proven useful for comparisons of the representativeness of waves in a survey, or of different surveys equipped with the same variables, see Schouten et al., (2012); Shlomo et al., (2012); Luiten & Schouten, (2013). The RISQ project (Representative Indicators for Survey Quality) has a well-developed functionality of translating auxiliary information to, computing and developing these quality indicators, exploring their characteristics, and exhibiting their use in fieldwork practice; see http://www.risq-project.eu/ for details.

1.3 Effective ASDs

ASDs include four key ingredients, each interacting with the others: auxiliary variables (create strata), design features or interventions (tailored or adapted to strata), quality or cost indicators (brace decisions), and optimization strategy (quality-cost trade-offs). Refer to Schouten et al., (2013) for an elaboration of each component.

The effectiveness of ASDs is subject to bias and variance in the quality indicators, and clearly, they are in turn subject to inaccuracy in survey design parameters, such as response propensities. Survey practitioners typically pursue expected response propensities by means of generalized linear models. Such models are built on the accumulating data from the current data collection wave, or on the historic data from previous data collection waves. They are explained further below, and can be further developed into more advanced methods later.

The current approach. The model coefficients are estimated based on data accumulated in the current wave; accordingly, predictions on response propensities are generated by applying the estimated coefficients to the remainder of this wave. In this wave, the fieldwork is still ongoing, resulting in a dataset that is not yest as comprehensive as those collected in previous phases. The incomplete data, in other words, are possibly under-representative of incoming data that are collected later in the wave. For example, a subgroup that participates willingly early may be unwilling to share their opinions later,

possibly because of losing interest in the survey topic. The research of Wagner & Hubbard, (2014) is evidence of the incomplete accumulated data, particularly at early time points, resulting in biased and volatile predictions of response propensities. Alternatively, such data reduces the prediction performance of categorical responses. Moreover, biased predictions or reduced performance can lead to inefficient decisions made via optimization strategies, such as allocation to nonresponse follow-up (Calinescu et al., 2013; Thompson & Kaputa, 2017).

The historic approach. Historic data obtained from prior fieldwork can also be employed. Using historic data in propensity models has the potential to improve predictions. The rationale for this approach is that historic data may be more representative of a complete wave, as opposed to the current approach of using accumulating data. Historic data are used to estimate expectations about model coefficients and predictions in a similar way to the current approach (See Schouten et al., 2013 & 2017). This approach is premised on the assumption that data from prior implementations and those from the current implementation comply with a similar data-generating process. This assumption is clearly contrary to survey circumstances in which the implementations are rather dissimilar over time. This difference may be caused by changes in design features or a time lag between the current and the last implementation.

Using either historic or current data alone for model-based predictions has limitations regarding accuracy and reliability, as either approach generates point estimates. Furthermore, as time passes, relying solely on historic data can lead to decreased relevance to predictions, as evidenced by the decline in response rates over the years. Researchers have been investigating new approaches to compensate for each approach's deficiencies.

The Bayesian approach. Historic and current data are adopted both in the prediction process. Prior beliefs concerning response propensities, rather than points estimates, are generated from historic data. These priors are updated to posteriors using current data. The resulting posteriors serve as priors for the next wave. As current data accumulate, the updating process is repeated; see Schouten et al., (2018) for a systematic study of the use of Bayesian methods for the prediction of response and cost parameters. The simulation results support the findings that combining historic and current data can improve

predictions. Emerging empirical evidence supports this discovery, when using other sources of historic data, such as published estimates (West et al., 2021). The Bayesian approach fully exploits the characteristics of historic data being representative and current data being informative. The resulting predictions are provided with a range of values in an effort to specify a particular probability of true but unknown values falling within this range.

The inherent replication, together with the Bayesian method, paves the way for consolidating the performance of ASD over time. Bayesian methods provide accurate and reliable predictions, formulating the basis of reliable decisions in time. The risk of nonresponse bias, i.e., representative indictors and their decompositions, can be monitored. Subgroups or individuals who have a low propensity to respond, are identified in the early phase of data collection. An intervention such as providing incentives or using a costly interview mode is thus required in response to recruit nonresponse. For those who do not need design changes, a different intervention may be called for, such as stopping data collection from the perspective of cost efficiency. Such decisions can be made based on quality metrics of nonresponse indicators, such as the root mean square error of the predictions versus that of the observations.

1.4 Contribution and outline of this dissertation

Predictions may be improved by incorporating additional sources of historic information. However, the effects of considering the different sources of historic information such as data collection expert, and the timeliness of historic data on predictions, are unknown. This raises the main research question considered in this dissertation:

How much improvement in ASD performance can be achieved through embedding expert knowledge and historic survey data?

Central in this dissertation is to reduce nonresponse error and improve the effectiveness of ASDs by making precise and timely predictions of response propensities. Adaptation is strongly sensitive to inaccurate response propensity. To take uncertainty into account, a novel methodology introduces a predictive model for response propensity parameters. This applies to infrequent or redesign surveys (see Chapter 2), and long-running surveys

(see Chapters 3 and 4), by leveraging historic data in a nuanced way. To anticipate the "state-of-the-art" interventions for the upcoming data collection round, the limited cost-prohibitive resources are optimally allocated across strata (see Chapter 4). An exploration is executed into the sensitivity of ASD performance to: (1) the pooling methods of experts and the choices of relevance (see Chapter 2), (2) the length of data collection (see Chapters 3 and 4), and (3) some specific budget levels (see Chapter 4).

Evaluations are made based primarily on data from the Energy Survey (Hernieuwbare energie in Dutch) and European Union Statistics on Income and Living Conditions Survey (SILC) in Chapter 2, and the Dutch Health Survey (GEZO in Dutch) in Chapters 3 and 4, from Statistics Netherlands.

A succinct summary of these three chapters, along with the associated research questions derived from the main research question, is provided below.

Chapter 2 addresses the need for a new method for accurate response propensity predictions for specific types of cross-sectional surveys, be they conducted infrequently or new due to redesigns. This chapter was inspired by the adoption of a Bayesian analysis when making predictions, as discussed by Wagner & Hubbard, (2014); Schouten et al., (2018), such that the bias of predictions made by relying on only historic data, can be mitigated. The following research question is addressed:

- How can we structure prior elicitation from historic data and expert knowledge?

We elicit expert knowledge from data collection staff, through a self-response questionnaire designed to gather predictions on the similarity between relevant past surveys and a new survey regarding expected response propensities, and on sample sizes of historic survey data sets. Historic data sets and expert respondent data translate into informative priors of response propensities for the new survey. Priors are updated to posteriors that serves as priors for the next round, based on data accumulated. We assess the performance of expert elicitation, against the worst-case scenario that has no access to historic surveys; thus, non-informative priors are set. This chapter considers quality metrics for nonresponse, providing ongoing feedback.

Chapter 3 studies time change in response propensities of Computer-Assisted-Web-Interview mode (CAWI) in the Dutch Health Survey (GEZO). In Chapter 2, we develop

an estimation strategy for response propensities in a short-term data collection climate, where those parameters are stable. This chapter proceeds to a long-term data collection climate to analyze the time-dependent variation in predictions. This approach starts from the fact that while survey response rates have been on the decline many years, research on the cause-and-effect relations between factors and responses is limited to date. Those changes call into question the tenuous future of adapting survey design. In this way, the following four research questions are addressed:

- What time-series components contribute most to variation in response propensities?
- What level of response propensity prediction accuracy can be achieved for the next upcoming time period?
- How does prediction accuracy vary over population strata?
- How does prediction accuracy depend on the length of the historic survey time series?

We disentangle the overall variation in response rates, and break it down into the essential constituents through multilevel models. The models are proposed in a Bayesian manner. In this way, we manage to make precise and timely predictions by learning from historic time series and updating outdated inferences, and consequently, to further reduce the risk of nonresponse error.

Chapter 4 focuses on ASD performance when optimizing cost-effective decisions in sequential mixed-mode surveys. In Chapter 3, we consider time change in response propensities regarding predictions for single-mode data collection. This chapter takes a further step by simultaneously addressing the timeliness of solid decisions, and the accuracy of each phase of prediction, to fill this gap in previous study by addressing the following three research questions:

- How can time-series models be constructed to improve response propensity prediction accuracy in a sequential mixed-mode design?
- How sensitive is ASD performance to the specific budget level?
- How does ASD performance depend on the length of historic data?

We adjust the models in Chapter 3 to suit data collected from mixed-mode survey designs. The developed models additionally allow for time-dependent variation in conditional

response propensities, and for between-mode correlations in propensities. By means of these predictions as input, we propose an allocation model for effective ASD. Then, a simple optimization setting is analyzed, where the availability of the interview mode, to be allocated to population strata, is constrained by a threshold on cost overrun. The objective is to minimize the expected nonresponse bias by recruiting a fraction of non-responders. The resulting problem is nonconvex and nonlinear. We show that the assignment task for ASD can be cast as a mathematical programming problem to explicitly account for time-dependent variation. Consequently, the solution is obtained by a nonlinear programming solver. We subsequently explore the sensitivity of the optimized performance to the budget level and the length of historic data sets.

Chapter 5 summarizes the outcomes from the analysis conducted in Chapters 2-4, and discusses the strengths and weaknesses of each methodology considered in this study. Furthermore, suggestions and recommendations are provided for future research, and implications for actual survey practice are discussed. I present my personal thoughts on the future of adaptive survey designs at the end of this chapter.

This dissertation advances on the methodology behind effective ASD, providing the new propensity models to use historic data sets in a nuanced manner for precise and timely predictions, and presenting a heuristic approach to the timeliness and effectiveness of interventions during data collection. I believe that the methodology proposed here lays the groundwork for the future challenges despite the empirical studies being focused on Dutch surveys.

Chapter 2

Data Collection Expert Prior Elicitation in Survey Design: Two Case Studies

This chapter is published in Journal of Official Statistics (JOS) as: Wu, S., Schouten, B., Meijers, R., & Moerbeek, M. (2022). Data Collection Expert Prior Elicitation in Survey Design: Two Case Studies. *Journal of Official Statistics, 38*(2), 637-662. http://dx.doi.org/10.2478/JOS-2022-0028.

Author contributions: Statistics Netherlands provided the survey data. SW, BS, RM, and MM contributed to the study concept. SW, BS, and RM designed the questionnaire, elicited expert data, and coded the data. SW performed the statistical analyses and wrote the paper. BS and MM critically reviewed the paper. BS wrote sections of the manuscript.

Abstract

Data collection staff involved in sampling designs, monitoring and analysis of surveys often have a good sense of the response rate that can be expected in a survey, even when this survey is new or done at a relatively low frequency. They make expectations of response rates, and, subsequently, costs on an almost continuous basis. Rarely, however, are these expectations formally structured. Furthermore, the expectations usually are point estimates without any assessment of precision or uncertainty.

In recent years, the interest in adaptive survey designs has increased. These designs lean heavily on accurate estimates of response rates and costs. In order to account for inaccurate estimates, a Bayesian analysis of survey design parameters is very sensible.

The combination of strong intrinsic knowledge of data collection staff and a Bayesian analysis is a natural next step. In this chapter, prior elicitation is developed for design parameters with the help of data collection staff. The elicitation is applied to two case studies in which surveys underwent a major redesign and direct historic survey data was unavailable.

Keywords: Nonresponse bias; Bayesian; response propensity; expert elicitation

2.1 Introduction

We propose a strategy to elicit prior distributions from survey data collection staff for key survey design parameters. We focus on expert prior elicitation for new surveys, with relatively little historic data, but our approach is also applicable to repeated surveys. We do so with an adaptation of survey design to relevant population subgroups in mind.

In monitoring survey design (e.g., Kreuter, 2013), and adapting survey design (e.g., Schouten et al., 2017), design parameters, such as contact propensities, participation propensities and costs, are crucial input to decision making for data collection staff. Such parameters need to be estimated or predicted at a subgroup/stratum level and, therefore, have a certain bias and imprecision. When evaluating survey design performance, it is important that uncertainty of these parameter estimates can be accounted for (see Burger

et al., 2017) in order to avoid false conclusions. This importance is even greater when starting to adapt survey design.

In repeated surveys, the natural strategy is to estimate standard errors of parameters using recent historic survey data, but it is unclear how to deal with uncertainty in the setting of new or low frequency surveys. In such surveys there is no direct historic survey data. A natural strategy then is to adopt a Bayesian analysis with expert prior elicitation, see Gelman et al., (2013). The elicitation is included to build informative prior distributions of design parameters, incorporating the knowledge from similar historic surveys and/or literature related to the new survey, and to update these during and after data collection.

Schouten et al., (2018) discuss and evaluate the construction of a general Bayesian analysis for response and cost. They show that misspecified priors may lead to weaker performance than non-informative priors, which include no prior knowledge. Prior elicitation is, therefore, an influential step. For repeated surveys that are conducted at a relatively high frequency, say every year, quarter or month, prior elicitation is straightforward, unless (major) design changes are introduced, such as a change of survey modes. For redesigned surveys or for new surveys, prior elicitation can be complex, because available historic survey data differs on one or more survey characteristics. Data collection staff frequently deal with this complexity and have found tactics to extract information from the historic survey data. We attempt to structure these tactics.

Quantification of the uncertainty by means of elicitation by experts who have access to historic datasets, is not novel. It has been the subject of research in biometrics and medical statistics, see O'Hagan et al., (2006). However, to date, application is scarce in the field of survey monitoring and analysis. Two recent examples are Coffey et al., (2020) and West et al., (2021). Coffey et al. (2020) invited data collection managers as experts and West et al., (2021) reported studies in the literature.

Expert prior elicitation depends heavily on the statistical skills of the experts. In biometrics and medical studies, experts are often viewed as relatively less trained in statistics (Gosling et al., 2007; Oakley & O'Hagan, 2007). The elicitation then focuses strongly on transforming properties of prior distributions, such as medians, means, quantiles and variances, to questions that can be answered by experts. Oakley & O'Hagan,

(2007) introduce an additional step in prior elicitation in which a prior is set on the prior itself by means of Gaussian processes and updated by the summaries provided by the experts. In settings where experts have no training at all in statistics, prior elicitation may even resort to game-like approaches that faciliate experts to express their beliefs (O'Hagan et al., 2006; Veen et al., 2017). These approaches also tend to rate the experts themselves on their amount of expertise and assign and estimate weights to each expert.

Survey data collection staff involved in response and cost predictions are usually trained statisticians with a good sense of probability distributions. This means that expert elicitation can, and must, be more advanced. In fact, in our experience, experts, as a standard practice, search for relevant historic survey data sets and estimate survey design parameters directly from these data. This means elicitation translates to collecting information on sample sizes of historic survey data sets and on similarity between these past surveys and the new survey. We must stress, however, that also data collection experts may over or underestimate importance of certain survey design features, as argued for example in Brownstein et al., (2019). Here, we do not distinguish between the various skills that data collection experts need to possess, which is a topic for further research.

To include such expert knowledge, power priors are an obvious option. Power priors were introduced by Ibrahim & Chen, (2000) and further discussed in Ibrahim et al., (2015). Historic datasets D_k, labelled $k = 1,2, \dots, K$, from previous studies are assigned a scalar quantity γ_k representing their similarity. In the derivation of the posterior, the scalar quantities are included as powers to the data likelihoods. Obviously, the prior elicitation then amounts to a selection of data sets and the choice of the associated powers. Rietbergen et al., (2011) discuss how to elicit γ_k. In their approach, the experts rank the historic studies based on their relevance and provide a prescribed fixed weight per study based on heterogeneity between study characteristics. In the survey context, study characteristics may reflect survey design features such as the target population, the survey topics, and the survey modes. Data collection staff have a good sense of the most influential features and already select historic surveys based on these features. However, there usually is not a structured approach and two experts may end up with different predictions.

To structure prior elicitation, we perform five steps. The first is to select design features for rating similarity of surveys. The second is to assign importance weights to these features. The third is that for each historic survey the design features are scored on their similarity to the new survey on each of the features. The fourth is to weight the scores of each historic survey to the γ_k in the power prior. The last step is to apply the informative priors to the new survey and update them during data collection.

Adaptive survey designs tend to focus on nonresponse bias reduction and ignore other errors, such as measurement error. Also, we will restrict ourselves to nonresponse bias. Nonresponse bias cannot be measured directly but proxy indicators have been developed that signal an increased risk of bias. The most studied are representativeness indicators (see Schouten et al., 2009), such as R-indicators and coefficients of variation of response propensities (CV). These are all functions of response propensities.

In this chapter, we evaluate the performance of the resulting priors against non-informative priors. To do so, we focus on relevant quality metrics for nonresponse that are also used in making adaptive survey design decisions. We focus on R-indicators and CV that measure variation in response propensities across relevant strata. To validate if making early decisions is profitable, we employ the root mean squared error (RMSE) that measures the accuracy of estimated indicators. Doing so, we do not directly look at gains in survey budgets, but it can be argued that improved accuracy early on in data collection may lead to smaller samples and/or shorter fieldwork periods.

To evaluate performance, we conduct an empirical evaluation study. Based on two case studies, the 2016 Dutch EU-SILC and the 2018 Dutch Energy follow-up survey, we empirically assess the strength of the expert knowledge. The priors for the two studies have been elicited by expert staff from the data collection department of Statistics Netherlands.

The remainder of this chapter is organized as follows. In Chapter 2.2, we describe the information that we elicit from experts and formulate the power priors that include expert judgments. In Chapter 2.3, we motivate our strategy to validate performance of the power priors for the quality indicators against noninformative priors. In Chapter 2.4, we empirically evaluate the performance for the two case studies. We close with a brief

discussion in Chapter 2.5. R code is available for the expert elicitation and posterior derivation steps at GitHub.

2.2 Methodology

In this chapter, the methodology to perform a Bayesian analysis is explained and prior elicitation is prepared. The Bayesian analysis is focused on response propensities in population strata. Two overall and one partial quality indicators: the R-indicator, the coefficient of variation of response propensities, and the partial coefficient of variation, are the main targets of the analysis.

2.2.1 Notation

For the design of a survey, response probabilities are primary input parameters for making design decisions about what sample units to assign to what treatments. Response rates, nonresponse bias and costs are all a function of response probabilities, so that they play a dominant role in accuracy-cost trade-offs. In this chapter, a response probability is defined as the variation in the 0-1 response outcome across replications of a survey that results from circumstances that cannot be controlled (e.g., weather, mood of the respondent) or that a survey institute is not attempting to control (e.g., mood of the interviewer, exact timing of call or visit). Obviously, individual response probabilities are unknown and need to be replaced by estimated probabilities given a model with a selection of available covariates for the whole sample. These are termed response propensities, and they depend on the model and covariates in the model.

In this chapter, in order to simplify both derivations and prior elicitation, the population and its sample are divided into disjoint groups, termed a stratification, and denoted by $\boldsymbol{G} = \{1,2,\dots,G\}$. ρ_g denotes the response propensity for the stratum g, where $g \in \boldsymbol{G}$ and $\rho_g \in (0,1)$. Since our ultimate goal is adaptation and adaptive survey design is essentially adjustment by design, the choice of strata can be made in a similar fashion to poststratification nonresponse adjustment (Bethlehem et al., 2011).

Data collection staff select $K \geq 1$ historic data sets with respect to a new survey. $\boldsymbol{D}_g^0 = \{D_{1,g}^0, D_{2,g}^0, \dots, D_{k,g}^0, \dots, D_{K,g}^0\}$ represents the sufficient statistics in the historic data sets for

stratum g. The superscript '0' in $\boldsymbol{D}_g^0$ denotes that it refers to baseline information. The element $D_{k,g}^0$ consists of two statistics, the number of observed respondents $r_{k,g}^0$ and the number of sample units $n_{k,g}^0$ for stratum g in the kth historic survey. Hence, $D_{k,g}^0 = (n_{k,g}^0, r_{k,g}^0)$. We assume throughout that the stratum classification of sample units itself is not subject to error and is the same across historic surveys.

During the new survey data collection, additional observations come in, which again consist of numbers of sample units and numbers of respondents in each stratum. Let the observed data at wave t be $D_g^t = (n_g^t, r_g^t)$, where $t \in \boldsymbol{T} = [1,2, \dots T]$. In this chapter, a wave is a new sample that receives the same data collection strategy, i.e., we consider final response propensities and do not look at intermediate response propensities during data collection.

In the Bayesian context, the response propensities ρ_g are viewed as random variables. At the start of survey data collection, a prior distribution is derived from the K historic data sets. This prior distribution is then updated with the accumulating wave-level data from the new survey.

Each historic survey data set will be assigned a scalar parameter between 0 and 1 indicating its similarity to the survey of interest. Let the similarity parameters be denoted as $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$, where γ_k is the parameter corresponding to historic survey k. When $\gamma_k = 0$, then the historic survey is deemed completely different and of no value to the new survey. When $\gamma_k = 1$, then the historic survey is deemed completely similar and of optimal value to the new survey.

2.2.2 Prior and posterior distributions

Let us assume for now that similarity parameters γ_k are available. In Chapter 2.3, we explain how to construct them. This chapter shows how prior distributions are derived and how they are updated to posterior distributions.

We start by looking at the likelihood of the response propensities, given the data (sample size and responses). When a sample of size n_g is drawn from stratum g and r_g sample

units respond, then r_g follows a Binomial distribution, $r_g \sim Bin(n_g, \rho_g)$. For historic data set k, one obtains the likelihood

$$L\left(\rho_g \middle| D_{k,g}^0\right) \propto \rho_g^{r_{k,g}^0}\left(1-\rho_g\right)^{n_{k,g}^0 - r_{k,g}^0}, \tag{2.1}$$

and for the *K* combined data sets, the likelihood is

$$L\left(\rho_g \middle| \boldsymbol{D}_g^0\right) = \prod_{k=1}^{K} L\left(\rho_g \middle| D_{k,g}^0\right) \propto \prod_{k=1}^{K} \rho_g^{r_{k,g}^0}\left(1-\rho_g\right)^{n_{k,g}^0 - r_{k,g}^0}. \tag{2.2}$$

Next, we choose a prior distribution. A practical choice is the Beta distribution, because it is the conjugate prior for ρ_g under the binomial distribution, i.e., when the prior for ρ_g is Beta, so is the posterior.

In the complete absence of historic information, the prior distribution might be non-informative. A non-informative prior for a response propensity ρ_g is the uniform distribution on the interval $[0,1]$, i.e., all values of the propensity are considered equally likely. The uniform distribution is a special case of the beta distribution when the shape parameters are equal to 1, i.e., a $Beta(1,1)$ distribution.

With the availability of the historic survey data, (2.2) can be used to formulate the informative $Beta(a_0, b_0)$ prior. It has the following shape parameters

$$a_0 = \left(\sum_{k=1}^{K} r_{k,g}^0\right) + 1, \tag{2.3}$$

$$b_0 = \left(\sum_{k=1}^{K} n_{k,g}^0 - r_{k,g}^0\right) + 1. \tag{2.4}$$

To come to (2.3) and (2.4), we assume that all historic surveys are perfectly similar to the new survey. This is not true in general and the impact of the data set likelihoods must be altered using the similarity parameter γ_k as the power. This conforms to the approach taken by Ibrahim & Chen, (2000). The power prior of the new survey raises the Binomial likelihood of historic data *k* in (2.1) to the powers represented by these similarity parameters,

$$\pi\left(\rho_g \middle| D_{k,g}^0, \gamma_k\right) \propto \rho_g^{\gamma_k r_{k,g}^0}\left(1-\rho_g\right)^{\gamma_k\left(n_{k,g}^0 - r_{k,g}^0\right)}. \tag{2.5}$$

Here, we let the powers be equal for all strata and, therefore, do not add a subscript g. However, our method could be extended to stratum-dependent powers when historic

surveys apply different strategies to different strata. The power γ_k reduces the strength of the kth historic data. When the power equals zero, then the power prior for the particular survey is non-informative. When the power is one, then the historic surveys are seen as copies of the current survey.

For the K combined historic survey data sets one obtains

$$\pi\left(\rho_g \middle| \boldsymbol{D}_g^0, \gamma_k\right) \propto \rho_g^{\sum_k \gamma_k r_{k,g}^0} \left(1-\rho_g\right)^{\sum_k \gamma_k \left(n_{k,g}^0 - r_{k,g}^0\right)}. \tag{2.6}$$

The Beta posterior distribution parameters change along with (2.6) to

$$a_0 = \left(\sum_{k=1}^K \gamma_k r_{k,g}^0\right) + 1, \tag{2.7}$$

$$b_0 = \sum_{k=1}^K \gamma_k (n_{k,g}^0 - r_{k,,g}^0) + 1. \tag{2.8}$$

When conducting the new survey, the Beta distribution shape parameters need to be updated with incoming data. After the first wave of the new survey, at time $t = 1$, the prior $Beta(a_0, b_0)$ is updated using the observed data D_g^1 in this time period to

$$\pi\left(\rho_g \middle| D_g^1, \boldsymbol{D}_g^0, \boldsymbol{\gamma}\right) \propto \rho_g^{\sum_k \gamma_k r_{k,g}^0 + r_g^1} \left(1-\rho_g\right)^{\sum_k \gamma_k \left(n_{k,g}^0 - r_{k,g}^0\right) + \left(n_g^1 - r_g^1\right)}. \tag{2.9}$$

Eq. (2.9) is repeated on a rolling basis, i.e., the posterior from the previous wave is used as the prior in the next wave to update the inference on ρ_g. In general, after t waves, the posterior becomes

$$\pi\left(\rho_g \middle| D_g^1, \ldots, D_g^t, \boldsymbol{D}_g^0, \boldsymbol{\gamma}\right) \propto \rho_g^{\sum_k \gamma_k r_{k,g}^0 + \sum_{s=1}^t r_g^s} \left(1-\rho_g\right)^{\sum_k \gamma_k \left(n_{k,g}^0 - r_{k,g}^0\right) + \sum_{s=1}^t \left(n_g^s - r_g^s\right)}. \tag{2.10}$$

From (2.10), it can be deduced that in wave t, the posterior is a $Beta(a_t, b_t)$ distribution, with

$$a_t = \sum_k \gamma_k r_{k,g}^0 + \sum_{w=1}^t r_g^w + 1, \tag{2.11}$$

$$b_t = \sum_k \gamma_k \left(n_{k,g}^0 - r_{k,g}^0\right) + \sum_{w=1}^t (n_g^w - r_g^w) + 1. \tag{2.12}$$

Thus far, a stratification of the population was assumed, i.e., a saturated model for estimating response probabilities. As a result, the updating procedure involves simple

computations, which makes the procedure computationally very attractive. However, the number of relevant covariates may be large and a non-saturated model with no or only part of the interactions may be preferred (see Schouten et al., 2018 for a Bayesian analysis with such models). In such a setting, the power prior approach may still be used to weight the impact of a historic data set, but the Beta distribution no longer is a conjugate prior-posterior. As a consequence, updating the prior becomes more complex and can only be done using numerical methods such as MCMC.

2.2.3 Nonresponse quality metrics

In order to validate that prior information from historic surveys and expert judgments add value, posterior distributions of quality indicators are monitored in the Bayesian analysis.

In monitoring and adapting survey design the interest is in response propensity variation, where the major objective is to reduce nonresponse bias. For that reason, underrepresented sample strata may be allocated more effort, while overrepresented strata may be allocated less effort. As nonresponse bias cannot be measured directly, the natural approach is to approximate the bias via some proxy indicators. Schouten et al., (2009) proposed to use the R-indicator (R) to measure the similarity between the response and a survey sample for a fixed set of auxiliary variables. A related measure is the coefficient of variation (CV) that also includes the response rate and has a direct relation to the bias of response means. We consider both metrics and also look at decompositions of the metrics through so-called partial R-indicator and partial CV. These indicators are used in nonresponse monitoring and adaptive survey designs (Schouten & Shlomo, 2017; Moore et al., 2018; Schouten et al., 2018). R code for the computation of (partial) CVs is available at www.risq-project.eu.

Let q_g be the stratum population distribution for a certain stratum g, $g = 1,2,\dots,G$. For the sake of simplicity, we assume they are constant over all data collection waves of the new survey and $\sum_g q_g = 1$. We repeat that a wave in this chapter is a new sample that receives the same data collection strategy. The indicator of representativeness, or R-indicator, is then defined as

$$R(\hat{\rho}) = 1 - 2\sqrt{\sum_{g=1}^{G} q_g\left(\hat{\rho}_g - \hat{\bar{\rho}}\right)^2}, \tag{2.13}$$

where $\hat{\rho}_g$ is the response propensity in stratum g, and $\hat{\bar{\rho}}$ is overall response rate explicit about the level of the population response propensity. Response is fully representative when (2.13) takes the value 1 and completely non-representative at the value 0. Lower standard deviation of response propensities means more representative response.

The overall coefficient of variation of response propensities, CV, is

$$CV(\hat{\rho}) = \frac{\sqrt{\sum_{g=1}^{G} q_g(\hat{\rho}_g - \hat{\bar{\rho}})^2}}{\hat{\bar{\rho}}}. \quad (2.14)$$

The overall CV is the response propensity standard deviation divided by $\hat{\bar{\rho}}$ defined in (2.13). It is an approximation to the nonresponse bias of response means. The larger the value of (2.14), the larger the risk of nonresponse bias.

The third indicator is the category-level partial $\mathrm{CV_u}$ which tightens the connection to the ultimate goal of adaptive survey design. For the sake of brevity, we do not look at partial R-indicators here. It measures the impact of single categories, in our case the population strata, on the overall CV. It is defined as

$$CV_u(\hat{\rho}_g) = \frac{\sqrt{q_g}(\hat{\rho}_g - \hat{\bar{\rho}})}{\hat{\bar{\rho}}}. \quad (2.15)$$

(2.15) can be negative and positive, implying the specific stratum is underrepresented or overrepresented, respectively. The more negative (2.15) is, the stronger the negative impact on representativeness of the stratum and the more effort the stratum needs. Since there are as many values for (2.15) as there are strata, we focus on the strata that need effort the most, i.e., that have the largest negative values. Let $b(w)$ be the stratum in wave w that has the largest negative value of (2.15), i.e., $CV_u(\hat{\rho}_{b(w)}) \leq CV_u(\hat{\rho}_g), \forall g$. Learning early on in data collection what strata need extra effort is crucial for implementation of adaptive survey designs.

Under the Beta distribution priors for the response propensities, the priors (and posteriors) of the quality indicators have no closed forms; they are complex functions of response propensities. The priors (posteriors) can, however, be approximated by drawing a large number of samples from the priors (posteriors) of stratum response propensities. Given the

advantage of our method, the conjugate distributions allow us to efficiently and rapidly obtain numerical iterations in Chapter 2.4.4.

2.3 Expert elicitation

In this chapter, the derivation of the survey similarity scores is presented. First, a general discussion is given on data collection staff as experts. Next, survey features are proposed that facilitate scoring of the similarity of two surveys. Finally, the weighting of the survey features is discussed.

2.3.1 Data collection staff as experts

The approach taken in this chapter closely resembles Rietbergen et al., (2011). Survey data collection staff assist prior elicitation in four ways:

1. The selection of the set of historic surveys included in the analysis
2. The construction of the list of design features on which surveys are compared to the new survey
3. The choice of weights for the features to construct an overall score
4. The actual scoring of the features for the selected historic surveys

Contributions 1 and 4 are conducted for each survey, while contributions 2 and 3 are performed only once and are used for all surveys. Contributions 2 and 3 may also be used by other institutions.

In daily practice, data collection experts select historic survey data sets in order to predict response rates and costs. This selection is to some extent subjective and usually not based on a fixed set of criteria. It must be assumed, however, that all of the selected historic surveys will show at least some similarity to the new survey. In theory, one could start from scratch, select all (recent) historic surveys undertaken by the survey organization and score all these surveys. In practice, this would imply a very heavy workload on data collection experts. For this reason, in the proposed methodology it is assumed that there is a pre-selection of historic surveys.

The other three contributions concern design features. A survey design has a number of features such as modes and topics. The more similar design features of two surveys are, the more likely it is that response rates will be similar. In Chapter 2.3.2, we describe the features that were chosen in close collaboration with data collection staff. For each historic survey, the similarity on each feature must be scored. The procedure to do this is explained in Chapter 2.3.3. In Chapter 2.3.4, the importance of the features is weighed, again after consulting data collection staff.

Our approach to elicit prior distributions follows how data collection staff work in practice, but it does not exactly mimic how they work. Due to time and workload constraints, staff often perform predictions individually. Also, their predictions usually concern point estimates only and not uncertainty of these estimates. We follow daily practice by involving multiple data collection experts and identifying a number of fixed steps that they can perform. These steps take less time than daily practice and are greatly appreciated by data collection staff at Statistics Netherlands.

