Optimal Stratification in Bayesian Adaptive Survey Designs

Yongchao Ma

Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, Heidelberglaan 8, Utrecht, 3584 CS, The Netherlands

E-mail: y.ma1@uu.nl

Abstract

In an increasing number of survey designs, adaptive data collection strategies for different members of the population are adopted to balance the data quality and cost. Stratifying the target population into subgroups in an effective manner plays a decisive role in identifying the optimal adaptive survey design. This paper presents a stratification method on the basis of which the optimal adaptive survey designs can be constructed under the Bayesian analysis to minimize nonresponse bias. The utility of this method compared to two other response- and cost-oriented stratification methods is assessed through a case study based on the Dutch Health Survey. The optimal adaptive survey designs based on the proposed method outperform in minimizing nonresponse bias, which indicates that the underlying stratification is the optimal stratification.

Keywords: Adaptive survey design; Stratification; Nonresponse; Response propensities; Survey costs

1 Introduction

Collecting official statistics survey data is getting more difficult. Costs have been increasing while budgets have been decreasing. Declining response rates yield nonresponse bias that jeopardizes the validity of inferences from data. As a result, statistical institutes are interested in adaptive survey designs that can balance cost and data quality. Analogous to the group sequential designs in clinical trials (Rosenblum et al. 2019), adaptive survey designs are based on the rationale that different data collection strategies may be effective for different members of the population (Groves and Heeringa 2006; Wagner 2008; Luiten and Schouten 2013).

One crucial step to adapt data collection strategies is to accurately estimate survey design parameters, such as response propensities and costs per interview, based on historic survey data and expert knowledge (Burger, Perryck and Schouten 2017). The optimal allocation of data collection strategies is determined by optimizing the quality and costs indicators that are functions of these design parameters. A natural approach to estimate design parameters and account for their uncertainty is through Bayesian analysis (Gelman et al. 2013). In such an analysis, the design parameters are treated as random variables and are assigned prior distributions that may be elicited from current knowledge about a survey; during data collection, the posterior distributions are derived and may serve as prior distributions in the next wave of the same survey (Schouten et al. 2018).

Another critical choice is the stratification of the population into subgroups. These subgroups are based on auxiliary data on the sample population, obtained from register data, paradata, or some other source. The stratification can be oriented at explaining heterogeneity in response behavior, key survey variables, or survey costs (Schouten, Peytchev and Wagner 2017). Different stratification objectives reflect the various concerns from survey stakeholders who decide whether to prioritize high response rates, acceptable survey estimates precision, or affordable survey costs. These two choices are requisites for the optimization of adaptive survey designs (Schouten et al. 2017). In this paper, the focus is on the adaptive survey design with the objective of minimizing nonresponse bias. The optimized survey design, i.e., the optimal allocation of data collection strategies to population subgroups, is expected to reduce nonresponse bias and satisfy other constraints. It brings two related questions:

- How to stratify the target population into subgroups effectively and efficiently?
- How to construct an optimal and robust adaptive survey design that minimizes nonresponse bias?

The answer to the second question implies the answer to the first one. The stratification underlying the optimal and robust adaptive survey design that minimizes nonresponse bias would be the optimal stratification.

Previous research has described several methods to form strata, most of which focus on explaining variation in response propensity and survey variables. Only a few studies discuss cost prediction, but no stratification is constructed to explain costs (Wagner 2019; Wagner, West, Elliott and Coffey 2020).

Response propensity variation. Särndal (2011) proposes to identify strata whose responses lead to a lack of balance on the auxiliary variables. Similarly, Schouten and Shlomo (2017) use partial representativeness indicators to build nonrespondent profiles to identify the lower-responding categories within the auxiliary variables. However, if the selected auxiliary variables are not related to the key survey variables, maximizing response rates does not necessarily reduce nonresponse bias (Rosen et al. 2014). For example, the web survey response is unbalanced across age strata, and the elderly can be identified as a target stratum with lower response propensities. Following up the elderly nonresponse by the face-to-face survey can thus improve the response rates. Suppose the obesity rate is a key survey variable and is not correlated with age, simply increasing the response rate of the elderly stratum has a limited effect on improving the estimate of obesity rate.

Survey variable variation. In order to address the precision of survey estimates, Wagner (2014) proposes to model the key survey variables by auxiliary data and identify influential units by regression diagnostic measures. Within each auxiliary variable, these units can be used to identify the categories where survey variables have wide variation relative to other categories. For example, assuming that a key survey variable, smoking prevalence, is modeled by age, a few influential observations in the youth stratum substantially distort the model estimate. It could be useful to have more young respondents to reduce the variance of the smoking prevalence estimate. However, unlike the strata differentiated by response behavior, the lack of knowledge

about the average stratum response propensity to different data collection strategies impedes efficient allocation of survey resources as well as cost control.

Methods oriented to response propensity variation can be used to balance survey response on the auxiliary variables that are related to survey variables. This is an efficient approach to increase response rates, but an indirect way to minimize nonresponse errors in survey estimates. On the other hand, methods oriented to survey variable variation can effectively detect strata from which nonresponse errors originate, but it is inefficient when the likelihood of recruiting units from these strata in a survey and the corresponding costs remain unclear.

This paper introduces a stratification method that determines the target strata that differ most in response behavior based on predicted survey variables. To demonstrate the effectiveness and efficiency of this method, this paper also develops a criterion to quantify the utility of different stratification methods in minimizing response variation with respect to survey variables. Before the start of data collection, the values of survey variables for sample units are unknown. Nevertheless, they may be predicted by the available auxiliary data from the sampling frame or administrative data. The resulting strata can account for the heterogeneity in both response behavior and key survey variables. In this sense, balancing the response rate across strata reduces the nonresponse error directly. For example, suppose smoking behavior is a key survey variable, a stratum may consist of sample units with probabilities of smoking higher than 0.3. If units in this stratum are less likely to respond to the survey, the smoking prevalence may be underestimated from the survey outcome.

This paper has six main sections: Section 2 describes the data collection strategies to be adapted and relates them to the optimization problem by functions of response propensities and costs, that is, the quality and cost indicators of interest. Section 3 describes the models for the input variables, which are the survey variable, of the classification tree algorithm. Based on the output stratification scheme, the response propensities and costs are estimated. All the model parameters receive prior distributions. Section 4 describes the optimization approach under a Bayesian analysis of the quality and cost indicators and develops a criterion to compare optimized survey designs based on different stratification methods. Section 5 shows an application to this methodology in the Dutch Health Survey and demonstrates that the proposed stratification method leads to optimal stratification. In the end, this paper concludes with a discussion of future

research in section 6.

2 Adaptive Strategies and Optimization Problem

This section describes data collection strategies in adaptive designs and formulates the optimization problem with the objective of minimizing nonresponse bias subject to several constraints.

2.1 Data Collection Strategies to Adapt

A survey design has different features, such as sample design, mode of administration, number of phases, type of questionnaire, and interviewer. A data collection strategy is the sequence of actions corresponding to the choices made for the design features. For example, a strategy can be to first invite the sample unit to participate in a web survey, and then attempt a face-to-face visit if no response is received.

Let the survey design consist of T phases, labeled $t = 1, 2, \ldots, T$. Let $\mathcal{S}_t$ denote the collection of all possible actions in phase t and let s_t represent the action taken in phase t. The aggregation of data collection strategies from phase one to T is defined as

$$\mathcal{S}_{1,T} := \{(s_1, \ldots, s_T) : s_t \in \mathcal{S}_t, t = 1, 2, \ldots, T\},$$

and let $s_{1,T} \in \mathcal{S}_{1,T}$ denote one possible strategy, that is, sequence of actions from phase one to T. In nonadaptive survey designs, a single strategy $s_{1,T}$ is implemented over the entire sample. In adaptive survey designs, a set of strategies are implemented with part of the design features varying for different sample units. Stratification separates the different sample units into subgroups that differ most in their response to different strategies. For example, people aged 60 and older may be less likely to respond to a web questionnaire but may respond to a face-to-face interview. Different groups of sample units may receive different treatments.

Let the target population consist of G strata, labeled $g = 1, 2, \ldots, G$. Every unit in a stratum g is eligible to be approached by any one of the strategies in $\mathcal{S}_{1,T}$. Let s_g represent the strategy assigned to a stratum g. The aggregation of allocations of data collection strategies over G strata

is defined as

$$\mathcal{S}_{1,T}^{G} = \underbrace{\mathcal{S}_{1,T} \times \mathcal{S}_{1,T} \times \cdots \times \mathcal{S}_{1,T}}_{G} := \{(s_1, \ldots, s_G) : s_g \in \mathcal{S}_{1,T}, g = 1, 2, \ldots, G\}$$

and $(s_{1,1,T}, \ldots, s_{G,1,T}) \in \mathcal{S}_{1,T}^{G}$ is one allocation when all strata receive the same strategy, that is, sequence of the strategy $s_{1,T}$ assigned to the first till the G-th stratum. To simplify the notation, let $s_\phi \in \mathcal{S}_{1,T}^{G}$, $\phi = 1, 2, \ldots, |\mathcal{S}_{1,T}^{G}|$, denote one possible allocation.

2.2 Optimization Objective and Constraints

The objective of the optimization is to find the best survey design $s_{\phi_{\text{opt}}}$, i.e., to assign strategies to strata considering the quality of the collected data and other constraints. The data quality and other constraints are formulated as indicators composed of design parameters. The design parameters considered here are the response propensity and the survey cost.

Under a specific allocation of strategies s_ϕ, the cost is an observed quantity that can be modeled straightforwardly, whereas the response propensity is a close estimate of the unobserved, nonzero response probability ρ (Bethlehem, Cobben and Schouten 2011; Schouten, Calinescu and Luiten 2013). The response probability reflects all the uncertainty arising from the impact of many factors such as the mood of the respondent or interviewer. The focus of this paper, however, is to reduce nonresponse bias in survey estimates $\bar{y}$ of the population mean $\bar{Y}$, that is, to suppress the correlation between response probability ρ_y and the survey variables of interest Y

$$\text{Bias}(\bar{y}) \approx \frac{Cov(\rho_y, Y)}{\bar{\rho}_y} = \frac{Corr(\rho_y, Y) S_{\rho_y} S_Y}{\bar{\rho}_y},$$

where $\bar{\rho}_y$ denotes the mean response probability in the population, and S_{ρ_y} is the standard deviation of the response probabilities, and S_Y is the standard deviation of the values taken for the survey variables. See Bethlehem (1988) for details. van Berkel, van der Doef and Schouten (2020) links the upper limit for the bias to the coefficient of variation of the response probabilities $\text{CV}(\rho_y)$

$$|\text{Bias}(\bar{y})| \leq \frac{S_{\rho_y} S_Y}{\bar{\rho}_y} = \text{CV}(\rho_y) S_Y.$$

