

# A Bayesian approach for combining probability and non-probability samples surveys

# A Bayesian approach for combining probability and non-probability samples surveys

## *Un approccio Bayesiano per combinare indagini da campioni probabilistici e non-probabilistici*

Salvatore Camilla, Biffignandi Silvia, Sakshaug Joseph, Struminskaya Bella, Wisniowski Arkadiusz

**Abstract**
Our paper proposes a method of combining probability and non-probability samples to improve analytic inference on logistic regression model parameters. A Bayesian framework is considered where only a small probability sample is available and the information from a parallel non-probability sample is provided naturally through the prior. A simulation study is run applying several informative priors. Comparisons on the performance of the models are studied with reference to their mean-squared error (MSE). In general, the informative priors reduce the MSE or, in the worst-case scenario, perform equivalently to non-informative priors.

**Abstract**
*Si propone di combinare campioni probabilistici e non per migliorare l'inferenza sui parametri del modello di regressione logistica con approccio Bayesiano. Si assume che sia disponibile un piccolo campione probabilistico e le informazioni provenienti da un grande campione non-probabilistico vengono fornite tramite la distribuzione a priori. Viene condotto uno studio tramite simulazione in cui si confrontano varie distribuzioni a priori informative. In generale, l'utilizzo di prior informative riduce l'errore quadratico medio o, nel caso peggiore, la performance è la stessa.*

**Key words:** Selection Bias, Data Integration, Bayesian Inference

[1] Salvatore Camilla, University of Milano-Bicocca, c.salvatore4@campus.unimib.it;
Biffignandi Silvia, CESS, biffisil@teletu.it;
Sakshaug Joseph, German Institute for Employment Research, joe.sakshaug@iab.de;
Struminskaya Bella, Utrecht University, b.struminskaya@uu.nl;
Wisniowski Arkadiusz, University of Manchester, a.wisniowski@manchester.ac.uk;

# 1 Introduction

Probability-based surveys (PS) are known to have higher data quality but are expensive and subject to relatively small sample sizes. Nonresponse is also becoming a relevant problem both for the sample size and the quality of the data. On the other hand, non probability surveys (NPS) are appealing since they are convenient but suffer from large selection biases. Accuracy of estimates and the inferential framework are still not methodologically defined. Nevertheless, due to large numbers of NPS available, and the problems arising in PS surveys the attention to study methods about how to use NPS and how to improve estimates in PS is growing and the issue more relevant. One natural strand of research is on the integration of PS and NPS. For example, Couper (2013), Miller (2017), Beaumont (2020) talk about exploiting advantages as well as overcoming respective disadvantages of both survey approaches. The most common approach is to adjust for selection bias in NPS estimates using reference PS or census data. A new recent approach proposed is to blend PS surveys with other NPS data sources (see Rao, 2021 for an extensive review).

Integrating both sample types is an ongoing topic of methodological research. We propose a method of combining probability and non-probability samples to improve analytic inference on model parameters. Specifically, we consider a Bayesian framework, where inference is based on a (small) PS and available information from a parallel NPS is provided naturally through the prior.

## *1.1 Research Aim*

Sakshaug et al. (2019) and Wisniowski et al. (2020) proposed a Bayesian data integration approach where inference is based on the PS and the available information from the NPS are supplied through the prior. This framework is studied for the analysis of continuous data. Nevertheless, categorical data, and particularly binary indicators, are of primary interest in surveys, especially in the field of social science, marketing research and psychological analyses. Our original contribution is to develop the abovementioned framework for the analysis of categorical data. In this paper we consider only a binary outcome. The aim is to improve inference about regression coefficients. To evaluate the proposed methodology, we conduct a simulation study assuming different selection scenarios (both missing-at-random MAR and non-missing-at-random MNAR), selection probabilities and sample sizes.

The aim is to compare the performance of some informative priors against a reference non-informative one in terms of mean-squared error (MSE).

The rest of this article is organised as follows. Section 2 introduces the methodological framework. The simulation results are presented and discussed in Section 3. In Section 4 conclusions are drawn.