2.3.2 Features for deriving similarity between surveys

In collaboration with data collection staff at Statistics Netherlands, the following eight features are selected as essential when comparing survey designs:

1. *Topics/themes of the survey*: The more similar the topics of the survey to the new survey, the more similar participation rates should be;
2. *Target population*: Response rates depend on characteristics of persons and households. If target populations differ, then response rates will be different due to the different composition of characteristics. If the target population of the new survey is a subset of the historic survey, then in some cases the subset can be selected and target populations can be made the same. If the historic survey target population is itself a subset, then such harmonization is not possible;
3. *Time elapsed since last fieldwork*: Response rates change in time and the older the historic data, the more change should be expected;
4. *Unit of observation*: Two observation units are distinguished, persons and households. When the unit is different, then response rates will differ;

5. *Mode strategy (including contact and reminder)*: Survey modes and the order in which they are presented to respondents is an influential design feature in both contact and participation rates. The more similar the set of modes and the order in which they are offered, the more similar response rates will be;
6. *Incentive strategy*: The type and amount of incentive are influential for participation rates. The more similar the amount, the more similar participation rates are expected to be;
7. *Respondent effort*: Respondent burden affects participation rates, especially when the burden is salient to sample units. However, when it is not salient at the start, break-off rates are higher for longer surveys;
8. *Bureau effect relative to Statistics Netherlands*: All else being equal, response rates do vary between survey institutions. This so-called bureau effect is, therefore, included as a design feature.

One important remark is in place: Response propensities are needed at the level of a pre-specified set of population strata. In order to form the strata, the sample needs to be enriched with auxiliary variables that define the strata. Within the same survey institution, sometimes the same auxiliary variables can be linked or the same population tables can be derived. However, if some or all auxiliary variables or tables are missing, then stratum response propensities cannot be estimated. If variables are missing, then a potential solution is to choose constant response propensities for the categories of these variables in the prior distribution. In this chapter, it is assumed that historic survey data sets have no missing auxiliary variables.

The list of design features may be altered, if deemed necessary. One may, for example, add the timing and number of calls or visits to sampled persons/households. We omitted some of the obvious features here, because they are fixed in survey designs at Statistics Netherlands.

2.3.3 Scoring the survey features

Operationalization of the similarity between two surveys in terms of a [0,1] score is not straightforward for most of the eight features. The easiest to score may be feature three, *Time elapsed,* as this is quantitative. However, rather than making the operationalization

of scores as objective as possible, which is deemed very hard, it was decided to ask three experts independently and request them to reach a consensus. We constructed similarity parameters $\gamma_k \in [0,1]$ for each historic survey as follows:

a. Ask three experts to independently derive similarity scores for each of the eight features.
b. Ask the three experts to meet and reach a consensus on each of the eight features. Let $\gamma_{k,l}$ be the consensus score for survey k for feature l.
c. Construct the overall score by weighting the eight features using weights w_l, with $\sum_{l=1}^{8} w_l = 1$. The survey score becomes $\gamma_k = \sum_{l=1}^{8} w_l \gamma_{k,l}$.

We stress that it is the scoring of the similarity between historic surveys and a new survey that balances the information contained in historic data against the information contained in new data. The higher the scores, the stronger the impact of the historic data.

2.3.4 Weighting the similarity scores on the survey features

The design feature similarity scores can be weighted according to their impact on contact and/or participation rates. In this chapter, two sets of weights are evaluated. With the first set of weights all features are treated as equally important, i.e., $w_l = \frac{1}{8}$. The second set of weights was constructed by asking data collection staff.

Three experts were asked to score the importance of the features on a scale of 1 to 5: not important, mildly important, moderately important, important and very important. Table 2.1 presents the feature-level scores of each expert, the average scores over three experts and the resulting weights. The importance weight is the ratio of the feature-level average score to the overall average score.

Table 2.1 Feature importance weights.

	Expert 1	Expert 2	Expert 3	Average	Weights
Topic	2	3	1	2.0	0.08
Population	3	4	4	3.7	0.15
Time	4	4	2	3.3	0.13
Observation	3	4	4	3.7	0.15
Mode	5	5	5	5.0	0.20
Incentive	4	4	4	4.0	0.16
Response Effort	2	1	1	1.3	0.05
Bureau Effect	2	1	3	2.0	0.08

The experts agreed that mode strategy is the most important feature, followed by incentive strategy, target population, and observation units.

2.4 Two cases studies to investigate the incorporation of historic surveys and expert elicitation

The effect of the power prior is compared to a non-informative prior in a Bayesian framework using two case studies, the Dutch Energy and the Dutch EU-SILC. In both cases, the survey design was new and no direct historic information was available. The interest is in the added benefit from the inclusion of historic data and expert elicitation.

First, an RMSE evaluation criterion is introduced to evaluate the gains from a power prior in Chapter 2.4.1. Second, the Energy and SILC data are described briefly in Chapter 2.4.2. Next, in Chapter 2.4.3, the scores over similarity criteria are presented and powers for the historic data are derived. Finally, the posterior credible regions of quality indicators, R-indicator and CV, are illustrated as a function of the data collection wave in Chapter 2.4.4.

2.4.1 The evaluation criterion

In this chapter, we explain how we assess and compare the performance of non-informative and informative priors. Our strategy consists of evaluating the prediction of the three metrics, R-indicator, CV and CV_u, defined in (2.13) to (2.15). To evaluate prediction accuracy, we consider the root mean square error (RMSE) of the predicted

indicators against their realizations per data collection wave. The criterion RMSE is defined for the overall metrics as

$$RMSE\left(\theta;\pi_{0,T}\right)=\frac{1}{T}\sum_{t=1}^{T}\sqrt{(\hat{\theta}_t-E_{\pi_{0,t-1}}(\theta))^2+var_{\pi_{0,t-1}}(\theta)}, \qquad (2.16)$$

where θ is the parameter of interest, i.e., R-indicator or CV, $\hat{\theta}_t$ is the realized value of this parameter in wave t, $\pi_{0,0}$ is the prior based on historic survey data, and $\pi_{0,t-1}$ is the posterior based on historic data and new data up to wave $t-1$. As noted, T is the present wave where the latest sample is released and (2.16) is a rolling average of the RMSE until that wave. Smaller RMSE implies that the accuracy has improved and decisions about allocation of effort and budget can be made at an earlier stage in data collection.

For the evaluation of the strata that have the smallest $\mathrm{CV_u}$, we have to make an intermediate step. Let B be the number of bootstrap samples and $A_{g,B}^{t}$ be the number of these samples from the posterior distribution based on data up to wave t where stratum g has the smallest $\mathrm{CV_u}$. Let $p_{g,B}^{t}=\frac{A_{g,B}^{t}}{B}$. Finally, let $1_b(g)$ be the binary indicator that equals one when $g=b$. We use the following RMSE criterion to assess the predictions of the strata that need extra effort,

$$RMSE\left(p_{g,B}\right)=\frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(p_{g,B}^{t-1}-1_{b(t)}(g)\right)^2+\frac{1}{G}\sum_{g=1}^{G}p_{g,B}^{t-1}(1-p_{g,B}^{t-1})}\,. \qquad (2.17)$$

For observed data in wave t we can derive $b(t)$ is the stratum that is most underrepresented. The RMSE is small when expected values based on all data up to the previous wave are close to the realized values and the posterior based on these data has a small variance.

The posterior terms in (2.16) and (2.17), expectation and variance, are estimated by empirically drawing 10,000 samples from the beta posterior of stratum response propensities using (2.11) and (2.12), and then computing the quality indicators for each iteration through the formula of quality indicators under formulas (2.13) to (2.15).

2.4.2 The two case studies

For illustration, we apply our proposed method to two surveys, the 2018 Energy Survey (EN18) and 2016 EU-SILC Survey (SILC16).

Energy Survey: The Energy Survey is conducted every six years and contains detailed questions on households' use of electricity, gas, and water facilities and energy savings measures that they have implemented in their houses. In the 2018 edition, the survey sampled from respondents to the Dutch Housing Survey 2018 (HS18). The HS18 is a more general survey on housing conditions that is fielded bi-annually. EN18 is an extension to HS18, but respondents to HS18 are not pre-notified of the EN18 and EN18 sample units get a separate invitation letter. The EN18 sampling design was a stratified simple random sample where strata were formed based on dwelling type, dwelling age and household income. These same strata are also used in the response propensity estimation, leading to 30 strata. The sample size of EN18 was 75,918 and the sample size of HS18 was 90,121.

In 2018, Statistics Netherlands conducted the Energy Survey for the first time. In the two previous rounds in 2006 (EN06) and 2012 (EN12), the survey was conducted by another institution. Statistics Netherlands decided to use the same survey modes as in HS18: web, telephone and face-to-face. HS18 had a mixed-mode survey design, where web was offered first and web non-respondents were assigned to telephone, when a phone number was available, and otherwise to face-to-face. The HS18 mode of response was used in EN18 for web and telephone. When a sampled HS18 respondent had used web, then EN18 sent a web invitation. When a sampled HS18 respondent was interviewed over the phone, then EN18 made phone calls too. The exception was face-to-face. Since this is an expensive mode, sampled HS18 respondents that were interviewed face-to-face, were first sent a web invitation letter. Only if they did not respond, a face-to-face interviewer was sent. The EN18 design had never been implemented before at Statistics Netherlands and was chosen because of cost reasons.

Given that EN18 was new to Statistics Netherlands and was implemented as a follow-up survey to HS18 in an unprecedented design, the EN18 response propensities were deemed very unpredictable. This is the reason, the EN18 is selected as a case study.

As historic survey data, EN06, EN12, and HS18 were available. EN06 and EN12 had a similar size and stratification. In addition, given the follow-up nature of EN18, also another survey was selected, the 2016 Dutch Survey on Care (SC16). This survey sampled from respondents to the Dutch Health Survey. The SC16 had a sample size of 10,414.

SILC: The EU Statistics on Income and Living Conditions survey (EU-SILC) is a rotating panel survey with one new panel group each year. The total duration of the survey is four years with one survey per year. EU-SILC is a survey that is mandatory within the European Statistical System (ESS) and conducted in 31 ESS countries. Topics are various forms of income and assets, housing conditions and health conditions. Derived statistics concern poverty rates and the ability of the household to make ends meet. The survey went through a major redesign around 2005 in which the panel design was introduced. The Dutch EU-SILC has been running in a more or less similar design since its introduction up to 2016. Respondents to the fifth wave of the Dutch Labor Force Survey (LFS) were invited to participate in EU-SILC. The motivations for this design were cost savings, overlap between LFS and EU-SILC statistics and the availability of rich administrative data on income. EU-SILC used only the telephone mode up to 2015. In 2015, it was decided that EU-SILC is to be based on new, separate samples and to be disconnected from the LFS. A sequential mixed-mode design with web followed by telephone was introduced. As response rates were uncertain, the sample was randomized into two parts. One part received no incentive, and one part received a conditional incentive of 10 Euro. In this case study, both samples are considered and scored separately.

The strata of interest for EU-SILC are 20 groups based on a mix of household size and income deciles. The income deciles are derived from administrative data in the previous year.

Two historic surveys were selected by data collection staff: the 2016 Dutch Labor Force Survey (LFS16) and the 2015 Dutch Household Budget Survey (HBS15). Both are conducted by Statistics Netherlands. LFS16 employed the same mixed-mode design, but added face-to-face to web non-respondents without a known phone number. HBS15 used the same design as SILC16 but is a diary survey. LFS16 was selected because the topics, survey modes, and unit of observation (the household) are very similar. The HBS15 was

selected because the topics and modes were similar and because it was the only survey that used incentives in a household setting. The overall sample sizes were 7,954 (7,955) for SILC16 without (with) incentive, 24,882 for LFS16, and 8,182 for HBS15.

The SILC16 survey was selected as a case study in this chapter as it was a relatively predictable design. The survey had been conducted by Statistics Netherlands for many years and the survey design resembled that of other surveys.

The number of waves varies over the two studies. Waves are chosen such that they correspond to time points where data collection may be adapted. For the EN18 case study 15 waves are chosen corresponding to different sample portions fielded throughout the period February 2018 to September 2018. For the SILC16 case study three waves are chosen, corresponding to the three data collection months: April, May, and June 2016.

For all historic datasets, Statistics Netherlands data collection department provided sample and response sizes at stratum level. For each case study, three data collection staff members scored the surveys on the eight criteria.

2.4.3 Similarity scores for the two case studies

This chapter presents the scores that the data collection staff members assigned to the historic surveys. For each historic survey and survey design feature, only the consensus score of the similarity over three data collection staff members is shown.

Table 2.2 gives the feature-level similarity scores on the EN18 case study for the EN06, EN12, SC16 and HS18, and the combined scores by two types of weights. The EN06 and EN12 have perfect scores for the criteria related to topic, target population, observation unit and respondent burden, but have low scores on the other criteria, especially time elapsed. The other two surveys score relatively well on these other criteria. The feature scores are combined in two ways: one is by weighting all features equally and one is by using the weights from expert staff in Table 2.1. For each historic survey, the expert-based weight yields lower score than the equal weight.

Table 2.2 Similarity scores per survey feature for the four surveys in the EN18 case study.

	EN06	EN12	SC16	HS18
Topic	1.0	1.0	0.0	0.7
Population	1.0	1.0	0.4	0.9
Time	0.1	0.2	0.4	0.7
Observation	1.0	1.0	1.0	1.0
Mode	0.1	0.1	0.5	0.0
Incentive	0.0	0.0	0.3	0.3
Response Effort	1.0	1.0	0.1	0.1
Bureau Effect	0.0	0.0	0.4	0.5
Equal weights	0.525	0.538	0.388	0.525
Expert weights	0.463	0.476	0.447	0.525

Table 2.3 Similarity scores per feature for the two surveys in the SILC16 case study. The incentive strategy criterion is scored for the SILC16 without incentive and with incentive.

	LFS16	HBS15
Topic	0.3	0.3
Population	0.7	0.6
Time	1.0	0.6
Observation	0.0	0.0
Mode	0.6	0.1
Incentive	0.2 (without)	0.5 (without)
	0.0 (with)	1.0 (with)
Response Effort	0.3	0.0
Bureau Effect	0.7	0.2
Equal weights	0.475 (without)	0.288 (without)
	0.450 (with)	0.350 (with)
Expert weights	0.482 (without)	0.308 (without)
	0.450 (with)	0.388 (with)

Table 2.3 shows the similarity scores in the SILC16 case study for the two historic surveys LFS16 and HBS15. Recall from the Chapter 2.4.2, the sample was randomized into two parts. For the incentive strategy criterion two scores are given, one for SILC16 without incentive and one for SILC16 with incentive. The only perfect score is for LFS16 as it was conducted very close in time to SILC16. When topic is considered, observation unit and

respondent effort criteria both historic surveys score weakly. The expert weighting has a strong impact on the similarity scores.

2.4.4 Posterior distributions for the aggregate quality indicators

In this chapter, our primary objective is to evaluate whether our method outperforms the non-informative prior in predicting quality indicators, and to look into whether the performance of our method would depend on the approach to pool the historic-specific criteria. They are illustrated by the RMSE evaluation criterion, applied to both case studies (EN18 and SILC16), with credible regions for overall indicators. Overall R-indicator and CV in (2.13) and (2.14) are a function of wave for either informative or non-informative priors. The survey, SILC16, is investigated under two scenarios, with incentive and without incentive. Recall from Chapter 2.4.2 that the sample was randomized into two parts.

For the EN18, Figure 2.1 displays the overall indicators predicted by our method using expert elicitation and relevant historic surveys in contrast with the noninformative method as well as the realized indicators (target predictions). The horizontal line is the indicator realization of the population. For ease of explanation, our proposed priors are called expert priors and the non-informative prior is called the standard prior.

For each wave, 95% credible regions summarize the posterior expectations and their uncertainty for either expert priors or the standard. In early data collection waves, either expert prior has a small uncertainty on predictions versus the standard prior. For example, in wave 1, expert priors predict R-indicator with 1% uncertainty while for the standard prior it is 3%, and uncertainty levels are 1.3% and 6% for expert priors and standard prior in predicting overall CV. The standard method can predict R-indicator with increasing precision when more data collection waves are released as the width of the credible intervals decreases (by at most 2%), however, there is no significant reduction in the precision of posterior predictions for either expert prior, only a 0.3% decline. Moreover, none of the priors can completely adapt to the upcoming data, while the standard method pushes its prediction toward the target R-indicators slightly better than expert priors in early waves. The difference of posterior expected R-indicators between priors is 5% at

most in wave 1 and declines fast to 1% in wave 4. Later on in predicting R-indicators, the standard prediction can catch up fast with expert prior predictions.

We propose overall CV as an alternative indicator to evaluate our method, a better indicator providing a link to actual non-response bias (Schouten et al., 2009). Either prior pushes CV predictions to target CVs and attempt to predict with less uncertainty against the standard method in early waves, 1 through 5. As data are accumulated, the prediction uncertainty shows a noticeable decline for the standard prior but expert priors remain unchanged. In the late data collection from wave 8 onwards, the interquartile ranges increasingly overlap between expert and standard priors, signifying that the standard method can compete with expert priors when predicting overall CV.

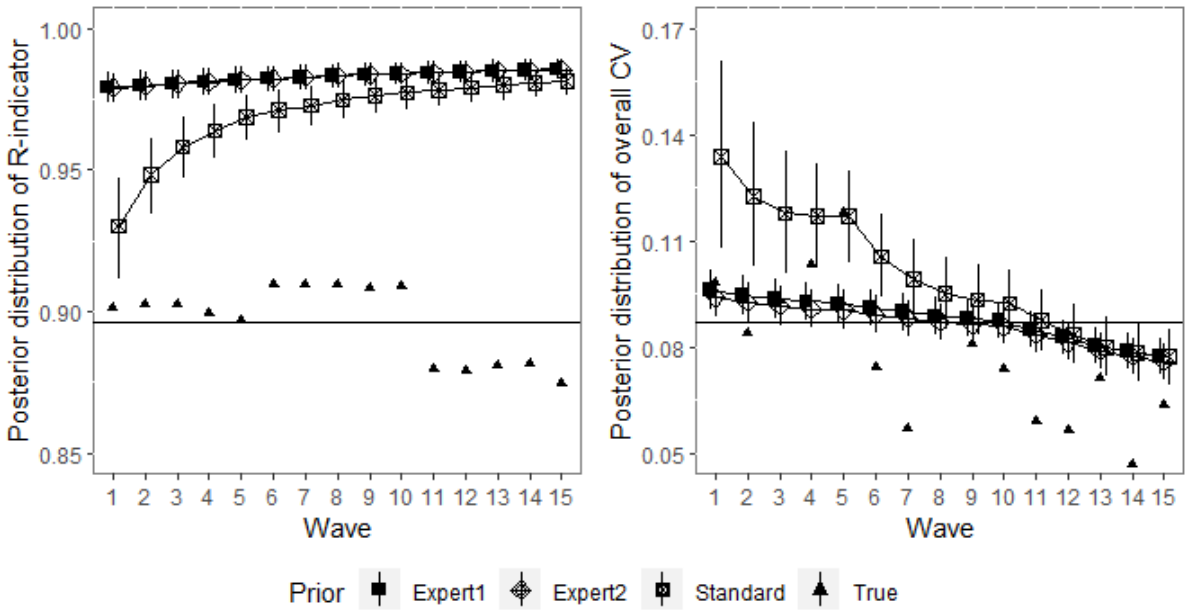

Figure 2.1 The 95% credible regions of the posterior distribution of R-indicator and overall CV in the EN18 case study. In each wave, the posterior distributions are constructed by updating three different priors: the power priors (Expert1 with equal weights and Expert2 with expert weights), and the non-informative prior (Standard). The observed values are presented as well at wave (True) or overall (horizontal line) level.

Important to note that the credible regions of either R-indicator or CV from experts' priors are not wide enough to the observed values ("True"). This means the observed values are extremely unlikely, according to experts, and they have a low probability of occurrence.

As we see in Table 2.4, the power priors (Expert 1 and Expert 2) predict a lower risk of nonresponse bias than the non-informative prior (Standard), where the expectation of R-indicator is closer to 1 and overall CV is closer to 0. Additionally, the power priors have little uncertainty about the expectation since the credible interval is much narrower than the non-informative prior. This is obvious because historic surveys provide information on the likelihood of response propensities.

Table 2.4 The expectations and 95% credible regions (in brackets) of R-indicator and overall CV in Wave 0 from three different prior distributions in EN18 case study.

	R-indicator	Overall CV
Expert 1 with equal weights	0.978 ([0.972,0.984])	0.092 ([0.098,0.104])
Expert 2 with expert weights	0.978 ([0.972,0.984])	0.090 ([0.096,0.101])
Standard	0.426 ([0.298,0.557])	0.341 ([0.562,0.860])

Understanding to what extent expert priors add value to predict target indicators is revealed by RMSE in Table 2.5.

In early waves, RMSE of predicted overall R-indicator and CV by expert priors are closer to 0, and additionally they are smaller than the standard prior. This is not surprising because the non-informative prior is entirely vague from the onset in the sense that it provides little information on the shape of unknown response propensities, and thus it causes large variance of predictions. Either expert prior continues to be superior to the standard prior relative to measure RMSE of overall R-indicator and CV, but overall CVs (Columns Expert and Standard) become competitive in Wave 12 to Wave 15. With more samples released over waves, the difference in RMSE is increasingly small, implying that the standard method better predicts overall R-indicator and CV, and more importantly the effect of expert priors diminishes. This is fairly straightforward that when more data come in, the posteriors for non-informative and informative prior will converge to each other at some point, because the likelihood dominates the posteriors instead. The results for the overall CV show weak evidence to support our argument whether our method is superior to the standard prior to a new survey. In contrast, R-indicator reveals expert priors making prediction better than the standard prior. The result is mixed because CV is aggregated over all strata. The effect within stratum has a different impact on response behavior and even propensity variation, where some are underrepresented, and others are

overrepresented. This can be measured by the unconditional partial CV in (2.15). A consistent improvement in RMSE of partial CV at wave level proves that our expert priors outperform the standard prior. There has a minor difference between expert prior with equal weights and expert prior with varying weights, indicating our method is insensitive to the pooled method to combine historic-level criteria.

Table 2.5 RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for overall R-indicator, overall CV, and partial CV for the EN18 case study.

Wave	*R-indicator*			*Overall CV*			*Partial CV*		
	Expert		*Standard*	*Expert*		*Standard*	*Expert*		*Standard*
	Equal	Varying		Equal	Varying		Equal	Varying	
1	0.033	0.031	0.461	0.014	0.014	0.525	0.167	0.153	0.170
2	0.031	0.029	0.242	0.027	0.027	0.082	0.142	0.138	0.167
3	0.029	0.027	0.166	0.017	0.017	0.060	0.141	0.138	0.163
4	0.026	0.025	0.126	0.006	0.006	0.044	0.137	0.136	0.160
5	0.024	0.022	0.102	0.010	0.010	0.029	0.130	0.128	0.159
6	0.024	0.022	0.088	0.035	0.035	0.072	0.141	0.140	0.157
7	0.023	0.022	0.076	0.051	0.051	0.075	0.142	0.146	0.156
8	0.023	0.021	0.067	0.018	0.018	0.035	0.138	0.142	0.156
9	0.022	0.020	0.060	0.025	0.025	0.038	0.129	0.134	0.155
10	0.021	0.019	0.055	0.031	0.031	0.042	0.121	0.125	0.153
11	0.021	0.019	0.054	0.045	0.045	0.055	0.127	0.131	0.160
12	0.021	0.020	0.053	0.045	0.045	0.051	0.132	0.135	0.166
13	0.021	0.020	0.052	0.027	0.027	0.031	0.136	0.139	0.172
14	0.022	0.021	0.052	0.048	0.048	0.051	0.138	0.142	0.176
15	0.023	0.022	0.052	0.029	0.029	0.032	0.141	0.145	0.179

Figure 2.2 shows the comparison of expert priors with two weights to the standard prior in the SILC16 under two scenarios (with incentive and without incentive). The results of R-indicators show that predictions from either expert prior is superior to the standard prior in wave 1 as the standard variance is 5 times larger than the expert prior variance, although the standard posterior mean reaches slightly the observation than either expert prior. This advantage of either expert prior continues in wave 2 but the standard competes with them as wave 3 showed. Either expert prior has obvious benefit versus the standard prior on the measure of overall CV. However, this advantage gradually declines and the standard method catches up as accumulating data.

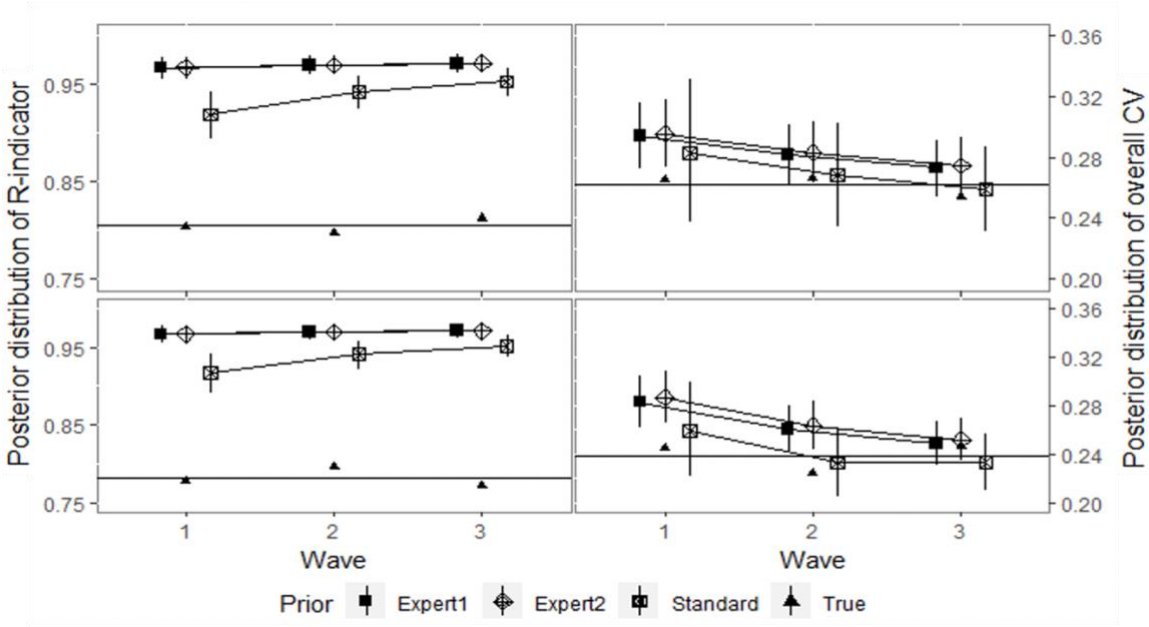

Figure 2.2 The 95% credible regions of the posterior distribution of R-indicator and CV in the SILC16 case study. In each wave, the posterior distributions are constructed by updating three different priors: the power priors (Expert 1 with equal weights and Expert2 with expert weights), and the non-informative prior (Standard). The observed values are presented as well at wave (True) or overall (horizontal line) level. The top/bottom panel corresponds to without/with incentive.

In Wave 0 (Table 2.6), the power priors behave much better than the non-informative prior in predicting the overall variation in response propensities and in the uncertainty measurement, regardless of whether there has incentive in the SILC16. R-indicators of either power prior take the values closer to one, and overall CVs take smaller values than the non-informative prior.

As Tables 2.7 and 2.8 show, the RMSE of the predicted R-indicator from expert priors is smaller in first two waves, and in wave 3, the standard is competitive to or slightly weaker than expert priors. When predicting overall CV, only in wave 1 either expert prior shows its superior performance, because the advantage of expert priors is overwhelmed by the size of upcoming data in one wave. However, the benefit to use our method is strongly supported by the RMSE of R-indicator and partial CV under either scenario. RMSE from either expert prior is consistently closer to 0 and better than the standard prior.

Table 2.6 The expectations and 95% credible regions (in brackets) of R-indicator and overall CV in Wave 0 from three different prior distributions in the SILC16 case study.

	R-indicator		Overall CV	
	No incentive	With incentive	No incentive	With incentive
Expert 1 with equal weights	0.964 ([0.950,0.975])	0.963 ([0.950,0.0.975)	0.318 ([0.293,0.344])	0.326 ([0.300,0.352])
Expert 2 with expert weights	0.964 ([0.951,0.975])	0.964 ([0.951,0.976])	0.320 ([0.294,0.344])	0.329 ([0.304,0.355])
Standard	0.427 ([0.310,0.555])	0.426 ([0.309,0.555])	0.601 ([0.422,0.909])	0.603 ([0.424,0.922])

Table 2.7 RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for three indicators in the SILC16 without incentive.

Wave	R-indicator			Overall CV			Partial CV		
	Expert		*Standard*	*Expert*		*Standard*	*Expert*		*Standard*
	Equal	Varying		Equal	Varying		Equal	Varying	
1	0.161	0.161	0.382	0.043	0.042	0.359	0.091	0.086	0.214
2	0.166	0.165	0.252	0.033	0.032	0.194	0.067	0.064	0.145
3	0.163	0.163	0.212	0.030	0.030	0.137	0.056	0.053	0.120

The results for the EN18 and SILC16 show that data collection staff provide accurate estimates on the variations of response. The empirical studies also advocate that prior elicitation from the staff experts working on surveys is of significant value to predict response propensities when a new survey has never been conducted before or when a survey is redesigned. The additional value of expert priors can be proved when predicting

quality indicators and monitoring data collection: overall R-indicator and unconditional partial CV.

Table 2.8 RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for three indicators in the SILC16 with incentive.

Wave	R-indicator			Overall CV			Partial CV		
	Expert		*Standard*	*Expert*		*Standard*	*Expert*		*Standard*
	Equal	Varying		Equal	Varying		Equal	Varying	
1	0.187	0.186	0.356	0.074	0.075	0.378	0.113	0.114	0.214
2	0.179	0.179	0.239	0.064	0.065	0.209	0.126	0.128	0.177
3	0.166	0.185	0.216	0.047	0.048	0.146	0.133	0.134	0.166

We also apply our method to predict the level of response propensities, results shown in Appendix B. The experts are uncertain about the level of response propensities. In this chapter, our primary concern is with the variation of response propensities more than with the level as adaptation is based on underrepresentation of certain strata.

2.5 Discussion

Our two most important goals were to set up a structured expert prior elicitation procedure in the context of surveys and the evaluation of the utility of this procedure relative to non-informative priors. In other words, can data collection staff knowledge be transformed and be utilized, so that survey design, and in particular adaptive survey design, can profit? With this procedure, we explicitly focus on data collection staff, both as informers and as users.

To include expert knowledge, we set up a procedure that takes a power prior as a key ingredient. The powers are derived by scoring a number of historic surveys on their similarity to the new survey. Scores are based on the identification of a number of relevant survey design features, which are weighted based again on expert opinions. In our case studies, we invited three data collection experts and averaged their scores. The power prior is updated using incoming survey data during data collection. The performance is evaluated based on three quality indicators: overall R-indicator and overall (partial) coefficient of variation of response propensities. The prior and posterior of the variation in

response propensities and coefficient of variation of the response propensities have no closed form. However, since Beta distributions are conjugate for the response propensities, it is relatively straightforward to construct the posteriors empirically. Consequently, the first three objectives of the chapter are achieved to monitor and adapt the survey data collection for the new survey to the subsequent phases. Our prior elicitation procedure has been set up in collaboration with data collection staff. We paid a lot of attention to making the procedure transparent, but also manageable. We view bridging the gap between data collection and methodology as the most important achievement of this study.

Our other objective was to assess the benefit of incorporation of the historic prior information into the new survey against the settings without prior knowledge. In the evaluation, a fully non-informative prior implies that no historic surveys and no expert knowledge can be used to specify a prior for a new survey. To achieve this goal, the root mean square error (RMSE) of the posterior of quality indicators is evaluated. The evaluation was made based on two case studies, the EN18 and SILC16, using the observed indicators and the posterior prediction with a series of samples released in time. Both case studies show that the approach to weight the survey features have no influence on the comparison of a Bayesian analysis against a non-Bayesian analysis, because the RMSE is only slightly distinct between equal weights and expert weights. The evaluation study shows either power prior can be vastly superior to a non-informative prior on predicting the variation in response propensities as well as the coefficient of variation of them throughout the course of new survey data collection. The advantage holds to predict CV but in late waves the non-informative prior can compete with expert priors. So far, we conclude that the power prior clearly has added value, but its prediction performance is closely related with the choice of historic surveys, the selected criteria, and the prescribed feature-level weights elicited from the staffs. Therefore, we propose to carefully use a power prior for predicting response propensities and related indicators when a survey is brand new, which ought necessarily to be compared with non-informative predictions.

