A Bayesian analysis of survey design parameters

*Lisette Bruin, Barry Schouten, Nino Mushkudiani,*

# 1. Introduction

Over the last two decades, there has been a strongly increasing interest in survey data collection monitoring, analysis, and intervention or adaptation. The main causes for this are the diversification of data collection that followed the emergence of online communication, the lack of predictability of survey response rates despite years of research into survey design, the gradual increase in costs per respondent when response rates are kept at traditional levels, and the availability of a wide range of data collection process data (termed paradata) see Kreuter (2013).

The lack of predictability of data caused different streams of data collection. One of these consists of adaptive or responsive survey designs that adapt or tailor strategies and effort to known and relevant characteristics of sampled units from the target population, see Groves and Heeringa (2006), Wagner (2008) and Schouten, Calinescu and Luiten (2013). In order to adapt, accurate estimates of survey design parameters are not just needed at the overall population level, but also at the deeper level of population subgroups.

A natural approach to deduce inaccuracy of survey design parameters and to make analyses and design optimization more robust is to incorporate historic survey data and expert judgment through a Bayesian analysis. In this paper we set up a general model for survey design parameters. To these survey design parameters are then assigned prior distributions, which are updated and transformed to posterior distributions during data collection. These posterior distributions may be used as prior distributions in new waves of the same or of similar surveys. We propose to use Gibbs samplers to obtain draws from the posterior distributions of response propensities and cost functions. As an important by-product, many quality and cost indicators can be computed from these basic survey design parameters.

# 2. Adaptive strategies and design parameters

To set up a general model that would describe all possible data collection designs in relation to available covariates from frame data, administrative data and paradata, is too complicated a goal. Here we intend to present models that have many features of designs but in their simplest forms. The included features are more than one data collection phase, baseline covariates as well as paradata, cost functions, and dependence on actions/decisions in earlier phases.

First we introduce some notations. Let the survey design consist of a maximum of phases that are labelled . We define as the collection of all possible actions in phase and let represent the action in phase . For different phases, the collections of actions may be different. The action sets may contain , which, if selected, implies that no attempt is made to obtain a response. We define the collection of survey strategies

and let denote one possible strategy, i.e. sequence of actions. For a strategy , we denote the actions in phase til by the vector .

For a subject , we let be the vector of auxiliary variables that is linked from frame data, administrative data or paradata, consists of the following entries

 ,

where contains the auxiliary variables available at the start of data collection, and are the auxiliary variables that are observed for the fielded sample units in phase . In the optimization of the adaptive survey designs (ASD), actions in phase can only be chosen based on to .

The design of each survey has a range of features, e.g. advance letter, contact protocol, screener interview, number of phases, reminder protocol, use of incentive, mode of administration (web, telephone, face-to-face, mail), interviewer, refusal conversion procedure and type of questionnaire. The total of choices made for the design features (e.g. incentive, phases, first web mail then telephone interview) will define the data collection strategy or simply strategy. In non-adaptive surveys, these features are implemented uniformly over the whole sample, i.e. there is one strategy. In adaptive surveys, part of the design features may be implemented differently for different sample units, i.e. there is a set of strategies, see Groves and Heeringa (2006), Wagner (2008), Coffey, Reist and White (2013).

On the other hand ASD either maximize a quality objective subject to cost constraints and other quality constraints or minimize a cost objective subject to quality constraints. The quality and cost constraints depend on the setting in which the survey is conducted. Three sets of survey design parameters suffice to compute most of the quality and cost constraints:

1. Response propensities per unit per strategy;
2. Expected costs per sample unit per strategy;
3. Adjusted mode effects per unit per strategy;

In this paper we restrict attention to nonresponse error. The last set of parameters, the adjusted mode effects , are not considered here. There are two options in defining and modeling survey design parameters. Design parameters can be detailed to subgroups or to individual cases. Below we focus on individual design parameters.

# 3. Modeling survey design parameters

We introduce basic models for response propensities and costs. Therefore, we break down these parameters into their basic components, like the contact and participation propensities. For these basic components we will, first, make some general assumptions. We assume that making contact, obtaining participation and the costs associated with an individual sample unit are independent of contact, participation and the costs of any other individual sample unit.

Some side remarks are in place. Below we introduce prior distributions for parameters that are shared by multiple design parameters. As a result, the propensities and costs of sample units may become dependent because they share the same underlying parameters. The assumptions then read as independence given the values of these parameters and likelihood functions can still be factorized as the product of individual likelihoods conditional on the parameters. Second, the three assumptions essentially ignore any impact of scale on data collection; it is assumed that there is a stable workload. Importantly, we allow for associations between contact propensities over phases, between participation propensities over phases, and between cost functions over phases.

