Optimal Stratification in Bayesian Adaptive Survey Designs

Yongchao Ma

Department of Methodology and Statistics

Utrecht University

Email: y.ma1@uu.nl

January 21, 2021

1. INTRODUCTION

Adaptive survey designs are based on the rationale that different data collection strategies may be effective for different members of the population. 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. 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. This paper demonstrates a comprehensive stratification method that combines the interpretation of response behavior and key survey variables, aiming at effectively and efficiently reducing nonresponse bias in survey estimates. A design parameter is estimated based on the stratification scheme generated by this method. In future research, more design parameters and the quality and cost indicators will be estimated to further the optimization of adaptive designs. The utility and generalizability of this stratification method will also be evaluated.

This paper has the following sections: In section 2, I introduce the methodological elements of adaptive survey designs, including data collection strategies to be adapted, the stratification of the target population, and one of the requisites for optimal allocation of strategies to strata. In section 3, I show an application and preliminary results of this methodology to the Dutch Health Survey, and I close with a discussion on future research in section 4.

2. METHODOLOGY

This section describes data collection strategies in adaptive designs, followed by a review of previous research and a proposal of a new method for stratification with the purpose of reducing nonresponse bias. To facilitate further optimization of the survey quality, a design parameter is differentiated by strata and strategies within the Bayesian framework.

2.1 Adaptive strategies

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.

2.2 Stratification

In nonadaptive survey designs, a single strategy 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.

With the objective of minimizing systematic bias in survey estimates, previous research have described several methods to form subgroups.

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 groups, and the elderly can be identified as a target group 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 target survey variable and is not correlated with age, simply increasing the response rate of the elderly group 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 target survey variable, smoking prevalence, is modeled by age, a few influential observations in the youth group 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 based on 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 based on survey variable variation can effectively detect strata from which nonresponse errors originate, but it is inefficient when the likelihood of recruiting them in a survey and the corresponding costs remain unclear.

To stratify the target population effectively and efficiently given practical objectives and constraints on data quality and costs, this paper proposes to determine the target strata that differ most in response behavior based on predicted 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 subgroups 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.

For a subject i in the sample, let $x_{0,i} = (x_{0,1,i}, \ldots, x_{0,M,i})'$ be the vector of M auxiliary variables available at the start of data collection. Then the J survey variables $\hat{y}_{0,i} = (\hat{y}_{0,1,i}, \ldots, \hat{y}_{0,J,i})'$ can be predicted as follows:

$$\hat{y}_{0,j,i} = g^{-1}\left(\beta_{0,j} + \sum_{m=1}^{M} \beta_{m,j} x_{0,m,i}\right), \tag{1}$$

for $m \in \{1, \ldots, M\}$ and $n \in \{1, \ldots, J\}$ where $\beta_{0,j}$ is the regression intercept for the j-th survey variable, and $\beta_{m,j}$ is the corresponding regression coefficient of the m-th auxiliary variable; the regression parameters are derived from the models of survey variables in a previous round of a repeated survey or another similar survey. $g^{-1}(.)$ is the inverse of the link function that depends on the probability distribution of the survey variable. A survey variable that is not continuous, should be dichotomized first.

Based on the predicted survey variables, stratification can be performed with a classification tree algorithm. The target strata are formed by means of response propensities under each strategy (van Berkel, van der Doef and Schouten 2020). For example, under a strategy of using web questionnaires, the algorithm identifies two strata that may have approximately the same response propensities, but different response propensities if they are approached by face-to-face visits. The minimum size of the stratum is set to ensure reliable estimates of response propensities.

2.3 Design parameter

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, for example, response propensities and costs. In this paper, I concentrate on response propensities, $\rho_{T,g}(s_{1,T})$, per stratum g and strategy $s_{1,T}$.

Given that the unit did not respond in earlier phases and is eligible for follow-up, let $\lambda_{t,i}(s_{1,T})$

be the individual response propensities in phase, $t = 1, 2, \ldots, T$, under strategy $s_{1,T}$. Let the target population be divided into G strata, labeled $g = 1, 2, \ldots, G$, a sample of size n is drawn from the population. It is assumed that all units in a stratum g of size n_g have the same response propensity in each phase under strategy $s_{1,T}$

$$\lambda_{t,g}(s_{1,T}) = \frac{1}{n_g} \sum_{i=1}^{n} \delta_{g,i} \lambda_{t,i}(s_{1,T}), \tag{2}$$

where $\delta_{g,i}$ indicates if unit i is in stratum g.

