Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Modelling time change in survey response rates: A Bayesian approach with an application to the Dutch Health Survey

by Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek
and Barry Schouten

Release date: June 30, 2023

Statistics Canada Statistique Canada

Canadä

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2023
Vol. 49, No. 1, pp. 163-190
Statistics Canada, Catalogue No. 12-001-X

Modelling time change in survey response rates: A Bayesian approach with an application to the Dutch Health Survey

Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek and Barry Schouten[1]

Abstract

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

Key Words: Response propensity; Time series; Multilevel model; Bayesian analysis.

1. Introduction

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

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

1. Shiya Wu, Utrecht University, Department of Methodology and Statistics. E-mail: s.wu@uu.nl; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods; Mirjam Moerbeek, Utrecht University, Department of Methodology and Statistics; Barry Schouten, Statistics Netherlands, Department of Statistical Methods and Utrecht University, Department of Methodology and Statistics.

should be incorporated into propensity models in order to avoid suboptimal designs (Burger, Perryck and Schouten, 2017).

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

So far, the approaches assume RPs are stable in a relatively short period. In a fixed survey climate, these parameters remain stable and their estimates ultimately converge with the accumulation of historic data. In survey practice however, those parameters change gradually over time, which means that predictions may not converge. For example, seasonal variation and downward trends in response rates can be observed. Thus, the benefit of prior elicitation could potentially be undone when ignoring time change. Recent articles by Mushkudiani and Schouten (2019), and Fang, Burger, Meijers and van Berkel (2020) describe what time-dependent factors significantly affect the parameter estimation accuracy, but the impact on prediction accuracy is still unknown, which is the topic of this paper.

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

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

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

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

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

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

This paper first introduces several time-related factors of great relevance to variation with a hypothetical illustrative example in Section 2, then goes on to the differential model compositions in the general form of the Bayesian multilevel time-series model in Section 3. Section 4 optimizes the model performance based on an empirical analysis of GEZO. We discuss our findings and end up with the brief overview of future work in Section 5.

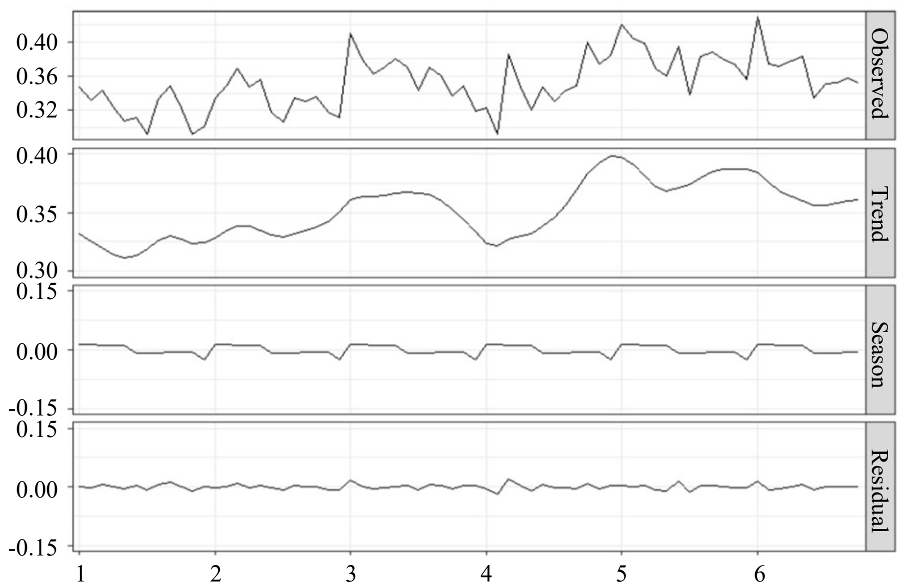
2. Time series components of survey response rates

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

We focus on population subgroups, or strata, as indexed by $g \in \{1, \ldots, G\}$, since we aim ultimately to let the proposed models inform adaptive design decisions. The strata are formed with the help of auxiliary variables that are linked to the sample and are, thus, available for all sample units. A time-series RP $\rho_{g,t}$ in stratum g and time t is a sequence of random variables. Assuming the availability of historic survey data up to time t, we are interested in measuring variation caused by time-related factors for the most up-to-date RP predictions. To achieve this goal, we first propose potential time-dependent factors. As an illustrative example of a time series divided into the following components: trend, seasonality, and so on, Figure 2.1 compares the overall response rate to the following time-dependent variation:

- *Trend*. The trend describes the long-term movement of the observed time series without the seasonal variation. It shows the general tendency of the population-level response rates over years, which can be linear or nonlinear. Hence, the growth or the fall of the long-term forecasts can be studied by this trend. As seen in Figure 2.1, the long-term direction does not behave like a cyclic fluctuation. Of greater importance for model development is to separate the total trend into a global trend shared by all strata, and local, i.e., stratum-specific trends.
- *Seasonality*. Seasonal variation in the overall responses describe periodic movements that recur regularly and do not influence annual averages. The periodic fluctuations possess a systematic and calendar-related nature that can be predicted and attributed to a fixed season per year. For instance, the response rate would be higher in the early period of the year while relatively lower in the middle year or in December.
- *Residual fluctuation*. The residual variation is the part of the signal obtained after excluding all of the above components. This part is usually modeled as white noise, i.e., as independent normally distributed fluctuations.

Figure 2.1 The observed series of simulated overall response rates over years versus its decomposition.



In addition, there may also be some additional time-dependent components not revealed in Figure 2.1 that nevertheless have a strong impact on the reliability of stratum RP predictions. For this reason, we also consider extra stratum-related time-dependent components:

- *Stratum*. Different subgroups have different response behaviors, such as, young subgroups are more likely to respond to the web survey than old subgroups due to the latter having potentially less access to or unfamiliarity with the internet. This variation in subgroups leads to a differential stratum-level trend and could potentially also contribute to differential seasonal movement.

