Journal of Official Statistics
vol. 38, 3 (September 2022)

Construction of Databases for Small Area Estimation.....673-708
Emily Berg

Hierarchical Bayesian Model with Inequality Constraints for US County Estimates.....709-732
Lu Chen, Balgobin Nandram and Nathan B. Cruze

Timely Estimates of the Monthly Mexican Economic Activity.....733-765
Francisco Corona, Graciela González-Farías and Jesús López-Pérez

Small Domain Estimation of Census Coverage – A Case Study in Bayesian Analysis of Complex Survey Data.....767-792
Joane S. Elleouet, Patrick Graham, Nikolai Kondratev, Abby K. Morgan and Rebecca M. Green

Identifying Data Quality Challenges in Online Opt-In Panels Using Cognitive Interviews in English and Spanish.....793-822
Yazmín García Trejo, Mikelyn Meyers, Mandi Martinez, Angela O'Brien, Patricia Goerman and Betsarí Otero Class

Measuring and Mapping Micro Level Earning Inequality towards Addressing the Sustainable Development Goals – A Multivariate Small Area Modelling Approach.....823-845
Saurav Guha and Hukum Chandra

Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM) 847-873
Daniel Hopp

Variable Inclusion Strategies through Directed Acyclic Graphs to adjust Health Surveys subject to Selection Bias for Producing National Estimates.....875-900
Yan Li, Katherine E. Irimata, Yulei He and Jennifer Parker

Pseudo Bayesian Mixed Models under Informative Sampling.....901-928
Terrance D. Savitsky and Matthew R. Williams

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 673–708, http://dx.doi.org/10.2478/JOS-2022-0031

Construction of Databases for Small Area Estimation

Emily Berg[1]

The demand for small area estimates can conflict with the objective of producing a multi-purpose data set. We use donor imputation to construct a database that supports small area estimation. Appropriately weighted sums of observed and imputed values produce model-based small area estimates. We develop imputation procedures for both unit-level and area-level models. For area-level models, we restrict to linear models. We assume a single vector of covariates is used for a possibly multivariate response. Each record in the imputed data set has complete data, an estimation weight, and a set of replicate weights for mean square error (MSE) estimation. We compare imputation procedures based on area-level models to those based on unit-level models through simulation. We apply the methods to the Iowa Seat-Belt Use Survey, a survey designed to produce state-level estimates of the proportions of vehicle occupants who wear a seat-belt. We develop a bivariate unit-level model for prediction of county-level proportions of belted drivers and total occupants. We impute values for the proportions of belted drivers and vehicle occupants onto the full population of road segments in the sampling frame. The resulting imputed data set returns approximations for the county-level predictors based on the bivariate model.

Key words: Imputation; area-level; unit-level; bootstrap; bivariate.

1. Introduction

We consider the problem of constructing a data set that provides good estimates for pre-specified small areas and for a pre-specified set of response variables. We approach this objective through donor imputation and replication variance estimation. We define procedures to select donors and construct corresponding weights such that the resulting small area estimate is close to the empirical best linear unbiased predictor (EBLUP) for a specified small area model. We also consider linear combinations of small area parameters, such as the average of two small area means and the overall mean for the full population.

Our interest in this topic grew out of a study of procedures used for the National Resource Inventory (NRI), a longitudinal survey of non-federal land in the United States (Nusser and Goebel 1997). The NRI product includes a record-level database, in addition to a set of estimates. In the 1997 NRI, small area estimates were constructed for urban areas and roads. Imputation was used to ensure that the record-level database provided good estimates for small area parameters of interest. We develop procedures that are more general than those used for the NRI.

In this article, we apply one of the proposed imputation procedures to data from the Iowa Seat-Belt Use Survey. We develop a bivariate small area model for the proportions of

[1]Iowa State University of Science and Technology, Department of Statistics Snedecor Hall, Ames. Iowa, 50011-2140, U.S.A. Email: emilyb@iastate.edu

© Statistics Sweden

drivers and vehicle occupants who are wearing a seat-belt. We produce a database that approximates model-based small area estimates.

The problem of constructing a database that re-produces model-based small area estimates is of interest beyond the NRI and seat-belt use survey applications. Many large-scale survey operations produce a data set in addition to a set of estimates and standard errors. The data set is typically equipped with a set of weights and replicate weights designed to produce direct estimates and associated standard errors for domains with adequately large sample sizes. Provision of a unit-level database enhances transparency and flexibility for a data user. Direct estimates for small areas constructed from the database may be unreliable due to small sample sizes.

For production of small area estimates, the data producer may consider two possible approaches. One is to provide a separate set of tables containing model-based small area estimates. At the other extreme, the data producer may expect the data user to implement specialized small area models using the unit-level data. Both of these approaches have limitations.

The approach of providing tables with model-based small area estimates and associated root mean square errors is restrictive. Use of tabular output to construct estimates for linear combinations of small area parameters may be difficult. Conflicts between tabular output and direct estimates can cause confusion for a user who is provided with both tables of estimates and unit-level data.

User-specific analyses also have limitations for small area estimation. Routines for implementing small area models in standard software packages, such as SAS and R, are limited to a relatively narrow class of models. Different data users may obtain different small area estimates from the same data set. The user may lack access to population-level auxiliary information needed for model-based small area estimation.

We propose an approach that offers a compromise between the two extremes of tabular estimates and an unstructured data set. We construct a database that preserves specified small area estimates. This permits the data producer to absorb responsibility for small area modeling, while still allowing the user access to unit-level data. Different users of the unit-level data will obtain the same model-based small area estimates. We formalize the problem of imputation for the purpose of producing a database that provides estimates of pre-specified small area parameters.

We suppose a model-based procedure exists to estimate small area means. We consider a framework where a single vector of covariates is used for a possibly multivariate response. We focus on linear small area models. The goal of the survey is not simply to produce a set of estimates but is also to produce a database that an analyst can use to construct estimates for desired parameters of interest. To accomplish this goal, we augment the set of observed values with a set of imputed values. The imputed values are observed response variables obtained from donors. We use donor imputation because donor imputation can be robust to model assumptions and can preserve multivariate associations in the data. We also construct a revised set of weights. Appropriately weighted sums of observed and imputed values re-produce model-based small area estimates. We also use the weights to construct estimators of linear combinations of small area means. We further provide a set of replicate weights such that replicate variance estimators re-produce standard estimators of the mean square error of the small area predictor.

We illustrate imputation for small area estimation through simulation and a data analysis. We apply the methods to data from the 2019 Iowa Seat-Belt Use Survey. This annual survey provides estimates of seat-belt use rates at the state level. Additionally, the survey produces an estimation dataset with information for sampled road segments. We develop a bivariate unit-level model for estimation of seat-belt use rates at the county level. We use the bivariate model as the basis for an imputation procedure. We construct imputed values and weights for every road segment in the sampling frame. Weighted sums of imputed values approximate the model-based small area estimates. We also investigate imputation for a bivariate unit-level model through simulation. In the simulations, we compare procedures based on univariate models to procedures based on bivariate models.

We aim for generality, but our procedures do not encompass all situations of possible interest. We focus on linear models. The approach based on unit-level models requires specification of a vector of variables of interest. The procedures do not immediately apply to arbitrary definitions of domains. We do not consider the case of nonsampled areas in detail. We require that the same vector of covariates be used for all response variables.

1.1. Related Literature

A close alternative to our approach is Chandra and Chambers (2009). They propose a set of weights intended to yield efficient estimates for full population parameters as well as reasonable small area estimates. We compare our imputation approach to the approach of Chandra and Chambers (2009) through simulation.

Our approach also relates to a procedure to construct a transparent data file for the 2000 Census of Population (Isaki et al. 2000). A statistical estimation procedure was used to combine estimates based on the Census of Population short form with data from a follow-up survey to estimate the under-count. Pseudo-values were computed so that the resulting data file would reproduce the combined estimates.

The method of weight-splitting (Randrianasolo and Tille 2013) also addresses an objective similar to ours. Weight-splitting is a way to produce a set of weights for small area estimation. The resulting weights preserve a benchmarking restriction that the sum of the estimates for the small areas is equal to the estimate for the full population. The weight-splitting procedure uses D sets of weights to produce estimates for D small areas. In contrast, we augment the data set with imputed values, allowing us to utilize a single vector of estimation weights.

1.2. Outline

We define the algorithmic steps required to construct imputed values and weights that reproduce small area estimates. The two main categories of small area models are unit-level models (Battese et al. 1988) and area-level models (Fay and Herriot 1979). Rao and Molina (2015) discuss these two categories of small area models in detail. An analyst may want to use either of these two categories of models. The desire to construct a database to support small area estimation need not limit the type of model used. We define imputation procedures for area-level and unit-level models in Sections 2 and 3, respectively. We evaluate the procedures through simulation in Section 4. We apply one of the imputation methods to estimate county level seat-belt use rates using data from the Iowa Seat-Belt

Use Survey in Section 5. We summarize and discuss limitations of our work in Section 6. The Appendix (Section 7) contains details of estimation procedures and a table summarizing the proposed imputation procedures.

2. Estimators for Area Level Models as Imputed Estimators

In this section, we define the steps to construct a database that reproduces estimates based on an area level model for direct estimators of means. As a precursor to the definition of the imputation procedure, we overview the widely-used Fay and Herriot (1979) model. The Fay-Herriot model is called an "area-level model" because the response variable in the model is a direct estimator of a small area parameter.

We establish notation needed to define the area-level model. Estimates are desired for D areas indexed by $k = 1, \ldots, D$. Let $j = 1, \ldots, N_k$ index the population elements for area k. Denote a univariate response variable by y_{kj}. Assume the population size N_k is known. A sample of n_k elements is selected for area k, and the sampled elements are denoted by $j = 1, \ldots, n_k$. We also use A_k to denote the index set of the n_k elements in the sample for area k. We assume that all areas have at least one sampled element. Denote the selection probability for element j in area k by π_{kj}. Let x_k be a fixed covariate for area k.

Define a direct estimator of a small area mean by $\hat{Y}_k = \sum_{j \in A_k} w_{kj} y_{kj}$, where $w_{kj} = N_k^{-1} \pi_{kj}^{-1}$. Denote the design-variance of $\hat{Y}_k$ by σ_{ek}^2. Assume that the survey provides an estimate of σ_{ek}^2. As is common in the literature, we operate as if σ_{ek}^2 is the known design-variance of $\hat{Y}_k$.

A standard Fay-Herriot model for the direct estimator of the mean for area k is given by

$$\hat{Y}_k = x_k' \beta + b_k + e_{k,} \tag{1}$$

where $b_k \sim N(0, \sigma_b^2)$, and $e_k \sim N(0, \sigma_{ek}^2)$. Recall that σ_{ek}^2 denotes the design variance of $\hat{Y}_k$. In practice, σ_{ek}^2 is estimated from the unit-level data. We treat the estimated σ_{ek}^2 as if it were the true σ_{ek}^2. The assumption that σ_{ek}^2 is known is required to identify the parameters of the model (1). The best linear unbiased estimator (BLUE) of β is

$$\hat{\beta}_{BLUE}\left(\sigma_b^2\right) = \left(\sum_{i=1}^{D} x_i x_i' \left(\sigma_b^2 + \sigma_{ei}^2\right)^{-1} \right)^{-1} \sum_{i=1}^{D} x_i \left(\sigma_b^2 + \sigma_{ei}^2\right)^{-1} \hat{Y}_i,$$

where the notation $\hat{\beta}_{BLUE}\left(\sigma_b^2\right)$ emphasizes dependence of the BLUE on the unknown σ_b^2.

The parameter representing the small area mean of interest is $\theta_k = x_k' \beta + b_k$. The best linear unbiased predictor (BLUP) of θ_k is given by

$$\hat{\theta}_k^B = \hat{\theta}_k^B \left\{ \hat{\beta}_{BLUE}\left(\sigma_b^2\right), \sigma_b^2 \right\} = (1 - \gamma_k) x_x' \hat{\beta}_{BLUE}\left(\sigma_b^2\right) + \gamma_k \hat{Y}_k,$$

where $\gamma_k = \sigma_b^2 \left(\sigma_b^2 + \sigma_{ek}^2\right)^{-1}$. The notation $\hat{\theta}_k^B \left\{ \hat{\beta}_{BLUE}\left(\sigma_b^2\right), \sigma_b^2 \right\}$ emphasizes dependence of the BLUP on the unknown σ_b^2. The design-variance σ_{ek}^2 generally decreases as n_k increases such that γ_k, approaches 1 as the area sample size increases. For a simple random sample, σ_{ek}^2 is of the form $\sigma_{ek}^2 = \sigma_e^2 / n_k$, in which case the connection between γ_k, and n_k is obvious.

The estimated best linear unbiased predictor (EBLUP) is defined by replacing the unknown σ_b^2 with an estimator $\hat{\sigma}_b^2$. We express the EBLUP for model (1) as $\hat{\theta}_k^{EB} = \hat{\theta}_k\left(\hat{\beta}, \hat{\sigma}_b^2\right)$, where $\hat{\beta} = \hat{\beta}_{BLUE}\left(\hat{\sigma}_b^2\right)$. A number of procedures to estimate σ_b^2 exist. We use a

moment-type estimator called the Fay-Herriot estimator that is implemented in the `SAE R` package. Given an estimate $\hat{\sigma}_b^2$, the EBLUP of θ_k is defined as

$$\hat{\theta}_k^{EB} = \hat{\theta}_k^B\{\hat{\beta}_{BLUE}(\hat{\sigma}_b^2), \hat{\sigma}_b^2\} = (1 - \hat{\gamma}_k)x_k'\hat{\beta}_{BLUE}(\hat{\sigma}_b^2) + \hat{\gamma}_k\hat{Y}_k, \tag{2}$$

where $\hat{\gamma}_k = \hat{\sigma}_b^2(\hat{\sigma}_b^2 + \sigma_{ek}^2)^{-1}$.

It is widely known that the mean square error (MSE) of the EBLUP decomposes into a sum of two terms as $mse\{\hat{\theta}_k^{EB}\} = M_{1k} + M_{2k}$, where $M_{1k} = \gamma_k\sigma_{ek}^2$, $M_{2k} = E[(\hat{\theta}_k^{EB} - \tilde{\theta}_k^B)^2]$, and $\tilde{\theta}_k^B = \gamma_k\hat{Y}_k + (1 - \gamma_k)x_k'\beta$ (Rao and Molina 2015, 273). The first term, M_{1k}, is $V\{b_k|\hat{Y}_k\} = \gamma_k\sigma_{ek}^2$. The second term accounts for estimation of β and σ_b^2.

Jiang et al. (2002) propose a jackknife estimator of M_{2k} of the form

$$\hat{M}_{2k}^{JK} = \left(\frac{D-1}{D}\right)\sum_{r=1}^{D}\left(\hat{\theta}_k^{(r)} - \hat{\theta}_k^{EB}\right)^2, \tag{3}$$

where $\hat{\theta}_k^{(r)} = \hat{\theta}_k^{EB}\{\hat{\beta}_{BLUE}^{(r)}(\hat{\sigma}_{b(r)}^2), \hat{\sigma}_{b(r)}^2\}$, and $(\hat{\beta}_{BLUE}^{(r)}(\hat{\sigma}_{b(r)}^2), \hat{\sigma}_{b(r)}^2)$ denotes an estimate of $(\beta', \hat{\sigma}_b^2)$ constructed with area r omitted. An estimator of $mse\{\hat{\theta}_k^{EB}\}$ is $\hat{mse}_k = \hat{\gamma}_k\sigma_{ek}^2 + \hat{M}_{2k}^{JK}$. We refer the reader to Jiang et al. (2002) and Rao and Molina (2015) for further detail on the jackknife MSE estimator (3).

The usual formulation of the EBLUP for the area level model (1) summarized here expresses the EBLUP as a function of the direct estimators $\{\hat{Y}_k: k = 1, \ldots, K\}$. Standard procedures involve first calculating $\hat{Y}_k$ from the unit-level data and then using $\hat{Y}_k$ to construct the EBLUP. Because $\hat{Y}_k$ is a weighted sum of unit-level data, one can express the EBLUP as a weighted sum of unit-level observations. Our database construction process will utilize an expression the EBLUP as a weighted sum of the unit-level data.

We define a database composed of real observations, imputed values, and weights such that appropriately weighted sums of observed and imputed values reproduce the EBLUP, $\hat{\theta}_k^{EB}$ defined in Equation (2). In this direction, we first express the EBLUP for the Fay-Herriot model as a weighted sum of sampled units in Subsection 2.1. We then define imputed values and weights to reproduce the EBLUP in Subsection 2.2. In Subsection 2.2, we also define replicate weights to reproduce the jackknife variance estimator (3).

2.1. Estimator for Area-Level Model As a Weighted Sum of Sampled Values

The model (1) is called an area level model because the response variable in the model is a direct estimator for an area. The EBLUP of Equation (2) for the Fay-Herriot model is an area-level estimator in the sense that one can construct the EBLUP using only the direct estimators $\{\hat{Y}_k: k = 1, \ldots, D\}$. The EBLUP (2) is implicitly a function of the unit-level data because Y_k is a function of the unit-level data. Specifically, $\hat{Y}_k = \sum_{j=1}^{nk} w_{kj}y_{kj}$, where $w_{kj} = \pi_{kj}^{-1}N_k^{-1}$. We now express the EBLUP of Equation (2) for the Fay-Herriot model as a weighted sum of all elements in the sample. In this direction, we require two indexes for the area and a separate index for the unit within the area. We use the index of k to denote the area for which the predictor is constructed. We use the label of i to index other areas used in construction of the predictor for area k. We use j to denote a unit within an area. One can write the EBLUP for area k as

$$\hat{\theta}_k^{EB} = \sum_{i=1}^{D} \sum_{j=1}^{n_i} \left\{ \left(1 - \hat{\gamma}_k\right) x_k' A_i w_{ij} + I[i = k] \hat{\gamma}_i w_{ij} \right\} y_{ij}, \tag{4}$$

where $A_i = M^{-1} x_i (\sigma_{ei}^2 + \hat{\sigma}_b^2)^{-1}$, and $M = \sum_{i=1}^{D} (\sigma_{ei}^2 + \hat{\sigma}_b^2)^{-1} x_i x_i'$. Expression (4) motivates a weight of the form

$$\delta_{kij} = \left(1 - \hat{\gamma}_k\right) x_k' A_i w_{ij} + I[i = k] \hat{\gamma}_k w_{ij}, \tag{5}$$

where $I[\cdot]$ is the indicator function that equals 1 if the argument is true and 0 otherwise. The form of the weight in Equation (5) combined with the expression of the predictor in Equation (4) imply that one can express the EBLUP as a weighted sum of sampled units as

$$\hat{\theta}_k^{EB} = \sum_{i=1}^{D} \sum_{j=1}^{n_i} \delta_{kij} y_{ij}. \tag{6}$$

The weight δ_{kij} depends on the estimation domain k. Therefore, we cannot apply a uniform weighting adjustment to construct small area predictors for all D areas. Simple use of the weight δ_{kij} would require D sets of weights to construct small area estimates for D domains, similar to Randrianasolo and Tillé (2013). We can overcome this issue by constructing imputed values for each area. We define imputed values and weights to reproduce $\hat{\theta}_k^{EB}$ in Subsection 2.2.

2.2. *Area-Level Imputation Procedure (Area-Imp)*

We define imputed values and weights that reproduce the EBLUP for the Fay-Herriot model. We call this first imputation procedure "Area-Imp." The imputation procedure has two parts. The first part involves defining imputed values to reproduce the small area predictors. To accomplish this first step, we augment the data set for each area k with the $n - n_k$ observations in the sample that are not in area k. We call the imputed values constructed in step 1 the "prediction imputed values," as they are defined so that weighted sums of observed and imputed values for an area reproduce the EBLUP for the area. The second step involves defining additional imputed values for each area for the purpose of variance estimation. We call the imputed values constructed in the second step the "variance estimation imputed values," as they are constructed for the purpose of variance estimation. We let R denote the number of variance estimation imputed values. After completing both imputation steps, the final data set has a total of $D(n + R)$ records (rows), where D is the number of areas, n is the number of elements in the original sample, and R is the number of variance estimation imputed values. The size of the resulting data set can be unwieldy if D is very large. In Section 3, we define procedures that require fewer imputed values.

First, we define the prediction imputed values. Conceptually, the imputed data set for area k consists of the n_k elements in domain k and all $n - n_k$ elements that are not in domain k. We now formalize the process of augmenting the sample for area k with the prediction imputed values. We continue with our convention of using the first index label to denote the area that is receiving imputed values (recipients) and the second index label to denote

the area providing the imputed value (donor). Let $\tilde{y}_{kij}$ denote the imputed response value from donor y_{ij} to area k. Elements in area k donate to themselves, such that $\tilde{y}_{kkj} = y_{kj}$. We now define an indexing system for the observed and imputed elements in area k. Arrange the n elements in the original data set in increasing order by area label and then by ascending order based on unit label within area. Then, element y_{ij} is in position $s(ij)$, where $s(ij) = \{\sum_{\ell<i} n_\ell\} + j$, $s(ij) \in \{1, \ldots, n\}$, and it is understood that $\sum_{\ell<1} n_\ell = 0$. Denote the n imputed values for area k by $\check{y}_{k1}, \ldots, \check{y}_{kn}$ where $\check{y}_{ks} = y_{kij}$ for the element in position $s(ij)$. (Note that here, we use $s(ij)$ instead of s_{ij} to emphasize that the position s is a function of the original indexing (ij).) Define δ_{ks} to be the weight associated with imputed value $\check{y}_{ks}$, such that $\delta_{ks} = \delta_{kij}$ for $s = s(ij)$. All unit-level variables in addition to variable y come from the donor. Precisely, let v_{ij} denote the $(q-1)$—dimensional vector of variables of interest in the survey, except for y_{ij}. The imputed value $\check{v}_{ks} = v_{ij}$ for element (ij) in position s. The covariate used in the Fay-Herriot model is an area-level covariate. Therefore, we think the natural choice for the imputed covariate value associated to $(\check{y}_{ks}, \check{v}'_{ks})'$ is x_k. At this stage, the imputed data set has Dn rows. The total number of columns at this step is $p + q + 1$, where p is the dimension of x_k, q is the dimension of $(\check{y}_{ks}, \check{v}'_{ks})'$, and the additional column is for the weight δ_{ks}.

To estimate the MSE of the EBLUP, we estimate M_{1k} and M_{2k} by adding variance estimation imputed values and replicate weights. To estimate M_{1k}, we define R additional imputed values for each area, for a total of DR variance estimation imputed values. We define the R imputed values such that the expected value of the sum of squared differences between the R imputed values and the EBLUP is $M_{1k} = \hat{\gamma}_k \sigma^2_{ek}$. The r^{th} variance estimation imputed value for area k is defined by

$$\check{y}_{k(n+r)} = \hat{\theta}^{EB}_k + \sqrt{D/(D-1)}\,\frac{\hat{\gamma}_k(b^{*(r)}_k + e^{*(r)}_k) - b^{*(r)}_k}{\sqrt{R}}, \tag{7}$$

where $b^{*(r)}_k \overset{iid}{\sim} (0, \hat{\sigma}^2_b)$, $e^{*(r)}_k \overset{iid}{\sim} (0, \hat{\sigma}^2_{ek})$ for $r = 1, \ldots, R$, and $\hat{\theta}^{EB}_k$ is the original EBLUP of (2). Observe that $E_*[D^{-1}(D-1)\sum_{r=1}^{R}(\check{y}_{k(n+r)} - \hat{\theta}^{EB}_k)^2 | \hat{Y}_1, \ldots, \hat{Y}_D] = \hat{\gamma}_k \sigma^2_{ek}$, where $E_*[\,\cdot\,]$ denotes expectation relative to the distribution of $(b^{*(r)}_k, e^{*(r)}_k)$. The R variance estimation imputed values for area k contribute only to variance estimation, not to prediction. The estimation weight $\delta_{kj} = 0$ for $j = n+1, \ldots, n+R$.

At this stage, the imputed data set has a total of $D(n+R)$ records (rows), composed of the $Dn - n$ prediction imputed values, RD variance estimation imputed values, and the original n observations. Let $\check{A}_k = \{j = 1, \ldots, n+R\}$ denote the index set for the observed and imputed elements for area k. By construction, the estimator defined by

$$\hat{\theta}_k = \sum_{j \in \check{A}_k} \delta_{kj} \check{y}_{kj}$$

is the EBLUP of Equation (2). Recall that $\delta_{ks} = \delta_{kij}$ for index (ij) satisfying $s = s(ij)$, and $\delta_{ks} = 0$ for $s = n + r$ for $r = 1, \ldots, R$. The δ_{kij} are defined in Equation (5) to reproduce the EBLUP of Equation (2). The use of δ_{kj} instead of δ_{kij} is a simplification of the notation from a triple-subscript to a double-subscript.

We now define the replicate weights needed for MSE estimation. We define $D + R$ vectors of replicate weights. The first D replicates are used to reproduce the jackknife

estimator (3) of the second term (M_{2k}) in the MSE of the EBLUP. The second R replicates approximate M_{1k}. For $r = 1,\ldots, D$, define

$$\delta_{kij}^{(r)} = \begin{cases} (1 - \hat{\gamma}_k^{(r)}) x_k' A_i^{(r)} w_{ij} + I[i = k]\hat{\gamma}_k^{(r)} w_{ij}, & \text{if } i \neq r \\ I[i = k]\hat{\gamma}_k^{(r)} w_{ij}, & \text{if } i = r \end{cases}$$

where $\hat{\gamma}_k^{(r)} = \hat{\sigma}_{b(r)}^2/(\hat{\sigma}_{b(r)}^2 + \sigma_{ei}^2)$, $A_i^{(r)} = M_{(r)}^{-1} x_i(\sigma_{ei}^2 + \hat{\sigma}_{b(r)}^2)^{-1}$, $M_{(r)} = \sum_{i:i\neq r}(\sigma_{ei}^2 + \hat{\sigma}_{b(r)}^2)^{-1}$ $x_i x_i'$, and $\hat{\sigma}_{b(r)}^2$is the estimator of σ_b^2 obtained with area r omitted. Associate the weight $\delta_{ks}^{(r)} = \delta_{kij}^{(r)}$ to the element with $s = s(ij)$. Set $\delta_{kj}^{(r)} = 0$ for $r = 1,\ldots, D$, and $j = n + 1,\ldots, n + R$, the indexes for the R variance estimation imputed values. We define R additional replicates to obtain a replicate estimate of the first term, $M_{1k} = \gamma_k \sigma_{ek}^2$. For $r = D + 1,\ldots, D + R$, set $\delta_{kj}^{(r)} = I[j > n \text{ and } j - n = r]$. The data set now has $D(n + R)$ rows and $p + q + 1 + D + R$ columns, where q is the dimension of $(y_{ks}, \check{v}_{ks}')'$, p is the dimension of x_k, D is the number of areas, R is the number of variance estimation imputed values, and the additional column is for the estimation weight.