Our study has a number of simplifications which are the subject of future research. First, besides the response propensities, it is crucial to model cost as another important design parameter playing a decisive role in a survey design. It is a challenge to realize the cost

model for each stratum, because the stratum costs depend on the response propensity and the mode strategy. Furthermore, it is hard to isolate the actual realized cost of an individual survey and a single stratum in that survey. Nonetheless, with some simplifications it is possible to model stratum costs as a function of stratum numbers of calls and visits, stratum interview durations, and stratum contact, refusal and participation propensities. See Schouten et al., (2018). Expert elicitations then amount to prediction of number of visits and calls and interview durations, since propensities are already part of the current approach. This may require additional or different survey design features than those selected in this chapter. It is an important topic for further refinement and extension. Second, although we assume that the Bayesian analysis is independent of time, time change may play a crucial role. The necessary extension is to incorporate the effect of time on stratum response propensities. The model proposed in this chapter must be expanded such that response propensities may change gradually over time. Third, measurement error is ignored, while in mixed-mode surveys, such as the case studies in this chapter, it can, and most likely does, play an influential role. Fourth, we suppose a fully saturated model, i.e., a full stratification in disjoint groups, when modelling nonresponse. While this may ultimately be easier in adaptive survey design, our methodology should be extended to parsimonious models that omit some or all interactions.

The simplifications will form the basis for extensions of the proposed procedure. Data collection staff usually have a good view on costs and time change in response propensities. However, measurement error, typically, is not analyzed by data collection staff. For this purpose, we need to find other experts.

Another follow-up research question is the impact of the choice of experts. In this study, we could not assess the impact of the choice of experts as part of the expert elicitation was performed through joint meetings in which they reached consensus. It should, however, be evaluated in future studies. Plus, the evaluation of experts' elicitation on estimating costs is an important topic for the future.

In the proposed method, we regard the timeliness of historic data sets as a similarity criterion. Assessment of the criterion obviously depends on the time-length of the historic data, e.g., the last quarter or the last year. The longer the time length the harder it is to

provide a single value as the data contain both very recent and relatively old data. Extra uncertainty is introduced by assuming a constant timeliness. Therefore, the issue involving how far a researcher should go back for picking up historic data sets should be addressed as a future topic.

Chapter 3

Modelling Time Change in Survey Response Rates: A Bayesian Approach with an application to the Dutch Health Survey

This chapter has been accepted by Survey Methodology Journal (SMJ) and will be published on the 2023 issue: Wu, S., Boonstra, H.J., Moerbeek, M., & Schouten, B. (2023). Modelling Time Change in Survey Response Rates: A Bayesian Approach with an application to the Dutch Health Survey.

Author contributions: SW, HJB, and BS contributed to concept and design of the study. Statistics Netherlands organized the data collection and database. SW performed the statistical analyses. SW wrote the first draft of the manuscript. HJB, BS, and MM critically reviewed the paper. HJB and BS wrote some sections of the paper. All authors contributed to manuscript revision, read, and approved the submitted version.

Abstract

Precise and unbiased estimates of response propensities (RPs) play a decisive role in the monitoring, analysis, and adaptation of data collection. In a fixed survey climate, those parameters are stable and their estimates ultimately converge when sufficient historic data is collected. In survey practice, however, response rates gradually vary in time. Understanding time-dependent variation in predicting response rates is key when adapting survey design. This chapter illuminates time-dependent variation in response rates through multi-level time-series models. Reliable predictions can be generated by learning from historic time series and updating with new data in a Bayesian framework. As an illustrative case study, we focus on Web response rates in the Dutch Health Survey from 2014 to 2019.

Keywords: Time series analysis; multilevel model; Bayesian analysis; response propensity predictions

3.1 Introduction

Over the last two decades, responsive and adaptive design (Chun et al., 2018) have attracted considerable interest in assembling survey design features ahead of or during data collection, with an ultimate goal of survey cost-quality optimization by a search for efficient resource allocation. The emergence of Web surveys, the availability of process data, and the increase in survey costs have driven research regarding the monitoring (Kreuter, 2013) and adaptation (Schouten et al., 2017) of data collection. However, a thorough understanding of how design features and time change affect important parameters in response and cost models is imperative to apply adaptation. For example, a critical factor is the likelihood of a participant to engage in a survey, i.e., their response propensity, which can be sensitive to factors both dependent on and independent from the nature of the survey itself. Additionally, the cost of the survey is a complex calculation that covers everything from planning the survey, to performing it and the data workup afterwards, and it can directly impact the type of survey performed, which can in turn influence response propensities (RPs). For this reason, the development of such parameter measurements is necessary before the data collection operation begins.

The last decade has seen a renewed importance in the predictability of RP for responsive and adaptive design. In survey methodology, using propensity scores (Rosenbaum & Rubin, 1983) is the common way to tailor differential features to sampled cases for desired cost- or quality-related goals. In a changing data collection climate, the performance and structure of a survey design hinge heavily on propensity models that may lead to inefficient decisions. For instance, by relying only on process or response data in the early stages of a responsive survey, the estimates of RP may produce biased estimates of the final RP by the end of the data collection (Wagner & Hubbard, 2014). Also, the uncertainty of RP estimates should be incorporated into propensity models in order to avoid suboptimal designs (Burger et al., 2017).

Accurate estimates of RP are thus the crux of survey operations. For this reason, survey researchers apply historic data to estimate the coefficients of a propensity model, and then use those estimated coefficients for the upcoming rounds of a survey. Bayesian analysis (Gelman et al., 2013) is a natural approach to utilize both historic and new data for improving predictions. Prior beliefs generated from historic data are evolved into posteriors, which serve as the priors for the subsequent analysis as the upcoming data accumulates. Schouten et al., (2018) were the first to apply a general Bayesian method to analyze RP and cost in the Dutch Health Survey. They discuss that misspecification of the priors may weaken prediction performance. As a result, prior elicitation becomes an influential step. The incorporation of expert beliefs is a prerequisite for such prior elicitation. This has a long history in biometric and medical literature, but the application is in its infancy in the context of surveys. Recent examples have been West et al., (2021), who reviewed empirical evidence for survey propensity prediction, Coffey et al., (2020), who consulted data collection managers about the estimated coefficients, and Wu et al., (2022), who used data collection staff as experts for relevant historic leverage under criteria for a new or redesigned survey.

So far, the approaches assume RPs are stable in a relatively short period. In a fixed survey climate, these parameters remain stable and their estimates ultimately converge with the accumulation of historic data. In survey practice however, those parameters change gradually over time, which means that predictions may not converge. For example, seasonal variation and downward trends in response rates can be observed. Thus, the

benefit of prior elicitation could potentially be undone when ignoring time change. Recent articles by Mushkudiani & Schouten, (2019) and Fang et al., (2021) describe what time-dependent factors significantly affect the parameter estimation accuracy, but the impact on prediction accuracy is still unknown, which is the topic of this chapter.

This chapter provides new insights into flexible time series models in a structural fashion for RPs in adaptive survey designs. We attempt to interpret time change in survey RPs that correlate significantly with nonresponse biases when nonresponse is subject to time change. Our approach applies to repeated cross-sectional surveys with multiple data collection phases.

Our main objective is to make reliable predictions for RP across relevant population strata. Note that population strata in which response propensities can differ herein can be subpopulations of interest either. They are called strata throughout, even though they do not necessarily coincide with sampling strata. Our main objective is additionally to examine the prediction performance so that we can measure how time alters the RP. This general question can be reduced to four concrete aspects:

1) What time-series components contribute most to variation in RPs?
2) What level of RP prediction accuracy can be achieved for the next upcoming time period?
3) How does prediction accuracy vary over population strata?
4) How does prediction accuracy depend on the length of the historic survey time series?

The abundant knowledge of historic survey time series allows us to learn the effects of time-related factors on RP. We consider two levels, time and strata, which make up multiple components involved in a time-series model. The components describe variation over time or strata or over both, and they can be analyzed individually as well as collectively. Several survey methodology studies employ such a multilevel time-series model approach in official statistics; e.g., Boonstra & van den Brakel, (2019 & 2022) estimate monthly and quarterly regional unemployment rates using a Bayesian hierarchical model to borrow strength over time, space, and from auxiliary series. Such usage originates from the small area estimation literature (Rao & Molina, 2015)

In this chapter, we use the Dutch Health Survey (GEZO) to evaluate our approach regarding the four research questions above. This survey has had a stable design since 2011 and we focus on the time series from 2014 to 2019.

To optimize predictions, we compare a collection of model compositions by different information criteria to obtain a balance between goodness of fit and model complexity. To evaluate the "optimal" model, we will assess its predictive performance and accuracy by its ability to correctly capture the magnitude and variation of RPs. Important to note, we focus on the achievement of reliable inference over time, rather than on minimizing nonresponse error, which is one of the objectives adaptive survey designs pursue.

This chapter first introduces several time-related factors of great relevance to variation with a hypothetical illustrative example in Chapter 3.2, then goes on to the differential model compositions in the general form of the Bayesian multilevel time-series model in Chapter 3.3. Chapter 3.4 optimizes the model performance based on an empirical analysis of GEZO. We discuss our findings and end up with the brief overview of future work in Chapter 3.5.

3.2 Time Series Components of Survey Response Rates

It is well-known that response propensity (RP) changes gradually in time. Failing to incorporate this temporal dependence in design decisions can lead to ineffective survey designs. In this chapter, we use an illustrative example for introducing some time-related factors linked with considerable variation in RP.

We focus on population subgroups, or strata, as indexed by $g \in \{1, \dots, G\}$, since we aim ultimately to let the proposed models inform adaptive design decisions. The strata are formed with the help of auxiliary variables that are linked to the sample and are, thus, available for all sample units. A time-series RP $\rho_{g,t}$ in stratum g and time t is a sequence of random variables. Assuming the availability of historic survey data up to time t, we are interested in measuring variation caused by time-related factors for the most up-to-date RP predictions. To achieve this goal, we first propose potential time-dependent factors. As an illustrative example of a time series divided into the following components: trend,

seasonality, and so on, Figure 3.1 compares the overall response rate to the following time-dependent variation:

- *Trend*. The trend describes the long-term movement of the observed time series without the seasonal variation. It shows the general tendency of the population-level response rates over years, which can be linear or nonlinear. Hence, the growth or the fall of the long-term forecasts can be studied by this trend. As seen in Figure 3.1, the long-term direction does not behave like a cyclic fluctuation. Of greater importance for model development is to separate the total trend into a global trend shared by all strata, and local, i.e., stratum-specific trends.
- *Seasonality*. Seasonal variation in the overall responses describe periodic movements that recur regularly and do not influence annual averages. The periodic fluctuations possess a systematic and calendar-related nature that can be predicted and attributed to a fixed season per year. For instance, the response rate would be higher in the early period of the year while relatively lower in the middle year or in December.
- *Residual fluctuation*. The residual variation is the part of the signal obtained after excluding all of the above components. This part is usually modeled as white noise, i.e., as independent normally distributed fluctuations.

In addition, there may also be some additional time-dependent components not revealed in Figure 3.1 that nevertheless have a strong impact on the reliability of stratum RP predictions. For this reason, we also consider extra stratum-related time-dependent components:

- *Stratum*. Different subgroups have different response behaviors, such as, young subgroups are more likely to respond to the web survey than old subgroups due to the latter having potentially less access to or unfamiliarity with the internet. This variation in subgroups leads to a differential stratum-level trend and could potentially also contribute to differential seasonal movement.
- *Sampling variation*. Sampling variation complicates the estimation of RPs, especially for strata with small sample sizes. The sampling variation is taken into account by adopting a binomial likelihood.

- *Unexpected events*. Unexpected events, such as web servers being down temporarily, will appear as outliers and may violate the existing pattern. They correspond to irregular movements during short periods. The resulting variation does not follow a particular model, is unpredictable, and can become influential in predicting future RPs.
- *Intervention*. Design change, such as introducing incentive, is used widely to conduct intervention on purpose, in order to stimulate responses for an improvement in data collection quality, and even to efficiently allocate limited resources for a reduction in survey cost. Intervention has a permanent impact on response propensities. The influence can be predictable, but only can be studied at the expense of wasting the potential value of rich historic data and of a long time period of data collection since then the implement of intervention. The resulting variation is less likely to affect seasonal patterns, while it can bring similar impacts on responses for some strata.

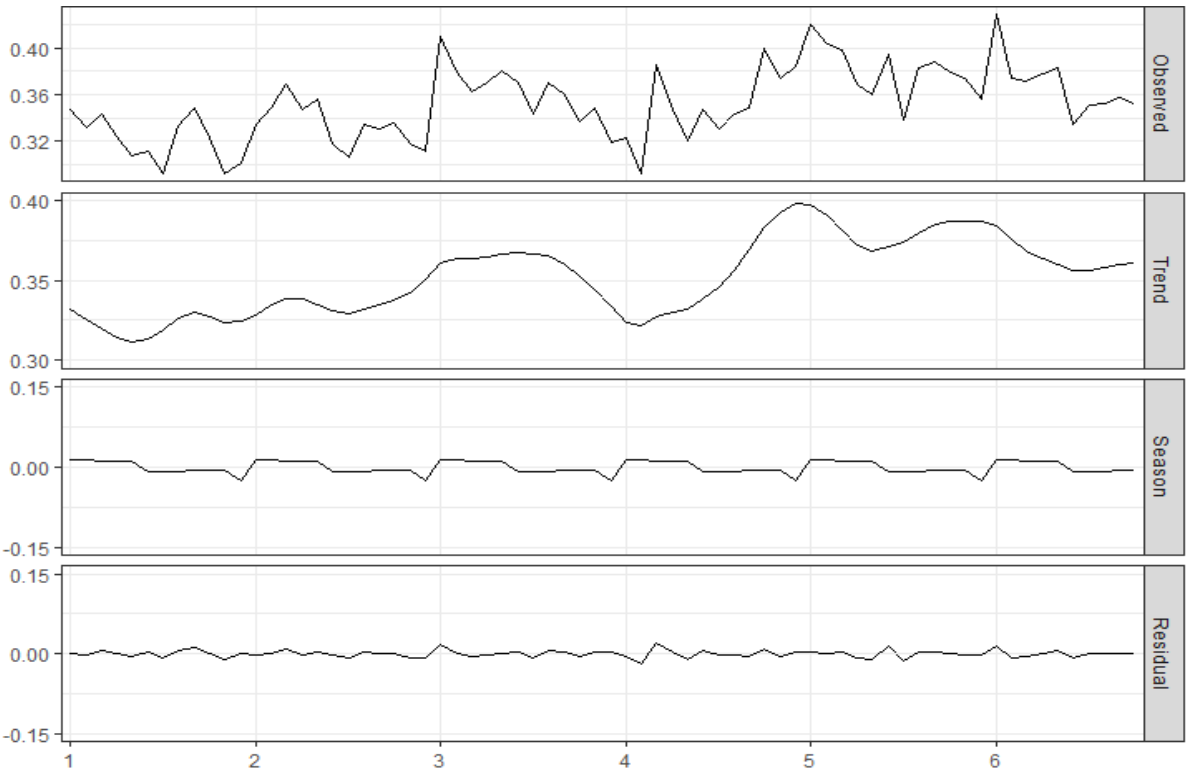

Figure 3.1 The observed series of simulated overall response rates over years versus its decomposition.

All components together, except for the sampling variation, form the signal, i.e., the latent true but unknown RPs. The mathematical formulations corresponding to each component are introduced in the following chapter that proposes the structural time series model; See Durbin & Koopman, (2012) and Harvey, (1990) for general background information on those time series components and models.

3.3 Methods

In this chapter, we translate the time series components discussed in Chapter 3.2 to multilevel time series models and devise the estimation strategy. We adopt a Bayesian approach in order to account for the uncertainty within the historic survey data and to update response propensity (RP) predictions in time. The use of multilevel models is widespread in small area estimation, in which interest focusses on reliable estimation for domains such as geographic areas, time periods, demographic subgroups, or a combination thereof, whose sample sizes are often too small to provide reliable direct estimates, see Rao & Molina, (2015) for an overview. Early references to the literature of small area studies using time series multilevel models include Pfeffermann & Burck, (1990); Rao & Yu, (1994); Datta et al., (1999); You et al., (2003). In most such studies, including Boonstra & van den Brakel, (2019), a Gaussian sampling distribution is assumed, possibly after a suitable transformation of the data. A notable difference of our current application to RPs is that we use a binomial sampling distribution, which is a natural distribution to describe the response process given the number of sampled individuals in each demographic subgroup and time period. Such binomial time series models have been considered by Franco & Bell, (2015). Their approach bears a resemblance to our strategy, whereas ours involves more different types of time series components in the model specification, such as seasonality.

We begin the discussion of our method by first introducing the notation used throughout the chapter. Next, we describe our model and the strategy used for estimating the RP, and we conclude with outlining the criteria used to evaluate the performance and applicability of prediction models for RPs in the Bayesian framework.

3.3.1 The multi-level time series model specification

The objective is to predict stratum-level RPs at a certain point in time. The population or a sample is partitioned into strata based on several auxiliary variables, i.e., stratified, equivalent to a cross-classification of selected variables. Here, we assume the stratification is specified prior to fitting the models. The categories of each variable may be merged to ensure sufficient sample sizes.

Let sample size in stratum g at wave t be $n_{g,t}$ and the number of respondents be $r_{g,t}$, where $g \in \{1, \dots, G\}$ and $t \in \{1, \dots, T\}$. The number of strata G is typically in the order of 10 to 20, and T refers to survey waves, each of which is a new replication of the survey starting from a fresh sample. We assume that all sampled units are independent in their response behavior within and between strata. For stratum g and time t, response $r_{g,t}$ follows a binomial distribution conditionally on RP $\rho_{g,t}$ and sample size $n_{g,t}$, i.e., $r_{g,t}|n_{g,t}, \rho_{g,t} \sim \text{Binom}(n_{g,t}, \rho_{g,t})$. Because RP is constrained to fall between 0 and 1, we transform the 0-1 scale to the real line $\mathbb{R}$ by utilizing a logit link function, where other link functions are usable as well. The function provides a nonlinear transformation and produces a latent variable $\theta_{g,t}$, which follows the log-odds function,

$$\theta_{g,t} = \text{logit}(\rho_{g,t}) = \ln\left(\frac{\rho_{g,t}}{1-\rho_{g,t}}\right).$$

We can reverse the transformation to compute $\rho_{g,t}$,

$$\rho_{g,t} = \frac{\exp(\theta_{g,t})}{1+\exp(\theta_{g,t})}.$$

For any stratum g and any time t, the linear predictor $\theta_{g,t}$ can take the most general form that can be linear, additive, multilevel and comprised of several time series components. As outlined in Chapter 3.2, there are demographic variables defining the strata, an overall trend, seasonal variation, stratum-specific trends, and a residual variation. Therefore, the multilevel model becomes:

$$\theta_{g,t} = \boldsymbol{\beta}'\boldsymbol{x}_g + \gamma t + \boldsymbol{\delta}'\boldsymbol{s}_t + v_g + u_t + z_{g,t} + w_{g,t}, \tag{3.1}$$

where the p-vector of regression effects $\boldsymbol{\beta}$ is associated with time-independent covariates $\boldsymbol{x}_g$. In the application we focus on later in this chapter, all covariates are binary as we only

consider categorical variables. However, in more general usage, the entries could be ordinal or numerical variables, such as contact attempts, and even they could vary over time.

Scalar γ is the slope parameter for the overall linear time trend. Vector $\boldsymbol{\delta}$ contains seasonal effects with vector $\boldsymbol{s}_t$ selecting the season corresponding to month t. The seasonal effects are either common to all strata, or they can be stratum-specific. In this chapter, we define seasons as a division of months in a calendar year, i.e., sets {1,2}, {3,4,5}, {6,7,8}, {9,10,11} and {12} as Winter, Spring, Summer, Fall and Christmas.

The first three terms are modelled as fixed effects while the last four terms are modelled as random effects in (3.1). The first of these random terms is the random intercepts for strata assumed to be normally distributed with mean 0 and variance σ_v^2 as

$$v_g \sim N(0, \sigma_v^2) \tag{3.2}$$

identically and independently for $g = 1, \dots, G$. Secondly, a global time trend is defined by a random effect vector $\boldsymbol{u} = (u_1, \dots, u_T)$ distributed as

$$\boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{V}_u). \tag{3.3}$$

Covariance matrix $\boldsymbol{V}_u$ describes the covariance structures between any u_i and u_j. One can assume either a first-order random walk (RW1, known as a local level trend) or a second-order random walk (RW2, the so-called smooth trend). The time-dependence structures are more conveniently expressed by the precision matrix, $\boldsymbol{Q}_u = \boldsymbol{V}_u^{-1}$. The precision matrix is preferred over the covariance matrix, since it is sparse and allows for efficient computation for hierarchical posterior inference in a Bayesian analysis, see e.g., Rue & Held, (2005). The matrix $\boldsymbol{Q}_u$ for RW1 and RW2 is a tridiagonal matrix and a pentadiagonal matrix respectively. Assumed a band matrix is $Q = (q_{i,j})$, 1 has one non-zero bands along the main diagonal such that $q_{i,j} = 0$ if $|i - j| > 1$, while 2 has two non-zero bands such that $q_{i,j} = 0$ if $|i - j| > 2$. See Appendix E for their definitions. Note that the precision matrices $\boldsymbol{Q}_u$ are singular, leading to an improper prior. This is not a problem, as constraints can be imposed on these random effects to ensure that all model coefficients remain identifiable. Under RW1 and RW2 $\boldsymbol{u}$, the constraint is $\sum_t u_t = 0$. Under RW2 $\boldsymbol{u}$,

the constraint $\sum_t tu_t = 0$ is additionally imposed, so that the corresponding overall level and linear slope are captured by the model's intercept and fixed effect γ.

We also consider distributions other than the normal distribution in (3.3). In particular, we consider Laplace, Student-t and horseshoe priors as alternatives. Such priors can be framed as scale-mixtures of the normal distribution, see West, (1987); Carvalho et al., (2010); Polson & Scott, (2010).

The third random effect term $\boldsymbol{z}_g = (z_{g,1}, \dots, z_{g,T})$ denotes stratum-specific trends distributed as

$$\boldsymbol{z}_g \sim N(0, \sigma_z^2 \boldsymbol{V}_z), \tag{3.4}$$

for $g = 1, .., G$. Covariance matrix $\boldsymbol{V}_z$ describes a RW1 over the months. The corresponding precision matrix is the same as described above, and a sum-to-zero constraint is imposed on each trend vector $\boldsymbol{z}_g$, as the stratum-specific levels are already captured by the random intercepts v_g. Important to note is that the trends $\boldsymbol{z}_g$ share a common covariance parameter σ_z^2. One could consider a separate variance parameter per stratum but we found it resulted in overfitting.

The last term $w_{g,t}$ in (3.1) represents white noise and allows for remaining unstructured variation in RPs over time and strata, i.e., at the most detailed level. For any stratum g and time t, these components are independently and identically distributed as

$$w_{g,t} \sim N(0, \sigma_w^2), \tag{3.5}$$

using a single variance parameter σ_w^2.

(3.1) describes the most general model considered combining all underlying components. Chapter 3.4 investigates this encompassing model as well as models built from various subsets of the components described in (3.2) - (3.5).

3.3.2 The estimation strategy

In this chapter, we adopt a hierarchical Bayesian approach to estimate model coefficients and predict RPs. Since the posterior distributions are unavailable in closed form a Gibbs

sampler is used as implemented in the *mcmcsae* R package (Boonstra, 2021). We begin this chapter by specifying the priors assigned to the model parameters.

For the fixed effects $\boldsymbol{\beta}$ we assume a weakly informative prior

$$\boldsymbol{\beta} \sim N(0{,}100\boldsymbol{I}_{\beta}),$$

with identity matrix $\boldsymbol{I}_{\beta}$. Standard errors for $\boldsymbol{\beta}$ are taken as 10, which is sufficiently large concerning the scale of RPs relative to the covariate scales. Similarly, the linear time trend γ, and seasonal effects δ, are assigned weakly informative priors also, with the same standard error.

For the random-effect components, the variance parameters in (3.2) - (3.5) are assigned inverse χ^2 priors, conditionally on auxiliary parameters ξ, with 1 degree of freedom and a scale parameter ξ^2. For example, $\sigma_v^2|\xi_v \sim \text{Inv}{-}\chi^2(1, \xi_v^2)$. The hyperparameters ξ are assigned $N(0{,}1)$ priors. Combining the normal ξ with the conditional inverse chi-squared variances results in marginal half-Cauchy priors for each standard deviation parameter σ_v, σ_u, σ_z and σ_w. As Gelman, (2006) and Polson & Scott, (2010) suggest, the half-Cauchy priors for standard deviations, or the more general half-t family of priors, generally perform better than the commonly used inverse gamma priors for variance parameters, which can be too informative.

The (hyper)parameter vector, denoted by ψ,

$$\psi = (\beta, \gamma, \delta, v, u, z, w, \sigma_v^2, \sigma_u^2, \sigma_z^2, \sigma_w^2, \xi_v, \xi_u, \xi_z, \xi_w)$$

includes all parameters in (3.1), the variance parameters associated with random effect terms as well as the introduced auxiliary parameters. The likelihood function can be written as

$$p(r|n, \psi) \propto \prod_{g,t} \rho_{g,t}^{r_{g,t}} \left(1 - \rho_{g,t}\right)^{n_{g,t} - r_{g,t}}, \tag{3.6}$$

where $\rho = logit^{-1}(\theta(\psi))$ and θ is the linear predictor function of vector ψ as expressed in (3.1). Based on Bayes' theorem, the posterior of vector ψ is proportional to the product of the prior and the likelihood, i.e., $p(\psi|n, r) \propto p(\psi)p(r|n, \psi)$. The Gibbs sampler then generates samples from the joint posterior, and the posterior estimates of RP $\rho_{g,t}$ comes as a by-product of these samples — per sample, RPs can be computed using reversed logit

transformation. Repeated samples are drawn from the full conditional posterior of each (hyper)parameter. See Appendix F for more information on the full conditional posterior distributions.

Three Markov Chains are produced by the Gibbs Sampler using the *mcmcsae* package (Boonstra, 2021) programmed in R (R Core Team, 2000). Each chain consists of 1500 draws that are sequentially generated; however only the last 1,000 draws are kept for the estimation algorithm. Convergence of the MCMC sample is assessed using trace and autocorrelation plots. The Gelman-Rubin potential scale reduction factor (Gelman & Rubin, 1992) is evaluated to diagnose the mixing of the chains. In particular, the autocorrelation of sequential draws is reduced, as the blocked Gibbs sampler updates all fixed and random coefficients simultaneously. In addition, the approach includes a novel data augmentation approach for sampling from binomial logistic models (Polson et al., 2013) which is known to lead to an efficient and relatively fast converging sampler.

3.3.3 Performance criteria

To guide the model building using the model components and priors described in Chapters 3.3.1 and 3.3.2, and to assess the models' adequacy, we employ three criteria for model assessment and one for model predictive performance.

The common and popular selection criteria in Bayesian hierarchical settings are the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010 & 2013) and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). They are chosen in the pursuit of a reasonable balance between model fit, model complexity and efficient computation (See Appendix E for their definitions). Models with lower DIC/WAIC are preferred. Next, we use posterior predictive p-values to check model adequacy, i.e., simulating draws from the posterior predictive distribution and comparing them to the observed data, see e.g., Gelman et al., (1996). This evaluates whether the multilevel model can reproduce data similar to the observations. The p-values are defined as

$$p = \Pr\left(S(r^{rep}) \geq S(r)|r\right), \tag{3.7}$$

where S is a test statistic and r^{rep} denotes a replicated dataset generated from the posterior predictive distribution based on the fitted model, $p(r^{rep}|r) = \int p(r^{rep}|\rho, n)p(\rho|r, n)d\rho$.

The p-values are estimated from the MCMC output, and values close to 0 or 1 are indicative of a poor fit regarding statistic S. Here we consider two test statistics:

1. $S(r) = \bar{r}$, the unweighted mean of the replicate data-vector.
2. $S(r) = \frac{1}{GT-1}\sum_{g,t}\left(r_{g,t} - \bar{r}\right)^2$, the unweighted variance of the replicate data-vector and $\bar{r}$ is the mean of $r_{g,t}$.

To assess the models' prediction performance, we define a predictive measure: the root mean squared error (RMSE) in stratum g at month t as the square root of the sum of two terms: 1) the quadratic differences between the posterior means of $\rho_{g,t}$ and the observed response rate (RR), and 2) the posterior variances of $\rho_{g,t}$. The general form of the expression in stratum g at month t is

$$\mathrm{RMSE}(g,t) = \sqrt{\left(E_{\pi_t}(\rho_{g,t}) - \hat{\rho}_{g,t}\right)^2 + \mathrm{var}_{\pi_t}(\rho_{g,t})}, \tag{3.8}$$

where $\hat{\rho}_{g,t}$ is the realized value of RP and estimated by the observed RR, and π_t is the posterior predictive distribution of the RPs, when employing historic data up to and including $t-1$ and new data in t for RP prediction. For ease of notation, the two terms under the square root in (3.8) are referred to as the bias term $B(g,t)$ and the standard deviation $\mathrm{SD}(g,t)$. The bias term in (3.8) will, in general, be larger than zero due to random variation in the sampling of strata and in the response of sample units. For this reason, we benchmark the RMSE against an empirical lower bound denoted by $\mathrm{RMSE}_{\min}$. The lower bound estimate is called the Monte Carlo approximation to the posterior mean of the binomial standard deviations, which is a function of the kth iteration from the posterior draws of $\rho_{g,t}$,

$$\mathrm{RMSE}_{\min}(g,t) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}, \tag{3.9}$$

where k runs over MCMC draws and $n_{g,t}$ is the size of stratum g sample in month t. (3.8) and (3.9) give one-month assessments per stratum g. They need to be aggregated across strata and in time to get meaningful overall assessments.

In any particular month, a stratum with a larger sample size should impose more weight on the reliable predictions. The weights $d_{g,t}$ are defined as the sample proportion, i.e.,

$$d_{g,t} = \frac{n_{g,t}}{\sum_g n_{g,t}} \text{ subject to } \sum_g d_{g,t} = 1.$$

Thus, the sub-terms

$$B(t) = \sqrt{\sum_g d_{g,t}\left(E_{\pi_t}(\rho_{g,t}) - \hat{\rho}_{g,t}\right)^2}$$

and

$$\mathrm{SD}(t) = \sqrt{\sum_g d_{g,t}\mathrm{var}_{\pi_t}(\rho_{g,t})}$$

in month t should be the square root of the sum of the weighted individual measures $B(g,t)$ and $\mathrm{SD}(g,t)$ by $d_{g,t}$ over strata, while the lower bound over strata in time t becomes

$$\mathrm{RMSE}_{\min}(t) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\sum_g d_{g,t}\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

Also, the stratum-specific sub-terms

$$B(g,T) = \frac{1}{T}\sum_t \sqrt{\left(E_{\pi_t}(\rho_{g,t}) - \hat{\rho}_{g,t}\right)^2}$$

and

$$\mathrm{SD}(g,T) = \frac{1}{T}\sum_t \sqrt{var_{\pi_t}(\rho_{g,t})}$$

in a time period $T = \{t|t_1, \ldots, t_T\}$ are the average of the individual measures $B(g,t)$ and $\mathrm{SD}(g,t)$ over months where t indicates a month, while stratum-specific lower bound over time period T becomes the average of the individual measures $\mathrm{RMSE}_{\min}(g,t)$, i.e.,

$$\mathrm{RMSE}_{\min}(g,T) = \frac{1}{T}\sum_t \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

Furthermore, the overall sub-terms or term in a time period T becomes the average of the weighted sub-terms $B(t)$, $\mathrm{SD}(t)$ and $\mathrm{RMSE}_{\min}(t)$ over months, i.e.,

$$B(T) = \frac{1}{T}\sum_t \sqrt{\sum_g d_{g,t}\left(E_{\pi_t}(\rho_{g,t}) - \hat{\rho}_{g,t}\right)^2},$$

$$\mathrm{SD}(T) = \frac{1}{T}\sqrt{\sum_g d_{g,t}\mathrm{var}_{\pi_t}(\rho_{g,t})}$$

and

$$\mathrm{RMSE}_{\min}(T) = \frac{1}{T}\sum_t \frac{1}{K}\sum_{k=1}^{K} \sqrt{\sum_g d_{g,t}\frac{\rho_{g,t}^{(k)}\left(1-\rho_{g,t}^{(k)}\right)}{n_{g,t}}}.$$

3.4 Analysis of Results

In this chapter, we introduce the Dutch Health Survey (GEZO) as a case study to demonstrate how the multi-level time series models can be built and how we update RPs in time. We address the four research questions in corresponding chapters.

