sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 637–662, http://dx.doi.org/10.2478/JOS-2022-0028

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

Shiya Wu[1], *Barry Schouten*[2], *Ralph Meijers*[3], *and Mirjam Moerbeek*[1]

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

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

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

Key words: Nonresponse bias; Bayesian; response propensity; expert elicitation.

1. Introduction

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

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

In repeated surveys, the natural strategy is to estimate standard errors of parameters using recent historic survey data, but it is unclear how to deal with uncertainty in the

[1] Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. Emails: s.wu@uu.nl and M.Moerbeek@uu.nl
[2] Department of Process Development and Methodology, Statistics Netherlands, P.O. Box 24500, 2490 HA Den Haag, The Netherlands. Email: jg.schouten@cbs.nl
[3] Department of Traffic and Transport of Division Social Statistics, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: rj.meijers@cbs.nl
Acknowledgments: This work was supported by the China Scholarship Council (CSC). This work contains micro data from Statistics Netherlands, which has the copyright.

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

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

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

Expert prior elicitation depends heavily on the statistical skills of the experts. In biometrics and medical studies, experts are often viewed as relatively less trained in statistics (Gosling et al. 2007; Oakley and O'Hagan 2007). The elicitation then focuses strongly on transforming properties of prior distributions, such as medians, means, quantiles and variances, to questions that can be answered by experts. Oakley and O'Hagan (2007) introduce an additional step in prior elicitation in which a prior is set on the prior itself by means of Gaussian processes and updated by the summaries provided by the experts. In settings where experts have no training at all in statistics, prior elicitation may even resort to game-like approaches that faciliate experts to express their beliefs (O'Hagan et al. 2006; Veen et al. 2017). These approaches also tend to rate the experts themselves on their amount of expertise and assign and estimate weights to each expert.

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

To include such expert knowledge, power priors are an obvious option. Power priors were introduced by Ibrahim and Chen (2000) and further discussed in Ibrahim et al. (2015). Historic data sets D_k, labelled $k = 1, 2, \ldots, K$, from previous studies are assigned a scalar

quantity γ_k representing their similarity. In the derivation of the posterior, the scalar quantities are included as powers to the data likelihoods. Obviously, the prior elicitation then amounts to a selection of data sets and the choice of the associated powers. Rietbergen et al. (2011) discuss how to elicit γ_k. In their approach, the experts rank the historic studies based on their relevance and provide a prescribed fixed weight per study based on heterogeneity between study characteristics. In the survey context, study characteristics may reflect survey design features such as the target population, the survey topics, and the survey modes. Data collection staff have a good sense of the most influential features and already select historic surveys based on these features. However, there usually is not a structured approach and two experts may end up with different predictions.

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

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

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

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

The remainder of this article is organized as follows. In Section 2, we describe the information that we elicit from experts and formulate the power priors that include expert judgments. In Section 3, we motivate our strategy to validate performance of the power priors for the quality indicators against noninformative priors. In Section 4, we empirically evaluate the performance for the two case studies. We close with a brief discussion in Section 5. R code is available for the expert elicitation and posterior derivation steps at *GitHub*.

2. Methodology

In this section, the methodology to perform a Bayesian analysis is explained and prior elicitation is prepared. The Bayesian analysis is focused on response propensities in

population strata. Two overall and one partial quality indicators: the R-indicator, the coefficient of variation of response propensities, and the partial coefficient of variation, are the main targets of the analysis.

2.1. Notation

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

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

Data collection staff select $K \geq 1$ historic data sets with respect to a new survey. $\boldsymbol{D}_g^0 = \left\{D_{1,g}^0, D_{2,g}^0, \ldots, D_{k,g}^0, \ldots, D_{K,g}^0\right\}$ represents the sufficient statistics in the historic data sets for stratum g. The superscript '0' in $\boldsymbol{D}_g^0$ denotes that it refers to baseline information. The element $D_{k,g}^0$ consists of two statistics, the number of observed respondents $r_{k,g}^0$ and the number of sample units $n_{k,g}^0$ for stratum g in the kth historic survey. Hence, $D_{k,g}^0 = \left(n_{k,g}^0, r_{k,g}^0\right)$. We assume throughout that the stratum classification of sample units itself is not subject to error and is the same across historic surveys.