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

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

3. Methods

In this section, we translate the time series components discussed in Section 2 to multilevel time series models and devise the estimation strategy. We adopt a Bayesian approach in order to account for the uncertainty within the historic survey data and to update response propensity (RP) predictions in time. The use of multilevel models is widespread in small area estimation, in which interest focusses on reliable estimation for domains such as geographic areas, time periods, demographic subgroups, or a combination thereof, whose sample sizes are often too small to provide reliable direct estimates, see Rao and Molina (2015) for an overview. Early references to the literature of small area studies using time series multilevel models include Pfeffermann and Burck (1990), Rao and Yu (1994), Datta, Lahiri, Maiti and Lu (1999), You, Rao and Gambino (2003). In most such studies, including Boonstra and van den Brakel (2019), a Gaussian sampling distribution is assumed, possibly after a suitable transformation of the data. A notable difference of our current application to RPs is that we use a binomial sampling distribution, which is a natural distribution to describe the response process given the number of sampled individuals in each demographic subgroup and time period. Such binomial time series models have been considered by Franco and Bell (2015). Their approach bears a resemblance to our strategy, whereas ours involves more different types of time series components in the model specification, such as seasonality.

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

3.1 The multi-level time series model specification

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

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

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

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

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

For any stratum g and any time t, the linear predictor $\theta_{g,t}$ can take the most general form that can be linear, additive, multilevel and comprised of several time series components. As outlined in Section 2.2, there are demographic variables defining the strata, an overall trend, seasonal variation, stratum-specific trends, and a residual variation. Therefore, the multilevel model becomes:

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

where the p-vector of regression effects $\boldsymbol{\beta}$ is associated with time-independent covariates $\mathbf{x}_g$. In the application we focus on later in this paper, all covariates are binary as we only consider categorical variables. However, in more general usage, the entries could be ordinal or numerical variables, such as contact attempts, and even they could vary over time.

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

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

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

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

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

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

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

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

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

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

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

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

using a single variance parameter σ_w^2.

(3.1) describes the most general model considered combining all underlying components. Section 4 investigates this encompassing model as well as models built from various subsets of the components described in (3.2)-(3.5).

3.2 The estimation strategy

In this section, we adopt a hierarchical Bayesian approach to estimate model coefficients and predict RPs. Since the posterior distributions are unavailable in closed form a Gibbs sampler is used as implemented in the *mcmcsae* R package (Boonstra, 2021). We begin this subsection by specifying the priors assigned to the model parameters.

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

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

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

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

The (hyper) parameter vector, denoted by ψ,

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

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

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

where $\rho = \text{logit}^{-1}\left(\theta\left(\psi\right)\right)$ and θ is the linear predictor function of vector ψ as expressed in (3.1). Based on Bayes' theorem, the posterior of vector ψ is proportional to the product of the prior and the likelihood, i.e., $p\left(\psi \mid n, r\right) \propto p\left(\psi\right) p\left(r \mid n, \psi\right)$. The Gibbs sampler then generates samples from the joint posterior, and the posterior estimates of RP $\rho_{g,t}$ comes as a by-product of these samples – per sample, RPs can be computed using reversed logit transformation. Repeated samples are drawn from the full conditional posterior of each (hyper) parameter. See Appendix D for more information on the full conditional posterior distributions.

Three Markov Chains are produced by the Gibbs Sampler using the *mcmcsae* package (Boonstra, 2021) programmed in R (R Core Team, 2020). Each chain consists of 1,500 draws that are sequentially generated; however only the last 1,000 draws are kept for the estimation algorithm. Convergence of the MCMC sample is assessed using trace and autocorrelation plots. The Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992) is evaluated to diagnose the mixing of the chains. In particular, the autocorrelation of sequential draws is reduced, as the blocked Gibbs sampler updates all fixed and random coefficients simultaneously. In addition, the approach includes a novel data augmentation approach for sampling from binomial logistic models (Polson, Scott and Windle, 2013) which is known to lead to an efficient and relatively fast converging sampler.

3.3 Performance criteria

To guide the model building using the model components and priors described in Sections 3.1 and 3.2, and to assess the models' adequacy, we employ three criteria for model assessment and one for model predictive performance.

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

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

where S is a test statistic and r^{rep} denotes a replicated dataset generated from the posterior predictive distribution based on the fitted model, $p\left(r^{\text{rep}} \mid r\right) = \int p\left(r^{\text{rep}} \mid \rho, n\right) p\left(\rho \mid r, n\right) d\rho$. The p-values are estimated from the MCMC output, and values close to 0 or 1 are indicative of a poor fit regarding statistic S. Here we consider two test statistics:

1. $S\left(r\right) = \bar{r}$, the unweighted mean of the replicate data-vector.

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

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

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

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

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

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

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

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

Thus, the sub-terms

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

and

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

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

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

Also, the stratum-specific sub-terms

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

and

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

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

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

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

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

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

and

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

4. Analysis of results

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

4.1 The Dutch Health Survey

The GEZO has been conducted annually since 1981 by Statistics Netherlands as a repeated cross-sectional survey in which a sample of households was interviewed with the aim of providing an overview of developments in the health, medical consumption, lifestyle and preventive behavior of the Dutch population. The sampling frame is formed by first drawing a sample from municipalities and then from all people who live in the selected municipalities. As of 2010, the survey changed to a mixed-mode survey involving an initial web and the follow-up telephone (or face-to-face) interview. Non-respondents to web were contacted via telephone if their telephone numbers were known at the register, and otherwise a face-to-face interview was arranged. Over these years, the sample size was increased to 15,000 and the overall response rate was increased by 25%. From 2014 onwards, the mix of the follow-ups was changed to a face-to-face interview. In 2018, however, a part of the web non-responses was approached via a face-to-face interview in a more effective way. The propensity to respond to personal interviews in a time series strongly depends on web response outcomes, so in a sense modeling follow-up propensity is conditional on web RP model. This issue needs consideration more than interpreting time change in web RP and is beyond the main aim of this paper. For the sake of simplicity, our concern is to model web response propensity in this paper as a fundamental start, hence modeling follow-up RP in a time series is more suited to future research. As an important note here, only the web GEZO data from 2014-01 up to and including 2019-10 are analyzed in this study. We employ three auxiliary variables that stem from administrative frame or registers. The prescribed auxiliary variables are age, gender and ethnicity, which divides the population or its sample into 20 disjoint strata (see Appendix A for more information).