Define the replicate estimate by

$$\hat{\theta}_k^{(r)} = \sum_{j\in\check{A}_k} \delta_{kj}^{(r)} \check{y}_{kj}.$$

For $r = 1,\ldots, D$, $\hat{\theta}_k^{(r)}$ is the predictor of area k that would be obtained if area r were omitted from the process of estimating β and σ_b^2. Therefore, $D^{-1}(\mathrm{D}-1)\sum_{r=1}^{D}\hat{\theta}_k^{(r)} - \hat{\theta}_k)^2$ is the jackknife estimator (3) defined by Jiang et al. (2002). For $r = D + 1,\ldots, D + R$, $\hat{\theta}_k^{(r)} = \check{y}_{k(n+r)}$. We define the MSE estimator by

$$\hat{V}_{area,1,k} = \frac{D-1}{D}\sum_{r=1}^{D+R}\left(\hat{\theta}_k^{(r)} - \hat{\theta}_k\right)^2. \tag{8}$$

One can expand $\hat{V}_{area,1,k}$ as $\hat{V}_{area,1,k} = D^{-1}\ (\mathrm{D}-1)\ [\hat{V}_{area,1,k,M1} + \hat{V}_{area,1,k,M2}]$, where $\hat{V}_{area,1,k,M1} = \sum_{r=D+1}^{D}(\hat{\theta}_k^{(r)} - \hat{\theta}_k)^2$ and $\hat{V}_{area,1,k,M2} = \sum_{r=1}^{D}(\hat{\theta}_k^{(r)} - \hat{\theta}_k)^2$. By construction, $\mathrm{E}_*[\hat{V}_{area,1,k,M1}|\hat{Y}_1,\ldots,\hat{Y}_D] = \hat{\gamma}_k\sigma_{ek}^2$, and $\hat{V}_{area,1,k,M2} = \hat{M}_{2k}^{JK}$, where $\hat{M}_{2k}^{JK}$ is defined in Equation (3). The MSE estimator (8) is similar but not identical to that of Jiang et al. (2002). Jiang et al. (2002) use the analytic form for M_{1k}, while (8) uses a MC approximation. Jiang et at. (2002) incorporate a jackknife estimator of $E[\hat{\gamma}_k - \gamma_k]$. The jack knife estimator (8) omits this correction for the bias of the estimator of M_{1k}. Therefore, the jackknife estimator (8) is not second-order unbiased. The general problem of constructing second-order unbiased MSE estimators is of theoretical and methodological importance in small area estimation literature. Our experience (i.e., Lyu et al 2020) is that MSE estimators with bias of order $O(D^{-1})$ are often adequate for practical purposes if the number of areas is sufficiently large. In the simulations of Section 4, we demonstrate that the MSE estimator (8) tends to slightly over-estimate the MSE of the predictor, even though the estimator (8) is not second-order unbiased.

We use a method of moments estimator of σ_b^2 developed in Fay and Herriot (1979). This procedure is iterative, and the iteration to estimate of may not converge. If a replicate

estimate of σ_b^2 is not defined, then we use the estimate of σ_b^2 from the full data set. We did not encounter convergence problems when working with the full data set.

One must select a value for the number of variance estimation imputed values, R. The choice of R will not impact the bias of the MSE estimator, but increasing R will decrease the variance of the variance estimator. We recommend conducting a sensitivity analysis to determine a value for R. We offer specific choices of R in the simulations and data analysis.

The estimation weight δ_{kj} is in the scale of a mean for an area. One can define a weight in the scale of totals as $\delta_{kj,T} = \delta_{kj}N_k$. Here, N_k is the population size for the area receiving the donors. One can work with δ_{kj} or $\delta_{kj,T}$. We prefer to operate with δ_{kj}. We work in the scale of means because parameters of interest are often means, and we think that means are often more natural to interpret than totals.

In small area estimation, it is common to benchmark the small area estimates to an estimator for the full population. One can enforce benchmarking restrictions by augmenting the vector x_k with an additional covariate, as described in Berg and Fuller (2018) and in Wang et al. (2008). One can then construct benchmarked small area estimates by applying the proposed procedure with the augmented covariate vector.

2.3. Multivariate Response Variables and Area Level Models

The imputation procedure of Subsections 2.1–2.2 is defined for a single response variable y_{kj}. In practice, multiple response variables are often of interest. We now generalize the procedure for a situation where response variables $y_{kj1},\ldots,y_{kjQ}$ are of interest. We assume the same covariate x_k is used for all response variables. We let $\hat{\sigma}_{b,q}^2$ denote the estimator of $\sigma_{b,q}^2$, the between-area variance for the q^{th} response variable. We then implement the imputation procedure using $\hat{\sigma}_b^2 = Q^{-1}\sum_{q=1}^{Q}\hat{\sigma}_{b,q}^2$ and $\sigma_{ek}^2 = Q^{-1}\sum_{q=1}^{Q}\sigma_{eqk}^2$, where σ_{eqk}^2 is the design variance (treated as known) of the direct estimator of the mean for the q^{th} response variable.

We consider the effect of using an incorrect estimate of σ_b^2 to construct the predictor. Consider a predictor of θ_k of the form

$$\hat{\theta}_k(g) = g\hat{Y}_k + (1-g)x_k'\beta,$$

where $0 < g < 1$. The MSE of $\hat{\theta}_k(g)$ for the parameter $\hat{\theta}_k$ is

$$MSE\left(\hat{\theta}_k(g)\right) = E\left[\left(\hat{\theta}_k(g) - \theta_k\right)^2\right] = (g-1)^2\sigma_{bq}^2 + g^2\sigma_{eqk}^2.$$

Straightforward algebra shows that $MSE(\hat{\theta}_k(g)) < \sigma_{eqk}^2$ if and only if $\sigma_{bq}^2/\sigma_{eqk}^2 < (1+g)/(1-g)$. This inequality provides an analyst with a way to check if using the average of the variance components will increase the MSE of the predictor to the point that the predictor is no longer more efficient than the direct estimator. For each variable q, one can compare $(1+g)/(1-g)$ to the ratio $\hat{\sigma}_{bq}^2/\sigma_{eqk}^2$ where $g = \hat{\sigma}_b^2/(\hat{\sigma}_b^2 + \sigma_{ek}^2)$. If $(1+g)/(1-g) < \hat{\sigma}_{bq}^2/\sigma_{eqk}^2$ for some q or k, then one may want to consider a multivariate model.

For multivariate response variables, we use the variable-specific variances to construct the variance estimation imputed values. We use $\hat{\sigma}_{bq}^2$ and σ_{eqk}^2 to define the variance estimation imputed value. Specifically, we define the variance estimation imputed value for variable q by

$$\check{y}_{qk(n+r)} = \hat{\theta}_k^{EB} + \sqrt{D/(D-1)}, \frac{\hat{\gamma}_{k,av}\left(b_{qk}^{*(r)} + e_{qk}^{*(r)}\right) - b_{qk}^{*(r)}}{\sqrt{R}}$$

where $b_{qk}^{*(r)} \overset{ind}{\sim} N(0, \hat{\sigma}_{bq}^2)$, $e_{qk}^{*(r)} \overset{ind}{\sim} N(0, \sigma_{eqk}^2)$, and $\hat{\gamma}_{k,av} = \hat{\sigma}_b^2(\hat{\sigma}_b^2 + \sigma_{ek}^2)^{-1}$.

Approaches that are more sophisticated than using the average of random effect variances to generalize the procedure for a multivariate model exist. Chandra and Chambers (2009) discuss weight adjustments for multivariate response variables. The unit-level framework of Section 3 provides an avenue for multivariate modeling. We illustrate the use of a bivariate unit-level model in the simulations and data analysis.

2.4. *Estimating Linear Combinations of Small Area Parameters*

We use the weights in the intuitive way to construct estimates of linear combinations of small area parameters. Define an estimator of a total for an area by $\hat{\theta}_{k,T} = \sum_{j \in \check{A}_k} N_k \delta_{kj} \check{y}_{kj}$. Express a linear combination of small area means by $\theta_L = \sum_{i=1}^{D} \ell_i \theta_i$, where $\ell_1, \ldots, \ell_D$ are specified constants. The estimate of the linear combination of means is $\hat{\theta}_L = \sum_{i=1}^{D} \ell_i \sum_{j \in \check{A}_i} \delta_{ij} \check{y}_{ij}$. Replicate estimates of the linear combinations are constructed as

$$\hat{\theta}_L^{(r)} = \sum_{i=1}^{D} \ell_i \sum_{j \in \check{A}_i} \delta_{ij}^{(r)} \check{y}_{ij}$$

for $r = 1, \ldots, D+R$. The replicate estimator of the variance for the linear combination is then defined as

$$\hat{V}_R = \frac{D-1}{D} \sum_{r=1}^{D+R} \left(\hat{\theta}_L^{(r)} - \hat{\theta}_L\right)^2.$$

An important linear combination of small areas is the full population mean. To estimate the full population mean, we apply the procedure for linear combinations with $\ell_i = N_i(\sum_{i=1}^{D} N_i)^{-1}$.

3. Unit-Level Predictor as an Imputed Estimator

In Section 2, we defined imputation procedures that reproduce small area estimates under area-level models. We now develop imputation procedures for unit-level small area models, the other main class of small area models. In the unit-level model, the response variable is an individual element instead of a direct estimator for an area.

The seminal unit-level model is the linear, unit-level model. Battese et al. (1988) introduce the linear, unit-level model for the purpose of estimating county crop areas. Similar to the Fay-Herriot model, one can express the EBLUP for the unit-level linear model as a weighted sum of sampled elements. Therefore, one can develop weighting and imputation procedures for the linear unit-level model using procedures completely analogous to the procedures of Section 2.

We pursue a different approach for imputation under the unit level model with two goals. One is to reduce the number of imputed values. The second is to define a procedure that

applies to nonlinear models. We define the procedures with enough generality to encompass nonlinear forms; however, we only implement the procedures for linear models.

We consider a general, unit-level model. In working with the unit-level model, we introduce notation that differs from the notation used for the area-level model. Let q denote the dimension of a vector of unit-level covariates, x_{ij}. Assume that x_{ij} is known for all $j \in U_i$. Let y_{ij} be a p-dimensional response, where $y_{ij} = (y_{1ij}, \ldots, y_{pij})'$. Note that we continue to assume that the same x_{ij} is common to all elements of y_{ij}. Assume that $y_{ij} \in A_i$ is observed through the sample. Define a general unit-level model by

$$y_{ij}|b_i \sim f(y_{ij}|x_{ij}, b_i, \theta), \text{ and } b_i \sim f(b_i|\Sigma_{bb}), \tag{9}$$

where $f_b(b_i \mid \Sigma_{bb})$ denotes the pdf of the distribution of a p—dimensional random effect with mean 0 and covariance matrix Σ_{bb}. A common estimator of (θ, Σ_{bb}) is the maximum likelihood estimator defined as

$$\left(\hat{\theta}, \hat{\Sigma}_{bb}\right) = argmax_{\theta, \Sigma_{bb}} \prod_{i=1}^{D} \int_{R^p} \prod_{j=1}^{n_i} f_y(y_{ij}|x_{ij}, b_i, \theta)\, f_b\ (b_i, \Sigma_{bb}) db_i.$$

The parameter of interest in small area estimation is the finite population mean defined as

$$\bar{y}_{N_i} = \frac{1}{N_i} \sum_{j \in U_i} y_{ij}. \tag{10}$$

The key to prediction under the unit-level model is the observation that the minimum mean square error (or, Bayes) predictor of a non-sampled element y_{ij} is the conditional expectation of y_{ij} given the other observations in area i. The Bayes predictor of y_{ij} for a non-sampled element (i, j) is defined as $\hat{y}_{ij}^B(\theta, \Sigma_{bb}) = (\hat{y}_{1ij}^B(\theta, \Sigma_{bb}), \ldots, \hat{y}_{pij}^B(\theta, \Sigma_{bb}))'$, where

$$\hat{y}_{1ij}^B(\theta, \Sigma_{bb}) = E\left[y_{kij}|y_i; x_{ij}, \theta, \Sigma_{bb}\right] = \int_{R^p} \int_{R^p} y_{kij}\, f\left(y_{ij}|b_i, x_{ij}, \theta, \Sigma_{bb}\right) f\left(b_i|y_i, \theta, \Sigma_{bb}\right) db_i dy_{ij}$$

$$= \frac{\int_{R^p} \int_{R^p} y_{kij} f\left(y_{ij}|x_{ij}, b_i, \theta, \Sigma_{bb}\right) \prod_{\ell=1}^{n_i} f\left(y_{i\ell}|b_i, x_{i\ell}, \theta\right) f\left(b_i|\Sigma_{bb}\right) db_i dy_{ij}}{\int_{R^p} \prod_{\ell=1}^{n_i} f\left(y_{i\ell}|b_i, x_{i\ell}, \theta\right) f\left(b_i|\Sigma_{bb}\right) db_i},$$

and $y_i = \{y_{ij} : j \in A_i\}$. The Bayes predictor of the small area mean in Equation (10) is then

$$\bar{y}_{N_i}^B(\theta, \Sigma_{bb}) = \frac{1}{N_i} \left\{ \sum_{j \in A_i} y_{ij} + \sum_{j \in \bar{A}_i} \hat{y}_{ij}^B\left(\theta, \Sigma_{bb}\right) \right\}, \tag{11}$$

where $\bar{A}_i$ is the index set of elements in U_i but not in A_i. The Bayes predictor $\hat{y}_{ij}^B\left(\theta, \Sigma_{bb}\right)$ is intractable because it is a function of the unknown $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_{bb}$. In practice, one substitutes the unknown $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_{bb}$ defining the Bayes predictor with estimators to obtain the the empirical Bayes predictor (EB). Formally, the EB predictor is defined as

$$\hat{y}_{ij}^{EB} = \hat{y}_{ij}^B\left(\hat{\theta}, \hat{\Sigma}_{bb}\right) = \left(\hat{y}_{1ij}^B\left(\hat{\theta}, \hat{\Sigma}_{bb}\right), \ldots, \hat{y}_{pij}^B\left(\hat{\theta}, \hat{\Sigma}_{bb}\right)\right)', \tag{12}$$

where $\hat{\theta}$ and $\hat{\Sigma}_{bb}$ are appropriately defined estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_{bb}$. The EB predictor of the small area mean is defined as

$$\hat{\bar{y}}_{N_i}^{EB} = \bar{y}_{N_i}^{B}\left(\hat{\theta}, \hat{\Sigma}_{bb}\right) = \frac{1}{N_i}\left\{\sum_{j\in A_i} y_{ij} + \sum_{j\in \bar{A}_i} \hat{y}_{ij}^{EB}\right\}. \tag{13}$$

The MSE of the EB predictor (13) for the unit-level model has a decomposition analogous to the MSE of the EBLUP for the Fay-Herriot model. The MSE of $\hat{\bar{y}}_{N_i}^{EB}$ decomposes into a sum of two terms as

$$mse(\hat{\bar{y}}_{N_i}^{EB}) = V\{\bar{y}_{N_i} + y_i\} + E\left[\left(\hat{\bar{y}}_{N_i}^{EB} - \bar{y}_{N_i}^{B}\left(\theta, \sum\nolimits_{bb}\right)\right)\left(\hat{\bar{y}}_{N_i}^{EB} - \bar{y}_{N_i}^{B}\left(\theta, \sum\nolimits_{bb}\right)\right)'\right]. \tag{14}$$

The first term in the sum of Equation (14) is the MSE of the Bayes predictor (11). The second term accounts for the variance of the estimators $\hat{\theta}$ and $\hat{\sum}_{bb}$. The second term in the MSE decreases as the number of areas increases. We call the first term, $V\{\bar{y}_{Ni}|y_i\}$, the leading term in the MSE. The leading term is typically the dominant term because $V\{\bar{y}_{Ni}|y_i\}$ does not decrease as D increases.

The crux of our imputation procedure is to define imputed values and weights so that weighted sums of imputed values approximate the EB predictor. We define the basic imputation and replication variance estimation procedure in Subsection 3.1. In Subsection 3.2, we define a calibration adjustment that removes the increase in MSE due to random imputation. We consider multivariate response variables in Subsection 3.3. We explain one way to adapt the procedure to account for an informative sample design in Subsection 3.4. In Subsection 3.5, we define estimators and MSE estimators for linear combinations of small area parameters.

3.1. *Basic Imputation and Replication Variance Estimation Procedure for the Unit-Level Model (Unit-Imp)*

The structure of the imputation procedure is similar to fractional imputation (Kim 2011; Kim and Shao 2013). For each nonsampled element (ij), we construct M imputed values and a set of weights. We denote imputed value m for non-sampled unit (ij) by $y_{ij,imp}^{(m)}$. We select $y_{ij,imp}^{(m)}$ independently and with replacement from the full set of sampled elements. Formally, $y_{ij,imp,\ldots,}^{(1)} y_{ij,imp}^{(M)}$ are *iid* multinomial random vectors such that

$$P\left(y_{ij,imp}^{(M)} = y_{ij}\right) = \frac{1}{n}, \tag{15}$$

where $n = \sum_{i=1}^{D}\sum_{j=1}^{n_i} 1$ is the total sample size. The imputed value for the covariate associated with non-sampled element (i, j) is its observed x_{ij}. Element (i, j) inherits all response variables from the donor but keeps its own covariate. We assign a weight to element (i, j) with selected imputed value $y_{ij,imp}^{(M)}$ given by

$$\delta_{ij,m} \propto f\left(y_{ij,imp}^{(m)}|b_i^{(m)}, x_{ij}, \hat{\theta}\right)/f_0\left(y_{ij,imp}^{(m)}\right), \tag{16}$$

where $f_0(y)$ is an estimate of the marginal distribution of $\{y_{ij}: i = 1,\ldots, D; j = 1,\ldots, n_i\}$, $b_i^{(m)} \overset{iid}{\sim} f(b_i|y_i; \hat{\Sigma}_{bb}; \hat{\theta})$, $y_{ij,imp}^{(m)}$ is the m^{th} imputed value selected for the element with covariate value x_{ij}, and $\sum_{m=1}^{M}\delta_{ij,m} = 1$. We divide by $f_0(y)$ because we select the donors from a discrete

uniform distribution that differs from the the marginal distribution of y_{ij}. We give specific choices for $f_0(y)$ in the simulation and data analysis. To create a complete vector of weights, we set $\delta_{ij} = 1$ for $j \in A_i$. The predicted area mean for the vector of response variables is then

$$\hat{\bar{y}}_{N_i} = \frac{1}{N_i}\left\{\sum_{j\in A_i}\delta_{ij}y_{ij} + \sum_{j\in\bar{A}_i}\sum_{m=1}^{M}\delta_{ij,m}y_{ij,imp}^{(m)}\right\}. \tag{17}$$

If the sampling fraction is small, one may construct an imputed value for all elements in the population and define the predictor by

$$\hat{\bar{y}}_{N_{i,\infty}} = \frac{1}{N_i}\sum_{j\in U_i}\sum_{m=1}^{M}\delta_{ij,m}y_{ij,imp}^{(m)}. \tag{18}$$

We use a predictor of the form (18) in the data analysis of Section 5 because the sampling rates in small areas are less than 1%.

Remark 1: In this we propose to select the imputed values with equal probability and then utilize weighting adjustments to account for variation in the conditional distributions. Alternatively, one can select a donor with probability proportional to $\delta_{ij,m}$ and assign equal weight to each donor.

Remark 2: To our knowledge, the imputation procedure defined by Equations (15) and (16) is novel in the small area estimation and missing data literature. The procedure has connections to hot-deck imputation (Kim and Shao 2013, 178–179) and to the fractional imputation procedure of Yang et al. (2013). We do not present a proof that the imputed estimator (17) is an approximation for the EB predictor (13). We justify the imputation procedure of Subsection 3.1 in Appendix (Subsection 7.4). We evaluate the properties of the proposed imputation method through simulation.

We define a replication procedure for estimating the two terms in the MSE (14). The structure of the replication procedure parallels that used for the Fay-Herriot model. We use simulation to approximate the first term in the MSE. We generate variance estimation imputed values from the conditional distribution of the finite population mean given the observed data. The sample variance of the resulting variance estimation imputed values is an approximation for the leading term in the MSE (14). We use the parametric bootstrap to define a set of replicate weights such that a resulting replication variance estimator approximates the second term in the MSE. Lyu et al. (2020) use a parametric bootstrap MSE estimator of this form. The entire procedure involves constructing B variance estimation imputed values and $2B$ sets of replicate weights. The first B replicate weights permit estimation of the leading term in the MSE and the second set is for the variance due to parameter estimation.

We first construct B variance estimation imputed values and associated replicate weights for the purpose of estimating the first term in the MSE (14). For consistency of notation, we create an additional index label for the variance estimation imputed value. We augment the imputed data set for area i with an artificial unit indexed by the pair (ij) with $j = N_i + 1$, where N_i is the population size for area i. We define B variance estimation imputed values denoted by $y_{iN_i+1}^{(m)}$ for $m = 1,\ldots, B$. We generate the mth variance estimation imputed value as

$$y_{iN_i+1}^{(m)} = \bar{y}_{N_i,imp}^{(m)} - \bar{y}_i^{(\cdot)} + \hat{\bar{y}}_{Ni}, \tag{19}$$

where $\bar{y}_{N_i,imp}^{(m)} \sim f(\bar{y}_{Ni}|y_i; \hat{\theta}, \hat{\Sigma}_{bb})$ and $\bar{y}_i^{(\cdot)} = B^{-1}\Sigma_{m=1}^{B}\bar{y}_{N_i,imp}^{(m)}$. Next, assign estimation weights $\delta_{iN_i+1,m}$ for $m = 1, \ldots, B$ to the variance estimation imputed values. We want the variance estimation imputed values to contribute only to the variance estimator and not to the estimator. We therefore set $\delta_{iN_i+1,m} = 0$ for $m = 1, \ldots, B$. We define replicate weights associated with $y_{iN_i+1}^{(m)}$ such that a replicate variance estimator reproduces the sample variance of $\{y_{iN_i+1}^{(m)} : m = 1, \ldots, B\}$. Denote the b^{th} weight associated with the m^{th} imputed value by $\delta_{iNi+1,m}^{(b)}$ for $b = 1, \ldots, B$. We set $\delta_{iNi+1,m}^{(b)} = 1$ if $b = m$ and set $\delta_{iNi+1,m}^{(b)} = 0$ otherwise. To create a rectangular data set, we define $\delta_{ij,m}^{(b)} = 0$ for $j \in \bar{A}_i$, $b = 1, \ldots, B$, and $m = 1, \ldots, M$. We also define $\delta_{ij}^{(b)} = 0$ for $j \in A_i$.

We now define B additional replicate weights for the purpose of estimating the second term in the MSE. Let $\{\hat{\theta}^{(b)}, \hat{\Sigma}_{bb}^{(b)} : b = 1, \ldots, B\}$ be replicate estimates of θ and Σ_{bb} obtained from a parametric bootstrap procedure. (See, for example, Molina and Rao (2010) for an explanation of how to obtain the bootstrap estimates.) For $j \in \bar{A}_i$, define the replicate weight by

$$\delta_{ij,m}^{(B+b)} \propto f\left(y_{ij,imp}^{(m)}|b_{i,m}^{(b)}, x_{ij}, \hat{\theta}^{(b)}\right)/f_0\left(y_{ij,imp}^{(m)}\right),$$

where $b_{i,m}^{(b)} \overset{iid}{\sim} f(b_i|y_i; \hat{\Sigma}_{bb}^{(b)}, \hat{\theta}^{(b)})$ and $\Sigma_{m=1}^{M}\delta_{ij,m}^{(B+b)} = 1$ for $b = 1, \ldots, B$. To complete the rectangular set of replicate weights, define $\delta_{ij}^{(B+b)} = 1$ for $j \in A_i$ and $\delta_{iN_i+1,m}^{(B+b)} = 0$ for $b = 1, \ldots, B$.

We now have $2B$ sets of replicate weights. The replicate estimator of the MSE for area i is

$$\hat{MSE}_{uni,i,rep} = B^{-1}\sum_{b=1}^{2B}\left(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{Ni}\right)\left(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{Ni}\right)', \tag{20}$$

where

$$\hat{\bar{y}}_{N_i}^{(b)} = \frac{1}{N_i}\left\{\left(\sum_{j\in A_i}\delta_{ij}^{(b)}y_{ij}\right) + \left(\sum_{j\in\bar{A}_i}\sum_{m=1}^{M}\delta_{ij,m}^{(b)}y_{ij,imp}^{(m)}\right) + \left(\sum_{m=1}^{B}\delta_{iN_i+1,m}^{(b)}y_{iN_{i+1}}^{(m)}\right)\right\}.$$

The MSE estimator decomposes into a sum of two terms as $\hat{MSE}_{uni,i,rep} = V_{1i} + V_{2i}$, where $V_{1i} = B^{-1}\sum_{b=1}^{B}(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{N_i})(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{N_i})'$, and $V_{2i} = B^{-1}\sum_{b=B+1}^{2B}(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{N_i})(\hat{\bar{y}}_{N_i}^{(b)} - \hat{\bar{y}}_{N_i})'$. For $b = 1, \ldots, B$, $\hat{\bar{y}}_{N_i}^{(b)} = \sum_{m=1}^{B}\delta_{iN_i+1,m}^{(b)}y_{iN_i+1}^{(m)} = y_{iN_i+1}^{(b)} = y_{N_iimp}^{(b)} - y_i^{(\cdot)} + \hat{\bar{y}}_{N_i}$. Therefore, $V_{1i} = B^{-1}\sum_{b=1}^{B}(\bar{y}_{N_iimp}^{(b)} - \bar{y}_i^{(\cdot)})(\bar{y}_{N_iimp}^{(b)} - y_i^{(\cdot)})'$. For $b = B+1, \ldots, 2B$, $\hat{\bar{y}}_{N_i}^{(b)} = N_i^{-1}\{(\sum_{j\in A_i}\delta_{ij}^{(b)}y_{ij}) + (\sum_{j\in\bar{A}_i}\sum_{m=1}^{M}\delta_{ij,m}^{(b)}y_{ij,imp}^{(m)})\}$. Therefore, V_{2i} is a generalization of the parametric bootstrap estimator of the second term in the MSE of Equation (14) used in Lyu et al.(2020).

