Model Assisted Data Integration: An unbiased sampling strategy to use nonprobability data

Martin Hyllienmark, Statistics Sweden, martin.hyllienmark@scb.se

Gustaf Strandell, Statistics Sweden, gustaf.strandell@scb.se

The full paper can be found at: Hyllienmark, M., & Strandell, G. (2026). Model Assisted Data Integration: An unbiased sampling strategy to use nonprobability data. arXiv preprint arXiv:2604.00956.

Abstract.

Considerable efforts have been put into exploring how to use of found or non-probability data to minimize cost and response burden. However, to be able to safely utilize such data, rigorous theoretical foundations is needed, where one main concern is the of lack control due to not having access to the selection mechanism for the data. Several methods have been proposed in the literature to deal with this, though often relying on assumptions that may be difficult or impossible to verify in practice. This paper introduce the Model Assisted Data Integration sampling strategy. The proposed sampling strategy includes an estimator that has the desired properties: it is design-unbiased, has a design-unbiased variance estimator and is suitable for the intense production cycle of the statistical agency. The estimator uses nonprobability data combined with a probability sample that has a sampling design which aims to include individuals not captured by the nonprobability data. The estimator can use arbitrary machine learning models to produce unbiased estimates. A main conclusion of the paper is that the proposed sampling strategy can produce estimates with much lower variances compared to traditional survey estimators, and we use real empirical data to illustrate this point.

Keywords: Nonprobability data, design unbiased, estimator, probability sample, data integration

1. Introduction

Nonprobability data or found data has been investigate broadly during the last decades. For the statistical offices, the hope is to utilize such data to lower costs, increase quality and reduce response burden (Tam & Holmberg, 2020). The step from research or pilot study to production at the statistical office has often proved to be a long and slow one (Holmberg, 2025). Although given much attention in research, one such challenge is undercoverage of non-probability data with respect to the population of interest.

We will use the term nonprobability data (NPD) to refer, somewhat loosely, any data source that does not have known non-zero inclusion probabilities or is a register with complete coverage of the target population. There is a growing necessity for methods and survey designs in which data from different sources can be combined, including traditional surveys, statistical registers and non-probability data sources (Holmberg, 2025).

The use of machine learning for variance reduction in survey sampling is a fairly new and exciting area of research (Ranalli, 2025). Lundy and Rao (2022) demonstrated that having a wider range of models to choose from can come with benefits such as greater variance reduction and a reduced dependence on properties of the auxiliary information. If the model can sort out the data itself, a higher degree of automation will be possible. The move from linear regression to machine learning does however increase the risk of overfitting of the model to the data available (Ranalli, 2025).

In this paper we present and evaluate a sampling strategy that we believe can enable the use of more NPD in the production of official statistics. The estimator is based on a particular form of the data integrated estimator by Kim and Tam (2021), although it provides a further variance reduction by using an arbitrary machine learning model and externally available auxiliary information. The sampling strategy requires the availability of auxiliary information at the population level before sampling, and the presence of an identifier in the non-probability data making it possible to match it to other data. These conditions are satisfied by many both existing and up-coming data sources at several NSOs. It is design-unbiased and we propose a design-unbiased variance estimator which is suitable for any choice of model, making it very safe to use. We demonstrate the proposed estimator with a random forest model since it usually performs well without the need of tuning hyperparameters, hence it is suitable for the industrialized conditions

under which official statistics is produced. It is however possible to apply any arbitrary model to the estimator.

2. The model assisted data integration estimator

Let U be a finite population with labels $\{1, \dots, i, \dots, N\}$ and let $(\boldsymbol{x_i}, y_i)$ be a vector of auxiliary variables $\boldsymbol{x_i}$ assumed here to be available for each $i \in U$ and the value of a study variable y_i associated with the ith unit. We shall consider the estimation of the population total $t_y = \sum_U y_i$. Let A be a subset of U of size N_A. In this paper we assume that the study variable y_i is known for all $i \in A$ but we make no assumptions on the selection process of A. Thus A is NPD and no design weights are available. Let B be the complement of A in U, with size N_B. Let s_b be a probability sample drawn from B with inclusion probabilities π_{iB} and second order inclusion probabilities $\pi_{ijB} > 0$. In other words, this is a traditional probability sample drawn from the subsection of U not covered by the NPD.

