

INTEGRATING PROBABILITY AND NONPROBABILITY SAMPLES FOR SURVEY INFERENCE

ARKADIUSZ WIŚNIEWSKI*

JOSEPH W. SAKSHAUG

DIEGO ANDRES PEREZ RUIZ

ANNELIES G. BLOM

Survey data collection costs have risen to a point where many survey researchers and polling companies are abandoning large, expensive probability-based samples in favor of less expensive nonprobability samples. The empirical literature suggests this strategy may be suboptimal for multiple reasons, among them that probability samples tend to outperform nonprobability samples on accuracy when assessed against population benchmarks. However, nonprobability samples are often preferred due to convenience and costs. Instead of forgoing probability sampling entirely, we propose a method of combining both probability and nonprobability samples in a way that exploits their strengths to overcome their weaknesses within a Bayesian inferential framework. By using simulated data, we

ARKADIUSZ WIŚNIEWSKI is with the Department of Social Statistics, University of Manchester, Humanities Bridgeford Street, Oxford Road, Manchester M13 9PL, UK. JOSEPH W. SAKSHAUG is with the Department of Statistical Methods Research, Institute for Employment Research, Regensburger Strasse 104, Nuremberg 90478; Department of Statistics, Ludwig Strasse 33, Ludwig Maximilian University of Munich, Munich 80539; and the School of Social Sciences, University of Mannheim, B6 30-32, 68131 Mannheim, Germany. DIEGO ANDRES PEREZ RUIZ is with the Department of Mathematics, University of Manchester, Alan Turing Building, Oxford Road, Manchester M13 9PL, UK. ANNELIES G. BLOM is with the Collaborative Research Center 884 “Political Economy of Reforms” and the School of Social Sciences, University of Mannheim, B6 30-32, Mannheim 68131, Germany. The authors gratefully acknowledge the financial support of the British Academy / Leverhulme Small Research Grant No. SG171279, the German Institute for Employment Research (IAB), and the German Research Foundation (Grant No. SFB884-Z1 and SFB884-A8). The authors thank Daniela Ackermann-Piek, Carina Cornesse, and Susanne Helmschrott for collecting, granting access to, and giving support with analysing the GIP and non-probability survey data used in this project. We also thank two anonymous reviewers and the Associate Editor who helped to improve the manuscript.

*Address correspondence to Arkadiusz Wiśniowski, Department of Social Statistics, University of Manchester, Humanities Bridgeford Street, Oxford Road, Manchester M13 9PL, UK; E-mail: a.wisniewski@manchester.ac.uk.

evaluate supplementing inferences based on small probability samples with prior distributions derived from nonprobability data. We demonstrate that informative priors based on nonprobability data can lead to reductions in variances and mean squared errors for linear model coefficients. The method is also illustrated with actual probability and nonprobability survey data. A discussion of these findings, their implications for survey practice, and possible research extensions are provided in conclusion.

KEYWORDS: Bayesian inference; Data integration; Online panels; Quota sampling; Web surveys.

1. INTRODUCTION

For more than a decade, the survey research industry has witnessed an increasing competition between two distinct sampling paradigms: probability and nonprobability sampling. Probability sampling is characterized by the process of drawing samples from a population using random selection, with every population element having a known (or knowable) nonzero inclusion probability. In contrast, nonprobability sampling involves some form of arbitrary selection of elements into the sample for which inclusion probabilities are unknowable (and possibly zero for some population elements). Both paradigms have strengths and weaknesses. The primary appeal of probability sampling is its theoretical basis in design-based inference, which permits unbiased estimation of the population mean along with measurable sampling error. However, in practice, unbiased estimation is not assured as response rates in probability surveys can be quite low. Another challenge of probability sampling is the need for large sample sizes for robust estimation, which can be problematic for survey organizations working with small- to medium-sized budgets.

A key advantage of nonprobability sampling, relative to probability sampling, is costs. Nonprobability samples can be drawn and fielded in a number of relatively inexpensive ways. The most common way is through volunteer web panels where survey firms entice large numbers of volunteers to take part in periodic surveys in exchange for money or gifts (Callegaro, Baker, Bethlehem, Göritz, Krosnick et al. 2014). However, nonprobability sampling has limitations. For instance, the lack of an underlying mathematical theory akin to probability sampling is problematic with respect to achieving accuracy and measuring uncertainty (sampling error) for estimates derived from nonprobability samples (Baker, Brick, Bates, Battaglia, Couper et al. 2013). Most benchmarking studies show that nonprobability samples tend to be less accurate than probability samples for descriptive population estimates (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Yeager, Krosnick, Chang, Javitz, Levendusky et al. 2011; Blom, Ackermann-Piek, Helmschrott, Comesse, and Sakshaug 2017; Dutwin and Buskirk 2017; MacInnis, Krosnick, Ho, and Cho 2018; Pennay, Neiger, Lavrakas, and Borg 2018). Multivariate

estimates (e.g., regression coefficients), on the other hand, tend to be less susceptible to discrepancies between probability and nonprobability samples (Ansolabehere and Rivers 2013; Pasek 2016).

A variety of methods have been proposed to improve the accuracy of nonprobability sample survey estimates, such as quota sampling, sample matching, and weighting techniques. These methods attempt to adjust the composition of the nonprobability sample to that of a reference probability sample or population figures based on a set of adjustment variables (Lee 2006; Rivers 2007; Lee and Valliant 2009; Rivers and Bailey 2009; Valliant and Dever 2011; Ansolabehere and Rivers 2013). These methods generally assume that the adjustment variables—which often comprise demographic characteristics—explain the selection mechanism that led to inclusion in the nonprobability sample. This can be a strong assumption if the target variable of interest subject to selection bias is not strongly related to the adjustment variables. Moreover, these methods do not solve the problem of quantifying uncertainty in nonprobability estimates, which is why measures of variability are infrequently reported alongside nonprobability survey estimates.

Because neither sampling paradigm is a panacea, efforts have been undertaken to combine both probability and nonprobability samples to produce a single inference that compensates for the limitations of each standalone paradigm. Elliott and Haviland (2007) evaluate a composite estimator to supplement a standard probability sample with a nonprobability sample. They show that the estimator, based on a linear combination of both sample types and a bias function, can produce estimates with a smaller mean squared error (MSE) relative to a probability-only sample. DiSogra, Cobb, Chan, and Dennis (2012), with further enhancements by Fahimi, Barlas, Gross, and Thomas (2014), propose a method they coin “blended calibration,” where a probability sample weighted to known population totals is combined with an unweighted nonprobability sample, and the combined sample is calibrated to differentiator variables in the probability-only sample. The authors present evidence that this method produces estimates with smaller bias and MSE relative to more traditional sample adjustment methods. Elliott (2009) (see also Elliott and Valliant 2017) proposes a pseudo-design-based estimation procedure that uses a probability sample to estimate pseudo-inclusion probabilities for elements of a nonprobability sample. Both samples are then combined to derive estimates that are shown to have improved accuracy and smaller MSE compared with estimates derived from a probability-only sample.

A limitation of these studies is the necessity of a large probability sample to produce robust calibration weights or pseudo-inclusion probabilities. Given the high costs of probability samples, it is unlikely that survey firms can always afford to field a large probability sample in parallel with a nonprobability sample. A more practical scenario—one we consider in the present investigation—is to field a small probability sample survey and integrate it with a parallel nonprobability sample survey to improve the efficiency (i.e., reduce the variance)

of the small sample estimates. Such a strategy is commonly used in small area estimation where relatively inexpensive auxiliary information collected from external data sources is used to improve the efficiency of estimates derived from small samples for specified domains (Rao 2003).

In this article, we consider a Bayesian approach for integrating a relatively small, presumably unbiased, probability sample with a parallel, relatively less expensive, but possibly biased, nonprobability sample. The Bayesian framework offers several advantages in this context. First, the Bayesian framework offers a natural apparatus for integrating data from different sources with varying levels of quality. Second, the Bayesian framework provides an intuitive structure that incentivizes the higher-quality data (in our case, the probability sample survey data) by giving decreasing weight to the lower-quality “prior information” (in our case, the nonprobability sample data) as the number of high-quality (i.e., probability sample) observations increases. Further, the Bayesian framework is capable of quantifying the uncertainty in estimation—a feature which is currently lacking for estimates based on nonprobability samples. A related idea of integrating probability and nonprobability samples is also explored in Sakshaug, Wiśniowski, Perez Ruiz, and Blom (2019) who describe a simulation-based approach rather than the simple and direct analytical derivations presented in this article.