A lower CV of the response probabilities means a smaller nonresponse bias, and under a given survey design s_ϕ, the response probabilities can be estimated conditional on the sample of size n

and the individual characteristics on survey variables Y_i

$$\hat{\rho}_{y,i} = \rho_{y,i}(s_\phi) = \Pr(u_{i,s_\phi} = 1|Y_i),$$

for $i = 1, 2, \ldots, n$, where u_{i,s_ϕ} represents the response outcome; $\hat{\rho}_{y,i} = \rho_{y,i}(s_\phi)$ is referred to as the response propensity and can be estimated by a logit or a probit model (Bethlehem et al. 2011). In deriving the design parameters at the stratum level, it is assumed that all units in a stratum g have the same response propensity and cost under a specific strategy. For all units in a survey sample, under one allocation of strategies s_ϕ, let $\rho_y(s_\phi)$ be the vector of response propensities

$$\rho_y(s_\phi) = (\rho_{1,1}(s_\phi), \ldots, \rho_{1,n_1}(s_\phi), \ldots, \rho_{G,1}(s_\phi), \ldots, \rho_{G,n_G}(s_\phi))',$$

and let $C(s_\phi)$ be the vector of costs

$$C(s_\phi) = (C_{1,1}(s_\phi), \ldots, C_{1,n_1}(s_\phi), \ldots, C_{G,1}(s_\phi), \ldots, C_{G,n_G}(s_\phi))',$$

where $\rho_{g,i}(s_\phi) = \rho_{g,j}(s_\phi)$ and $C_{g,i}(s_\phi) = C_{g,j}(s_\phi)$, $\forall i \neq j \in \{1, \ldots, n_g\}$.
The coefficient of variation of the response propensities, $\mathrm{CV}(\rho_y(s_\phi))$, is a quality indicator of the survey design s_ϕ, as well as the objective function in the mathematical optimization problem. The optimization amounts to choosing the optimal design from $\mathcal{S}_{1,T}^G$ with the minimal CV subjected to some constraints on other quality and cost indicators such as response rate, RR, and required budget per respondent, B

$$\begin{aligned} \underset{s_\phi \in \mathcal{S}_{1,T}^G}{\text{minimize}} \quad & \mathbb{E}\left(\mathrm{CV}(\rho_y(s_\phi))\right) \\ \text{subject to} \quad & \mathbb{E}\left(\mathrm{RR}(\rho_y(s_\phi))\right) \geq \mathrm{RR}_{\text{lower}}, \\ & \Pr\left(\mathrm{B}(C(s_\phi), \rho_y(s_\phi)) \geq \mathrm{B}_{\text{upper}}\right) \leq 0.10, \end{aligned}$$

where the boundary values of the constraints are determined by the survey stakeholder; details of the computation of the indicators will follow.
By solving the optimization problem, the optimal survey design is expected to reduce response variation relative to the survey variables, with the implication that the response propensity $\rho_y(s_\phi)$ is the projection onto the space spanned by the vector of stratum membership indicators associated with the survey variables. Hence, the stratification related to the survey variables is expected to perform best.

The following sections detail the proposed method of generating strata related to survey variables and the derivation of the design parameters at the stratum level, followed by the optimization problem solving. To demonstrate the advantages of the proposed stratification method, the optimal design outputs based on different stratification methods will be compared in terms of their utility in minimizing nonresponse bias against the same criterion.

3 Stratification and Survey Design Parameters Estimation

This section briefly describes two stratification methods oriented to explain response behavior and costs, and elaborates on the proposed method that can explain the heterogeneity in both response behavior and survey variables. The ensuing estimation of survey design parameters such as response propensities and costs is compatible with any stratification scheme so that the same optimization problem can have different sets of solutions corresponding to different schemes.

3.1 Classification Tree for Stratification

The stratification of the target population can be performed with a classification tree algorithm that determines the splitting values to form subgroups that correspond to the objective of adaptive designs. For a subject i in the sample, let g_i be the vector of binary indicators of the stratum membership $g_i = (g_{1,i}, \ldots, g_{G,i})'$ where the target G strata are formed to explain, for example, the response outcomes under each strategy. Under a strategy of using web questionnaires, the algorithm may identify two strata that have approximately the same response rate, but different response rates if they are approached by face-to-face visits.

In an adaptive design with emphasis on reducing survey costs, the stratification method targets the detection of heterogeneity in survey costs, which will be referred to as CostX. The algorithm takes the auxiliary data as input, and clusters units into subgroups that differ most in cost-related variables such as the number of in-person visits. This method is used in this paper to represent the cost-oriented stratification, although no studies have been conducted to implement this stratification in survey design.

When the objective of adaptive design is to minimize nonresponse bias, the heterogeneity in response behavior becomes of interest. A stratification method, to be called ResponseX, is to use

the auxiliary data as input to the algorithm to account for differences in responses to different strategies such as face-to-face interviews; the resulting subgroups, however, are not related to the survey variables of interest. This response-oriented stratification has been implemented in the Dutch Health Survey (van Berkel et al. 2020).

To allow stratification to account for the heterogeneity in both response behavior and survey variables of interest, in the proposed method, termed Response$\hat{Y}$, the survey variables predicted by auxiliary data serve as inputs to the algorithm to explain the response behavior. In contrast to the ResponseX method, this novel method is also response-oriented but determines stratification based on the survey variables that are not observed at the beginning of data collection, which entails prediction of selected survey variables prior to estimating design parameters.

3.2 Predicting Key Survey Variables

For a subject i, $i = 1, 2, \ldots, n$, in the sample, let $x_i = (x_{1,i}, \ldots, x_{m,i})'$ be the vector of m auxiliary variables available at the start of data collection. Then the j survey variables $y_i = (y_{1,i}, \ldots, y_{j,i})'$ can be predicted as follows:

$$\hat{y}_i = f^{-1}(\beta x_i), \tag{1}$$

where $\beta = (\beta_{1,1}, \ldots, \beta_{m,j})'$ is the vector containing the corresponding regression coefficient for m-th auxiliary variable in modeling j-th survey variable; $f^{-1}(.)$ is the inverse of the link function that depends on the probability distribution of the survey variable.

The regression parameters in (1) can be derived from the models of survey variables in a previous round of the repeated survey or another similar survey. For a subject i in the sample, let y_i be the vector of observed survey variables,

$$y_i = (y_{C,1,i}, \ldots, y_{C,j_C,i}, \ldots, y_{D,1,i}, \ldots, y_{D,j_D,i})',$$

where $y_{C,i} = (y_{C,1,i}, \ldots, y_{C,j_C,i})'$ contains the j_C continuous survey variables, and $y_{D,i} = (y_{D,1,i}, \ldots, y_{D,j_D,i})'$ are the j_D dichotomous survey variables.

The observed continuous survey variables can be modeled using an identity link function

$$y_{C,i} = \beta_C x_i + \varepsilon_i, \tag{2}$$

and the observed probability of success of dichotomous survey variables can be modeled using a binomial link function

$$\Pr\left(y_{D,i}=1|x_i\right)=\Phi\left(\beta_D x_i\right), \tag{3}$$

where $\beta_C=(\beta_{C,1,1},\ldots,\beta_{C,m,j_C})'$ and $\beta_D=(\beta_{D,1,1},\ldots,\beta_{D,m,j_D})'$ are the vectors containing the regression coefficient for m-th auxiliary variables in modeling j_C-th continuous and j_D-th dichotomous survey variable. In modeling continuous ones, the error terms of the uncertainty $\varepsilon_i\sim\mathcal{N}\left(0,\sigma_y^2\right)$.

To keep the scope manageable, a survey variable that is neither continuous nor dichotomous is dichotomized first. Because predictions of multinomial models can be complex, they are left to future research.

For each sample unit, the models predict the values of continuous survey variables and the probability of success of dichotomous ones given the observed auxiliary data. The classification tree algorithm takes these predicted survey variables as inputs to explain the response outcomes and outputs a stratification scheme. All sample units are divided into several subgroups that are homogeneous in response behavior with respect to the survey variables of interest. The units that are predicted to be in different ranges of values or probabilities of the survey variables may respond differently to the data collection strategies. Design parameters, such as response propensities and costs of each stratum, are estimated before embarking on the optimization problem, i.e., finding the optimal allocation of strategies to these subgroups.

3.3 Estimating Survey Design Parameters

Adaptive survey designs determine the allocation of data collection strategies by maximizing a quality objective subject to constraints such as budgets. The quality and cost constraints are formulated as indicators composed of design parameters. In this paper, two sets of survey design parameters are the focus:

1. response propensities, $\rho_{y,T,i}(s_{1,T})$, per unit i and strategy $s_{1,T}$.

2. costs, $C_{T,i}(s_{1,T})$, per unit i and strategy $s_{1,T}$.

Instead of modeling response propensities and costs on the stratum level, models for response propensities and costs are constructed on the individual sample unit level. It offers more flexibility, and stratum propensities and costs may be derived from the individual ones.
A probit model is applied to estimate the cumulative individual response propensities from phase one through phase T, that is, to model the response outcomes at the end of phase T. Each sample unit i has a response ability represented as a continuous latent variable $Z_{T,i}(s_{1,T})$, and a response is obtained when this latent variable is larger than zero,

$$u_{T,i}(s_{1,T}) = \begin{cases} 1, & Z_{T,i}(s_{1,T}) > 0, \\ 0, & Z_{T,i}(s_{1,T}) < 0, \end{cases}$$

where $u_{T,i}(s_{1,T})$ is the indicator of the response of unit i at the end of phase T under the strategy $s_{1,T}$, and $Z_{T,i}(s_{1,T}) \sim \mathcal{N}\left(\mu(s_{1,T}), \sigma(s_{1,T})\right)$ for some $\mu(s_{1,T})$ and $\sigma(s_{1,T})$

$$\rho_{y,T,i}(s_{1,T}) = \Pr\left(Z_{T,i}(s_{1,T}) > 0\right).$$

Given that strategy $s_{1,T}$ is allocated to a unit, let $\alpha_T(s_{1,T})$ be a vector of regression coefficients through phase T for the binary indicators of the stratum membership g_i. The model can be written as follows:

$$Z_{T,i}(s_{1,T}) = \alpha_T(s_{1,T})g_i + \varepsilon^Z_{T,i}, \tag{4}$$

where $\varepsilon^Z_{T,i} \sim \mathcal{N}(0,1)$ is an error term for the uncertainty of response of the subject.
Similarly, a linear model is used to model the cumulative individual costs from phase one through phase T. Given that strategy $s_{1,T}$ is allocated to a unit, let $\gamma_T(s_{1,T})$ be a vector of regression coefficients through phase T for the binary indicators of the stratum membership g_i. The model can be written as follows:

$$C_{T,i}(s_{1,T}) = \gamma_T(s_{1,T})g_i + \varepsilon^C_{T,i}, \tag{5}$$

where $\varepsilon^C_{T,i} \sim \mathcal{N}\left(0, \sigma^2(s_{1,T})\right)$ is an error term for the uncertainty of the survey cost of the subject. The error terms are modeled as independent normal, but other distributions may be applicable when costs are skewed or attain values close to zero. For self-administered modes, like web or mail, the cost is more or less fixed for all sample units, which means there is no need to model the costs of these modes.