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

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

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

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

1. Start with the baseline model (auxiliary variables only).
2. Add fixed effects sequentially, linear time trend and seasonal trends, to the baseline.
3. Investigate whether the model in 2 continues to improve with global time effects or global seasonality.
4. Investigate whether stratum-specific time trends or seasonal effects further improve the model.
5. Determine whether a white noise term for unexplained variation is needed.
6. Explore robustness for outliers through different prior specifications or time-dependent structure of global random intercepts over time.
7. Evaluate the model using a number of diagnostics.

Table 4.1 shows the selection results. The fixed-effect models (M1 to M3) behave worse than the mixed effect models relative to the trade-off between fitness and complexity, as the latter ones yield lower ICs (DIC, WAIC). Comparing M2/M3 to M1 implies that time slope λ or seasonality δ causes a decrease in ICs. However, the model further improves by introducing global trend u_t, as a significant decrease in ICs in M4 relative to M3 is observed. As M5 and M6 show, the improvement continues with the addition of random intercepts for strata v_g and stratum-specific time trends $z_{g,t}$. Although white noise $w_{g,t}$ seems to add only very little in M7 overall, the posterior predictive p-values for variances imply that it is worth to include white noise. Further, we found that using a local level trend (RW1) or smooth trend (RW2) as the global trend u_t makes hardly any difference concerning ICs for models M6 to M11.

Finally, the 4th column of Table 4.1 shows the prior distribution used for the global trend coefficients u_t. The non-normal priors that have been attempted do not further improve ICs, but because of heavier tails they help to combat an outlier in the data, an exceptional issue in February 2017.

Table 4.1
Summary of the multilevel time-series models considered

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

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

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

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

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

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

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

Table 4.2 shows that the posterior uncertainty in the overall RP predictions decrease steadily but slowly and converge to around 0.027. Because of the sampling variation that is inherent to the bias term, the pattern for bias is erratic and shows at best a modest decrease. Relative to the realized response rates, the greatest deviation of posterior means is around 0.07 in January and June, and the smallest deviation around 0.04 in March, May, August and October. The RMSE results vary along with the bias term across months, as the estimated SDs are much smaller than the estimated biases. The RMSE has a maximum value of 0.084 in January, and is likely caused by the outlier months in early 2017. Although the model reacts to this disruption, it has a negative impact on the performance of the resulting predictions in this month. Aside from January, in some months the estimated RMSE is close to the benchmark $RMSE_{min}$. It can be concluded that the estimated accuracy lies relatively close to the maximal possible accuracy.

Table 4.2
One month ahead prediction of three measures of RPs over strata: bias, standard deviation (SD), and the root mean square error (RMSE) compared to the benchmark ($RMSE_{min}$)

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

Notes : The column indicates the present month for evaluating prediction performance.
Response propensities (RPs).

4.4 How does prediction accuracy vary over population strata?

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

Similar to Section 4.3, we limit the analysis to the assumptions. The preferred model is selected from Section 4.1, and the historic time series is fixed to be 60 months (2014 to 2018). For each stratum, the model is fully trained by the fixed historic data and makes inference on predictions in the remaining months in 2019.

Table 4.3 shows prediction criteria for each stratum together with the benchmark. The estimated bias terms vary widely between strata. The greatest departure of posterior expectation from the realized response rates occurs in stratum 8, 10, 12, and 18, all with biases larger than 0.1. Compared to biases, there is a relatively smooth change in the estimated SDs around 0.03, where stratum 4 has smallest uncertainty about the posterior estimates with 0.018.

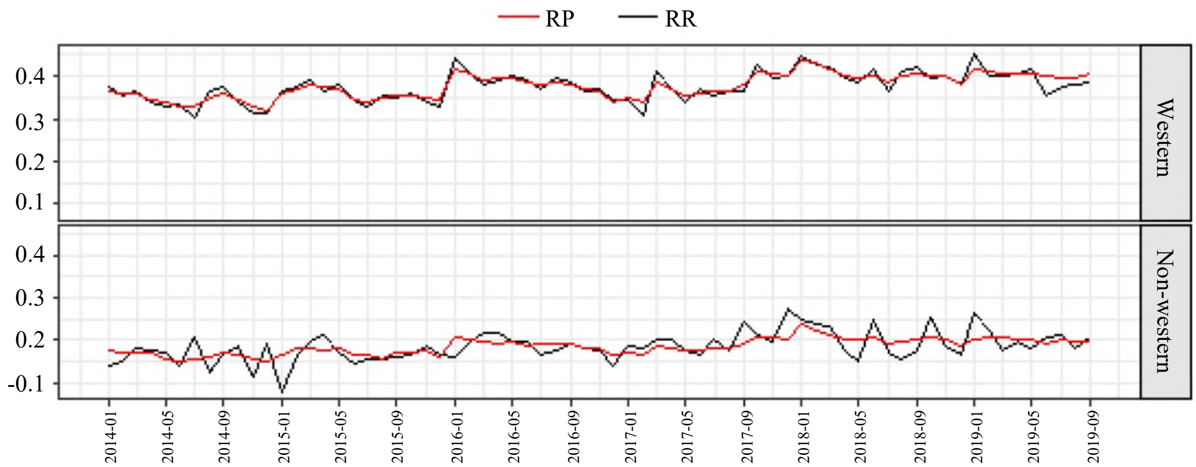
Some strata with greater biases may have less accuracy in posterior estimates of RPs than strata with less biased propensity. Similarly, the more biased the prediction is, the greater the RMSE is estimated to be. This is because the estimated biases are much greater than the estimated SDs. It is not surprising that stratum 10 has the greatest value in RMSE, where the prediction is the most biased and has the least precision. The RMSE results can catch up with, and even can be comparable/superior to the benchmark. For example, when the model generates prediction for stratum 20, more significant gains can be achieved than other strata.