3.4.1 The Dutch Health Survey

The GEZO has been conducted annually since 1981 by Statistics Netherlands as a repeated cross-sectional survey in which a sample of households was interviewed with the aim of providing an overview of developments in the health, medical consumption, lifestyle and preventive behavior of the Dutch population. The sampling frame is formed by first drawing a sample from municipalities and then from all people who live in the selected municipalities. As of 2010, the survey changed to a mixed-mode survey involving an initial web and the follow-up telephone (or face-to-face) interview. Non-respondents to web were contacted via telephone if their telephone numbers were known at the register, and otherwise a face-to-face interview was arranged. Over these years, the sample size was increased to 15,000 and the overall response rate was increased by 25%. From 2014 onwards, the mix of the follow-ups was changed to a face-to-face interview. In 2018, however, a part of the web non-responses was approached via a face-to-face interview in a more effective way. The propensity to respond to personal interviews in a time series strongly depends on web response outcomes, so in a sense modeling follow-up propensity is conditional on web RP model. This issue needs consideration more than interpreting time change in web RP and is beyond the main aim of this chapter. For the sake of simplicity, our concern is to model web response propensity in this chapter as a

fundamental start, hence modeling follow-up RP in a time series is more suited to future research. As an important note here, only the web GEZO data from 2014-01 up to and including 2019-10 are analyzed in this study. We employ three auxiliary variables that stem from administrative frame or registers. The prescribed auxiliary variables are age, gender and ethnicity, which divides the population or its sample into 20 disjoint strata (see Appendix C for more information).

The GEZO conducted over many years is a relatively consistent survey design. This feature makes exploring time-dependence in RPs valid because of the abundant time series. Our interest focuses on monthly response data, i.e., sample size and the number of respondents of each stratum. Predictions are made monthly but also can be aggregated quarterly or annually.

3.4.2 What time series components contribute most to variation in RPs?

We address this question in two steps: First, we go through model combinations and then we compare their performance. The comparison of multiple models is made from two views: (1) "what combination fits best to response data?" and (2) "what combination makes the most reliable predictions?" We use information criteria and posterior predictive p-values to measure the performance of each model, and thus search for an "optimal" model. The model is preferred when it has lower information criteria and predictive p-values closer to 0.5.

Since trying all combinations of components in (3.1) places a heavy burden on computation, it is important to apply an efficient search for the "optimal" model. To do so, we fit the models to response data using the following strategy:

1. Start with the baseline model (auxiliary variables only).
2. Add fixed effects sequentially, linear time trend and seasonal trends, to the baseline.
3. Investigate whether the model in **2** continues to improve with global time effects or global seasonality.
4. Investigate whether stratum-specific time trends or seasonal effects further improve the model.

5. Determine whether a white noise term for unexplained variation is needed.
6. Explore robustness for outliers through different prior specifications or time-dependent structure of global random intercepts over time.
7. Evaluate the model using a number of diagnostics.

Table 3.1 shows the selection results. The fixed-effect models (M1 to M3) behave worse than the mixed effect models relative to the trade-off between fitness and complexity, as the latter ones yield lower ICs (DIC and WAIC). Comparing M2/M3 to M1 implies that time slope λ or seasonality δ causes a decrease in ICs. However, the model further improves by introducing global trend u_t, as a significant decrease in ICs in M4 relative to M3 is observed. As M5 and M6 show, the improvement continues with the addition of random intercepts for strata v_g and stratum-specific time trends $z_{g,t}$. Although white noise $w_{g,t}$ seems to add only very little in M7 overall, the posterior predictive p-values for variances imply that it is worth to include white noise. Further, we found that using a local level trend (RW1) or smooth trend (RW2) as the global trend u_t makes hardly any difference concerning ICs for models M6 to M11.

Finally, the 4th column of Table 3.1 shows the prior distribution used for the global trend coefficients u_t. The non-normal priors that have been attempted do not further improve ICs, but because of heavier tails they help to combat an outlier in the data, an exceptional issue in February 2017.

T-distributed and horseshoe priors are likely to accommodate and be robust against the outlier better than normal and Laplace priors, as shown by comparing their posterior means of global trend u_t in Appendix D. Besides, the local level trend of M8 seems to outweigh slightly the smooth trends of M11. P-values of the mean of M8 bring the value closer to 0.5 than M9 and M11.

Table 3.1 Summary of the multilevel time-series models considered.

Model	Fixed	Random	Prior	DIC	pDIC	WAIC	pWAIC	PPP	
								Mean	Variance
M1	β	-	-	7,511	7	7,518	13	0.501	0.006
M2	β,λ	-	-	7,415	8	7,421	15	0.503	0.051
M3	β,λ,δ	-	-	7,368	12	7,378	22	0.498	0.092
M4	β,δ	u_t	Normal	7,255	43	7,280	68	0.484	0.168
M5	β,δ	u_t, v_g	Normal	6,916	56	6,925	65	0.491	0.172
M6	β,δ	u_t,v_g,$z_{g,t}$	Normal	6,790	98	6,781	90	0.494	0.356
M7	β,δ	u_t,v_g,$z_{g,t}$,$w_{g,t}$	Normal	6,790	131	6,769	110	0.517	0.425
M8	β,δ	u_t,v_g,$z_{g,t}$,$w_{g,t}$	Laplace	6,790	133	6,768	111	0.503	0.397
M9	β,δ	u_t,v_g,$z_{g,t}$,$w_{g,t}$	T-distributed	6,790	130	6,769	110	0.492	0.391
M10	β,δ	u_t,v_g,$z_{g,t}$,$w_{g,t}$	Horseshoe	6,790	131	6,769	110	0.518	0.411
M11	β,λ,δ	u_t,v_g,$z_{g,t}$,$w_{g,t}$	Laplace	6,806	150	6,779	122	0.519	0.413

Notes: “-” indicates no random effects or prior. Deviance Information Criterion (DIC); Widely Applicable Information Criterion (WAIC); Posterior predictive p-values (PPP).

To determine which model is flexible to the outlier and which one generates reliable estimation throughout the series, we look further into the discrepancy between observations and model-based estimates, specially M8, M9 and M11. The comparison demonstrates that the three models have limited ability to combat the outlier. The lower quantiles attempt to reach to the outlier but cannot cover it. In addition, the Laplace prior has slightly smaller uncertainty about the posterior estimates than the T-distributed prior, but it has similar size in uncertainty to the smooth trend model (See Appendix D).

3.4.3 What level of response propensity prediction accuracy can be achieved for the next upcoming new time period?

To answer the second research question, we estimate the level of and variation in overall response predictions for the forthcoming data collection wave. The estimated level is the deviation of the expected posterior propensity prediction from the realized response rate, while the estimated variation refers to the prediction accuracy in the overall RP. Also, we measure the balance between the level and variation and compare it with the benchmark in (3.9). The assessment allows us to validate if gains can be achieved from our method. Actions can be taken to adapt/maintain data collection in the following wave once the gain is known upon historic series.

We stress that the analysis is made based on the "optimal" model, M8. For all strata in a new sample per month, the months since January 2014 up to but not including the present month are viewed as the historic time series, which are used for training M8. Then we use the fresh sample from the present month for the estimated predictive criteria. The historic time series is accumulated and predictive criteria are updated with the new wave. The rolling assessment ends with 2019/09 as one month must be left for the prediction exercise because 2019/10 is the last month of data available. To lend robustness to the impact of historic size on predictive performance, we let historic time series start with 60 months (from 2014/01 to 2018/12) as the default initial trial.

Table 3.2 shows that the posterior uncertainty in the overall RP predictions decrease steadily but slowly and converge to around 0.027. Because of the sampling variation that is inherent to the bias term, the pattern for bias is erratic and shows at best a modest decrease. Relative to the realized response rates, the greatest deviation of posterior means

is around 0.07 in January and June, and the smallest deviation around 0.04 in March, May, August and October. The RMSE results vary along with the bias term across months, as the estimated SDs are much smaller than the estimated biases. The RMSE has a maximum value of 0.084 in January, and is likely caused by the outlier months in early 2017. Although the model reacts to this disruption, it has a negative impact on the performance of the resulting predictions in this month. Aside from January, in some months the estimated RMSE is close to the benchmark $RMSE_{min}$. It can be concluded that the estimated accuracy lies relatively close to the maximal possible accuracy.

Table 3.2 One month ahead prediction of three measures of RPs over strata: bias, standard deviation (SD), and the root mean square error (RMSE) compared to the benchmark ($RMSE_{min}$).

	2019									
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**
bias	0.078	0.064	0.045	0.062	0.046	0.077	0.063	0.049	0.058	0.048
SD	0.030	0.031	0.031	0.029	0.028	0.029	0.028	0.027	0.028	0.027
RMSE	0.084	0.071	0.055	0.069	0.054	0.082	0.068	0.056	0.065	0.055
$RMSE_{min}$	0.055	0.056	0.055	0.055	0.055	0.055	0.048	0.048	0.049	0.049

Notes: The column indicates the present month for evaluating prediction performance.

3.4.4 How does prediction accuracy vary over population strata?

This research question concerns the different strata and how well the model performs in predicting RP per stratum. For this purpose, we consider the stratum-level RMSE as well as its two components, i.e., bias and standard deviation. The evaluation measures are taken as the average over the ten months ahead predictions. Month 2019/10 is the last month available. Looking ahead by almost a year allows data collection staff to plan adaptive designs well ahead of time.

Similar to Chapter 3.4.3, we limit the analysis to the assumptions. The preferred model is selected from Chapter 3.4.2, and the historic time series is fixed to be 60 months (2014/01 to 2018/12). For each stratum, the model is fully trained by the fixed historic data and makes inference on predictions in the remaining months in 2019.

Table 3.3 shows prediction criteria for each stratum together with the benchmark. The estimated bias terms vary widely between strata. The greatest departure of posterior expectation from the realized response rates occurs in stratum 8, 10, 12, and 18, all with biases larger than 0.1. Compared to biases, there is a relatively smooth change in the estimated SDs around 0.03, where stratum 4 has smallest uncertainty about the posterior estimates with 0.018.

Table 3.3 The average of ten months ahead prediction of three measures in each stratum: bias, standard deviation (SD), and the root mean square error (RMSE) that is compared to the benchmark ($RMSE_{min}$).

	bias	**SD**	**RMSE**	**$RMSE_{min}$**
1	0.045	0.030	0.060	0.046
2	0.066	0.030	0.077	0.094
3	0.028	0.025	0.039	0.039
4	0.049	0.018	0.053	0.061
5	0.035	0.026	0.045	0.035
6	0.047	0.021	0.053	0.064
7	0.062	0.032	0.073	0.054
8	0.105	0.030	0.111	0.154
9	0.047	0.031	0.060	0.045
10	0.165	0.035	0.173	0.160
11	0.044	0.031	0.057	0.048
12	0.134	0.030	0.138	0.092
13	0.030	0.027	0.044	0.042
14	0.081	0.022	0.086	0.074
15	0.044	0.028	0.056	0.038
16	0.067	0.022	0.072	0.072
17	0.030	0.031	0.048	0.053
18	0.114	0.029	0.120	0.146
19	0.031	0.029	0.046	0.041
20	0.095	0.030	0.105	0.172

Some strata with greater biases may have less accuracy in posterior estimates of RPs than strata with less biased propensity. Similarly, the more biased the prediction is, the greater the RMSE is estimated to be. This is because the estimated biases are much greater than

the estimated SDs. It is not surprising that stratum 10 has the greatest value in RMSE, where the prediction is the most biased and has the least precision. The RMSE results can catch up with, and even can be comparable/superior to the benchmark. For example, when the model generates prediction for stratum 20, more significant gains can be achieved than other strata.

The predictive performance shows a significant difference between strata when there is only one different characteristic. For example, stratum 20 RMSE is 0.06 lower than stratum 10 RMSE. This seems to imply that female groups may have smaller bias or variance than male groups when non-western people above the age of 64 are considered. Given the age and ethnicity of groups and compared with non-western groups (even rows), RMSE results are much lower in western groups (odd rows). To validate this supposition, some strata are combined into subgroups with less detailed characteristics. As Figure 3.2 shows, the model yields better predictions for western group than non-western groups, as expected posterior estimates reach mostly the observed response per month. The comparative performance for age/gender groups are presented in Appendix D.

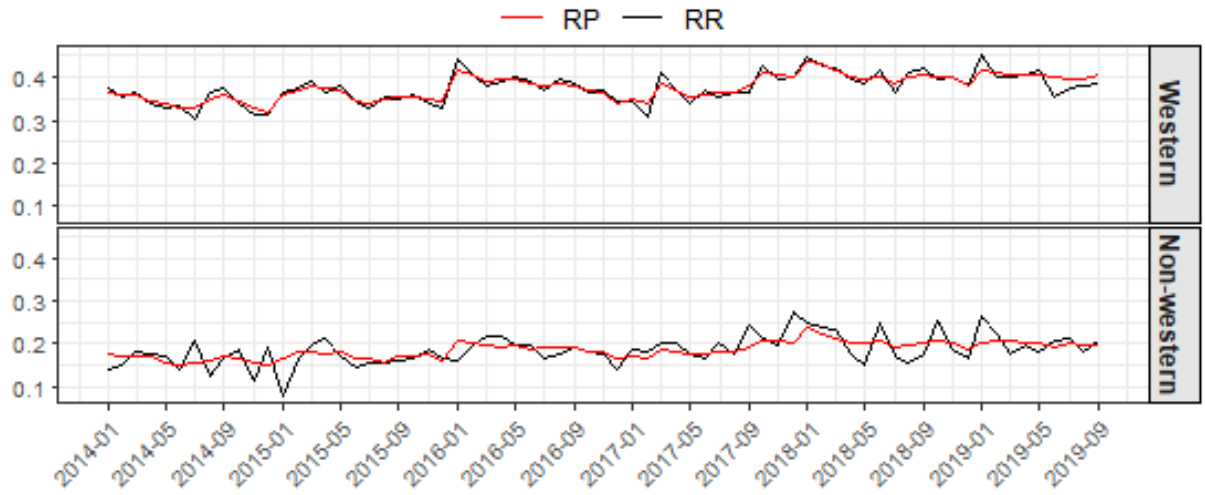

Figure 3.2 Monthly posterior means of RP aggregated over Ethnic groups versus observed response rates (RR) of Ethnic groups. Month 2014/01 to 2018/12 for the estimated model and Month 2019/01 to 2019/10 for RP predictions.

3.4.5 How does prediction accuracy depend on the length of the historic survey time series?

The primary concern of this question is to find out how robust the prediction performance is to the amount of historic time series that is used for model training and predicting. For this purpose, we continue with the average of three-month ahead predictions of RMSE and its two terms, bias and SD, at the overall level at any given time point. We call this length-based average the quarterly average. To explore the impact of historic data size, we perform 3-split time series cross validation on dataset, i.e., successively add three months of new data to the training dataset used for model-based predictions. This analysis is iterated on a rolling basis and the step-by-step strategy is laid down as follows,

1. Select the model components based on the whole time series.
2. Select the baseline set of historic time periods of length t. Partition the window into the training set D_o of the first $t-3$ time periods and the test set D_t of the last three periods.
3. Data D_o trains the selected model, by simulating from the posterior distribution of all model parameters, given D_o.
4. Based on the simulated model in **3,** posterior predictive means and variances are computed for the RPs for each stratum and each time point in the test set D_t.
5. Based on individual RPs predictions in **4,** compute overall Bias, and SD, by using the sample proportion $d_{g,t}$ in stratum g at time t as weights, then (3.8) computes the quarterly average of RMSE.
6. Expand the window to $t+1$, moving one time period forward. Repeat **3-5** for updating the predictions of RMSE and its two terms.
7. Repeat **6** until length t is the last available time period.

We stress that it is necessary to use at least two years as the initial training length when seasonal components are included. In our case, time periods are months and the time series runs until 2019/07.

In Fig 3.3 RMSE, Bias and SD estimates are length-dependent and computed over strata for three months ahead. The baseline window of training data for fitting the optimal model are 2014/01 to 2015/12. Along the time series window, bias results are approximately greater than two times the SDs. Consequently, RMSE results are dominated by biases and show the same volatility. At the end, their estimates are approximately 0.06, whereas SD estimates undergo a slight increase. The latter is somewhat surprising, as one would normally expect that using a longer time series to estimate the model would decrease the posterior standard errors of prediction. It turns out, however, that two events had a large impact on the prediction performance. First, early in 2017, data collection experienced an interruption caused by technical issues with the web server. This incident had a large immediate impact on RPs and consequently also on model prediction performances. Second, in 2018, conditional incentives were introduced and the survey questionnaire was made smartphone proof. This design intervention had a more gradual and longer lasting impact.

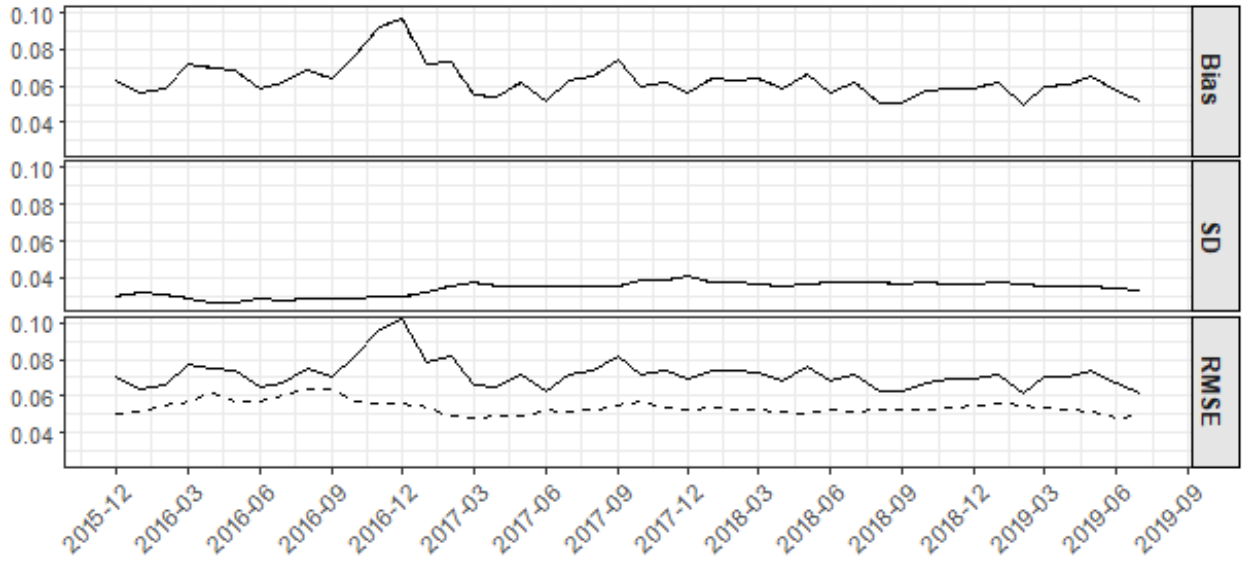

Figure 3.3 One-step forward moving average of quarterly Bias (upper panel), quarterly SD (middle panel) and quarterly RMSE (bottom panel). RMSE (solid) against benchmark RMSE (dashed) when the length of training data set moves on x axis.

The Bias and RMSE results undergo a big increase from 2016/10 to 2017/02. When the training window arrives at time point 2016/10, the test window starts to include 2017/01data where RPs dropped. Their climbing curves continues and reach maxima around 0.1, when the training window moves to 2016/12 and the test window first moves

to the "stable" month 2017/03 where the Bias and RMSE curves drop back around 0.05. During these months a slight, gradual increase in SD can be observed as well.

The inclusion of these outlier months affects prediction accuracy during 2017. Between month 2017/03 and 2017/12 the Bias and RMSE curves are more volatile, and decline only after 2018/01. The SDs have a rising tendency from 2017/03 to 2017/12 and hardly decrease.

In 2018, the impact of the design intervention on RPs was much more modest than in 2017, but since it is structural it does affect prediction accuracy in 2018 and 2019. The two events, one technical incident and one design change, are realistic in survey practice and when ignored can have a devastating impact. We set an example of how one might deal with them. The extra efforts are:

- **Discard method**. For the original data, clear the response numbers $r_{g,t}$ in 2017/01 and 2017/02 and treat them as missing data. Impute these missing $r_{g,t}$ by the posterior means of simulated responses from the posterior predictive distribution. Note that in Chapter 3.4.2 we argued that using specific non-normal priors for time series components can also limit the effect of outliers. If an outlier is quite extreme and known to occur at a specific time, it may however be better to discard it.
- **Intervention method**. Include an intervention term in the model or capturing the possible structural change. Add intervention binary variables to the original data series and let them be 0-1, where in our case they would take the value 1 and become active from 2018/01. The potential intervention-related effects could be either a single fixed effect, stratum-specific random effect or both.

Results from applying these two methods separately are shown in Fig 3.4. The two methods have a clear effect on predictions. In the period of 2016/12 to 2017/05 where the training time series window stops, the posterior means of the discard method show a declining trend, relative to the original model's posterior means ("Whole" in Fig 3.4). However, from 2017/10 to 2018/03 the difference in the mode-based means and observations becomes small for the discard method. Also, the discard method decreases uncertainty about posterior means as the credible band becomes narrower since from

2016/12. The intervention-related impact on overall RP cannot be estimated well using just a few new months of data.

While, it was not our intention to provide a detailed account of modelling options for incidental and structural changes, the time series model we propose can be modified in a relatively straightforward and flexible way. Replication with long survey time series is warranted to get a sense of what options are superior.

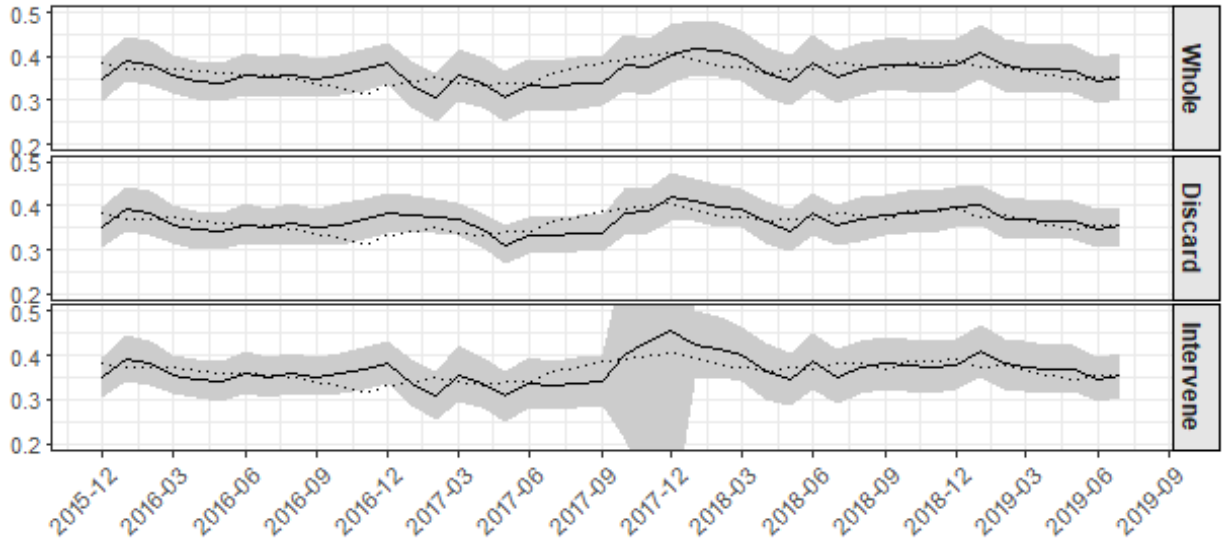

Figure 3.4 One-step forward moving average of quarterly posterior estimates of overall RP under three scenarios: (1) the original time series (top panel); (2) the new time series by discarding early 2017 data (middle panel), and (3) the new time series by adding intervention-related effects (bottom panel). Compared to the moving average of observed response rates quarterly (dashed), the model-based estimates are summarized as the posterior means (solid) with 95% credible region (band in grey). X axis labels the length of training data from 2014 to that time point.

3.5 Discussion

Accurate and reliable prediction of response propensities (RPs) is the key to improving and optimizing adaptive survey designs. Such inference can be complicated due to seasonal variation and time-related trends that may be specific to population strata. In this chapter, we introduce a Bayesian multilevel time series model for stratum-level RP predictions. The model is flexible enough to include seasonal variation, various forms of

trends, design changes and stratum-dependency, so that it can facilitate preparation for adaptive survey design in a changing survey climate. They are elicited from historic survey data and updated with new survey data.

In this chapter, we apply the method to a general population repeated survey, the Dutch Health Survey at Statistics Netherlands, to provide empirical support in a realistic case. The major focus is on improving stratum-level RP predictions that are subject to time-related factors. Based on various model combinations made of these factors, one of our concrete objectives is to search for the highest-performance model that makes a trade-off between model fitness and computational ease. The optimal model is selected based on criteria that assess both performance (high IC, p-value ≈ 0.5) and predictive ability (low RMSE_{RP}). These measures provide valuable insight into the relative gain achieved by adding new factors. This flexible approach allows other survey researchers to consider different time-related factors and ultimately choose the preferred model in their settings.

The remaining objectives of this chapter center on evaluating the prediction performance of stratum-level/overall RPs based on the preferred model. We use predictive metrics, specifically the root-mean-square error (RMSE), to assess uncertainty in predictions. This allows us to directly compare: (1) overall predicted response in first forthcoming data, (2) annual-averages of predicted response for each stratum, and (3) quarter-averages of overall predicted response. We evaluate the role of length of the historic survey time series in both the ultimately preferred model and a model that is re-optimized when data come in. Doing so we can find out when is a suitable time to start implementing an adaptive survey design. Note that when the survey design is made adaptive, it becomes less evident how to learn about the time change in model parameters. Also, the time series model itself may need to be updated depending on the type of survey adaptation.

While our attempt is a first step to adaptive survey designs, there are, however, various methodological and practical considerations that should be addressed. First, our approach is applied to a frequently-repeated cross-sectional survey. Historic data in such surveys has rich resources for relatively robust estimates of model coefficients and for making reliable predictions. When a survey is novel or conducted infrequently at a statistics bureau, our approach may be less powerful. Second, we assume that stratification is done through a fully-saturated model, i.e., strata are pre-specified by some auxiliary variables

that are strong predictors for web responses. How does the prediction performance change when adding less influential auxiliary variables? It is important to assess the sensitivity of reliable predictions to the choice of auxiliary variables. Also, we assume that strata are fixed throughout the time series. In survey practice, the selected auxiliary data may gradually change over time, and thus also the relevance of certain strata. Hence, it is essential to consider auxiliary-related change in stratification when predicting responses. Third, we assume the design of a survey should be the same over time, i.e., the model assumptions must be valid over the whole time series. If an intervention or another self-reported mode (e.g., smartphone) is introduced, variation in responses caused by this must be included explicitly. The advanced method is needed because there is no prior historic knowledge for a design change before it happens. A large jump can be caused by the inclusion of such a change in the model and, before the model can be informative about the effects of the change on RPs, a sufficiently long historic sampling must first be acquired.

We see also some limitations to the proposed methodology. In one particular year of the Dutch Health Survey data, we find a sudden increase in the standard deviations of predicted response propensities and overall quality indicators. The increase was the result of the intervention (smartphones were introduced as devices as well as conditional incentives). The results show that the model can be sensitive to design change. Hence, accounting for design changes is necessary and will temporarily reduce prediction performance.

Future research needs to address conditional response predictions in mixed-mode survey designs. In this chapter, we focus only on single mode response predictions. Such considerations are worthwhile for optimizing decisions of adaptive survey designs, for example, whether to switch to a cheap or expensive mode given the budget. Our method paves the way for the development of such conditional models.

Currently, the proposed model is designed for repeated cross-sectional surveys, but one may extend to other survey and sampling designs such as rotating panels. Such an extension would imply that panel response/attrition propensities are added to the model vector, and that the correlation structure among the propensities needs to be revisited.

Chapter 4

Robust adaptive survey design for time changes in mixed-mode response propensities

Wu, S., Boonstra, HJ., Moerbeek, M., & Schouten, B. (2023). *Submitted to Survey Methodology Journal (SMJ).*

Author contributions: SW, HJB, and BS contributed to the study concept and design. Statistics Netherlands provided the survey data. BS organized and coded a part of raw data collected from individuals. SW prepared the aggregated data, performed the analyses and wrote the first draft of the article. MM, BS and HJB critically reviewed the paper.

Abstract

Adaptive survey designs (ASDs) tailor recruitment protocols to population subgroups that are relevant to a survey. In recent years, effective ASD optimization has been the topic of research and several applications. However, the performance over time is sensitive to time changes in response propensities. How adaptation strategies can adjust to such variation over time is not yet fully understood. In this chapter, we propose a robust optimization approach in the context of sequential mixed-mode surveys employing Bayesian analysis. The approach is formulated as a mathematical programming problem that explicitly accounts for uncertainty due to time change. ASD decisions can then be made by considering time-dependent variation in conditional mode response propensities and between-mode correlations in response propensities. The approach is demonstrated in a case study: the 2014-2017 Dutch Health Survey. In the comparisons, we evaluate the sensitivity of ASD performance to 1) the budget level and 2) the length of historic time-series data.

Keywords: Response propensity model; time series analysis; allocation optimization; adaptive survey design

4.1 introduction

Adaptive survey designs (ASDs, Wagner, 2008; Schouten et al., 2017) have rapidly become an interesting alternative to conventional surveys; a single survey protocol is no longer offered to all individuals or subgroups but rather can be tailored to efficiently attain their responses based on population characteristics. This shift was accelerated by persistent declines in response rates, limited budgets, the variety of data sources, the emerging new means to collect data (e.g., smartphones), the multiple survey modes and so on. Combined, these factors put survey quality at risk, where most designs focus on nonresponse, and lead to an increase in survey costs.

A key element in ASD is the optimization strategy, i.e., the set of decision rules. Such strategies rely on input on response propensities and other survey design parameters. The main approaches to optimization include case prioritization (Peytchev et al., 2010; Wagner, 2013; Wagner & Hubbard, 2013), trial and error, and mathematical and statistical

optimization (Calinescu & Schouten, 2016; Schouten et al., 2013). See Schouten et al., (2017) for the advantages and disadvantages of each approach. These researchers mostly ignored inaccuracy in response propensities estimated from historic data. In mathematical programming, objectives can be parameterized as functions of response propensities acting as one of the main inputs to optimization. Error would be introduced in making decisions when response propensities change over time. As a result, inaccuracy treats any ASD as suboptimal, or even worse makes, it ineffective. Placing ASD optimization in a Bayesian context is natural to address this issue, yet the relevant survey methodology research is still in its infancy. Recently, Ma et al., (2021) developed a methodology to efficiently optimize a stratification by holding out for accurate estimates of response propensities in a Bayesian manner, given the most recent historic data.

Time changes in response propensities and inaccurate estimates from historic data endanger the robustness of ASD optimization. See Schouten et al., (2017) and Chun et al., (2018) for more discussion. Recently, ASD researchers started to focus on developing response propensity models and improving prediction accuracy; Schouten et al., (2018) pioneered Bayesian updating methods to combat this bias by statistically leveraging accumulated survey data and historic data generated from past implementations of the same survey. Being the most informative, prior beliefs gathered from past survey data can enhance current data for prediction purposes. Clearly, translating external data sources to prior beliefs is a requisite for the development of response propensity models. To do so, using a literature review (West et al., 2021) and eliciting expert knowledge (Coffey et al., 2020 and Wu et al., 2022) are recent approaches to source prior information. Survey researchers treat the matter of historic data timeliness incompletely and regard response propensities at different survey phases overall, whereas some facts, such as consistently reduced response rates over years, indicate that accurate estimates of response propensities are dependent on time, and response propensities in sequential designs are likely to correlate. The most closely related work by (Wu et al., 2023) explored deconstructing time changes in response propensities at multiple levels to study the length of historic survey data, ensuring stabilized prediction accuracy. While modeling time-related effects involved only the Computer-assisted web interview (CAWI) data collection phase in the Dutch Health Survey (GEZO) that follows up CAWI nonrespondents by Computer-assisted personal interviewing (CAPI), the present study additionally evaluates conditional

prediction accuracy based on accurate CAWI predictions and potential correlation in between.

Taken together, this chapter aims to make two contributions in sequential mixed-mode (MM) designs: predicting each survey mode response propensity as accurately as possible and making adaptive decisions in the Bayesian context as optimally as possible. To fulfil this ambition by leveraging historic time-series data in the evaluation, we raise three research questions:

- How can time-series models be constructed to improve response propensity prediction accuracy in a sequential mixed-mode design?
- How sensitive is ASD performance to the specified budget level?
- How does ASD performance depend on the length of historic data?

In response to the first question, we extend the multilevel time-series models for a single mode proposed by Wu et al., (2023) to each mode with the application of a case study, the GEZO survey. Such an extension considers the in-between correlation of response propensities of CAWI and CAPI. Considering the alternative of model components, the models fit to 2014-2017 survey data are compared with each other via information criteria (Spiegelhalter et al., 2002; Watanabe, 2010 & 2013) for selecting the "best" representative. Concerning the second and third questions, a strategy that accommodates uncertainty about input to optimization when optimizing probabilistic allocations is in great demand. The survey design performance is monitored through an indirect indicator of the risk of nonresponse bias, i.e., the coefficient of variation of response propensities (CV). We benchmark the ASD performance against the performance of CAWI-only and nonadaptive designs to ensure that the determined allocations can improve the ASD performance. To determine the sensitivity of ASD performance, we conduct two experiments. The first experiment, in which a quarter is fixed as the optimization target and the budget level is gradually decreased, enables comparison of posterior CV estimates to probe the sensitivity of the ASD performance to the specific level. In the second experiment, the budget level is set, and the time series window of historic data moves forward to the next new data collection quarter. This training data set is used in updating the model's estimates and reoptimizing the allocations for the following quarter. The

dependence of ASD performance on the length of historic data can be observed by comparison of posterior CVs.