3.4 Bayesian Analysis

The models for both survey variables and design parameters are estimated in a Bayesian manner. For design parameter models, the added benefit is that prior distributions may be elicited from historic survey data of a repeated or similar survey, or expert judgment (Schouten et al. 2018). Besides, it is feasible to account for the uncertainty in survey design parameters and any function of these parameters, which enhances the monitoring of adaptive survey designs. For example, it is possible to monitor the uncertainty of the $\mathrm{CV}\left(\rho_y(s_{\phi_{\mathrm{opt}}})\right)$ of the optimal design. A similar motivation to assign prior distributions to regression parameters in survey variable models is to capture the uncertainty in the survey variables on which the stratification is based.
For the classification tree algorithm, not only the parameters of the tree, but also the tree structure becomes random under the Bayesian setting, which means the tree structure has a posterior distribution. In practice, however, one tree is necessary, which contrasts with the Bayesian viewpoint. Therefore, a conventional classification tree algorithm is applied to generate a single tree in this paper.

3.4.1 Prior distributions

The model parameters in (2), (3), (4), and (5) receive prior distributions. For a dichotomous outcome, a probit model is applied and its regression slope parameters $\theta_D \in \{\beta_D, \alpha_T(s_{1,T})\}$ receive normal prior distributions.

$$\theta_D \sim \mathcal{N}\left(\mu_{\theta_D}, \Sigma_{\theta_D}\right). \tag{6}$$

For a continuous outcome, a linear model is applied and its regression slope parameters $\theta_C \in \{\beta_C, \gamma_T(s_{1,T})\}$ receive normal prior distributions:

$$\theta_C \sim \mathcal{N}\left(\mu_{\theta_C}, \Sigma_{\theta_C}\right), \tag{7}$$

and the regression dispersion parameter receives inverse Gamma prior distributions. The error term variances $\zeta \in \{\sigma_y^2, \sigma^2(s_{1,T})\}$ are modeled as

$$\zeta \sim \Gamma^{-1}(a_\zeta, b_\zeta). \tag{8}$$

The hyperparameters in (6), (7), and (8) can be elicited from historic or similar survey data or expert knowledge.

3.4.2 Posterior distributions

The aim is to first derive the posterior predictive distributions of the key survey variables y_i given the observed auxiliary data. Then the posterior distributions of response propensities and costs are derived given the observed response outcomes, realized costs, and the stratification scheme formed by expected values of key survey variables. The expressions for the full conditional distributions of the regression coefficients are derived to apply Markov Chain Monte Carlo (MCMC) methods to generate draws from the posterior distributions.
The observed data consists of the following:

1. The survey outcome vector: y_i
2. The response outcome up to per phase per sample unit: $u_{T,i}$
3. The realized costs up to per phase per sample unit: $c_{T,i}$
4. The auxiliary vector: x_i

The posterior predictive distributions of the key survey variables $p(y_i|y, x)$ come as by-products of Gibbs samplers applied to the posterior distributions of survey variable model parameters $p(\beta_C, \beta_D|y, x)$. The classification tree algorithm takes the expected values of the key survey variables $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ as inputs to form G strata, which results in a categorical stratum indicator variable g_i for each sample unit.
In the following, let $\rho_y(s_{1,T})$ and $C(s_{1,T})$ be the shorthand for the vectors of response propensities and costs over all sample units for a particular data collection strategy. In the same manner, let u_T, c_T, and g denote the vectors of response outcomes, realized costs, and stratum indicator variables over sample units. To shorten the expressions, let α, γ, and σ^2 be the vectors of regression slope and dispersion parameters over phases. For simplicity, the dependence on the hyperparameters is not a consideration.
The joint posterior distribution of interest is

$$p(\rho_y(s_{1,T}), C(s_{1,T})|u_T, c_T, g). \tag{9}$$

The joint density follows from integration over all possible combinations of regression parameters α, γ, σ^2 and cannot be written in closed form. A Gibbs sampler can be applied to the joint density

of the regression parameters α, γ, σ^2

$$p(\alpha, \gamma, \sigma^2 | u_T, c_T, g). \tag{10}$$

An approximation to the joint density in (9) comes as an important by-product of a Gibbs sampler applied to (10); per draw the response propensities and costs for a particular strategy can be computed by (4) and (5). A Gibbs sampler for (10) involves repeated draws from the conditional densities of each regression parameter, given the observed data and the other regression parameters. Appendix A contains expressions for the full conditionals of each regression parameter.

Based on the individual response propensities and cost parameters, their counterparts on the stratum level can be derived. It is assumed that all units in a stratum have the same response propensity and cost through phase T under the strategy $s_{1,T}$. Let $\rho_g(s_{1,T})$ and $C_g(s_{1,T})$ be the shorthand for the vectors of stratum response propensities and costs for a particular data collection strategy

$$\rho_g(s_{1,T}) = \frac{1}{\sum_{i=1}^{n} \delta_{g,i} d_i} \sum_{i=1}^{n} \delta_{g,i} d_i \rho_y(s_{1,T}), \tag{11}$$

$$C_g(s_{1,T}) = \frac{1}{\sum_{i=1}^{n} \delta_{g,i}} \sum_{i=1}^{n} \delta_{g,i} C(s_{1,T}), \tag{12}$$

where $\delta_{g,i}$ indicates if a sample unit i is in the stratum g, and d_i denotes the design or inclusion weight for the unit i. For the optimization purpose, the focus is on functions of the design parameters corresponding to the quality objectives and constraints, which will be developed in the following.

4 Optimization and Criterion for Determining Optimal Stratification

This section elaborates the quality and cost indicators in the optimization problem presented in section 2.2, and briefly outlines the optimization practice under a Bayesian analysis. Also, this section introduces a criterion to judge the utility of optimal design solutions based on different stratification schemes in minimizing nonresponse bias.

4.1 Quality and Cost Indicators

This paper considers three functions of design parameters: the response rate, the required budget per respondent, and the coefficient of variation of the response propensities. See Nishimura, Wagner and Elliott (2016) for a discussion of other indicators. In the optimization problem, for each allocation of strategies $s_\phi \in \mathcal{S}_{1,T}^G$, let $\rho_g(s_\phi)$ and $C_g(s_\phi)$ denote the vectors of stratum response propensity and cost respectively for units in the stratum g, $g = 1, 2, \ldots, G$, under one allocation of strategies s_ϕ. For a sample unit i, $i = 1, 2, \ldots, n$, let d_i denote the design or inclusion weight and let $\delta_{g,i}$ indicate if a unit i is in the stratum g. It is assumed that a unit i, $i = 1, 2, \ldots, n_g$, in a stratum g has the same response propensity and cost as other units in the same stratum.

The overall weighted response rate, RR, under the allocation of strategies s_ϕ is

$$\mathrm{RR}(\rho_y(s_\phi)) = \frac{1}{\sum_{i=1}^n d_i} \sum_{g=1}^{G} \sum_{i=1}^{n} \delta_{g,i} d_i \rho_g(s_\phi), \tag{13}$$

and the required budget per respondent, B, associated with s_ϕ is

$$\mathrm{B}(C(s_\phi), \rho_y(s_\phi)) = \frac{1}{n} \sum_{g=1}^{G} \frac{n_g C_g(s_\phi)}{\rho_g(s_\phi)}, \tag{14}$$

and the coefficient of variation of the response propensities, CV, is

$$\mathrm{CV}(\rho_y(s_\phi)) = \frac{\sqrt{\frac{1}{\sum_{i=1}^n d_i} \sum_{g=1}^{G} \sum_{i=1}^{n} \delta_{g,i} d_i (\rho_g(s_\phi) - \mathrm{RR}(\rho_y(s_\phi)))^2}}{\mathrm{RR}(\rho_y(s_\phi))}, \tag{15}$$

where $\sum_{i=1}^n d_i = N$ for many customary sampling designs.

The prior and posterior distributions for these three functions are determined by the prior and posterior distributions of the response propensities and costs. Since these functions are random variables under Bayesian analysis and it is not feasible to evaluate their posteriors for each design s_ϕ, this paper chooses an alternative approach to mimic their posterior distributions.

4.2 Optimization under the Bayesian Analysis

The objective of this mathematical optimization problem is to search for an optimal design $s_{\phi_{\mathrm{opt}}}$ from the feasible region $\mathcal{S}_{1,T}^G$, i.e., the optimal allocation of strategies across strata, which will

yield minimal variation in response outcomes with respect to the range of predicted values or probabilities of the survey variables.

Since the survey design parameters are estimated through Bayesian analysis, the objective function and constraints are random. Besides, the number of design solutions increases exponentially as the number of strata grows. It is infeasible to evaluate the posterior distributions of quality and cost indicators for all $|\mathcal{S}_{1,T}^{G}|$ possible adaptive design solutions given that no explicit closed forms exist for the distributions. Numerical approximation with the Gibbs sampler is also a computationally intensive task.

This paper constructs artificial posterior distributions of the objective functions and constraints rather than evaluating, for example, the model of response propensities $\rho_y(s_\phi)$ for every design solution to get their posterior distributions. Given the observed data, let $\rho_g(s_{1,t}) = \{\rho_{1,g}(s_{1,1}), \ldots, \rho_{T,g}(s_{1,T})\}$ and $C_g(s_{1,t}) = \{C_{1,g}(s_{1,1}), \ldots, C_{T,g}(s_{1,T})\}$ be the collections of estimates of response propensities and costs respectively derived from (11) and (12) for each stratum g under each strategy $s_{1,t}$.

According to a particular allocation of strategies s_ϕ, for a stratum g, the response propensity $\rho_g(s_\phi)$ can be any value in $\rho_g(s_{1,t})$, and the cost $C_g(s_\phi)$ can be any value in $C_g(s_{1,t})$. For example, when the first stratum receives a strategy $s_{1,T}$, its response propensity $\rho_1(s_\phi)$ takes the estimate $\rho_{T,1}(s_{1,T})$. After assigning the response propensities and costs for all strata, inserting them in (13), (14), (15) gives the corresponding values of response rate, required budget per respondent, and coefficient of variation of response propensities. In this way, the artificial posterior distributions of the objective functions and constraints can be constructed without simulating and modeling the response outcomes u_{s_ϕ} and costs c_{s_ϕ} under the particular allocation of strategies s_ϕ. After repeating this process for all the design solutions $s_\phi \in \mathcal{S}_{1,T}^{G}$, the optimization problem can be solved and outputs optimal survey designs based on the specified stratification scheme. If other stratification methods are applied, the stratification scheme and the estimates of response propensities $\rho_g(s_{1,t})$ and costs $C_g(s_{1,t})$ will change, as well as the optimization result. To demonstrate the advantages of the proposed stratification method, the utility of designs optimized based on different stratification schemes in reducing nonresponse bias should be comparable.

4.3 Determining Optimal Stratification

Different optimal designs based on different stratification schemes cannot be compared directly regarding their expected values of the coefficient of variation of response propensities, $\mathbb{E}\left(\mathrm{CV}(\rho_y(s_{\phi_{\mathrm{opt}}}))\right)$, because their response propensities consist of different stratum propensities that are estimated from different models. When the stratification scheme varies, for a subject i, models for design parameters in (4) and (5) have different inputs g_i, as the same sample unit may be classified into different homogeneous groups of units. In consequence, the derivation of stratum design parameters in (11) and (12) also have different grouping indicators $\delta_{g,i}$. For optimal designs based on different stratification schemes, therefore, it is crucial to re-evaluate their response propensity models with the same inputs which are the predicted survey variables. The survey variables are predicted by auxiliary data that remain unchanged regardless of changes in the stratification scheme.