Table 4.3
The average of ten months ahead prediction of three measures in each stratum: bias, standard deviation (SD), and the root mean square error (RMSE) that is compared to the benchmark ($\mathrm{RMSE}_{\min}$)

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

The predictive performance shows a significant difference between strata when there is only one different characteristic. For example, stratum 20 RMSE is 0.06 lower than stratum 10 RMSE. This seems to imply that female groups may have smaller bias or variance than male groups when non-western people above the age of 64 are considered. Given the age and ethnicity of groups and compared with non-western groups (even rows), RMSE results are much lower in western groups (odd rows). To validate this supposition, some strata are combined into subgroups with less detailed characteristics. As Figure 4.1 shows, the model yields better predictions for western group than non-western groups, as expected posterior estimates reach mostly the observed response per month. The comparative performance for age/gender groups are in Appendix B.

Figure 4.1 Monthly posterior means of RP aggregated over Ethnic groups versus observed response rates (RR) of Ethnic groups.



Note: Month 2014-01 to 2018-12 for the estimated model and Month 2019-01 to 2019-10 for RP predictions.

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

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

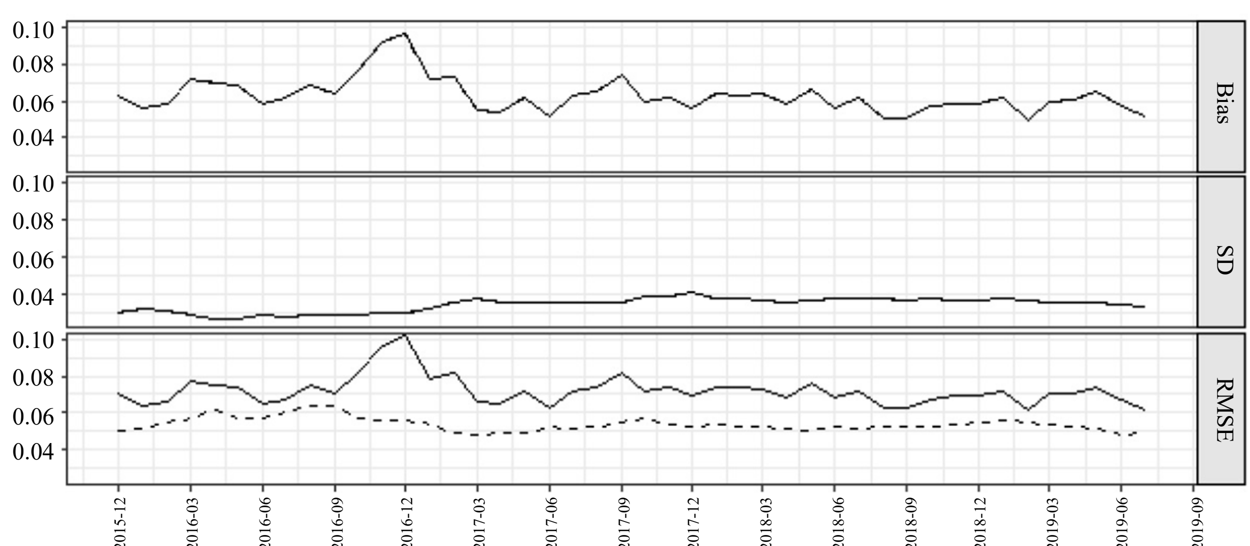
1. Select the model components based on the whole time series.
2. Select the baseline set of historic time periods of length t. Partition the window into the training set D_o of the first $t-3$ time periods and the test set D_t of the last three periods.
3. Data D_o trains the selected model, by simulating from the posterior distribution of all model parameters, given D_o.

4. Based on the simulated model in 3, posterior predictive means and variances are computed for the RPs for each stratum and each time point in the test set D_t.
5. Based on individual RPs predictions in 4, compute overall Bias, and SD, by using the sample proportion $d_{g,t}$ in stratum g at time t as weights, then equation (3.8) computes the quarterly average of RMSE.
6. Expand the window to $t+1$, moving one time period forward. Repeat 3-5 for updating the predictions of RMSE and its two terms.
7. Repeat 6 until length t is the last available time period.

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

In Figure 4.2 RMSE, Bias and SD estimates are length-dependent and computed over strata for three months ahead. The baseline window of training data for fitting the optimal model are 2014-01 to 2015-12. Along the time series window, bias results are approximately greater than two times the SDs. Consequently, RMSE results are dominated by biases and show the same volatility. At the end, their estimates are approximately 0.06, whereas SD estimates undergo a slight increase. The latter is somewhat surprising, as one would normally expect that using a longer time series to estimate the model would decrease the posterior standard errors of prediction. It turns out, however, that two events had a large impact on the prediction performance. First, early in 2017, data collection experienced an interruption caused by technical issues with the web server. This incident had a large immediate impact on RPs and consequently also on model prediction performances. Second, in 2018, conditional incentives were introduced and the survey questionnaire was made smartphone proof. This design intervention had a more gradual and longer lasting impact.

Figure 4.2 One-step forward moving average of quarterly Bias (upper panel), quarterly SD (middle panel) and quarterly RMSE (bottom panel). RMSE (solid) against benchmark RMSE (dashed) when the length of training data set moves on x axis.



The Bias and RMSE results undergo a big increase from 2016-10 to 2017-02. When the training window arrives at time point 2016-10, the test window starts to include 2017-01 data where RPs dropped. Their climbing curves continues and reach maxima around 0.1, when the training window moves to 2016-12 and the test window first moves to the "stable" month 2017-03 where the Bias and RMSE curves drop back around 0.05. During these months a slight, gradual increase in SD can be observed as well.