The outline of this chapter is as follows: We begin by developing the time-series approach to modeling the sequential response propensities between survey modes and to measuring the mode-correlated impact on propensity predictions in Chapter 4.2. Chapter 4.3 constructs an optimization problem subject to reasonable budget overrun, enabling us to reduce nonresponse bias through CV to the greatest extent. Then, the chapter that follows describes a sequential web-face-to-face mixed-mode survey, i.e., the Dutch Health Survey (GEZO), and we apply the proposed model to it to evaluate the model performance and determine a certain adaptation. The last chapter discusses the advantages and disadvantages of our method and concludes with some thoughts on future research.

4.2 Methods

In this chapter, a multivariate time series model is developed for response propensities in sequential mixed-mode designs. Notably, we extend the approach of Wu et al., (2023). The updated approach provides the necessary background for setting up an optimization in the Bayesian context in Chapter 4.3.

4.2.1 Modeling response propensities in sequential mixed-mode designs

The models of Wu et al., (2023) generate precise estimates of response propensities for survey designs with a sole mode or for the first mode of mixed-mode surveys during fieldwork. Here, the objective evolves into making reliable predictions for each mode of mixed-mode surveys to broaden the model's appeal. Notably, discrete-valued time series data, including the size of a sample and the number of respondents to each mode, are considered from a multinomial distribution, while Wu et al., (2023) considered a binomial distribution for data.

Response propensity is the theoretical propensity of a sampled subject with a set of known characteristics being a responder in a specific mode interview. This subject can be either an individual or a well-defined stratum. Of interest, a stratification is made by cross-classifying several auxiliary variables that are regarded as strong predictors of survey

variables. Within a stratum, units have homogeneous demographic attributes, such as age. Such a stratification can vary with time or design change (See Schouten et al., 2017 for more discussion on a stratification), but we assume that it is specified before fitting the model and that it stabilizes during data collection.

To model mode-level RP with ease, in this chapter, we first blur the subscripts indicating a specific stratum in the propensity parameter and indicating a specific time point, but the next chapter must specify this subscript to decompose a time series to some fixed or random effects at the stratum, time, and/or mode level.

Assume that a mixed-mode survey is provided with *M* modes for data collection. The first *M*-1 modes correspond to survey modes implemented in practice, while the *M*th mode corresponds to nonresponse, implying no response to the first *M*-1 modes. Let a random sample of size n be known before data collection starts, and let r_j denote the observed number of respondents in the *j*th mode, where $j \in \{1, \dots, M\}$. Consider a multinomial distribution in M modes with response propensity ρ_j for the *j*th mode, where $\rho_j \in [0,1]$. Clearly, ρ_M is the nonresponse propensity when $j = M$; however, it is no longer explicitly modeled later.

Vector $\boldsymbol{r} = (r_1, \dots, r_M)$ follows a multinomial distribution with sample size n and response propensity $\boldsymbol{\rho} = (\rho_1, .., \rho_M)$, i.e., the joint distribution of $\boldsymbol{r}$ having the likelihood as a multivariate generalization of a binomial distribution,

$$mult(\boldsymbol{r}|n, \boldsymbol{\rho}) = \frac{n!}{\prod_{j=1}^{M} r_j!} \prod_{j=1}^{M} \rho_j^{r_j}. \tag{4.1}$$

Linderman et al., (2015) used a stick-breaking transformation to reformulate the multinomial distribution as a product of binomial distributions, where the constructed parameters are dependent. This offered a chance to rewrite the m-dimensional (4.1) recursively in terms of $M - 1$ binomials. In the stick-breaking representation, the propensity vector $\boldsymbol{\rho}$ serves as a stick that is recursively split into two pieces to create binomial variable $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_1, \dots, \tilde{\rho}_{M-1})$. To provide a derivation, let the jth mode response variable r_j follow a binomial density with parameters n_j and $\tilde{\rho}_j$, i.e., $bin(r_j|n_j, \tilde{\rho}_j)$, where

n_j and $\tilde{\rho}_j$ represent the remaining size of the sample and the fraction of the remaining probability approached by the jth mode,

$$n_j = n - \sum_{k<j} r_k, \tag{4.2}$$

$$\tilde{\rho}_j = \frac{\rho_j}{1-\sum_{k<j}\rho_k}, \tag{4.3}$$

where $j \in \{2, \dots, M\}$. When $j = 1$, parameter $n = n_1 = \sum_{j\in\{1,\dots,M\}} r_j$ and parameter $\tilde{\rho}_j = \rho_1$.

The exponential term in (4.1) can be rewritten using ρ_j rather than $\tilde{\rho}_j$ to represent the exponential term in the binomial density explicitly by substituting (4.3),

$$\tilde{\rho}_j^{r_j}\left(1-\tilde{\rho}_j\right)^{n_j-r_j} = \rho_j^{r_j}\frac{1}{\left(1-\sum_{k<j}\rho_k\right)^{n_j}}\left(1-\sum_{k\le j}\rho_k\right)^{n_j-r_j}. \tag{4.4}$$

Note that r_j sums to n over j and $n_j = n_{j-1} - r_{j-1}$ for any $j \in \{2, \dots, M\}$. This simplification means that pairs of terms $\left(1-\sum_{k<j}\rho_k\right)^{n_j}$ will cancel out and immediately offer the product of (4.4)**,** leading to the same format as the multinomial exponential term in the following,

$$\prod_{j=1}^{M-1}\tilde{\rho}_j^{r_j}\left(1-\tilde{\rho}_j\right)^{n_j-r_j} = \prod_{j=1}^{M}\rho_j^{r_j}. \tag{4.5}$$

The normalization constants follow the same logic and work out correctly as well, combined with exponential terms in (4.5), so (4.1) can be rewritten as

$$mult(\boldsymbol{r}|n,\boldsymbol{\rho}) = \prod_{j=1}^{M-1} bin(r_j|n_j,\tilde{\rho}_j). \tag{4.6}$$

We use the stick-breaking representation of the multinomial model for practical reasons: for this representation, there is a simple and efficient Gibbs sampler for the multinomial (multilevel) model. As explained in Linderman et al., (2015), it uses the same Pólya-Gamma data augmentation method (Polson et al., 2013) that was used for the binomial models in Wu et al., (2023), and the stick-breaking representation allows sampling the model coefficients for all M-1 modeled categories in a block, thereby improving the convergence of the Gibbs sampler. This representation also has a drawback: the definition

of $\tilde{\rho}_j$ makes interpretation of the underlying model coefficients more difficult, particularly the interpretation of correlation coefficients in the models detailed in Chapter 4.2.2.

In the following, we employ a structural time-series model to decompose an observed time series into some underlying time-related components.

4.2.2 Multinomial multilevel time series model

To measure the dependence of response propensities among the modes, we develop the models of Wu et al., (2023) by introducing a new hierarchical parameter indicative of correlation coefficients. Such dependence spread over time-series components of interest is similar to those adopted in Chapter 3; revisit that chapter for more details on each component's definition and for technical details.

To describe each model component at the most detailed level, let the dependent propensity parameter vector of the sequential modes be associated with a specific stratum and time point, i.e., $\widetilde{\boldsymbol{\rho}}_{g,t} = \{\tilde{\rho}_{g,t,j} | j \in \{1,..,M-1\}\}$, where the jth entry denotes the propensity parameter of the jth mode in stratum g at time t, as defined in (4.3). The numbers of strata, time points and survey modes are G, T, and $M-1$, respectively.

We let a latent variable $\theta_{g,t,j} = logit(\tilde{\rho}_{g,t,j})$. A logit link function is defined to take a linear combination of some model components and convert the constrained scale of a probability between 0 and 1 to the real line $\mathbb{R}$. Therefore, the multinomial likelihood function (4.1) can be rewritten by substituting the inverse transformation for $\tilde{\rho}_{g,t,j}$,

$$mult(\boldsymbol{r}|n,\boldsymbol{\rho}) \propto \prod_{j=1}^{M-1}\left(\frac{e^{\theta_{g,t,j}}}{1+e^{\theta_{g,t,j}}}\right)^{r_j}. \tag{4.7}$$

The multilevel models considered for modeling the linear predictor $\theta_{g,t,j}$ take the general form of additive decomposition, which refers to a function of the sum of time-series components. Thus, the signal $\theta_{g,t,j}$ has the form

$$\theta_{g,t,j} = \beta_j + \beta'_{xj}x_g + \delta'_s s_t + u_{t,j} + v_{g,j} + z_{g,t,j} + e_{g,t,j}. \tag{4.8}$$

The first three and the last four terms are modeled as fixed effects and random effects, respectively.

The first regression fixed effects β_j are mode-specific intercepts, measuring the main effect on $\theta_{g,t,j}$. The second fixed effect β_{xj} measures the mode-specific regression coefficients associated with p-vector covariate x_g, while the third fixed effect δ_s measures the season-specific regression coefficients associated with q-vector s_t. Their estimators can interpret the between-mode differential in the effect of stratum g possessing some specific demographic characteristics that make up x_g and of time t belonging to the specific season. Currently, all strata share common seasonal and mode effects. In a broader sense, these fixed effects can be stratum-specific. In the present application throughout this chapter, the entries x_g and s_t are binary in the usage of categorical variables, but they can extend to ordinal or numerical and even time-variant variables if needed.

Each random effect term in (4.8) implicitly allows correlation between survey modes. Refer to Wu et al., (2023) for a description of the random effect components. As stressed, these terms are now crossed with the mode, i.e., separate variance parameters for each mode and correlation parameters among the modes are introduced. The global time trend $\boldsymbol{u}$, random intercept for strata $\boldsymbol{v}_g$, and stratum-specific trend $\boldsymbol{z}_g$ conform to this rule. White noise random effects $e_{g,t,j}$ are also crossed with mode, but we use a single common variance parameter, and no correlation is allowed.

We adopt a Bayesian approach to the estimated model in (4.8) to obtain reliable predictions of the response propensities at the mode, stratum and time levels. As noted, the priors are the same for coefficients corresponding to different modes. For notational convenience, we suppress subscripts g, t and j in each model component term. Fixed effects β and δ are provided with weakly informative priors normally distributed with zero mean and diagonal variance matrix, where the standard error takes a very large value of 10. Each random effects vector is assumed to be distributed as a Gaussian prior with mean $\mathbf{0}$ and covariance denoted as the Kronecker product of covariance matrices A and V. The fully parameterized covariance matrix $\boldsymbol{V}$, conditional on the introduced auxiliary parameter vector $\boldsymbol{\xi}$, is distributed as a scaled-inverse Wishart prior (Gelman & Hill, 2007; O'Malley & Zaslavsky, 2012). More technical details about the prior specification, the estimation strategy and the full conditionals of each hyperparameter can be found in Wu et al., (2022) and Boonstra et al., (2019).

A more parsimonious model can be obtained by omitting the mode-oriented interaction and replacing the fully parameterized covariance matrix V with uncorrelated forms, such as a diagonal matrix, if there is little interest in the between-mode effects on propensity predictions. This model is called the no-correlation model relative to (4.8), which provides a reference to evaluate the model performance when incorporating mode correlations.

4.3 Optimizing mode allocation under the Bayesian multilevel time series model

This chapter explores and exploits an allocation problem accounting for such uncertainty to grasp the timeliness and implementation of adaptive survey designs. Chapter 4.3.1 outlines the main ingredients for the construction and operation of this problem in a Bayesian framework. A strategy is proposed in Chapter 4.3.2 to assess the gain of adaptive allocations against nonadaptive allocations concerning nonresponse bias risk reduction by monitoring a measure of bias risk.

4.3.1 Main ingredients

Generally, mathematical optimization involves the selection of the "best available" values of some objective function relative to a number of constraints by choosing input values from an allowed set. Establishing optimization models entails three major elements: decision variables to optimize the goal, objectives to be minimized or maximized, and constraints on the decision variables. Because of optimization on the Bayesian setting, we emphasize that all mentioned statistical parameters are considered to be random variables with values that change over time. Consequently, objective functions and constraint functions are also random variables. In the following, the main ingredients are first introduced for a non-Bayesian setting and are then developed for the Bayesian setting.

Decision variables are symbolic representations of an intervention decided by the decision maker. They represent unknown parts of an objective function that can be manipulated and may take on any possible value within an allowed set if specified. In this chapter, an intervention is supposed to allocate interview modes to strata when preceding modes fail to obtain their data. Therefore, decision variables refer to allocation probabilities that indicate how likely nonrespondents are to be approached via a follow-up mode. Allocation

probability $s_{g,t} \in [0,1]$ makes a decision on the size of follow-up candidates in stratum g at time point t, where $s_{g,t} = 0$ implies that the collection of stratum g data should be halted up to the present mode, and $s_{g,t} = 1$ means that all stratum g nonrespondents in the preceding modes should be allocated to the next mode in an effort to recruit the toughest cases to attain their representativeness.

The objective function defines the evaluation criterion for the solution of decision variables. The objective is formulated as a mathematical function of decision variables and a global optimum to be found. The search for the global optimum may be hampered by the existence of multiple local optima. A unique global optimum can be guaranteed to be found under certain properties of the objective function, such as linearity and convexity. In the setting of this chapter, the objective functions do not have these properties. Therefore, numerical methods are necessary. Our optimization goal is to minimize the expected risk of nonresponse error via optimal allocation. Since nonresponse cannot be observed directly, this chapter considers a proxy indicator of nonresponse error that is a function of response propensities. Herein, the indicator is the coefficient of variation of response propensities, or CV for short (See Schouten et al., 2009). The true population CV bounds the absolute standardized bias of respondent means. The estimated CVs can be close to this upper bound when auxiliary variables are strongly related to nonresponse. See Moore et al., (2018) and Nishimura et al., (2016) for other proxy indicators. The overall indicator is the standard deviation divided by the weighted response rate,

$$CV(s,t) = \frac{\sqrt{\sum_g d_{g,t}(\rho_{g,t} - \bar{\rho}_t)^2}}{\bar{\rho}_t}. \tag{4.9}$$

Weight $d_{g,t}$ is the sample proportion of the stratum g size at time t against the overall size at time t, that is, $d_{g,t} = \frac{n_{g,t}}{\sum_g n_{g,t}}$. This notation implicitly assumes that the sampling design leads to equal inclusion weights, but if not, the design weights should also be incorporated. This addition is straightforward but makes the notation intractable. Mixed response propensity $\rho_{g,t}$ denotes the overall propensity over modes, which is the sum of the marginal response propensity of the starting mode and the joint response propensities of mode $j \geq 2$ supposing that stratum g did not respond to the last $j - 1$ modes (Here, we implicitly assume that all nonrespondents in a mode are eligible for follow-up. In practice,

some types of non-response, such as due to physical or mental illness, may not be eligible),

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j\in\{2,..,m-1\}} \rho_{g,t,j} \prod_{i\le j-1}(1-\rho_{g,t,k}). \quad (4.10)$$

The individual (conditional) propensities $\rho_{g,t,j}$ for any mode j are estimated by multinomial models assumes that all nonrespondents to the preceding modes will be recruited by mode j for a nonadaptive survey; however, this can be modified to an adaptive case by reducing joint propensity to decision variable $s_{g,t,j} \in [0,1]$, so the updated equation becomes

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j\in\{2,..,m-1\}} s_{g,t,j}\rho_{g,t,j} \prod_{i\le j-1}(1-\rho_{g,t,k}). \quad (4.11)$$

Clearly, (4.10) is equivalent to (4.11) when all $s_{g,t,j} = 1$.

The denominators of (4.9), called the weighted response rates over strata, indicate the estimated level of unknown propensities, which are defined as the weighted sum of mixed propensities of (4.10) and (4.11). We call CV nonadaptive when all $s_{g,t,j} = 1$ and adaptive when at least one $s_{g,t,j} \neq 1$.

Constraints are functional inequalities or equations that represent logical restrictions on what values of decision variables are allowed. For example, constraints might ensure a thorough search of feasible solutions from a finite solution space. In the survey design context, a constraint can be a limit placed either on the survey quality, such as solutions making the overall response rate greater than 0.5, or on the survey cost, such as the overall cost of interviewers reaching nonrespondents being lower than a specified amount. In this chapter, we focus on cost constraints regarding the workload of approaching nonresponse candidates by means of a follow-up mode. To constrain the workload, let the budget level be h; the adaptive workload should not exceed the nonadaptive workload under the budget level h,

$$p\left(\sum_g s_{g,t}\left(n_{g,t} - r_{g,t}\right) \ge \sum_g h\left(n_{g,t} - r_{g,t}\right)\right) \le \alpha, \quad (4.12)$$

Here, acceptable budget overrun α is used in opting for the best sound values of $\boldsymbol{s}$, and p is the probability of the adaptive workload exceeding the budget-constrained nonadaptive

workload. When the values of $s_{g,t}$ satisfy constraint (4.12), the corresponding solution of the decision variable is called acceptable; otherwise, the solution will contradict the rule. It is natural to specify lower and upper bounds on decision variable $\boldsymbol{s} = \{s_{g,t} | \forall g, t\}$, which are referred to as box constraints,

$$0 \leq \boldsymbol{s} \leq 1. \tag{4.13}$$

Therefore, the optimization problem in a sample allocation application for the non-Bayesian setting is formulated to detect a vector $\boldsymbol{s}$ that minimizes objective (4.9) subject to constraints (4.12) and (4.13) given parameters (n, r). As stated above, (4.9) and the workload in (4.12) are random variables in the Bayesian case, so we take expectations of the posterior distributions. Given that no explicit expression exists for the posterior distributions, they are approximated. We then obtain

$$\hat{E}\big(CV(s,t)\big) = \frac{1}{K}\sum_k \frac{\sqrt{\sum_g d_{g,t}\left(\rho_{g,t}^{(k)} - \bar{\rho}_t^{(k)}\right)^2}}{\bar{\rho}_t^{(k)}}, \tag{4.14}$$

where $\hat{E}\big(CV(s,t)\big)$ refers to the estimated posterior expectation at time t, $\rho_{g,t}^{(k)}$ is the *k*th iterated estimate from the posterior predictive function of $\rho_{g,t}$, and subscript *k* runs over MCMC draws. The excess probability p in (4.12) is estimated empirically by the ratio of the number of times the excess workload exceeds *K*,

$$\frac{\sum_k \mathbf{1}_{\sum_g s_{g,t}\left(n_{g,t} - r_{g,t}^{(k)}\right) \geq \sum_g h(n_{g,t} - r_{g,t}^{(k)})}}{K} \leq \alpha. \tag{4.15}$$

$\mathbf{1}_{\sum_g s_{g,t}\left(n_{g,t} - r_{g,t}^{(k)}\right) \geq \sum_g h(n_{g,t} - r_{g,t}^{(k)})}$ is an indicator function that takes a value of one when inequality as a subscript is met for the *k*th iteration and is zero otherwise. Therefore, Bayesian optimization aims to minimize objective (4.14) subject to constraints (4.13) and (4.15).

Benchmark: In the Bayesian optimization problem, we set a benchmark to evaluate ASD performance from two viewpoints: improving quality and saving money. Specifically, promoting sample representativeness by recruitment can improve data collection quality, while distributing cost-prohibitive resources to where they are most needed can save money. This goal can be achieved, for example, by switching from a single mode to mixed but optimally reallocated modes or switching from full mixed modes to partial mixed

modes. By letting decision variables $\boldsymbol{s} = 0$ or $\boldsymbol{s} = 1$, the optimization problem proposed above can settle those reallocations. To do so, the performance of the single-mode design and the full mixed-mode design are standards of and compared with the ASD performance.

4.3.2 Static ASD optimization

In allocating survey resources, ASDs based solely on information available in registry and frame data before the start of data collection are termed *static,* while ASDs based on paradata (data collected during data collection) are termed *dynamic*. These *dynamic* ASDs reflect the dynamic nature of the optimization since optimization is performed at each data collection phase, i.e., after each mode is completed.

For *dynamic* ASDs in the current context, decisions on assigning interviewers to strata are made dependent on intermediate survey results from the preceding modes. The correlations between response propensities to different modes are employed to update the prior distributions for the interviewer response propensities. Theoretically, the evaluation can identify the priorities of refusers in strata to be interviewed and inform the interviewer workload. In reality, there may be insufficient time to compute reallocated interviewers' workload in time because of geographical clustering. Additionally, reallocation requires complex logistics in case management; we leave this point to the discussion chapter.

Therefore, in this chapter, we focus on the *static* ASD. Chapter 4.3.3 constructs the strategy to account for uncertainty in making decisions and to specify the optimization routine to determine the optimal allocations for the Bayesian optimization problem in Chapter 4.3.1.

4.3.3 The optimization strategy

To solve the formulated optimization problem in Chapter 4.3.1, we propose a two-step strategy at time t,

1. *Construct the posterior distribution of the response numbers* $r_{g,t}$. Let historic time series data sets up to time $t-1$ be data used for model training, and let data sets at time t be test data for prediction. All model coefficients and

hyperparameters in (4.8) can be estimated by the size of a sample $\boldsymbol{n}_{1:t-1} = \{\boldsymbol{n}_{g,1:t-1}|\forall g\}$ and the response numbers in all modes $\boldsymbol{r}_{1:t-1} = \{\boldsymbol{r}_{g,1:t-1,j}|\forall g,j\}$. Under the estimated model, predictions can be obtained on dependent propensities $\widetilde{\boldsymbol{\rho}}_t = \{\tilde{\rho}_{g,t,j}|\forall g,j\}$ and $\boldsymbol{r}_t = \{\boldsymbol{r}_{g,tj}|\forall g,j\}$, given data $\boldsymbol{n}_t = \{n_{g,t}|\forall g\}$. Since the posterior distributions for $\widetilde{\boldsymbol{\rho}}_t$, as well as the posterior predictive distributions for $\boldsymbol{r_t}$, have no closed form, we resort to Markov Chain Monte Carlo (MCMC) simulation techniques, in particular, the Gibbs sampler. See Wu et al., (2022) and Boonstra & van den Brakel, (2022) for more details on the Gibbs sampler used, including the full conditionals and Markov chains.

2. *Determine optimal allocations*. Specify budget level h and overrun level α. Set multiple starting vectors of stratum allocations $\boldsymbol{s}$, each vector viewed as an initial state and each having a finite number of well-defined successive states. For any stratum g, assume K iterations of estimates of $\boldsymbol{r}_{g,t} = \{r_{g,t,m}|\forall j\}$ and $\widetilde{\boldsymbol{\rho}}_{g,t} = \{\tilde{\rho}_{g,t,j}|\forall j\}$ generated from the posteriors in **1**. These posterior estimates and given parameters h and α are separately substituted into (4.14) and (4.15) to compute the posterior expectation $\hat{E}\big(CV(s,t)\big)$ and the posterior probability of workload excess. To detect the optima, starting from each initial state, such a computation proceeds through its successive states, produces output, and eventually terminates at the final state. Discard constraint-violated states and their output, and preserve constraint-met states and their output. Within these results, sum the minimum of $\hat{E}\big(CV(s,t)\big)$ and its corresponding allocations optima.

Solving this mathematical program is a computationally intensive task. Therefore, the methods in step **1** are implemented in R using the *mcmcsae* package (Boonstra, 2021), while the methods in step **2** are implemented in R using the *auglag* (Augmented Lagrangian Minimization Algorithm) function of the *Alabama* package (Varadhan, 2015).

4.3.4 Performance evaluation

This chapter introduces an evaluation criterion to assess the prediction accuracy. The criterion can shed light on the gain in nonresponse risk reduction from different models or

from different survey designs. This gain is quantified by the root mean square error (RMSE) of the posterior distribution of a parameter τ, e.g., response propensity or CV, relative to the "true", with the latter estimated via observations.

We consider performance over rolling windows of three months. This choice is motivated by the three-month fieldwork duration of the application in this chapter but can be changed to any length. In time window $q = \{t, t+1, t+2 | \forall t\}$, the RMSE of the gth stratum is then defined as

$$RMSE(\tau, q) = \sqrt{B^2(\tau, q) + SD^2(\tau q)}, \tag{4.16}$$

where the first term is called the bias term, represented as the quadratic difference between the posterior mean of estimated CV and the observed CV,

$$B^2(q) = \sum_g d_{g,q} \left(E_{\pi_q} CV(s_q, q) - \widehat{CV}(s_q, q) \right)^2. \tag{4.17}$$

and the second term is the posterior variance of CV, which is a quadratic form of the standard deviation (SD),

$$SD^2(q) = \sum_g d_{g,q} \, Var_{\pi_q} CV(s_q, q). \tag{4.18}$$

Weight $d_{g,q} = \frac{n_{g,q}}{\sum_g n_{g,q}}$ is the ratio of the stratum g size to the sample size in window q. The posterior distributions π_q of CV and allocation $\boldsymbol{s}_q$ are derived from the computing strategy in Chapter 4.3.3.

These criteria depend strongly on sample size sampling variation, especially for surveys with small sample sizes. Empirical data subject to sampling variation are used to evaluate the performance. While surveys with large sample sizes provide rich information and thus their performance can be evaluated precisely, for small surveys, a contradiction to time change becomes acute, i.e., they take longer to make a precise evaluation. Noisy criteria performance makes it more difficult to draw a sound conclusion about putting the adaptation into practice.

4.4 The Dutch Health Survey Case Study

This chapter explores and exploits the application of multinomial time-series models in Chapter 4.2 and the optimization approach in a Bayesian framework in Chapter 4.3 to the Dutch Health Survey (GEZO for short). Chapter 4.4.1 briefly introduces the background of GEZO. We illustrate how time changes in sequential propensities can be modeled, how the performance of optimal allocations depends on the budget level, and how optimal decisions depend on the length of the historic data separately in the following.

4.4.1 The Dutch Health Survey

The GEZO survey is conducted annually by Statistics Netherlands, providing a thorough overview of developments in medical contacts, lifestyle, health, and preventative behavior of the Dutch population, including all individuals living in private households. A two-stage sampling frame forms the sample, which first draws a sample from municipalities and next from people who live in the selected municipalities, selected with equal probabilities. The survey changed to a mixed-mode design after 2014. The observation method involves online and face-to-face interviews. First, computer-assisted web interviewing (CAWI) is used to request the participation of sample units from the population. Next, nonrespondents are recruited to participate in a computer-assisted personal interview (CAPI). As of 2018, however, adaptation is adopted to stabilize the interviewers' workload. Only a portion of CAWI nonrespondents is reapproached for a CAPI to reduce survey costs and improve the representativeness. Higher response rates in CAWI sample units lead to a smaller chance they are reapproached.

In this chapter, we focus on a time series of data collected from 2014 to 2017, involving 48 months. Note that data collected early in 2017 were "abnormal" because of technical issues with the web server, resulting in an interruption in data collection. This comes with practical reasons: Statistics Netherlands has implemented static ASDs since 2018, and the adaptation may waste the potential value of historic data used to improve prediction accuracy (Wu et al., 2023).

Additionally, sample units are stratified into 13 disjoint strata by two auxiliary variables from the administrative frame or registers: age and ethnicity. See the stratification in

Appendix G. Note that this stratification is fixed throughout this chapter, and our time-series strata are different from the ASD strata (See van Berkel et al., 2020).

4.4.2 How can a time-series model be constructed under a sequential mixed-mode design?

This chapter elaborates the approach to estimate sequential propensities accurately by building multilevel time-series models using model component candidates. In this chapter, three levels are the most relevant: strata, time and mode. They formulate different model components; consequently, the time-series model in (4.8) can differ in a combination of components. Additionally, the model specifications can be classified into four scenarios: with/without season and with/without correlations of propensities between modes. The model with seasonal inclusiveness considers seasonality of great significance as a strong predictor contributing to reliable propensity predictions; otherwise, seasonality is incidental to reliable predictions for no-season models. When the correlation between CAWI and CAPI in propensity predictions is considered, each component in the model describes the between-mode correlation effect on variation in predictions; by contrast, CAPI propensities are independent of CAWI propensities when the model ignores the between-mode correlation effects.

Therefore, this research question is a matter of comparing models under each scenario to select the "favorite" combination of model components, and of comparing the preferred models among scenarios. In the first comparison, the "favorite" model of each scenario can best fit the sequential response data and make the most reliable predictions. The second comparison focuses on the performance of the model considering seasonality and correlations relative to that of the models without either seasonality or correlations. Thus, information criteria (ICs) are applied to measure those models' performance. The higher the ICs of the model are, the more competent the model is. Additionally, since it is meaningless to try all combinations of components, we apply the same strategy proposed in Wu et al., (2023) for an efficient comparison. The advanced strategy herein fits the models to 2014-2017 mixed-mode response data such that simultaneous evaluations can be undertaken for different scenarios:

1. Set up two baseline models and start with each of them. The common model components are fixed mode effects and fixed effects of mode-specific auxiliary variables β. The difference between the two models is whether seasonality δ is included.
2. Add a single random effect, i.e., {global time trend u_t}, {random intercepts for strata v_g}, or {white noise e_{gt}}, to models in **1**. Each random effect is correlated with modes or independent of modes. Examine whether the no-season or season-inclusive model in **1** is enhanced by each of the three random effects.
3. Add a combination of two random effects, i.e., {global time trend u_t, random intercepts for strata v_g} or {random intercepts for strata v_g, stratum-specific time trend z_{gt}}, to the models in **1**. Each random effect is correlated with modes or independent of modes. Examine whether the updated models outperform the models in **2**.
4. Add a combination of three random effects, i.e., $\{u_t, v_g, e_{gt}\}$, or $\{u_t, v_g, z_{gt}\}$, to the models in **1**. Each random effect is correlated with modes or independent of modes. Examine whether the updated models outperform the models in **3**.
5. Add all random effects to models in **1**. Each random effect is correlated with modes or independent of modes. Examine whether the complete combination makes the model performance best.

As seen in each row of Table 4.1, the models with and without seasonality are evenly matched at fitting and predicting. The IC results of the two baseline models (Model 1) show that the with-season model performs slightly better than the no-season model. This advantage continues with the addition of some random effects (see Models 1, 3 and 6), as the inclusion of seasonality δ yields lower ICs. The opposite of the with-season models having slightly worse performance can be seen in Models 2, 4, 7, 8, and 9. Notably, the results of Model 5 show mixed traits. The DIC results favor modeling seasonal effects in accurate propensity predictions, but WAIC cannot conform to this.

Concerning the balance between model complexity and model fitness, the mode-independent models perform barely as well as the mode-correlated models, even though they slightly outperform (the no-season model of Row M5 and Row M6).

Table 4.1 ICs and effective number when evaluating the model fit and complexity. Fit the 2014-2017 data to a mode-independent model ("IND") and a mode-correlated model ("COR"). Each model starts with fixed effect only and then complicate with random effects added.

Model	Fixed effect	Random effect	DIC		p_{DIC}		WAIC		p_{WAIC}	
			IND	*COR*	*IND*	*COR*	*IND*	*COR*	*IND*	*COR*
1	β	-	6808		18		6817		28	
	β,δ		6800		25		6813		38	
2	β	u_t	6504	6504	67	67	6509	6509	72	72
	β,δ		6506	6505	70	70	6510	6510	74	74
3	β	v_g	6775	6774	26	26	6786	6785	36	36
	β,δ		6768	6767	32	32	6781	6780	46	45
4	β	e_{gt}	6602	6601	322	321	6479	6479	200	200
	β,δ		6609	6608	322	321	6488	6487	201	200
5	β	u_t,v_g	6488	6490	147	146	6472	6475	131	130
	β,δ		6474	6472	77	77	6477	6476	80	81
6	β,δ	v_g,z_{gt}	6494	6495	143	144	6481	6482	130	129
	β,δ		6488	6490	147	146	6472	6475	131	130
7	β	u_t, v_g,e_{gt}	6449	6448	195	194	6398	6396	143	142
	β,δ		6453	6452	191	191	6404	6402	142	142
8	β	u_t, v_g,z_{gt}	6396	6395	109	108	6381	6380	94	94
	β,δ		6398	6397	112	111	6382	6381	96	95
9	β	u_t, v_g, z_{gt},e_{gt}	6398	6398	137	137	6371	6371	110	110
	β,δ		6400	6397	132	134	6375	6372	108	108

With random effects considered, the mixed models become better because they cause a decrease in ICs, in contrast to the models including fixed effects only (Model 1). Comparing Model 2-4 to Model 1 entails either the no-season or with-season model is improved by introducing a single random effect, where global time trend u_t induces the greatest decrease in ICs, followed by white noise e_{gt} and a random intercept for strata. Such improvement persists in ICs when applying the combinations of two random effects, as indicated by the comparisons of Model 5 to Models 2 and 3 and Model 6 to Model 3. Apparently, Model 5 has the most significant decrease in ICs thus far. Models 7 and 8 show that the models can be enhanced further with the addition of white noise e_{gt} and stratum-specific time trend z_{gt} to Model 5, and Model 8 makes ICs decrease more than does Model 7. Including white noise e_{gt} is of value to improved performance, as it adds little in lowering the WAIC of Model 9 despite the scarce contribution made to DIC.