## 2 Methodology

We rely on the Bayesian framework which offers a unified approach for integrating multiple data sources of different sizes and quality in a natural way, that is, through the prior structure. We consider a logistic regression to model a binary outcome with covariates. We assume to have a small PS survey and information from a NPS are provided through the prior. Following this approach, biased NP data are incorporated in the estimation process, and posterior estimates are likely to have more bias but possibly less variance than the one obtained using the refence prior.

In the full paper, we also present a real-data analysis study where the potential cost reductions are demonstrated.

### *2.1 The priors*

We propose and test the performance of several informative priors which can be grouped in two categories, distances priors and the power prior.

Distances priors are normally distributed and centred around the maximum likelihood estimates (MLEs) of regression coefficients using only NPS-data, while the scale parameter is a function of the distance between MLEs using the PS and NPS-data only. The smaller is the difference, the more informative is the prior. Hereafter, we refer to the basic Distance prior (Dist) which is representative of this class and its performance is good even in the worst-case scenario. In the full paper, more priors are presented and evaluated. We also consider a mixed version of the distance priors, i.e., only for the intercept the prior is replaced by the reference one.

The Power prior is based on the idea that the prior is proportional to an "initial prior", that we set equal to the reference one, and to the likelihood of the NPS-data. The likelihood is scaled by a parameter which regulates the influence of the NPS data. We set this parameter equal to the p-value resulting from the Hotelling T-test for differences between the two vectors of MLEs from the PS and NPS respectively.

The reference prior is a weakly informative prior proposed by Gelman et al. (2008). It is based on a Student t-distribution with 3 degrees of freedom, centered around 0 and with scale equal to 2.5.

To approximate the posterior distribution, we use the R packages rstan (Stan Development Team, 2021) and rstanarm (Goodrich et al., 2020) based on the No-U-Turn sampler, which is a variant of the Hamiltonian Monte Carlo algorithm.

### *2.2 The simulation framework*

We consider a simulation framework where we take into consideration different models to generate the population, various PS and NPS sizes and several combinations of selection scenarios and selection variables in order draw biased NPS data. Here the results for some selected cases are reported. In our full research study the extensive simulation framework and the complete combination set of the scenarios are considered. We consider the population to be generated from a logistic model with two binary predictors $X_1 \sim Ber(0.5)$ and $X_2 \sim Ber(0.5)$. We assume the coefficients to be $\beta_{MIX} \in (0.5,-1.3,0.9)$ so that the proportion of the outcome variable is almost balanced (0.57). Other results are discussed in the full study.

Under this model, we simulate a population of size N = 1,000,000 from which the PS is drawn with simple random sampling without replacement (srswor) design. We consider different probability sample sizes, from 50 to 1000. We draw a NPS of size 1000 from a simulated NP-panel considering five selection scenarios with different selection probabilities. Here, we present three scenarios and only two selection probabilities which refer to two extreme cases: no bias and high level of bias. The scenarios are the following: (1) $p$ depends on $Y$ (MNAR); (2) p depends on $X_1$ and $X_2$ (MAR); and (3) $p$ depends on $Y$, $X_1$ and $X_2$ (MNAR).

## 3 Results and Discussion

Given the framework presented in the previous section, we repeat the simulation 100 times and to compare the performance of the informative priors against the reference non-informative one, we consider the MSE of the posterior estimates, given by the square of the posterior bias plus the posterior variance.

Figure 1 shows the median MSEs for the selected scenarios and priors. If there is no bias, the reduction in MSEs using informative prior is remarkable, especially when the PS size is small, e.g., lower than 200 cases. When using mixed prior, due to the model formulation, the MSEs for the intercept are always close to the reference prior values.

If the selection mechanism is MAR, using informative priors and controlling for all the selection variables results in an impressive MSE reduction with respect to the reference prior, regardless of the level of selection bias. The power prior performs well when the level of bias is small and especially for small PS sizes (50-100 cases).