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

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

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

2.2. Prior and Posterior Distributions

Let us assume for now that similarity parameters γ_k are available. In Section 3, we explain how to construct them. This section shows how prior distributions are derived and how they are updated to posterior distributions.

We start by looking at the likelihood of the response propensities, given the data (sample size and responses). When a sample of size n_g is drawn from stratum g and r_g sample units respond, then r_g follows a Binomial distribution, $r_g \sim Bin(n_g, \rho_g)$. For historic data set k, one obtains the likelihood

$$L\left(\rho_g | D^0_{k,g}\right) \propto \rho_g^{r^0_{k,g}} \left(1 - \rho_g\right)^{n^0_{k,g} - r^0_{k,g}}, \tag{1}$$

and for the K combined data sets, the likelihood is

$$L\left(\rho_g | \boldsymbol{D}^0_g\right) = \prod\nolimits_{k=1}^{K} L\left(\rho_g | D^0_{k,g}\right) \propto \prod_{k=1}^{K} \rho_g^{rk,g0} \left(1 - \rho_g\right)^{n^0_{k,g} - r^0_{k,g}}. \tag{2}$$

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

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

With the availability of the historic survey data, Equation (2) can be used to formulate the informative $Beta(a_0, b_0)$ prior. It has the following shape parameters

$$a_0 = \left(\sum\nolimits_{k=1}^{K} r^0_{k,g}\right) + 1, \tag{3}$$

$$b_0 = \left(\sum\nolimits_{k=1}^{K} n^0_{k,g} - r^0_{k,g}\right) + 1. \tag{4}$$

To come to Equations (3) and (4), we assume that all historic surveys are perfectly similar to the new survey. This is not true in general and the impact of the data set likelihoods must be altered using the similarity parameter γ_k as the power. This conforms to the approach taken by Ibrahim and Chen (2000). The power prior of the new survey raises the Binomial likelihood of historic data k in Equation (1) to the powers represented by these similarity parameters,

$$\pi\left(\rho_g | D^0_{k,g}, \gamma_k\right) \propto \rho_g^{\gamma_k r^0_{k,g}} \left(1 - \rho_g\right)^{\gamma_k \left(n^0_{k,g} - r^0_{k,g}\right)}. \tag{5}$$

Here, we let the powers be equal for all strata and, therefore, do not add a subscript g. However, our method could be extended to stratum-dependent powers when historic surveys apply different strategies to different strata. The power γ_k reduces the strength of the kth historic data. When the power equals zero, then the power prior for the particular

survey is non-informative. When the power is one, then the historic surveys are seen as copies of the current survey.

For the K combined historic survey data sets one obtains

$$\pi\left(\rho_g|\boldsymbol{D}_g^0, \gamma_k\right) \propto \rho_g^{\Sigma_k \gamma_k r_{k,g}^0}\left(1-\rho_g\right)^{\Sigma_k \gamma_k\left(n_{k,g}^0 - r_{k,g}^0\right)}. \tag{6}$$

The Beta posterior distribution parameters change along with Equation (6) to

$$a_0 = \left(\sum\nolimits_{k=1}^{K} \gamma_k r_{k,g}^0\right) + 1, \tag{7}$$

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

$$\pi\left(\rho_g|D_g^1, \boldsymbol{D}_g^0, \gamma\right) \propto \rho_g^{\Sigma_k \gamma_k r_{k,g}^0 + r_g^1}\left(1-\rho_g\right)^{\Sigma_k \gamma_k\left(n_{k,g}^0 - r_{k,g}^0\right) + \left(n_g^1 - r_g^1\right)}. \tag{9}$$

Equation (9) is repeated on a rolling basis, that is, the posterior from the previous wave is used as the prior in the next wave to update the inference on ρ_g. In general, after t waves, the posterior becomes

$$\pi\left(\rho_g|D_g^1, \ldots, D_g^t, \boldsymbol{D}_g^0, \gamma\right) \propto \rho_g^{\sum\limits_k \gamma_k r_{k,g}^0 + \Sigma_{s=1}^t r_g^s}\left(1-\rho_g\right)^{\sum\limits_k \gamma_k\left(n_{k,g}^0 - r_{k,g}^0\right) + \Sigma_{s=1}^t\left(n_g^s - r_g^s\right)}. \tag{10}$$