To clarify the result of imputation and replicate weight construction, we state the dimensions of the data set at this stage. The number of rows in the data set is now $n + M(N - n) + D(B)$. To establish the number of columns, assume that the number of variables is $q + p$, where q is the dimension of x_{ij} and p is the dimension of y_{ij}. The number of columns in the data set is then $1 + 2B + q + p$, where the 1 is from the estimation weight $\delta_{ij,m}$, and the $2B$ is from the replicate weights.

In our simulations and data analysis, we do not construct the rectangular database for the unit-level procedures. We construct the imputed values, the non-zero weights, and the non-zero replicate weights. We do not construct the zero weights. We calculate the predictors as in (17). We calculate the MSE estimates as in (20) with $\hat{\bar{y}}_{N_i}^{(b)} = y_{N_i+1}^{(b)}$ for $b = 1, \ldots, B$, and $\hat{\bar{y}}_{N_i}^{(b)} = N_i^{-1} \{\sum_{j \in A_i} y_{ij} + \sum_{j \in \bar{A}_i} \sum_{m=1}^{M} \delta_{ij,m}^{(b)} y_{ij,imp}^{(m)}\}$ for $b = B+1, \ldots, 2B$.

One must select values for B and for M. The choices of B and M will impact the variance of the estimator and MSE estimator but will not impact the bias. The value for B is essentially the bootstrap sample size. We recommend taking B to be as large as possible, subject to computational and data storage restrictions. In the simulations, we use $B = 20$. In a study of the double-bootstrap, Erciulescu and Fuller (2016) consider both 200 and 100 for the bootstrap sample size. In Subsection 3.2 below, we define a calibration procedure that permits use of a relatively small value for M.

3.2. Adjustment of Unit Level Weights to Improve Eciency (Unit-Imp-Adj)

The weighted sum of randomly generated imputed values is a Monte Carlo (MC) mean. Two implementations of the imputation procedure will yield different predictors due to the MC variance. Eliminating the increase in MSE from random imputation is desirable for the purposes of improving consistency across implementations of the procedure and for improving the efficiency of the predictor. Increasing the number of imputed values reduces the imputation variance but also increases the computational demands and requirements for storage space. One way to remove the imputation variance is to use calibration to force the weighted sum of imputed values to equal the EB predictor (Kim and Shao 2013, 89). We adopt the calibration approach of Kim and Shao (2013) to eliminate the imputation variance without increasing the number of imputed values.

We obtain adjusted weights $\delta_{ij,m}^{adj}$ satisfying $\sum_{m=1}^{M} \delta_{ij,m}^{adj} y_{ij,imp}^{(m)} = \hat{y}_{ij}^{EB}$, where $\hat{y}_{ij}^{EB}$ the EB predictor (12). The adjusted weight is defined

$$\delta_{ij,m}^{adj} = \delta_{ij,m} \left[1 + \left\{ \left(1, \left(\hat{y}_{ij}^{EB} \right)' \right) - \left(1, \left(\hat{y}_{ij,imp}^{init} \right)' \right) \right\} M_{ij,imp,w}^{-} m_{ij,imp,m} \right],$$

where

$$\hat{y}_{ij,imp}^{init} = \sum_{m=1}^{M} \delta_{ij,m} y_{ij,imp}^{(m)}, \; m_{ij,imp,m} = \left(1, \left(y_{ij,imp}^{(m)} \right)' \right)', \; M_{ij,imp,w}$$

$$= \sum_{m=1}^{M} \delta_{ij,m} m_{ij,imp,m} m_{ij,imp,m}'$$

and

$$M_{ij,imp,w}^{-} = \begin{cases} M_{ij,imp,w}^{-} & \text{if the inverse exists} \\ \left(M_{ij,imp,w} + 0.01 I \right)^{-1} & \text{otherwise,} \end{cases} \tag{21}$$

where $\boldsymbol{I}$ is an identity matrix with the same dimensions as $\mathbf{M}_{ij,imp,w}$. The use of the soft calibration operator in the second line of Equation (21) averts computational problems but may result in a situation where the calibration constraint is not exactly satisfied. We use 0.01 for the tuning parameter because this value had little impact on the results of our simulations and data analysis. In a different application of the proposed procedures, the analyst may want to investigate different values for the tuning parameter. The small area predictor based on the adjusted weight is

$$\hat{\bar{y}}_{N_i} = \frac{1}{N_i}\left\{\sum_{j\in A_i} y_{ij} + \sum_{j\in \bar{A}_i}\sum_{m=1}^{M} \delta_{ij,m}^{adj} y_{ij,imp}^{(m)}\right\}$$

or

$$\hat{\bar{y}}_{N_i\infty} = \frac{1}{N_i}\sum_{j\in U_i}\sum_{m=1}^{M} \delta_{ij,m}^{adj} y_{ij,imp}^{(m)}$$

if the finite population correction factor is ignored.

We modify the replication variance estimator to account for the decrease in MSE due to the calibration step. The variance estimation imputed values remain the same as in Section 3.1. The first B sets of replicate weights also remain unchanged. The $\delta_{iN_i+1,m}^{(b)}$ for $m = 1,\ldots, B$ and $b = B+1,\ldots, 2B$ remain zero, as in Subsection 3.1. As in Subsection 3.2, $\delta_{ij}^{(b)} = 1$ for $j \in A_i$ and $b = B+1,\ldots, 2B$. We calibrate the replicate weights $\{\delta_{ij,m}^{(b)}: m = 1,\ldots, M;\ b = B+1,\ldots, 2B\}$. We modify $\delta_{ij,m}^{(b)}$ as

$$\delta_{ij,m}^{(b)adj} = \delta_{ij,m}^{(b)}\left[1 + \left\{\left(1, \left(\hat{y}_{ij}^{EB(b)}\right)'\right) - \left(1, \hat{y}_{ij,imp}^{init(b)}\right)\right\}\left[M_{ij,imp,w}^{(b)}\right]^{-} m_{ij,imp,m}\right],$$

where

$$\hat{y}_{ij}^{EB(b)} = E\left[y_{ij}|y_i; x_{ij}, \hat{\theta}^{(b)}, \hat{\sum}_{bb}^{(b)}\right], \hat{y}_{ij,imp}^{init(b)} = \sum\nolimits_{m=1}^{M} \delta_{ij,m}^{(b)} y_{ij,imp}^{(m)},$$

$$m_{ij,imp,m} = \left(1, \left(\hat{y}_{ij,imp}^{(m)}\right)'\right)',$$

and

$$M_{ij,imp,w}^{(b)} = \sum_{m=1}^{M} \delta_{ij,m}^{(b)} m_{ij,imp,m} m_{ij,imp,m}'.$$

3.3. Multivariate Response Variables (Unit-Imp-Adj-Aug)

The unit-level model (9) is general enough to allow for a multivariate vector of response variables. The definition of the weight also accounts for the joint distribution of a vector valued response. In this sense, the proposed imputation procedure for the unit-level model automatically allows for multivariate response variables.

In practice, the analyst may prefer to utilize separate univariate unit-level models. Estimation for a univariate model can be simpler than estimation for a multivariate model. Different response variables may be uncorrelated. The additional complexity from multivariate modeling may be judged not worth the effort if the dependence among different response variables is not important.

We define a simple procedure for a situation in which the analyst chooses to adopt separate univariate models for two response variables. The operation is similar in structure to the calibration procedure of Subection 3.2 but serves a different purpose. We propose to use calibration to generate a set of weights that preserve the EB predictors based on two separate univariate models for two response variables.

Suppose estimates are of interest for a pair of response variables, (y_{1ij}, y_{2ij}). First, construct imputed values and weights for y_{1ij} according to the procedure of Subsection 3.1.

Denote the weight by $\delta_{ij,m}$. Then, select the imputed value for y_{2ij} from the same donor that provides the imputed value for y_{1ij}.

We define an augmented adjusted weight by extending the calibration procedure to preserve small area means for y_{2ij}. Let $\hat{y}_{2ij}^{EB}$ be the EB predictor of a non-sampled y_{2ij} based on a specified univariate model. Let $\hat{y}_{2ij,imp}^{init} = \sum_{m=1}^{M} \delta_{ij,m} y_{2ij,imp}^{(m)}$. We define an augmented adjusted weight by

$$\delta_{ij,m}^{adj} = \delta_{ij,m}\left(1 + \left(\left(1, \hat{y}_{1ij}^{EB}, \hat{y}_{2ij}^{EB}\right) - \left(1, \hat{y}_{1ij,imp}^{init}, \hat{y}_{2ij,imp}^{init}\right)\right) M_{ij,imp,aug}^{-} m_{ij,imp-aug,m}\right),$$

where $m_{ij,imp-aug,m} = (1, y_{1ij,imp}^{(m)}, y_{2ij,imp}^{(m)})'$ and $M_{ij,imp,aug} = \sum_{m=1}^{M} m_{ij,imp-aug,m}$ $\delta_{ij,m}$ $m'_{ij,imp,aug,m}$. The modification to the replicates is analogous to that of Subsection 3.2. We call the resulting predictor "Unit-Imp-Adj-Aug."

3.4. Informative Sampling

We next consider the case in which the sample design is informative for the unit-level model. We adopt the approach of Pfeffermann and Sverchkov (2007), focusing on the case in which all areas are selected. The imputation procedure is the same as that of Subsection 3.1. The only difference is that the weights are based on the complement distribution defined as $f(y_{ij} \mid b_i, x_{ij}, I_{ij} = 0)$, where $b_i \overset{iid}{\sim} f(b_i|y_i)$ and I_{ij}. is the sample inclusion indicator for unit j in area i. In Appendix (Subsection 7.2), we spell out the complement distribution for the unit-level linear model with normally distributed errors and the specific model for the sampling weights used in Pfeffermann and Sverchkov (2007).

3.5. Linear Combinations of Small Area Parameters

Given imputed values and weights, the procedure to estimate linear combinations of small area parameters is intuitive. The method is analogous to that of Subsection 2.4 for the area level model. Estimates of linear combinations of small area parameters are appropriately defined weighted sums of observed an imputed values. Replicate estimates are appropriately defined aggregates obtained by replacing estimation weights with replicate weights. Although the methods are intuitive, we define them specifically in this section. For simplicity, we consider a univariate y_{ij}.

Let the linear combination of small area parameters to predict be defined by

$$\bar{y}_L = \sum_{i=1}^{D} \ell_i \sum_{j=1}^{N_i} y_{ij},$$

where $\ell_1, \ldots, \ell_D$ are specified constants. Let $\{w_{ij,m} : j \in \bar{A}_i$, and $j = N_i + 1\}$ be the weights from one of the imputation methods defined in Subsections 3.1–3.4. The predictor of $\bar{y}_L$ is

$$\hat{\bar{y}}_L = \sum_{i=1}^{D} \ell_i \left\{ \sum_{j \in A_i} y_{ij} + \sum_{j \in \bar{A}_i} \sum_{m=1}^{M} w_{ij,m} y_{ij,imp}^{(m)} + \sum_{m=1}^{B} w_{iN_i+1,m} y_{iN_{i+1}}^{(m)} \right\}. \tag{22}$$

If the finite population correction factor is ignored, then sums over the sample and complement index sets are replaced by sums over the imputed values for the full population. The replication variance estimator is defined as

$$\hat{V}_{rep}\left(\hat{\bar{y}}_L\right) = B^{-1}\sum_{b=1}^{2B}\left(\hat{\bar{y}}_L^{(b)} - \hat{\bar{y}}_L\right)^2,$$

where $\hat{\bar{y}}_L^{(b)}$ has the form of Equation (22) with the b^{th} set of replicate weights in place of $w_{ij,m}$.

4. Simulation Study

We use a bivariate model for the simulation. The bivariate model reflects the structure of the model used for the data analysis in Section 5. The use of a bivariate model also permits us to assess the properties of the procedure for multivariate response variables. We define a simulation framework with a single covariate $x_{ij.}$ and two response variables y_{1ij} and y_{2ij}. The simulation model is

$$(y_{1ij}, y_{2ij})' = B'x_{ij} + u_i + e_{ij}, \tag{23}$$

where

$$B = \begin{pmatrix} \beta_{01} & \beta_{11} \\ \beta_{02} & \beta_{12} \end{pmatrix},$$

$$x_{ij} = \left(1, x_{ij}\right)', x_{ij} \overset{iid}{\sim} N\left(3, 1.6\right), u_{ij} \overset{iid}{\sim} BVN\left(0, \Sigma_{uu}\right), e_{ij} \overset{iid}{\sim} BVN\left(0, \Sigma_{ee}\right),$$

$$\Sigma_{uu} = \text{diag}\left(\sigma_{u1}, \sigma_{u2}\right)\begin{pmatrix} 1 & \rho_u \\ \rho_u & 1 \end{pmatrix}\text{diag}\left(\sigma_{u1}, \sigma_{u2}\right)$$

and

$$\Sigma_{ee} = \text{diag}\left(\sigma_{e1}, \sigma_{e2}\right)\begin{pmatrix} 1 & \rho_e \\ \rho_e & 1 \end{pmatrix}\text{diag}\left(\sigma_{e1}, \sigma_{e2}\right).$$

The population has ten areas, five with population size $N_{\text{i}} = 1{,}000$ and five with population size $N_i = 500$. We set $(\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma^2_{u1}, \sigma^2_{u2}, \rho_u, \sigma^2_{e1}, \sigma^2_{e2}, \rho_e)$, = (-1.50, 1.00, 1.50, 0.50, 0.50, 0.75, 0.20, 0.50, 0.75, 0.10). The sample design is stratified sampling with areas as strata and a sampling rate of 1%.

4.1. Comparison of Alternative Procedures

We implement the area-level imputation procedure that is called "Area-Imp" and is defined in Subsection 2.2. The Area-Imp procedure begins with a weight of the form $w_{ij} = N_i^{-1}\pi_{ij}^{-1}$. This weight simplifies to $w_{ij} = n_i^{-1}$ for the stratified simple random sample design used for the simulation. An application of a Fay-Herriot area level model to each of the two response variables separately furnishes two estimates of σ^2_b. We use the average of these estimates for the imputation procedures, as explained in Subsection 2.3. If we instead

applied the Area-Imp procedure to each response separately, Area-Imp would reproduce the EBLUP for the Fay-Herriot model.

We consider unit-level procedures based on univariate and bivariate models. For the univariate models, we use y_1 as the univariate response variable. We take the corresponding value of y_2 from the donor. We first implement the procedure of Subsection 3.1, where the weights are constructed under a univariate model for y_1. We refer to this univariate version of the unit-level procedure of Subsection 3.1 as "Unit-Imp." For the Unit-Imp procedure, we approximate $f_0(y)$ by the density of a normal distribution with mean $\bar{y}_{1,n} = n^{-1}\sum_{i=1}^{D}\sum_{j=1}^{n_i} y_{1ij}$ and variance $S^2_{1,n} = (n-1)^{-1}\sum_{i=1}^{D}\sum_{j=1}^{n_i}(y_{1ij} - \bar{y}_{1,n})^2$. To remove the increase in MSE from random imputation, we implement the adjusted unit-level procedure of Subsection 3.2. We refer to this univariate version of the estimator of Subsection 3.2 as "Unit-Imp-Adj." We implement the augmented calibration procedure of Subsection 3.3 to improve predictions for y_2, and we refer to this procedure as "Unit-Imp-Adj-Aug." We also implement bivariate versions of the methods of Subsections 3.1 and 3.2. For the bivariate case, the model for imputation is the bivariate model used to generate the data. For the bivariate version of the imputation method of Subsection 3.1, we take $f_0(y_1,y_2)$ to be the density of a bivariate normal distribution with mean vector $(\bar{y}_{1,n}, \bar{y}_{2,n})$ and covariance matrix S_n, where $\bar{y}_{2,n} = n^{-1}\sum_{k=1}^{D}\sum_{j=1}^{n_i} y_{2ij}$, and $S_n = (n-1)^{-1}\sum_{i=1}^{D}\sum_{j=1}^{n_i}((y_{1ij}, y_{2ij})' - (\bar{y}_{1,n}, \bar{y}_{2,n})')((y_{1ij}, y_{2ij})' - (\bar{y}_{1,n}, \bar{y}_{2,n})')'$. We explain how we estimate the parameters of the bivariate model in the Appendix (Subsection 7.1). We label the bivariate estimator with no adjustment as Unit-Imp-Biv. We label the bivariate estimator that incorporates the adjustment of Subsection 3.2 as "Unit-Imp-Biv-Adj."

We compare the imputed estimators to three other estimators. Area-EBLUP is the EBLUP based on a univariate Fay-Herriot model fit to each response variable separately. As explained previously, Area-EBLUP would equal Area-Imp if separate estimators of σ_b^2 (instead of the average) were used for Area-Imp. The second predictor is the estimator of Chandra and Chambers (2009). The Chandra and Chambers procedure begins with an expression of the EBLUP for the overall population mean under the unit-level linear model as a weighted sum of sampled units. These weights for the total are then applied to construct estimates for arbitrary domains. The basic procedure of Chandra and Chambers is univariate in that a different set of weights is produced for each response variable. Chandra and Chambers (2009) define a multivariate adaptation of their procedure. For simplicity, our implementation of Chandra and Chambers constructs two separate sets of weights, each based on a univariate model for y_1 or y_2. We refer to the estimators based on the Chandra and Chambers procedure as "Unit-Weights-CC." We also report results for a "direct" estimator defined as the sample mean for the area.

Table 1 contains the average MC MSE's of the alternative predictors of the small area means as well as the MC MSE's of the estimators of the overall population mean (y_{pop}). The prefix of the name of a method indicates if the method uses unit-level or area-level information. The Appendix (Subsection 7.3) contains a table with the name of each imputation procedure and a summary of how the imputation procedure works. Estimates of the overall population means are defined as in Subsections 2.4 and 3.5, respectively, for the area-level and unit-level procedures. The estimate of the overall population mean based on the EBLUP is $\hat{y}_{pop}^{EBLUP} = \sum_{i=1}^{D} w_{i,N}\hat{\theta}_i^{EBLUP}$, where $w_{i,N} = N^{-1}N_i$, $N = \sum_{i=1}^{D} N_i$,

Table 1. Average MC MSE's of alternative small area predictors and estimators of overall population mean (y_{pop}).

Method	y_1			y_2		
	$n_i = 5$	$n_i = 10$	y_{pop}	$n_i = 5$	$n_i = 10$	y_{pop}
Area-Imp	0.313	0.173	0.029	0.198	0.107	0.012
Area-EBLUP	0.309	0.180	0.029	0.186	0.100	0.012
Unit-Weights-CC	0.416	0.209	0.007	0.221	0.104	0.008
Unit-Imp	0.095	0.050	0.007	0.723	0.657	0.009
Unit-Imp-Adj	0.089	0.046	0.007	0.732	0.656	0.009
Unit-Imp-Adj-Aug	0.089	0.046	0.007	0.119	0.065	0.007
Unit-Imp-Biv	0.098	0.051	0.007	0.141	0.077	0.008
Unit-Imp-Biv-Adj	0.091	0.046	0.007	0.123	0.066	0.008
Direct	0.430	0.224	0.029	0.218	0.106	0.012

and $\hat{\theta}_i^{EBLUP}$ is the EBLUP of θ_i from the Fay-Herriot model. The MC sample size is 100, and the averages are across areas with the same sample size.

The Unit-Weights-CC method is efficient for the overall population mean. This is expected because the Unit-Weights-CC weights return the EBLUP of the overall population mean. The Unit-Weights-CC estimator is more efficient than the direct estimator but less efficient than Area-EBLUP for the small areas. We do not consider the Unit-Weights-CC method further.

We next consider Area-Imp. As expected, Area-Imp has nearly the same efficiency as Area-EBLUP. The difference between Area-Imp and Area-EBLUP is that Area-Imp uses the average of the two estimates of the variance components, while Area-EBLUP uses the variable-specific variance components. An investigation not reported here confirms that the Area-Imp procedure returns the standard EBLUP if the same estimator of σ_b^2 used for imputation is used for the EBLUP. Use of the average of the two estimates of the variance components has variable effects on the properties of the predictors. For y_1, Area-Imp is more efficient than Area-EBLUP when $n_{\text{i}} = 10$. For y_2, the use of the average causes Area-Imp to be less efficient than the direct estimator for $n_i = 10$.

The unit-level EBLUPs for y_1 are more efficient than the area level predictors for y_1. This is not surprising (Chen 2019) because the error variance for y_1 for the area level model is $n_i^{-1}\left(\beta_{11}^2\sigma_x^2 + \sigma_{e1}^2\right)$, while the error variance for y_1 for the unit-level model is $\sigma_{e1}^2 n_i^{-1}$. The finding that the unit-level model is more efficient than the area-level model is also consistent with results in Hidiroglou and You (2016) and Mauro et al. (2017). The adjustments to remove the imputation variance (Unit-Imp-Adj) improve the unit-level predictors of the small area means slightly. It is disappointing that carrying y_2 from the donor that supplied y_1 leads to a substantial loss of efficiency for the small area mean of y_2, as seen in the rows labeled Unit-Imp and Unit-Imp-Adj. The additional calibration to the EBLUP of y_2 (Unit-1-Adj-Aug) improves the efficiency for y_2. The difference between Unit-Imp-Biv and Unit-Imp-Biv-Adj is that Unit-Imp-Biv-Adj incorporates the adjustment to remove the imputation variance. The use of random imputation has only a minor effect on the efficiency of the predictor. We suspect that the effect of random imputation is more important for y_2 than for y_1 because $V\{u_{2i}\} > V\{u_{1i}\}$ and $V\{e_{2ij}\} >$

$V\{e_{1ij}\}$. The bivariate model predictors have nearly the same efficiency as the predictors Unit-Imp-Adj-Aug. For this simulation configuration, estimating the additional correlation parameters ρ_u and ρ_e leads to a slight increase in the MSE.

Table 2 contains the MC means of the replication MSE estimators for the imputed predictors. We include only the predictors that performed the best among the unit-level and area-level procedures considered in Table 1. We use $R = 30$ to construct the replicates for the area-level model and a bootstrap sample size of $B = 20$ for the unit-level procedures. The MSE estimator for the Area-Imp procedure over-estimates the MSE for y_2. In results not presented here, we have observed this same phenomenon with the standard jackknife MSE estimator for the EBLUP. The MSE estimator for Unit-Imp-Biv under-estimates the MSE for the small areas because the MSE estimator does not incorporate an estimate of the imputation variance. We consider the mean square error estimators for the unit-level predictors of the small area means to be reasonable approximations to the mean square errors of the predictors.

4.2. *Linear Combinations of Area Means*

Equipped with a complete data set and a set of weights, the analyst has the capability to construct estimates for any specified domain of interest. We consider linear combinations of domain means. We use the estimation and variance estimation procedures defined in Subsections 2.4 and 3.5 for area-level and unit-level models, respectively. For variables $q \in \{1, 2\}$, we define the parameters of interest by

$$\bar{y}_{q,126} = \frac{N_1\bar{y}_{q,N_1} + N_2\bar{y}_{q,N_2} + N_6\bar{y}_{q,N_6}}{N_1 + N_2 + N_6} \text{ and by } \bar{y}_{q,rest} = \frac{\Sigma_{i\notin\{1,2,6\}}N_i\bar{y}_{q,N_i}}{\Sigma_{i\notin\{1,2,6\}}N_i}.$$

We construct estimates of $\bar{y}_{q,126}$ and of $\bar{y}_{q,rest}$ using the weights and imputed values constructed with the "Area-Imp," "Unit-Imp-Adj-Aug," "Unit-Imp-Biv," and "Unit-Imp-Biv-Adj" procedures. We compare the MC MSE's of the imputation-based estimates to the MC MSE's of direct estimates in Table 3. The sample sizes are larger for the combined domains than for the individual small areas. Therefore, the MSE's in Table 3 are smaller than the MSE's for the small areas in Table 1. The Area-Imp predictor is slightly more efficient than the direct estimator for y_1 and has efficiency close to the direct estimator for y_2. The increase in MSE from random imputation for the linear combinations is noticeable, as for the individual small areas. With the calibration adjustment, the MSE based on the bivariate model is nearly the same as the MSE based on the univariate model.

Table 2. MC means of MSE estimators for imputed small area predictors and overall population means.

	y_1			y_2		
Method	$n_i = 5$	$n_i = 10$	y_{pop}	$n_i = 5$	$n_i = 10$	y_{pop}
Area-Imp	0.289	0.171	0.027	0.264	0.131	0.024
Unit-Imp-Adj-Aug	0.085	0.047	0.007	0.135	0.072	0.011
Unit-Imp-Biv	0.084	0.047	0.012	0.131	0.073	0.012
Unit-Imp-Biv-Adj	0.082	0.045	0.007	0.129	0.071	0.011

Table 3. MC MSE's of imputation based estimators and of direct estimators of specified linear combinations of small area parameters.

Method	$y_1(q=1)$		$y_2(q=2)$	
	$\bar{y}_{126}$	$\bar{y}_{rest}$	$\bar{y}_{126}$	$\bar{y}_{rest}$
Area-Imp	0.067	0.043	0.035	0.021
Unit-Imp-Adj-Aug	0.016	0.010	0.023	0.013
Unit-Imp-Biv	0.018	0.012	0.031	0.015
Unit-Imp-Biv-Adj	0.016	0.010	0.024	0.014
Direct	0.081	0.046	0.034	0.021

The MC means of MSE estimates are presented in Table 4. The area-level procedure tends to over-estimate the MSE for the linear combination. The estimated MSE for the unit-level predictor constructed under the univariate model for y_1 over-estimates the MSE for y_2. Use of the bivariate model rectifies this problem.