The aims of this research are threefold. First, we evaluate the extent to which informative nonprobability-based prior information can reduce the variability of regression estimates derived from small- and medium-sized probability samples. Further, we evaluate the trade-off between variance and bias, particularly when the nonprobability samples are subject to large biases. Lastly, we evaluate the MSE of estimates for different sample sizes and bias parameters to determine whether the method is likely to be practically useful from an error perspective. The evaluation is performed using simulation studies and a real-world application involving a probability sample web survey and eight nonprobability sample web surveys conducted in parallel by different survey vendors using the same questionnaire.

The remainder of the article is organized as follows: In section 2, we describe the methodology and notation. In section 3, we describe the setup and present results of the simulation assessment. In section 4, the data sources are described and results of the application are presented. Final conclusions and limitations of the method are discussed in section 5.

2. METHODOLOGY AND MODELING APPROACH

We consider the situation in which the researcher’s primary interest is to produce estimates of effect sizes for covariates specified in a model for predicting a certain outcome variable. For that purpose, we employ a linear regression model in which we estimate the coefficients. Let us denote the $n \times 1$

vector of the response variable by $\mathbf{y} = (y_1, \dots, y_n)^T$ and an $n \times p$ design matrix $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ containing the fixed and nonstochastic covariates that are of substantive interest. We assume that

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \tau^{-1}I_n), \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a column vector of length p , τ is a precision (inverse variance, $\tau = 1/\sigma^2$) parameter, and I_n is the $n \times n$ identity matrix. This specifies the likelihood $p(\mathbf{y}|\boldsymbol{\beta}, \tau, X)$ for the outcome variable.

For a fully Bayesian model specification, we require a prior distribution for model parameters $\boldsymbol{\beta}$ and τ in (1) to produce a marginal posterior distribution for the coefficient of interest, which might be one (or many) of the elements belonging to vector $\boldsymbol{\beta}$. The posterior, $p(\boldsymbol{\beta}, \tau|\mathbf{y}, X)$, is obtained by applying Bayes theorem (Bayes 1763):

$$p(\boldsymbol{\beta}, \tau|\mathbf{y}, X) \propto p(\mathbf{y}|\boldsymbol{\beta}, \tau, X) \times p(\boldsymbol{\beta}, \tau), \quad (2)$$

where $p(\boldsymbol{\beta}, \tau) = p(\boldsymbol{\beta}|\tau)p(\tau)$ is the prior distribution for the model parameters. This specification is scale-dependent because the prior for $\boldsymbol{\beta}$ coefficients depends on the scale of the outcome variable as measured by τ . We employ a conjugate prior specification, that is, a probability distribution that leads to the posterior distribution belonging to the same distributional family as the prior (Carlin and Louis 2008).

The conjugate prior for $p(\tau)$ is a gamma distribution with shape parameter a and rate b :

$$p(\tau) \propto \tau^{a-1} e^{-b\tau}. \quad (3)$$

For $a = b = 0$, this distribution reduces to Jeffrey's noninformative invariant prior (Jeffreys 1946; Zellner 1971), which can be interpreted as an improper uniform distribution for $\log \sigma$. This noninformative prior for precision is used throughout this article.

The conjugate prior for the regression coefficients vector $\boldsymbol{\beta}$ is a multivariate normal distribution:

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_\beta, \tau^{-1}k_0V), \quad (4)$$

where subscript p denotes the length of the vector $\boldsymbol{\beta}$, k_0 is a scalar, and V is a $p \times p$ matrix of hyperparameters. This specification is an extension of the version presented by Ntzoufras (2011) as it allows a separate specification of the matrix and scalar scaling factors for the prior variance of the regression coefficients.

Then, the joint posterior distribution is of the normal-gamma form (Ntzoufras 2011, p. 11). Our interest is in the marginal distribution of $\boldsymbol{\beta}$, which is a multivariate noncentred t distribution with expectation

$$E(\boldsymbol{\beta}|\mathbf{y}) = W\hat{\boldsymbol{\beta}} + (I_p - W)\boldsymbol{\mu}_\beta, \quad (5)$$

$$W = [X^T X + (k_0 V)^{-1}]^{-1} X^T X, \quad (6)$$

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}, \quad (7)$$

and variance

$$\text{Var}(\boldsymbol{\beta}|\mathbf{y}) = [X^T X + (k_0 V)^{-1}]^{-1} (SS + 2b)(n + 2a - 2)^{-1}, \quad (8)$$

$$SS = \text{RSS} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta)^T [(X^T X)^{-1} + k_0 V]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\mu}_\beta), \quad (9)$$

$$\text{RSS} = (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T (\mathbf{y} - X\hat{\boldsymbol{\beta}}), \quad (10)$$

where I_p is the $p \times p$ identity matrix. Equation (7) is a maximum likelihood estimator (MLE) for the linear regression coefficients. Matrix W in (6) can be viewed as a “weight” assigned to a data-driven MLE and $I_p - W$ a weight for the prior. If the scaling factor, $k_0 V$, is very large (implying large prior variance for $\boldsymbol{\beta}$), the weight assigned to the prior mean $\boldsymbol{\mu}_\beta$ will be very small. Alternatively, if the prior for $\boldsymbol{\beta}$ is relatively tight, more weight is assigned to the prior mean. The RSS in (10) is the residual sum of squares in the classical regression analysis, whereas SS is the sum of RSS and a measure of the distance between the MLE for $\boldsymbol{\beta}$ and the prior mean (Ntzoufras 2011, p. 13).

The marginal posterior for the precision is a gamma distribution with shape parameter $n/2 + a$ and rate $SS/2 + b$. With a noninformative Jeffrey’s prior, i.e., $a = b = 0$ in (3), the expected value of the posterior is equivalent to the MLE for precision.

2.1 Specification of the Prior Distributions

We consider five specifications of the prior in (4) for the regression coefficients. The first one is a reference specification, hereafter referred to as “noninformative,” in which we “let the (probability) data speak for themselves”. That is, we assume

$$\begin{aligned} \boldsymbol{\mu}_\beta &= \mathbf{0}, \\ k_0 &= n, \\ V &= I_p. \end{aligned} \quad (11)$$

This specification leads to a posterior (2) with the expectation of $\boldsymbol{\beta}$ in (5) being

$$E(\boldsymbol{\beta}|\mathbf{y}) = W\hat{\boldsymbol{\beta}}, \quad (12)$$

with $W = (X^T X + (1/n)I_p)^{-1} X^T X$. This prior allows a probability sample of size n to dominate the inference. With n becoming larger, the posterior converges to the MLE. Interestingly, the expected value of the posterior

(equation 12) is a *ridge* estimator (Hoerl and Kennard 1970; Amemiya 1985). By using (8) and (9–10), the posterior variance of β is

$$\text{Var}(\beta|\mathbf{y}) = \frac{1}{n-2} \left(X^T X + \frac{1}{n} I_p \right)^{-1} \{RSS + \hat{\beta}^T [(X^T X)^{-1} + n I_p]^{-1} \hat{\beta}\}, \quad (13)$$

which, for sufficiently large n , yields values equivalent to the MLE variance.

The second prior specification, denoted as “conjugate,” is informative in the sense that it utilizes a nonprobability sample to inform the posterior. Here, we assume

$$\begin{aligned} \mu_\beta &= \hat{\beta}_{NP}, \\ k_0 &= \frac{1}{\log n_{NP}} \mathbb{1}(p^{\text{Hotelling}} < 0.05) + \frac{1}{n_{NP}} \mathbb{1}(p^{\text{Hotelling}} \geq 0.05), \\ V &= I_p, \end{aligned} \quad (14)$$

where $\hat{\beta}_{NP}$ is an MLE of the coefficients in a regression model based on a nonprobability sample of size n_{NP} , and $\mathbb{1}(w)$ is an indicator function taking the value one if w is true and zero otherwise. Additionally, $p^{\text{Hotelling}}$ is a p -value from a Hotelling T^2 test for equality of two vectors (Anderson 1984; Khuri 2003, p. 48). The Hotelling T^2 test is a multivariate generalization of the Student t test. In our application of the test, the null hypothesis is $H_0 : \hat{\beta}_{NP} = \hat{\beta}$, and we assess whether the two vectors of coefficients are equal. The implicit assumption is that the results based on the probability sample are unbiased with respect to the target population. Hence, if the difference between all coefficients based on the nonprobability sample, $\hat{\beta}_{NP}$, is statistically significant at the 5 percent level ($p^{\text{Hotelling}} < 0.05$ and $\hat{\beta}_{NP}$ is “biased” with respect to the probability sample-based $\hat{\beta}$), then the prior variance for β is relatively large: the scaling factor k_0 is $1/\log n_{NP}$, which is larger than the $1/n_{NP}$ proposed for the cases when the difference between probability and nonprobability coefficients is not statistically significant ($p^{\text{Hotelling}} \geq 0.05$).