From Equation (10), it can be deduced that in wave t, the posterior is a $Beta\left(a_t, b_t\right)$ distribution, with

$$a_t = \sum_k \gamma_k r_{k,g}^0 + \sum_{w=1}^{t} r_g^w + 1, \tag{11}$$

$$b_t = \sum_k \gamma_k\left(n_{k,g}^0 - r_{k,g}^0\right) + \sum_{w=1}^{t}\left(n_g^w - r_g^w\right) + 1. \tag{12}$$

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

2.3. *Nonresponse Quality Metrics*

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

In monitoring and adapting survey design the interest is in response propensity variation, where the major objective is to reduce nonresponse bias. For that reason,

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

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

$$R(\hat{\rho}) = 1 - 2\sqrt{\sum_{g=1}^{G} q_g \left(\hat{\rho}_g - \widehat{\bar{\rho}}\right)^2}. \tag{13}$$

where $\hat{\rho}_g$ is the response propensity in stratum g, and $\widehat{\bar{\rho}}$ is overall response rate explicit about the level of the population response propensity. Response is fully representative when Equation (13) takes the value 1 and completely non-representative at the value 0. Lower standard deviation of response propensities means more representative response.

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

$$CV(\hat{\rho}) = \frac{\sqrt{\sum_{g=1}^{G} q_g \left(\hat{\rho}_g - \widehat{\bar{\rho}}\right)^2}}{\widehat{\bar{\rho}}}. \tag{14}$$

The overall CV is the response propensity standard deviation divided by $\widehat{\bar{\rho}}$ defined in Equation (13). It is an approximation to the nonresponse bias of response means. The larger the value of Equation (14), the larger the risk of nonresponse bias.

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

$$CV_u\left(\hat{\rho}_g\right) = \frac{\sqrt{q_g}\left(\hat{\rho}_g - \widehat{\bar{\rho}}\right)}{\widehat{\bar{\rho}}}. \tag{15}$$

Equation (15) can be negative and positive, implying the specific stratum is underrepresented or overrepresented, respectively. The more negative Equation (15) is, the stronger the negative impact on representativeness of the stratum and the more effort the stratum needs. Since there are as many values for Equation (15) as there are strata, we focus on the strata that need effort the most, that is, that have the largest negative values.

Let $b(w)$ be the stratum in wave w that has the largest negative value of Equation (15), that is, $CV_u\left(\hat{\rho}_{b(w)}\right) \leq CV_u\left(\hat{\rho}_g\right), \forall g$. Learning early on in data collection what strata need extra effort is crucial for implementation of adaptive survey designs.

Under the Beta distribution priors for the response propensities, the priors (and posteriors) of the quality indicators have no closed forms; they are complex functions of response propensities. The priors (posteriors) can, however, be approximated by drawing a large number of samples from the priors (posteriors) of stratum response propensities. Given the advantage of our method, the conjugate distributions allow us to efficiently and rapidly obtain numerical iterations in Subsection 4.4.

3. Expert Elicitation

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

3.1. Data Collection Staff As Experts

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

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

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

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

The other three contributions concern design features. A survey design has a number of features such as modes and topics. The more similar design features of two surveys are, the more likely it is that response rates will be similar. In Subsection 3.2, we describe the features that were chosen in close collaboration with data collection staff. For each historic survey, the similarity on each feature must be scored. The procedure to do this is explained in Subsection 3.3. In Subsection 3.4, the importance of the features is weighed, again after consulting data collection staff.