4.3. Simulation for an Informative Sample Design

We simplify the simulation model to a univariate model for the purpose of illustrating the procedures under informative sampling. We generate y_{1ij} as

$$y_{1ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij},$$

where $\beta_0 = -1.5$, $\beta_1 = -1$, $u_i \overset{iid}{\sim} N(0, 0.25)$ and $e_{ij} \overset{iid}{\sim} N(0,1)$ for $i = 1, \ldots, D$ and $j = 1, \ldots, N_i$. We use D $=$ 20 areas, of which 10 have $N_i = 2,000$ and the other 10 have $N_i = 3,000$. We generate x_{ij} as for simulation model (23). The sample design is Poisson sampling with

$$\pi_{ij} \propto 0.01 N_i \exp(0.2(y_{1ij} - \beta_0 - x_{ij}\beta_1) + 0.2\delta_{ij}),$$

where $\delta_{ij} \overset{iid}{\sim} N(0,1)$ independent of $(x_{ij}, y_{ij})'$ and $\sum_{j=1}^{N_i} \pi_{ij} = 0.01 N_i$. The MC sample size is 100.

We construct small area predictors using Area-Imp and using the univariate unit-level model for y_{1ij}. For Area-Imp, we set $w_{ij} = N_i^{-1}\pi_{ij}^{-1}$. We modify the weights for the unit-level model to account for informative sampling using the procedure of Subsection 3.4 with additional detail provided in the Appendix (Subsection 7.2). For the informative sample design, we approximate $f_0(y)$ as a discrete uniform distribution over the set $\{y_{1ij} : i = 1, \ldots, D, j \in A_i\}$. Table 5 compares the average MC MSE's of the Area-Imp,

Table 4. MC means of estimated MSEs of imputation based estimators of specified linear combinations of small area parameters.

Method	$y_1(q=1)$		$y_2(q=2)$	
	$\bar{y}_{126}$	$\bar{y}_{rest}$	$\bar{y}_{126}$	$\bar{y}_{rest}$
Area-Imp	0.076	0.042	0.059	0.037
Unit-Imp-Adj-Aug	0.019	0.011	0.323	0.088
Unit-Imp-Biv	0.020	0.010	0.030	0.016
Unit-Imp-Biv-Adj	0.019	0.010	0.029	0.016

Table 5. MC MSE's of predictors for the informative sample design.

	y_1	
	$n_i = 20$	$n_i = 30$
Direct	0.216	0.164
Area-Imp	0.140	0.114
Unit-Imp	0.068	0.051
Unit-Imp-Adj	0.045	0.032
Unit-EBLUP-SRS	0.087	0.072

Unit-Imp, and Unit-Imp-Adj-Aug predictors to the MC MSE's of two alternative predictors. The direct estimator is the Hajek estimator of the small area mean. We also calculate the EBLUP for the unit-level model under the incorrect assumption of a noninformative sample design. We call the EBLUP that ignores the design Unit-EBLUP-SRS.

The results are as expected. The Area-Imp procedure improves upon the efficiency of the direct estimator and unit-level modeling further improves upon the efficiency of the Area- Imp procedure. The increase in MSE from random imputation is more important in Table 5 than in Table 1 because we use the discrete uniform distribution instead of the normal distribution for $f_0(y)$. Ignoring the design leads to a biased estimator of the intercept. As a result, the MSE of Unit-EBLUP-SRS exceeds the MSEs of Unit-Imp and Unit-Imp-Adj. This simulation study illustrates estimation for an informative design using a simple, univariate model. In future research, we may expand the study for informative sampling to include a vector of response variables and MSE estimation.

5. Application to Seat-Belt Survey Data

We construct a record-level database that supports small area estimation using data from the 2019 Iowa Seat-Belt Use Survey. The survey produces state-level estimates of the proportions of occupants, drivers, and passengers who wear a seat-belt. The sample for the Iowa Seat-Belt Use Survey is a two-stage sample, where primary sampling units are 15 counties and secondary sampling units are road segments within counties. The secondary sampling units are selected according to a stratified two-stage design. The strata are three road types, defined as primary, secondary, and local. The size measure for probability proportional to size (PPS) sampling is the vehicle miles traveled (VMT) for the road segment. The road type and VMT are obtained from the 2010 Iowa Department of Transportation (DOT) Geographic Information Management System. The precise definition of VMT is as follows: "VMT − A measure of total vehicle activity. For a given segment of roadway, the VMT is obtained by multiplying the annual average daily travel (AADT) by the length of the roadway segment. For example, on a 5-mile highway segment traveled by 5,000 vehicles daily (an average obtained over a year), the VMT would be 25,000." (DOT 2010). In the survey, each sampled road segment is observed over a 45-minute period during which data collectors record the number of passengers and drivers who are wearing or not wearing a seat-belt. The survey is designed with the aim of obtaining a standard error for the overall estimate of the state-level seat-belt use rate that does not exceed 2.5%. County estimates are not produced as part of the standard operation.

We produce a database that provides county-level estimates of the proportions of belted drivers and total vehicle occupants. We restrict attention to the 15 sampled counties. We specify a bivariate unit-level small area model. We use the method of Subsection 3.1 to define imputed values and weights for every road segment in the population. The weighted sum of the imputed values approximates the small area estimates based on the bivariate model.

5.1. Model for Small Area Estimates

We estimate the county proportions of drivers and total occupants who are wearing a seat-belt. Let $\hat{p}_{dij}$ and $\hat{p}_{tij}$, respectively, be the proportions of drivers and total occupants on road segment j of county i who are wearing a seat-belt. The population parameters of interest are defined as

$$\left(\theta_{di}, \theta_{ti}\right) = \left(\sum_{j=1}^{N_i} v_{ij}\right)^{-1} \sum_{j=1}^{N_i} v_{ij}\left(\hat{p}_{dij}, \hat{p}_{tij}\right),$$

where v_{ij} is the VMT for road segment j of county i. We only observe $\left(\hat{p}_{dij}, \hat{p}_{tij}\right)$ for the road segments in the sample. The VMT (v_{ij}) and the road type (primary, secondary, and local) are known for all road segments in the population. Therefore, the VMT and road type are possible covariates to use in a small area model for $(\hat{p}_{dij}, \hat{p}_{tij})$.

To motivate a model form, we conduct an exploratory analysis. We fit a preliminary model with $(\hat{p}_{dij}, \hat{p}_{tij})$ as the response and road type as the covariate. The coefficient for VMT is not significant in a model that contains road type. The residuals from the preliminary model violate the normality assumption. To improve the normal approximation, we define a transformed vector by $(\tilde{p}_{dij}, \tilde{p}_{tij})' = (sin^{-1}(\sqrt{\hat{p}_{dij}}), sin^{-1}(\sqrt{\hat{p}_{tij}}))'$. We find that the mean of $(\hat{p}_{dij}, \hat{p}_{tij})$ is related to the road type. After the transformation, the coefficient for VMT remains insignificant when road type is included in the model. The variance of the residuals from a constant variance model with road type as the covariate decreases as the vehicle miles traveled increases. We therefore incorporate the vehicle miles traveled in the variance to reflect the presence of a mean-variance relationship in the data. We reflect these characteristics in the model for the transformed proportions.

The bivariate unit-level model is defined as

$$\begin{pmatrix} \tilde{p}_{dij} \\ \tilde{p}_{tij} \end{pmatrix} = \begin{pmatrix} \beta_{0d} + \beta_{1d} I\left[R_{ij} = \text{Primary}\right] + \beta_{2d} I\left[R_{ij} = \text{Secondary}\right] \\ \beta_{0t} + \beta_{1t} I\left[R_{ij} = \text{Primary}\right] + \beta_{2t} I\left[R_{ij} = \text{Secondary}\right] \end{pmatrix} + \begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} + \begin{pmatrix} e_{1ij} \\ e_{2ij} \end{pmatrix}, \tag{24}$$

where R_{ij} is the road type of road segment j in county i, and $(u_{i1}, u_{i2})' \overset{iid}{\sim} N(0, \sum_{uu})$ independent of $(e_{1ij}, e_{2ij})' \overset{ind}{\sim} N(0, W_{ij}^{-0.5} \sum_{ee} W_{ij}^{-0.5})$ with $W_{ij} = (v_{ij}, v_{ij})$. We parametrize $\sum_{uu}$ and $\sum_{ee}$ as

$$\sum_{uu} = \operatorname{diag}(\sigma_{u1}, \sigma_{u2}) \begin{pmatrix} 1 & \rho_u \\ \rho_u & 1 \end{pmatrix} \operatorname{diag}(\sigma_{u1}, \sigma_{u2})$$

and

$$\sum_{ee} = \operatorname{diag}(\sigma_{e1}, \sigma_{e2}) \begin{pmatrix} 1 & \rho_e \\ \rho_e & 1 \end{pmatrix} \operatorname{diag}(\sigma_{e1}, \sigma_{e2}).$$

Denote the vector of fixed parameters of model (24) by

$$\theta = \left(\beta_{0d}, \beta_{1d}, \beta_{2d}, \beta_{0t}, \beta_{1t}, \beta_{2t}, \sigma^2_{u1}, \sigma^2_{u2}, \rho_u, \sigma^2_{e1}, \sigma^2_{e2}, \rho_e\right)'.$$

The Appendix (Subsection 7.1) explains how we estimate θ. Let $\hat{\theta}$ denote the estimator of θ, where $\hat{\theta} = (\hat{\beta}_{0d}, \hat{\beta}_{1d}, \hat{\beta}_{2d}, \hat{\beta}_{0t}, \hat{\beta}_{1t}, \hat{\beta}_{2t}, \hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}, \hat{\rho}_u, \hat{\sigma}^2_{e1}, \hat{\sigma}^2_{e2}, \hat{\rho}_e)'$. Table 6 gives the estimates of the regression coefficients. The local roads tend to have the lowest seat-belt use rates, while the primary roads tend to have the highest seat-belt use rates. The estimates of the variance parameters are $(\hat{\sigma}^2_{e1}, \hat{\sigma}^2_{e2}, \hat{\rho}_e) = (12.61, 11.21, 0.87)$ and $(\hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}, \hat{\rho}_u) = (0.0010, 0.00037, 0.52)$. The estimates of σ^2_{ek} are several orders of magnitude greater than the estimates of σ^2_{uk} because of the use of the inverse of the vehicle miles traveled as the weight. Based on the estimates of p_u and p_e, the bivariate correlation structure appears important.

To assess the adequacy of the model assumptions for this data set, we define standardized residuals. We define residuals for assessing the model for the transformed proportions by

$$r_{dij} = \frac{\tilde{p}_{dij} - \left(\hat{\beta}_{0d} + \hat{\beta}_{1d} I\left[R_{ij} = \text{Primary}\right] + \hat{\beta}_{2d} I\left[R_{ij} = \text{Secondary}\right]\right)}{\sqrt{\hat{\sigma}^2_{u1} + \hat{\sigma}^2_{e1}/(v_{ij})}}, \tag{25}$$

and

$$r_{tij} = \frac{\tilde{p}_{tij} - \left(\hat{\beta}_{0t} + \hat{\beta}_{1t} I\left[R_{ij} = \text{Primary}\right] + \hat{\beta}_{2t} I\left[R_{ij} = \text{Secondary}\right]\right)}{\sqrt{\hat{\sigma}^2_{u2} + \hat{\sigma}^2_{e2}/(v_{ij})}}. \tag{26}$$

The left panel of Figure 1 contains normal probability plots of the residuals, and the right panel depicts the residuals plotted against the predicted values. The top plots are for the proportion of drivers, and the bottom plots are for the proportion of total occupants. The residuals show modest outliers relative to a normal distribution but do not indicate severe departures from the model assumptions. Overall, the model seems to fit the data adequately.

Table 6. Estimates and standard errors for regression coefficients based on univariate linear mixed models applied to drivers and total vehicle occupants separately.

	Drivers ($k = d$)		Total ($k = t$)	
	Estimate	Std. error	Estimate	Std. error
β_{0k}	1.323	0.023	1.318	0.020
β_{1k}	0.087	0.022	0.064	0.021
β_{2k}	0.039	0.024	0.028	0.022

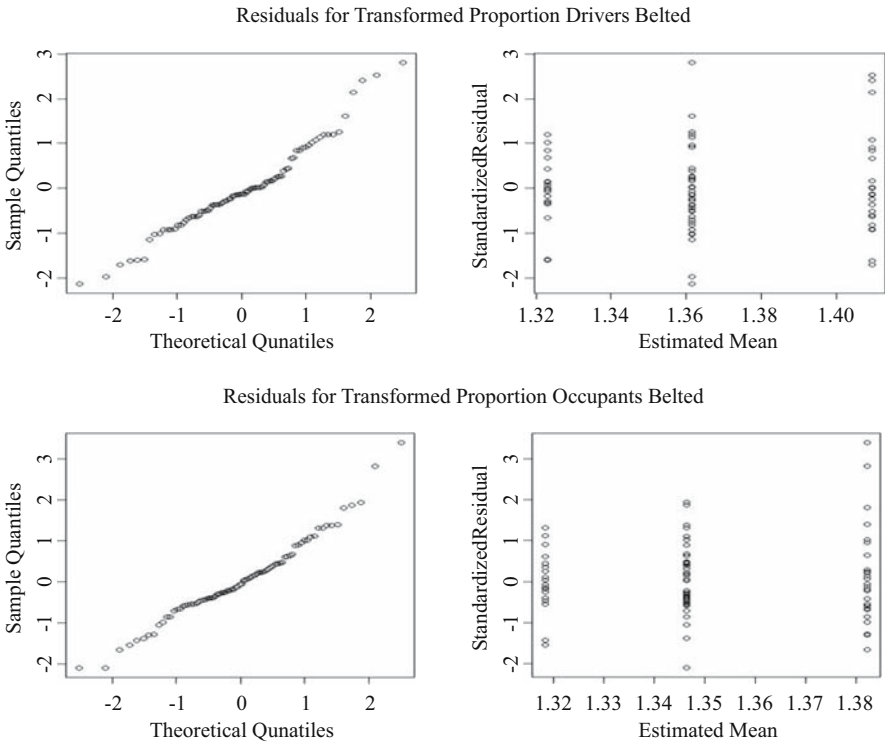


Fig. 1. Left: Normal probability plots of residuals. Right: Plots of residuals against predicted values. Residuals for proportions of drivers and total occupants are defined in Equations (25) and (26), respectively.

5.2. *Imputation and Weight Construction for Seat-Belt Survey Analysis*

We implement a bivariate version of the procedure of Subsection 3.1, termed "Unit-Imp-Biv." We select imputed values as described in Subsection 3.1. The weight for imputed value m for element (ij) is given by

$$\delta_{ij,m} = f\left(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)} | u_i^{(m)}, R_{ij}, \boldsymbol{v}_{ij}; \theta\right) / f_0\left(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)}\right),$$

were

$$u_i^{(m)} \sim f\left(u_{i1}, u_{i2} | \tilde{p}_{di}, \tilde{p}_{ti}, \hat{\theta}\right),$$

and $(\tilde{p}_{di}, \tilde{p}_{ti})$ denotes the observed vectors of transformed proportions county i. We approximate the marginal distribution of $(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)})$ with a mixture of three normal distributions. We then define $f_0(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)})$ as

$$f_0\left(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tif}^{(m)}\right) \propto \sum_{k \in \{\text{Primary, Secondary, Load}\}} \phi_k \exp\left(\left(\tilde{p}_{ij}^{(m)} - \mu_k\right)' S^{-1}\left(\tilde{p}_{ij}^{(m)} - \mu_k\right)\right),$$

where $\tilde{p}_{ij}^{(m)} = (\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)})'$, $\mu_k = (\sum_{i=1}^{D} \sum_{j=1}^{n_i} I[R_{ij} = k])^{-1} \sum_{k=1}^{D} \sum_{j=1}^{n_i} I[R_{ij} = k]$ $(\tilde{p}_{dij}, \tilde{p}_{tij})'$, $S = (n-1)^{-1} \sum_{i=1}^{D} \sum_{j=1}^{n_i} (\tilde{p}_{ij}, \bar{p}_{..})(\tilde{p}_{ij}, \bar{p}_{..})'$, $\bar{p}_{..} = n^{-1} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \tilde{p}_{ij}$, and $\phi_k = n^{-1}$ $\sum_{i=1}^{D} \sum_{j=1}^{n_i} I(R_{ij} = k)$.

For this data set the covariate is categorical with three categories, and this motivates our choice for the approximation of the marginal distribution. Other than the approximation of the marginal distribution, use of a continuous covariate or a covariate with more than two categories would present no new issues.

We construct $M = 5$ imputed values for each (ij). For this application, the sampling rates for the counties are all less than 1%, so we ignore the finite population correction factor and construct imputed values for all road segments in the population. The predictor of $(\theta_{di},\theta_{ti})$ is calculated as

$$\left(\hat{\theta}_{di},\, \hat{\theta}_{ti}\right) = \frac{1}{\sum_{j=1}^{N_i} \upsilon_{ij}} \sum_{j=1}^{N_i} \upsilon_{ij} \sum_{m=1}^{5} \delta_{ij,m} \left(\hat{p}_{dij}^{(m)}, \hat{p}_{tij}^{(m)}\right),$$

where $\hat{p}_{dij}^{(m)} = sin(\hat{p}_{dij}^{(m)})^2$ and $\hat{p}_{tij}^{(m)} = sin(\hat{p}_{tij}^{(m)})^2$.

The employed "Unit-Imp-Biv" procedure accounts for the bivariate structure but does not remove the imputation variance. As a result of the transformation, calculation of the optimal EB predictor is challenging. Therefore, removing the imputation variance is more difficult for the data analysis than for the simulation study. For this reason, we use the procedure Unit-Imp-Biv for the data analysis instead of the Unit-Imp-Biv-Adj method that eliminates the effect of random imputation.

We use the bootstrap procedure of Subsection 3.1 to estimate the mean square error of the predictor. The definition of the variance estimation imputed values for the application reflects the use of the transformation. For $b = 1,\ldots, B$, define

$$\bar{y}_{N_{i,imp}}^{(b)} = \frac{1}{\sum_{j=1}^{N_i} \upsilon_{ij}} \sum_{j=1}^{N_i} \upsilon_{ij} \left(sin\left(\tilde{p}_{dij}^{(b)}\right)^2, sin\left(\hat{p}_{tij}^{(b)}\right)^2 \right),$$

where

$$\begin{pmatrix} \tilde{p}_{dij}^{(b)} \\ \\ \tilde{p}_{tij}^{(b)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{0d} + \hat{\beta}_{1d} I[R_{ij} = \text{Primary}] + \hat{\beta}_{2d} I[R_{ij} = \text{Secondary}] \\ \\ \hat{\beta}_{0t} + \hat{\beta}_{1t} I[R_{ij} = \text{Primary}] + \hat{\beta}_{2t} I[R_{ij} = \text{Secondary}] \end{pmatrix} + \begin{pmatrix} u_{i1}^{(b)} \\ \\ u_{i2}^{(b)} \end{pmatrix} + \begin{pmatrix} e_{1ij}^{(b)} \\ \\ e_{2ij}^{(b)} \end{pmatrix}$$

$$u_i^{(b)} \sim f\left(u_{i1}, u_{i2} | \tilde{p}_{di}, \tilde{p}_{ti}, \hat{\theta}\right).$$

and $(e_{1ij}^{(b)}, e_{2ij}^{(b)})' \overset{ind}{\sim} N(0, W_{ij}^{-0.5} \sum_{ee} W_{ij}^{-0.5})$. We then define the variance estimation imputed value by $\hat{y}_i^{(b)} - \hat{y}_{N_i imp}^{(b)} - \bar{y}_i^{(\cdot)} + \hat{\bar{y}}_{N_i}$, where $\bar{y}_i^{(\cdot)} = B^{-1} \sum_{b=1}^{B} \bar{y}_{N_i imp}^{(b)}$ and $\hat{\bar{y}}_{N_i} = \left(\hat{\theta}_{di}, \hat{\theta}_{ti}\right)'$. We implement the parametric bootstrap (Rao and Molina 2015, 183) to obtain bootstrap estimates $\hat{\theta}^{(b)}$ for $b = 1,\ldots, B$. We define the bootstrap weight as

$$\delta_{ij,m}^{(B+b)} = f\left(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)} | u_{i,m}^{(b)}, R_{ij}, \upsilon_{ij}; \hat{\theta}^{(b)}\right) / f_0\left(\tilde{p}_{dij}^{(m)}, \tilde{p}_{tij}^{(m)}\right),$$

where

$$u_{i,m}^{(b)} \sim f\left(u_{i1}, u_{i2} | \tilde{p}_{di}, \tilde{p}_{ti}, \hat{\theta}^{(b)}\right).$$

We use the variance estimation imputed values and the bootstrap weights to construct an MSE estimator. For $b = B + 1, \ldots, 2B$, we define the replicate estimate by

$$\hat{y}_i^{(b)} = \frac{1}{\sum_{j=1}^{N_i} \upsilon_{ij}} \sum_{j=1}^{N_i} \sum_{m=1}^{M} \delta_{ij,m}^{(b)} \upsilon_{ij} \left(sin\left(\tilde{p}_{dij}^{(m)}\right)^2, sin\left(\tilde{p}_{tij}^{(m)}\right)^2 \right).$$

The MSE estimate for area i is defined as

$$\hat{M}_i = B^{-1} \sum_{b=1}^{2B} \left(\hat{y}_i^{(b)} - \left(\hat{\theta}_{di}, \hat{\theta}_{ti}\right)' \right) \left(\hat{y}_i^{(b)} - \left(\hat{\theta}_{di}, \hat{\theta}_{ti}\right)' \right)'. \tag{27}$$

5.3. *Results for Imputed Data Set*

Table 7 illustrates the structure of the final imputed data set. To preserve the confidentiality of the collected data, the the county name is suppressed, the street names are artificial, and the values for the proportions are generated from a distribution. We impute five pairs of proportions for each road segment. The weights for South Street sum to 1.01 instead of 1 due to rounding.

Table 8 contains the estimated county level proportions of drivers and total occupants wearing a seat-belt with corresponding standard errors. The standard errors are the square roots of the diagonal elements of the MSE estimator defined in Equation (27). The estimated standard errors vary between 1% and 2%, approximately. Most of the standard errors are below the 2.5% threshold established for the state level estimates. The estimated seat-belt use rates are typically in the range of 94%–96%. The estimates of the seat-belt use rates for the collection of 15 counties are 95.2% (SE = 1.7%) for drivers and 94.8% (SE = 1.7%) for total vehicle occupants. These estimates of overall seat-belt use rates are reasonably close to the direct seat-belt use rate estimate of 94.6% for the state of Iowa. Further analysis of small area estimation using this data set is an area for future work.

Table 7. Example of structure of imputed values for seat belt use survey: Five imputed values for each of three road segments with weights. (County label suppressed for confidentiality.)

Street Name	VMT	Road.Type	Weight	$\hat{p}_{drivers}$	$\hat{p}_{total}$
MAIN ST	25.37	Secondary	0.08	0.92	0.94
MAIN ST	25.37	Secondary	0.55	0.95	0.95
MAIN ST	25.37	Secondary	0.04	0.89	0.89
MAIN ST	25.37	Secondary	0.21	0.92	0.94
MAIN ST	25.37	Secondary	0.12	0.99	0.97
SOUTH ST	97.02	Local	0.01	0.94	0.92
SOUTH ST	97.02	Local	0.37	0.99	0.97
SOUTH ST	97.02	Local	0.56	0.87	0.88
SOUTH ST	97.02	Local	0.00	1.00	1.00
SOUTH ST	97.02	Local	0.07	0.97	0.97
6TH ST NW	9.11	Local	0.19	0.96	0.96
6TH ST NW	9.11	Local	0.50	0.95	0.93
6TH ST NW	9.11	Local	0.07	0.94	0.91
6TH ST NW	9.11	Local	0.09	0.90	0.83
6TH ST NW	9.11	Local	0.15	0.96	0.95

Table 8. County estimates of proportions of belted drivers and total occupants with standard errors. County names suppressed for confidentiality.

	Drivers		Total	
County	Est	SE	Est	SE
1	0.9396	0.0130	0.9370	0.0096
2	0.9449	0.0113	0.9408	0.0091
3	0.9493	0.0110	0.9432	0.0124
4	0.9417	0.0101	0.9396	0.0076
5	0.9571	0.0222	0.9533	0.0213
6	0.9667	0.0199	0.9547	0.0214
7	0.9528	0.0144	0.9471	0.0230
8	0.9584	0.0093	0.9550	0.0100
9	0.9547	0.0113	0.9486	0.0121
10	0.9484	0.0105	0.9431	0.0121
11	0.9514	0.0069	0.9467	0.0057
12	0.9545	0.0132	0.9498	0.0124
13	0.9526	0.0076	0.9507	0.0081
14	0.9495	0.0133	0.9454	0.0113
15	0.9561	0.0285	0.9492	0.0250

6. Discussion

We define algorithmic steps to construct imputed values and weights that produce specified small area estimates. In simulations, the area-level procedures and multivariate unit-level procedures work well for multivariate response variables. We use a unit-level bivariate model to construct an augmented database that supports county-level estimation for the proportions of drivers and total occupants who are wearing a seat-belt.

We develop imputation procedures for the two main classes of small area models: unit-level models and area-level models. The imputation procedures for the two model categories have different strengths and weaknesses. In the simulations, the procedures for area-level models adapt naturally multi-dimensional survey variables. In contrast, the procedures based on unit-level models are tailored to a specified response variable or set of response variables. A disadvantage of the procedure based on the area-level model is that the resulting weights may be negative. The weights based on the procedure defined in Subsection 3.1 for unit-level models are guaranteed to remain positive. However, the calibration-adjusted weights for unit-level models may be negative. The area-level procedures are limited to linear models. In principle, the unit-level procedures are general enough to encompass nonlinear forms, but we do not consider nonlinear models. The computational complexity will increase with nonlinear models. We illustrate the methods for linear models, and further study of the unit-level procedure in the context of a nonlinear model would be an interesting area for future work.