The rationale for using one of the two scaling factors is that we prefer to avoid any potential biases that might be present in the nonprobability samples. Again, the bias considered here is assessed in relation to the probability sample. If the nonprobability MLE coefficients show significant differences in comparison with the probability ones, then the prior variance will not tend to extremely small values with an increase in nonprobability sample size. On the other hand, when the MLE coefficients from the nonprobability sample are unbiased compared with the probability sample, then as the nonprobability sample size increases, the smaller the prior variance becomes, which, in turn, allows further shrinkage of the posterior variance of β , comparing with (13). Our approach to assessing bias in coefficients is dictated by typical applications, in which practitioners rarely have information about the gold standard

against which they could assess bias. However, if such information is available, then the procedure can easily be modified to allow setting k_0 depending on the true bias of the nonprobability coefficients and correcting for it in the prior mean μ_β .

The third specification of the prior is a Zellner's g -prior (referred to as "Zellner;" see Zellner 1986; Liang, Paulo, Molina, Clyde, and Berger 2008) where the Fisher's information matrix (inverse covariance matrix of covariates) is used to rescale the prior variance. This permits accounting for the scale and correlation structure of the covariates in the posterior, which might be relevant if strong multicollinearity is present and various scales are used for different variables. The prior is specified with hyperparameters

$$\begin{aligned}\mu_\beta &= \hat{\beta}_{NP}, \\ k_0 &= n_{NP}^2 \mathbb{1}(p^{\text{Hotelling}} < 0.05) + \mathbb{1}(p^{\text{Hotelling}} \geq 0.05), \\ V &= (X_{NP}^T X_{NP})^{-1},\end{aligned}\tag{15}$$

where X_{NP} is a matrix of covariates from the nonprobability sample, and k_0 is a scaling hyperparameter g as in the original specification. Zellner's g -prior is a convenient and widely used prior in the linear regression setting, mainly due to its simplicity, computational efficiency, and understandable interpretation (Liang et al. 2008, p. 411). It requires selecting only one hyperparameter, k_0 , as a measure of prior dispersion of the model parameters. Various recommendations for this choice have been proposed, including k_0 being set to the sample size, squared number of model dimensions, or values obtained from using empirical Bayes methods (Liang et al. 2008, p. 413).

To create the prior based on nonprobability data, we propose using the scaling factors V and k_0 that depend on the potential bias in the MLE of the coefficients based on nonprobability sample data, which is assessed against the MLE based on the probability sample data. The two scaling factors are a function of the nonprobability design matrix and sample size, respectively.

We suggest using the nonprobability-based MLE for the prior mean and the Fisher's information matrix of the nonprobability covariates X_{NP} for V . In the case of bias in the MLE estimates, we recommend using $k_0 = n_{NP}^2$ which implies smaller variance and a more informative prior compared with, for example, probability sample size n as recommended by Kass and Wasserman (1995). When bias is not present in the nonprobability MLE, then the even more informative prior depends only on the Fisher's information matrix from the nonprobability sample and $k_0 = 1$.

The next two informative prior specifications are variations of the previously described conjugate and Zellner's priors. The key part of them is the distance between the MLEs based on the probability and nonprobability data, $\hat{\beta}$ and $\hat{\beta}_{NP}$, respectively, and the standard error of the MLEs based on nonprobability data, $\hat{\sigma}_{\beta, NP}$ (cf. Sakshaug et al. 2019). These distance and standard error

are used to rescale the variance of the prior distribution for the regression coefficients. Thus, the fourth prior specification (referred to as “conjugate-distance”) is specified as

$$\begin{aligned}\boldsymbol{\mu}_\beta &= \hat{\boldsymbol{\beta}}_{NP}, \\ k_0 &= \frac{1}{\log n_{NP}}, \\ V &= \text{diag}\{\max[(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{NP})^2, \hat{\boldsymbol{\sigma}}_{\beta, NP}^2]\},\end{aligned}\tag{16}$$

whereas the fifth proposed specification (referred to as “Zellner-distance”) is an extension of the Zellner prior from (15):

$$\begin{aligned}\boldsymbol{\mu}_\beta &= \hat{\boldsymbol{\beta}}_{NP}, \\ k_0 &= n_{NP}, \\ V &= V_d(X_{NP}^T X_{NP})^{-1} V_d, \\ V_d &= \text{diag}\{\max[\sqrt{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{NP})^2}, \hat{\boldsymbol{\sigma}}_{\beta, NP}]\}.\end{aligned}\tag{17}$$

Here, $\text{diag}(\mathbf{v})$ denotes a diagonal matrix with elements of \mathbf{v} on its diagonal, $\max(\mathbf{a}, \mathbf{b})$ is an element-wise maximum of two vectors \mathbf{a} and \mathbf{b} , and $\hat{\boldsymbol{\sigma}}_{\beta, NP}$ is a vector of ML standard errors of the regression coefficients estimated using nonprobability data. Since nonprobability samples are usually larger than the probability ones, an implicit assumption here is that $\hat{\boldsymbol{\sigma}}_{\beta, NP}$ are much smaller than the respective standard errors based on probability data. This specification implies that the prior variance does not depend on the binary outcome of the Hotelling T^2 test and allows the prior variance for each coefficient to be rescaled by its own factor that depends on the distance between the coefficients or the standard error of the coefficient based on nonprobability data—which-ever is larger for a given coefficient.

The squared distance introduced to the prior variance permits shrinking the prior distribution in cases where the difference is relatively small and allows for larger variability if discrepancies between probability and nonprobability data arise. Further, the probability-based MLE is used only in relation to the nonprobability-based MLE to construct the prior for coefficient variability, rather than central tendency. Hence, the shrinkage in the posterior estimator depends on the relative distance between probability and nonprobability samples (measured by a difference between MLEs) or nonprobability MLE standard error, rather than a probability sample alone.

The role of the nonprobability-based MLE standard errors is to serve as “safety vents” should the distance between the MLEs be extremely small. In such cases, it prevents the prior from being overly tight, which may cause numerical matrix inversion problems. If one decides that there is no risk of

extremely small differences arising between $\hat{\beta}$ and $\hat{\beta}_{NP}$, then the prior can easily be transformed to rely on the distance between the two coefficients only.

3. MODEL ASSESSMENT USING SIMULATED DATA

To assess the proposed methods, we generate both probability and nonprobability data from a simple linear regression model with assumed known coefficients. We also introduce bias in one of the coefficients in the simulated nonprobability samples. We then estimate the parameters using these generated probability samples and, as described in the previous section, priors constructed based on nonprobability samples as specified in (11) and (14–17). Lastly, we assess how bias introduced in the nonprobability data affects the posteriors for the coefficients.

The simulated data were created by assuming various sets of coefficients and using data generating models: $x_1 \sim N(0, \sigma_1)$, $x_2 \sim N(5, \sigma_2)$, $y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma_y)$. We also assume that x_1 and x_2 are correlated with correlation $\rho = 0.1$. Here, we present two scenarios. In the first one (denoted as scenario A), we assume $\sigma_1 = \sigma_2 = \sigma_y = 1$, $\beta_0 = 1$, $\beta_1 = 0.5$ and $\beta_2 = 0.1$. This scenario reflects a situation in which the effect of the covariate of interest, say x_2 , on the outcome is relatively small compared with the other covariate. In scenario B, we assume $\sigma_1 = 4$, $\sigma_2 = 0.5$, $\sigma_y = 2$, $\beta_0 = 1$, $\beta_1 = 0.5$ and $\beta_2 = 2$, which reflects a relatively strong effect of x_2 .

Next, by using both scenarios, we generate one hundred sets of probability data with sample sizes $n \in \{50, 100, 150, 200, 300, 500, 750, 1,000\}$. For each of the probability sets, we generate fifty nonprobability samples of sizes $n_{NP} \in \{10^3, 10^4, 5 \cdot 10^4\}$ by using the same models but with bias in coefficient β_2 induced by multiplying it by a factor $f \in \{0.5, 1, 1.5, 2, 3\}$. This bias in the coefficient leads to bias in the generated response; factor $f = 1$ implies the nonprobability sample is unbiased. These assumptions about bias are rather extreme, implying that the nonprobability samples are severely biased, as compared with the probability data (e.g., in scenario B, the expectation of the response variable under no bias is $E(y) = 11$, but when $f = 3$, this almost triples to $E(y) = 31$).