Given an optimal survey design is conducted, the response outcome $u_{i,s_{\phi_{\mathrm{opt}}}}$ is reassigned to each unit. For example, the first stratum of size n_1 receives strategy $s_{1,T}$, the response outcomes of units in that stratum take observed outcomes $u_{T,i}$, where $i = 1, 2, \ldots, n_1$. Let $Z_i(s_{\phi_{\mathrm{opt},i}})$ be the vector of latent response ability for each sample unit, and $\alpha_{s_{\phi_{\mathrm{opt}}}}$ be the vector of regression coefficients for the expected values of survey variables. For convenience, the subscript of the expected values corresponding to the sample unit i is omitted. The stratification-assessed response propensity model can be written as follows:

$$Z_i(s_{\phi_{\mathrm{opt},i}}) = \alpha_{s_{\phi_{\mathrm{opt}}}} \mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right) + \varepsilon^{Z}_{i,s_{\phi_{\mathrm{opt}}}}, \tag{16}$$

where $\varepsilon^{Z}_{i,s_{\phi_{\mathrm{opt}}}} \sim \mathcal{N}(0,1)$ is an error term for the uncertainty of response of the subject.

The individual response propensities can then be derived as $\rho_{y,i}(s_{\phi_{\mathrm{opt},i}}) = \Pr\left(Z_i(s_{\phi_{\mathrm{opt},i}}) > 0\right)$. The regression coefficients receive normal prior distributions $\alpha_{s_{\phi_{\mathrm{opt}}}} \sim \mathcal{N}\left(\mu_\alpha(s_{\phi_{\mathrm{opt}}}), \Sigma_\alpha(s_{\phi_{\mathrm{opt}}})\right)$, and their posterior distributions $p(\alpha_{s_{\phi_{\mathrm{opt}}}}|u_{i,s_{\phi_{\mathrm{opt}}}}, \mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right))$ are derived from a Gibbs sampler. The Gibbs sampler involves repeated draws from the conditional densities of each regression parameter, given the reassigned response outcomes, the expected values of survey variables, and the other regression parameters. Appendix A contains expressions for the full conditionals of each regression parameter. The posterior distribution of individual response propensities $p(\rho_{y,i}(s_{\phi_{\mathrm{opt},i}})|u_{i,s_{\phi_{\mathrm{opt}}}}, \mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right))$ comes as a by-product; per draw the response propensities for a

particular optimal design can be computed by (16).

The assessment criterion, which is the coefficient of variation of the individual response propensities under a specific optimal design, can be derived as follows:

$$\mathrm{CV}\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right), s_{\phi_{\mathrm{opt},i}}\right) = \frac{\sqrt{\frac{1}{\sum_{i=1}^{n} d_i}\sum_{i=1}^{n} d_i \left(\rho_{y,i}(s_{\phi_{\mathrm{opt},i}}) - \frac{1}{\sum_{i=1}^{n} d_i}\sum_{i=1}^{n} d_i \rho_{y,i}(s_{\phi_{\mathrm{opt},i}})\right)^2}}{\frac{1}{\sum_{i=1}^{n} d_i}\sum_{i=1}^{n} d_i \rho_{y,i}(s_{\phi_{\mathrm{opt},i}})}, \quad (17)$$

where d_i represents the invariant design or inclusion weight for the unit i.

This CV criterion explicitly depends on the expectations of the survey variables given the auxiliary vector, which means it will yield a different value for any other choice of survey and auxiliary variables. Since the predicted survey variables are random under the Bayesian analysis, the CV could also be dependent on repeated draws from the posterior predictive distributions of the survey variables. The stratification-assessed response propensity model itself, however, will also be random, which makes it infeasible to derive the posterior of individual response propensities.

This criterion is valid because the prediction of the survey variables does not depend on the grouping pattern of the units. More importantly, it directly measures the variation in the response outcomes across the range of predicted values or probabilities of the survey variables. Different optimal design solutions based on different stratification schemes can be compared against this criterion. The solution that incurs the minimum CV is the optimal design solution, and the corresponding stratification is subsequently the optimal stratification.

5 A Case Study on the Dutch Health Survey

This section applies the proposed and two other response- and cost-oriented stratification methods described in section 3.1 to Dutch Health Survey and shows that the proposed Response$\hat{Y}$ stratification method is the optimal approach for minimizing nonresponse bias. The optimized adaptive survey designs constructed based on the optimal stratification are considered to be optimal and robust.

5.1 Key Survey Variables

The Dutch Health Survey is to provide a complete overview of developments in the health, medical contacts, lifestyle, and preventive behaviour of the Dutch population (CBS n.d.). The target population is all persons living in private households. For respondents under the age of 12, the survey questions are answered by a parent or guardian. For the Health Survey stakeholders, stability and low nonresponse bias in key survey variables are the highest priority; surveys are used to compare health statistics over time. This priority translates to our objective of efficiently balancing nonresponse against predicted survey variables.

This paper uses data collected in the last three quarters of 2017, and the first quarter of 2018. The sample units under the age of 12 are excluded. Three key survey variables are selected:

1. *How is generally (1: your health 2: your child's health)?*
   This variable is measured on a 5-point scale: *Very good, Good, Goes well, Bad, Very bad*. The first two categories are recoded as *healthy*, and the last three are *unhealthy*.

2. *Do you ever smoke?*
   This is a dichotomous variable: *Yes/No*.

3. *Obesity*.
   It is transformed from body mass index (BMI) calculated by height and weight measurements (weight in kilos divided by squared height in meters):
   How tall (1: are you 2: is your child)? It is the length in centimeters, without shoes;
   What is the weight in kilos (1: you 2: your child)? (We mean the weight before pregnancy.) It's the weight in whole kilos, without clothes.
   For adults, a BMI over 30 indicates *obesity*. For teenagers, the criteria depend on different ages.

5.2 Survey Design and Adaptive Strategies

The Dutch Health Survey has a sequential mixed-mode design with web interview as the starting mode. Follow-up of web mode nonresponse is done through short face-to-face (F2F-short) or extended face-to-face (F2F-extended) interview. The F2F-extended means an additional round of

face-to-face visits for those sample units that are soft refusals after the first three face-to-face visits.

The survey design consists of three phases with different modes: phase 1 (Web), phase 2 (F2F-short), and phase 3 (F2F-extended). Let s_1, s_2, and s_3 denote the actions, which is the choice of mode, taken in each phase. The survey variables can be observed in any phase. It is possible for a unit to receive all three modes of the interview, going through all the phases. Reaching a phase means that the unit has been through any previous phase. Therefore, a unit may receive one of three strategies, $\mathcal{S}_{1,3} = \{s_{1,1}, s_{1,2}, s_{1,3}\}$, where $s_{1,1} = \{(s_1)\}$, $s_{1,2} = \{(s_1, s_2)\}$, and $s_{1,3} = \{(s_1, s_2, s_3)\}$.

5.3 Stratification and Design Parameters

The classification tree algorithm is implemented in R with the `rpart` package (Therneau and Atkinson 2019), and three key survey variables are modeled by Bayesian probit regression with the `MCMCpack` package (Martin, Quinn and Park 2011). In order to integrate the optimization with the estimation of survey design parameters, I programmed the Gibbs sampler in R to model the response propensities and costs.

This case study applies the proposed Response$\hat{Y}$ stratification method. As for the response-oriented ResponseX and the cost-oriented CostX methods, applying them to the same data, the former one yields ten strata and with the latter, seven strata; the detailed stratification schemes and their estimated design parameters are not displayed.

Regarding the Response$\hat{Y}$ method, before the stratification, the posterior predictive distributions of survey variables are derived given the categorical auxiliary variables, including age, sex, income, marital status, level of education, migration background, receiving rent benefit or not, type of household, and urbanization level of the area of residence. The results of survey variable models can be found in Appendix B.

Expected values for general health, smoking, and obesity are treated as predicted values for the survey variables and used as inputs to the classification tree algorithm; since the survey variables are dichotomous, the predicted values are the probabilities of "success", i.e., the probabilities of being in the category coded as 1. The homogeneous subgroups are created to explain whether a

sample unit responds to the starting mode (Web) in phase 1. The algorithm determines where to split between 0 and 1 predicted probabilities of success of survey variables. Table 1 shows the resulting strata. The smoking probabilities split at 0.21, leading to group 1 consisting of sample units with smoking probabilities greater than or equal to 0.21. Other units with smoking probabilities less than 0.21 are further divided into eight groups based on obesity and health probabilities.

TABLE 1

Stratification based on the Predicted Probabilities of Success of Survey Variables.

Stratum	Smoking probabilities	Health probabilities	Obesity probabilities
1 (5841)	≥ 0.21		
2 (720)	< 0.21	< 0.56	
3 (1370)	< 0.21	≥ 0.86	< 0.06
4 (626)	≥ 0.13 & < 0.21	≥ 0.86	≥ 0.06
5 (371)	≥ 0.08 & < 0.13	≥ 0.86	≥ 0.06
6 (188)	< 0.08	≥ 0.86	≥ 0.06
7 (1240)	≥ 0.16 & < 0.21	≥ 0.56 & < 0.86	
8 (825)	< 0.16	≥ 0.56 & < 0.63	
9 (2016)	< 0.16	≥ 0.63 & < 0.86	

Note: Stratum size in parentheses. Total sample size is 13197.

The response propensities and costs per stratum and strategy are estimated to facilitate further analysis of quality and cost indicators and optimization of the adaptive designs. In the Bayesian analysis, non-informative priors are specified for the regression parameters of the models for response propensities and costs. The convergence properties of the Gibbs sampler are presented in Appendix C. Table 2 shows the estimated response propensities per stratum and strategy along with the corresponding 95% credible intervals.

Figure 1 further illustrates the difference in response propensities among nine strata under three strategies. Different groups of units, such as the first and the sixth stratum, may have varying tendencies to respond to the Web mode. It is also possible that units in two strata are close in their tendency to respond to Web mode but behave differently in response to F2F-short mode, as units

TABLE 2

Estimated Response Propensities per Stratum and Strategy.

Stratum	Web	Web → F2F-short	Web → F2F-extended
1	29.9% (0.288, 0.311)	51.3% (0.500, 0.526)	55.9% (0.546, 0.571)
2	31.2% (0.280, 0.346)	55.3% (0.516, 0.588)	57.1% (0.533, 0.607)
3	39.6% (0.370, 0.422)	62.7% (0.602, 0.651)	68.1% (0.655, 0.705)
4	39.6% (0.358, 0.436)	60.9% (0.570, 0.647)	67.3% (0.636, 0.709)
5	47.1% (0.422, 0.521)	67.4% (0.627, 0.720)	71.7% (0.672, 0.761)
6	54.8% (0.474, 0.618)	70.3% (0.636, 0.766)	73.4% (0.667, 0.795)
7	44.2% (0.414, 0.469)	60.9% (0.582, 0.637)	64.4% (0.618, 0.670)
8	46.6% (0.431, 0.498)	66.6% (0.634, 0.697)	67.7% (0.646, 0.707)
9	56.2% (0.541, 0.583)	71.3% (0.693, 0.733)	74.0% (0.721, 0.760)

Note: 95% Credible intervals in parentheses.

in the first two strata do. Smaller strata tend to have wider credible intervals, but the response propensities in the Web mode can still be distinguished from the propensities in the other two modes. Since there are not many units reaching the F2F-extended mode in phase 3, the classification tree algorithm cannot distinguish them from units reaching the F2F-short mode in phase 2 when stratifying.