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

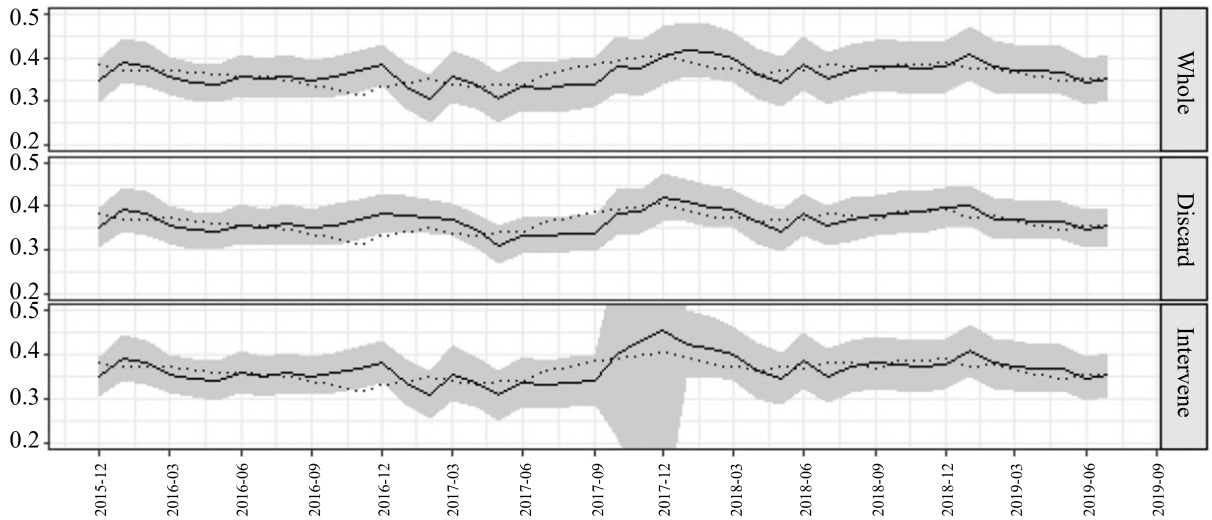
In 2018, the impact of the design intervention on RPs was much more modest than in 2017, but since it is structural it does affect prediction accuracy in 2018 and 2019. The two events, one technical incident and one design change, are realistic in survey practice and when ignored can have a devastating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

- *Discard method*. For the original data, clear the response numbers $r_{g,t}$ in 2017-01 and -02 and treat them as missing data. Impute these missing $r_{g,t}$ by the posterior means of simulated responses from the posterior predictive distribution. Note that in Section 4.2 we argued that using specific non-normal priors for time series components can also limit the effect of outliers. If an outlier is quite extreme and known to occur at a specific time, it may however be better to discard it.
- *Intervention method*. Include an intervention term in the model orcapturing the possible structural change. Add intervention binary variables to the original data series and let them be 0-1, where in our case they would take the value 1 and become active from 2018-01. The potential intervention-related effects could be either a single fixed effect, stratum-specific random effect or both.

Results from applying these two methods separately are shown in Figure 4.3.

The two methods have a clear effect on predictions. In the period of 2016-12 to 2017-05 where the training time series window stops, the posterior means of the discard method show a declining trend, relative to the original model's posterior means ("Whole" in Figure 4.3). However, from 2017-10 to 2018-03 the difference in the mode-based means and observations becomes small for the discard method. Also, the discard method decreases uncertainty about posterior means as the credible band becomes narrower since from 2016-12. The intervention-related impact on overall RP cannot be estimated well using just a few new months of data.

Figure 4.3 One-step forward moving average of quarterly posterior estimates of overall RP under three scenarios: (1) the original time series (top panel); (2) the new time series by discarding early 2017 data (middle panel), and (3) the new time series by adding intervention-related effects (bottom panel).



Note: Compared to the moving average of observed response rates quarterly (dashed), the model-based estimates are summarized as the posterior means (solid) with 95% credible region (band in grey). X axis labels the length of training data from 2014 to that time point.

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

5. Discussion

Accurate and reliable prediction of response propensities (RPs) is the key to improving and optimizing adaptive survey designs. Such inference can be complicated due to seasonal variation and time-related trends that may be specific to population strata. In this paper, we introduce a Bayesian multilevel time series model for stratum-level RP predictions. The model is flexible enough to include seasonal variation, various forms of trends, design changes and stratum-dependency, so that it can facilitate preparation for adaptive survey design in a changing survey climate. They are elicited from historic survey data and updated with new survey data.

In this paper, we apply the method to a general population repeated survey, the Dutch Health Survey at Statistics Netherlands, to provide empirical support in a realistic case. The major focus is on improving stratum-level RP predictions that are subject to time-related factors. Based on various model combinations made of these factors, one of our concrete objectives is to search for the highest-performance model that makes a trade-off between model fitness and computational ease. The optimal model is selected based on criteria that assess both performance (high IC, p-value ≈ 0.5) and predictive ability (low $\mathrm{RMSE}_{\mathrm{RP}}$).

These measures provide valuable insight into the relative gain achieved by adding new factors. This flexible approach allows other survey researchers to consider different time-related factors and ultimately choose the preferred model in their settings.

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

While our attempt is a first step to adaptive survey designs, there are, however, various methodological and practical considerations that should be addressed. First, our approach is applied to a frequently-repeated cross-sectional survey. Historic data in such surveys has rich resources for relatively robust estimates of model coefficients and for making reliable predictions. When a survey is novel or conducted infrequently at a statistics bureau, our approach may be less powerful. Second, we assume that stratification is done through a fully-saturated model, i.e., strata are pre-specified by some auxiliary variables that are strong predictors for web responses. How does the prediction performance change when adding less influential auxiliary variables? It is important to assess the sensitivity of reliable predictions to the choice of auxiliary variables. Also, we assume that strata are fixed throughout the time series. In survey practice, the selected auxiliary data may gradually change over time, and thus also the relevance of certain strata. Hence, it is essential to consider auxiliary-related change in stratification when predicting responses. Third, we assume the design of a survey should be the same over time, i.e., the model assumptions must be valid over the whole time series. If an intervention or another self-reported mode (e.g., smartphone) is introduced, variation in responses caused by this must be included explicitly. The advanced method is needed because there is no prior historic knowledge for a design change before it happens. A large jump can be caused by the inclusion of such a change in the model and, before the model can be informative about the effects of the change on RPs, a sufficiently long historic sampling must first be acquired.