We evaluate the performance of each of the priors by calculating the mean-squared error for the posterior characteristics of the parameters β . For a given posterior distribution, we define the MSE as

$$MSE(p(\beta, \tau|\mathbf{y}, X)) = \mathbb{E}[(p(\hat{\beta}, \tau|\mathbf{y}, X) - \beta^*)^2], \quad (18)$$

where β^* is the vector of true parameters used to generate data. The MSE can be decomposed into variance and squared bias:

$$MSE(p(\beta, \tau|\mathbf{y}, X)) = Bias^2(p(\beta, \tau|\mathbf{y}, X)) + Var(p(\beta, \tau|\mathbf{y}, X)). \quad (19)$$

We calculate the bias as the difference between the mean of the posterior means for all iterations, $\bar{p}(\boldsymbol{\beta}, \tau | \mathbf{y}, X)$ and the true coefficients:

$$\text{Bias}(p(\boldsymbol{\beta}, \tau | \mathbf{y}, X)) = \bar{p}(\boldsymbol{\beta}, \tau | \mathbf{y}, X) - \boldsymbol{\beta}^*, \quad (20)$$

with variance being the average posterior variance of the coefficient across all iterations.

In [figure 1](#), we present the bias as measured by the difference between the posterior mean and the true coefficient, posterior variance, and mean-squared error (i.e., the sum of squared bias and posterior variance) of the coefficients β_0 , β_1 , and β_2 in scenario A. All methods except for Zellner lead to reductions in variance even with a large level of bias induced in β_2 (i.e., $f = 3$). Further, substantial reductions in MSE compared with noninformative results are observed when using conjugate (for β_0 and β_2) and conjugate-distance (β_0 , β_1 , and β_2) priors, which stems from the low posterior bias produced by these methods. For the Zellner method, however, gains in efficiency are minimal, and for Zellner-distance, they are offset by a larger bias observed even for large probability sample sizes, except for the cases with relatively low levels of induced bias (i.e., $f \in (0.5, 1.5)$).

In [figure 2](#), we present the analogous results for the scenario B simulation. Here, we observe that gains in efficiency (reduction in MSE) are obtained when using the conjugate-distance method, even for large (in terms of absolute value) values of induced bias (e.g., $f = 3$). Both conjugate and Zellner-distance methods lead to large bias in β_0 and β_2 which, for Zellner-distance, cannot be offset by relatively small gains in efficiency if induced bias is present ($f \neq 1$). The Zellner method does not lead to large bias, but gains in efficiency are observed only when no induced bias is present ($f = 1$). This is also indicated by the increasing proportions of the Hotelling test rejecting the null hypothesis stating the probability and nonprobability-based coefficients are the same (section 2.1) in scenario B compared with scenario A (see [figure A.1](#) of the appendix). In [figure A.2](#) of the appendix, we also present the coefficients of variability for the two simulations.

Overall, the conjugate and conjugate-distance methods seem to yield gains in efficiency and produce the posterior estimates least sensitive to bias induced in nonprobability samples, especially for coefficients directly affected by this bias (β_0 and β_2). The Zellner and Zellner-distance methods are much more susceptible to potential bias in the nonprobability samples. Nevertheless, one should keep in mind that when the bias implied by any informative prior distribution is very large, the posterior will be overwhelmed by it, especially when nonprobability samples are much larger than probability samples. Also, if the posterior distribution for a coefficient based on the probability data alone is tight (or when using MLEs, the standard errors are small), then even small amounts of bias in the prior based on nonprobability sample may lead to reduced rather than improved efficiency. Thus, the proposed methods are recommended when precision of the

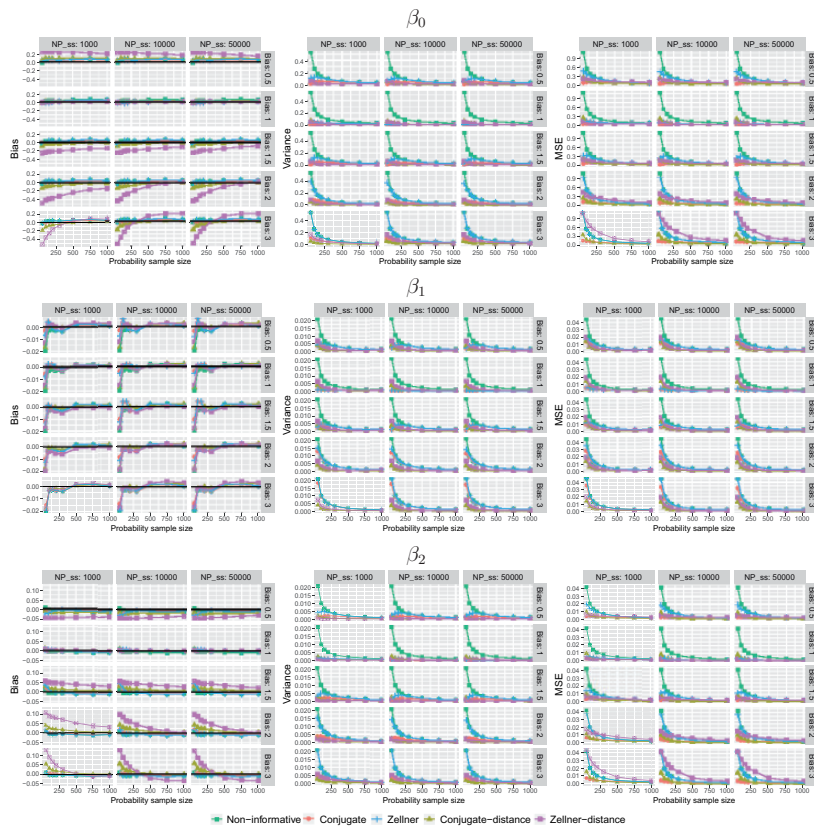


Figure 1. Simulation Scenario A. Bias, variance, and mean squared error (MSE) for β_0 , β_1 , and β_2 coefficients. NP_ss denotes nonprobability sample size (n_{NP}).

probability-based estimates is relatively low and it is desirable to increase it. Typically, such situations occur when small probability samples are available.

4. APPLICATION TO THE GERMAN INTERNET PANEL

4.1 Probability and Nonprobability Data

Next, we evaluate the method and four informative prior specifications using actual survey data. The probability survey data come from the German Internet Panel (GIP). The GIP is an ongoing, probability-based longitudinal survey designed to represent the population of Germany 16–75 years of age. The survey is funded by the German Research Foundation (DFG) as part of the Collaborative Research Center 884 “Political Economy of Reforms” based at

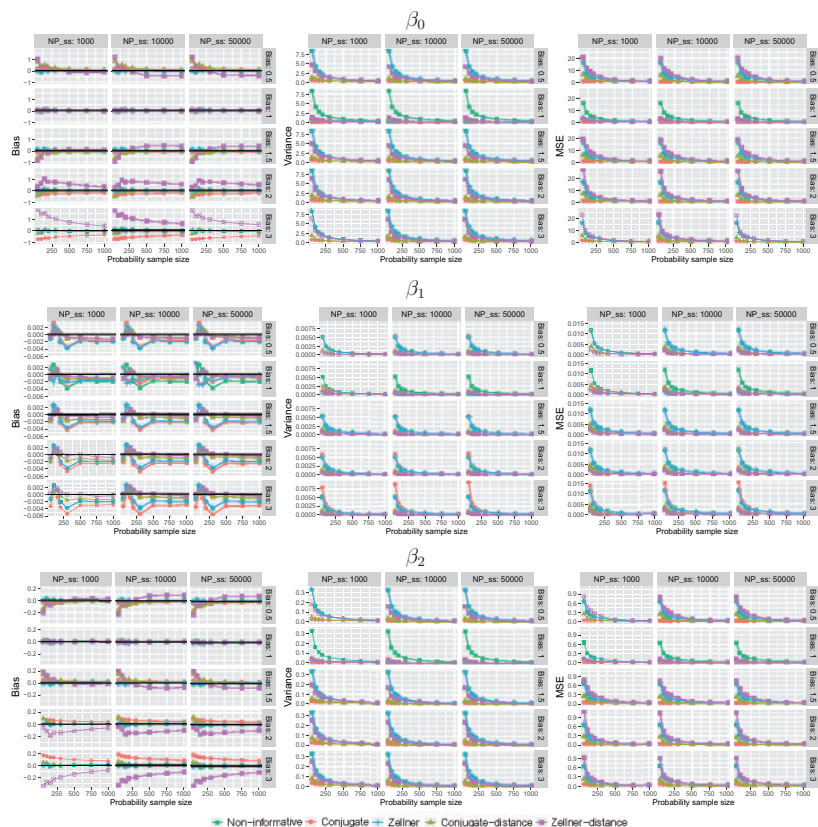


Figure 2. Simulation Scenario B. Bias, variance, and mean squared error (MSE) for β_0 , β_1 , and β_2 coefficients. NP_ss denotes nonprobability sample size (n_{NP}).