Our focus on standard errors instead of confidence intervals is motivated by the needs of the NRI and Seat-Belt Use surveys, the two surveys that provide primary motivation for our work. The products of these two survey efforts are estimates and standard errors (not confidence intervals). One can construct normal theory confidence intervals from the provided standard error estimates, as is common in small area estimation. We refer the

reader to Datta et al. (2002) for a rigorous discussion of confidence interval calibration in the context of the Fay-Herriot model. We view the problem of constructing imputed values and replicate weights to reproduce calibrated confidence intervals as an area for future research.

We focus on the case in which all areas have at least one sampled element. Defining the steps to extend the procedures to non-sampled areas is conceptually straightforward. Estimators for non-sampled areas are fully model-dependent. A non-sampled area does not contain any data for assessing if the model assumptions hold for the area. Caution is therefore warranted when constructing predictors for non-sampled areas. We take a prudent approach in this article and focus on estimation for the case in which every area has at least one sampled element.

A limitation of our approach is that the imputation procedure is tied to a specified small area model. The definition of the small area model includes a definition of a set of small domains. In the simulations, we demonstrate that the imputation procedure produces reasonable estimates for linear combinations of small area parameters, as well as for the specified domains. In practice, a user may define an estimation domain that is not a linear combination of domains included in the model. One can use the database to construct an estimate for such a domain. The efficiency of the estimator is unknown. The survey-statistician may have a general idea for a collection of possible domains that may be of interest to a variety of users. One approach then is to define the small areas in the model to be the intersection of possible domains of interest. Then, user-specified domains are likely to be linear combinations of small areas included in the model. An extension of the procedure to models with nested random effects may improve estimates for linear combinations. An effort to reduce the dependence of the procedure on a model that includes specified domains is a possible direction for future work.

An important limitation that the simulation study reveals deserves reiteration. The unit-level procedures may be inefficient for variables excluded from the model. The weights based on a univariate unit-level model are not automatically suited for estimation of a vector of response variables. Constructing general-purpose weights based on unit-level models is an avenue for future research.

We demonstrate the validity of the procedures through a specialized application, where the number of response variables is only two. Users are likely to encounter many challenges when applying these methods in general. The limitations that we have discussed point to some of the anticipated challenges. In the seat-belt survey application, only two response variables are of interest. Applying the methods to high-dimensional vectors of response variables may lead to difficulties. Obtaining an efficiency gain for all elements of a high-dimensional response may be difficult. One may suffer a loss of efficiency for some components of a vector-valued response. As discussed in Subsection 2.2, the calibration constraint is not satisfied if the required matrix inverse does not exist. We think that this issue is likely to become more important as the number of response variables in the model increases. We assume that the same covariate is used for all response variables. The analyst may have different covariates for different elements of a vector-valued response. In this case, we recommend including all covariates in the common covariate vector. This will guard against bias at the possible expense of an increase in variance. The analyst may want to use a nonlinear model, while we focus on

linear models. The user may want to construct an estimate for a domain that is not included in the model. This is operationally possible, but the statistical properties of the resulting predictor are unknown.

Despite the limitations, this work offers a contribution of value to the literature on small area estimation. The problem of constructing multi-purpose databases that support small area estimation is of general interest. Our contribution is to define the operational procedures to construct weights and imputed values that approximate specified small area estimates. The development in this article may prove useful for further efforts to construct a database that supports small area estimation.

7. Appendix

7.1. Estimators of Parameters for Bivariate Model

We explain how we estimate the parameters for the type of bivariate model used in the simulation and the data analysis. We define a two-step procedure to estimate the correlation parameter. The procedure uses a model for the difference between the two response variables. An alternative is to estimate all parameters simultaneously using REML. We prefer the two-step procedure described below for computational simplicity. We use a general notation with response random variable $(Y_{1ij}, Y_{2ij})'$ for elements $j = 1, \ldots, n_i$ and $i = 1, \ldots, D$. The bivariate normal model implies two univariate mixed effects models for each of Y_{1ij} and Y_{2ij} separately. Specifically, the implied univariate model for Y_{kij} $(k = 1, 2)$ is defined as

$$Y_{kij} = x'_{ij}\beta_k + u_{ki} + e_{kij}, \tag{28}$$

where $u_{ki} \overset{iid}{\sim} N(0, \sigma^2_{uk})$ for $i = 1, \ldots, D$, and $e_{kij} \overset{ind}{\sim} N(0, \sigma^2_{ek} w^{-1}_{ij})$. Using REML under the assumptions of the univariate model (28), we obtain the estimate $(\hat{\sigma}^2_{uk}, \hat{\sigma}^2_{ek}, \hat{\beta}'_k)'$ of $(\sigma^2_{uk}, \sigma^2_{ek}, \beta_k{}')'$ for $k = 1,2$.

We estimate ρ_e and ρ_u using the two univariate models as well as a model for the difference, $Y_{2ij} - Y_{1ij}$. The bivariate model implies a linear mixed model for $Y_{2ij} - Y_{1ij}$. To define the model for the differences, let $D_{ij} = Y_{2ij} - Y_{1ij}$. Then, under the assumed bivariate model, D_{ij} satisfies

$$D_{ij} = x'_{ij}(\beta_2 - \beta_1) + b_i + \eta_{ij},$$

where $b_i \overset{iid}{\sim} N(0, \sigma^2_b)$, $\eta_{ij} \overset{iid}{\sim} N(0, \sigma^2_\eta w^{-1}_{ij})$

$$\sigma^2_b = \sigma^2_{u1} + \sigma^2_{u2} - 2\rho_u \sigma_{u1} \sigma_{u2} \tag{29}$$

and

$$\sigma^2_\eta = \sigma^2_{e2} + \sigma^2_{e1} - 2\rho_e \sigma_{e2} \sigma_{e1}. \tag{30}$$

We let $\hat{\sigma}^2_\eta$ and $\hat{\sigma}^2_b$ be REML estimators of σ^2_η and σ^2_b obtained from the linear mixed effects model for D_{ij}. Equations (29) and (30) define relations between the correlations and the other variance parameters. We use these relations to define estimators of ρ_u and ρ_e. We define estimates of ρ_e and ρ_u as

$$\hat{\rho}_{e1} = -\frac{\hat{\sigma}_{\eta}^2 - \hat{\sigma}_{e2}^2 - \hat{\sigma}_{e1}^2}{2\hat{\sigma}_{e2}\hat{\sigma}_{e1}},$$

and

$$\hat{\rho}_{u1} = -\frac{\hat{\sigma}_{b}^2 - \hat{\sigma}_{u2}^2 - \hat{\sigma}_{u1}^2}{2\hat{\sigma}_{u2}\hat{\sigma}_{u1}}.$$

The estimates ρ_{u1} and ρ_{e1} are consistent for ρ_u and ρ_e but may be outside of the boundary of the parameter space for a correlation. We define alternate estimates of the correlation that are guaranteed to remain in the parameter space. Let $\hat{u}_{1i}$ and $\hat{u}_{2i}$ be the best linear unbiased predictors (BLUPs) of the random effects in the univariate models for Y_{1ij} and Y_{2ij}, respectively. Define a residual by

$$\hat{e}_{kij} = \frac{Y_{kij} - x_{ij}'\beta_k - \hat{u}_k}{\sqrt{\sigma_{ek}^2 w_{ij}^{-1}}}.$$

Define

$$\hat{\rho}_{u0} = \frac{\sum_{i=1}^{D}(\hat{u}_{i1} - \bar{u}_1)(\hat{u}_{i2} - \bar{u}_2)}{\sqrt{\sum_{i=1}^{D}(\hat{u}_{i1} - \bar{u}_1)^2 \sum_{i=1}^{D}(\hat{u}_{i2} - \bar{u}_2)^2}}.$$

and

$$\hat{\rho}_{e0} = \frac{\sum_{i=1}^{D}\sum_{j=1}^{n_i}\left(\hat{e}_{1ij} - \bar{e}_1\right)\left(\hat{e}_{2ij} - \bar{e}_2\right)}{\sqrt{\sum_{i=1}^{D}\sum_{j=1}^{n_i}\left(\hat{e}_{1ij} - \bar{e}_1\right)^2 \sum_{i=1}^{D}\sum_{j=1}^{n_i}\left(\hat{e}_{2ij} - \bar{e}_2\right)^2}}.$$

The estimates $\hat{\rho}_{u0}$ and $\hat{\rho}_{e0}$ are not necessarily consistent but are in the parameter space for a correlation.

We combine the two estimates of the correlation parameters to obtain the final estimates of ρ_u and ρ_e. We define the estimate of ρ_u by

$$\hat{\rho}_u = \begin{cases} \hat{\rho}_{u1} & \hat{\rho}_{u1} \in (-1, 1) \\ \hat{\rho}_{u0} & \text{otherwise.} \end{cases}$$

Likewise, we define the estimate of ρ_e by

$$\hat{\rho}_e = \begin{cases} \hat{\rho}_{e1} & \text{if } \hat{\rho}_{e1} \in (-1, 1) \\ \hat{\rho}_{e0} & \text{otherwise.} \end{cases}$$

A bootstrap estimate of σ_{uk}^2 can be singular. We apply a positive lower bound to the diagonal elements of the bootstrap estimates of the area random effect covariance matrix. For $k = 1, 2$, let $\hat{\sigma}_{uk,0}^{(b)2}$ be the estimate of σ_{uk}^2 obtained by fitting the univariate model to the bootstrap data set. If $\hat{\sigma}_{uk,0}^{(b)2} = 0$, then define the bootstrap estimate by $\hat{\sigma}_{uk}^{(b)2} = 0.5\sqrt{2\hat{\sigma}_{uk}^4/D}$. This lower bound is tied to the recommendation in Wang and Fuller (2003). We do not reestimate the correlation parameter in the bootstrap. We define $\left(\hat{\rho}_e^{(b)}, \hat{\rho}_u^{(b)}\right) = \left(\hat{\rho}_e, \hat{\rho}_u\right)$ for $b = 1, \ldots, B$.

7.2. Details of Imputation and Estimation for an Informative Sample Design

We define the complement distribution for the unit-level linear model with normally distributed errors. The results below are available in Pfefferman and Sverchkov (2007). We include them here for ease of reference. Assume the sample distribution satisfies

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + e_{ij},$$

for $i = 1, \ldots, D$ and $j = 1, \ldots, n_i$, where the area random effect $b_i \sim N(0, \sigma_b^2)$ and the error terms $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Assume the sampling weight satisfies

$$E[\pi_{ij}^{-1} | y_{ij}, x_{ij}, j \in A_i] \propto \exp(by_{ij} + ax_{ij}),$$

where A_i is the index set for the sample in area i. Then, assuming the sampling fractions are small, the complement distribution satisfies

$$f(y_{ij} | \tilde{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{y}_i, j \notin A_i) \propto \exp(-0.5(y_{ij} - b\sigma_e^2 - \beta_0 - \beta_1 x_{ij} - \tilde{b}_i)^2 / b\sigma_e^2), \qquad (31)$$

where $\tilde{b}_i \overset{iid}{\sim} N(\hat{b}, \gamma_i \sigma_e^2 / n_i)$, $\hat{b}_i = \gamma_i(\bar{y}_i - \beta_0 - \beta_1 \bar{x}_i)$, $\gamma_i = \sigma_b^2 (\sigma_b^2 + \sigma_e^2 n_i^{-1})^{-1}$, and $(\bar{x}_i, \bar{y}_i) = n_i^{-1} \sum_{j=1}^{n_i} (x_{ij}, y_{ij})$.

The method of Pfefferman and Sverchkov (2007) can be used for estimation. The complement distribution defines the weights for imputed values.

7.3. Names of Imputation Procedures

Table 9. Names and descriptions of imputation procedures.

Name	Description	Subsection defined
Area-Imp	Imputation procedure for area-level model	2.2
Unit-Imp	Imputation procedure for univariate unit-level model	3.1
Unit-Imp-Adj	Imputation procedure for univariate unit-level model with adjustment to remove imputation variance	3.2
Unit-Imp-Adj-Aug	Imputation procedure for univariate unit-level model with adjustment to remove imputation variance and weight adjustment to preserve EB predictor for a second variable	3.3
Unit-Imp-Biv	Imputation procedure for bivariate unit-level model	3.1
Unit-Imp-Biv-Adj	Imputation procedure for bivariate unit-level model with adjustment to remove imputation variance	3.2

7.4. Justification for Imputation Procedure of Subsection 3.1 (Unit-Imp)

We justify the unit-level imputation procedure of Subsection 3.1. For simplicity, we consider a univariate variable y_{ij}. The integral of interest is

$$E[g(y) | \boldsymbol{y}_i, \boldsymbol{x}_{ij}] = \int g(y) f(y | b_i, \boldsymbol{x}_{ij}) f(b_i | \boldsymbol{y}_i) db_i dy.$$

We can approximate this integral through a two-step simulation procedure. For $m = 1, \ldots, M$, generate

(i) $b_i^{(m)} \sim f(b_i|y_i)$.
(ii) $y^{(m)} \sim f(y|b_i^{(m)})$.
Then, approximate the integral as

$$\hat{E}[g(y)|y_i, x_{ij}] = \frac{1}{M}\sum_{m=1}^{M} g(y^{(m)}).$$

We want to use donor imputation instead of simulating from a parametric model. Therefore, we replace step (2) with an importance sampling step. Instead of step (2), we implement step (2*):

(2*) Generate $y^{(m)} \sim$ Multinomial $(\mathbf{y}, n^{-1}1'_n)$. This results in a realization from the marginal distribution of y_{ij}. Assign an importance weight (Kim and Shao 2013, 45) of

$$w_m = \frac{f(y^{(m)}|b_i^{(m)}, x_{ij})/f_0(y^{(m)})}{\sum_{m=1}^{M} f(y^{(m)})|b_i^{(m)}, x_{ij}/f_0(y^{(m)})}.$$

Approximate the integral as

$$\hat{E}[g(y)|y_i, x_{ij}] = \sum_{m=1}^{M} w_m g(y^{(m)}).$$

This procedure is similar to analogous integral approximations in Kim and Shao (2013) and Kim (2011).

7. References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association* 83(401): 28–36. DOI: https://doi.org/10.1080/01621459.1988.10478561.

Berg, E. and W.A. Fuller. 2018. "Benchmarked small area prediction." *Canadian Journal of Statistics* 46(3): 482–500. DOI: https://doi.org/10.1002/cjs.11461.

Chandra, H. and R. Chambers. 2009. "Multipurpose Weighting for Small Area Estimation." *Journal of Official Statistics* 25(3): 379–395. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/multipurpose-weighting-for-small-area-estimation.pdf (accessed July 2022).

Chen, Z. 2019. *Model-based analysis in survey: an application in analytic inference and a simulation in small area estimation.* Available at: https://dr.lib.iastate.edu/handle/20.500.12876/16847 (accessed July 2022).

Datta, G.S., M. Ghosh, D.D. Smith, and P. Lahiri. 2002. "On an asymptotic theory of conditional and unconditional coverage probabilities of empirical bayes confidence intervals." *Scandinavian Journal of Statistics* 29(1): 139–152. Available at: http://www.jstor.org/stable/4616705 (accessed July 2022).

DOT. 2020. Available at: https://iowadot.gov/maps/Data/Vehicle-miles-traveled (accessed April 2022).

Erciulescu, A.L. and W.A. Fuller. 2016. "Small area prediction under alternative model specifications." *Statistics in Transition new series* 17(1): 9–24. DOI: https://doi.org/10.21307/stattrans-2016-003.

Fay, R.E. and Herriot, R.A. 1979. "Estimates of income for small places: an application of james-stein procedures to census data." *Journal of the American Statistical Association* 74(366a): 269–277. DOI: https://doi.org/10.2307/2286322.

Hidiroglou, M.A. and Y. You. 2016. "Comparison of unit level and area level small area estimators." *Survey Methodology* 42(1): 41–61. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016001/article/14540-eng.pdf?st=OLmHgT9E Accessed 7/6/22.

Isaki, C., Ikeda, M. and Fuller, W. 2000. "An estimation file that incorporates auxiliary information." *Journal of Official Statistics* 16(2): 155–172. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-estimation-file-that-incorporates-auxiliary-information.pdf.

Jiang, P. Lahiri and S.-M. Wan. 2002. "A unified jackknife theory for empirical best prediction with m-estimation." *The Annals of Statistics* 30(6): 1782–1810. DOI: https://doi.org/10.1214/aos/1043351257.

Kim, J.K. 2011. "Parametric fractional imputation for missing data analysis." *Biometrika* 98(1): 119–132. DOI: https://doi.org/10.1093/biomet/asq073.

Kim, J.K. and J. Shao. 2013. *Statistical methods for handling incomplete data*, CRC press.

Lyu, X., Berg, E.J. and H. Hofmann. 2020. "Empirical bayes small area prediction under a zero-inflated lognormal model with correlated random area effects." *Biometrical Journal* 62(8): 1859–1878. DOI: https://doi.org/10.1002/bimj.202000029.

Mauro, F., Monleon, V.J., H. Temesgen, and K.R. Ford. 2017. "Analysis of area level and unit level models for small area estimation in forest inventories assisted with lidar auxiliary Information." *PLoS One* 12(12): e0189401. DOI: https://doi.org/10.1371/journal.pone.0189401.

Nusser, S.M. and J.J. Goebel. 1997. "The national resources inventory: a long-term multiresource *monitoring* programme." *Environmental and Ecological Statistics* 4(3): 181–204. DOI: https://doi.org/10.1023/A:1018574412308.

Pfeffermann, D. and M. Sverchkov. 2007. "Small-area estimation under informative probability sampling of areas and within the selected areas." *Journal of the American Statistical Association* 102(480): 1427–1439. DOI: https://doi.org/10.1198/016214507000001094.

Randrianasolo, T. and Y. Tillé. 2013. "Small area estimation by splitting the sampling weights." *Electronic Journal of Statistics* 7: 1835–1855. DOI: https://doi.org/10.1214/13-EJS827.

Rao, J.N. and I. Molina. 2015. *Small area estimation*, John Wiley & Sons.

Wang, J. and W.A. Fuller. 2003. "The mean squared error of small area predictors constructed with estimated area variances." *Journal of the American Statistical Association* 98(463): 716–723. DOI: https://doi.org/10.1198/016214503000000620.

Wang, J., W.A. Fuller and Y. Qu. 2008. "Small area estimation under a restriction." *Survey methodology* 34(1): 29. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10619-eng.pdf?st=mAh7Jalj (accessed July 2022).

Yang, S., J.K. Kim, and Z. Zhu. 2013. "Parametric fractional imputation for mixed models with nonignorable missing data." *Statistics and Its Interface* 6(3): 339–347. DOI: https://dx.doi.org/10.4310/SII.2013.v6.n3.a4.

Received June 2021
Revised December 2022
Accepted April 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 709–732, http://dx.doi.org/10.2478/JOS-2022-0032

Hierarchical Bayesian Model with Inequality Constraints for US County Estimates

Lu Chen[1], *Balgobin Nandram*[2], *and Nathan B. Cruze*[3]

In the production of US agricultural official statistics, certain inequality and benchmarking constraints must be satisfied. For example, available administrative data provide an accurate lower bound for the county-level estimates of planted acres, produced by the U.S. Department of Agriculture's (USDA) National Agricultural statistics Services (NASS). In addition, the county-level estimates within a state need to add to the state-level estimates. A sub-area hierarchical Bayesian model with inequality constraints to produce county-level estimates that satisfy these important relationships is discussed, along with associated measures of uncertainty. This model combines the County Agricultural Production Survey (CAPS) data with administrative data. Inequality constraints add complexity to fitting the model and present a computational challenge to a full Bayesian approach. To evaluate the inclusion of these constraints, the models with and without inequality constraints were compared using 2014 corn planted acres estimates for three states. The performance of the model with inequality constraints illustrates the improvement of county-level estimates in accuracy and precision while preserving required relationships.

Key words: Administrative data; bayesian diagnostic; benchmarking; crop acreage estimates; small area estimation; sub-area models; survey data.

1. Introduction

The National Agricultural statistics Service (NASS), the primary statistical data collection agency within the U.S. Department of Agriculture (USDA), conducts the County Agricultural Production Survey (CAPS) annually. CAPS provides county-level estimates by commodity crop for the following estimands: planted acres, harvested acres, yield and production. 'Crop type by county' represents a planned domain, in the sense that the CAPS multivariate-probability-proportional-to-size design and sample selection is specifically intended to support NASS's county-level data products. However, the number of survey reports obtained for each domain can vary widely due to issues of survey nonresponse, genuine differences in planting decisions each year, and in the inherent complexity of sampling for the breadth of crops of interest nationwide. The current method of producing

[1] National Institute of Statistical Sciences, 1750 K Street, NW, Suite 1100, Washington. D.C., 20006-2306, U.S.A. Email: lchen@niss.org
[2] Worcester Polytechnic Institute and USDA National Agricultural statistics Service, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609. U.S.A. Email: balnan@wpi.edu
[3] USDA National Agricultural statistics Service, 1400 Independence Avenue, SW, Washington, D.C. 20250-2054. Email: nathan.cruze@gmail.com
Acknowledgments: The findings and conclusions in this article are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy. This research was supported by the intramural research program of the USDA, NASS. Dr. Nandram's work was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). Dr. Cruze's contributions to this work were made during his tenure at USDA, NASS.

© Statistics Sweden

official county-level crop estimates is an expert assessment conducted by NASS's Agricultural statistics Board (ASB), which incorporates multiple sources of information. The information includes CAPS estimates and administrative data whenever it is available. These county-level estimates are key indicators to farmers, ranchers and a number of federal and state agencies for decision making. Two USDA agencies, the Farm Service Agency (FSA) and the Risk Management Agency (RMA), consider the estimates as part of their processes for distributing farm subsidies and insurance respectively.

Given the importance of the crops county estimates program, NASS engaged a panel of experts under the National Academies of Sciences, Engineering, and Medicine for guidance and recommendations on implementing models for integrating multiple sources of information to provide county-level crop estimates with measures of uncertainty. The panel's recommendations were issued in a publicly available report; see National Academies of Sciences, Engineering, and Medicine NASEM 2017. See also Cruze et al. (2019) for a recapitulation of some of the panel's findings. In the traditional process of setting official statistics, the ASB has relied on standard processes, multiple data sources, historical performance of these sources, and expert judgment. The ASB analyzes the survey estimates and integrates them with multiple data sources through a series of informal composite estimators. (See NASEM 2017, 27–28; Cruze et al. 2019, sec. 2.) Final estimates are checked for *coherence with external administrative totals* that are interpreted as minimum amounts of activity known to have taken place in the county, and the estimates are rounded in accordance with NASS rounding rules. In a statistical sense, the ASB results are not reproducible and measures of uncertainty have not been produced with the traditional data product.

In recent years, small area models have gained increased attention by academic researchers and government agencies. Small area estimation models can "borrow strength" from related areas across space and/or time or through auxiliary information to provide "indirect" but reliable estimates for small areas while also increasing precision. One challenge of a model-based approach is to provide reliable and coherent estimates that satisfy important relationships nested among estimates and administrative data. The NASS county-level official estimates of planted acres should be greater than or equal to the corresponding available administrative totals that represent known minimum amounts of planting activity within the county, while also satisfying benchmarking constraints so that county-level estimates add up to the state-level estimates. In this article, hierarchical Bayesian models with constraints for small area estimation are discussed and applied to NASS's planted area estimates of corn for grain with reference to the 2014 crop year. With the goal of improving transparency of processes and quantifying the uncertainty associated with each estimate, NASS implemented the described model-based approach for estimating county-level planted area totals for thirteen commodity crops nationwide beginning with the 2020 crop year.

Two major types of small area models, area-level and unit-level models, have been developed using both frequentist and Bayesian methods. Pfeffermann (2013) and Rao and Molina (2015) provide a comprehensive overview of the development, methods and application of small area estimation including various types of area-level and unit-level models. For continuous responses, the first and most common model is the Fay-Herriot (FH) model (Fay and Herriot 1979) in small area estimation. It is an area-level model

based on a "normal-normal-linear" assumption. That is, the direct estimates and area-level random effects are both assumed to follow normal distribution and a linear regression function relates the true estimates of interest to covariates. The popular unit-level model, nested-error regression (NER) model, was proposed by Battese et al. (1988) when data are available on the individual sampled units. The NER model is also developed under the normality assumption.

The objective of NASS crops county estimates program is to incorporate different sources of auxiliary information with survey estimates in the model to provide coherent and reliable estimates with associated measures of uncertainty. The modeling strategies in both frequentist and Bayesian methods could operate in similar way. However, Bayesian approaches are more straightforward for obtaining estimates for any known functions of the model parameters. In addition, Bayesian methodology is well suited for inequality constrained problems as it naturally provides a framework that allows complex constraints via hierarchical models. Recent studies and papers related to the NASS crops county estimates program have shown that hierarchical Bayesian small area models can incorporate auxiliary sources of data to improve county-level survey estimation of crop totals with measures of uncertainty. Battese et al. (1988) introduced the unit-level models for small area estimation based on nested error linear regression. They combined survey indications with satellite data. Erciulescu et al. (2019) proposed and implemented a double shrinkage hierarchical Bayesian sub-area level model to provide the acreage estimates with associated measures of uncertainty. The paper discussed the results when integrating different data sources and showed that the county-level model-based acreage estimates decreased the coefficients of variation relative to the survey ones. Erciulescu et al. (2020) discussed the challenges of missing data, either survey responses or administrative data, when fitting hierarchical Bayesian sub-area level model to obtain the crops total estimates for the whole nation. In these two papers, the state-to-county benchmarking constraint is included.

Increasingly, constrained estimation problems have found application and international importance in the small area estimation literature. Sen et al. (2018) proposed the method to conduct inference for a constrained posterior and project samples to the constrained space through a minimal distance mapping. Instead of placing a prior on the constraint space and conducting posterior computation, a general formulation of projected posteriors in a Bayesian decision-theoretic framework is provided. Cruze et al. (2019) identified constraints among estimates and administrative data as a necessity and allowed for the possibility of different constraints by small area. Whereas the inequality constraint problems were not addressed in the aforementioned, NASS-authored literature, Nandram et al. (2022) addressed the inequality constraint problem and proposed several hierarchical Bayesian models for NASS crops county-level planted area estimates which have ultimately been used in practice by NASS effective with the 2020 crop year. They discussed the methodologies of fitting constrained models and provided a simulation study to show the performance of all models.