Let $\mu(\boldsymbol{x}, A)$ be a model trained on A based on the observed study variable y_i and the auxiliary information $\boldsymbol{x}_i$ from units in A. No assumptions are made about the structure of $\mu(\boldsymbol{x}, A)$, it can be any arbitrary model. The proposed model assisted data integration ($MADI$) estimator has the following structure

$$\hat{t}_{y,MADI} = \sum_{i \in A} y_i + \sum_{i \in B} \mu_{MADI}(\boldsymbol{x_i}, A) + \sum_{i \in s_b} \frac{e_i}{\pi_{iB}}$$

where $e_i = y_i - \mu_{MADI}(\boldsymbol{x_i}, A)$. Since s_b is a probability sample, we can aggregate it using the design weights π_{iB} to estimate the population total. In words this estimator combines the total of the known y_i in A, a model estimate for the total of y_i in B, and a third term that estimates the error in the model term. Since $\mu_{MADI}(\boldsymbol{x_i}, A)$ does not depend on a probability sample it is unbiased

$$E\left(\hat{t}_{y,MADI}\right) = E\left(\sum_{i \in A} y_i + \sum_{i \in B} \mu_{MADI}(\boldsymbol{x_i}, A) + \sum_{i \in s_b} \frac{e_i}{\pi_{iB}}\right) = t_{y,A} + t_{\mu,B} + t_{y,B} - t_{\mu,B} = t_y$$

where $t_{\mu,B}$ correspond to the total of y in B based on the predictions from $\mu(\boldsymbol{x}, A)$. The variance for the proposed estimator is given by

$$V\left(\hat{t}_{y,MADI}\right) = \sum_{i \in B} \sum_{j \in B} \Delta_{ijB} \frac{e_i}{\pi_{iB}} \frac{e_j}{\pi_{jB}}$$

where $\Delta_{ijb} = \pi_{ijB} - \pi_{iB}\pi_{jB}$, and an unbiased estimator for the variance is given by

$$\hat{V}(\hat{t}_{y,MADI}) = \sum_{i\in b}\sum_{j\in b}\frac{\Delta_{ijB}}{\pi_{ijB}}\frac{e_i}{\pi_{iB}}\frac{e_j}{\pi_{jB}}$$

It is important to note that since $\mu_{MADI}(\boldsymbol{x_i}, A)$ is trained on the NPD, which perceptually is much larger compared to a traditional probability sample, the risk of having too few data points for a given model is mitigated. For example, a neural network model with several dimensions is most often not possible to use in traditional surveys since the number of required observations is usually too few. Furthermore, it also allows for more extensive use of auxiliary information in the modeling process. As illustrated by e.g. Lundy & Rao, 2022, a GREG model has a limit to the number of possible auxiliary variables the model can handle in relationship to the number of sampled units. Under the assumption that the NPD is of decent size, this issue is counteracted with $MADI$.

3. Simulations

We shall investigate what sample size is needed given a coefficient of variation (square roof of the variance over the parameter) of 1%. Since we have shown that the $MADI$ sampling strategy is unbiased, we chose to use the CV here instead of the RMSE since it provides a more standardized measurement.

The simulation is based on data from the Swedish Income and Taxation (IaT) register from year 2021. The register includes variables related to different types of income, i.e., income related taxes, pensions, and social insurance benefits. For computational reasons, we constructed a population by drawing a random subset of size $N = 9\,838$, excluding extreme outliers. The study variable y is Annual Determined Earned Income measured in Swedish Krones. We selected twelve explanatory variables, where X_1 represents age, X_2 represents gender, and X_3 to X_{12} are associated with taxes.