the University of Mannheim. Sample members are selected through a multi-stage stratified area probability sample. At the first stage, geographic districts are selected from a database of 52,947 districts in Germany. A random sample of listed households are then drawn, and all age-eligible members of the sampled households are invited to join the panel (Blom, Gathmann, and Krieger 2015). The GIP is designed to cover both the online and offline populations and provides internet service and/or internet-enabled devices to facilitate participation for the offline population (Blom, Herzing, Cornesse, Sakshaug, Krieger et al. 2016). The first GIP recruitment process occurred in 2012, achieving an 18.5 percent recruitment rate (based on Response Rate 2; AAPOR 2016). A second recruitment effort occurred in 2014 and achieved a recruitment rate of 20.5 percent (AAPOR Response Rate 2). Panelists receive a survey invitation every two months to complete a web survey of approximately 20–25 minutes. The interview covers a range of social and political topics. The questionnaire

Table 1. List of Nonprobability Surveys

Survey	N	Quota variables	Total cost (€)
1	1,012	Age, gender, region, education	0 (pro bono)
2	1,000	Age, gender, region	5,392.97
3	999	Age, gender, region	5,618.57
4	1,000	Age, region	7,061.11
5	994	Age, gender, region	7,411.00
6	1,002	Age, gender, region, education	7,636.22
7	1,000	Age, gender, region	8,380.46
8	1,038	Age, gender, region	10,676.44

module we utilize was administered in March 2015. During this month, completed interviews were obtained from 3,426 out of 4,989 (or 68.7 percent of) panelists. After removing responses with partial missing information and adjusting the age range of the GIP to match the age range of the nonprobability samples (discussed below), we utilize a sample size of 2,681 cases.

As part of a separate methodological project studying the accuracy of non-probability surveys (Blom et al. 2017), the GIP team sponsored several non-probability web survey data collections that were conducted in parallel with the GIP in March 2015. The nonprobability web surveys were carried out by different commercial vendors in response to a call for bids. The key stipulation of the call was that each vendor should recruit a sample of approximately one thousand respondents that is representative of the general population 18–70 years of age living in Germany. The call for bids did not explicitly specify how representativeness should be achieved. A total of seventeen bids were received, of which seven were selected based on technical and budgetary criteria. An eighth commercial vendor voluntarily agreed to perform the data collection without remuneration after learning about the study aims. Additional information about the eight nonprobability surveys can be found in table 1, including the quota sampling variables used and costs. For confidentiality reasons, we do not disclose the names of the survey vendors and identify them simply as survey one, survey two, etc., ordered from least expensive (pro bono) to most expensive (10,676.44 EUR).

In this application, the outcome variable is body mass index (BMI), a commonly studied variable and disease risk factor. A histogram of BMI, as derived from the GIP, is provided in appendix B. The variable has been constructed by using respondents’ self-reported height and weight. Body mass index follows an approximate normal distribution with a slight right skew. In practice, height and weight (and hence, BMI) may not be properly recorded due to measurement error (e.g., respondents adding to their height or reducing their weight). We neglect this limitation in the application. Mean BMI for the GIP and eight nonprobability surveys are shown in figure 3. We observe that all of the nonprobability surveys

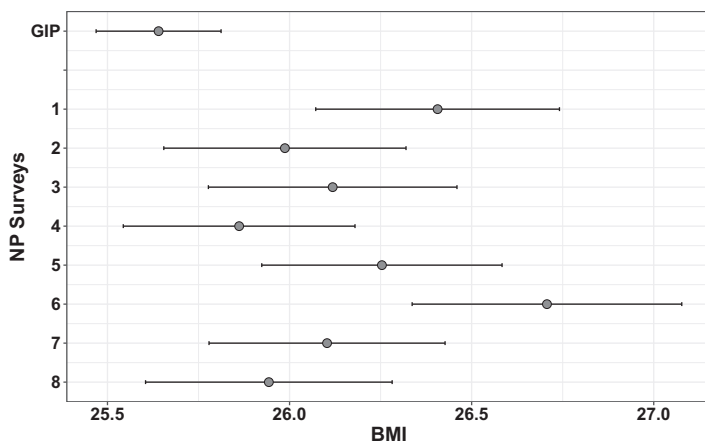


Figure 3. Means and 95 Percent Confidence Intervals for Body Mass Index in the German Internet Panel and Nonprobability Surveys.

produce a larger mean BMI than the probability-based GIP survey. The most severe bias is apparent in nonprobability surveys six and one, respectively.¹

Following from the simulation assessment in Section 3, our interest lies in estimating coefficients from a linear regression model using BMI as the outcome variable. The model is fitted using the following covariates: age (continuous), sex (1 = female, 0 = male), self-reported health status (1 = fair/poor/very poor, 0 = very good/good), marital status (1 = single, 0 = other), education (mutually exclusive binary variables, that is, 1 = Hauptschulabschluss [basic secondary level] and 0 otherwise, 1 = Mittlere reife [extensive secondary level] and 0 otherwise, 1 = Fachhochschule [vocational school] and 0 otherwise, with university degree being the baseline category), and regularly employed full or part time (1 = no, 0 = yes). Each respondent characteristic was measured in both the GIP and nonprobability surveys using the same questionnaire items. In addition, we include a survey weight variable produced by the GIP team that includes a raking adjustment to population benchmarks as a covariate in the model. Incorporating the survey weight variable as a covariate in the analysis of survey data and Bayesian inference has been considered in previous work (Rubin 1985; Pfeffermann, 1993; Skinner 1994). We do not have access to further sample design information (clustering, stratification) for the GIP. Incorporating such information in the modeling approaches is a topic we leave for future work. Both continuous covariates (age and weights) are standardized to put coefficients on a comparable scale.

1. Similar conclusions can be drawn when bootstrapped confidence intervals are produced for the nonprobability surveys.

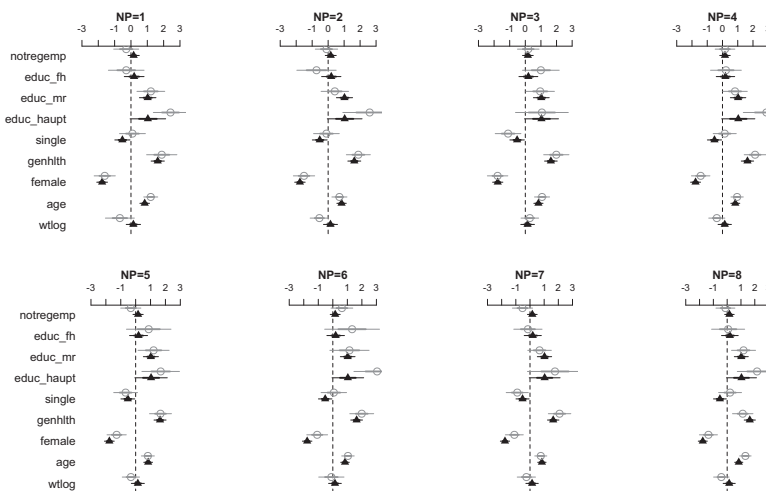


Figure 4. MLE of Regression Coefficients and 95 Percent Confidence Intervals for Body Mass Index in the German Internet Panel (Black) and Eight Nonprobability Surveys (Gray). Abbreviations: notregemp, not in employment; genhlth, self-reported health; educ_fh, vocational school; educ_mr, extensive secondary school; educ_haupt, basic secondary school; single, marital status; female, sex; age, standardized age; wtlog, standardized weights (log-scale).

Maximum likelihood estimators of the regression coefficients of BMI on the covariates, fitted using eight nonprobability surveys, are plotted in [figure 4](#), with the MLE of the coefficients fitted using 2,681 available cases in the GIP data (in gray). The fitted coefficients are similar between the GIP and the eight nonprobability surveys. The largest differences occur for education where the lowest education classification (hauptschulabschluss) has a larger positive effect on BMI in the nonprobability surveys than in the GIP survey. Females are negatively associated with BMI, and this effect is slightly weaker in the nonprobability surveys compared with the GIP. Other variables that tend to have a significant effect on BMI include age (+) and fair-poor general health status (+). As a whole, the coefficients are roughly comparable between surveys, and there is no strong indication of bias in the nonprobability surveys.