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

3.2. Features for Deriving Similarity Between Surveys

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

1. *Topics/themes of the survey*: The more similar the topics of the survey to the new survey, the more similar participation rates should be,
2. *Target population*: Response rates depend on characteristics of persons and households. If target populations differ, then response rates will be different due to the different composition of characteristics. If the target population of the new survey is a subset of the historic survey, then in some cases the subset can be selected and target populations can be made the same. If the historic survey target population is itself a subset, then such harmonization is not possible,
3. *Time elapsed since last fieldwork*: Response rates change in time and the older the historic data, the more change should be expected,
4. *Unit of observation*: Two observation units are distinguished, persons and households. When the unit is different, then response rates will differ,
5. *Mode strategy (including contact and reminder)*: Survey modes and the order in which they are presented to respondents is an influential design feature in both contact and participation rates. The more similar the set of modes and the order in which they are offered, the more similar response rates will be,
6. *Incentive strategy*: The type and amount of incentive are influential for participation rates. The more similar the amount, the more similar participation rates are expected to be,
7. *Respondent effort*: Respondent burden affects participation rates, especially when the burden is salient to sample units. However, when it is not salient at the start, break-off rates are higher for longer surveys, and
8. *Bureau effect relative to Statistics Netherlands*: All else being equal, response rates do vary between survey institutions. This so-called bureau effect is, therefore, included as a design feature.

One important remark is in place: Response propensities are needed at the level of a pre-specified set of population strata. In order to form the strata, the sample needs to be enriched with auxiliary variables that define the strata. Within the same survey institution, sometimes the same auxiliary variables can be linked or the same population tables can be derived. However, if some or all auxiliary variables or tables are missing, then stratum response propensities cannot be estimated. If variables are missing, then a potential solution is to choose constant response propensities for the categories of these variables in

the prior distribution. In this article, it is assumed that historic survey data sets have no missing auxiliary variables.

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

3.3. *Scoring the Survey Features*

Operationalization of the similarity between two surveys in terms of a [0,1] score is not straightforward for most of the eight features. The easiest to score may be feature three, *Time elapsed,* as this is quantitative. However, rather than making the operationalization of scores as objective as possible, which is deemed very hard, it was decided to ask three experts independently and request them to reach a consensus. We constructed similarity parameters $\gamma_k \in [0, 1]$ for each historic survey as follows:

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

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

3.4. *Weighting the Similarity Scores on the Survey Features*

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

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

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

4. Two Case Studies to Investigate the Incorporation of Historic Surveys and Expert Elicitation

The effect of the power prior is compared to a non-informative prior in a Bayesian framework using two case studies, the Dutch Energy and the Dutch EU-SILC. In both

Table 1. Feature importance weights.

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

cases, the survey design was new and no direct historic information was available. The interest is in the added benefit from the inclusion of historic data and expert elicitation.

First, an RMSE evaluation criterion is introduced to evaluate the gains from a power prior in Subsection 4.1. Second, the Energy and SILC data are described briefly in Subsection 4.2. Next, in Subsection 4.3, the scores over similarity criteria are presented and powers for the historic data are derived. Finally, the posterior credible regions of quality indicators, R-indicator and CV, are illustrated as a function of the data collection wave in Subsection 4.4.

4.1. The Evaluation Criterion

In this section, we explain how we assess and compare the performance of non-informative and informative priors. Our strategy consists of evaluating the prediction of the three metrics, R-indicator, CV and CV_u, defined in Equations (13) to (15). To evaluate prediction accuracy, we consider the root mean square error (RMSE) of the predicted indicators against their realizations per data collection wave. The criterion RMSE is defined for the overall metrics as

$$RMSE\left(\theta; \pi_{0,T}\right) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{(\hat{\theta}_t - E_{\pi_{0,t-1}}(\theta))^2 + var_{\pi_{0,t-1}}(\theta)}, \tag{16}$$

where θ is the parameter of interest, i.e., R-indicator or CV, $\hat{\theta}_t$ is the realized value of this parameter in wave t, $\pi_{0,0}$ is the prior based on historic survey data, and $\pi_{0,t-1}$ is the posterior based on historic data and new data up to wave $t-1$. As noted, T is the present wave where the latest sample is released and Equation (16) is a rolling average of the RMSE until that wave. Smaller RMSE implies that the accuracy has improved and decisions about allocation of effort and budget can be made at an earlier stage in data collection.