As Model 9 shows, the mode-correlated and mode-independent models (Columns COR and IND), when seasonality is overlooked, perform similarly in terms of ICs, yet for the with-season models, modeling correlations (Column COR) come first in IC scores relative to the IND column. However, it is difficult to conclude that the with-season model has an absolute advantage over the no-season model in terms of model fitness and complexity. To identify whether seasonal effects play a vital role in adaptive allocations, we consider both the no-season and with-season models (Row M9 correlated with modes) in Chapter 4.4.4.

4.4.3 How sensitive is ASD performance to the specified budget level?

This research question is concerned with how, given a budget level, we adapt allocations for CAWI nonresponses across strata to lower the risk of nonresponse the most. It also raises the question of whether such a reduction can be sustained across different budget levels.

We answer this question by first minimizing (4.14) subject to (4.13) and (4.15) for the next data collection quarter when the budget level is specific, then by comparing the optimum (4.13) to the realized CV under the same budget level, and finally by comparing the optimum (4.13) under different budget levels. We focus on the next quarter because in the static case, the number of CAWI respondents is unknown until data are collected and

because the sufficient sample of a quarter can ensure the prediction precision. Referring to the optimization strategy in Chapter 4.3.3, the evaluation procedure in quarter q is

1. Let budget level h begin at 100% and then successively decrease in steps of 10%, i.e., $h \in \{1,0.9,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1\}$.
2. Identify the forthcoming quarter q and set data in q as the test data set.
3. Set time-series data up to quarter $q-1$ as the training data set for estimating the selected models. The models are made of the components in Model 9 that are viewed as the "best" representation.
4. Use the sample size in q to simulate CAWI responses. For each stratum and each month within q, 3,000 draws are generated from posterior predictive distributions.
5. Based on the simulated model in **3**, individual posterior predictions of CAWI and conditional CAPI are generated separately 3,000 times for each stratum and each month in q.
6. Substitute the specified budget level h and the CAWI responses simulated in **4** into cost constraint (4.15).
7. Compute mixed propensities by substituting level h and individual predictions in **5** into (4.11).
8. Initialize three starting solutions of allocations probabilities, $s \in \{0,0.5,1\}$, each of which applies to 13 strata simultaneously.
9. Start from each initial point in **8** to find the optimal solutions for each stratum by solver *auglag* based on **6** and **7**.
10. Link the identified solutions to the actual sample for computing posterior CV predictions and CV realizations.
11. Conduct comparison by repeating **2**-**10** for each budget level in **1**.

Note that $s = 0$ indicates no CAPI follow-up, 0.5 means half of CAWI nonresponses are assigned to CAPI, and 1 represents full CAPI follow-up. To distinguish different mode strategies and ease notation, the CAWI-only, nonadaptive, and adaptive designs are denoted w, w-Ap, and w-Np throughout.

In Figure 4.1, the posterior CV predictions of 2017 Q1 are summarized for each budget level. See Table G.2 in Appendix G for the bias-adjusted CV results. We benchmark the w-Ap performance as a function of budget level h against the performance of w and w-Np. For brevity, CVs for the CAWI-only, nonadaptive, and adaptive are simplified to CV(w), CV(w-Np) and CV(w-Ap).

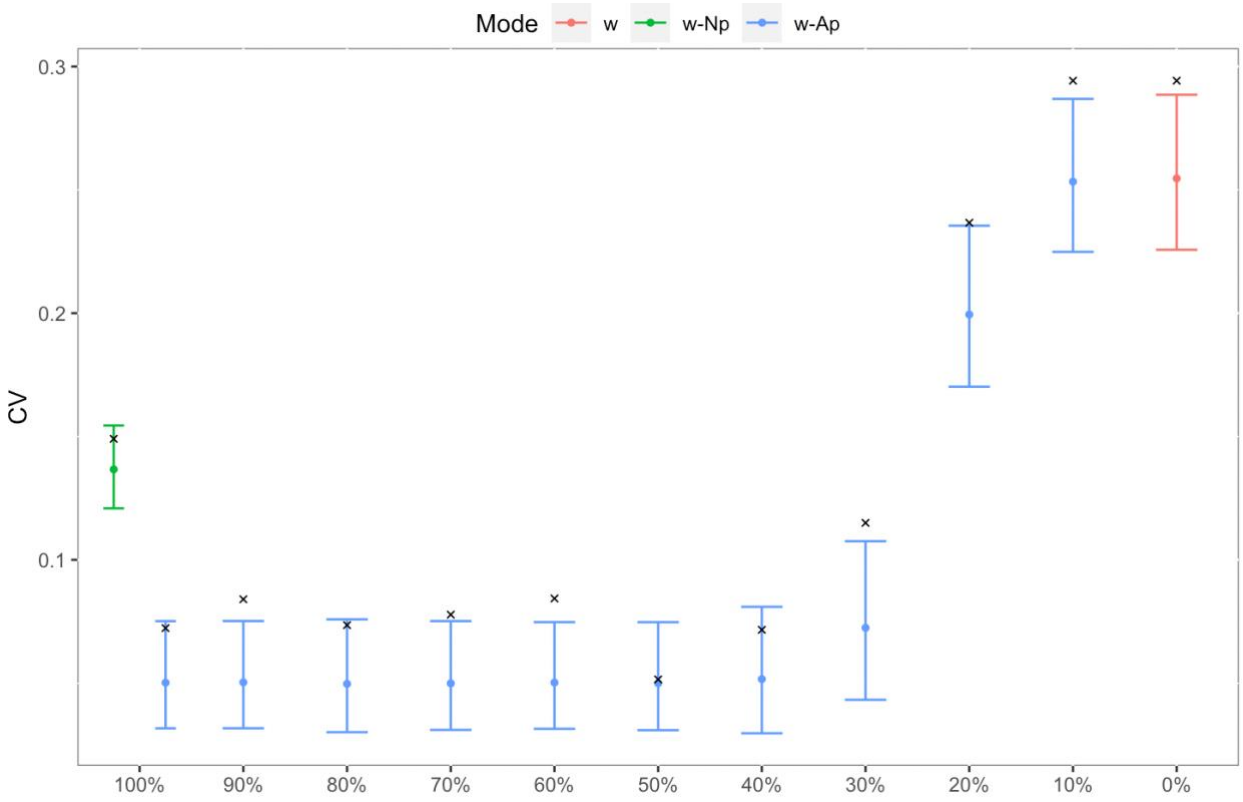

Figure 4.1 Comparison of CVs of model-based RP predictions to bias-adjusted CV observations in 2017 Q1. The CV estimates are made separately for CAWI-only ("w"), nonadaptive ("w-Np") and adaptive ("w-Ap"). The posterior CV predictions are summarized by the 95% credible region, while the observations are marked by scatter points ("x").

Comparison of CV(w-Np) to CV(w) indicates that recruiting CAWI-nonresponses via CAPI can decrease nonresponse, as the 95% credible region (CI) of posterior CV(w-Np) is much narrower than that of CV(w), and the 97.5% quantile of posterior CV(w-Np) is far below the 2.5% quantile of CV(w). When h = 100%, a further decrease in the overall variation can be achieved by the optimized allocations of the adaptive survey. Posterior predictions and observations of CV(w-Ap) deviate from 0.1 and move toward 0 relative to

the 2.5% quantile of CV(w-Np), yet the broader CI for the adaptive approach indicates that prediction accuracy is compromised moderately. Because CIs scarcely alter when the budget is cut from 90% to 50%, the uncertainty reduction associated with CV(w-Ap) is unlikely to increase more than 100%. This implies that in the interval of levels 100% to 50%, the low budget performs as well on the estimated nonresponse risk as does the high budget. The upper limits of CIs appear to approach and even run beyond the observed CVs; for instance, at the 50% level, the observation overlaps with the posterior mean.

The nonresponse risk rises with continued shrinkage of the budget since the estimates of CV(w-Ap) increase and point to an increased risk of nonresponse bias. For budget levels smaller than 50%, the allocation scheme identified puts more uncertainty on the posterior estimates of overall variation. In addition, the lower limits move toward and even far beyond 0.1 when the level is 20% or 10%, for which the solver ends up with false local "optimum" due to the violated convergence criteria. For the 10% level, the allocation scheme is especially of less interest and loses its edge, as seen by the exact same CV(w-Ap) as CV(w). To determine which budget level is preferred most, we adopt a criterion, i.e., the relative cost defined as the overall cost of the adaptive size for CAPI relative to the nonadaptive size constrained by the budget level. See Table G.4 for the results of the relative cost under different levels in Appendix G.

Optimized reallocations make adaptation performance consistent across relatively large budget levels (100%-50%). Additionally, adaptation, although it loses precision slightly, wins nevertheless at the smaller estimated nonresponse risk compared to the w and w-Np designs (red and green error bars). Up to a 40% budget level, performance reverses and moves in the opposite way, implying that the nonresponse risk grows sharply.

4.4.4 How does the performance of adaptive designs depend on the length of historic data?

This question is a matter of examining how the accumulating historic time series influences the adaptive design performance, that is, the nonresponse risk measured by CV and the bias-variance balance measured by RMSE. To answer this question, we explore the performance of w, w-Np, and w-Ap designs at the calendar quarter level. Additionally, we benchmark the adaptive performance against the performance of w and w-Np.

We compare and evaluate the models with/without the inclusion of seasonality and budget levels of 50% and 30%. Chapter 4.4.2 hints that seasonality is an ignorable factor since the models with and without this score have similar fitness and complexity. Chapter 4.4.3 implies that in a specific time window, budget level 50% promotes ASD performance most cost-effectively, and the ASD loses its absolute advantage for smaller values. The performance's sensitivity to the time-series length is less clear if the models consider seasonality and/or the budget level is less than 50%, so it is premature to skip them in the analysis. By crossing the two conditions, comparisons can be made simultaneously in the four scenarios of the models: (1) with the inclusion of seasonality and level 50%, (2) with the inclusion of seasonality and level 30%, (3) without the inclusion of seasonality and level 50%, and (4) without the inclusion of seasonality and level 30%.

To explore the sensitivity to the historic time-series length, the analysis is performed on a rolling basis by adding one month at a time. Recall that the initial historic time-series length should be at least one year for the models without the inclusion of seasonality (scenarios 3 and 4) but at least two years for the models with the inclusion of seasonality (scenarios 1 and 2). For each, the training process ends in 2017 Q3 because one quarter should be left for prediction.

In Figure 4.2, the uncertainty about the estimated CVs that is assessed by the 95% credible region, and the posterior means together, are compared to the CV observations over quarters and between different designs. In the w and w-Np designs, the observed CVs fall within the intervals or are very close to the intervals' limits in most quarters, with the exception of CV(w) at 2017 Q3. The ASD results in panels A1, A2, and B2 support this finding. Additionally, observations far outside of the CIs appear in 2015 Q2 of Panel B2 and 2017 Q1 of Panels A2 and B2. The exception implies that in corresponding quarters, it is less convinced of the evaluated adaptive performance duplicating the performance in practice.

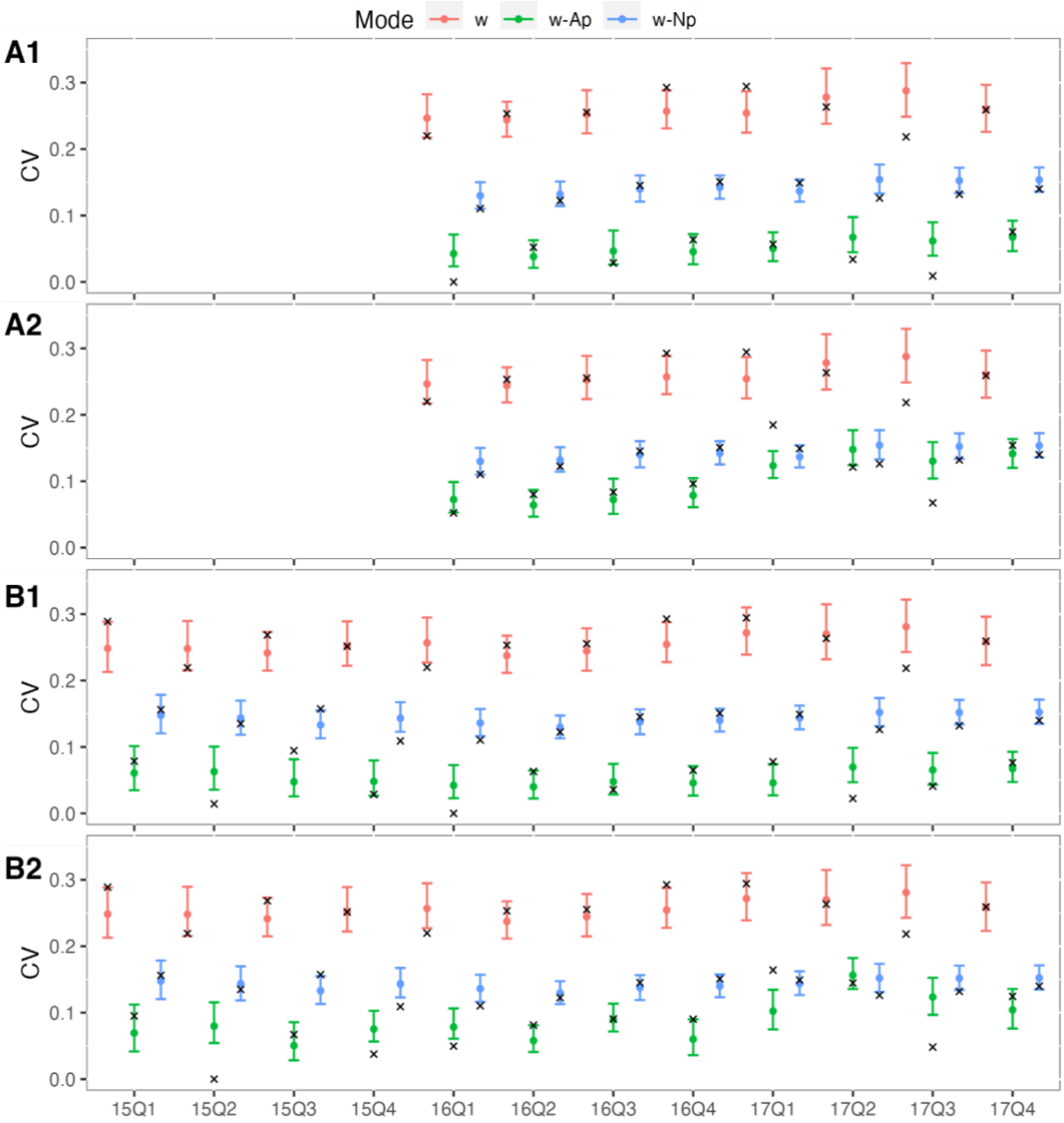

Figure 4.2 Under a given budget level, the posterior CVs for the adaptive, nonadaptive, and CAWI-only against the observations over quarters. 95% credible regions with posterior expectations are summarized for CAWI-only (w), nonadaptive (w-p), and adaptive (w-Ap). Observations are denoted by the black cross "x" points. Panels "A" are plotted for with-season models, and panels "B" are plotted for no-season models. Panels "A1" and "B1" correspond to budget level 50%, while panels "A2" and "B2" correspond to budget level 30%. The quarter on the x-axis denotes the present quarter for prediction purposes.

As mentioned before in Chapter 4.4.3, the performance for adaptive designs under level 50% is consistently superior to the performance under level 30% across quarters, as shown

by comparing A1 to A2 or B1 to B2. At 50%, the estimated CV(w-Ap) is more precise because of the narrower credible regions, which can be seen implicitly. Moreover, we can observe the absolute advantage of adaptive designs in outperforming nonadaptive designs since the upper limits of CV(w-Ap) deviate substantially from the lower limits of CV(w-Np), but at 30%, they are competitive, for example, 2017 Q1 and Q2 (Panels A2 and B2).

As historic data accumulate, it is conjectured that the models can be optimized further, the CV prediction accuracy can show a consistent increase, and the resulting performance can be improved. Clearly, this is the case in panel B1 until 2016Q4, but from that point onwards, there appears to be no room for improvement in performance, and the posterior estimates of CV(w-Ap) shift to approximately 0.1. The results, similar to those in panels A1 and B1, since 2016 Q1 signify that modeling seasonality contributes little to prediction accuracy.

Jumping to conclusions on the ASD performance's robustness is dogmatic for two reasons. First, some strata may benefit more than others. Their individual CV estimates may be less biased and vary toward the seasonality-inclusive option despite little difference in the overall variation. Second, the sample sizes in some strata are quite small. In early data collection phases, they may have volatile behavior in the error-variation balance. The historic time-series length determined based on those results is not a guarantee of robustness.

Therefore, we evaluate individual strata performance measured according to the criteria in Chapter 4.3.4. As above, we apply the sliding time window approach moving forward on a time series to the evaluation. To illustrate, this is used in nonadaptive designs. With an application to ASDs, allocations must be reoptimized for the upcoming time window using the strategy in Chapter 4.3.3.

Provided that the number of historic time periods within a time window is called the window width, the time window frames time-series data with width t, split into a training data set with width $t-3$ and a test data set with width 3. The latter corresponds to a window q. The window slides as the width are increased to include the next upcoming new time period such that the window moves one step forward. In quarter q, we can evaluate the prediction performance for each stratum by substituting individual posterior

response propensity estimates and individual realizations into (4.16) - (4.18), that is, $RMSE(g,q)$, $B(g,q)$ and $SD(g,q)$. Since this analysis is iterated on a rolling basis, a sufficiently long time series allows for thorough comprehension of how each stratum prediction performance changes with time.

Figure 4.3 shows that when comparing red and black curves, the introduction of seasonality is unlikely to be a trigger for an effective reduction in bias and variance. This is solidly true for almost all quarters, with the exception of the quarter involving months 2017-01 to 2017-03 in some strata (such as stratum 8) for the bias and RMSE estimates. As observed in panels a and b, the estimated variation in response propensity decreases smoothly overall, in sharp contrast to the estimated level of response propensity having volatile behavior. The volatility differs by strata. The estimated bias results of some strata (strata 1-8) fluctuate approximately 0.05 across quarters until the quarter starting in 2017-01. After that point, they experience a transient increase caused by the technical issue at that time (See Chapter 3 for more discussion and a possible remedy). When the training data are extended to include "normal" data, the biases can quickly decrease to 0.05. Note that stratum 8 acts in the opposite manner. In contrast, strata 9-13, which have relatively small sample sizes, obtain relatively more biased response propensity expectation estimates in most quarters.

Ultimately, the analysis results suggest that when modeling a short time series, the seasonal effects, when they are assumed to be the same for different modes, can be less important to the improvement of ASD performance. With more data available for training, the ASD performance can be consistently improved until a time point, implying that a stopping rule of data collection may be implemented and an effort-based strategy for strata of small sample sizes may be adopted.

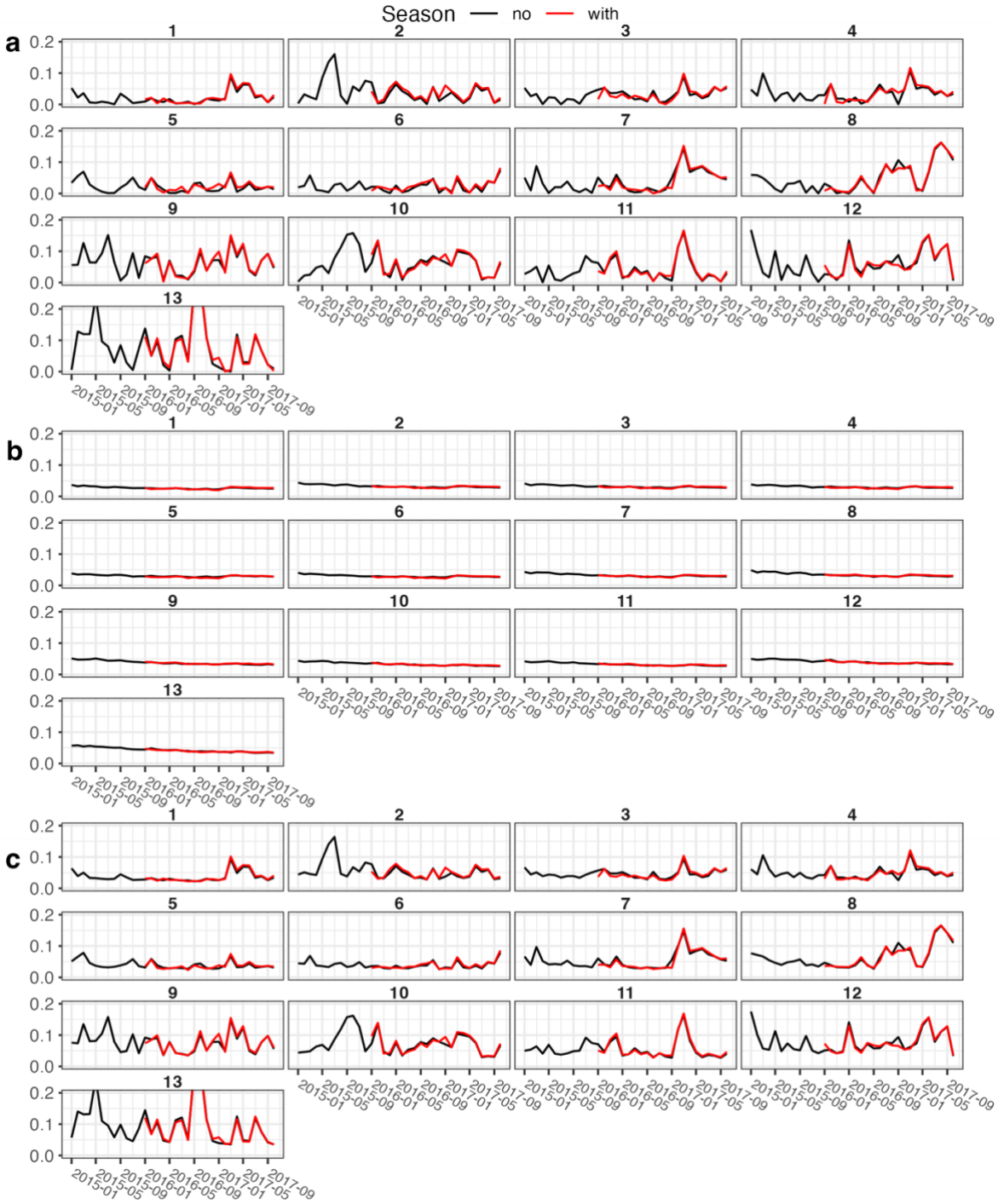

Figure 4.3 One-step forward moving averages of estimated bias (panel a), SD (panel b), and RMSE (panel c) of **r**esponse propensity at the stratum level. The "black" curve represents to the no-season model, while the "red" curve refers to the with-season model. Both models include the correlations between CAWI and CAPI regarding propensity predictions. The X-axis labels the first time point in the artificial quarter used for prediction.

4.5 Discussion

Given the survey budget, adaptive survey design (ASD) seeks the optimal match between respondent behavior and design features, i.e., a set of decision rules, which can be determined through optimization approaches (See Schouten et al., 2017 for each approach of pro and cons). Serving as the main input for ASD optimization, accurate estimates of survey design parameters, such as response propensities, are required for reliable strategies. Strictly speaking, inaccuracy jeopardizes ASD performance and design due to the suboptimal and ineffective decisions made in the optimization approach. Adverse impacts are apparent when response propensities change gradually over time.

In this chapter, we discuss a methodology to evaluate the impact of temporal factors (e.g., seasonality) on the accuracy of sequential response propensity predictions in a mixed-mode survey with replication, and to investigate the manner and timeliness of applying the optimal allocation scheme to population strata. We introduced a Bayesian multinomial time-series model for sequential response propensities and an optimization model for ASDs. The propensity model has a general form that describes multiple time-related and strata-related factors, and accounts for the dependence of the current mode's response propensities on the preceding modes' response behaviors. The optimization model, on the other hand, enables the inclusion of uncertainty in the follow-up workload, and describes the way to allocate reviewers to each stratum for the greatest decrease in nonresponse risk. Most cross-sectional mixed-mode or unique-mode surveys conducted over many years can fit into this framework. Furthermore, we constructed an analysis for the GEZO survey to examine the highest performance of the propensity model. Owing to diverse model compositions, information criteria measuring the fitness and complexity of the propensity model are adopted to compare the performance of different models. We are thus able to meet the first objective of this chapter to select and construct the "favorite" time-series model (Model 9 having lowest information criteria scores) that contributes most to prediction accuracy for a sequential mixed-mode survey.

The second and third objectives are to examine the sensitivity of ASD performance to the specified budget level and the length of historic data separately. In the evaluation, ASD performance must be reoptimized when the budget level and/or the length of historic data

is updated. Then, we benchmarked ASD performance against CAWI-only and nonadaptive design performance. This analysis is essentially a comparison of the reduction in the nonresponse risk if a fraction of CAWI nonresponses (with no follow-up at all and full follow-up as special cases) is assigned to interviewers. To make this comparable for a range of scenarios, we utilize the posterior summary of the nonresponse of variation, that is, the credible region and expectation. The evaluation examines, in a specific time window, the improvement in performance under different budget levels. Additionally, the evaluation examines, for a specific budget level, the improvement over rolling time windows. The evaluation shows that ASD performance is quite solid for budget levels greater than 50%, but is susceptible to budget levels less than 50%. Additionally, the evaluation shows that without taking seasonality into account, ASD performance is obviously enhanced in the early stage of accumulating data. After that, this trend slows and even stops despite increasing, performing almost the same as the with-season model, and consequently hints at seasonality being of little use to further improvement in prediction accuracy and ASD performance.

Our models have two conceptual simplifications and one limitation that ask for further research and replication in other mixed-mode surveys.

Ignoring mode-specific seasonal effects is our first simplification. This consideration can ease the complexity of the model specification but leads to seemingly offset seasonal effects on propensity predictions. However, we believe that one can conveniently accommodate seasonal effects specific to each model to the adjusted model if seasonality is supposed to be a strong predictor for propensity predictions. To ensure prediction accuracy and reliability, we consider only two candidate data collection strategies (CAWI and CAPI) as the second simplification.

Clearly, the number of CAPI visits to sample units may be further tailored, and the optimization may include the actual number of visits. Response propensities after each visit can be modeled and estimated simultaneously, and the predictions of a follow-up mode are correlated only with its nearest predecessor. Such an application is easy, but it entails careful checking of the predictions' reliability.

Finally, as seen in the GEZO case study, our propensity model is sensitive to structural design changes (See Wu et al., 2023 for more discussion) that result in a temporary misspecification in the prior distributions of response propensities. The passive impact is fairly obvious for the first several data collection waves since 2018, and may become weak but at the cost of a long data collection process. Fortunately, our model possesses some robustness when faced with unexpected/outlier occurrences, such as temporary downtime of a web server. Without careful consideration of misspecification when modeling, abundant historic data would lose its value in providing informative knowledge about time changes in response propensities. In an effort to characterize abrupt change, it is of first importance to pinpoint the affected strata that have changed the most in terms of what features describe the change and with what magnitude the change occurs. The types, such as time duration, persist with changes, and the effects of structural changes across strata can be detected by noncontaminated models (i.e., present "optimal" models without any modification). Alteration can be moderated by extra hierarchical model parameters to represent the moving averages of changes in bias and/or variation estimates. Design changes, therefore, can update outdated models for timely and accurate predictions. We leave this extension to further research.

This chapter briefly touches on the construct of adapting optimal allocations to population strata under a Bayesian analysis. This application is meant for a static survey in which the population is stratified based on only auxiliary data before the start of data collection. In a dynamic environment, however, both the stratification and assignment of interviewers to strata hinge additionally on intermediate survey data in the present mode's predecessor. The correlations between response propensities are then also employed to update the prior distributions for the upcoming mode response propensities. This would allow, for example, for the inclusion of paradata-type auxiliary information in choosing strata. In a face-to-face interviewer setting, such a dynamic approach is not operationally straightforward, as interviewer workloads become known only at a point in time close to fieldwork. To a lesser extent, this is true for telephone follow-up, but there is no geographical clustering needed to limit travel times and costs. A possible approach would be to randomly subsample nonrespondents to fixed workloads. In the GEZO survey, the monthly sample sizes are too small to make robust decisions after each month, which is why months were pooled to quarters. The pooling further complicates dynamic allocation.

Nonetheless, abstracting to general mixed-mode designs and, even more generally, to general (sequential) data collection phases, it would be worthwhile to extend our approach to dynamic designs.

Chapter 5

Discussion

Structured research on effective ASD optimization has emerged in recent years. Optimization strategies are the final step of ASD. They essentially translate the objective of quality-cost trade-offs into intervention and adaptation, more specifically, the optimal allocation of resources to strata. Accurate design parameters, such as response propensities, play a decisive role in the performance of adaptation strategies. Efforts have been made to improve response propensity predictions by using Bayesian analysis and incorporating historic information into the prediction process. The advance in this direction disregards the fact that response propensities can change over time, called time changes, or roughly involve its effects as a fixed amount. Research is lacking on understanding these ill-defined temporal effects and their knock-on effects on optimum allocation.

This chapter is structured as follows. In Chapter 5.1, we summarize the main findings on the reduction of the risk of nonresponse bias, designed to address the research questions. Next, we provide recommendations for future research in Chapter 5.2, based on the deficiencies identified in the current study.

5.1 Dissertation Findings

The gist of this dissertation is to, within a Bayesian framework, improve predictions of response propensities, thereby gaining effective ASDs, in an increasingly difficult survey climate. To make predictions with precision, we leverage historic survey data in a delicate way. Specifically, efforts are made to introduce expert knowledge and to consider the timeliness of historic data, when developing propensity models. The study examines how the model prediction performance is associated with each attempt, and how the estimated nonresponse indicators, as a function of the propensities, are used to intervene in and monitor the data collection process. We also examine how time alter the decision-making process. These explorations are conducted based on three Dutch general population surveys, each administered by the Statistics Netherlands. The following summarizes the main points of findings from Chapters 2, 3 and 4 separately.

Chapter 2 investigates how ASDs can profit from data collection staff (as experts) knowledge. Intrinsic expert knowledge is worth improving response propensities among

previous studies that cannot yet be fully elicited. Thus, we set up a structured elicitation procedure that translates such expert knowledge to the informative (expert) prior of the response propensities, and evaluate the benefit of this procedure. We find that compared to the settings in which there are no historic data and no expert knowledge (the non-informative prior used), the expert prior has superior performance that predicts a lower risk of nonresponse bias. This outperformance is especially noticeable in early data collection phases, but in late phases, it becomes implicit. We also find that the way knowledge, which is elicited at various criteria levels, is pooled matters slightly for the comparison result. Providing criteria with equal or unequal importance weights improve the competitiveness of the expert priors against the non-informative prior. Either expert prior can lead to a lower predicted variation in response propensities. The choice of historic surveys and experts, the selected criteria, and the criterion-level weights all influence the prediction performance of expert priors. Therefore, we advise researchers to put effort into each element.

Chapter 3 investigates time-dependent variation in predicting response propensities. This temporal dependence emerges from joint efforts of diverse factors: seasonal effects, various forms of trends, interventions and stratum dependency. Previous studies fail to determine how and to what extent each factor exerts influence on predictions. In this study, we look at all factors to gain insight into the prediction process. Our results show that the optimal model combination made of these factors performs best in both fitting to response data and in contributing to reliable predictions. In addition, clear evidence shows that using a longer time series in the estimated model can lead to a modest decrease in both the bias and the root mean square error (RMSE) in overall predictions, although this process is full of twists and turns. Rather, the estimated variance shows an increasing trend from the moment at which the design intervention is carried out. Last, the gains in terms of the timeliness of precise predictions vary among subgroups. It is significant to persons who either have Western ethnicities or are in the 18-64 age range. To conclude, this study demonstrates that explaining time-dependent variation is a promising tool for determining a well-chosen time to start implementing effective interventions to particular strata. On the other hand, little is known about such variation especially when the survey design is made adaptive. To this end, there still has much work to do on growing our

models and on being informative about the effects of design changes and their types on response propensities.

Chapter 4 investigates how adaptation strategies can adjust to time-dependent variation in response propensities. In regard to optimizing resource allocation, most researchers rely on the assumption that survey design parameters remain static from start to finish, and their estimates eventually converge on the basis of abundant historic data acquired. Addressing uncertainty, due to time change, can be a valuable addition to effective ASD optimization research. We can thus obtain thorough insight not only into what strata require interventions, but also into how cost-efficiently the limited resources are allocated among them. This enriches ASD decisions by allowing time change to be understood. Therefore, we propose a method for the timeliness of decisions-making by considering time-dependent variation. We also examine, based on optimized allocations, the sensitivity of ASD performance to the specified budget level and the length of historic time-series data separately. Our results indicate that the performance is fairly robust when the budget level decreases to 50%. Adapting optimal allocations to strata promotes a further reduction in the nonresponse risk, compared to that of the case when adjusting the uni-mode design to the full follow-up design. Additionally, we show that the performance can be enhanced in the early stage of data collection, but can hardly be improved further in the middle and late stages.