4.2 Computation and Evaluation

We compare the performance of the five specifications of the prior (noninformative, conjugate, Zellner, conjugate-distance, and Zellner-distance; see section 2 for details) by estimating posteriors under different probability sample size scenarios and using different nonprobability surveys to construct the prior distributions. We consider the probability sample sizes $n \in \{50, 100, 150, 200\}$,

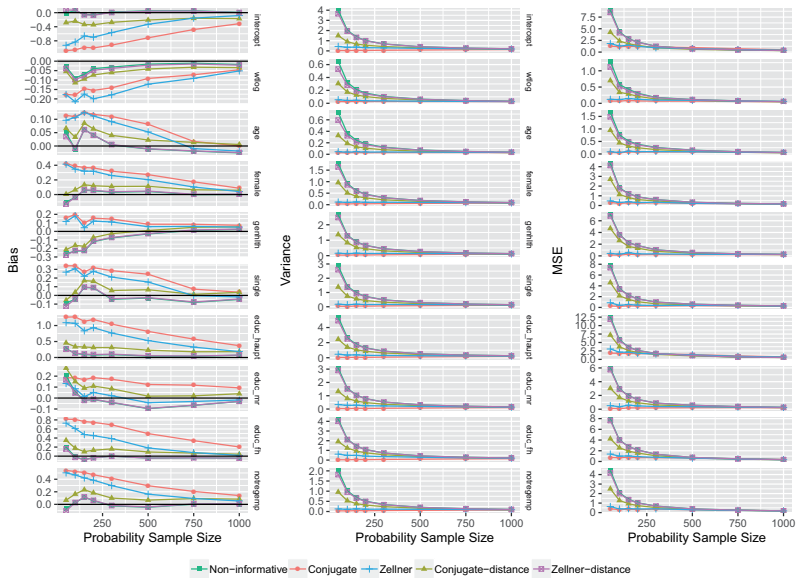


Figure 5. Bias, Variance and MSE for Regression Coefficients (Averaged Over One Hundred Samples) in Bayesian Model of Body Mass Index on Respondent Covariates in the German Internet Panel. Prior distributions constructed using non-probability survey six. Age and weight (wtlog) covariates are standardized.

300, 500, 750, 1,000}. We use these various sample sizes to simulate a situation in which a survey sponsor can only afford a small to medium size probability sample alongside a potentially larger nonprobability sample. These probability sample sizes are drawn at random from the full GIP sample and are constructed cumulatively so that the same cases used in the smaller sample size scenarios are included in the larger sample size scenarios. The size of the nonprobability sample used to construct the informative prior distributions is not manipulated, and the full sample size (roughly one thousand respondents; see table 1) is used.

Next, we evaluate the five prior specifications by calculating the bias, variance, and MSE (sum of the variance and squared bias) for the estimated regression coefficients as in (18–19) described in section 2. To calculate bias in (20), we use a vector of MLEs calculated with the entire GIP sample, which we assume to be unbiased. This strong assumption can be relaxed if more reliable sources of data on a given characteristic (in our case, BMI) are known.

The entire evaluation procedure (including drawing of the probability samples) is repeated one hundred times to produce one hundred sets of variance, bias, and MSE values for each sample size scenario. In the following results section, we report the average of these values across the one hundred iterations for each sample size.

4.3 Results

The results for bias, variance, and MSE of each regression coefficient are presented in [figure 5](#). For brevity, we present here the results only for the case where one of the more biased nonprobability surveys, survey six, is used to construct the prior distributions for the four informative models. The results using each of the other nonprobability surveys can be found in appendix figures C.1 (bias), C.2 (variance), and C.3 (MSE).

The second column in [figure 5](#) shows that conjugate and Zellner prior specifications yield coefficient estimates with considerably smaller variances relative to the estimates derived under the noninformative (reference) prior, particularly for the smaller probability sample sizes. We observe moderate reductions in variance for the conjugate-distance prior and significantly smaller reductions under the Zellner-distance prior. The largest reductions occur for the smallest sample sizes (fifty and one hundred cases). These reductions continue at a decreasing rate until a probability sample size of about five hundred cases is reached; at this point, the variances converge under all five priors. Further, the variances produced under the conjugate and Zellner priors for the smallest probability sample sizes are roughly equivalent to the variances produced under the noninformative prior for the much larger probability sample sizes (starting with five hundred or more cases). The same variance reduction patterns are observed when each of the other seven nonprobability surveys are used to construct the priors (see appendix figure C.2).

Next, we turn to bias. [Figure 5](#) (first column) shows that a few of the coefficients are biased when the informative priors are used. This is particularly true for the variables sex and education, which were previously identified as being biased in the nonprobability surveys (see MLEs in [figure 4](#)). A relatively smaller bias is observed for marital status (single) and health status (genhlth). As expected, the biases are most prominent for the smallest probability sample sizes, where the influence of the informative priors is at their peak. All informative priors have very similar impact on bias, although one can see a slight indication that the conjugate and Zellner priors yield a slightly larger and more persistent bias than the other two priors, but this depends on the coefficient. As with the variance results, the bias induced by the informative priors tends to converge to the noninformative reference model as the probability sample size increases and the influence of the priors weaken.

Lastly, we examine the joint impact of variance and bias in the form of mean squared errors for the regression coefficients. [Figure 5](#) (third column) shows the MSE results, which share a similar pattern to the variance results described previously. That is, the MSEs produced by the model with the informative priors are considerably smaller than the MSEs under the noninformative reference model for the smallest probability sample sizes. The reductions in MSE are smallest for the Zellner-distance prior. This pattern holds even for the biased covariates identified earlier, except for the Zellner-distance prior. Thus,

it is apparent that reductions in variance outweigh the increase in bias under the informative prior models. As with the variance results, the MSE gap between the values yielded by using informative and noninformative priors closes as the probability sample size increases and reaches an equilibrium at about five hundred cases. For nearly all coefficients, the MSEs under the informative priors for the smallest probability sample sizes are approximately equivalent to the MSEs produced under the noninformative (reference) prior for the largest probability sample size ($n = 1,000$).

5. DISCUSSION

In this article, we evaluate a method of integrating relatively small probability samples with nonprobability samples to improve the efficiency (i.e., reduce variability) and reduce the mean squared error for estimated regression coefficients. We exploit a Bayesian framework and consider four prior specifications constructed using nonprobability sample survey data that inform posterior estimations based on small probability sample survey data. This approach may be considered a reversal of the more traditional approach of adjusting a nonprobability sample toward a supplementary probability sample (or other high-quality benchmark source), as we deliberately “borrow strength” from the nonprobability survey data and use these data to influence the probability sample survey estimates.

Using simulations and a real-data application, we show that the nonprobability-based informative priors can yield reductions in variance and, subsequently, MSE for linear regression coefficients estimated from small probability samples. The reductions in variance/MSE yielded such values that were consistent with the variance/MSE values of much larger probability-only samples, suggesting that smaller probability sample sizes coupled with relatively inexpensive nonprobability samples could produce levels of error that are comparable with large and more expensive probability samples.

However, using potentially biased nonprobability survey data to inform estimations has the potential to introduce bias in the final estimations and offset gains in efficiency. This issue was not encountered in the case study where only modest biases were present in the nonprobability surveys and these biases did not affect the efficiency gains. Further, the simulation study showed that large biases can potentially overwhelm the method and produce larger MSEs, depending on the particular prior specification. In particular, the Zellner-distance typically led to this behavior. The conjugate and conjugate-distance methods were reasonably robust to large amounts of bias induced in the simulated nonprobability samples. Nevertheless, one should be cautious in implementing the proposed methods when fitting regression models with covariates that are particularly susceptible to very large biases in nonprobability surveys.

Some additional limitations of the method should be noted. First, the proposed method relies on the assumption that the probability sample is unbiased, particularly when compared with the nonprobability sample. This assumption may not hold in some situations in which one is modeling rare or difficult to measure phenomena (e.g., homelessness, HIV). We acknowledge that even high-quality probability samples can be subject to measurement issues due to nonsampling error and might sometimes provide biased estimates relative to a nonprobability sample designed specifically for measuring these phenomena (e.g., respondent-driven sampling). Second, the method presented here is limited to continuous outcome variables. Extending the method to handle other variable types (e.g., ordinal, nominal, count) is possible and will be considered in future work. Further research extensions include accounting for complex sample design features, model misspecification, and alternative prior specifications. The four informative priors considered here are only a small subset of possible specifications. Alternative specifications that, for instance, can correct for bias in the nonprobability samples could be considered, as well. Further, a strength of the study application was a demonstration of the method using several nonprobability surveys collected by different commercial vendors. This demonstration noted some variability in the results that warrants further investigation. Future work ought to investigate the properties of nonprobability surveys that make them potentially good (or bad) candidates to construct useful prior distributions.

Based on the results of our evaluation, we provide the following advice to practitioners. For practitioners who utilize the probability sampling approach but cannot afford to continue fielding large sample sizes, we view this method as a means to reduce the probability sample size without inducing more variability in the estimates. Indeed, adopting this method and allocating some resources to field a parallel nonprobability data collection, which need not be so large itself (e.g., 1,000 cases), is likely to produce similar measures of variability (and MSE) compared with a large probability-only sample at a reduced cost. The risk of inducing bias by introducing nonprobability data collection is likely to be small as the proposed prior specifications, particularly the conjugate methods, were reasonably robust to varying levels of bias.