In this article, models with inequality constraints are discussed and implemented to address the needs and challenges of inequality and benchmarking constraints that NASS official statistics must satisfy. The models with inequality constraints of Nandram et al. (2022) are applied to 2014 NASS CAPS data. In Section 2, data sources and some particular needs of the NASS crops county estimates for total planted acres are presented.

In Section 3, hierarchical Bayesian models with inequality constraints are proposed to produce reliable and coherent county-level estimates and associated measures of uncertainty. External ratio benchmarking is applied to the county-level estimates so that they sum to state targets. The results are contrasted with those obtained from unconstrained models. In Section 4, a case study based on three different states shows the model-based estimation results and highlights the different performances of the constrained models and the unconstrained models. Conclusions and future work are presented in Section 5.

2. Data Sources and Requirements

2.1. Survey Data

Although NASS has been producing official county-level agricultural estimates since 1917, it was in 2011 that NASS completely implemented the large-scale probability survey, CAPS, to provide county-level official estimates for many principle small grains and row crops in several states.

The CAPS survey uses a Multivariate Probability Proportional to Size (MPPS) sample design. The target population for CAPS is all agricultural operators with cropland and/or storage capacity in any of the eligible states. The NASS list frame includes all known agricultural establishments. The list frame for CAPS consists of those NASS list frame records with positive planted acres or storage capacity of the desired commodities in the previous year (NASEM 2017, 111–117). Sample size is dependent upon the number of operations in the universe list and the variability of data among operations on a given list. Sample sizes vary widely among states, and the number of obtained reports will vary by state, commodity crop, and county, but there is some effort to treat 'county by crop' as a planned domain, and construct samples accordingly.

The list of crops and states in CAPS may change year to year depending on the requirement of coverage for federally mandated program crops and others. Figure 1 shows

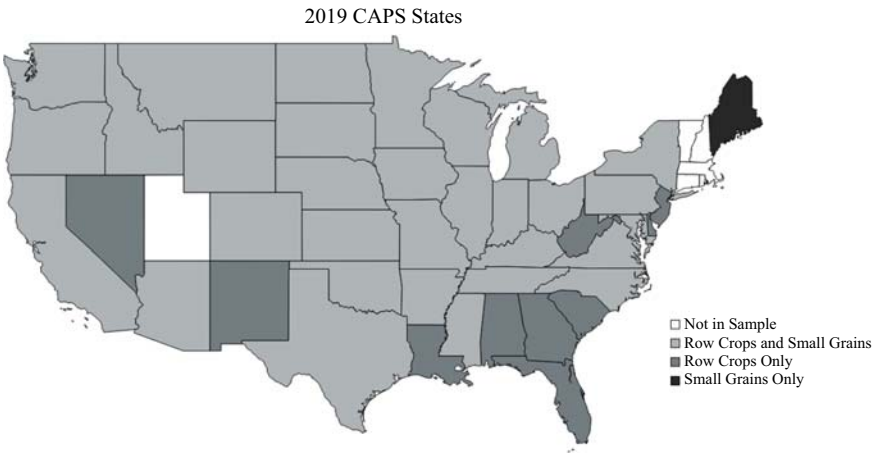


Fig. 1. 2019 row crops and small grains CAPS states.

the 2019 CAPS states. Four different colors indicate the category that each state is in for the 2019 CAPS. The state in black was for small grains CAPS only. States in dark gray were for row crops CAPS only. States in light gray were for both small grains and row crops CAPS. States in white were not included in 2019 CAPS. The row crops CAPS (e.g., corn, soybeans) was conducted in 41 states. The small grains CAPS (e.g., barley, oats) was conducted in 32 states. No other states were included in 2019 CAPS.

As discussed in the introduction, the smallest area at which CAPS produces estimates is the county. Historically, NASS has also produced estimates for an intermediate domain called the agricultural statistics district (ASD). Each ASD is comprised of contiguous counties in the state. Both county-level and ASD level survey estimates and associated variance estimates are available in CAPS summary. The state-level planted acreage estimates are published before the completion of data collection for the CAPS. Therefore, when setting the county level estimates, an external state benchmarking constraint exists. In the traditional estimation process, the ASB reviews the direct estimates based on CAPS and relevant auxiliary information to set county estimates that aggregate to those state targets.

2.2. Auxiliary Data

NASS obtains auxiliary sources of information on crop acres from FSA and RMA. Both agencies have farmer-reported administrative data on planted acres. While FSA and RMA programs are popular, they are not compulsory. The activity of some parts of the population may be absent in either record. The participation rates can vary by crop, by state, and even by locality within state. For example, the rates of enrollment in FSA programs for corn are typically higher in so-called corn-belt states, for example Illinois, than in some other states such as Ohio or Pennsylvania.

As described in NASEM (2017, 20), "FSA defines the common land unit (CLU) as an individual, contiguous farming parcel, which is the smallest unit of land that has a permanent, contiguous boundary; common land cover and land management; and a common owner or common producer association." FSAmaintains a database of these digitized, geolocated field boundaries for the entire United States. The size and location of the fields are known with accuracy. The *contents* contained within these field boundaries remain empty until planted acreages are reported by farmers to FSA each year. Farmers who opt to participate in FSA programs must certify their planting activity by prescribed due dates. The process of certification typically entails a visit by farmers to an FSA office, where the farmer is assisted in identifying the fields (CLUs) operated on maps. The farmer then provides the acreages by type of crop planted in a standard form containing all associated identifiers with the parcels operated. Deliberate misreporting is dissuaded under penalty of "loss of program benefits for noncompliance." The FSA handbook on *Acreage and Compliance Determinations* (USDA FSA 2018) details the procedures for certification, as well as quality assurance and compliance procedures at length.

As overseer of the Federal Crop Insurance Corporation, RMA receives administrative data on planted acres as farmers enroll in insurance coverage through approved insurance providers or file claims that are associated with these programs. Farmers may choose not participate in any crop insurance programs, or they may not insure all commodity crops

they choose to grow. The participation rates in crop insurance can vary by state and commodity. For example, in RMA's own analysis of the 2015 federal crop insurance portfolio market penetration, defined as the percent of national planted area totals (estimates produced by NASS) that are also insured acres under these programs, it was found that 89% of corn acres, 90% of soybean acres, 73% of barley acres, and just 17% of oats acres were insured (USDA RMA 2017). Accordingly, NASS treats the RMA administrative data on planted acres as a useful lower bound on the planted acreage.

Non-probability sources like the FSA and RMA programs administrative data are not free of nonsampling errors. Foremost, neither of these collections represents a registry of the activity of all population units, and therefore, totals obtained at any geographic level from either source may subject to some degree of under coverage or provide underestimates of population totals. Good (2014) discussed this in the context of comparing national totals from NASS and from FSA. The aforementioned RMA analysis pointed to likely undercoverage nationwide. As a department, USDA, takes steps to mitigate other types of nonsampling error that could affect the quality of the reported data, through minimizing opportunities or incentives to misreport, through ongoing quality assurance procedures, and in the case of FSA, through geospatially resolving the collected data to the county, and more specifically, to the field, where the crop was planted. With these combinations of factors and the need to produce estimates that are coherent given these other USDA data sources, NASS has interpreted the administrative data as *informative lower bounds* in the construction of official county estimates produced under the traditional ASB process and seeks to retain that feature in any candidate model for planted area. While there can be significant overlap of FSA and RMA data, not all operations will participate in both. Because NASS treats both FSA and RMA data as the lower bounds of the county-level planted acreage estimates, the definition of the lower bound in the constrained models is the maximum of both sources of administrative data. That is, where FSA and RMA acreages may differ, the larger is taken as a firm lower bound NASS estimates should respect.

2.3. *Important Relationships for Planted Acres*

In the production of the official statistics for total acres reported by NASS, benchmarking *and inequality constraints* should be satisfied. NASS sometimes describes its procedures as 'top-down', meaning that national and state estimates are published before the sub-state ASD and county estimates, even while additional CAPS data collection may be ongoing (NASEM 2017, 24). In practical terms, it means that official county-level estimates will have an external target for benchmarking county totals to the published state total. Additionally, NASS's official estimates of planted acres should cover corresponding available administrative data: FSA and RMA planted acreage data within any given geographic boundary, such as the US, a state, and county. The differences between NASS official statistics and FSA administrative data of total planted acreage for corn, soybeans, barley and oats at US level from 2012 to 2019 are displayed in Figure 2. Each plot shows that the differences between NASS official estimates and FSA data are all positive at the US level. However, the county-level survey estimates of the planted acreage do not always satisfy the constraints.

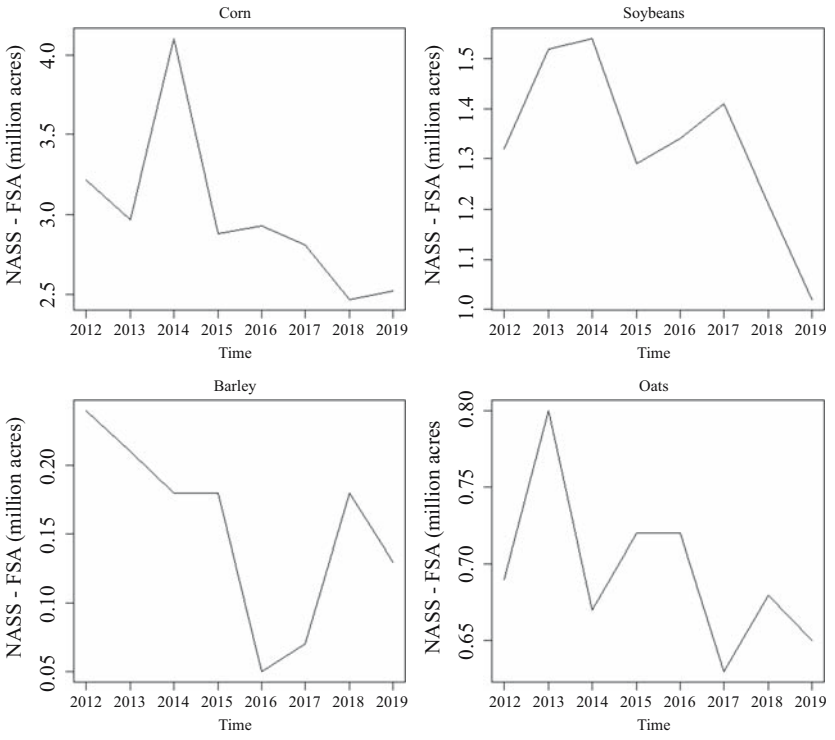


Fig. 2. The differences of US-level planted acreage estimates of several commodities between NASS and FSA.

Figure 3 indicates that the points in the plot of the survey estimates in log10 scale versus the FSA data in log10 scale are scattered around the 45 degree line. Some of the survey estimates are one or two standard deviations below the corresponding FSA or RMA data. This introduces difficulties for models without constraints to preserve the relationships. However, inequality constraints must be incorporated into the model so that all known relationships are satisfied at all levels before NASS can rely on model-based estimates as the foundation for the final official estimates.

3. Models

Bayesian area-level and sub-area level models are popular in small area estimation. They are excellent reproducible tools that combine survey data and auxiliary data to produce reliable estimates for areas. In this paper, models with constraints are considered based on Nandram et al. (2022). For comparison, the model without constraints are from Erciulescu et al. (2020).

Two model assumptions are made for both constrained and unconstrained model. First, it is assumed that the sampling variances are known and valid estimates from the survey summary in both area-level and subarea level sampling models. The modeling strategies are developed to deal with the crop county estimates including different commodities in all states covered by CAPS. Whereas Erciulescu et al. (2019) developed and compared models for direct estimates scaled by the sample sizes with a hierarchy for sampling variances, here

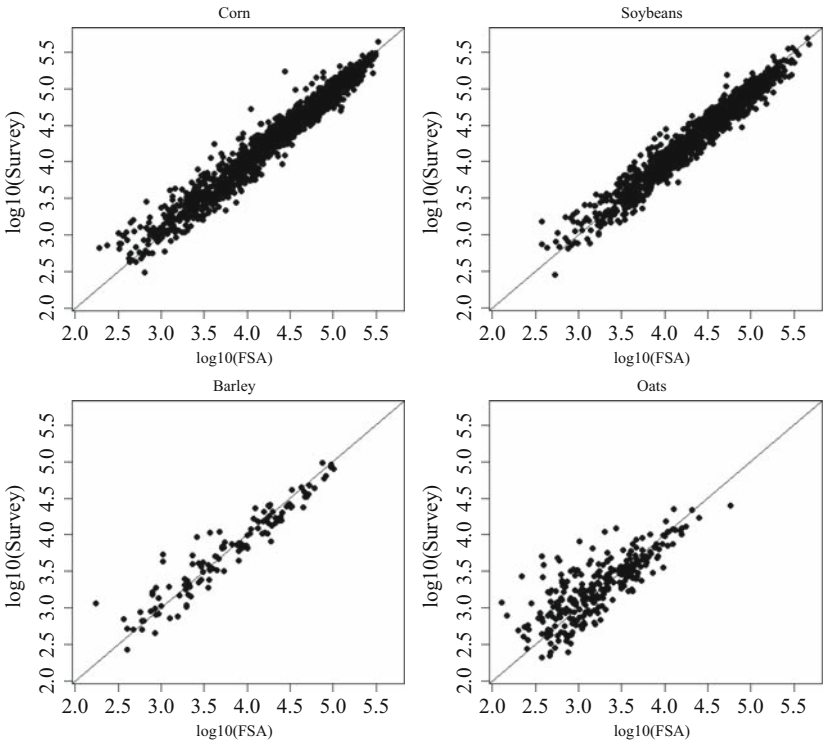


Fig. 3. The county-level planted acreage estimates (log10 scale) of several commodities for CAPS and FSA in all eligible counties.

we adopted the models for the direct estimates described in Erciulescu et al. (2020). Adding an extra model assumption for the sampling variance is not feasible, and so we assume that the sampling variances are fixed to avoid computational difficulties. Second, assuming normality of direct estimates is a practical method with good performance that provides estimates for counties with sample sizes as small as one and zero. This is impossible under the model specification in Erciulescu et al. (2019) because the sample sizes are denominators in the models. On the other hand, in our case, each county has its own unique inequality constraint and the sum needs to satisfy another benchmarking constraint.

In this section, models, with and without constraints, are presented and applied in a case study of 2014 corn data. They are illustrated for one state and one commodity, that is, all parameters are state and commodity-specific. The area-level model without inequality constraints was first introduced by Fay and Herriot (1979), where an area represents a county. The sub-area level models without inequality constraints were discussed by Fuller and Goyeneche (1998) and Torabi and Rao (2014) as an extension of FH model. Nandram et al. (2022) propose and discuss both area and sub-area level models to address the inequality constraints into the models.

3.1. Models Without Constraints

Erciulescu et al. (2020) discussed and applied a hierarchical Bayesian sub-area model to estimate the number of planted and harvested acres. In their paper, the state benchmarking

constraint is handled by ratio benchmarking in the output analysis, but inequality constraints are not addressed either in the model or in the output analysis. In this article, this model without inequality constraints is referred to the unconstrained model and several comparisons between this type of model and models with inequality constraints (constrained models) will be presented in Section 4.

In the sub-area level models, an area is an ASD and a subarea is a county. Let $i = 1,\ldots,m$ be an index for m ASDs in the state and $j = 1,\ldots,n_i$ be an index for the county in the i^{th} ASD. The survey estimate of planted acreage in county j in district i is denoted by $\hat{\theta}_{ij}$ and the associated survey variance is $\hat{\sigma}^2_{ij}$. The total number of counties in a state is $\sum_{i=1}^{m} n_i$. The auxiliary data used in the models are x_{ij}, including an intercept.

The sub-area hierarchical Bayesian model is

$$\begin{aligned}
\hat{\theta}_{ij}|\theta_{ij},\hat{\sigma}^2_{ij} &\overset{ind}{\sim} N(\theta_{ij},\hat{\sigma}^2_{ij}), i = 1,\ldots,m,\\
\theta_{ij}|\beta,\sigma^2_{\mu} &\overset{ind}{\sim} N(x'_{ij}\beta + v_i,\sigma^2_u), j = 1,\ldots,n_i,\\
v_j|\sigma^2_v &\overset{ind}{\sim} N(0,\sigma^2_v),
\end{aligned} \tag{1}$$

where $(\beta,\sigma^2_{\mu},\sigma^2_v)$ is a set of nuisance parameters. The county-level FSA and RMA planted acreage data are highly correlated. To avoid the multicollinearity problem, we choose to use the maximum of these two data sources. Thus, the vector of regressors for the county j with in the district i consists of $x_{ij} = (1,\max(\mathrm{FSA}_{ij},\mathrm{RMA}_{ij}))'$.

Note that the above sub-area level model without area level (ASD) effects, v_i, reduces to the basic area-level FH model without constraints, that is,

$$\begin{aligned}
\hat{\theta}_{ij}|\theta_{ij},\hat{\sigma}^2_{ij} &\overset{ind}{\sim} N(\theta_{ij},\hat{\sigma}^2_{ij}), i = 1,\ldots,m,\\
\theta_{ij}|\beta,\sigma^2_{\mu} &\overset{ind}{\sim} N(x'_{ij}\beta,\sigma^2_{\mu}), j = 1,\ldots,n_i.
\end{aligned} \tag{2}$$

A diffuse prior is adopted to the coefficients β, that is, a bivariate normal prior distribution with fixed and known mean and variance and covariance matrix $\beta \sim MN(\hat{\beta}, 1000\ \hat{\sum}_{\hat{\beta}})$. Here, $\hat{\beta}$ are the least squares estimates of β obtained from fitting a simple linear regression model of the county-level survey estimates on the auxiliary data x_{ij} and $\hat{\sum}_{\hat{\beta}}$ is the estimated covariance matrix of $\hat{\beta}$. The proper diffused prior on β is used because it provides a degree of computational stability. The prior distributions for σ^2_{μ} and σ^2_v are Uniform $(0,10^{10})$ and Uniform $(0,10^{10})$. The discussion in Browne and Draper (2006) motivates the use of an uniform prior distribution for the random-effect variance components.

3.2. Models With Constraints

Because of the advantage of shrinkage estimation in small area models without constraints discussed in Subsection 3.1, smaller survey estimates are likely to be pulled upwards. This will help to meet the bounds, but it does not solve the problem. As discussed in Subsection 2.2, the county-level estimates must be larger than the corresponding FSA and RMA

planted acres data. If the model does not incorporate inequality constraints, the final estimates do not necessarily cover the lower bounds in all cases. The inequality constraints need to be incorporated in the models. In this section, the hierarchical Bayesian models with inequality constraints by Nandram et al. (2022) are discussed.

First, inequality constraints between the true parameter θ_{ij} of interest and administrative values need to be included in the model, that is,

$$\theta_{ij} \geq c_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i, \tag{3}$$

where the c_{ij} is fixed known quantity.

In our application on planted acres, $c_{ij} = \max(\text{FSA}_{ij}, \text{RMA}_{ij})$ is the maximum value between FSA and RMA corresponding values in the same county. Notice that in Figure 3, some of the survey estimates are one or two standard deviations below their corresponding c_{ij}, thereby creating some difficulties for the model estimates to do the same. The benchmarking constraint creates an additional challenge because the state target may be only slightly larger than the state total from administrative data, $c = \sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij}$. This may be a tight condition, as discussed in Cruze et al. (2019).

In addition, under NASS's top-down approach, the benchmarking constraint needs to be considered as well. In this article, we fit Bayesian models using Markov chain Monte Carlo (MCMC) simulation. After model fitting, a series of MCMC samples are obtained to construct the posterior summaries of interest. We perform ratio benchmarking in each iteration of the MCMC samples. Erciulescu et al. (2020) discussed and applied the ratio benchmarking adjustment method at the (MCMC) iteration level in the output analysis to address the county-state benchmarking constraint. It provides a suitable benchmarking adjustment to ensure consistency of county-level estimates with the state target efficiently.

Let $\tilde{\theta}_{ij}^{B}$ be the adjusted model estimate for county j in district i. Let $\theta_{ij,k}$ denote the draw of θ_{ij} and $\theta_{ij,k}^{B}$ denote the adjusted (after benchmarking) draw, where k denotes the draw from the posterior distribution and $k = 1, \ldots, K$. Let a be the benchmarking state target.

The arithmetic mean of the MCMC samples is used to construct the point estimates of interest. After the ratio benchmarking adjustment,

$$\tilde{\theta}_{ij}^{B} = \frac{1}{K}\sum_{k=1}^{K}\theta_{ij,k}^{B} = \frac{1}{K}\sum_{k=1}^{K} r_k \theta_{ij,k}, \tag{4}$$

where r_k is the adjusted ratio at iterate level and the ratio r_k is

$$r_k = a \times \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij,k}\right)^{-1}. \tag{5}$$

Therefore, the following relationship holds for state benchmarking constraint,

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij}^{B} = a. \tag{6}$$

However, we need to make sure the adjusted final estimate $\tilde{\theta}_{ij}^{B}$ can satisfy inequality constraint as well. Given Equation (3), the inequality constraint can be preserved for $\theta_{ij,k}$ in each k^{th} iteration. If $r_k \geq 1$ for each k, the following relationship follows from combining

Equations (3) and (4):

$$\tilde{\theta}_{ij}^{B} = \frac{1}{K}\sum_{k=1}^{K} r_k \theta_{ij,k} \geq \frac{1}{K}\sum_{k=1}^{K} \theta_{ij,k} \geq \frac{1}{K}\sum_{k=1}^{K} c_{ij} \geq c_{ij}. \tag{7}$$

Therefore, when $r_k \geq 1$ $\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij,k} \leq a\right)$ for each iteration k, it follows that that all model estimates are raked up, and no individual county's inequality constraint will be violated.

Based on the discussion above, θ_{ij} should be drawn subject to the constraints

$$\theta_{ij} \geq c_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i, \sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij} \leq a \tag{8}$$

to address both inequality and benchmarking constraints in the models.

According to the constraints Equation (8),

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij} \leq \sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij} \leq a. \tag{9}$$

Therefore, the support of θ_{ij} given $\theta_{(ij)}$ is

$$\mathcal{T} = \left\{\theta_{ij} : \max\left(c_{ij,}\sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij} - \sum_{i'=1,i'\neq i}^{m}\sum_{j'=1,j'\neq j}^{n_i}\theta_{i'j'}\right) \leq \theta_{ij} \leq a - \sum_{i'=1,i'\neq i}^{m}\sum_{j'=1,j'\neq j}^{n_i}\theta_{i'j'}\right\}, \tag{10}$$

where the lower bound c_{ij} is known and fixed and $i = 1, \ldots, m;\ j = 1, \ldots, n_i$.

To preserve the relationships, the constraint Equation (10) is added to the FH model and the sub-area model in the priors to get the joint posterior density of θ_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. This problem falls under the general heading of constraint problems in statistics (Nandram et al. 1997).

Therefore, the sub-area hierarchical Bayesian model with constraints is proposed as

$$\begin{aligned}
\hat{\theta}_{ij}|\theta_{ij}, \hat{\sigma}_{ij}^2 &\overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2), j = 1, \ldots, n_i, \\
\theta_{ij}|\beta, \sigma_\mu^2 &\overset{ind}{\sim} N(x_{ij}'\beta + v_i, \sigma_\mu^2),\ \theta_{ij} \in \mathcal{T}, \\
v_i|\sigma_v^2 &\overset{ind}{\sim} N(0, \sigma_v^2), i = 1, \ldots, m,
\end{aligned} \tag{11}$$

where $\mathcal{T}$ denotes the support in Equation (10) of θ_{ij} such that both the benchmarking constraint and inequality constraints are satisfied. Here, $(\beta,\sigma_\mu^2,\sigma_v^2)$ is a set of nuisance parameters and $x_{ij} = (1, x_{ij1}, \ldots, x_{ijp})$ is the vector of covariates and the intercept. In the constrained model, the vector of regressors for the county j with in the district i are the same with those in the unconstrained model, that is, $x_{ij} = (1, \max(\mathrm{FSA}_{ij}, \mathrm{RMA}_{ij}))'$. Note that the above sub-area level model without sub-area level (ASD) effects, v_i, reduces to the area-level FH model with constraints, that is,

$$\hat{\theta}_{ij}|\theta_{ij}, \hat{\sigma}_{ij}^2 \overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2), j = 1, \ldots, n_i, \tag{12}$$

$$\theta_{ij}|\beta, \sigma_\mu^2 \overset{ind}{\sim} N(x'_{ij}\beta, \sigma_\mu^2),\ \theta_{ij} \in \mathcal{T}.$$

A diffuse prior is adopted to the coefficients β, the same as the prior mentioned in Subsection 3.1. The prior distributions in subarea-level model for σ_μ^2 and σ_v^2 are Uniform $(0, 10^{10})$ and Uniform $(0, 10^{10})$, respectively and the prior distribution in area level model for σ_μ^2 is Uniform $(0, 10^{10})$. Notice that without benchmarking constraint based on ratio benchmarking, the θ_{ij} s are not correlated a priori, but they are correlated a posteriori because of the common parameters over areas. With benchmarking constraint, they are correlated because they must add up to the state target a.

The methodology for creating the state targets guarantees that a state target a is greater than or equal to the administrative state total c. That is, $a = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{\theta}_{ij}^{B} \geq \sum_{i=1}^{m}\sum_{j=1}^{n_i}c_{ij} = c$. Therefore, there are feasible solutions to the inequality constraint problem in Equation (8), and a feasible solution clearly depends on the state target and the FSA and RMA data. As discussed in Subsection 2.2, most of the survey estimates are within two standard deviations of the bounds, but many of the smaller ones are much further below the bounds. If the model does not incorporate inequality constraints, the final model estimates do not necessarily cover the lower bounds in all cases. Therefore, inequality constraints need to be incorporated in the models to provide not only reliable but also coherent estimates.