It is conspicuous that the variation in overall response propensities can be significant if a single strategy is implemented uniformly over the entire sample. The optimization aims to find an optimal way to distribute different strategies among these unit groups, which will reduce the impact of nonresponse on the estimates of survey variables of interest. Since the objective is to

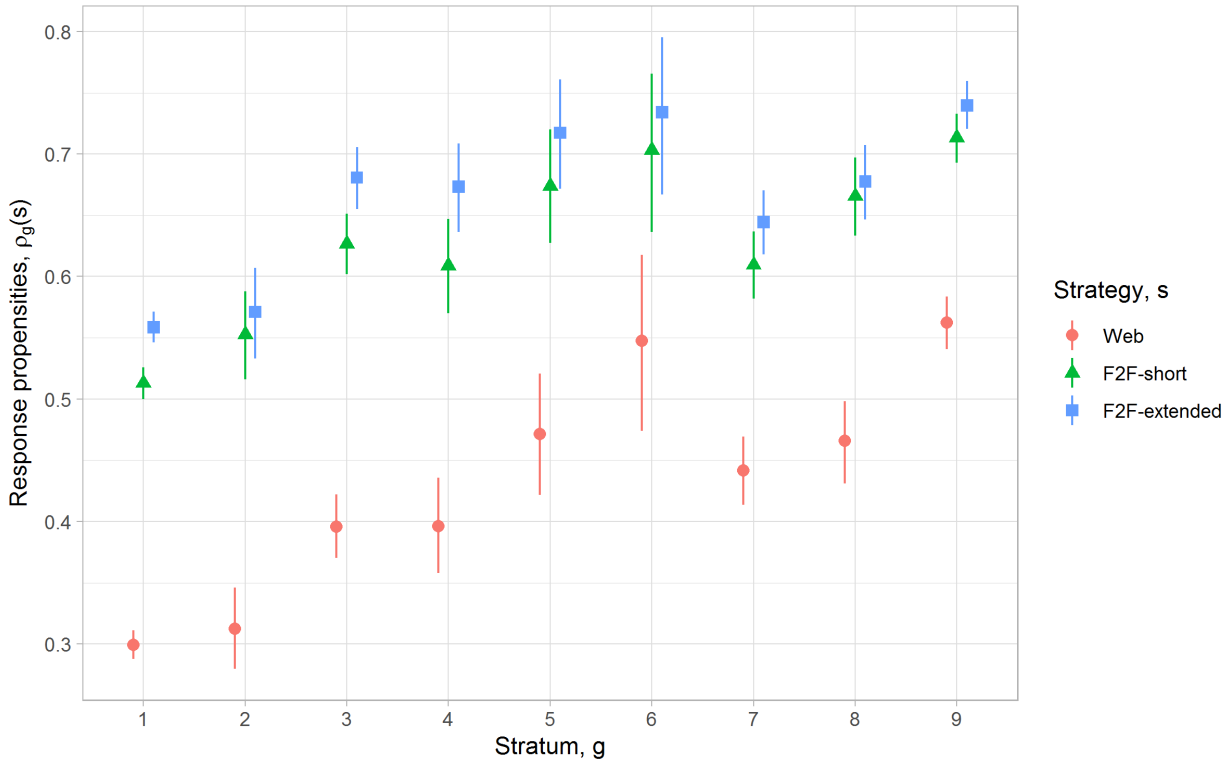

FIGURE 1: Estimated Response Propensities per Stratum and Strategy.

balance the responses, the estimates of costs per stratum and strategy are not shown to save space.

5.4 Optimization, Optimal Stratification, and Optimal Adaptive Designs

With the estimated response propensities and costs per stratum and strategy, the quality and cost indicators such as coefficient of variation of response propensities, response rate, and required budget per respondent can be derived to solve the optimization problem. The application of the proposed and two other response- and cost-oriented stratification methods leads to three sets of solutions whose utility in balancing the responses with respect to the predicted survey variables will be assessed to determine the optimal stratification. The adaptive survey design solutions based on such optimal stratification are regarded as the optimal solutions.

5.4.1 Optimization problem

As a sample unit may receive one of the three strategies and can belong to one of the nine groups when the Response$\hat{Y}$ method is applied, the collection of survey designs, i.e., the allocations of strategies, $\mathcal{S}_{1,3}^{9}$, has a total of 19683 solutions. Similarly, based on the ResponseX and CostX methods, the collections of survey designs have $\mathcal{S}_{1,3}^{10} = 59049$ and $\mathcal{S}_{1,3}^{7} = 2187$ solutions, respectively. The optimization corresponds to selecting the optimal designs from a solution set such that the objective function of the coefficient of variation of response propensities is optimized subject to the constraints on the response rate and required budget per respondent. For the Health Survey stakeholders, the lower limit of satisfactory response rate, $\mathrm{RR}_{\mathrm{lower}}$, is set at 50%, and the upper limit of required budget per respondent, $\mathrm{B}_{\mathrm{upper}}$, is set at 42 EUR. These conditions allow the filtering of the designs that satisfy the constraints, from which the objective function selects the designs with the minimal CV.

5.4.2 Optimal stratification

Since it is not feasible to directly compare the optimization outputs based on different stratification methods, the top five optimal designs of each solution set are re-evaluated against the criterion developed in section 4.3. In the Bayesian analysis, non-informative priors are specified for the regression parameters of the stratification-assessed response propensity models. The coefficients of variation of the response propensities for the fifteen design solutions are derived without the grouping structure, and their expectations are plotted along with the 95% credible intervals in Figure 2.

It is shown that most optimal solutions selected based on the ResponseX and CostX methods tend to yield higher variation in response outcomes with respect to the predicted survey variables compared to those based on the proposed Response$\hat{Y}$ method. Furthermore, there is no overlap in the 95% credible intervals of the expectations of coefficient of variation between the Response$\hat{Y}$ method and others. Only one solution based on the CostX method performs no less well. In general, the proposed stratification method in this paper is superior to two other response- and cost-oriented methods in terms of minimizing the impact of nonresponse bias on the survey estimates of interest.

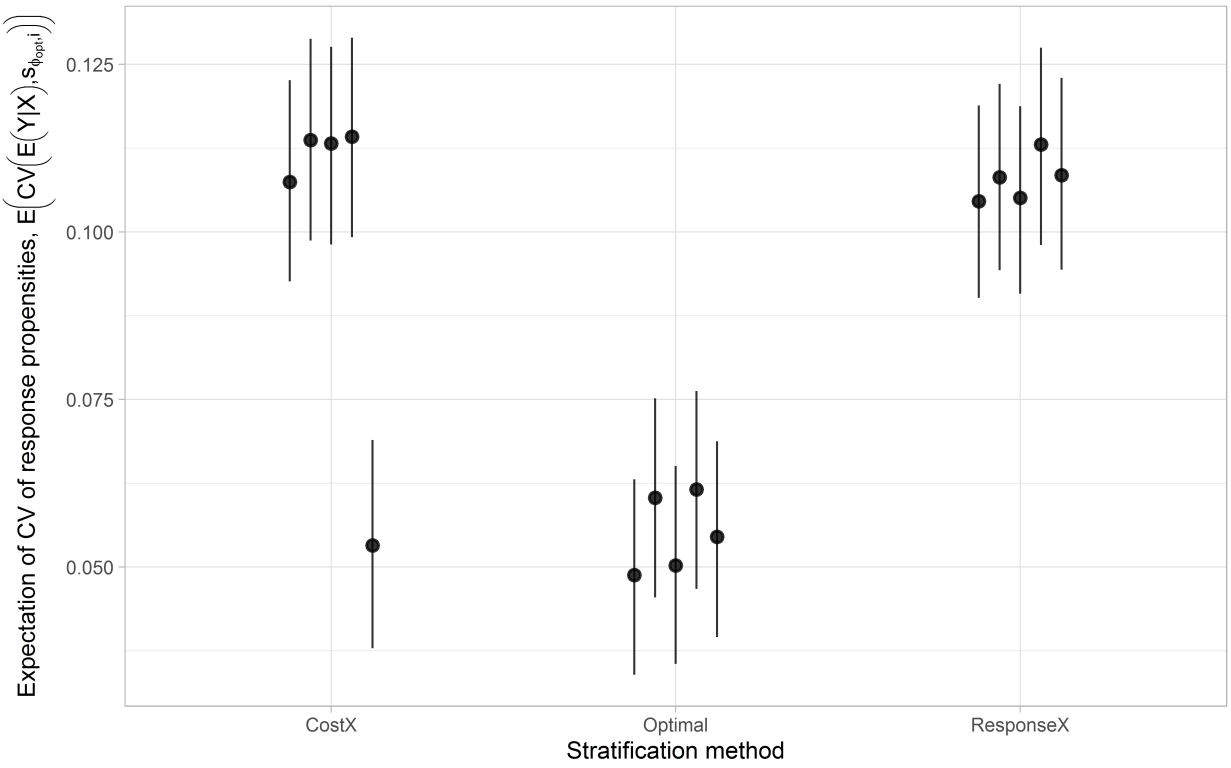

FIGURE 2: Expected Values of the Assessment Criterion, i.e., Coefficient of Variation (CV) of Response Propensities of Optimal Design Solutions Based on Different Stratification Methods.

5.4.3 Optimal adaptive survey designs

To eventually address the optimization problem, Table 3 provides the strategy allocations as well as the expected values of the quality and cost indicators for the five optimal adaptive survey designs based on the optimal stratification method Response$\hat{Y}$.

Meanwhile, it is important to emphasize that under the Bayesian analysis, the objective function and constraints are random, which means that no single design solution is superior, but rather there is a range of solutions with similar performance. In Figure 3, for the top 150 design solutions, their expectations of coefficient of variation of the response propensities are plotted against the expectations of budget per respondent and response rate. While the top five solutions stand out, their 95% credible intervals of coefficient of variation still partially overlap with those of a range of solutions. Some of these solutions also yield similar budgets per respondent and response rates. Overall, this figure illustrates the uncertainty of the optimal designs and suggests the need for further research on the sensitivity of the optimal designs.

TABLE 3

Optimal Designs with Minimal Coefficient of Variation (CV) Subject to Constraints on Budget per Respondent (B) and Response Rate (RR).