The response propensity from phase 1 through t of a stratum g under strategy $s_{1,t}$, is denoted by $\rho_{t,g}(s_{1,t})$. When all nonresponses are followed up in subsequent phases, the response propensity through all T phases of data collection equals

$$\rho_{T,g}(s_{1,T}) = \lambda_{1,g}(s_{1,1}) + \sum_{t=2}^{T} \left(\left(\prod_{l=1}^{t-1} (1 - \lambda_{l,g}(s_{1,l})) \right) \lambda_{t,g}(s_{1,t}) \right). \tag{3}$$

Instead of modeling response propensities on the stratum level, I construct a Bayes model on the individual sample unit level. It offers more flexibility, and stratum propensities may be derived from the individual ones.

Schouten et al. (2018) elaborated on the derivation of the posterior distributions of the individual response propensities $\rho_{T,i}(s_{1,T})$ per strategy given the observed data. Since these propensities are complex functions, Markov Chain Monte Carlo (MCMC) methods are used to generate draws from the posterior distributions. This paper applies the same settings for most parts of the model and prior specification of individual response propensities in each phase. To be consistent with the stratification method, this paper estimates the response propensities using predicted survey variables rather than auxiliary data. For simplicity, this paper does not decompose the response propensities into contact and participation propensities. To keep the scope manageable, paradata obtained during the data collection and dependence of response in a certain phase on past actions are left for future research.

3. A CASE STUDY

3.1 Dutch Health Survey

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 selects three key survey variables:

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.*

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

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 long (1: are you 2: is your child)? It is the length in centimeters, without shoes*; *How many kilos does it weigh (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 *obese.* For children, the criteria depend on different ages.

3.2 Survey design

This paper focuses on the mix of survey modes. It is assumed that a sequential mixed-mode design is used 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 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 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 $s_{1,3} = \{(s_1, s_2, s_3)\}$.

3.3 Stratification based on predicted survey variables

The classification tree algorithm is implemented in R with the `rpart` package. First, three key survey variables are predicted by the categorical auxiliary variables, including age, sex, income, marital status, level of education, migration background, receiving rent benefit, type of household, urbanization level of the area of residence. The key survey variables are modeled by stepwise probit regression. Table 1 shows the results of three models.

[Table 1 about here.]

Not all auxiliary variables are significant predictors of key survey variables. The pseudo R^2 shows that the prediction on general health performs well, whereas the prediction on obesity is relatively weak.

Based on these models, probabilities are generated for general health, smoking, and obesity. The homogeneous subgroups are created based on these predicted probabilities. The algorithm determines where to split between 0 and 1 probabilities. To ensure sufficient units in each subgroup, a minimum of 500 units per subgroup is set.

[Table 2 about here.]

Table 2 shows the resulting strata. The smoking probability splits at 0.24, leading to group 1 composed of units whose smoking probabilities higher than or equals 0.24. The other units whose smoking probability lower than 0.24 are further split into three groups based on obesity and health probabilities.

Within each group, units are homogeneous in response behavior to a data collection strategy. The response behavior to different strategies varies across groups. The allocation of strategies to these groups is subjected to practical objectives and constraints on data quality and costs.

3.4 Response propensities

The response propensities per stratum and strategy should be estimated to facilitate further analysis of quality indicators and optimization of the adaptive designs. In the Bayesian analysis, I specify non-informative priors for the regression parameters of the models for response propensities in each phase. Table 3 shows the estimated response propensities per group for three strategies, and the corresponding 95% credible intervals. For example, the average probability that a person in group 1 responds to a web survey is 29.6%, and 95% of 5000 draws fall between 0.284 and 0.309.

[Table 3 about here.]

4. DISCUSSION

This paper proposes a stratification method to divide the target population into subgroups based on survey variables predicted by auxiliary data. A subsequent Bayesian analysis of response propensities is a requisite for further analysis of quality indicators and optimization of adaptive designs. To implement a complete process of selecting optimal designs, several elements should be added to this paper. The limited budget to conduct a survey requires the analysis of cost parameters in each step of data collection. By setting constraints on the total cost, the optimal designs are subjected to the required budget. Besides, to be able to incorporate dynamic adaptive survey designs, paradata obtained during data collection may also be included in the response propensity models. While building Bayesian models, it is crucial to elicit prior knowledge from an appropriate amount of historic survey data.