The nonprobability dataset was generated directly from U and remained fixed for all estimators.

Two different designs are used in the simulation, denoted SRS_U and SRS_B. In both cases we use a simple random sample, the difference between the two is that in SRS_U the sample is drawn from U while in SRS_B the sample is drawn from $B = U\backslash A$. In SRS_U we

draw a sample of size n where each unit has the inclusion probability $\pi_i = \frac{n}{N}$. This design represents a traditional survey where an estimate is calculated based on a sample drawn from a population. Similarly, in SRS_B a SRS of size n_b is drawn with inclusion probabilities $\pi_{iB} = \frac{n_b}{N_B}$. This design is used to illustrate the scenario where we have information about the study variable in an incomplete register or where we have access to other nonprobability data that do not cover the whole population.

We employ the following sampling strategies in the simulations

$SRS_U - HT$: Simple random sample from U coupled with the HT-estimator
$SRS_U - GREG$: Simple random sample from U coupled with the GREG estimator
$SRS_B - MADI_{RF}$: Simple random sample from B coupled with a $MADI$-estimator and a RF model

Our aim is to illustrate that $MADI$ performs well even when a missing not at random mechanism has generated the NPD. To this end, we shall alter two parameters in the simulation. The first parameter is the percentage of U that is generated in A. Let subsets $A_{kl} \subseteq U$ for $k = 1,2, \dots ,8$ and $l = 1, \dots ,9$ be proportionally to U such that $|A_{kl}| = 0.1l * |U|$, i.e. the size of A_{kl} range from 10% of the population size up to 90% of the population size. The second parameter k shall alter the way that A is generated from U which allows us to study how the $SRS_B - MADI_{RF}$ performs under different levels of selection bias in the NPD.

In Table 1 below we illustrate the mean value of the study variable y_i in A_{kl} and B_{kl} for all possible k shown by rows and l illustrated on the columns. This should give an indication of the level of selection bias in each scenario. We observe an exceptionally high difference in mean values between A_{kl} and B_{kl} for $k = 1,2$ which is expected since these scenarios use cutoff points based on the study variable. The difference in the other scenarios is less pronounced but still visible for all A_{kl}.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
A_{1l}	702818	609254	551557	507715	470249	436196	404410	373785	340951
B_{1l}	263018	231426	202169	173212	143764	113206	79799	39964	2606
A_{2l}	2524	39964	79799	113206	143764	173212	202169	231426	263018

B_{2l}	340769	373785	404410	436196	470249	507715	551557	609254	702818
A_{3l}	402603	399047	397125	381699	373889	364999	352552	338558	323371
B_{3l}	296383	283991	268392	257216	240124	220012	200714	180836	159766
A_{4l}	263877	273093	267340	272268	273918	275202	276593	282424	288113
B_{4l}	311800	315487	324003	330164	340095	354718	377987	405313	477015
A_{5l}	226549	232548	230905	238657	242258	252213	263167	278913	297507
B_{5l}	315949	325626	339616	352569	371756	389205	409319	419355	392485
A_{6l}	386215	383985	376508	378903	371579	364791	357929	343788	320649
B_{6l}	298204	287757	277226	259080	242435	220324	188166	159918	184252
A_{7l}	254864	257267	259471	260613	264043	273180	276490	286037	294922
B_{7l}	312802	319445	327375	337934	349971	357752	378226	390863	415744
A_{8l}	322552	339419	340915	331382	335557	330436	329925	326606	320559
B_{8l}	305279	298902	292478	290758	278457	271860	253521	228630	185066

Table 1. Each cell shows the mean of y_i for A_{kl} and B_{kl} where l is represented in different columns as the proportional size of A_{kl} from U and k is represented by rows indicating the selection process of A_{kl}.

The coefficient of variation (*CV*) is calculated as a function of the theoretical variance. The sample size needed for each strategy to achieve a *CV* of 1% for each A_{kl} is presented in Table 2. Note that the sample size needed for $SRS_B - MADI_{RF}$ is based on the population size in B.