We define our approach only for contact propensity, a part of response propensity. Models for the response propensity and cost functions can be defines similarly.

Let be the propensity of a contact in phase under strategy given that the unit did not respond in earlier phases and is eligible for follow-up. We assume that design features in subsequent phases have no impact on making contact. The outcome(s) of the previous phase(s) can be included in the auxiliary vector, when contact propensities are considered to be dependent on whether there was a noncontact or a refusal. is the propensity of a participation in phase of subject under strategy given contact (and given that the unit did not respond in earlier phases and is eligible for follow-up). Then the response propensity in phase of a subject under strategy , , is .

When in subsequent phases all nonresponse receives a follow-up, then

We model the propensities using a probit model, i.e. using a binomial link function. Each sample unit has a certain contactability represented as a latent variable and contact is obtained when this latent variable is larger than zero and , for some so that

For , let be the regression coefficient in phase corresponding to the -th entry in the auxiliary vector given that is applied to a unit. Obviously, , when . The model could be written as

where is an error term for the uncertainty of contact of the subject.

To be able to include dynamic adaptive survey designs, we need to include paradata. However to keep the model simple, we assume that there is just one phase, say , in which paradata is collected. Up to phase only the auxiliary variables in can be used to model the propensities. After phase , the auxiliary variables obtained in phase can also be included in the model. Second, we consider the dependence on past actions. It is unrealistic to assume there is no such dependence in most settings. Past actions could be included by introducing a fixed or random effect per possible history. We add the history as a random effect here. Third, since we suggest to add a dependence on the history of actions as a random effect, the regression coefficients become necessarily dependent on the phase and not on the past. The model becomes

 (1)

where is a random effect.

The analysis become Bayesian by assigning prior distributions to the regression coefficients and random effects in (1). Our aim is the derivation of the posterior distributions of the individual response propensities and the individual cost parameters per strategy given observed data. These overall parameters are, in general, complex functions of the underlying survey design parameters per phase. We derived expressions for the posterior distributions of the regression coefficients and random effects when it was possible, otherwise derived these numerical approximations and applied Markov Chain Monte Carlo methods to generate draws from the posterior distributions.

Data collection may apply randomization in order to learn about multiple strategies simultaneously. Here, we assume that the observed data may contain randomization over strategies but that randomization is only at the outset. Hence, strategy allocation probabilities may depend on auxiliary information known at the start of data collection, but not on paradata coming in during data collection. So in addition to the outcomes, costs and auxiliary vectors, we observe

* The series of actions, or simply strategy, that were applied per sample unit . Since we assume that future actions do not impact the outcomes or costs of earlier phases, we do not need to include the planned actions for phases not actually conducted. The latter occurs when there is a response or a nonresponse not eligible for follow-up;

In the following, we use and for the vector of response propensities and cost parameters over all sample units for a particular strategy. In the same fashion, we use , , , and to denote the vectors of outcomes, realized costs components and auxiliary variables over sample units. Note that may in fact be a matrix, when the auxiliary variables are a vector by themselves. With we denote the vector of used strategies for all sample units. To shorten expressions, we use for the vectors of regression slope parameters, random effects and regression dispersion parameters over phases and actions, but elaborate when needed. For the sake of convenience, we use to express joint and marginal density functions; we omit the reference to the random variables to which they apply and ignore differences between discrete and continuous probability distributions. Finally, in the density functions, we omit the dependence on the hyperparameters. A straightforward solution is to perform a Gibbs sampler to the joint density of the regression parameters

 . (3)

A Gibbs sampler for (3) requires repeated draws from the conditional densities of each regression parameter given the observed data and the other regression parameters, the so-called full conditionals. Literature provides a range of options to sample from these conditional distributions, see Albert and Chib (1993) and Gelman et al (2003).

## 3.1 Functions of survey design parameters

In the monitoring and optimization of data collection, the focus is on functions of the design parameters that correspond to overall quality or cost objectives. We consider three such functions here for the sake of brevity, the response rate, the total costs and the coefficient of variation of the response propensities; the analysis of other functions can often be done in an analogous way.

Let represent the design or inclusion weight for sample unit , . The response rate, *RR,* for strategy can be written as

 , (4)

the total costs, or required budget, *B*, associated with are

 , (5)

and the coefficient of variation, *CV,* is

 . (6)

For the *CV*, we explicitly denote the dependence on the covariate vector ; for any other choice of auxiliary variables it will, generally, attain a different value. The response rate and total costs do not depend on the choice of .

Obviously, the prior and posterior distributions for these three functions are determined by the prior and posterior distributions of the components of the response propensities and cost functions. They have even more complex forms than the individual response propensities and cost parameters. However, they can again be approximated as a by-product of the Gibbs sampler. For every draw of the individual response propensities and cost parameters, we compute (4) to (6).