For practitioners who defend the nonprobability sampling approach, the primary concern would be the cost of fielding an additional (albeit, small) probability sample. In our view, the benefits of collecting a small parallel probability sample outweigh the added costs for three reasons. First, the empirical literature continues to show that probability samples produce more accurate estimates than nonprobability samples, thus introducing some probability data collection is likely to be perceived as a more scientific and credible approach. Second, the Bayesian framework is a natural system for incorporating multiple data sources with varying levels of quality, is principally structured to incentivize sparse, yet high quality observations in the posterior estimations, and is a commonly accepted method in official statistics for estimating small domains.

Third, collecting a small probability sample need not be so costly if the data are collected via an existing probability-based omnibus survey or relatively inexpensive probability-based mail survey. While it would be important to achieve a high response and minimize the risk of nonresponse bias, this might be more feasible with a small sample than with a large one.

In conclusion, it is interesting to know that probability and nonprobability samples can be integrated in a way that exploits their advantages to compensate for their weaknesses and improve estimation of model parameters. It is further possible that the method can lead to cost savings for a fixed variance (or MSE) if the nonprobability sample units are significantly cheaper to interview than the probability sample units. Another attractive quality of the method is its computational efficiency. It is easily implemented using freely available software such as R and any statistical software that allows Bayesian inference and specification of the prior distributions for the linear regression model. The R code used to obtain results is available in the online [supplementary material](#).

Appendixes

A. Hotelling Test Results for the Simulation Study

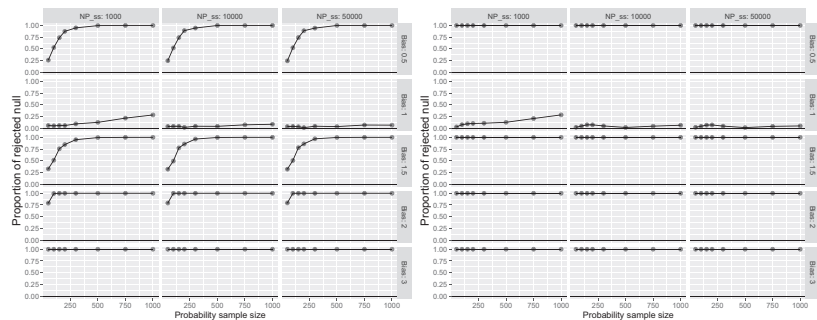


Figure A.1 Proportion of rejected null hypotheses in the simulation scenario A (left panel) and scenario B (right panel).

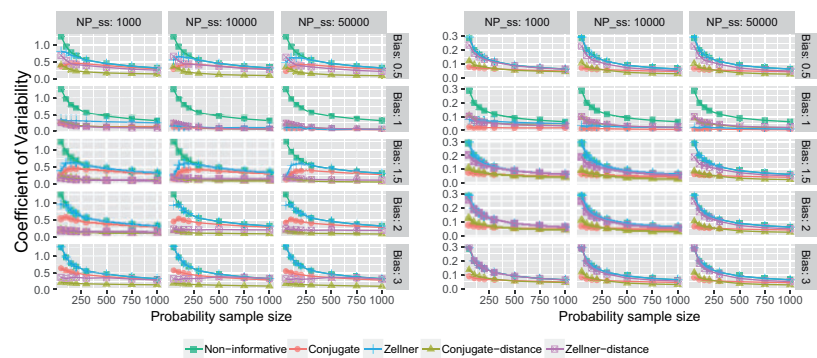


Figure A.2 Coefficients of variability in the simulation scenario A (left panel) and scenario B (right panel). Coefficient of variability is defined as the standard deviation divided by the expectation of the posterior distribution.

B. Distribution: Body Mass Index

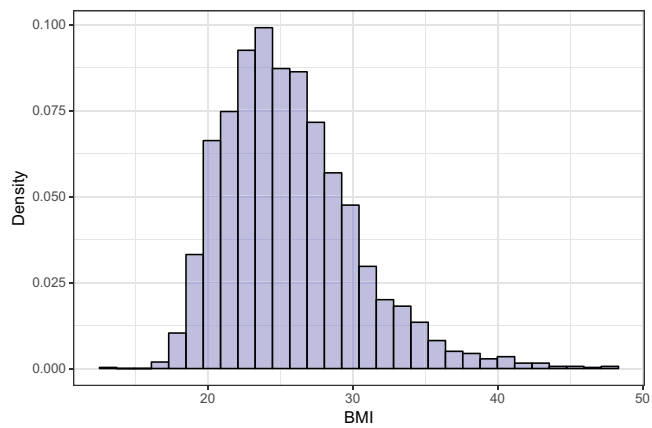


Figure B.1 Histogram of body mass index in the German Internet Panel.

C. Application Using GIP Data

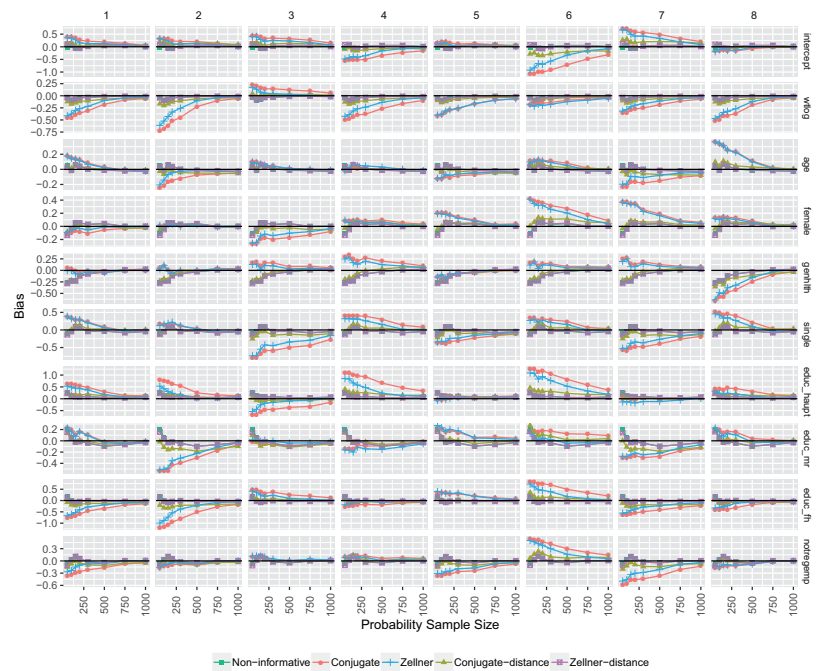


Figure C.1 Bias for regression coefficients (averaged over one hundred samples) in Bayesian model of body mass index on respondent covariates in the German Internet Panel. Prior distributions constructed using one of eight nonprobability surveys (in columns). Age and weights (wtlog) covariates are standardized.

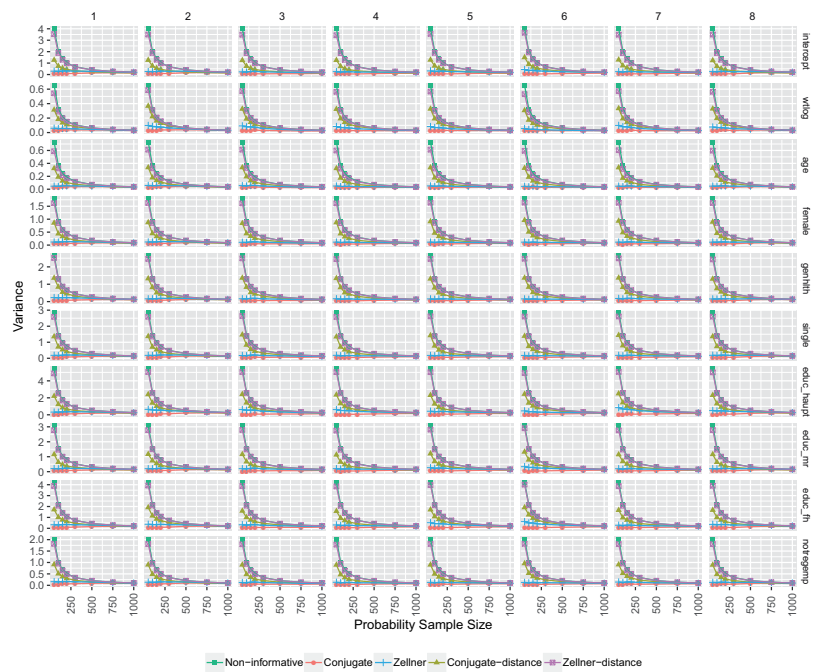


Figure C.2 Variance for regression coefficients (averaged over one hundred samples) in Bayesian model of body mass index on respondent covariates in the German Internet Panel. Prior distributions constructed using one of eight non-probability surveys (in columns). Age and weights (*wtlog*) covariates are standardized.