The first two rows in Table 3 illustrate $SRS_U - HT$ and $SRS_U - GREG$. Since these estimates do not depend on A_{kl}, they are constant over l and k. The $SRS_U - HT$ requires 3044 sampled units and the $SRS_U - GREG$ requires 269 sampled units. The table further shows the $SRS_B - MADI_{RF}$ for each A_{kl} where k are presented on different rows and l are presented on columns. We observe that for $k = 3, \dots, 8$, $SRS_B - MADI_{RF}$ outperform both the $SRS_U - GREG$ and $SRS_U - HT$ for all sizes of l except for A_{68} in this illustration.

The other situation where the $SRS_U - GREG$ requires a lower sample size than $SRS_B - MADI_{RF}$ is when the selection bias is extreme as illustrated by $k = 1,2$. This illustration does not consider the issue $SRS_U - GREG$ has with invertible matrices for small samples in combination with many auxiliary variables. In a separate simulation not shown in this paper, it was concluded that $SRS_U - GREG$ often yielded invertible matrices when $n < 400$.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
$SRS_U - HT$	3044	3044	3044	3044	3044	3044	3044	3044	3044
$SRS_U - GREG$	269	269	269	269	269	269	269	269	269
$A_{1l} - MADI_{RF}$	3081	3030	2820	2850	2857	2627	2386	1788	984
$A_{2l} - MADI_{RF}$	2210	1646	1266	988	775	626	518	417	255
$A_{3l} - MADI_{RF}$	122	106	97	91	102	112	113	137	139
$A_{4l} - MADI_{RF}$	125	117	97	83	79	73	64	65	42
$A_{5l} - MADI_{RF}$	123	109	80	82	76	65	70	73	76
$A_{6l} - MADI_{RF}$	151	142	140	136	157	157	222	286	211
$A_{7l} - MADI_{RF}$	229	163	119	130	107	87	103	90	80
$A_{8l} - MADI_{RF}$	126	96	89	77	73	75	89	108	91

1. Conclusion

In this paper we propose the $MADI$ sampling strategy that can incorporate information from a non-probability dataset. Ranalli (2025) highlight the importance of unbiased and reliable methods when using such data to avoid uncertainties introduced by having data not collected by a survey with associated design weights. We have shown both theoretically the $MADI$ sampling strategy is unbiased and that it has an unbiased variance estimator. The simulation showed that the proposed strategy performs well for realistic scenarios compared to standard alternatives. The estimator is competitive in the sense that it does not rely on MAR in the NPD which is quite common in other estimators that incorporate such data. The estimator is also easy to implement in statistical production, especially for statisticians with a background in survey theory.

We have limited ourselves to illustrate that the $MADI$ sampling strategy works under certain conditions in this paper. However, further research should look at how the sampling strategy performs with the implementations of other models. A clear advantage of $MADI$ is that if the dataset A is relatively large, the number of available models becomes larger compared to a traditional survey dataset. For example, a neural network of several dimensions requires a vast amount of data to accurately estimate all model parameters, this is simply not feasible with the amount of data gathered from most surveys.

Reference list

Holmberg, A. (2025). Future Pathways Embracing Multisource Statistics and Novel Data Sources at National Statistical Offices. *Journal of Official Statistics*, *41*(3), 795-803.

Kim, J. K., & Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 89(2), 382-401.

Lundy, E. R., & Rao, J. N. K. (2022). Relative performance of methods based on model-assisted survey regression estimation: A simulation study. *SURVEY METHODOLOGY*, *48*(1), 49-71.

Ranalli, M. G. (2025). Machine learning methods for estimation in official statistics. Journal of Official Statistics, 41(3), 912-920.

Tam, S. M., & Holmberg, A. (2020). New Data Sources for Official Statistics–A Game Changer for Survey Statisticians?. The Survey Statistician, 81, 21-35.