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

$$RMSE\left(p_{g,B}\right) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(p_{g,B}^{t-1} - 1_{b(t)}(g)\right)^2 + \frac{1}{G}\sum_{g=1}^{G}p_{g,B}^{t-1}\left(1 - p_{g,B}^{t-1}\right)}. \quad (17)$$

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

The posterior terms in Equations (16) and (17), expectation and variance, are estimated by empirically drawing 10,000 samples from the beta posterior of stratum response propensities using Equations (11) and (12), and then computing the quality indicators for each iteration through the formula of quality indicators under formulas (13) to (15) in Subsection 2.3.

4.2. The Two Case Studies

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

4.2.1. Energy Survey

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

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

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

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

4.2.2. SILC

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

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

Two historic surveys were selected by data collection staff: the 2016 Dutch Labor Force Survey (LFS16) and the 2015 Dutch Household Budget Survey (HBS15). Both are conducted by Statistics Netherlands. LFS16 employed the same mixed-mode design, but added face-to-face to web non-respondents without a known phone number. HBS15 used the same design as SILC16 but is a diary survey. LFS16 was selected because the topics, survey modes, and unit of observation (the household) are very similar. The HBS15 was selected because the topics and modes were similar and because it was the only survey that used incentives in a household setting. The overall sample sizes were 7,954 (7,955) for SILC16 without (with) incentive, 24,882 for LFS16, and 8,182 for HBS15.

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

The number of waves varies over the two studies. Waves are chosen such that they correspond to time points where data collection may be adapted. For the EN18 case study 15 waves are chosen corresponding to different sample portions fielded throughout the

period February 2018 to September 2018. For the SILC16 case study three waves are chosen, corresponding to the three data collection months: April, May, and June 2016.

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

4.3. *Similarity Scores for the Two Case Studies*

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

Table 2 gives the feature-level similarity scores on the EN18 case study for the EN06, EN12, SC16 and HS18, and the combined scores by two types of weights. The EN06 and EN12 have perfect scores for the criteria related to topic, target population, observation unit and respondent burden, but have low scores on the other criteria, especially time elapsed. The other two surveys score relatively well on these other criteria. The feature scores are combined in two ways: one is by weighting all features equally and one is by using the weights from expert staff in Table 1. For each historic survey, the expert-based weight yields lower score than the equal weight.

Table 3 shows the similarity scores in the SILC16 case study for the two historic surveys LFS16 and HBS15. Recall from the Subsection 4.2, the sample was randomized into two parts. For the incentive strategy criterion two scores are given, one for SILC16 without incentive and one for SILC16 with incentive. The only perfect score is for LFS16 as it was conducted very close in time to SILC16. When topic is considered, observation unit and respondent effort criteria both historic surveys score weakly. The expert weighting has a strong impact on the similarity scores.

4.4. *Posterior Distributions for the Aggregate Quality Indicators*

In this subsection, our primary objective is to evaluate whether our method outperforms the non-informative prior in predicting quality indicators, and to look into whether the performance of our method would depend on the approach to pool the historic-specific

Table 2. Similarity scores per survey feature for the four surveys in the EN18 case study.

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

Table 3. Similarity scores per feature for the two surveys in the SILC16 case study. The incentive strategy criterion is scored for the SILC16 without incentive and with incentive.

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

criteria. They are illustrated by the RMSE evaluation criterion, applied to both case studies (EN18 and SILC16), with credible regions for overall indicators. Overall R-indicator and CV in Equations (13) and (14) are a function of wave for either informative or non-informative priors. The survey, SILC16, is investigated under two scenarios, with incentive and without incentive. Recall from Subsection 4.2 that the sample was randomized into two parts.

For the EN18, Figure 1 displays the overall indicators predicted by our method using expert elicitation and relevant historic surveys in contrast with the noninformative method