5.2 Prospects for Follow-up Research

This section focuses on two important avenues for future research on ASD — practical implementation and methodology. We hope that these suggestions will pave the way for a more detailed exploration of effective ASD in the Bayesian framework.

5.2.1 Future Research on Practical Implementation

The present dissertation provides valuable insight into the improvement in response propensity predictions and into the timeliness of decisions-making. Despite this, two things are worth further exploration before the methodology can be used widely.

Require replication. The proposed models in Chapters 2, 3 and 4 are designed for improving response propensity predictions. As shown in those chapters, these models are empirically evaluated, and their outperformance in reducing nonresponse errors is rather attractive to survey practitioners. It is, however, less persuasive to generalize this study's findings based only on a few case studies. To broaden the scope of the investigation, we strongly advise testing and examining more, i.e., extending to other survey and sampling designs. We also recommend investigating how to make the models work in practice, as they have been evaluated by using historic survey data.

Maintain monitoring of all strata. In Chapters 3 and 4, we examine how sensitive prediction and ASD performance are to the length of historic time series data. The study of time change contributes to effective decisions-making early for situations, such as determining a suitable time to stop data collection for certain strata, and timely distributing resources profitably over strata. If doing, the controversial issue that has been raised year after year is on the table once again, how do we optimize ASD and keep an eye on learning simultaneously? There is no silver bullet that allows these two conflicting things to function together. Stopping data collection from some strata after a time point implies that their auxiliary information and response outcomes may no longer be recorded in register frames. There is no guarantee that the value-added for the design performance to which time-series models contribute is future-proof, due to those missing data. Survey partitioners cannot monitor the data collection quality of those skipped strata, just the same as before. They cannot also intervene in the first place when any change happens, such as a sudden outbreak of an epidemic. This becomes a matter of great concern. We thus advise to keep monitoring strata, even if clear evidence shows that making efforts for them is unnecessary. From a practical implementation perspective, this suggestion ensures at least a baseline level of observation.

5.2.2 Future Research on ASD methodology

For the time-series models presented thus far, it holds that understanding time change in responses is crucial to timely obtain accurate predictions and effective survey designs. The models, however, have simplifications and suffer drawbacks. Each of them merits further exploration.

Mode-specific seasonal effects. Seasonality is a significant predictor for responses. In Chapter 4, we mix seasonal variation over different survey modes, i.e., the seasonal effects on CAWI responses are the same as the effects on CAPI responses. This simplification eases the complexity of the model specification, whereas these effects on response predictions seem to be offset. This has been seen by comparing the prediction performance between the models including seasonal effects and those ignoring these effects. Therefore, understanding mode-specific seasonal effects, rather than using uniform effects for all modes, is essential, especially when seasonality is supposed to be a strong predictor for responses. Therefore, we advise that one can adjust the model specification for accommodating mode-specific seasonal effects.

Adapt the number of calls and/or visits. In Chapter 4, we consider only two candidate data collection strategies, i.e., CAWI and CAPI, and allocate the CAPI strategy to CAWI nonrespondents. This simplification ensures prediction accuracy and reliability. Such adaptation is just a simplified example of resource allocation. Clearly, a data collection strategy has some attributes that we can adapt to sample units, for example, the number of CAPI visits. Some researchers study the tailoring of CAPI-short (1-3 times) and CAPI-extended ($\geq$3 times). However, we advise generalizing adaptation strategies to choosing the number of calls and/or visits, and even general interview modes such as smartphones. This further tailoring requires minor adjustments to the optimization. As an example of how to tailor the number of visits, the actual number should be included in the optimization, response propensities after each visit should be modeled and estimated simultaneously, and the predictions of a follow-up mode should be correlated only with its nearest predecessor.

Understand structural design changes. As shown in Chapter 3, a main drawback of the time-series model is its sensitivity to structural design changes. The introduction of an intervention (e.g., incentive) and/or another self-reported mode (e.g., smartphone) results in a temporary misspecification in the prior distributions of the response propensities. This negative impact is likely to fade away only at the expense of long data collection. The misspecification also makes it worthless that abundant historic data promote informative knowledge about time change in responses. Thus, explicitly including variation in responses due to design changes is called for. There is no prior knowledge for design

changes before they happen. In response to model design changes, one possibility would be to pinpoint the strata affected most, the features causing the change, and the magnitude of the change. After that, one could adjust the time-series model to suit what is detected.

Extend to multi-purpose optimization. In Chapter 4, the optimization has a single objective, i.e., minimizing the proxy indicator of nonresponse (the posterior expectation of the coefficient of variation of the response propensities). Research in survey literature mostly focuses on the decrease in nonresponse error when designing adaptive surveys. Of equal importance is, however, to decrease the measurement error. An appealing extension would be to allow the optimization to have multiple objectives, i.e., minizine the nonresponse and measurement errors simultaneously.

Learn new design features. New digital technologies such as wearable and smartphone sensors, and new types of survey incentives such as personalized feedback, have become increasingly popular and are used as new data collection methods. These new design features affect response propensities in yet unknown ways. They may potentially, however, be powerful in attracting subpopulations that otherwise would not respond. Learning these new design features is about how they change response propensities, and about how this differential may be put to practice. For example, survey practitioners can make smartphone fitness more salient for some strata in letters, or assign new types of incentives to some strata that are more reluctant to respond. These attempts are intended as new options to vary in ASD.

Appendices

Appendix A: Stratification & Questionnaire

Table A.1 Definition of the SILC16 strata across age, household size, and income deciles.

Strata	Age	Persons in household	Decile of income
1	17+	1	1
2	17+	1	2
3	17+	1	3
4	17+	1	4
5	17+	1	5
6	17+	1	6
7	17+	1	7
8	17+	1	8
9	17+	1	9
10	17+	1	10
11	17+	2+	1
12	17+	2+	2
13	17+	2+	3
14	17+	2+	4
15	17+	2+	5
16	17+	2+	6
17	17+	2+	7
18	17+	2+	8
19	17+	2+	9
20	17+	2+	10

Table A.2 Definition of the EN18 strata across categorical variables, ownership, type of dwelling, and year of construction.

Strata	Ownership	Dwelling type	Year of construction
1	buy	single family	up to and including 1930
2	buy	multiple familiy	up to and including 1930
3	social rental	single family	up to and including 1930
4	social rental	multiple familiy	up to and including 1930
5	private rental	single family	up to and including 1930
6	private rental	multiple familiy	up to and including 1930
7	buy	single family	1931-1959
8	buy	multiple familiy	1931-1959
9	social rental	single family	1931-1959
10	social rental	multiple familiy	1931-1959
11	private rental	single family	1931-1959
12	private rental	multiple familiy	1931-1959
13	buy	single family	1960-1980
14	buy	multiple familiy	1960-1980
15	social rental	single family	1960-1980
16	social rental	multiple familiy	1960-1980
17	private rental	single family	1960-1980
18	private rental	multiple familiy	1960-1980
19	buy	single family	1981-1995
20	buy	multiple familiy	1981-1995
21	social rental	single family	1981-1995
22	social rental	multiple familiy	1981-1995
23	private rental	single family	1981-1995
24	private rental	multiple familiy	1981-1995
25	buy	single family	since 1996
26	buy	multiple familiy	since 1996
27	social rental	single family	since 1996
28	social rental	multiple familiy	since 1996
29	private rental	single family	since 1996
30	private rental	multiple familiy	since 1996

Questionnaire for Judgements of Data Collection Expert.

Goal

Provide expert input for predicting response propensities for the Energy 2018 (EN18) and the SILC 2016 survey (SILC16). If an answer in the range of 0 to 1 is required, 0 basically means the historic data is not useful at all for this aspect and 1 means there is total agreement.

Historic surveys

The EN18 survey:

1. Energy 2006.
2. Energy 2012.
3. Dutch Survey on Care 2016.
4. Dutch Housing Survey 2018.

The SILC16 survey:

1. Dutch Labor Force Survey 2016.
2. Dutch Household Budget Survey 2015.

Elicited information of expert at criterion level to each historic survey

Note that * indicates the EN18 survey or the SILC16 survey

Topics

To what degree do you think the *topic* of the historic survey is comparable in terms of expected response propensities to *, specifying it in the range of 0 to 1?

Target population

How well does the *target population* of the historic survey agree with the population of *, specifying it in the range of 0 to 1?

Time

How well does the *time* (month and year) of the historic survey agree with * in terms of expected response propensities, specifying it in the range of 0 to 1?

Unit of observation

How well is the *observation unit* (person versus household) of the historic survey comparable to * in terms of expected response behavior, specifying it as 1 or 0?

Design/Mode strategy

How well is the *design/mode strategy* (including contact and reminder strategy) of the historic survey comparable to * with respect to expected response propensities, specifying it in the range of 0 to 1?

Incentive strategy

How well is the jump in the overall response rate caused by adding *incentive* to historical surveys transferred to * with respect to expected response propensities, specifying it in the range of 0 to 1?

Respondent effort/burden to complete the survey

To what extent is the required **respondent effort** of the historic survey similar to *, specifying it in the range of 0 to 1?

Bureau effect of survey data collector relative to Statistics Netherlands (CBS)

Is there a *bureau effect* between the historic survey data and the current practice of CBS, specifying it in the range of 0 to 1?

Appendix B: The Level of Response Propensities

The weighted response rate over all the strata, RR, is then defined as

$$RR = \hat{\bar{\rho}} = \sum_{g=1}^{G} \hat{\rho}_g q_g,$$

where $\hat{\rho}_g$ is the response propensity in stratum g, and $\hat{\bar{\rho}}$ is overall response rate explicit about the level of the population response propensity.

Table B.1 RMSE of the predicted RR from informative prior (Expert) with two weights and the non-informative prior (Standard) for the EN18.

Wave	**Expert**		**Standard**
	Equal	*Varying*	
1	0.231	0.237	0.160
2	0.213	0.218	0.007
3	0.198	0.202	0.005
4	0.196	0.199	0.010
5	0.186	0.189	0.010
6	0.145	0.147	0.022
7	0.132	0.134	0.023
8	0.134	0.137	0.012
9	0.125	0.126	0.012
10	0.119	0.121	0.010
11	0.096	0.097	0.028
12	0.086	0.087	0.032
13	0.093	0.094	0.020
14	0.093	0.094	0.014
15	0.089	0.090	0.014

Table B.2 RMSE of the predicted RR from informative prior (Expert) with two weights and the non-informative prior (Standard) for the SILC16 under two scenarios.

Wave	*without*			*with*		
	Expert		*Standard*	*Expert*		*Standard*
	Equal	Varying		Equal	Varying	
1	0.064	0.065	0.147	0.148	0.149	0.084
2	0.063	0.064	0.012	0.122	0.123	0.009
3	0.044	0.045	0.009	0.117	0.118	0.013

Appendix C: GEZO Survey Stratification

Table C.1 Auxiliary variables form 20 strata and season is considered as an influential factor to predict response propensities.

Auxiliary Variable	**Category**
Gender	Male
	Female
Age	Youth (<= 17)
	Young (18-34)
	Middle-aged (35-54)
	Old (55-64)
	Retired (>= 65)
Ethnicity	Western (incl. native, first and second western generation)
	Non-western (incl. first and second non-western generation)
Variable	**Category**
	Winter (January-February)
	Spring (March-May)
Season	Summer (June-August)
	Autumn (September-November)
	Christmas (December)

Appendix D: Comparison between Models

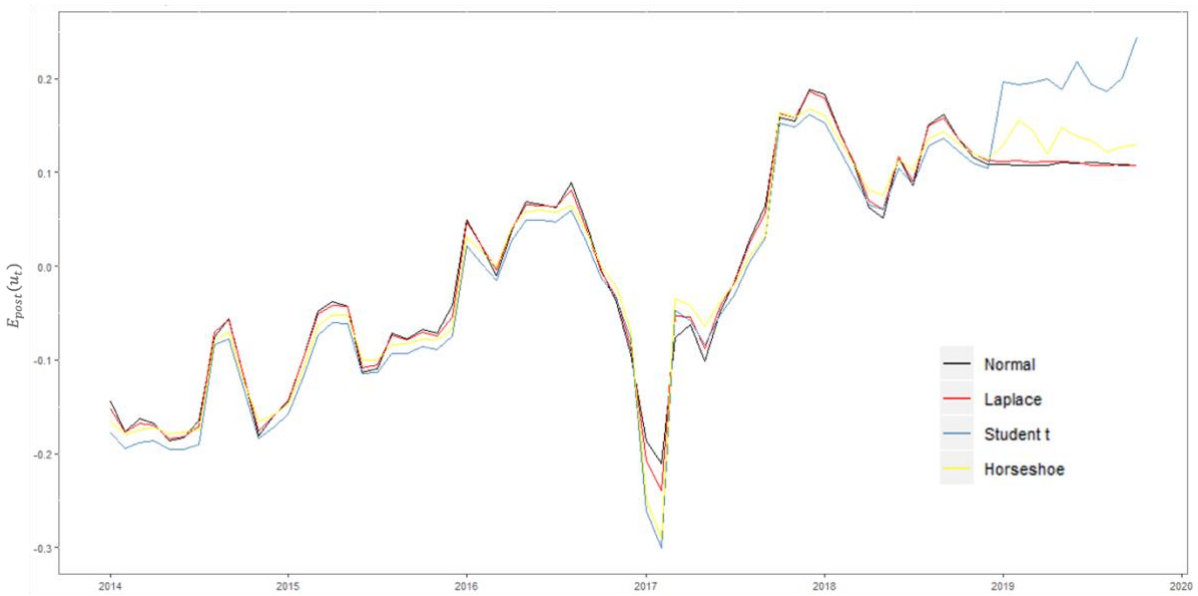

Figure D.1 The posterior means of global time trends u_t under M7 to M10.

Figure D.2 Compare the posterior predictions of RP over strata made by four models (M7 to M10) to the observed RR and make a choice on the most compatible model with the observed outliers. The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

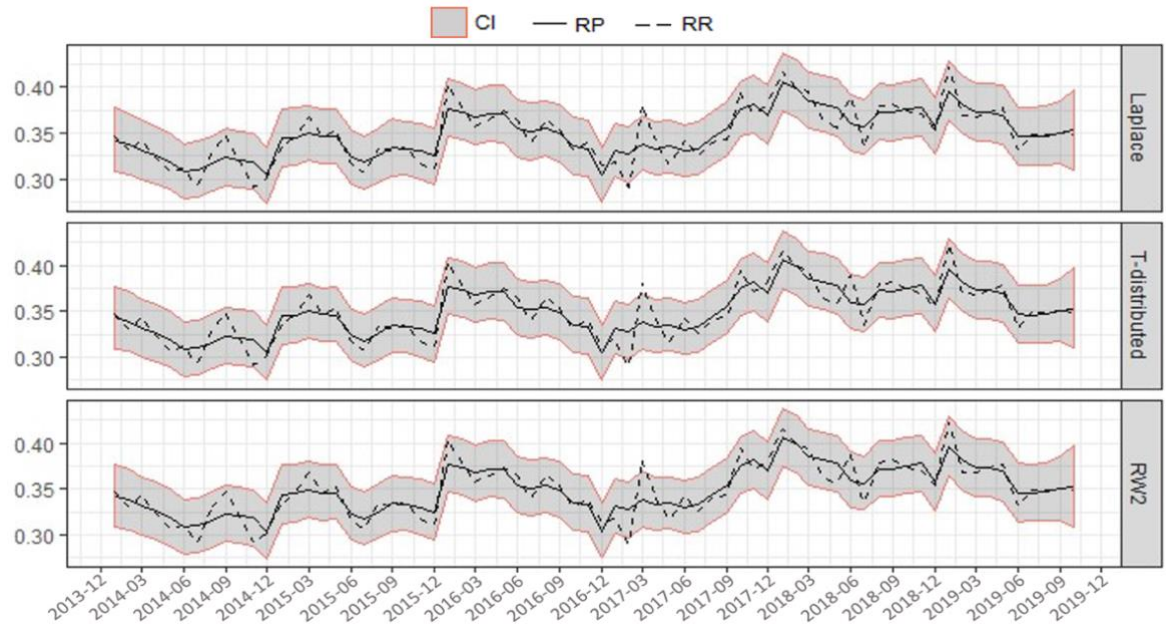

Figure D.3 The posterior predictions of RP over strata against the observed RR under M8, M9 and M11. The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

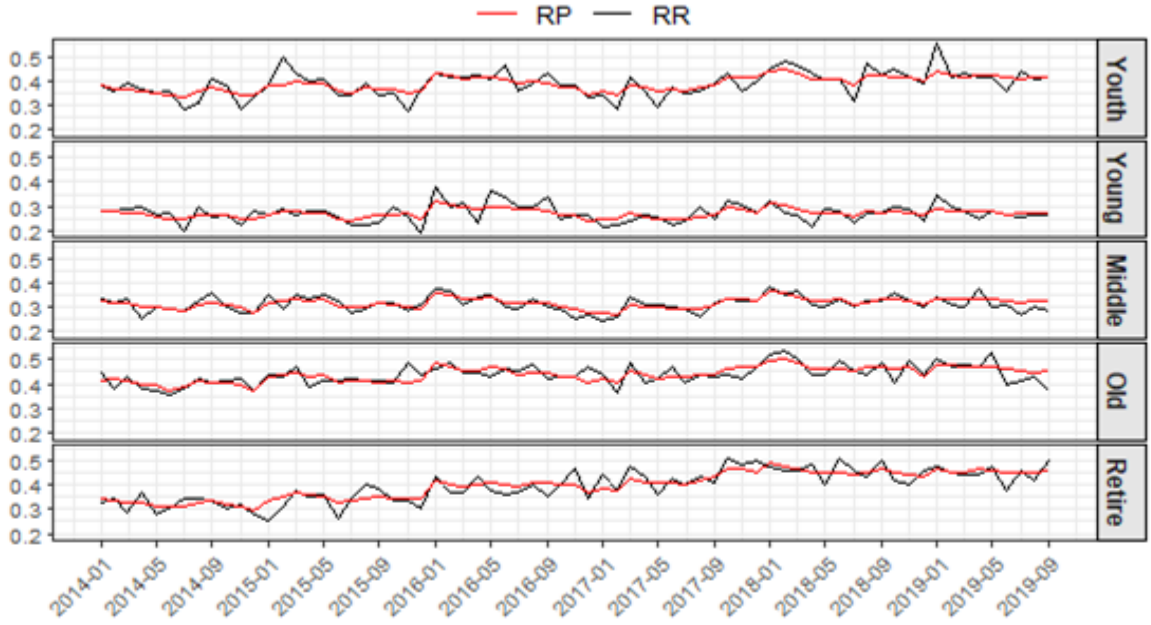

Figure D.4 Monthly posterior means of RP of Age groups versus observed response rates (RR) of Age groups. Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

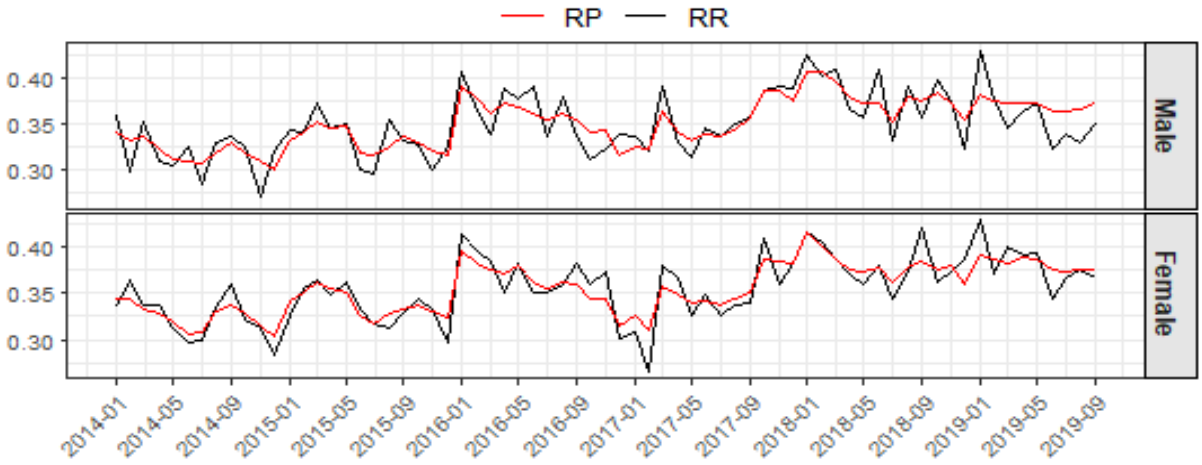

Figure D.5 Monthly posterior means of RP of Gender groups versus observed response rates (RR) of Gender groups. Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

Appendix E: Precision Matrix and Information Criteria

The precision matrix Q_u contains the neighbor structure of the trend innovations (e.g., Rue & Held, 2005). For first order random walk it is

$$\boldsymbol{Q}_u = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix},$$

and for second order random walk is

$$\boldsymbol{Q}_u = \begin{pmatrix} 1 & -2 & 1 & & & & & \\ -2 & 5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 \\ & & & & 1 & -4 & 5 & -2 \\ & & & & & 1 & -2 & 1 \end{pmatrix}.$$

The DIC is defined as

$$DIC = -2\big(\log p\big(r\big|n, E_{post}\theta\big) - p_{DIC}\big),$$

$$p_{DIC} = 2\big(\log p\big(r\big|n, E_{post}\theta\big) - E_{post}\log p(r|n,\theta)\big),$$

where $E_{post}(\theta)$ is the posterior mean of the latent parameter, $p(r|n, E_{post}(\theta))$ is the likelihood evaluated at the posterior mean of θ and p_{DIC} is an estimate of the effective number of model parameters. Models with lower DIC values are preferred.

The WAIC is defined as

$$WAIC = -2\sum_{i\in[1,GT]} log E_{post} p(r_i|n_i,\theta) + 2p_{WAIC},$$

$$p_{WAIC} = 2\big(\sum_{i\in[1,GT]}\big(\log E_{post} p(r_i|n_i,\theta) - E_{post}\log p(r_i|n_i,\theta)\big)\big).$$

Here $p(r_i|n_i,\theta)$ is the pointwise-likelihood for stratum-by-time combination i. Similar to DIC, models with lower WAIC values are preferred.

Appendix F: Full Conditional Distributions

The binomial multi-level time-series models are fit using a Gibbs sampler. For the derivation of the set of full conditional distributions we refer to the (appendix of the) technical report version of Boonstra & van den Brakel, (2022). There, the Gibbs sampler has been worked out for a general class of multilevel models, which encompasses the set of models discussed here, except for the fact that here we employ a binomial instead of Gaussian data distribution. Fortunately, the use of the scale-mixture data augmentation approach for binomial-logistic likelihoods (Polson et al., 2013) ensures that the same closed-form full conditional distributions as in the Gaussian case can be used with only minimal changes to their parameters, along with an additional full conditional distribution for the auxiliary latent scale factors. To start with the latter, the full conditional for scale factor ω_i is given by

$$p(\omega_i|r,.) = PG(\omega_i|n_i,\theta_i)$$

independently for all i. For notational simplicity we use index i instead of the double index g,t used in the main text, and r denotes the full observed response vector. Here θ_i is the linear predictor, and $PG(\omega_i|n_i,\theta_i)$ denotes the Pólya-Gamma distribution with parameters n_i and θ_i, see Polson et al., (2013). The coefficients' full conditionals change only in their parameters. For example, in the full conditional for a general random effects component, eq. (A.28) in the technical report, the precision matrix Σ^{-1} becomes $\Sigma^{-1} = diag(\omega)$ and the response vector y gets replaced by 'working response' $\frac{r-n/2}{\omega}$. The same holds true for the full conditionals of the fixed effects and auxiliary parameters ξ. All other full conditionals remain unchanged.

Appendix G: Bias-adjusted vs Unadjusted CV

Table G.1 A stratification made by auxiliary variables (Age and Ethnicity) for sequential CAWI-CAPI GEZO survey.

Strata	**Ethnicity**	**Age**
1	Western	0-17
2		18-24
3		25-34
4		35-54
5		55-64
6		65-74
7		75+
8	Non-western	0-17
9		18-24
10		25-34
11		35-54
12		55-64
13		65+

Table G.2 The bias-adjusted (adj) and unadjusted (unadj) CV observations and the standard errors (se) under bias adjustment in 2017 Q1 under different budget levels.

	w			**w-Np**			**w-Ap**		
	unadj	*adj*	*se*	*unadj*	*adj*	*se*	*unadj*	*adj*	*se*
0%*	0.305	0.294	0.023	-	-	-	-	-	-
100%	-	-	-	0.156	0.149	0.013	0.102	0.072	0.021
90%	-	-	-	-	-	-	0.093	0.084	0.021
80%	-	-	-	-	-	-	0.104	0.074	0.020
70%	-	-	-	-	-	-	0.092	0.078	0.021
60%	-	-	-	-	-	-	0.093	0.084	0.021
50%	-	-	-	-	-	-	0.106	0.052	0.020
40%	-	-	-	-	-	-	0.114	0.072	0.021
30%	-	-	-	-	-	-	0.136	0.115	0.022
20%	-	-	-	-	-	-	0.259	0.237	0.022
10%	-	-	-	-	-	-	0.305	0.294	0.023

**No budget indicates only the CAWI mode is used. "-" means no results*.

Table G.3 The bias-adjusted (adj) and unadjusted (unadj) CV observations and the standard errors (se) of bias adjustment under the 100% budget level in each quarter.

Year	Quarter	**w**			**w-Np**			**w-Ap**		
		unadj	*adj*	*se*	*unadj*	*adj*	*se*	*unadj*	*adj*	*se*
2016	**Q1**				-	-	-	-	-	-
	Q2	-	-	-						
	Q3	-	-	-	-	-	-	0.093	0.084	0.021
	Q4	-	-	-	-	-	-	0.104	0.074	0.020
2017	**Q1**	-	-	-	-	-	-	0.092	0.078	0.021
	Q2	-	-	-	-	-	-	0.093	0.084	0.021
	Q3	-	-	-	-	-	-	0.106	0.052	0.020
	Q4	-	-	-	-	-	-	0.114	0.072	0.021

Table G.4 The relative cost (c) of 2017 Q1 under different budget levels for adaptive surveys. The allocations in each case are determined by the optimization solver "auglag" starting with initial point 0 set to 13 strata. If the convergence (C) is TRUE, the local optimum can be found, and the corresponding allocations are returned; otherwise, the process results in a false local "optimum".

	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
c	0.425	0.473	0.532	0.608	0.709	0.851	0.983	1.311	1.966	0.977
C	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE

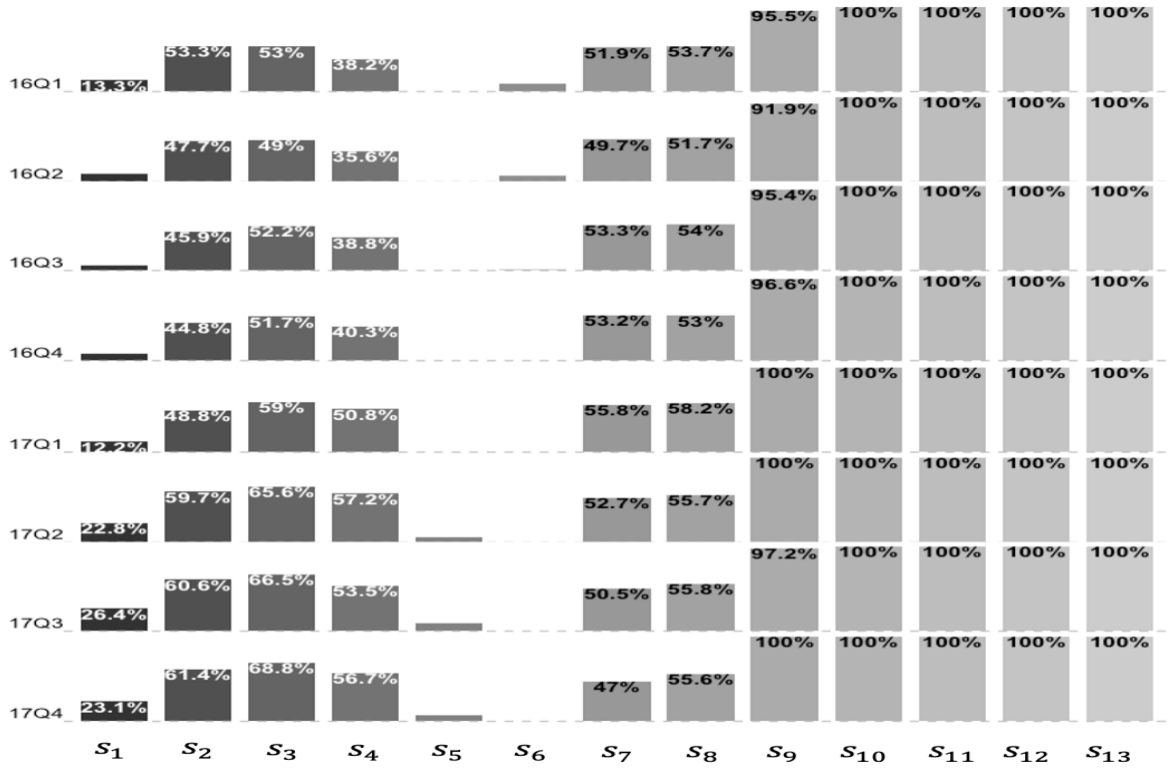

Figure G.1 Optimal allocation of each stratum to CAPI given starting point $s_g|s_0$ over quarters under $h = 100\%$. Different s_0 values result in convergent optima $s_g|s_0$. The dotted line indicates that $s_g|s_0 = 0$.

Reference list

Axinn, W. G., Link, C. F., & Groves, R. M. (2011). Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior. *Demography*, *48*(3), 1127–1149. doi: 10.1007/S13524-011-0044-1.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. In *Handbook of Nonresponse in Household Surveys*. doi: 10.1002/9780470891056.

Bethlehem, J. G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, *4*(3), 251–260. Online available on: (https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/reduction-of-nonresponse-bias-through-regression-estimation.pdf)

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. John Wiley & Sons.

Boonstra, H. J. (2021). *mcmcsae: MCMC small area estimation. R Package Version 0.6.0. https://cran.r-project.org/web/packages/mcmcsae/index.html*.

Boonstra, H. J., van den Brakel, J. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, *45(3)*, 395-425. Online available on: (https://www150.statcan.gc.ca/n1/pub/12-001-x/2019003/article/00005-eng.htm)

Boonstra, H. J., & van den Brakel, J. (2022). Multilevel time-series models for small area estimation at different frequencies and domain levels. *The Annals of Applied Statistics*, *16*(4), 2314–2338. doi: 10.1214/21-AOAS1592. *Technical Report 2018-12*, https://www.cbs.nl/en-gb/background/2018/50/models-for-estimation-at-various-aggregation-levels, *Statistics Netherlands.*

Brownstein, N.C., Louis, T.A., O'Hagan, A., & Pendergast, J. (2019). The role of expert judgment in statistical inference and evidence-based decision-making. *The American Statistician. 73*(1), 56-68. doi: 10.1080/00031305.2018.1529623.

Burger, J., Perryck, K., & Schouten, B. (2017). Robustness of adaptive survey designs to inaccuracy of design parameters. *Journal of Official Statistics*. *33(3)*, 687-708. doi: 10.1515/JOS-2017-0032.

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, *226*(1), 115–121. doi: 10.1016/j.ejor.2012.10.046.

Calinescu, M., & Schouten, B. (2016). Adaptive survey designs for nonresponse and measurement error in multi-purpose surveys. *Survey Research Methods*, *10*(1), 35–47. doi: 10.18148/srm/2016.v10i1.6157.

Calinescu, M., Schouten, B., & Bhulai, S. (2012). Adaptive survey designs that minimize nonresponse and measurement risk. *Statistics Netherlands Discussion Paper*. Online available on: (https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b7f035250e336e7732ffaa062d716e70eeb52c3c).

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. doi: 10.1093/biomet/asq017.

Chun, A. Y., Heeringa, S. G., & Schouten, J. G. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, *34*(3), 581–597. doi: 10.2478/jos-2018-0028.

Coffey, C. S., Levin, B., Clark, C., Timmerman, C., Wittes, J., Gilbert, P., & Harris, S. (2012). Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop. *Clinical Trials*, *9*(6), 671–680. doi: 10.1177/1740774512461859.

Coffey, S., West, B. T., Wagner, J., & Elliott, M. R. (2020). What do you think? Using expert opinion to improve predictions of response propensity under a bayesian framework. *Methoden, daten, analysen, 14(2)*. doi: 10.12758/mda.2020.05.