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

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

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

Acknowledgements

This project was generously supported by Chinese Scholarship Council (CSC), Statistics Netherlands (CBS) and Utrecht University. We would like to thank Anouk Roberts for fruitful discussion.

Appendix A

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

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

Appendix B

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

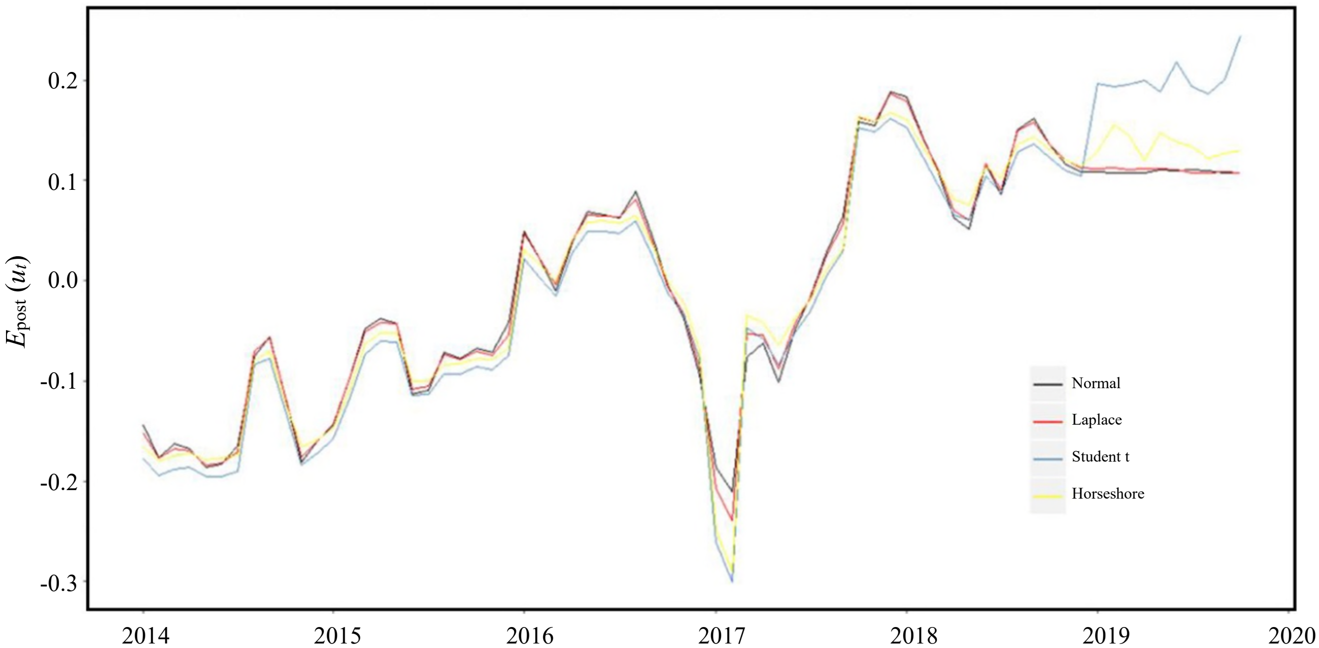
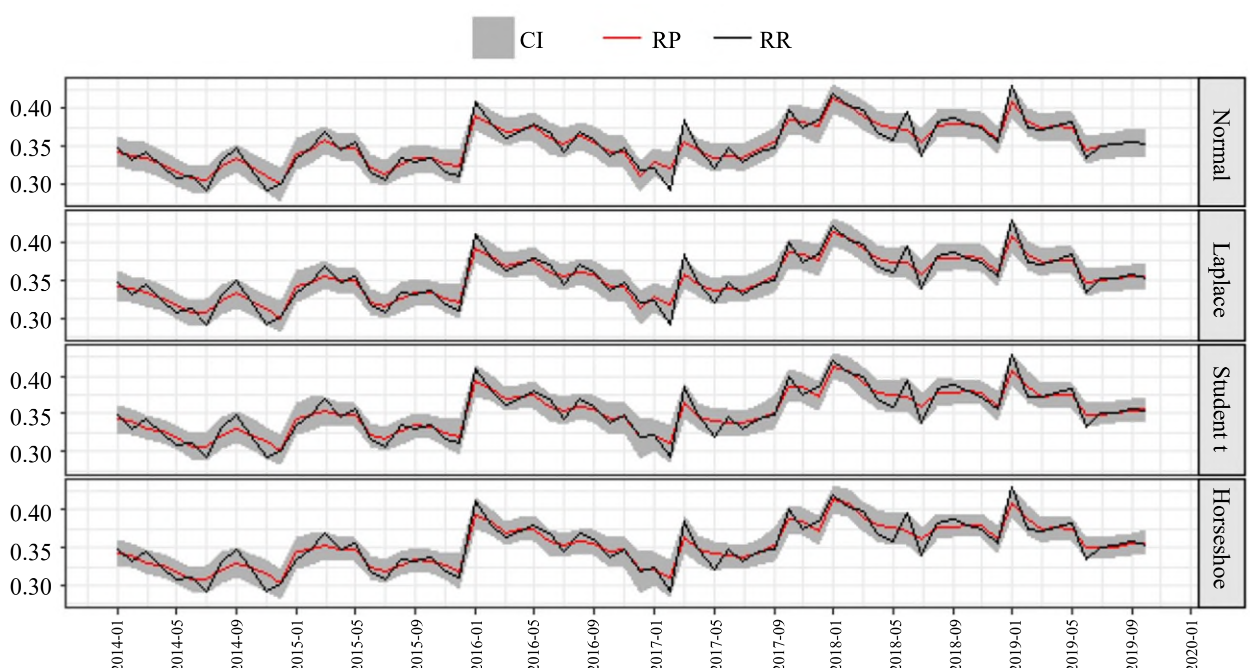
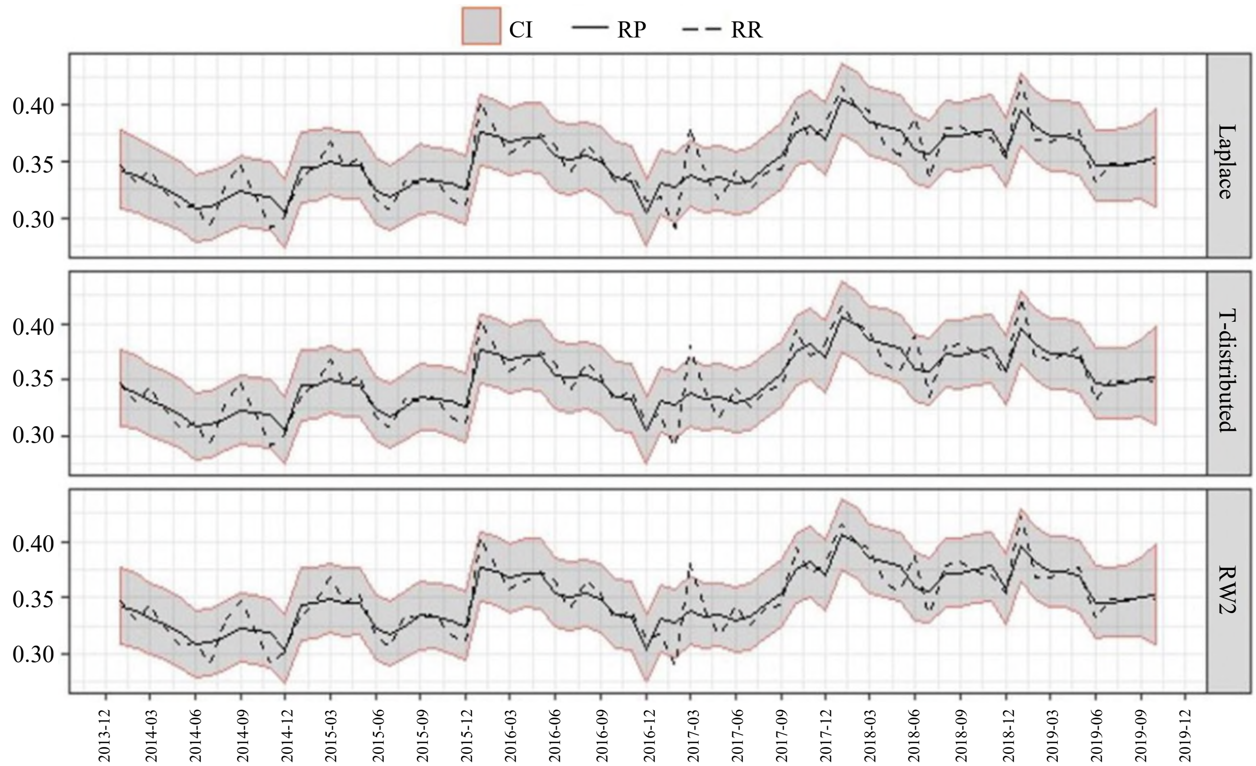