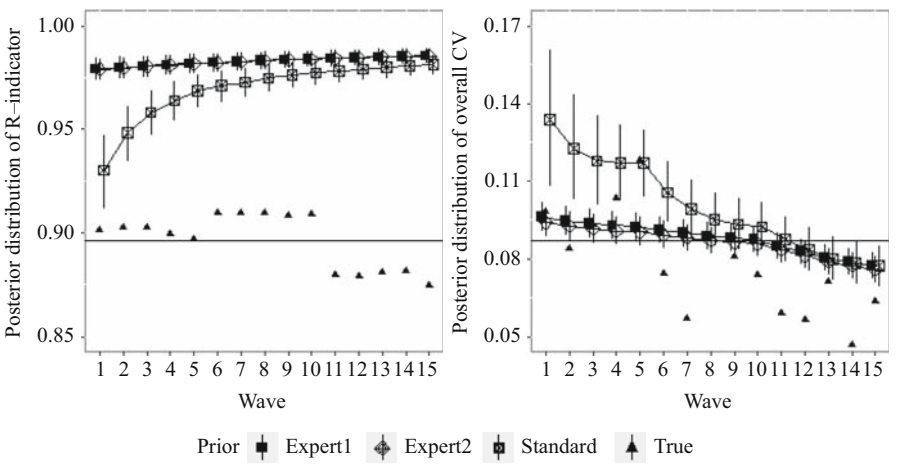

Fig. 1. The 95% credible regions of the posterior distribution of R-indicator and overall CV in the EN18 case study.

Note: In each wave, the posterior distributions are constructed by updating three different priors: the power priors (Expert 1 with equal weights and Expert 2 with expert weights), and the non-informative prior (Standard). The observed values are presented as well at wave (True) or overall (horizontal line) level.

as well as the realized indicators (target predictions). The horizontal line is the indicator realization of the population. For ease of explanation, our proposed priors are called expert priors and the non-informative prior is called the standard prior.

For each wave, 95% credible regions summarize the posterior expectations and their uncertainty for either expert priors or the standard. In early data collection waves, either expert prior has a small uncertainty on predictions versus the standard prior. For example, in wave 1, expert priors predict R-indicator with 1% uncertainty while for the standard prior it is 3%, and uncertainty levels are 1.3% and 6% for expert priors and standard prior in predicting overall CV. The standard method can predict *R*-indicator with increasing precision when more data collection waves are released as the width of the credible intervals decreases (by at most 2%), however, there is no significant reduction in the precision of posterior predictions for either expert prior, only a 0.3% decline. Moreover, none of the priors can completely adapt to the upcoming data, while the standard method pushes its prediction toward the target R-indicators slightly better than expert priors in early waves. The difference of posterior expected R-indicators between priors is 5% at most in wave 1 and declines fast to 1% in wave 4. Later on in predicting R-indicators, the standard prediction can catch up fast with expert prior predictions.

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

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

As we see in Table 4, the power priors (Expert 1 and Expert 2) predict a lower risk of nonresponse bias than the non-informative prior (Standard), where the expectation of R-indicator is closer to 1 and overall CV is closer to 0. Additionally, the power priors have little uncertainty about the expectation since the credible interval is much narrower than the non-informative prior. This is obvious because historic surveys provide information on the likelihood of response propensities.

Understanding to what extent expert priors add value to predict target indicators is revealed by RMSE in Table 5.

Table 4. The expectations and 95% credible regions (in brackets) of R-indicator and overall CV in Wave 0 from three different prior distributions in EN18 case study.

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

Table 5. RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for overall R-indicator, overall CV, and partial CV for the EN18 case study.

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

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

Figure 2 shows the comparison of expert priors with two weights to the standard prior in the SILC16 under two scenarios (with incentive and without incentive). The results of R-indicators show that predictions from either expert prior is superior to the standard prior in