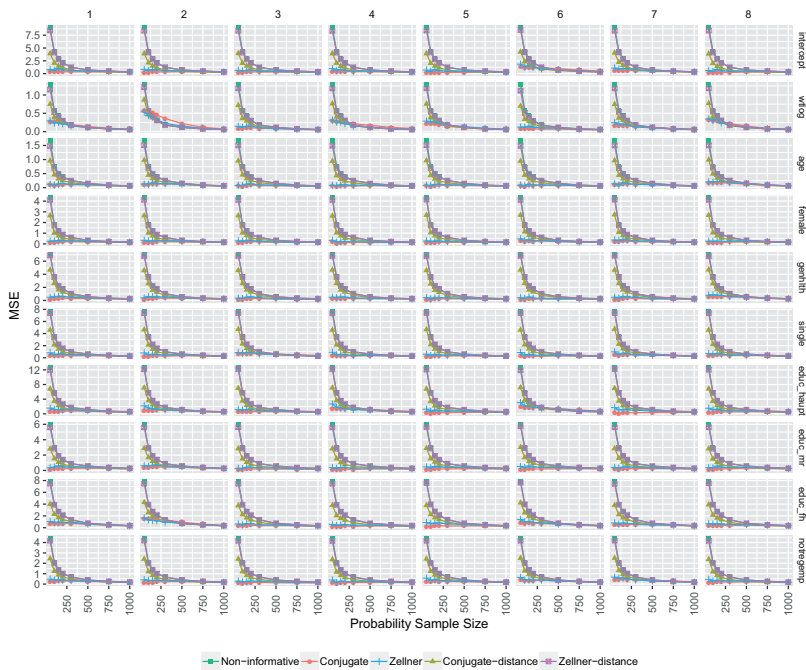


Figure C.3 Mean squared error for regression coefficients (averaged over one hundred samples) in Bayesian model of body mass index on respondent covariates in the German Internet Panel. Prior distributions constructed using one of eight nonprobability surveys (in columns). Age and weights (wtlog) covariates are standardized.

References

- AAPOR (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), Lenexa: American Association for Public Opinion Research.
- Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.
- Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons.
- Ansolabehere, S., and D. Rivers (2013), Cooperative Survey Research, *Annual Review of Political Science*, 16, 307–329.
- Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau (2013), Summary Report of the AAPOR Task Force on Non-Probability Sampling, *Journal of Survey Statistics and Methodology*, 1, 90–143.
- Bayes, T. (1763), An Essay towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions*, 53, 370–418.
- Blom, A. G., D. Ackermann-Piek, S. C. Helmschrott, C. Cornesse, and J. W. Sakshaug (2017), “The Representativeness of Online Panels: Coverage, Sampling and Weighting,” paper presented at the General Online Research Conference, Berlin, Germany.
- Blom, A. G., C. Gathmann, and U. Krieger (2015), Setting up an Online Panel Representative of the General Population: The German Internet Panel, *Field Methods*, 27, 391–408.

- Blom, A. G., J. M. E. Herzing, C. Cornesse, J. W. Sakshaug, U. Krieger, and D. Bossert (2016), Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German Internet Panel,” *Social Science Computer Review*, 35, 498–520.
- Callegaro, M., R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (2014), *Online Panel Research: A Data Quality Perspective*, New York: John Wiley & Sons.
- Carlin, B. P., and T. A. Louis (2008), *Bayesian Methods for Data Analysis*, Boca Raton: CRC Press.
- Chang, L., and J. A. Krosnick (2009), National Surveys via Rdd Telephone Interviewing versus the Internet Comparing Sample Representativeness and Response Quality, *Public Opinion Quarterly*, 73, 641–678.
- DiSogra, C., C. Cobb, E. Chan, and J. Dennis (2012), “Using probability-based online samples to calibrate non-probability opt-in samples,” paper presented at the 67th Annual Conference of the American Association for Public Opinion Research (AAPOR).
- Dutwin, D., and T. D. Buskirk (2017), Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples, *Public Opinion Quarterly*, 81, 213–239.
- Elliott, M. N., and A. Haviland (2007), Use of a Web-Based Convenience Sample to Supplement a Probability Sample, *Survey Methodology*, 33, 211–215.
- Elliott, M. R. (2009), Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights, *Survey Practice*, 2, 1–9.
- Elliott, M. R., and R. Valliant (2017), Inference for Nonprobability Samples, *Statistical Science*, 32, 249–264.
- Fahimi, M., F. M. Barlas, W. Gross, and R. K. Thomas (2014), “Towards a New Math for Nonprobability Sampling Alternatives,” paper presented at the 69th Annual Conference of the American Association for Public Opinion Research (AAPOR), Anaheim, CA.
- Hoerl, A. E., and R. W. Kennard (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12, 55–67.
- Jeffreys, H. (1946), An Invariant Form for the Prior Probability in Estimation Problems, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186, 453–461.
- Kass, R. E., and L. Wasserman (1995), A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion, *Journal of the American Statistical Association*, 90, 928–934.
- Khuri, A. I. (2003), *Advanced Calculus with Applications in Statistics*, Volume 486. New York: John Wiley & Sons.
- Lee, S. (2006), Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys, *Journal of Official Statistics*, 22, 329–349.
- Lee, S., and R. Valliant (2009), Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment, *Sociological Methods & Research*, 37, 319–343.
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008), Mixtures of g Priors for Bayesian Variable Selection, *Journal of the American Statistical Association*, 103, 410–423.
- MacInnis, B., J. A. Krosnick, A. S. Ho, and M.-J. Cho (2018), The Accuracy of Measurements with Probability and Nonprobability Survey Samples: replication and Extension, *Public Opinion Quarterly*, 82, 707–744.
- Malhotra, N., and J. A. Krosnick (2007), The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples, *Political Analysis*, 15, 286–323.
- Ntzoufras, I. (2011), *Bayesian Modeling Using WinBUGS*, Volume 698, New York: John Wiley & Sons.
- Pasek, J. (2016), When Will Nonprobability Surveys Mirror Probability Surveys? considering Types of Inference and Weighting Strategies as Criteria for Correspondence, *International Journal of Public Opinion Research*, 28, 269–291.
- Pennay, D. W., D. Neiger, P. J. Lavrakas, and K. A. Borg (2018), “The Online Panels Benchmarking Study: a Total Survey Error Comparison of Findings Form Probability-Based

- Surveys and Nonprobability Online Panel Surveys in Australia.” CSRM & SRC Methods Paper No. 02/2018, available at http://csrcm.cass.anu.edu.au/sites/default/files/docs/2018/12/CSRM_MP2_2018_ONLINE_PANELS.pdf (last accessed November 28, 2019).
- Pfeffermann, D. (1993), The Role of Sampling Weights When Modeling Survey Data, *International Statistical Review/Revue Internationale de Statistique*, 61, 317–337.
- Rao, J. N. (2003), *Small-Area Estimation*, New York: Wiley Online Library.
- Rivers, D. (2007), “Sampling for Web Surveys,” paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, UT, available at https://static.texastribune.org/media/documents/Rivers_matching4.pdf. (last accessed November 28, 2019).
- Rivers, D., and D. Bailey (2009), “Inference from Matched Samples in the 2008 US National Elections,” in *JSM Proceedings*, Survey Research Methods Section, p. 627–639, Alexandria, VA: American Statistical Association.
- Rubin, D. B. (1985), The Use of Propensity Scores in Applied Bayesian Inference, *Bayesian Statistics*, 2, 463–472.
- Sakshaug, J. W., A. Wiśniowski, D. A. Perez Ruiz, and A. G. Blom (2019), Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach, *Journal of Official Statistics*, 35, 653–629.
- Skinner, C. (1994), “Sample Models and Weights,” *Proceedings of the Section on Survey Research Methods*, American Statistical Association, p. 133142.
- Valliant, R., and J. A. Dever (2011), Estimating Propensity Adjustments for Volunteer Web Surveys, *Sociological Methods & Research*, 40, 105–137.
- Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpsen, and R. Wang (2011), Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples, *Public Opinion Quarterly*, 75, 709.
- Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Chichester: John Wiley & Sons.
- . (1986), “On Assessing Prior Distributions and Bayesian Regression Analysis with g -Prior Distributions,” in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, p. 233–243, North-Holland: Elsevier Science.