Figure B.2 Compare the posterior predictions of RP over strata made by four models (M7 to M10) to the observed RR and make a choice on the most compatible model with the observed outliers.



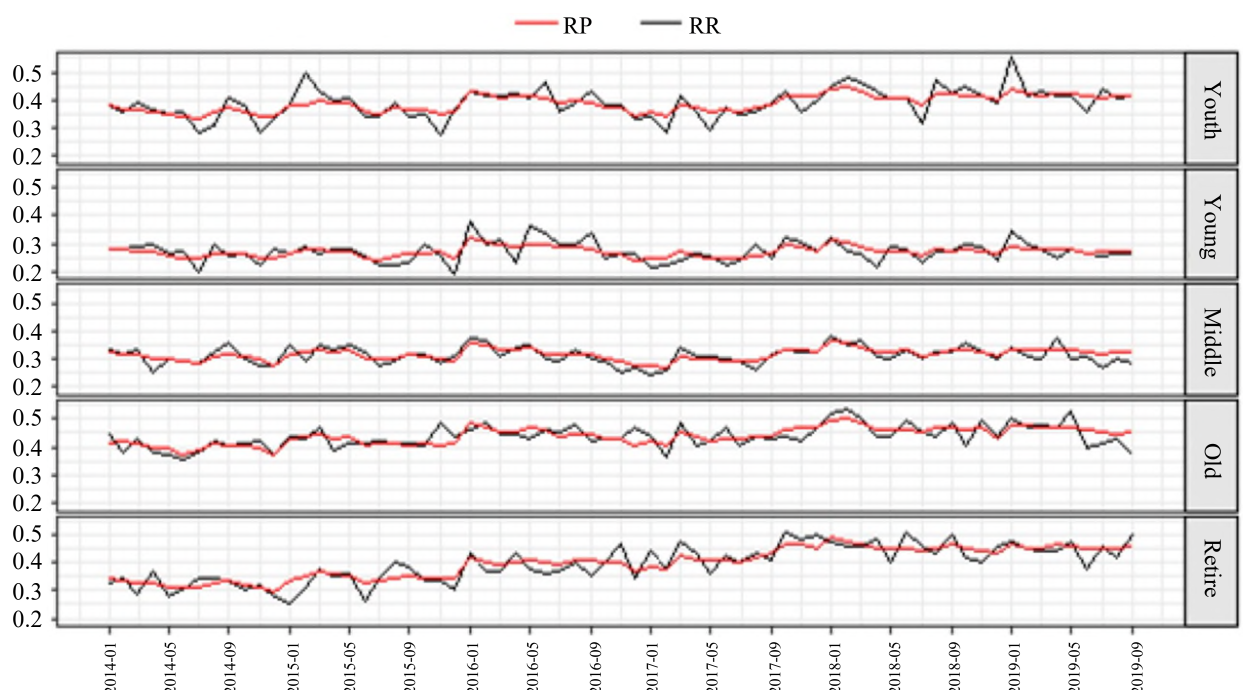
Note: The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

Figure B.3 The posterior predictions of RP over strata against the observed RR under M8, M9 and M11.