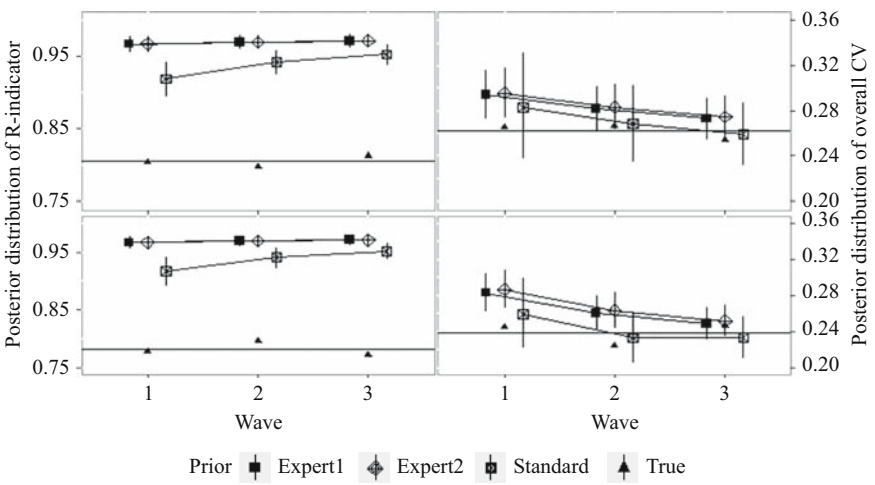

Fig. 2. The 95% credible regions of the posterior distribution of R-indicator and CV in the SILC16 case study. Note: In each wave, the posterior distributions are constructed by updating three different priors: the power priors (Expert 1 with equal weights and Expert 2 with expert weights), and the non-informative prior (Standard). The observed values are presented as well at wave (True) or overall (horizontal line) level. The top/bottom panel corresponds to without/with incentive.

wave 1 as the standard variance is five times larger than the expert prior variance, although the standard posterior mean reaches slightly the observation than either expert prior. This advantage of either expert prior continues in wave 2 but the standard competes with them as wave 3 showed. Either expert prior has obvious benefit versus the standard prior on the measure of overall CV. However, this advantage gradually declines and the standard method catches up as accumulating data.

In Wave 0 (Table 6), the power priors behave much better than the non-informative prior in predicting the overall variation in response propensities and in the uncertainty measurement, regardless of whether there has incentive in the SILC16. R-indicators of either power prior take the values closer to one, and overall CVs take smaller values than the non-informative prior.

As Tables 7 and 8 show, the RMSE of the predicted R-indicator from expert priors is smaller in first two waves, and in wave 3, the standard is competitive to or slightly weaker than expert priors. When predicting overall CV, only in wave 1 either expert prior shows

Table 6. The expectations and 95% credible regions (in brackets) of R-indicator and overall CV in Wave 0 from three different prior distributions in the SILC16 case study.

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

Table 7. RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for three indicators in the SILC16 without incentive.

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

Table 8. RMSE of the informative prior (Expert) with two weights and the non-informative prior (Standard) for three indicators in the SILC16 with incentive.

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

its superior performance, because the advantage of expert priors is overwhelmed by the size of upcoming data in one wave. However, the benefit to use our method is strongly supported by the RMSE of R-indicator and partial CV under either scenario. RMSE from either expert prior is consistently closer to 0 and better than the standard prior.

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

We also apply our method to predict the level of response propensities, results shown in Appendix (Subsection 6.2). The experts are uncertain about the level of response propensities. In this article, our primary concern is with the variation of response propensities more than with the level as adaptation is based on underrepresentation of certain strata.

5. Discussion

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

To include expert knowledge, we set up a procedure that takes a power prior as a key ingredient. The powers are derived by scoring a number of historic surveys on their similarity to the new survey. Scores are based on the identification of a number of relevant survey design features, which are weighted based again on expert opinions. In our case studies, we invited three data collection experts and averaged their scores. The power prior is updated using incoming survey data during data collection. The performance is evaluated based on three quality indicators: overall R-indicator and overall (partial) coefficient of variation of response propensities. The prior and posterior of the variation in response propensities and coefficient of variation of the response propensities have no closed form. However, since Beta distributions are conjugate for the response propensities, it is relatively straightforward to construct the posteriors empirically. Consequently, the first three objectives of the article are achieved to monitor and adapt the survey data collection for the new survey to the subsequent phases. Our prior elicitation procedure has been set up in collaboration with data collection staff. We paid a lot of attention to making the procedure transparent, but also manageable. We view bridging the gap between data collection and methodology as the most important achievement of this study.