4. Case Study

Three states, Illinois (IL), Ohio (OH), and Pennsylvania (PA), are considered in the case study. The four models discussed in Section 3 are compared: the sub-area level model with inequality constraint, the area-level model with inequality constraint, the sub-area level model without inequality constraint and the area-level model without inequality constraint. All models are fit using the administrative data sources described in Subsection 2.2.

All models produce 2014 CAPS estimates of planted acres for corn in IL, OH, and PA. FSA and RMA administrative data in IL usually have very high coverage rates of the planted acres for corn in each county. But in some specific counties in OH, both sets of administrative data have relatively low coverage rates for planted acres. In PA, administrative data in many counties have low coverage rates for planted acreage estimates. Therefore, these three states have different features. The model performance is evaluated for all different scenarios.

As mentioned in Subsection 2.2, the county-level survey estimates did not automatically cover all FSA and RMA administrative data. The relationship between survey estimates and the corresponding lower bounds based on administrative data (the maximum of FSA and RMA data) is displayed in Figure 4. The plotted pairs of survey estimates and administrative data are scattered around the 45 degree line. Around 31% of the county-level survey estimates cover FSA and RMA for IL. About 56% of the survey estimates cover FSA and RMA for OH. About 83% of the survey estimates cover FSA and RMA for PA.

In Subsection 4.1, a summary of the model fitting process is provided. Subsection 4.2 includes the internal checks for all four models. Several diagnostic tools are explored to

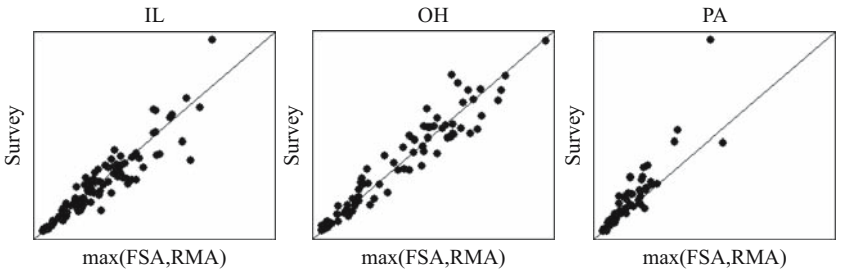


Fig. 4. The county-level planted acreage estimates of corn for CAPS and the lower bounds in IL, OH, and PA.

check the adequacy of the models. External checks between model estimates, survey estimates and official statistics from NASS are presented for all models in Subsection 4.3.

4.1. Model Estimation

All four models are applied to all counties with positive data in one state for which $\left(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij}\right)$ are available. In IL, there are 102 counties and 9 ASDs in the CAPS samples for planted acreage. In OH, there are 88 counties and 9 ASDs. In PA, there are 65 counties and 9 ASDs.

MCMC simulation method is used to fit all four hierarchical Bayesian models using R and JAGS (Plummer 2003). In each model, three chains are run for our MCMC simulation. Each chain contains 50,000 Monte Carlo samples, and the first 15,000 iterates are discarded as a burn-in to improve the mixing of each chain. After that point, 35,000 further iterations were produced for each of the three chains. In order to eliminate the correlations among neighboring iterations, those iterations are thinned by taking a systematic sample of 1 in every 35 samples. Finally 1,000 MCMC samples in each chain are obtained for constructing the posterior distributions of all the parameters, the nuisance parameters and the parameters for the planted acres.

Markov Chain Monte Carlo (MCMC) methods have been used to approximate the posterior marginals in Bayesian Hierarchical models and are computationally intensive if models are complicated and intractable. Computation time is one key factor when candidate models are evaluated for production especially for crops county estimates project involving multiple commodities for all related counties in US. As mentioned before, all models are fit by MCMC simulation using RJAGS. The computation time in reaching convergence for the different parameters in the unconstrained models is one to two minutes for each state and each commodity depends on the sizes of the data. But the computational time for the constrained models are five to six minutes since two inequality constraints nested with parameters are incorporated in the models. Their posterior distributions are more complicated than the unconstrained models, involving truncated normal distributions. Therefore, it takes more times to fit the constrained models than unconstrained ones. However, the computation time to produce county-level estimates with associated uncertainties is acceptable in current production procedure.

Convergence diagnostics are conducted. The convergence is monitored using trace plots, the multiple potential scale reduction factors ($\hat{R}$ close to 1) and the Geweke test of

stationarity for each chain (Gelman and Rubin 1992; Geweke 1992). Also, once the simulated chains have mixed, the effective number of independent simulation draws to monitor simulation accuracy is determined. Effective sample sizes and the $\hat{R}$ are shown in Tables 1 and 2, resulting in good convergence for all four models: area and sub-area models with inequality constraint (C) and without constraint (NC) for IL, OH, and PA. The values of $\hat{R}$ of most coefficient parameters are close to 1. The effective sample sizes of coefficient parameters in sub-area level models are 3,000 and those in area-level models are around 2,000 for IL. The effective sample sizes vary from 1,100 to 3,000 for OH. The effective sample sizes vary from 1,900 to 3,000 for PA.

4.2. *Internal Check*

Several diagnostic tools are available to check the adequacy of all four models to the observed data considered in this article. First, the fit of the models to the data is assessed using Bayesian predictive checks. If a model fit is adequate to all observations $\hat{\theta}$, replicated values θ_{rep} that generated data from the model would be similar to observations. We

Table 1. Sub-area level models: Effective sample sizes (ESS) and $\hat{R}$ for 2014 IL, OH and PA corn.

State	Parameters	ESS		$\hat{R}$	
		C Sub-area	NC Sub-area	C Sub-area	NC Sub-area
IL	β_0	3000	3000	1.001	1.001
	β_1	3000	3000	1.001	1.002
	σ^2_μ	1900	2100	1.004	1.012
	σ^2_v	3000	3000	1.001	1.003
OH	β_0	1800	2800	1.006	1.006
	β_1	3000	1400	1.006	1.002
	σ^2_μ	2000	1200	1.007	1.007
	σ^2_v	2300	1800	1.003	1.003
PA	β_0	2400	2800	1.004	1.003
	β_1	3000	3000	1.001	1.001
	σ^2_μ	1900	2000	1.019	1.007
	σ^2_v	2500	3000	1.011	1.009

Table 2. Area level models: Effective sample sizes (ESS) and $\hat{R}$ for 2014 IL, OH, and PA corn.

State	Parameters	ESS		$\hat{R}$	
		C Area	NC Area	C Area	NC Area
IL	β_0	1500	1700	1.010	1.002
	β_1	2000	1900	1.002	1.002
	σ^2_μ	2100	2500	1.008	1.004
OH	β_0	2300	3000	1.007	1.001
	β_1	1700	1100	1.009	1.017
	σ^2_μ	1900	1200	1.008	1.007
PA	β_0	2400	2400	1.009	1.011
	β_1	2600	2700	1.011	1.005
	σ^2_μ	2800	3000	1.001	1.003

calculate the Bayesian predictive p-value (BPP) to measure the adequacy of all models to the data from Rubin (1984), Meng (1994), and Gelman et al. (2013). The Bayesian posterior predictive *p*-value (BPP) is defined as

$$p = Pr\left(T\left(\theta^{rep}, \Omega\right) > T\left(\hat{\theta}, \Omega\right)|\hat{\theta}\right), \tag{13}$$

where discrepancy function, $T(\theta, \Omega)$, is selected as $T(\theta, \Omega) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\left(\theta_{ij} - E(\theta_{ij}|\hat{\theta})\right)^2}{Var\left(\theta_{ij}|\Omega\right)}$ and Ω are the nuisance parameters in each model. The *p*-value is the probability of the sum of square residuals based on replicated estimates larger than the one from observed data. If the value is extreme, smaller than 0.05 or larger than 0.95, it indicates a discrepancy between the model and the data, meaning the model is not adequate. The BPP for each model is presented in Table 3. For IL, the BPPs in the area-level and subarea level models with constraints are 0.663 and 0.504, respectively, which are not close to 0 or 1. The models without constraints have high BPP, 0.903 and 0.947, respectively. Those BPP are close to 0.95. Similar results show for OH in Table 5. Noticed that the sub-area level model without constraints for OH is 0.967, which is a borderline case. It indicated that the model is not adequate when comparing with survey estimates. The model's predictions are "biased" to be too high. For PA, all BPPs are smaller than 0.5 but they are not close to 0. However, models cannot be ranked based on BPPs.

Another goodness-of-fit measure for models is the deviance information criterion (DIC) (Spiegelhalter et al. 2002) shown in Table 3. It is not well suited to make the model selection based on DICs between constrained and unconstrained models. In particular, we consider the following type of comparison based on DIC only: between sub-area models and area models within either constrained models and unconstrained models. Table 3 shows the DICs from sub-area models are slightly smaller than those in area-level models. They indicate that the sub-area level models are better than the area-level models because sub-area models can borrow information from both area and sub-area levels.

Therefore, based on DIC diagnostics, sub-area level models are better than the area level models. To check model performance between sub-area level constrained and unconstrained models, external comparisons are discussed in the next section.

4.3. External Check

Internal checks show that sub-area level models have slightly smaller DICs than area-level models. Comparisons between area level and sub-area level unconstrained models and

Table 3. DICs and BPPs for constrained and unconstrained models.

Type	Model	DIC		BPP	
		C	NC	C	NC
IL	Sub-area	2334.6	2285.3	0.504	0.947
	Area	2335.7	2285.2	0.633	0.903
OH	Sub-area	1881.1	1766.7	0.331	0.967
	Area	1884.7	1776.4	0.248	0.908
PA	Sub-area	1613.8	1313.7	0.551	0.178
	Area	1618.4	1281.1	0.151	0.111

comparisons between area level and sub-area level constrained models are fine. However, the internal checks considered are not appropriate in terms of the model comparison of both sub-area level constrained and unconstrained models. Therefore, in this section, several external checks are conducted. In the guidance provided in NASEM (2017), the recommendation related the external comparisons is to use published estimates in assessing the quality and reasonableness of the model-based county-level estimates, especially at the research stage. In addition, before NASS can adopt a model-based approach to producing crops county estimates, the model must incorporate all known relationships. The inequality constraints check is another important factor in our evaluation.

First, the inequality check between the final model estimates of planted acres and the corresponding FSA and RMA administrative data is conducted for each model. Figure 5 shows the results in the unconstrained models for IL and OH. Counties in white indicate that the corresponding model estimates are smaller than FSA and RMA data. Counties in gray mean that their estimates are larger than the maximum of both FSA and RMA administrative data. In unconstrained model, 34 out of 102 counties in IL, 8 out of 88 counties in OH, and 3 out of 65 counties in PA do not satisfy with the constraints. However, based on the constrained model setting, all counties in both states satisfy the constraints after ratio benchmarking.

For the unconstrained model, the coverage rate on administrative data depends on the relationship between survey estimates and the administrative data. In PA, when many administrative data are smaller than survey estimates, only a few model-based estimates of planted acres are smaller than the administrative data.

Figure 6 shows comparison of constrained model estimates (denoted C) and unconstrained model estimates (denoted NC) relative to the FSA administrative totals in each county. In each panel, counties have been sorted on the horizontal axis in ascending order by number of CAPS survey reports collected and assigned a corresponding index value, for example, ranging from 1 to 102 for all counties in IL. Within each state, all modeled county estimates are benchmarked to the same fixed state total. On the vertical axis, values greater than one indicate that the estimated county acreage covers the corresponding FSA administrative data. Eliciting an acceptable tolerance below the FSA or RMA acreages has been difficult; at 640 acres (approximately 259 hectares) to the square mile, even some of the apparently modest differences below one become points of

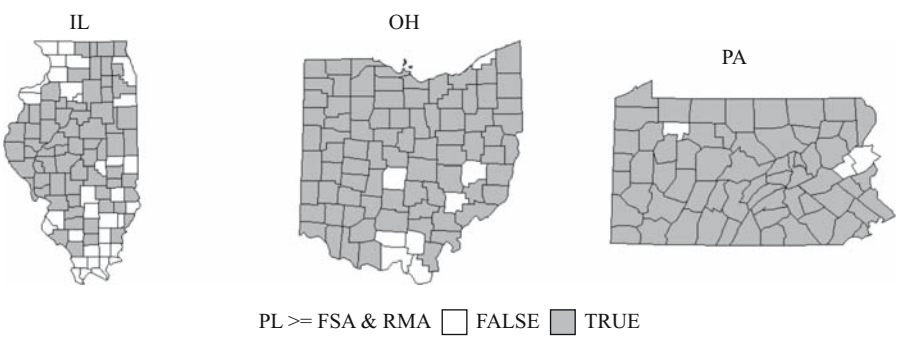


Fig. 5. Inequality check for unconstrained models for IL, OH, and PA.

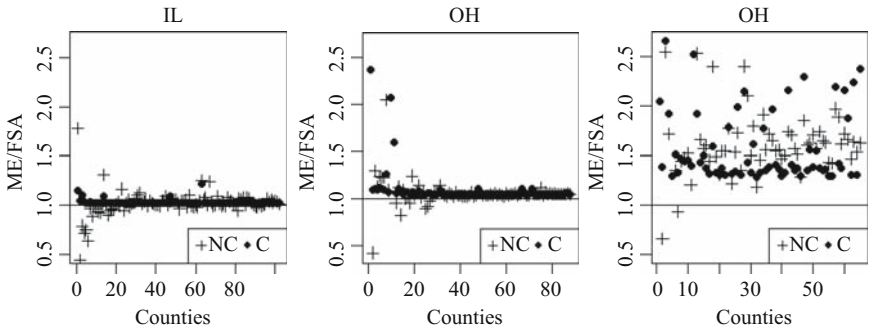


Fig. 6. Ratio of constrained model estimates (C) and unconstrained model estimates (NC) to FSA administrative acreage for IL, OH, and PA.

concern as the unconstrained model estimate begins to disagree with minimum amounts of planting activity on record in FSA's accurately geolocated database. The rightmost panel for PA points to some of the differences by state and commodity in NASS's estimation program. It speaks to the importance of quantifying the uncertainty of estimates when official statistics based on a blend of data may have properties more like administrative data in some scenarios, and more like survey or unconstrained model estimates in others.

In addition, both model-based estimates and survey estimates are compared to the published estimates. Let $\tilde{\theta}^{\text{NC}}_{\text{MERB}}$ be the unconstrained (NC) model estimates after ratio benchmarking and $\tilde{\theta}^{\text{C}}_{\text{MERB}}$ be the constrained (C) model estimates after ratio benchmarking. Let $\tilde{\theta}^{\text{DE}}_{\text{MERB}}$ be the survey (DE) estimates. The absolute relative differences between those estimates and published estimates,

$$\text{ARD} = \frac{|\tilde{\theta}^{t}_{\text{MERB}} - \text{Published}|}{\text{Published}}, \tag{14}$$

are calculated and presented, where $t =$ NC, C, DE. A small ARD is one key check on the performance of model-based point estimates. It is true that ARD will not be useful in the current year because the published estimates will not be available. However, it is a good check in a previous year when the official estimates are already decided and published. Note that we are using 2014 data and corresponding official estimates were published. Indeed, ARD is a check on models for future applications in the research stage recommended by NASEM (2017).

The posterior coefficients of variation (CV),

$$\text{CV} = \frac{\text{PSD}^{t}}{\tilde{\theta}^{t}_{\text{MERB}}}, \tag{15}$$

are calculated, where $t =$ NC, C, DE and the posterior standard deviation (PSD) is the corresponding posterior standard deviation of $\tilde{\theta}^{t}_{\text{MERB}}$, $t =$ NC, C, DE from different models and survey (see Table 4 and 5).

The sample sizes for planted acres in CAPS varies with county in each state. Many counties in these three states have relatively large sample sizes. However, many counties have small sample sizes as well. Small area models tend to improve the accuracy of estimates comparing to the accuracy of survey estimates, especially in areas with small

Table 4. 2014 IL corn planted acres: comparisons of ARDs and CVs among survey, sub-area unconstrained model and constrained model.

Sample size	Statistics	ARD (%)			CV (%)		
		DE	NC	C	DE	NC	C
Overall	Min	0.259	0.007	0.003	10.501	1.899	0.144
	Median	14.914	0.948	0.194	19.210	5.199	0.272
	Max	82.973	51.346	34.908	92.283	125.905	12.705
[0,30)	Min	0.259	0.622	0.273	25.315	20.544	1.466
	Median	16.585	13.530	0.978	42.421	34.905	2.187
	Max	66.174	51.346	34.908	92.283	125.905	12.705
[30,60)	Min	0.575	0.007	0.007	10.501	2.459	0.185
	Median	9.721	1.204	0.176	19.885	5.812	0.278
	Max	39.620	17.036	1.940	33.961	21.985	2.336
≥ 60	Min	7.474	0.096	0.003	9.108	1.899	0.144
	Median	33.990	0.646	0.196	15.731	3.151	0.214
	Max	82.973	2.032	1.199	53.570	5.522	1.740

Table 5. 2014 OH corn planted acres: comparisons of ARDs and CVs among survey, sub-area unconstrained model and constrained model.

Sample size	Statistics	ARD (%)			CV (%)		
		DE	NC	C	DE	NC	C
Overall	Min	0.002	0.103	0.093	8.754	1.043	0.473
	Median	12.942	2.394	2.575	22.292	3.670	0.797
	Max	114.123	95.376	49.858	100.000	104.411	89.816
[0,30)	Min	0.002	0.103	0.671	17.169	3.266	0.533
	Median	24.898	9.791	4.650	35.280	22.292	5.044
	Max	95.687	95.376	49.858	100.000	104.411	89.816
[30,60)	Min	1.574	0.136	0.093	10.224	1.206	0.473
	Median	12.699	2.266	2.191	19.468	2.546	0.660
	Max	114.123	10.968	14.864	33.072	29.994	10.548
≥ 60	Min	6.172	0.322	0.216	8.755	1.043	0.499
	Median	11.982	0.876	1.241	14.699	1.507	0.765
	Max	18.915	5.001	6.785	19.384	5.231	4.136

sample sizes. In order to examine the effect of sample size among our models, we split counties of IL, OH, and PA, respectively, into three groups according to their number of reports in CAPS: small sizes (less than 30); median sizes (between 30 and 60); large sizes (larger than 60). All statistics are shown in Table 4, 5, and 6 as well.

Among all counties in IL, the median ARD value between survey estimates and published estimates in IL is 14.914%. Substantial improvement can be noticed from both the constrained model and the unconstrained model. Again compared to published estimates, the median ARD value based on the constrained model is 194%, less than the median ARD value based on the unconstrained model, 0.948%. Moreover, the range of ARD values from the constrained model (0.003%, 34.908%) are much narrower than the range based on survey estimates (0.259%, 82.973%) and also less than those from the

Table 6. 2014 PA corn planted acres: comparisons of ARDs and CVs among survey, sub-area unconstrained model and constrained model.

Sample size	Statistics	ARD (%)			CV (%)		
		DE	NC	C	DE	NC	C
Overall	Min	0.128	0	0	9.685	2.874	2.874
	Median	12.537	12.272	11.198	22.644	14.237	11.584
	Max	73.300	733.127	33.318	70.941	75.132	44.113
[0,30)	Min	0.128	0	0	19.421	9.057	2.874
	Median	20.000	16.343	14.284	32.611	31.736	14.532
	Max	73.300	733.127	33.318	70.941	75.132	44.113
[30,60)	Min	2.520	1.71	0	14.165	6.778	2.795
	Median	11.200	15.943	10.910	21.143	12.924	11.484
	Max	54.600	41.624	19.247	38.203	27.496	21.941
≥ 60	Min	0.303	7.510	1.513	9.685	4.521	3.294
	Median	10.654	15.933	9.343	12.943	8.943	8.223
	Max	27.382	41.015	26.477	19.257	20.140	17.858

unconstrained model (0.007%, 51.349%). Therefore, for IL, the sub-area level model with constraints performs the best among the unconstrained model and survey estimates as measured by the ARD. In addition, Table 4 shows the ARD values based on the sample sizes of counties in IL. The ranges of ARD values based on both models are large for counties with small number of reports. ARD values from the constrained model are within 2% for median size counties but those from the unconstrained model are from 0.007% to 17.036 %. For large counties, the relative differences from all models are the smallest among all three types of counties. They are within 2% difference for constrained models and 3% from unconstrained model. As expected, all estimates are closer to the published estimates with increasing sample size. Overall, the comparisons of ARD values show that the constrained model increases the accuracy of the estimates significantly.

The CVs of the IL model and survey estimates are shown in Table 4. The sub-area level model can borrow information from both covariates and other counties in the district (sub-area) level. Therefore, the posterior CVs would have a greater reduction compared with the CVs of the survey estimates. The median CVs among all counties in IL are in decreasing order: survey, the unconstrained model and the constrained model. In the unconstrained model, the CVs of small size counties are the largest (20.544%, 125.905%).

The maximum estimated CVs exceeds the maximum of CVs from survey. The CVs of the constrained model are much smaller than those from survey and the unconstrained model. As expected, the CVs are smaller when sample sizes increase. In the model with inequality constraints, the maximum CVs is in the small size counties as well.

Table 5 shows the comparisons for OH. The median of ARDs between survey estimates and published estimates is 12.942%. Substantial improvement can also be noticed from both constrained and unconstrained models. The median ARD value between model-based estimates and the published estimates is around 2%. The smallest median of the relative differences is 2.394% in the unconstrained model. However, the range of ARD values from the constrained model is (0.093%, 49.858%), which is narrower than the one from the unconstrained model, (0.103%, 95.376%). Notice that the ranges of ARDs in OH are

larger than those in IL. The administrative data for OH are not stronger comparing with those in IL. In several counties, FSA and RMA administrative data have the undercoverage issue.

To examine the effect of sample sizes, OH is split into three groups and all statistics are presented in Table 5. The ranges of the ARD values based on models and the survey are relatively large in small size counties. Both model estimates are much closer to the published estimates. The estimates of the constrained model in small size counties are closest to the published estimates based on the range of the ARD values. However, the median ARD value from the constrained model is 1.241% for large size counties, which is larger than the one from the unconstrained model, 0.876%. The maximum ARD value is similar as well.

For the median size counties, constrained model tends to provide larger estimates compared with those from unconstrained model.

The CVs are compared among models and the survey estimates for OH as well. Similar to IL, the posterior CVs based on the models are small compared with the CVs from survey. The median CV in the unconstrained model is 3.67%, larger than the one in the constrained model. The maximum CV in the unconstrained model is the highest among models and survey. As expected, the CVs are smaller when sample sizes increase. The maximum of CVs is in small size counties as well. The CVs based on constrained model are much smaller than those of constrained model and survey. For OH, the range of CVs in model with inequality constraints are wider than those for IL.

Table 6 shows the comparisons for PA. The median of ARDs between survey estimates and published estimates is 12.272%. Slight improvement can be noticed from both constrained and unconstrained models in terms of median ARD value but big improvement from the constrained model when comparing with the maximum of ARD value. The median ARD value between model-based estimates and the published estimates is around 12.218% and 11.198% for unconstrained and constrained model respectively. However, the range of ARD values from the unconstrained model is (0%, 733.127%). That biggest ARD value, 733.127%, is in county with small sample sizes, far from the published estimate comparing with survey. the range of ARD values from the unconstrained model is (0%, 33.318%), which is narrower than both from survey and unconstrained model. Notice that the median of ARDs in PA is larger than those in IL and OH.

The administrative data for PA are not stronger comparing with those in IL and OH.

PA is also split into three groups and all statistics are presented in Table 6. The ranges of the ARD values based on models and the survey are relatively large in small size counties. Both model estimates are closer to the published estimates in small size counties. Models have better performance than the survey estimates when sample sizes are small. The estimates of the constrained model in small size counties are closest to the published estimates based on the range of the ARD values. However, the median ARD values from the constrained model in both medium and large size counties are only slightly smaller than the one in the survey. For unconstrained model, those are larger than the median ARD value in survey. As stated before, the administrative data in PA are not strong compared with those in IL and OH. If there was no inequality constraint, the model estimates would be affected by undercoverage from the administrative data when borrowing information from them.

The posterior CVs, based on the models, are smaller than those from the survey for PA. However, the reductions are much smaller than those in IL and OH. The loose lower bounds based on administrative data allow estimates to have more room to move. The median CV in the constrained model is 11.584%, smaller than the survey of 22.264% and the one in the unconstrained model, 14.237%. As expected, the CVs are smaller when sample sizes increase. The maximum of CVs is in small size counties as well.

5. Conclusion

NASS puts extensive research efforts on crops county estimate program aimed primarily to improve the precision of the estimates at county level while preserving the underlying relationships among the estimates and administrative data. Different small area estimation models are implemented to integrate multiple sources of auxiliary information with CAPS data. In this paper, models with inequality constraints are discussed and implemented to address the needs and challenges of the inequality and benchmarking constraints that NASS official statistics need to satisfy. That is, the county-level estimates of planted acreage should be greater than or equal to the corresponding administrative data while the total acreage of all available county-level estimates are equal to the state target.

We apply both sub-area and area-level models with inequality constraints to construct reliable and coherent county-level planted acreage estimates. In the case study of 2014 corn based on IL, OH, and PA, we show model diagnostics and provide internal checks between area-level models and sub-area level models. DICs indicate that the sub-area level models are slightly better than the area-level model. However, the residual-type internal checks are not very suitable for comparing the constrained and unconstrained models since our focus is to provide coherent estimates close to the official estimates but not to the survey estimates. For the model with inequality constraint, one would need to check it against external constraints.

Now more comparisons among both sub-area level model estimates and survey estimates are made. We pick three different states because their administrative data have different coverage rates. The results show that the performances of the constrained model are different among these three states. When many survey estimates are larger than the administrative data as shown in PA, the improvements are not that significant when comparing with IL and OH. However, the constrained model is still better than the survey and unconstrained model in terms of the external check. Inequality checks show that constrained model can preserve the relationships among estimates and administrative data. But this is not necessarily the case for the unconstrained model. It is true that including inequality constraints in some areas is unnecessary. But if we relax the inequality constraints for those counties that meet the constraint, they may not be satisfied in the model estimates. Not putting the constraint on the areas that are much higher than the lower bound is incoherent from a Bayesian view. Therefore, we have to put the constraints on all counties.