Stratum	Design 1	Design 2	Design 3	Design 4	Design 5
1	F2F-extended	F2F-extended	F2F-extended	F2F-extended	F2F-extended
2	F2F-extended	F2F-extended	F2F-short	F2F-short	F2F-extended
3	F2F-short	F2F-short	F2F-short	F2F-short	F2F-short
4	F2F-short	F2F-short	F2F-short	F2F-short	F2F-short
5	Web	F2F-short	Web	F2F-short	Web
6	Web	Web	Web	Web	F2F-short
7	F2F-short	F2F-short	F2F-short	F2F-short	F2F-short
8	Web	Web	Web	Web	Web
9	Web	Web	Web	Web	Web
$\mathbb{E}\,(\mathrm{CV})$	0.0727 (0.0584, 0.0874)	0.0729 (0.0587, 0.0869)	0.0730 (0.0589, 0.0876)	0.0734 (0.0592, 0.0873)	0.0777 (0.0636, 0.0923)
$\mathbb{E}\,(\mathrm{B})$	40.16 (39.07, 41.26)	41.08 (39.98, 42.20)	40.03 (38.94, 41.13)	40.95 (39.85, 42.08)	40.52 (39.43, 41.63)
$\mathbb{E}\,(\mathrm{RR})$	56.6% (0.557, 0.574)	57.1% (0.563, 0.580)	56.5% (0.556, 0.573)	57.0% (0.562, 0.579)	56.8% (0.560, 0.576)

Note: 95% Credible intervals in parentheses.

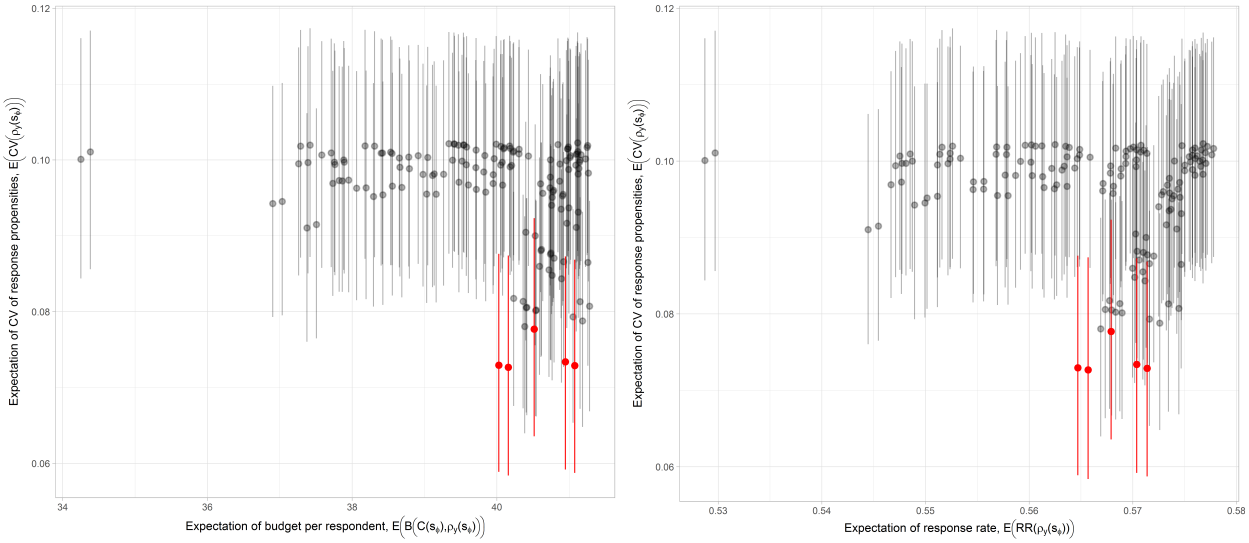

FIGURE 3: Expected Values of Coefficient of Variation (CV) of Response Propensities of Top 150 Design Solutions Satisfying the Constraints on Budget per Respondent (B) and Response Rate (RR). Top 5 Optimal Solutions Marked in Red.

6 Discussion

This paper answers the main research question by implementing an optimal approach to stratify the target population into subgroups and demonstrating its effectiveness and efficiency through a coefficient of variation criterion after the optimization of adaptive survey designs. The optimal adaptive survey designs constructed based on such stratification method perform robustly and minimizes nonresponse bias compared to the other two response- and cost-oriented stratification methods.

The entire process of conducting adaptive survey designs needs to be presented, including generating subgroups, estimating the survey design parameters such as response propensity and cost for each subgroup under each data collection strategy, and optimally filtering out the optimal design solutions, i.e., the optimal allocations of strategies to subgroups, subject to a number of specified constraints.

The optimal stratification method is implemented to generate subgroups based on predicted survey variables. Prior to stratification, this method entails predicting unobserved survey variables with available auxiliary data before data collection starts. Assisted by a wealth of administrative

data, many multi-phase, mixed-strategy surveys conducted by statistical institutes can employ this method. To minimize the impact of nonresponse on survey estimates, this method makes it possible to balance responses in subgroups consisting of ranges of predicted values or probabilities for the survey variables. Theoretically, optimized allocations of the data collection strategies to such subgroups would produce minimal response variation with respect to the survey variables. This paper accordingly develops a coefficient of variation criterion to quantify the utility of this method compared to other stratification methods that do not involve survey variables. The comparison is illustrated by a case study optimizing adaptive survey designs based on this novel stratification method and two other response- and cost-oriented methods. The case study shows that the optimal design solutions based on this novel stratification method yield less variation in response behavior with respect to the predicted survey variables.

The analysis performed in this paper is based on a Bayesian framework, except for the classification tree algorithm. The predicted survey variables take the expected values of their posterior predictive distributions given the model parameters repeatedly drawn in a Gibbs sampler. The classification algorithm takes the predicted survey variables as inputs and outputs several subgroups. Optimization is performed to search for the optimal allocations of the data collection strategies to these subgroups by comparing the quality indicators, such as the response rate or coefficient of variation of response propensities, and cost indicators. Since the quality and cost indicators are functions of the response propensities and costs, their posterior distributions can be derived as important by-products of the posterior distributions of the response propensities and costs whose estimates are provided by a Gibbs sampler. The posterior distribution of the coefficient of variation criterion for determining the optimal stratification is also derived in a similar manner.

It is to be emphasized that when solving the optimization problem under Bayesian analysis, the optimal design solutions are selected based on artificial posterior distributions of the quality and cost indicators. Since the size of a solution set grows exponentially when the number of subgroups increases, it is computationally laborious to derive the posterior distributions of the quality and cost indicators for every design solution by evaluating the corresponding response propensity and cost models. A more elegant optimization approach remains to be developed, which is not the focus of this paper. The alternative optimization approach adopted in this paper is

sufficient to output the optimal design solutions for the assessment of different stratification methods.

The novel stratification method excels in reducing nonresponse bias in survey estimates because it relies on the predicted survey variables to stratify, which means that this method is limited by the predictive power of the available auxiliary data for key survey variables. In an extreme case, there could be no available auxiliary variables that are predictors of the survey variables. It warrants sensitivity analyses and simulation studies in future research to examine the impact of inaccurate survey variable predictions on this stratification method. Moreover, this paper does not consider the strategy-dependent measurement error of the survey variables, which may impinge on the accuracy of predicted survey variables. Such an extension is worthwhile for mixed-mode survey design and is subject to future research.

This paper demonstrates the advantages of this novel stratification method over two other response- and cost-oriented methods, but space does not permit the assessment of more existing methods such as partial representativeness indicators and regression diagnostic measures (Schouten and Shlomo 2017; Wagner 2014). With the assessment criterion developed by this paper, however, future studies can specifically assess all the different stratification methods.

Ultimately, the optimal stratification allows parsimonious use of auxiliary information to account for heterogeneity in both response behavior and key survey variables. Such an application implies that the optimal allocations of strategies to the resulting strata should have enhanced capacity to reduce nonresponse bias effectively and efficiently.

Appendix A Full Conditionals

This appendix contains expressions for the full conditionals of the regression model parameters for survey variables, response propensities, and costs that are sampled in the Gibbs sampler.

A.1 Full conditionals for regression parameters of dichotomous variable models

Models for dichotomous survey variables and response propensities have the following general form:

$$Z_{D,i} = \Theta_D X_{D,i} + \varepsilon_{D,i},$$

where D represents one of three probit models (3), (4), and (16), for one of which $Z_{D,i}$ is a latent variable, and $X_{D,i}$ is a column vector of covariates, and the regression parameters are the slope parameters in the vector Θ_D. The latent variable $Z_{D,i}$, which is not of direct interest, is updated in the Gibbs sampler to derive the probability of success. Although three models have different outcomes and predictors, it is analogous for the derivation of full conditionals.
For the slope paramter $\Theta_D \in \{\theta_D, \alpha_{s_{\phi_{\text{opt}}}}\}$, the prior distribution is normal $\Theta_D \sim \mathcal{N}(\mu_{\Theta_D}, \Sigma_{\Theta_D})$. The full conditional distribution is also normal, denoted as,

$$(\Theta_D | u_D, Z_D, X_D) \sim \mathcal{N}(\mu_{FULL,D}, \Sigma_{FULL,D}). \tag{A1}$$

The expectation and covariance of the full conditional distribution for a model are derived as follows:

$$\Sigma_{FULL,D} = \left(\Sigma_{\Theta_D}^{-1} + X_D^T X_D\right)^{-1}, \tag{A2}$$

$$\mu_{FULL,D} = \Sigma_{FULL,D}\left(\Sigma_{\Theta_D}^{-1}\mu_{\Theta_D} + X_D^T Z_D\right). \tag{A3}$$

The latent variable follows the truncated normal distribution, $(Z_{D,i}|\Theta_D, X_D) \sim \mathcal{TN}(\Theta_D X_{D,i}, 1)$. When the outcome $u_{D,i} = 1$, then $Z_{D,i} > 0$ and $(Z_{D,i}|u_D, \Theta_D, X_D)$ is the normal distribution restricted to the positive real axis. For $u_{D,i} = 0$, $(Z_{D,i}|u_D, \Theta_D, X_D)$ is the normal distribution restricted to the non-positive real axis.

A.2 Full conditionals for regression parameters of continuous variable models

For continuous survey variable and cost models, the derivation of full conditionals is analogous. The model has the following general form:

$$Y_{C,i} = \theta_C X_{C,i} + \varepsilon_{C,i},$$

$$\varepsilon_{C,i} \sim \mathcal{N}(0, \zeta),$$

where C represents one of two linear models (2) and (5), for one of which $Y_{C,i}$ is either the continuous survey variable or the realized costs, and $X_{C,i}$ is a column vector of covariates, and $\varepsilon_{C,i}$ is phase-dependent error term. In this paper, the costs depend on the phase but not on the action itself. The regression slope parameters θ_C and the dispersion parameters ζ need to be updated.