After the optimization of adaptive designs, future research can further demonstrate the strengths and limitations of this stratification method. This method relies on the predictive power of auxiliary variables to the target survey variables, which varies in different surveys. In the case study, this method can perform well as the predictions on the Health Survey variables are satisfactory. The advantages can be substantiated by applying other methods in the Health Survey and comparing their performance with this method. However, in an extreme case, there could be no available auxiliary variables that are predictors of the survey variables. The utility of this method in reducing nonresponse bias might not outstrip other methods. More surveys should be involved to examine the applicability and corresponding conditions of this new method.

REFERENCES

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].

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.

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., 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*, Boca Raton, Florida: CRC Press.

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

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

List of Tables

Table 1: Probit models of survey variables predicted by auxiliary data

	General Health	Smoking	Obesity
Age (< 12)			
12-17	-0.289^{***}	3.137	-0.245
18-24	-0.392^{***}	4.083	0.013
25-29	-0.720^{***}	4.277	0.723^{***}
30-34	-0.852^{***}	4.437	0.905^{***}
35-39	-0.870^{***}	4.307	0.947^{***}
40-44	-1.036^{***}	4.443	0.933^{***}
45-49	-1.290^{***}	4.264	1.107^{***}
50-54	-1.240^{***}	4.218	1.040^{***}
55-59	-1.403^{***}	4.109	1.166^{***}
60-64	-1.583^{***}	4.102	1.071^{***}
65-69	-1.359^{***}	3.828	1.121^{***}
≥ 70	-1.521^{***}	3.569	0.939^{***}
Female (Male)		-0.271^{***}	
Income (Missing)			
0-20% perc	-0.049	0.232^{***}	0.017
20-40% perc	-0.162^{*}	0.242^{***}	0.099
40-60% perc	-0.032	0.170^{*}	-0.012
60-80% perc	0.256^{***}	0.088	-0.099
80-100% perc	0.445^{***}	-0.109	-0.194^{**}
Migration (Native)			
1^{st} generation non-western	-0.415^{***}		
1^{st} generation western	-0.173^{*}		
2^{nd} generation non-western	-0.211^{**}		
2^{nd} generation western	-0.197^{***}		
Marital status (Single)			
Married or partnership		-0.246^{***}	-0.056
Widowed		-0.149	0.096
Divorced		0.128	0.092
Urbanization level (Moderate)			
Non-urban		-0.056	
Highly urban		0.154^{***}	
Little urban		-0.010	
Very highly urban		0.034	
Household type (Single)			
Couple	0.298^{***}	-0.094	
Couple with offspring	0.348^{***}	-0.192^{***}	
Single with offspring	0.162^{*}	0.057	
Other	0.052	0.187	
Education level (Primary)			
VMBO	0.034	0.112	0.027
HAVO-VWO-MBO	0.148	-0.039	-0.116

(to be continued)

Table 1: Probit models of survey variables predicted by auxiliary data (continued)

	General Health	Smoking	Obesity
Bachelor HBO-WO	0.330^{***}	-0.439^{***}	-0.449^{***}
Master HBO-WO	0.328^{**}	-0.644^{***}	-0.604^{***}
Unknown	0.138^{*}	0.051	-0.121
Receive rent benefit (No)	-0.416^{***}	0.117	
Constant	1.510^{***}	-4.651	-1.831^{***}
Observations	7,472	6,341	7,118
Nagelkerke pseudo R^2	0.213	0.135	0.099
CoxSnell pseudo R^2	0.135	0.086	0.051

Note: Reference groups in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 2: Stratification of the population based on predicted survey variables

Stratum	Smoking probability	Obesity probability	Health probability
1 (3872)	≥ 0.24	-	-
2 (2442)	< 0.24	< 0.037	-
3 (544)	< 0.24	≥ 0.037	< 0.54
4 (4707)	< 0.24	≥ 0.037	≥ 0.54

Note: Stratum size in parentheses.

Table 3: Estimated response propensities per stratum and strategy

	1		2		3		4	
	n_g	$\rho_{T,g}$	n_g	$\rho_{T,g}$	n_g	$\rho_{T,g}$	n_g	$\rho_{T,g}$
Web	3872	29.6%	2442	43.7%	544	38.8%	4707	42.2%
		(0.284, 0.309)		(0.421, 0.453)		(0.365, 0.410)		(0.412, 0.433)
Web → F2F-short	2774	49.7%	1491	67.1%	407	61.2%	2481	63.8%
		(0.483, 0.512)		(0.657, 0.685)		(0.589, 0.633)		(0.628, 0.648)
Web → F2F-extended	687	62.1%	277	82.2%	33	70.4%	441	77.3%
		(0.602, 0.641)		(0.804, 0.838)		(0.672, 0.735)		(0.761, 0.786)

Note: 95% Credible intervals in parentheses.