In the worst cases, the informative priors perform similarly to the reference prior while for lower levels of selection bias the gain in MSE reduction is evident.

## 4 Conclusions

The presented framework contributes to the survey data integration literature by proposing a Bayesian data integration approach to improve analytic inference about model parameters by integrating a small PS with a parallel NPS survey.

The current simulation study that, opposite to previous studies, also entails populations with different characteristics and the formulation of various selection mechanisms to account for varying levels of selection bias and selection variables, demonstrates that our approach is robust even in worst-case scenarios. In such a situation, informative priors perform similarly to the reference prior. In the presence of high selection bias, the Distance prior performs better.

In the full research study, we also present a real-case study where potential costs savings are evaluated. We point out that this methodology is not only suitable and profitable for low-budget organisations that can only afford a small PS, but also in the case where a larger PS is available (e.g. greater than 200 units).

In conclusion, in present times when probability samples are suffering from decreasing response rates and high costs and researchers are shifting towards convenient non-probability samples, integrating both samples becomes attractive from both an error and cost perspective.

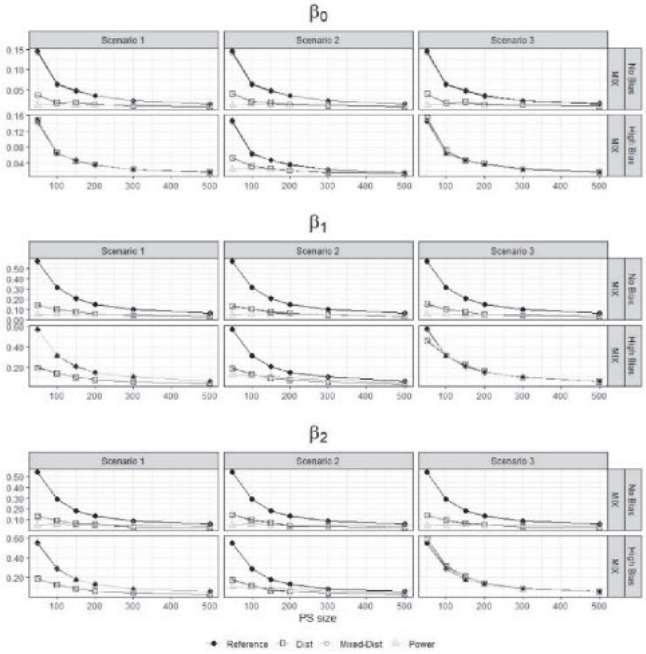


**Figure 1:** Median MSEs for regression coefficients over 100 iterations in alternative scenarios

## References


1. Beaumont, J.-F.: Are probability surveys bound to disappear for the production of official statistics? Survey Methodology 46 (1), 1–29 (2020).
2. Couper, M. P: Is the sky falling? new technology, changing media, and the future of surveys. Survey Research Methods 7 (3), 145–156 (2013).
3. Gelman, A., Jakulin, M. G. Pittau, and Y.-S. Su : A weakly informative default prior distribution for logistic and other regression models. The annals of applied statistics 2 (4), 1360–1383 (2008).
4. Goodrich, B., Gabry J., Ali I., and Brillema S.: rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 (2020).
5. Miller, P. V.: Is There a Future for Surveys? Public Opinion Quarterly 81 (S1), 205–212 (2017).
6. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 28 (2020).
7. Rao, J.: On making valid inferences by integrating data from surveys and other sources. Sankhya B 83 (1), 242–272 (2021).
8. Sakshaug, J. W., A. Wisniowski, D. A. P. Ruiz, and A. G. Blom.: Supplement-ing small probability samples with nonprobability samples: A bayesian approach. Journal of Official Statistics 35 (3), 653–681 (2019).
9. Wisniowski, A., J. W. Sakshaug, D. A. Perez Ruiz, and A. G. Blom.: Integrating probability and nonprobability samples for survey inference. Journal of Survey Statistics and Methodology 8 (1), 120–147 (2020).