Couper, M., & Wagner, J. (2011). Using paradata and responsive design to manage survey nonresponse. In *ISI World Statistics Congress, Dublin, Ireland.* Online available on: (https://2011.isiproceedings.org/papers/450080.pdf).

Couper, M. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the ASA at JSM1998*, 41–49.

Datta, G. S., Lahiri, P., Maiti, T., & Lu, K. L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S. *Journal of the American Statistical Association, 94*(448), 1074–1082. doi: 10.1080/01621459.1999.10473860.

De Heij, V., Schouten, B., & Shlomo, N. (2015). RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators, partial R-indicators and partial coefficients of variation. Representativity Indicators for Survey Quality; 2015[updated 2017 Dec7; cited 2019 Mar 5]. Online available on: (https://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-manual-v21.pdf).

de Leeuw, E. D., & Toepoel, V. (2018). Mixed-mode and mixed-device surveys. *The Palgrave Handbook of Survey Research*, 51–61. doi: 10.1007/978-3-319-54395-6_8.

De Leeuw, E., & de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse, 41*, 41-54. New York: John Wiley. Online available on: (https://www.researchgate.net/publication/284051397).

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods* (Vol. 38). Oxford University Press.

Fang, Q., Burger, J., Meijers, R., & van Berkel, K. (2021). The Role of Time, Weather and Google Trends in Understanding and Predicting Web Survey Response. *Survey Research Methods*, *15*(*1*), 1–25. doi: 10.18148/srm/2021.v15i1.7633.

Franco, C., & Bell, W. R. (2015). Borrowing information over time in binomial/logit normal models for small area estimation. *Statistics in Transition New Series, 16*(4), 563–584. doi: 10.21307/stattrans-2015-033.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal. 1*(3): 515–534. doi: 10.1214/06-BA117A.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, *6*(4),773–760. Online available on: (https://www.jstor.org/stable/24306036).

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. Online available on: (https://www.jstor.org/stable/2246093).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. doi: 10.1201/9780429258411.

Gosling, J. P., Oakley, J. E., & O'Hagan, A. (2007). Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Analysis, 2(*4*),* 693–718. doi: 10.1214/07-ba228.

Groves, R. M., & Couper, M. (2012). *Nonresponse in household interview surveys*. John Wiley & Sons.

Groves, R. M. (2005). *Survey Errors and Survey Costs*. John Wiley & Sons.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology*. John Wiley & Sons.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3), 439-457. doi: 10.1111/j.1467-985X.2006.00423.x.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly, 64*(3), 299–308. doi: 10.1086/317990.

Groves, R. M., & McGonagle, K. A. (2001). A Theory-Guided Interviewer Training Protocol Regarding Survey Participation. *Journal of Official Statistics*, *17*(2), 249–265. Online available on: (https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-theory-guided-interviewer-training-protocol-regardingsurvey-participation.pdf).

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public opinion quarterly*, *72*(2), 167–189. doi: 10.1093/poq/nfn01l.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Ibrahim, J. G., & Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science, 15*(1), 46-60. doi: 10.1214/ss/1009212673.

Ibrahim, J. G., Chen, M. H., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. *Statistics in medicine, 34*(28), 3724-3749. doi: 10.1002/sim.6728.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of official statistics*, *19*(2), 81–97. Online available on: (https://search.crossref.org/?q=Weighting+Methods&from_ui=yes).

Kreuter, F. (2013). Facing the Nonresponse Challenge. *The Annals of the American Academy of Political and Social Science*, *645*(1), 23-35. doi: 10.1177/0002716212456815.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213–236. doi: 10.1002/acp.2350050305.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Sage publications.

Linderman, S., Johnson, M. J., & Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, 3456-3464. Online available on: (https://proceedings.neurips.cc/paper/2015/file/07a4e20a7bbeeb7a736682b26b16ebe8-Paper.pdf).

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., & Langer Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 779–787. doi: 10.1093/poq/nfu054.

Luiten, A., & Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *176*(1), 169–189. doi: 10.1111/j.1467-985x.2012.01080.x.

Ma, Y., Mushkudiani, N., & Schouten, B. (2021). *Optimal Stratification in Bayesian Adaptive Survey Designs* (Master's thesis). Presented in the ninth conference of the European Survey Research Association. Online available on: (https://www.europeansurveyresearch.org/conf2021/uploads/57/3653/67/Ma_ESRA21.pdf).

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*(1), 79–104. doi: 10.1177/147078530805000107.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. Social science computer review, *31*(6), 725–743. doi: 10.1177/0894439313485201.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly, 81*(S1), 250–271. doi: 10.1093/poq/nfw060.

Moore, J. C., Durrant, G. B., & Smith, P. W. (2018). Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society. Series A* (*Statistics in Society*), *181*(1), 229–248. doi: 10.1111/rssa.12256.

Moore, J. C., Durrant, G. B., & Smith, P. W. (2021). Do coefficients of variation of response propensities approximate non-response biases during survey data collection? *Journal of the Royal Statistical Society. Series A* (*Statistics in Society*), *184*(1), 301–323. doi: 10.1111/rssa.12624.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), *65*(2), 331–355. doi: 10.1111/1467-9868.00389.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., & TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence, 88*, S24–S30. doi: 10.1016/j.drugalcdep.2006.09.008.

Musch, J., & Reips, U. D. (2000). A brief history of web experimenting. In *Psychological experiments on the Internet*, 61–87. Academic Press. doi: 10.1016/B978-012099980-4/50004-6.

Mushkudiani, N., & Schouten, B. (2019). Time-dependent survey design parameters: Choosing the length of historic survey data in a Bayesian analysis, application to the Dutch Health Survey. *Workshop paper for advances in adaptive and responsive survey design.*

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review, 84*(1), 43–62. doi: 10.1111/insr.12100.

Oakley, J. E., & O'Hagan, A. (2007). Uncertainty in prior elicitations: A nonparametric approach. *Biometrika, 94*(2), 427-441. doi: 10.1093/biomet/asm031.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). Uncertain judgements: Eliciting experts' probabilities.

O'Malley, A. J., & Zaslavsky, A. M. (2012). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association, 103*(484), 1405–1418. doi: 10.1198/016214508000000724.

Peytchev, A. (2011). Breakoff and Unit Nonresponse Across Web Surveys. *Journal of Official Statistics, 27*(1), 33. doi: 10.4135/9781526421036927734.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., & Lindblad, M. (2010). Reduction of Nonresponse Bias through Case Prioritization. In *Survey Research Methods*, *4*(1), 21–29. doi: 10.18148/srm/2010.v4i1.3037.

Pfeffermann, D., & Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, *16*(2), 217–237. Online available on: (https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf?st=c38tPbwk).

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics, 9*, 501-538. Retrieved from: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017.

Polson, N. G., & Scott, J. G. (2013). Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika*, *100*(2), 459–471. doi: 10.1093/biomet/ass081.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American statistical Association, 108*(504), 1339–1349. doi: 10.1080/01621459.2013.829001.

Rao, J. N., & Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, *22*(4), 511–528. doi: 10.2307/3315407.

Rao, J. N., & Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Raymer, J., Rees, P., & Blake, A. (2015). Frameworks for guiding the development and improvement of population statistics in the United Kingdom. *Journal of Official Statistics*, *31*(4), 699-722. doi: 10.1515/jos-2015-0041.

Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G., & Hoijtink, H. J. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, *32*(6), 848-855. doi: 10.1016/j.cct.2011.06.002.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi: 10.1093/biomet/70.1.41.

Rosenblum, M., Miller, P., Reist, B., Stuart, E. A., Thieme, M., & Louis, T. A. (2019). Adaptive design in surveys and clinical trials: similarities, differences and opportunities for cross-fertilization. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(3), 963–982. doi: 10.1111/rssa.12438.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields*: *theory and applications.* CRC press. doi: 10.1201/80203492024.

Sakshaug, J. W., & Antoni, M. (2019). Evaluating the utility of indirectly linked federal administrative records for nonresponse bias adjustment. *Journal of Survey Statistics and Methodology*, *7*(2), 227–249. doi: 10.1093/jssam/smy009.

Särndal, C. E., & Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, *2*(4), 361–387. doi: 10.1093/jssam/smu014.

Särndal, C. E., & Lundquist, P. (2019). An assessment of accuracy improvement by adaptive survey design. *Survey Methodology*, *45*(2), 317–338. Available online on:

(https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019002/article/00008-eng.pdf?st=MSHOHfm6).

Savitsky, T. D. (2019). Bayesian nonparametric functional mixture estimation for time-series data, with application to estimation of state employment totals. *Journal of Survey Statistics and Methodology, 7*(1), 3-33. doi: 10.1093/jssam/smy001.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, ∅., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response hrough R-indicators and partial R-indicators. *International Statistical Review*, *80*(3), 382–399. doi: 10.1111/j.1751-5823.2012.00189.x.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey methodology*, 39(1), 29–58. Available online on: (https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11824-eng.pdf?st=-qT1Y_ga).

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology, 35*(1), 101–113. Available online on: (https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2008000/article/10976-eng.pdf?st=KpFR-1NJ).

Schouten, B., Cobben, F., Lundquist, P., & Wagner, J. (2016). Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society. Series A (Statistics in Society), 179*(3), 727–748. doi: 10.1111/rssa.12152.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., & Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*, *6*(4), 431-464. doi: 10.1093/jssam/smy012.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. CRC Press. doi: 10.1201/9781315153964.

Schouten, B., & Shlomo, N. (2017). Selecting adaptive survey design strata with partial R-indicators. *International Statistical Review*, *85*(1), 143–163. doi: 10.1111/insr.12159.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics, 27*(2), 231-253.

Available online on: (https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf).

Schouten, B. (2004). Adjustment for bias in the integrated survey on household living conditions (POLS) 1998. Discussion paper 04001, *Statistics Netherlands*.

Shlomo, N., & Schouten, B. (2013). Theoretical Properties of Partial Indicators for Representative. Available online on: (https://hummedia.manchester.ac.uk/institutes/cmist/risq/shlomo-schouten-2013.pdf).

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference, 142*(1), 201–211. doi: 10.1016/j.jspi.2011.07.008.

Singer, E., Groves, R. M., & Corning, A. D. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly, 63*(2), 251–260. doi: 10.1086/297714.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639. doi: 10.1111/1467-9868.00353.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Thompson, K. J., & Kaputa, S. J. (2017). Investigating adaptive nonresponse follow-up strategies for small businesses through embedded experiments. *Journal of Official Statistics, 33*(3), 835–856. doi: 10.1515/jos-2017-0038.

Tourangeau, R., Rasinski, K., Jobe, J. B., Smith, T. W., & Pratt, W. F. (1997). Sources of error in a survey on sexual behavior. *Journal of Official Statistics, 13*(4), 341-366. Available online on: (https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/sources-of-error-in-a-survey-on-sexual-behavior.pdf).

van Berkel, K., van der Doef, S., & Schouten, B. (2020). Implementing adaptive survey design with an application to the Dutch Health Survey. *Journal of Official Statistics, 36*(3), 609–629. doi: 10.2478/jos-2020-0031.

Varadhan, R. (2015). alabama: Constrained nonlinear optimization. R package version 2015.3-1.

Veen, D., Stoel, D., Zondervan-wijnenburg, M., & Van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgment. *Frontiers in Psychology, 8*, 2110. doi: 10.3389/fpsyg.2017.02110.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly, 74*(2), 223–243. doi: 10.1093/poq/nfq007.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly, 76*(3), 555–575. doi: 10.1093/poq/nfs032.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods, 7*(1), 45–55. doi: 10.18148/srm/2013.v7i1.5037.

Wagner, J., & F. Hubbard. (2013). Using propensity models during data collection for responsive designs: Issues with estimation. Paper presented at 68th AAPOR conference, May 16-19, Boston, USA.

Wagner, J., & Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology, 2*(3), 323–342. doi: 10.1093/jssam/smu009.

Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias*. Doctoral dissertation, University of Michigan. Available online on: (https://deepblue.lib.umich.edu/bitstream/handle/2027.42/60831/jameswag_1.pdf).

Wagner, J., West, B. T., Elliott, M. R., & Coffey, S. (2020). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics, 36*(4), 907–931. doi: 10.2478/jos-2020-0043.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine*

Learning Research, *11*(12), 3571–3594. Available online on: (https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf).

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*(1), 867–897. Available online on: (https://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf).

West, B. T., Wagner, J., Coffey, S., & Elliott, M. R. (2020). The elicitation of prior distributions for Bayesian responsive survey design: Historical data analysis vs. literature review. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf.

West, B. T., Wagner, J., Coffey, S., & Elliott, M. R. (2021). Deriving priors for Bayesian prediction of daily response propensity in responsive survey design: Historical data analysis vs. literature review. *Journal of Survey Statistics and Methodology.* doi: 10.1093/jssam/smab036.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, *74*(3), 646–648. doi: 10.1093/biomet/74.3.646.

Wu, S., Boonstra, H. J., Moerbeek, M., & Schouten, B. (2023). Modelling Time Change in Survey Response Rates: A Bayesian Approach with an application to the Dutch Health Survey. *Accepted by Survey Methodology*.

Wu, S., Schouten, B., Meijers, R., & Moerbeek, M. (2022). Data Collection Expert Prior Elicitation in Survey Design: Two Case Studies. *Journal of Official Statistics*, *38*(2), 637–662. doi: 10.2478/jos-2022-0028.

You, Y., Rao, J. N., & Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, *29*(1), 25–32. Retrieved from: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf?st=yS5Ljmzz.

Zhang, S., & Wagner, J. (2022). The Additional Effects of Adaptive Survey Design Beyond Post-Survey Adjustment: An Experimental Evaluation. *Sociological Methods & Research* doi: 10.1177/00491241221099550.

Zijlstra, T., Wijgergangs, K., & Hoogendoorn-Lanser, S. (2018). Traditional and mobile devices in computer assisted web-interviews. *Transportation Research Procedia*, *32*, 184–194. doi: 10.1016/j.trpro.2018.10.033.

English summary

Survey practitioners keep steadily searching for methods to improve effectiveness of adaptive survey design. Methods most often come in conflict with the rare historic data sets for running an infrequent or new survey. Also, methods most often ignore the timeliness of historic data of an ongoing survey. In this dissertation, we developed and applied Bayesian methods in adaptive survey design, both for precise and reliable predictions to make about survey design parameters and for ensuring timeliness of scarce survey resources to allocate. We discuss the Bayesian framework for its ability to include external data through prior distributions and to learn how responses vary in time in order to improve prediction precision. We also discuss effective adaptive survey designs that timely tailor the follow-up strategy to approach nonrespondents in order to enhance the obtained response. A useful tool is provided here to decide such optimal allocations given a survey quality objective.

The models for adaptive survey designs are offered in Chapters 2 to 4. In Chapter 2 we first start our research in the context of response propensity estimation for running a survey that is conducted infrequently or newly at Statistics Netherlands. The Bayesian framework is developed for its ability to include data collection staff (expert) knowledge in prior distribution specification. This modelling flexibility is compared to the standard framework when relevant historic information is unavailable completely. We also explore two different ways to pool expert knowledge in the expert models. These expert models and the standard model are routinely applied to evaluate survey design quality based on two overall and one partial representative indicators (R-indicators, overall and partial coefficients of variation). We examine these approaches within two case studies where the accuracy of estimated indicators is measured through the root mean squared error. The accuracy is estimated to validate if making early decision was profitable. Results from various evaluations show that the pooling way led to similar performance, and also show that either of expert models is in early and middle data collection waves preferable over the standard model to estimate the accuracy. The expert models provide most accurate estimates on the variations of response. Results also reveal that the expert prior elicitation approach is recommended when a survey has never been conducted before or when a survey has gone through design changes.

Survey literature often assumes that in a relatively fixed survey climate, response propensities are quite stable and eventually converge when one obtains sufficient historic data. However, the reality that response rates vary in time is not in line with this common assumption. To fill this gap, we explore the use of Bayesian methods in the context of multilevel prediction models and apply them to a time series data in Chapter 3 where response rates are decomposed into multiple components at strata and time levels and each component is evaluated. The main motivation of this study is to grasp how time alters response propensities and thus to examine the prediction performance. Six-year time series data are studied and seven model components are specified (including three fixed effects and four random effects): time-independent covariates, overall linear time trend, seasonal effects, random intercepts for subgroups, a global time trend, stratum-specific trends, and effects for remaining unstructured variation in response propensities. These components form eleven models considered. Each model performance is assessed through information criteria in order to opt for the best-performance model contributing most to variation in response propensities. Given the "optimal" model specification, evaluation criteria that measure the level of and variation in overall and subgroup-level response propensity predictions are explored as well. Results show that with accumulating historic data, the overall variation decreases steadily while a modest but volatile decrease on the overall bias. At a time period, the predictions of some subgroups are more biased than others. Results also indicate that including data collected during outlier months and under design change to train a model can worse prediction accuracy. The latter negative impact (design change) is modest. As a heuristic, we show a simple suggestion about how to deal with them. Prediction accuracy can improve further, when the "abnormal" data during outlier months are replaced and imputed by the model, and when the impact since design change is introduced is estimated well using sufficient new data.

How adaptation strategies can adjust to such variation over time is another aspect of concern in survey design. Practical evidence is rare as this is not yet fully figured out. In order to grasp the sensitivity of the adaptation performance to time change in response propensities, we further explored the use of Bayesian multilevel time series models in mixed-mode surveys. A new parameter is, called correlation in response propensities

between two consecutive interview modes, introduced in order to make reliable and precise predictions at each mode. In Chapter 4, we also construct the resource allocation model that employs Bayesian analysis, with the flexibility of accommodating uncertainty due to time change. In this approach, a constraint on the workload of interviewers is set conditional on a budget level. The quality indicator, i.e., the overall coefficient of variation, is minimized based on this constraint. Then, given the estimated response propensities at each mode, we can estimate the proportion of nonrespondents of each subgroup to assign to the costly mode, if the previous modes fail to obtain their responses. In addition to explore the sensitivity of adaptation performance over time, we also investigate how sensitive the performance is to different budget levels. For each budget level, its estimated overall variation of adaptive design is compared with those of CAWI-only and non-adaptive designs. Results show there is a clear advantage of adaptive designs, in contrast to non-adaptive designs, at relatively high budget levels. However, no further improvement in estimated overall variation of adaptation can be seen for less than 50% budget level. Modelling the seasonal effects seem to be of little value to contribute to reduction in overall variation. On the one hand, this is possibly because we do not distinguish the seasonal effects between CAWI and CAPI. On the other hand, the study on individual subgroup performance implies that a stopping rule of data collection may be implemented and an effort-based strategy for small sample size subgroups may be adopted.

This dissertation presents the development of Bayesian methods behind effective adaptive survey design. We show that in addition to the Bayesian analysis with internal historic data, expert prior elicitation and a deeper understanding of time change are useful tool to improve survey design parameter prediction accuracy and to intervene timely and appropriately.

Nederlandse Samenvatting

Enquêteurs blijven voortdurend zoeken naar methoden om de doeltreffendheid van adaptieve enquêteontwerpen te verbeteren. De methoden komen meestal in conflict met de zeldzame historische gegevensreeksen voor het houden van een onregelmatige of nieuwe enquête. Ook gaan methoden meestal voorbij aan de actualiteit van historische gegevens van een lopende enquête. In dit proefschrift hebben wij Bayesiaanse methoden ontwikkeld en toegepast bij adaptief enquêteontwerp, zowel om nauwkeurige en betrouwbare voorspellingen te kunnen doen over enquêteontwerp-parameters als om de tijdigheid van schaarse toe te wijzen enquêtemiddelen te waarborgen. Wij bespreken het Bayesiaanse kader voor zijn vermogen om externe gegevens op te nemen via prior verdelingen en om te leren hoe antwoorden in de tijd variëren om de nauwkeurigheid van de voorspellingen te verbeteren. Wij bespreken ook doeltreffende adaptieve enquêteontwerpen die de follow-upstrategie tijdig aanpassen om niet-respondenten te benaderen teneinde de verkregen respons te verbeteren. Hier wordt een nuttig instrument aangereikt om dergelijke optimale toewijzingen te bepalen gegeven een doelstelling voor de kwaliteit van de enquête.

De modellen voor adaptieve enquêteontwerpen worden aangeboden in de hoofdstukken 2 tot en met 4. In hoofdstuk 2 beginnen we ons onderzoek eerst in de context van de schatting van de responsdichtheid voor een enquête die onregelmatig of nieuw bij het CBS wordt gehouden. Het Bayesiaanse raamwerk is ontwikkeld vanwege de mogelijkheid om (expert)kennis van dataverzamelaars mee te nemen in de specificatie van de prior distribution. Deze modelleringsflexibiliteit wordt vergeleken met het standaardkader wanneer relevante historische informatie volledig ontbreekt. We onderzoeken ook twee verschillende manieren om expertkennis in de expertmodellen te bundelen. Deze deskundigenmodellen en het standaardmodel worden routinematig toegepast om de kwaliteit van het enquêteontwerp te evalueren op basis van twee algemene en één partiële representatieve indicator (R-indicatoren, algemene en partiële variatiecoëfficiënt). Wij onderzoeken deze benaderingen in twee casestudies waarbij de nauwkeurigheid van de geschatte indicatoren wordt gemeten aan de hand van de gemiddelde kwadratische fout. De nauwkeurigheid wordt geschat om te valideren of het nemen van een vroegtijdige beslissing rendabel was. Uit de resultaten van verschillende evaluaties blijkt dat de

poolingmethode tot vergelijkbare prestaties leidt, en ook dat een van beide expertmodellen in vroege en middelste gegevensverzamelingsgolven de voorkeur verdient boven het standaardmodel om de nauwkeurigheid te schatten. De deskundigenmodellen leveren de nauwkeurigste schattingen op van de variaties in de respons. Uit de resultaten blijkt ook dat de aanpak van de voorafgaande elicitatie door deskundigen wordt aanbevolen wanneer een enquête nog nooit eerder is uitgevoerd of wanneer een enquête wijzigingen in de opzet heeft ondergaan.

In de enquêteliteratuur wordt er vaak van uitgegaan dat bij een betrekkelijk vast enquêteklimaat de reactiebereidheid vrij stabiel is en uiteindelijk convergeert wanneer men over voldoende historische gegevens beschikt. De realiteit dat de respons in de tijd varieert, strookt echter niet met deze gangbare veronderstelling. Om deze leemte op te vullen, onderzoeken wij het gebruik van Bayesiaanse methoden in de context van multilevel voorspellingsmodellen en passen deze toe op een tijdreeks van gegevens in hoofdstuk 3, waarin de responspercentages worden ontleed in meerdere componenten op strata- en tijdsniveau en elke component wordt geëvalueerd. De belangrijkste motivatie van deze studie is te begrijpen hoe de tijd de responsgevoeligheid verandert en aldus de voorspellingsprestaties te onderzoeken. Er worden tijdreeksgegevens van zes jaar bestudeerd en er worden zeven modelcomponenten gespecificeerd (waaronder drie vaste effecten en vier willekeurige effecten): tijdsonafhankelijke covariaten, een algemene lineaire tijdstrend, seizoenseffecten, willekeurige intercepts voor subgroepen, een algemene tijdstrend, stratum-specifieke trends en effecten voor de resterende ongestructureerde variatie in responsneigingen. Deze componenten vormen elf in aanmerking genomen modellen. De prestaties van elk model worden beoordeeld aan de hand van informatiecriteria, teneinde het best presterende model te kiezen dat het meest bijdraagt tot de variatie in de responsneiging. Gegeven de "optimale" modelspecificatie worden ook evaluatiecriteria onderzocht die het niveau van en de variatie in de voorspellingen van de totale en subgroepresponsneiging meten. Uit de resultaten blijkt dat met de accumulatie van historische gegevens de algemene variatie gestaag afneemt, terwijl de algemene vertekening bescheiden maar volatiel afneemt. In een bepaalde periode zijn de voorspellingen van sommige subgroepen sterker vertekend dan die van andere. De resultaten wijzen er ook op dat het opnemen van gegevens die zijn verzameld tijdens uitbijtermaanden en onder ontwerpwijzigingen om een model te trainen, de

voorspellingsnauwkeurigheid kan verslechteren. Het laatste negatieve effect (ontwerpwijziging) is bescheiden. Als heuristiek geven wij een eenvoudige suggestie hoe daarmee om te gaan. De voorspellingsnauwkeurigheid kan verder verbeteren wanneer de "abnormale" gegevens tijdens de maanden met een uitschieter worden vervangen en door het model worden toegerekend, en wanneer het effect sinds de invoering van de ontwerpwijziging goed wordt geschat met behulp van voldoende nieuwe gegevens.

Hoe aanpassingsstrategieën zich in de loop der tijd aan een dergelijke variatie kunnen aanpassen, is een ander aspect van de enquêteopzet. Praktische bewijzen zijn schaars omdat dit nog niet volledig is uitgezocht. Om inzicht te krijgen in de gevoeligheid van de aanpassingsprestaties voor veranderingen in de responsbereidheid in de tijd, hebben wij het gebruik van Bayesiaanse multilevel tijdreeksmodellen in gemengde enquêtes verder onderzocht. Er is een nieuwe parameter ingevoerd, namelijk de correlatie in responsneigingen tussen twee opeenvolgende interviewmodi, om betrouwbare en nauwkeurige voorspellingen voor elke modus te kunnen doen. In hoofdstuk 4 construeren wij ook het model voor de toewijzing van middelen dat gebruik maakt van Bayesiaanse analyse, met de flexibiliteit om onzekerheid als gevolg van tijdsverandering op te vangen. In deze aanpak wordt een beperking op de werklast van de interviewers vastgesteld op basis van een budgetniveau. De kwaliteitsindicator, d.w.z. de totale variatiecoëfficiënt, wordt op basis van deze beperking geminimaliseerd. Vervolgens kunnen wij, gezien de geschatte responsbereidheid in elke modus, het aandeel niet-respondenten van elke subgroep schatten dat aan de dure modus moet worden toegewezen indien de vorige modi geen respons opleveren. Behalve de gevoeligheid van de aanpassingsprestaties in de tijd onderzoeken wij ook hoe gevoelig de prestaties zijn voor verschillende begrotingsniveaus. Voor elk budgetniveau wordt de geschatte totale variatie van het adaptieve ontwerp vergeleken met die van CAWI-only en niet-adaptieve ontwerpen. Uit de resultaten blijkt dat er een duidelijk voordeel is van adaptieve ontwerpen, in tegenstelling tot niet-adaptieve ontwerpen, bij relatief hoge begrotingsniveaus. Bij een budgetniveau van minder dan 50% is er echter geen verdere verbetering van de geschatte totale aanpassingsvariatie te zien. De modellering van de seizoensgebonden effecten lijkt weinig bij te dragen tot een vermindering van de totale variatie. Enerzijds komt dit mogelijk doordat wij geen onderscheid maken tussen de seizoenseffecten van CAWI en CAPI. Anderzijds impliceert het onderzoek naar de prestaties van individuele subgroepen dat er

een stopregel voor de gegevensverzameling kan worden toegepast en een op inspanning gebaseerde strategie voor subgroepen met een kleine steekproefomvang.

Dit proefschrift presenteert de ontwikkeling van Bayesiaanse methoden voor een effectief adaptief enquêteontwerp. Wij tonen aan dat, naast de Bayesiaanse analyse met interne historische gegevens, het ontlokken van voorkeuren door deskundigen en een dieper begrip van tijdsverandering nuttige hulpmiddelen zijn om de nauwkeurigheid van de voorspelling van enquêteontwerpparameters te verbeteren en tijdig en passend in te grijpen.

Acknowledgements

Embarking on the journey to obtain a PhD degree is a challenging experience that demands curiosity and enthusiasm for discovery. Although there are moments of frustration and despair along the way, this journey is also full of sweet and tough moments. Every time when I look back at those years working at Centraal Bureau voor de Statistiek (CBS) and the department of Methodology & Statistics (MS), I am deeply grateful that many wonderful people have supported me and helped me get through the rough time.

First of all, I would like to express my deepest gratitude to my supervisors, Barry Schouten and Mirjam Moerbeek. Without their encouragement, I would not be in this position today. **Barry**, I have always appreciated by your endless patience and continuous guidance at each stage of my project. In particular, to all my emails when being stuck with research and any time whenever I need feedback, you are always there. I always feel inspired by your caring for detail, your broad expertise, your open-mindedness, and your perspectives on the project. These teaches me how to be precise in research and how to develop an idea into an article. During our periodic meetings and trips to Den Haag or Heerlen, I have learned so much about survey methodology and survey practice. I am also grateful for the generous grants you looked for me supporting the extension of this project. Thank you for understanding regarding my procrastination and all positive reminders when I get lost. **Mirjam**, I am grateful to have you as my supervisor. You consistently guide me step by step in academic life, giving useful suggestions on scientific writing, presentation, taking research-related courses etc. Thank you also for caring for me in personal life.

I would like to thank my co-authors Ralph and Harm Jan for contributions to my project. Your involvement and feedback have been useful in shaping our research and ensuing its quality. **Ralph**, thank you for your expertise and fruitful discussion on the expert elicitation questionnaire. **Harm Jan**, thank you for inviting me to join your lecture organized in Heerlen, inspiring discussions during our bi-weekly meetings, all valuable input, critical comments and dealing with the flood of emails about my questions. It is a great pleasure working with you.

I also would like to thank the regarding evaluation committee members, Jan van den Brakel, Irene Klugkist, Rens van de Schoot, Bella Struminskaya and James Wagner. Thank you for careful reading of my dissertation, and accepting a seat on the jury.

Many thanks to all my (former) colleagues at MS and CBS. I have thoroughly enjoyed talking with colleagues who share their passion for work and life. **Danielle**, you brighten up my cloudiest moments. The conference trip to Zagreb is full of the warmest memories. Thank you for the interesting discussions, all the happy chats, coffees, and you liking my cook, etc.

I want to thank my paranymphs, **Danielle** and **Xinwei**. Thank you for being willing to stand by my side through the last stretch of this journey. **Xinwei**, when I was in Enschede, I felt so lonely and confused about everything. Thank you for being my close friend, sharing you experience and hobbies with me, and helping me get my academic and personal life back on track when things are out of control. I always enjoy cooking together, sharing the delicious food, going out together, car rides, playing the board games, etc.

I own sincere thanks to my family as well. Thank you for your constant emotional support. To **Geyang**, I am deeply grateful for the way you have treated me as a peer and friend and encouraged my growth. Thank you for always pushing me to be my best self, and being a critical but empathetic listener to all my concerns and thoughts.

I also own a super big thank to my UU friends, Yue, Dan, Shuai, Zhongfang, Wenrui, Qiaorao, Qianqian, Qixiang, Yongchao, Haifang. Thank you for being an important part of my journey and for helping me to develop both personally and professionally. To **Yue**, for being like my sister, I still remembered we met at the conference hold by Chinese psychometric Association, and we really hit it off. You are always open to hearing my thoughts and ideas and can always get my point. To **Shuai**, I really enjoyed working together with you. I can be inspired by all your wonderful idea. To **Dan**, for generously sharing a lot of coupons, helping me out when I have technical problems, and the interesting discussions on writing R codes. To **Haifang**, for giving me a lot of help in academic and personal life, and being my photographer. In addition, thanks also to my other close friends in Europe and China, Lin, Tjerry, Xiaoxiao, Jiangyang, Xiaoling,

Yidong, Cui, and Zheng, Jiayun. To **Lin** and **Tjerry**, for the sweetest memories that we have created together in your house, Den Haag, Paris, etc. To **Cui** and **Zheng**, for being my travel pal. To **Xiaoling**, for the continuous encouragement, always cheering me up when I am depressed and being your baby son's godmother. To **Jiayun**, for being my "survey research" buddy since we met at the 8th ESRA conference.

Special thanks to China Scholarship Council for providing me financial support.

Lastly, I would like to specially mention my lovely husband **Xiaofei**. I am so lucky enough to have you as being my rock. Without you, I could not finish this journey. Let me quote from your favorite animation Toy Story, "You've got a friend in me. You got troubles. I've got em too. There isn't anything I wouldn't do for you. We stick together and see it through. Cause you've got a friend in me." This year, we will welcome a new family member, our baby girl. Thank you for choosing us as parents and accompanying me to the end of this journey.

About the author

Shiya Wu was born on November 1, 1989 in Hubei, China. In 2012, she completed her bachelor study (BSc.) in Environmental Engineering at Changchun University of Technology. In 2016 she completed the research master Statistics: Simultaneous Confidence Bands for Linear Regression with Covariates Restricted and obtained her MSc. from Northeast Normal University.

In October 2017, she started as a PhD Candidate at the department of Methodology and Statistics at Utrecht University and in survey methodology at Statistics Netherlands. She started the PhD project on developing the methodology behind effective adaptive survey designs under the supervision of prof. dr. J.G. Schouten and Dr. M. Moerbeek at Utrecht University. During her PhD, she also gave some presentations at conferences such as ERSA and IOPS.