For the slope parameter θ_C, the prior distribution is normal $\theta_C \sim \mathcal{N}\left(\mu_{\theta_C}, \Sigma_{\theta_C}\right)$. The full conditional distribution is also normal, denoted as,

$$(\theta_C | Y_C, X_C, \zeta) \sim \mathcal{N}(\mu_{FULL,C}, \Sigma_{FULL,C}). \tag{A4}$$

The expectation and covariance of the full conditional distribution for a model are derived as follows:

$$\Sigma_{FULL,C} = \left(\frac{1}{\zeta} X_C^T X_C + \left(\Sigma_{\theta_C}^2\right)^{-1}\right)^{-1}, \tag{A5}$$

$$\mu_{FULL,C} = \Sigma_{FULL,C} \left(\frac{1}{\zeta} X_C^T Y_C + \left(\Sigma_{\theta_C}^2\right)^{-1} \mu_{\theta_C}\right). \tag{A6}$$

For the dispersion parameter ζ, the prior distribution is inverse Gamma $\zeta \sim \Gamma^{-1}(a_\zeta, b_\zeta)$. The full conditional is also inverse Gamma:

$$(\zeta | \theta_C, Y_C, X_C) \sim \Gamma^{-1}(a_{FULL}, b_{FULL}), \tag{A7}$$

where the hyperparameters are derived as follows:

$$a_{FULL} = a_\zeta + \frac{n}{2}, \tag{A8}$$

$$b_{FULL} = b_\zeta + \frac{1}{2} \left(Y_C - X_C \theta_C\right)^T \left(Y_C - X_C \theta_C\right). \tag{A9}$$

Appendix B Results of Survey Variable Models

The following table shows the results of modeling three dichotomized survey variables which are general health, smoking, and obesity. The auxiliary data provides nine categorical predictors, including age, sex, income, marital status, level of education, migration background, receiving rent benefit, type of household, urbanization level of the area of residence. Not all of them are predictors of the survey variables. The best performing model for each survey variable is selected based on the deviance information criterion (Spiegelhalter, Best, Carlin and Van Der Linde 2002). The model fit is quantified by Bayesian R^2 (Gelman, Goodrich, Gabry and Vehtari 2019).

TABLE B.1

Probit Models of Survey Variables Predicted by Auxiliary Data.

	General Health	Smoking	Obesity
Age (12–17)			
18–24	-0.172 (-0.407, 0.069)	0.955 (0.722, 1.192)	0.270 (-0.059, 0.599)
25–29	-0.417 (-0.681, -0.161)	1.125 (0.857, 1.383)	1.009 (0.685, 1.314)
30–34	-0.575 (-0.842, -0.301)	1.274 (1.009, 1.534)	1.119 (0.797, 1.449)
35–39	-0.700 (-0.966, -0.426)	1.209 (0.950, 1.468)	1.253 (0.931, 1.560)
40–44	-0.855 (-1.118, -0.593)	1.265 (1.000, 1.529)	1.261 (0.952, 1.576)
45–49	-1.049 (-1.311, -0.794)	1.123 (0.862, 1.379)	1.286 (0.987, 1.586)
50–54	-1.074 (-1.327, -0.812)	1.064 (0.804, 1.329)	1.308 (1.005, 1.606)
55–59	-1.251 (-1.509, -0.992)	0.976 (0.709, 1.243)	1.356 (1.063, 1.655)
60–64	-1.460 (-1.723, -1.191)	0.941 (0.681, 1.208)	1.313 (1.019, 1.609)

(to be continued)

Table B.1: Probit Models of Survey Variables Predicted by Auxiliary Data (continued).

	General Health	Smoking	Obesity
65–69	-1.221 (-1.492, -0.943)	0.677 (0.396, 0.953)	1.365 (1.067, 1.666)
≥ 70	-1.416 (-1.683, -1.149)	0.460 (0.186, 0.736)	1.213 (0.920, 1.509)
Female (Male)	-0.056 (-0.127, 0.014)	-0.294 (-0.366, -0.222)	
Income (Missing)			
0–20% perc	-0.061 (-0.213, 0.089)	0.213 (0.056, 0.372)	0.108 (-0.063, 0.287)
20–40% perc	-0.151 (-0.302, 0.003)	0.280 (0.108, 0.450)	0.184 (-0.001, 0.377)
40–60% perc	-0.016 (-0.170, 0.138)	0.177 (0.009, 0.347)	0.060 (-0.126, 0.244)
60–80% perc	0.246 (0.086, 0.406)	0.080 (-0.088, 0.254)	-0.032 (-0.213, 0.153)
80–100% perc	0.447 (0.280, 0.611)	-0.102 (-0.277, 0.079)	-0.065 (-0.251, 0.126)
Migration (Native)			
1st generation non-western	-0.405 (-0.546, -0.270)	-0.137 (-0.283, 0.012)	
1st generation western	-0.093 (-0.247, 0.057)	0.116 (-0.036, 0.268)	
2nd generation non-western	-0.275 (-0.468, -0.069)	-0.123 (-0.325, 0.071)	
2nd generation western	-0.125 (-0.260, 0.008)	0.120 (-0.026, 0.259)	
Marital status (Single)			
Married or partnership	0.081 (-0.034, 0.197)	-0.296 (-0.398, -0.192)	

(to be continued)

Table B.1: Probit Models of Survey Variables Predicted by Auxiliary Data (continued).

	General Health	Smoking	Obesity
Widowed	0.136 (-0.030, 0.305)	-0.156 (-0.347, 0.031)	
Divorced	-0.045 (-0.179, 0.092)	0.101 (-0.033, 0.239)	
Urbanization level (Moderate)			
Non-urban		0.009 (-0.125, 0.138)	
Highly urban		0.121 (0.025, 0.219)	
Little urban		-0.026 (-0.132, 0.081)	
Very highly urban		0.052 (-0.062, 0.162)	
Household type (Single)			
Couple	0.255 (0.137, 0.381)	-0.079 (-0.197, 0.040)	
Couple with offspring	0.257 (0.130, 0.385)	-0.148 (-0.267, -0.029)	
Single with offspring	0.147 (-0.006, 0.300)	0.063 (-0.087, 0.217)	
Other	-0.042 (-0.442, 0.367)	0.169 (-0.197, 0.535)	
Education level (Primary)			
VMBO	-0.011 (-0.194, 0.170)	0.047 (-0.132, 0.227)	-0.036 (-0.236, 0.161)
HAVO-VWO-MBO	0.127 (-0.039, 0.298)	-0.077 (-0.246, 0.092)	-0.208 (-0.393, -0.023)
Bachelor HBO-WO	0.255 (0.064, 0.444)	-0.467 (-0.665, -0.267)	-0.448 (-0.658, -0.239)

(to be continued)

Table B.1: Probit Models of Survey Variables Predicted by Auxiliary Data (continued).

	General Health	Smoking	Obesity
Master HBO-WO	0.328 (0.108, 0.551)	-0.590 (-0.819, -0.361)	-0.734 (-0.998, -0.478)
Unknown	0.127 (-0.032, 0.289)	-0.007 (-0.172, 0.158)	-0.236 (-0.410, -0.059)
Receive rent benefit (No)	-0.436 (-0.563, -0.311)	0.195 (0.068, 0.323)	
Constant	1.375 (1.175, 1.572)	-1.466 (-1.681, -1.253)	-2.081 (-2.322, -1.856)
Bayesian R^2	0.134 (0.121, 0.148)	0.096 (0.084, 0.110)	0.038 (0.031, 0.046)
Observations	8331	8297	8178

Note: Reference groups in parentheses after predictors;

95% Credible intervals in parentheses below parameters.

Appendix C Convergence Properties of the Gibbs Sampler

The Gibbs sampler produces a sampling of a Markov chain that takes the posterior distribution of interest as its stationary distribution. In this paper, the Markov chain is initiated from two starting values, the maximum likelihood estimate and zero, respectively. Since the Markov chain may take time to converge to its stationary distribution, a burn-in period of 1000 iterations is discarded. After the burn-in period, the Markov chain moves another 5000 iterations through its parameter space. Figures C.1, C.2, and C.3 show Gibbs sampler runs for regression slope parameters in the survey variable models, the response propensity and cost models, and the stratification-assessed response propensity models. Two chains represented by different colors are mixing around a stationary point, leading to convergence.

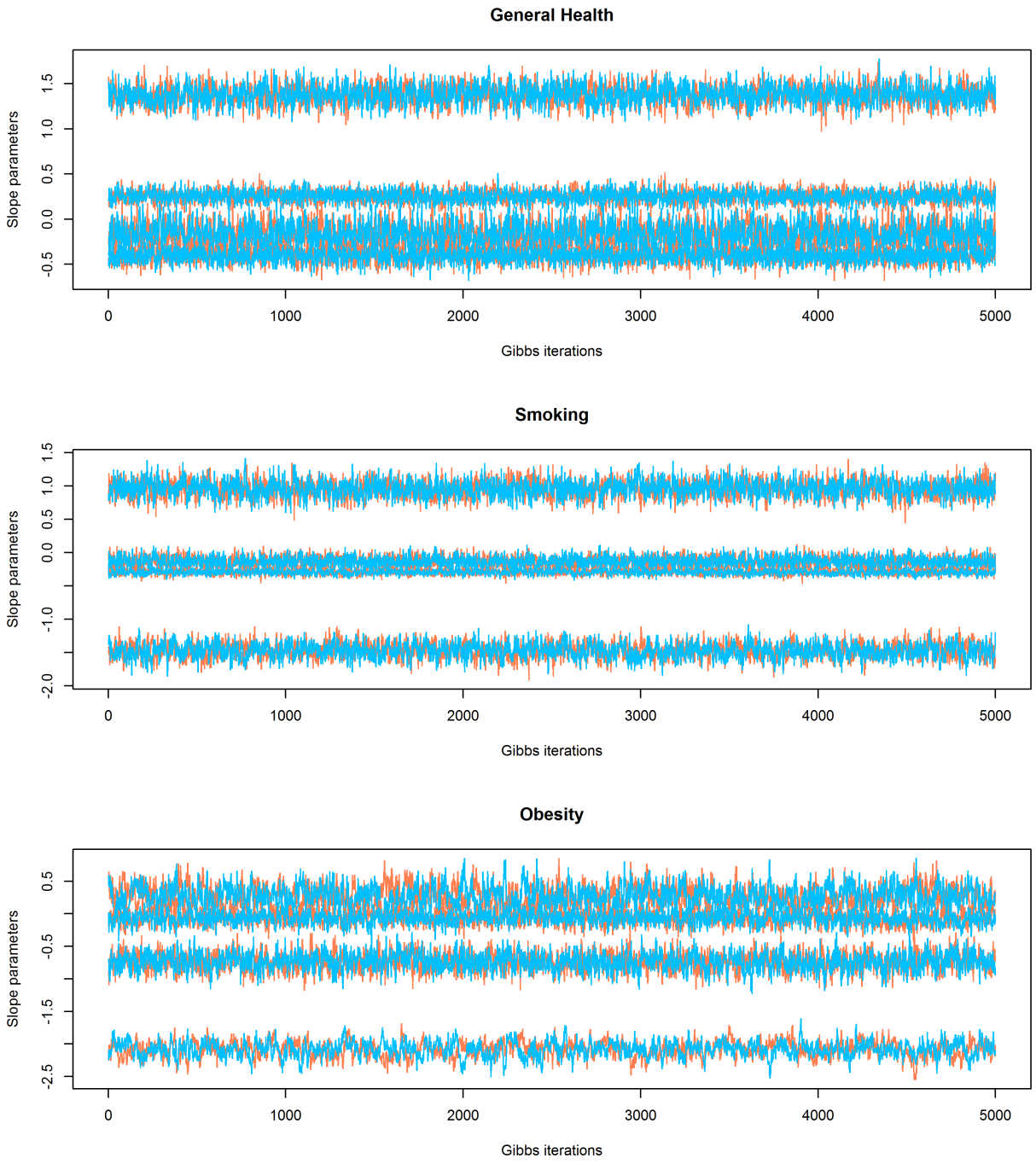

FIGURE C.1: Gibbs Sampler Draws for the Survey Variable Model Slope Parameters.