# 4. A simulation study

In the simulation study, we investigate the impact of prior distribution specification and of survey sample size on the shape of posterior distributions. Furthermore, we explore the convergence properties of the Gibbs sampler. We link the simulation study to the Dutch Health Survey and employ historic parameters in order to mimic a real setting as much as possible.

## 4.1 Design of the simulation study

We base our simulation study on the 2015 Dutch Health Survey (HS). The HS has a sequential mixed-mode survey design with Web followed by face-to-face interviewing, i.e. non-respondents to a Web survey invitation are re-allocated to interviewers. We consider three data collection phases: Web, short face-to-face, and extended face-to-face. The extended face-to-face corresponds to an additional round of face-to-face visits for those sample units that have not been contacted or that are soft refusals after three face-to-face visits. The first three visits are termed short face-to-face. We label the three phases as Web, F2F-short, F2F-extended and monitor marginal and cumulative response rates, coefficients of variation and costs after each of the phases.

For simplicity, we employ only three auxiliary variables: Two variables, gender and age, are linked from administrative data, and one variable, web break-off, is added from phase 1 paradata. Gender and age are crossed to form six strata, {0-29 years, 30-59 years, 60 years and older}×{female, male}. Web break-off is a 0-1 indicator for a broken-off Web response; it is not crossed with the gender-age variable but added as a main effect. We refer to the variables as GenderAge and BreakOff

From the HS, survey design parameters are estimated for the model GenderAge in phase 1 and for the model GenderAge + BreakOff in phases 2 and 3. For phase 1, contact and participation cannot be distinguished and participation propensities are set to one. For phases 2 and 3, we distinguish contact and participation propensities. Table 1 gives estimated response rates, coefficients of variation and total costs per phase and cumulatively. Contact and participation are modeled through probit regression and costs through linear regression.

*Table 1: Expected response rates (RR), coefficients of variation (CV) and total costs (B) per phase and cumulatively based on the simulation model.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Web* | *F2F-short* | *F2F-extended* |
| *Phase 2* | *Cumulative* | *Phase 3* | *Cumulative* |
| RR | 32.2% | 39.4% | 58.9% | 13.9% | 64.6% |
| CV | 0.108 | 0.113 | 0.043 | 0.215 | 0.048 |
| B | 3.0 | 20.1 | 17.3 | 16.6 | 23.7 |

## 4.2 Prior distribution specification

One of the goals of our simulation study is to analyze the impact of misspecified priors and non-informative priors on the shape of posterior distribution. Therefore, we want to compare these posterior distributions to the posterior distributions based on ‘true’ or informed priors. For each parameter we will, therefore, compare four prior specifications:

1. True prior: Expectations of regression coefficients are equal to their simulation model expected values, and variances of regression coefficients are based on one ‘historic’ wave of a size similar to the current wave;
2. Misspecified prior: Like the true prior, but with shifted expectations for one or more regression parameters;
3. Naïve prior: The same variances as the true prior, but with equal expectations for regression parameters over population strata;
4. Non-informative prior:Like the naive prior, but with very large variances for regression parameters.

In order to come up with prior distribution hyperparameters for non-saturated probit models, i.e. with some or all interaction terms missing, we proceeded as follows: Following the simulation model, we generated 200 independent data sets and fitted the probit models to these. The mean and sample covariance matrix were used as hyperparameters. Although this is beyond the scope of the current paper, such an approach may also work best when hyperparameters are elicited from expert knowledge.

As a non-informative prior for dispersion parameters in cost models, we used a distribution, despite the warning in Gelman (2006) that for small values of the variance the prior may not behave like a non-informative prior.

## 4.3 Simulation results

For all scenarios, we find that the burn-in period is very short and that Gibbs sampler draws seem to converge very quickly to draws from the stationary distribution, i.e. the joint distribution of interest. Figures 1 and 2 show Gibbs sampler runs for regression parameters in the phase 2 contact and participation models, respectively. Figure 3 shows the Gibbs sampler runs for the coefficient of variation after phases 1 and 3. Runs for other parameters show a similar picture.

*Figure 1: Gibbs sampler runs for the regression slope parameters in the phase 2 contact model for young males ( and for web break-off (.*



*Figure 2: Gibbs sampler runs for the regression slope parameters in the phase 2 participation model for young males ( and for web break-off (.*



*Figure 3: Gibbs sampler runs for the coefficient of variation after phase 1 and after all phases.*



Table 2 gives the expectations of the posterior distribution for response propensities per phase for samples of 5000 and 10000 units based on the non-informative prior and as approximated by the Gibbs sampler.