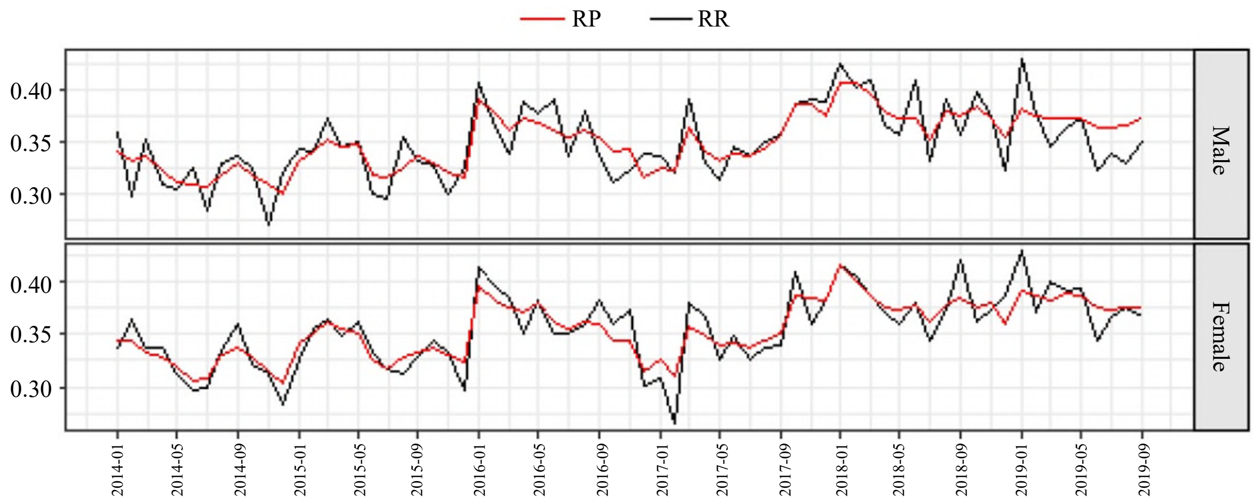
Note: The overall RP predictions are summarized as the posterior means (RP) and 95% credible region (CI).

Figure B.4 Monthly posterior means of RP of Age groups versus observed response rates (RR) of Age groups.



Note: Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

Figure B.5 Monthly posterior means of RP of Gender groups versus observed response rates (RR) of Gender groups.



Note: Month 2014-01 to 2018-12 for a model fit and Month 2019-01 to 2019-10 for RP predictions.

Appendix C

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

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

and for second order random walk is

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

The DIC is defined as

$$\mathrm{DIC} = -2\left(\log p\left(r \mid n, E_{\mathrm{post}}\,\theta\right) - p_{\mathrm{DIC}}\right),$$

$$p_{\text{DIC}} = 2\left(\log p\left(r \mid n, E_{\text{post}}\,\theta\right) - E_{\text{post}} \log p\left(r \mid n, \theta\right)\right),$$

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

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

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

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

Appendix D

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

$$p\left(\omega_i \mid r, .\right) = \text{PG}\left(\omega_i \mid n_i, \theta_i\right)$$

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

References

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

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

Boonstra, H.J., and van den Brakel, J.A. (2022). Multilevel time-series models for small area estimation at different frequencies and domain levels. Accepted for publication in *Annals of Applied Statistics*. Technical Report 2018-12, https://www.cbs.nl/en-gb/background/2018/50/models-for-estimation-at-various-aggregation-levels, Statistics Netherlands.

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

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

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

Coffey, S., West, B.T., Wagner, J. and Elliott, M.R. (2020). What do you think? Using expert opinion to improve predictions of response propensity under a Bayesian framework. *Methods Data Analyses*, 14(2), 159-194. https://doi.org/10.12758/mda.2020.05.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the US. *Journal of the American statistical Association*, 94(448), 1074-1082.

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Fang, Q., Burger, J., Meijers, R. and van Berkel, K. (2020). The role of time, weather and google trends in understanding and predicting Web survey response. *Survey Research Methods*, 15(1), 1-25. https://doi.org/10.18148/srm/2021.v15i1.7633.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533. https://doi.org/10.1214/06-ba117a.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472. https://doi.org/10.1214/ss/1177011136.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733-760.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*. CRC press.

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

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

Mushkudiani, N., and Schouten, B. (2019). Time-Dependent Survey Design Parameters: Choosing the Length of Historic Survey Data in a Bayesian Analysis, Application to the Dutch Health Survey. Workshop paper for advances in adaptive and responsive survey design.

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

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

Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339-1349. https://doi.org/10.1080/01621459.2013.829001.

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

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley StatsRef: Statistics Reference Online, 1-8.

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

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

Rue, H., and Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications. CRC press.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*. CRC press.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.R. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64, 583-616. https://doi.org/10.1111/1467-9868.00353.

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

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594. http://jmlr.org/papers/v11/watanabe10a.html.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

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

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

Wu, S., Schouten, B., Meijers, R. and Moerbeek, M. (2022). Data collection expert prior elicitation in survey design: Two case studies. *Journal of Official Statistics*, 38(2), 637-662. https://sciendo.com/es/article/10.2478/jos-2022-0028.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.