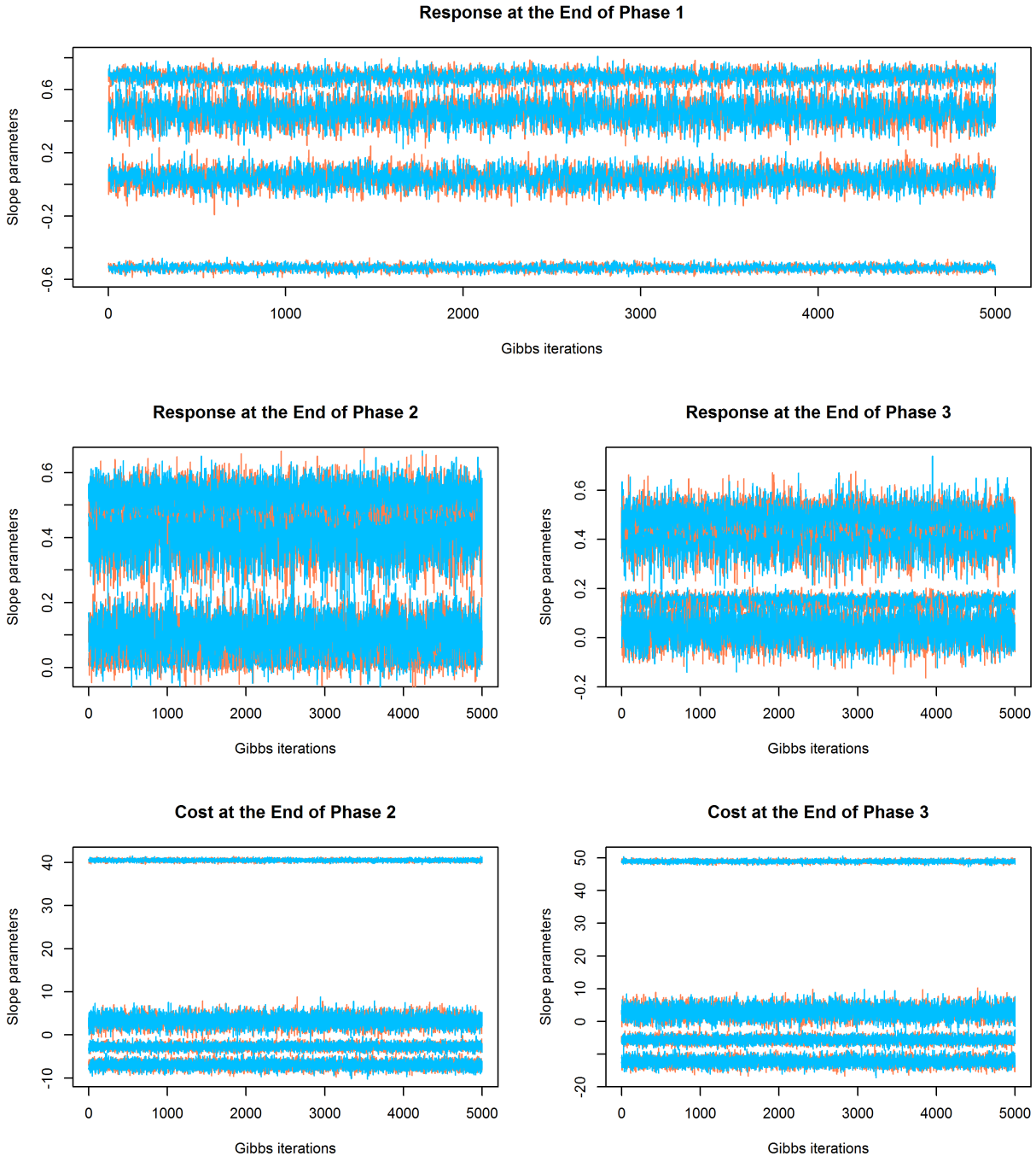

FIGURE C.2: Gibbs Sampler Draws for the Response Propensity and Cost Model Slope Parameters.

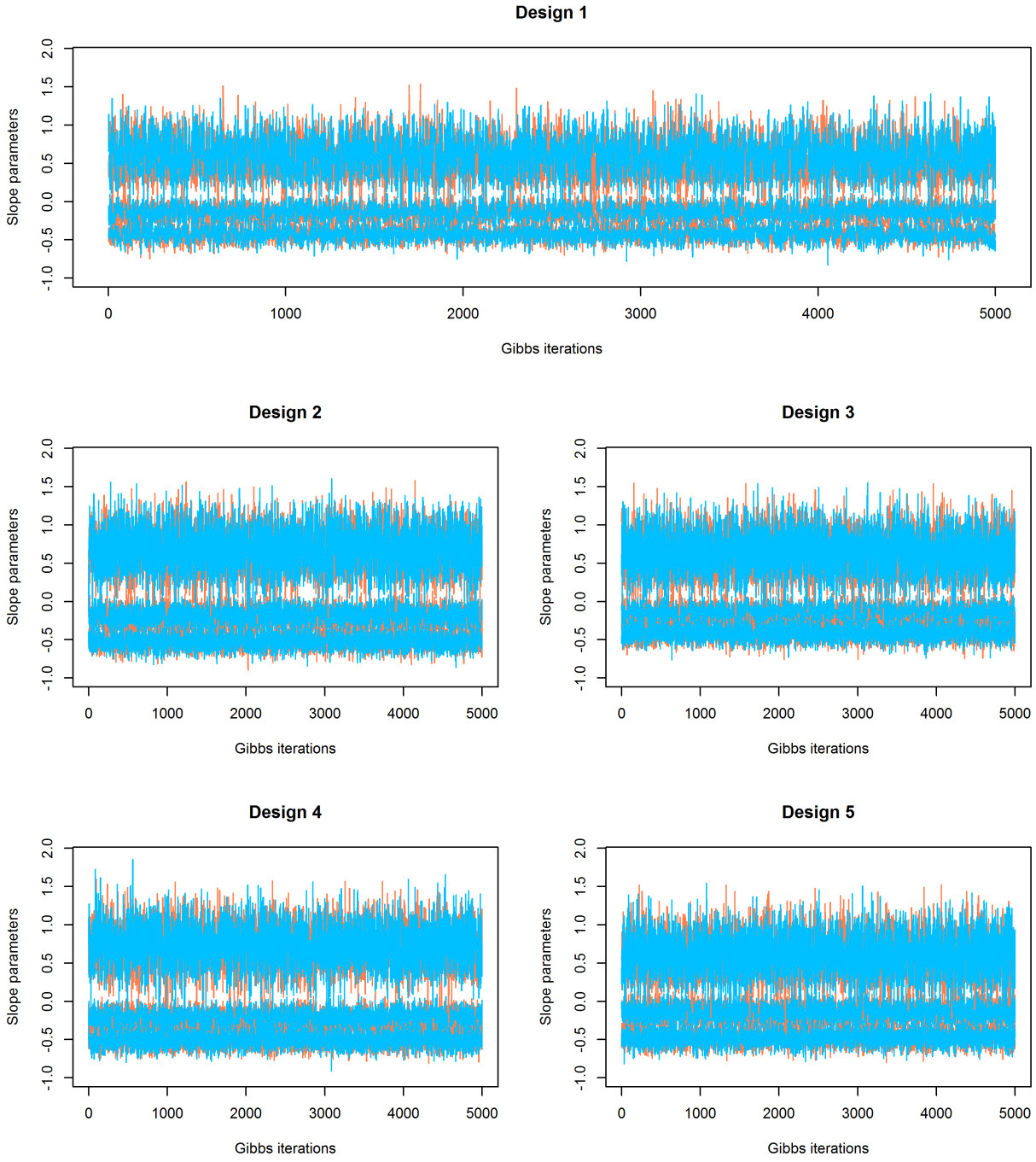

FIGURE C.3: Gibbs Sampler Draws for the Stratification-assessed Response Propensity Model Slope Parameters.

References

Bethlehem, J., Cobben, F., and Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*, Wiley Handbooks in Survey Methodology, Hoboken, NJ: John Wiley & Sons.

Bethlehem, J. G. (1988), "Reduction of nonresponse bias through regression estimation," *Journal of Official Statistics*, 4(3), 251.

Burger, J., Perryck, K., and Schouten, B. (2017), "Robustness of Adaptive Survey Designs to Inaccuracy of Design Parameters," *Journal of Official Statistics*, 33(3), 687–708.

CBS (n.d.), Health Survey As Of 2014. [online] Available at: <https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/health-survey-as-of-2014>[Accessed 3 January 2021].

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Boca Raton, FL: CRC Press.

Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A. (2019), "R-squared for Bayesian Regression Models," *The American Statistician*, 73(3), 307–309.

Groves, R. M., and 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.

Luiten, A., and 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.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011), "MCMCpack: Markov Chain Monte Carlo in R," *Journal of Statistical Software*, 42(9), 22.

Nishimura, R., Wagner, J., and Elliott, M. (2016), "Alternative Indicators for the Risk of Non-response Bias: A Simulation Study," *International Statistical Review*, 84(1), 43–62.

Rosen, J. A., Murphy, J., Peytchev, A., Holder, T., Dever, J., Herget, D., and Pratt, D. (2014), "Prioritizing Low Propensity Sample Members in a Survey: Implications for Nonresponse Bias," *Survey Practice*, 7(1), 1–10.

Rosenblum, M., Miller, P., Reist, B., Stuart, E. A., Thieme, M., and 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.

Särndal, C.-E. (2011), "The 2010 Morris Hansen Lecture. Dealing with survey nonresponse in data collection, in estimation," *Journal of Official Statistics*, 27(1), 1.

Schouten, B., Calinescu, M., and Luiten, A. (2013), "Optimizing quality of response through adaptive survey designs," *Survey Methodology*, 39(1), 29–58.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., and Wagner, J. (2018), "A Bayesian Analysis of Design Parameters in Survey Data Collection," *Journal of Survey Statistics and Methodology*, 6(4), 431–464.

Schouten, B., Peytchev, A., and Wagner, J. (2017), *Adaptive Survey Design*, Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, Boca Raton, FL: CRC Press.

Schouten, B., and Shlomo, N. (2017), "Selecting Adaptive Survey Design Strata with Partial R-indicators," *International Statistical Review*, 85(1), 143–163.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and 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.

Therneau, T., and Atkinson, B. (2019), *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

van Berkel, K., van der Doef, S., and Schouten, B. (2020), "Implementing Adaptive Survey Design with an Application to the Dutch Health Survey," *Journal of Official Statistics*, 36(3), 609–629.

Wagner, J. (2008), Adaptive Survey Design to Reduce Nonresponse Bias., PhD thesis, University of Michigan, Ann Arbor.

Wagner, J. (2014), Limiting the Risk of Nonresponse Bias by using Regression Diagnostics as a Guide to Data Collection,, in *Joint Statistical Meetings*.

Wagner, J. (2019), "Estimation of Survey Cost Parameters Using Paradata," *Survey Practice*, 12(1).

Wagner, J., West, B. T., Elliott, M. R., and 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.