Our other objective was to assess the benefit of incorporation of the historic prior information into the new survey against the settings without prior knowledge. In the evaluation, a fully non-informative prior implies that no historic surveys and no expert knowledge can be used to specify a prior for a new survey. To achieve this goal, the root

mean square error (RMSE) of the posterior of quality indicators is evaluated. The evaluation was made based on two case studies, the EN18 and SILC16, using the observed indicators and the posterior prediction with a series of samples released in time. Both case studies show that the approach to weight the survey features have no influence on the comparison of a Bayesian analysis against a non-Bayesian analysis, because the RMSE is only slightly distinct between equal weights and expert weights. The evaluation study shows either power prior can be vastly superior to a non-informative prior on predicting the variation in response propensities as well as the coefficient of variation of them throughout the course of new survey data collection. The advantage holds to predict CV but in late waves the non-informative prior can compete with expert priors. So far, we conclude that the power prior clearly has added value, but its prediction performance is closely related with the choice of historic surveys, the selected criteria, and the prescribed feature-level weights elicited from the staffs. Therefore, we propose to carefully use a power prior for predicting response propensities and related indicators when a survey is brand new, which ought necessarily to be compared with non-informative predictions.

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

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

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

In the proposed method, we regard the timeliness of historic data sets as a similarity criterion. Assessment of the criterion obviously depends on the time-length of the historic data, e.g., the last quarter or the last year. The longer the time length the harder it is to provide a single value as the data contain both very recent and relatively old data. Extra uncertainty is introduced by assuming a constant timeliness. Therefore, the issue involving how far a researcher should go back for picking up historic data sets should be addressed as a future topic.

6. Appendix

6.1. The Stratification of Two Case Studies

Table A1. Definition of the EN18 strata across categorical variables, ownership, type of dwelling, and year of construction.

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

Table A2. Definition of the SILC16 strata across variables, age, household size, and income deciles.

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

6.2. The Level of Response Propensities

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

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

where $\hat{\rho}_g$ is the response propensity in stratum g, and $\widehat{\bar{\rho}}$ is overall response rate explicit about the level of the population response propensity. Equation (16) applies to evaluate the performance of predicted RR of our method.

Table B1. RMSE of the predicted RR from informative prior (Expert) with two weights and the non-informative prior (Standard) for the EN18.

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

Table B2. RMSE of the predicted RR from informative prior (Expert) with two weights and the non-informative prior (Standard) for the SILC16 under two scenarios.

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

7. References

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. John Wiley and Sons.

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

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

Coffey, S., B.T. West, J. Wagner, and M.R. Elliott. 2020. "What do you think? Using expert opinion to improve predictions of response propensity under a bayesian framework." *Methods, Data, Analyses* 14(2). DOI: https://doi.org/10.12758/mda.2020.05.

Gelman, A., H.S. Stern, J.B. Carlin, D.B. Dunson, A. Vehtari, and D.. Rubin. 2013. *Bayesian data analysis.* Chapman and Hall CRC.

Gosling, J.P., J.E. Oakley, and A. O'Hagan. 2007. "Nonparametric elicitation for heavy-tailed prior distributions." *Bayesian Analysis* 2(4): 693–718. DOI: https://doi.org/10.1214/07-BA228.

Ibrahim, J.G., and M.H. Chen. 2000. "Power prior distributions for regression models." *Statistical Science*: 40–60. DOI: https://doi.org/10.1214/ss/1009212673.

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

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

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

O'Hagan, A., C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow. 2006. *Uncertain judgements: Eliciting experts' probabilities.* New York: Wiley.

Oakley, J.E., and A. O'Hagan. 2007. "Uncertainty in prior elicitations: A nonparametric approach." *Biometrika* 94(2): 427–441. DOI: https://doi.org/10.1093/biomet/asm031.

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

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the representativeness of survey response." *Survey Methodology 35*(1): 101–113. Available at: https://www150.statcan.gc.ca/n1/pub/11-522-x/2008000/article/10976-eng.pdf.

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

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive survey design*. CRC Press.

Schouten, B., and N. Shlomo. 2017. "Selecting Adaptive Survey Design Strata with Partial R-indicators." *International Statistical Review 85*(1): 143–163. DOI: https://doi.org/10.1111/insr.12159.

Veen, D., D. Stoel, M. Zondervan-wijnenburg, and R. van de Schoot. 2017. "Proposal for a five-step method to elicit expert judgment." *Frontiers in Psychology* 8: 2110. DOI: https://doi.org/10.3389/fpsyg.2017.02110.

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

Received March 2021
Revised September 2021
Accepted January 2022