*Table 2: Expectations of posterior distributions for phase response propensities per stratum approximated through the Gibbs sampler for n=5000 and n=10000. Standard deviations of posterior distributions are given within brackets.*

|  |  |  |
| --- | --- | --- |
| *Stratum* | *True* | *Gibbs sampler*  |
| *Web* | *F2F-short* | *F2F-extended* |
| Web | F2FS | F2FE | *n=5000* | *n=10000* | *n=5000* | *n=10000* | *n=5000* | *n=10000* |
| 1 | 0.280 | 0.448 | 0.180 | 0.294 (0.012) | 0.294 (0.011) | 0.449 (0.018) | 0.453 (0.014) | 0.154 (0.016) | 0.165 (0.012) |
| 2 | 0.310 | 0.326 | 0.099 | 0.302 (0.015) | 0.304 (0.012) | 0.345 (0.016) | 0.337 (0.011) | 0.110 (0.012) | 0.117 (0.010) |
| 3 | 0.400 | 0.288 | 0.072 | 0.456 (0.022) | 0.422 (0.014) | 0.315 (0.021) | 0.302 (0.016) | 0.128 (0.018) | 0.123 (0.013) |
| 4 | 0.300 | 0.406 | 0.150 | 0.305 (0.016) | 0.299 (0.013) | 0.400 (0.017) | 0.393 (0.014) | 0.160 (0.0160 | 0.148 (0.010) |
| 5 | 0.330 | 0.363 | 0.136 | 0.319 (0.013) | 0.336 (0.011) | 0.373 (0.016) | 0.378 (0.012) | 0.133 (0.013) | 0.134 (0.010) |
| 6 | 0.350 | 0.304 | 0.090 | 0.351 (0.015) | 0.362 (0.014) | 0.344 (0.018) | 0.342 (0.015) | 0.147 (0.015) | 0.127 (0.011) |
| 7 |  | 0.651 | 0.325 |  |  | 0.639 (0.025) | 0.658 (0.027) | 0.190 (0.027) | 0.256 (0.028) |
| 8 |  | 0.696 | 0.409 |  |  | 0.582 (0.033) | 0.601 (0.026) | 0.143 (0.022) | 0.203 (0.024) |
| 9 |  | 0.752 | 0.621 |  |  | 0.554 (0.034) | 0.584 (0.028) | 0.165 (0.025) | 0.221 (0.030) |
| 10 |  | 0.609 | 0.295 |  |  | 0.616 (0.029) | 0.621 (0.026) | 0.198 (0.026) | 0.241 (0.027) |
| 11 |  | 0.696 | 0.409 |  |  | 0.605 (0.032) | 0.636 (0.027) | 0.169 (0.023) | 0.226 (0.026) |
| 12 |  | 0.760 | 0.648 |  |  | 0.592 (0.029) | 0.628 (0.028) | 0.188 (0.026) | 0.227 (0.029) |

# 5. Discussion

We introduced a Bayesian model for survey design parameters related to response and costs. The model is general in that it describes multiple data collection phases, different types of auxiliary data, multiple nonresponse outcomes and dependence on previous actions. The model does, however, not fit all possible data collection designs, and any particular design may require adaptations of the model. Furthermore, we constructed an analysis strategy based on a Gibbs sampler in which all model parameters are repeatedly drawn. The Gibbs sampler provides estimates for the posterior distributions of the contact and participation propensities and the costs per sample unit. From the runs of the Gibbs sampler, also the posterior distributions for overarching quality indicators, like the response rate or coefficient of variation of the response propensities, and cost indicators can easily be derived as an important by-product.

In our simulation study we consider the convergence properties of the Gibbs sampler and the impact of specification and sample size. In a Bayesian analysis, such an evaluation is different in nature from a traditional analysis: Prior distributions for regression parameters reflect what we believe is the underlying mechanism to the generation of such parameters. If we specify a prior as having a (relatively) small variance, then automatically the variance of the posterior distribution becomes small as well; even when actual survey data seems to contradict the prior specification.

# References

Albert, J.H., Chib, S. (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, 669 – 680.

Coffey, S., Reist, B., White, M. (2013). Monitoring Methods for Adaptive Design in the National Survey of College Graduates,  In JSM Proceedings, Survey Methods Research Section,  Alexandria, VA:  American Statistical Association,  3085-3099.

Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, 1 (3), 515 – 534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman and Hall, second edition.

Groves, R.M., Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society: Series A*, 169, 439 – 457.

Kreuter, F. (2013), *Improving Surveys with Paradata. Analytic Uses of Process Informaton*, Edited monograph, John Wiley and Sons, Hoboken, New Jersey, USA.

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

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, PhD thesis, University of Michigan, USA.