In addition, the statistics of ARD values show that the constrained model provides estimates closer to the published values than those from the unconstrained model as well as those from the survey, especially for IL. FSA and RMA are very significant covariates for the estimates of planted acres. Moreover, the associated measures of uncertainty (CVs)

from models are significantly smaller than the CVs of the survey estimates. The basic sub-area models can reduce the CVs while borrowing strength from auxiliary information and all counties in one district and all districts in one state. In addition, for the constrained model, the prior information based on the lower bound information from FSA and RMA data and the upper bound related to the state target reduce the CVs of the model-based estimates since estimates can be drawn only in the restricted support. Therefore, the performance of the sub-area level model with inequality constraints illustrates significant improvement of county-level estimates of planted acres in accuracy and precision.

Major ongoing and future research related to sub-area level constrained model involves the investigation of different auxiliary information. The auxiliary information considered here is the key data sources of planted acres (the combination of FSA and RMA administrative data). Future efforts will be on searching and applying other useful data sources to strengthen the model. Remote sensing data, NASS cropland data layer (see, Boryan et al., 2011), and weekly weather data are available at the county level. Variable selections should be investigated for different states and commodities because weather conditions influence the planting progress and the planted acres within different time periods based on different states and commodities.

In addition, missing data problems are another challenge for the application of the constrained model. In this article, case studies related to IL, OH, and PA, which do not have missing data in 2014 corn, are provided. However, it is not always the case for other states or other commodities. As mentioned in Subsection 2.1, CAPS is conducted for different commodities among all eligible states. In some cases, the survey may not indicate any planted area with respect to a particular commodity, but administrative data might represent some positive acres or vice versa. Erciulescu et al. (2020) used the nearest neighbor methods to impute missing data for either survey or covariates. This approach of imputing and borrowing information from previous year or the average of several years estimates are being explored. How to deal with missing data and provide reliable and coherent predictions are ongoing research.

6. References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83(401): 28–36. DOI: https://doi.org/10.2307/2288915.

Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. "Monitoring US agriculture: the US Depart ment of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program." *Geocarto International* 26(5): 341–358. DOI: https://doi.org/10.1080/10106049.2011.562309.

Browne, W.J., and D. Draper. 2006. "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models." *Bayesian Analysis* 1(3): 473–514. DOI: https://doi.org/10.1214/06-BA117.

Cruze, N.B., A.L. Erciulescu, B. Nandram, W.J. Barboza, and L.J. Young. 2019. "Producing Official County-Level Agricultural Estimates in the United States: Needs

and Challenges." *Statistical Science* 34(2): 301–316. DOI: https://doi.org/10.1214/18-STS687.

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2019. "Model-Based County Level Crop Estimates Incorporating Auxiliary Sources of Information." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(1): 283–303. DOI: https://doi.org/10.1111/rssa.12390.

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2020. "Statistical Challenges in Combining Survey and Auxiliary Data to Produce Official Statistics." *Journal of Official Statistics* 36(1): 63–88. DOI: https://doi.org/10.2478/jos-2020-0004.

Fay, R.E., and R.A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74(366a): 269–277. DOI: https://doi.org/10.2307/2286322.

Fuller, W.A., and J.J. Goyeneche. 1998. "Estimation of The State Variance Component." Unpublished manuscript.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2013. *Bayesian Data Analysis.* CRC press. DOI: https://doi.org/10.1201/b16018.

Gelman, A., and D.B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical science* 7(4): 457–472. DOI: https://doi.org/10.1214/ss/1177011136.

Geweke, J. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In *Bayesian Statistics* 4: 169–193. DOI: https://doi.org/10.21034/sr.148.

Good, D. 2014. "Comparing NASS and FSA Planted Acreage Data." *Farmdoc Daily* 4(9). Available at: https://farmdocdaily.illinois.edu/2014/01/comparing-nass-fsa-planted-acreage-data.html (accessed January 2014).

Meng, X.-L. 1994. "Posterior Predictive *p*-Values." *The Annals of Statistics* 22(3): 1142–1160. DOI: https://doi.org/10.1214/aos/1176325622.

Nandram, B., N.B. Cruze, and A.L. Erciulescu. 2022. Bayesian Small Area Models under Inequality Constraints with Benchmarking and Double Shrinkage. *Research Report RDD-22-02, National Agricultural Statistics Service*, USDA. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/ResearchReport_constraintmodel.pdf /(ACCESSEDE jYKY 2022).

Nandram, B., J. Sedransk, and S.J. Smith. 1997. "Order-Restricted Bayesian Estimation of The Age Composition of A Population of Atlantic Cod." *Journal of the American Statistical Association* 92(437): 33–40. DOI: https://doi.org/10.2307/2291447.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2017. *Improving Crop Estimates by Integrating Multiple Data Sources.* National Academies Press. DOI: https://doi.org/10.17226/24892.

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28(1): 40–68. DOI: https://doi.org/10.1214/12-STS395.

Plummer, M. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling." 3rd International Workshop on Distributed Statistical Computing (DSC 2003); March 21, Vienna, Austria: 124. Avaiable at: https://www.r-project.org/conferences/ DSC-2003/Proceedings/Plummer.pdf.

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation.* 2015 John Wiley/Sons, Inc. DOI: https://doi.org/10.1002/9781118735855.

Rubin, D.B. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician." *The Annals of Statistics*: 1151–1172. DOI: https://doi.org/10.1214/aos/1176346 785.

Sen, D., S. Patra, and D. Dunson. 2018. "Constrained Inference Through Posterior Projections". arXiv preprint. DOI: https://doi.org/10.48550/arxiv.1812.05741.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B* 64(4): 583–639. DOI: https://doi.org/10.1111/1467-9868.00353.

Torabi, M., and J.N.K. Rao. 2014. "On Small Area Estimation under A Sub-Area Level Model." *Journal of Multivariate Analysis* 127: 36–55. DOI: https://doi.org/10.1016/j.jmva.2014.02.

USDA FSA (U.S. Department of Agriculture Farm Service Agency). 2018. *FSA Handbook: Acreage and Compliance Determinations.* 2-CP, Revision 16. Available at: https://www.fsa.usda.gov/Internet/FSA_File/2cp16-a1.pdf (accessed April 2018).

USDA RMA (U.S. Department of Agriculture Risk Management Agency). 2017 The Risk Management Safety Net: Market Penetration and Market Potential-Analysis of the Federal Crop Insurance Portfolio. Available at: https://www.rma.usda.gov/-/media/RMA/Publications/Market-Penetration-and-Market-Potential-2017.ashx?la = en (accessed September 2017).

Received January 2021
Revised May 2021
Accepted January 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 733–765, http://dx.doi.org/10.2478/JOS-2022-0033

Timely Estimates of the Monthly Mexican Economic Activity

Francisco Corona[1], *Graciela González-Farías*[2] *and Jesús López-Pérez*[1]

In this article, we present a new approach based on dynamic factor models (DFMs) to perform accurate nowcasts for the percentage annual variation of the Mexican Global Economic Activity Indicator (IGAE), the commonly used variable as an approximation of monthly GDP. The procedure exploits the contemporaneous relationship of the timely traditional macroeconomic time series and nontraditional variables as Google Trends with respect to the IGAE. We evaluate the performance of the approach in a pseudo real-time framework, which includes the pandemic of COVID-19, and conclude that the procedure obtains accurate estimates, for one and two-steps ahead, above all, given the use of Google Trends. Another contribution for economic nowcasting is that the approach allows to disentangle the key variables in the DFM by estimating the confidence interval for the factor loadings, hence allows to evaluate the statistical significance of the variables in the DFM. This approach is used in official statistics to obtain preliminary and accurate estimates for IGAE up to 40 days before the official data release.

Key words: Dynamic factor models; global mexican economic activity indicator; google trends; nowcasts; pseudo real-time.

1. Introduction

Currently, the large amount of economic and financial time series collected over several years by official statistics agencies allows researchers to implement statistical and econometric methodologies that can generate accurate models to understand and anticipate macroeconomic phenomena. One of the most important events to anticipate are the movements of the gross domestic product (GDP) because they allow policy, and investment decisions to be made with greater certainty, in the expected scenario. For instance, if an economic contraction is foreseeable, businesses can adjust their expansion plans, governments can apply countercyclical policy, and consumers can adjust their spending patterns. In this context, "nowcasting" or timely estimates of macroeconomic time series is very relevant for policy makers and market participants to make informed decisions as a function of the short term of the economy. This is more evident for central

[1] Instituto Nacional de Estadistica y Geografia, Research Department, Av. Adolfo López Mateos 160, Col. San Ángel Inn, Alcaldía Álvaro Obregón, 01060, CDMX, Mexico. Emails: franciscoj.corona@inegi.org.mx and jesus.lopezp@inegi.org.mx

[2] Centro de Investigación en Matemáticas, A.C., Guanajuato, Guanajuato, Mexico. Email: farias@cimat.mx

Acknowledgments: The authors thankfully acknowledge the comments and suggestions carried out by the authorities of INEGI Julio Santaella, Sergio Carrera and Gerardo Villaseñor Leyva. The seminars and meetings organized by them were very useful to improve this research. To Elio Villasenõr who provided the Twitter mobility index, Manuel Lecuanda by the discussion about the Google Trends topics to be considered and Víctor Muñiz about the comments for the machine learning methods implemented. To Miguel Bedolla for the English grammar suggestions. Partial financial support from CONACYT CB-2015-25996 is gratefully acknowledged by Francisco Corona and Graciela González-Farías.

© Statistics Sweden

banks, whose influential role during the pandemic has required them to adapt their communication policies to the incumbent state of the economy. For example, Benchimol et al. (2021) document such behavior using state-ot-the-art text-mining methodologies to analyze both pre and post pandemics communications. They find out that communications and actions have been more reactive to the COVID-19 crisis than to the global financial crisis and the dot-com crisis.

A joint econometric, statistical, and computational analysis with transcendental information is relevant to be able to tackle the complexity of generating accurate estimations for key variables of the economy, above all when unexpected phenomenon sharply affects the short run of the economy. Hence, this article proposes a novel approach to best exploit the available relevant information, which would allow to generate informative decisions for policy, decision making and is of interest for the society in general. We focus on Mexico's economy by using econometric models that are frequently used in the literature, but in our case, we estimate them by exploiting not only traditional information but also nontraditional variables, as has been recently used in other econometric applications.

In this sense, it is clear that as new economic and financial information is released, the estimates for a certain period are constantly being updated as well; thus, repetitive GDP estimations are made for the same quarter because unexpected economical or non-economical events can drastically affect the economy in the short term. Hence the need to use not only traditional economic and financial information, but also nontraditional high-frequency indicators, such as news, topics extracted from the Internet searches, social networks analysis, and so on.

The seminal work of Varian (2014) has become an obligatory reference for the inclusion of high-frequency information in econometric analysis. Buono et al. (2017) is also an important reference to characterize the types of nontraditional data and the econometric methods usually employed to extract information from the data. Both studies conclude that online search information is appropriate when the goal is to generate nowcasts of macroeconomic time series. More specific, Google Trends topics, an up-to-date source of online search information tool which provides an index of Internet searches or queries by category and geography, is frequently used to predict economic phenomena. See for example Stephens-Davidowitz and Varian (2014) which provides a full review of this tool, and other analytical tools from Google, as applied in the social sciences.

The inclusion of nontraditional information in prediction models has been well investigated in the literature, with evidence mostly in favor of its value in improving predictions. For example, Choi and Varian (2012), among others, have shown the effectiveness of Google Trends to improve predictions in real time for some United States (US) variables as motor vehicles and parts and unemployment benefits, the Hong Kong tourism and the Australian consumer confidence. Recent applications include Goldsmith-Pinkham and Sojourner (2020), who nowcast the number of workers filing unemployment insurance claims in the US, Caperna et al. (2022), who develop random forest models to nowcast country-level unemployment figures for the 27 European Union countries and Ali et al. (2021), who analyzes online job postings in the US childcare market under stay-at-home orders.

Note that the term nowcast is also relevant from an econometric and statistical standpoint, because we can use a rich variety of information to model, from a multivariate point of view, macroeconomic and financial events, plus specific incidents that can affect the dynamics of GDP in the short run. In this sense, these facts are related to the literature of dynamic factor models (DFMs), because a large amount of time series is useful to estimate underlying common factors.

In this order of ideas, we claim that Google search data, when combined with appropriate statistical tools has highly predictive power when applied in nowcasting models based on DFMs. In particular, we are interested in introducing Google search data along with DFMs to nowcast the monthly GDP proxy, officially published in Mexico by the Mexican National Institute of Statistics and Geography (INEGI) through the Global Economic Activity Indicator (IGAE), which covers approximately 95% of the total economy. It is released two months after the reference month (INEGI. a). Thus, by generating nowcasts for IGAE in an accurate and timely manner, the GDP's nowcasts can be drastically improved according to Gálvez-Soriano (2020).

Nowcasting the Mexican economy becomes relevant because of its importance in the world economy and in North America. According to data from the World Bank, in 2020 the Mexican GDP was ranked fifteenth in the world and fourth in the Americas. Moreover, since 1994 Mexico has been on a trade agreement with the US and Canada, which is currently known as the Agreement between the United States of America, the United Mexican States, and Canada, and formerly known as the North America Free Trade Agreement. Consequently, the development of new approaches to predict the Mexican economy activity is of relevance not only for Mexico but for at least the three countries.

DFM and other methodologies have been used to nowcast or predict the Mexican economy. To generate timely estimates, Delajara et al. (2016) use small-scale mixed-frequency DFM to nowcast, backcast and forecast GDP figures. Also, in one of the first works along this line Corona et al. (2017a), estimated common trends in a large and non-stationary DFM to predict the IGAE levels two steps ahead and concluded that error prediction was reduced with respect to some benchmarking univariate and multivariate time-series models. Caruso (2018) focuses on international indicators, mainly for the US economy, to show that its nowcasts of quarterly GDP outperform the predictions obtained by professional forecasters. Recently, Gálvez-Soriano (2020) concluded that bridge equations perform better than DFM and static principal components (PC) when making the nowcasts of quarterly GDP. It is worth to mention, because it is related to timely GDP estimation, Guerrero et al. (2013) in which, based on vector autoregression (VAR) models, they generate rapid GDP estimates (and its three grand economic activities) with a delay of up to 15 days from the end of the reference quarter, while the official GDP takes around 52 days after the quarter closes. Guerrero's work is the main reference for GDP flash estimates. "Estimación Oportuna del PIB Trimestral." (INEGI. b).

In our article, we deviate from prior studies relevant for Mexico, by including nontraditional information to capture more drastic frictions that occur in the very short run, in a time span of one or two months. We identified that previous works focus only on traditional information, which bounds their capacity to predict the recent historical declines attributed to COVID-19 and the associated economic closures since March 2020. A possible reason in the previous Mexican econometric analysis for not including nontraditional

information is the fact that traditional variables actually perform well when there are not unexpected events. Additionally, our econometric analysis focuses on obtaining nowcasts for the monthly Mexican economic activity and gives an interpretation to its results from an economic point of view by using timely macroeconomic and financial time series plus other nontraditional time series commonly used in big data analysis, specifically Google Trends.

We address the peculiarities previously mentioned for the Mexican economic activity by making use of previous knowledge to fill in the identified gaps by: (1) building a timely and correlated database using traditional economic and financial time series and real-time nontraditional information, determining the latter relevant variables with least absolute selection and shrinkage operator (LASSO) regression and other methods of time series selection variables; (2) estimating the common factors using the two-step (2SM) methodology of Doz et al. (2011) and (3) training and testing univariate time series models with the DFM's common factors to select the best nowcasts and combine the statistically equal better nowcasts to generate the current estimates. Additionally, to increase the interpretability of the results, we determine the confidence intervals for both the factor loadings and the factor itself to analyze the importance of each variable and the uncertainty attributed to the estimation.

By providing this new modeling strategy we consider the contributions of this article to be timeliness, openness and interpretability. First, given the timely availability of the information that our approach uses, we can generate nowcasts of the IGAE up to 40 days before the official data release; thus, our approach becomes an alternative to obtaining IGAE's preliminary estimates, which are very important in official statistics. Second, this article illustrates to practitioners the empirical strategy to generate IGAE nowcasts step-by-step, so that any user can replicate the results for other time series. Third, we can disentangle the contribution of each timely time series with respect to the economic activity, that is, the relative importance of traditional and nontraditional variables that encompasses the DFM. In this way, our objective is to generate nowcasts that yield an economic explanation of the nowcasts, while we do not pretend to carry out estimates with a pure statistical model or a non-interpretive approach. Consequently, we consider our approach to be both a predictive and an explicative model, in the sense of Shmueli (2010).

Another contribution of this article is to analyze and model the Mexican economic activity in pseudo real-time during the previous and current stage of the COVID-19 pandemic, by evaluating the performance of this approach with respect to that of other approaches. Furthermore, this article provides methodological support for the Mexican Indicador Oportuno de la Actividad Económica reported as experimental statistics by (INEGI. c).

The remainder of this article is structured as follows. Section 2 presents the methodology considered to generate the nowcasts. Section 3 describes the data and the descriptive analysis. Section 4 contains the empirical results. Finally, Section 5 concludes the article.

2. Methodology

This section describes how we employ DFM to generate the nowcasts of the IGAE. First, we describe how we build the correlated database by using traditional and nontraditional

time series. Then, we report how the stationary DFM shrinks the complete dataset in the 2SM strategy to obtain the estimated factor loadings and common factors and the Onatski (2010) procedure to detect the number of common factors. Finally, we describe the nowcasting approach.

2.1. Building the Database

2.1.1. LASSO Regression: Selecting the Google Topics

The goal of using LASSO regression in our research is to select Google topics with most predictive power with respect to IGAE because, we do not have *a priori* knowledge about which ones are the most relevant across time. In contrast, regarding traditional data there exists a vast literature that indicates which variables are related with the economic activity. Hence, we propose an iterative algorithm described in detail in this subsection.

LASSO regression was introduced by Tibshirani (1996) as a new method of estimation, in linear models, that minimizes the residual sum of the squares (RSS) subject to the sum of the absolute value of the coefficients being less than a constant. In macroeconomic applications, Aprigliano and Bencivelli (2013) use the LASSO regression to select the relevant economic and financial variables in a large data set with the goal of estimating a new Italian coincident indicator.

Let $W_t = (w_{1t}, \ldots, w_{Kt})'$ be a $K \times 1$ vector of stationary and standardized variables. Consider the following penalized RSS:

$$\min_{RSS} = (y - W\beta)'(y - W\beta) + \lambda \sum_{j=1}^{K} |\beta_j|, \tag{1}$$

where $y = (y_1, \ldots, y_T)'$ is a $T \times 1$ vector, $\beta = (\beta_1, \ldots \beta_K)'$ is a $K \times 1$ vector, $W = (W_1, \ldots, W_T)'$ is a $T \times K$ matrix and λ is a tuning parameter that controls the shrinkage of the estimates.

Tibshirani (1996) considers that this problem is equivalent to minimizing the sum of squares subject to $f(\beta) = \sum_{j=1}^{K} |\beta_j| \leq c$, where c is a predetermined parameter that indicates the degree of regularization; in this case, the solution of Equation (1) is not closed, and it is obtained by convex optimization techniques. The LASSO solution has the following implications: (a) when $\lambda \rightarrow 0$, we obtain solutions similar to OLS, and (b) when $\lambda \rightarrow \infty$, $\hat{\beta}_{\lambda}^{LASSO} \rightarrow 0$. Therefore, the LASSO regression can perform as a variable selection method in linear models. Consequently, if λ is large, more coefficients tend to zero, selecting the variables that minimize the error prediction.

LASSO regression is related to Ridge regression, but the latter focuses on determining the tuning parameter, λ, that controls the regularization effect. Where $f(b) = \sum_{j=1}^{K} |\beta_j^2$, solution is $\hat{\beta}_{\lambda}^{Ridge} = \left(W'W - \lambda I_p\right)^{-1} W'y$. In practice, this solution never sets the coefficients to exactly zero; therefore, ridge regression cannot perform as a variable selection method in linear models, although its prediction ability is better than OLS.

Hence, for a sample K topics on Google Trends, the relevant topics $l = 0, \ldots, \zeta$, with $\zeta \geq 0$ are selected with the LASSO regression as follows:

1. Split the data for $t = 1, \ldots, T^* - H_g$, where H_g is the number of the last months equivalent to the sample size in the CV.

2. For $h = 1$ and for the sample of size $T^* - H_g + h$, estimate $\hat{\beta}_{\lambda,h}^{LASSO}$ in a grid between 10^{10} and 10^{-10}, that is, the rank of possible values of λ by using a fixed window, two periods in test set sample and a training percentage of 0.75. Compute the following vector of indicator variables:

$$\hat{\beta}_{j,h} = \begin{cases} 1 & \text{if } \hat{\beta}_{\lambda,h}^{LASSO} \neq 0 \\ 0 & \text{if } \hat{\beta}_{\lambda,h}^{LASSO} = 0 \end{cases}$$

3. Repeat 2 until H_g.
4. Define the $H_g \times K$ matrix, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$, where $\hat{\beta}_j = (\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,Hg})'$ is an $H_g \times 1$ vector.
5. Select the l significant variables that satisfy the condition $\hat{\beta}_l = (\hat{\beta}_{l \in j} | 1\hat{\beta} > \varphi)$, where φ is the $1-\alpha$ sample quantile of $1\hat{\beta}$ with $\mathbf{1}$ being a vector $1 \times H_g$ of ones.

With this procedure, we select the topics that frequently reduce the prediction error –out-of-sample– for the IGAE estimates during the last H_g months. We estimate the optimum A by using the `glmnet` package, from the R program, while the CV in the LASSO procedure is implemented in the `caret` package.

Once the ζ topics have been selected, we can refine our search by estimating DFMs with them and the macroeconomic and financial variables with them. The intention is to select the combination of topics $\binom{\zeta}{j}$ for $l = 0, \ldots, \zeta$, that minimizes the out-of-sample error estimation in a CV, when the obtained common factor is included in a regression model. This procedure is computationally expensive but can guide us to the most relevant Google topics for our nowcasting objectives in an empirical way.

Specifically, in each $h = 1, \ldots, H_g$, we selected the most relevant topics through the following heuristic algorithm:

1. Set $h = 1$.
2. Selecting ζ relevant topics, consider the different combination of matrices $\binom{\zeta}{j}$ for $l = 0, \ldots, \zeta$. For example, for $l = 1$ we have ς matrices. The case of $l = 0$ is when we do not consider Google topics.
3. By using also the traditional information, we estimate dynamic factors for each combination of topics and they are introduced in linear regressions estimating the MAEs for the recent past months $h* = 1, \ldots, H_g^*$.
4. We select the combination -or matrix-that consider the topics that minimize the MAE.
5. Repeat the procedure until $h = H_g$.

This procedure allows us to obtain the relevant topics in the past for each h, consequently, this is essentially as doing nowcasts in pseudo real time and not only for the incumbent period.

2.1.2. Transformations

We propose to carry out transformations on the variables of the DFM to orient the estimated common factors so that they are highly correlated contemporaneously with the IGAE, so that the estimated common factors are more appropriate to our nowcasting goals.

Therefore, we define the annual percentage variation of IGAE as $y^* = (y_1, \ldots, y_{T^*})$, where $T^* = T-2$; hence, to predict y^*, the macroeconomic, financial and nontraditional time series are collected in $X_i^* = (x_{i1}^*, \ldots, x_{iT^*}^*)$, and are transformed such that they satisfy the following condition:

$$X_i = \left(f\left(X_i^*\right)\Big|_{corr}^{\max}\left(f\left(X_i^*\right), y^*\right)\right). \tag{2}$$

Hence, we select the $f(X_i^*)$ that maximizes the correlation with y^*. Consider $f(\cdot)$ as follows:

1. None (n)
2. Monthly percentage variation (m): $\left(\frac{X_t^*}{X_{t-1}^*} \times 100\right) - 100$
3. Annual percentage variation (a): $\left(\frac{X_t^*}{X_{t-12}^*} \times 100\right) - 100$
4. Lagged (l): X_{t-1}^*

Consequently, each time series is transformed to have maximum correlation with the target time series.

2.2. *Dynamic Factor Model*

We use and specify a DFM to jointly summarize, the dynamics of the traditional and nontraditional time series by estimating the common factors and their contributions. Once estimated, the common factors are used in statistical prediction models to generate the timely estimates of the IGAE.

Hence, we consider a stationary DFM where the observations, X_t, are generated by the following process:

$$X_t = PF_t + \varepsilon_t, \tag{3}$$

$$\Phi(L)F_t = \eta_t, \tag{4}$$

$$\Gamma(L)\varepsilon_t = a_t, \tag{5}$$

where $X_t = \left(x_{1t}, \ldots, x_{Nt}\right)'$ and $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$ are $N \times 1$ vectors of the variables and idiosyncratic noises observed at time t. The common factors, $F_t = (F_{1t}, \ldots, F_{rt})'$, and the factor disturbances, $\eta_t = (\eta_{1t}, \ldots, \eta_{rt})'$, are $r \times 1$ vectors, with $r(r < N)$ being the number of static common factors, which is assumed to be known. The $N \times 1$ vector of idiosyncratic disturbances, a_t, is distributed independently of the factor disturbances, η_t, for all leads and lags, denoted by L, where $LX_t = X_{t-1}$. Furthermore, η_t and a_t, are assumed to be Gaussian white noises with positive definite covariance matrices $\Sigma_\eta = \text{diag}(\sigma_{\eta 1}^2, \ldots \sigma_{\eta r}^2)$ and Σ_a, respectively. $P = (p_1, \ldots, p_N)'$, is the $N \times r$ matrix of factor loadings, where, $p_i = (p_{i1}, \ldots, p_{ir})'$ is an $r \times 1$ vector. Finally, $\Phi(L) = \mathrm{I} - \sum_{i=1}^{k} \Phi L^i$ and $\Gamma = I - \sum_{j=1}^{s} \Gamma L^j$, where Φ and Γ are $r \times r$ and $N \times N$ matrices containing the VAR parameters of the factors and idiosyncratic components with k and s orders, respectively. For simplicity, we assume that the number of dynamic factors, r_1, is equal to r.

Alternative representations in the stationary case are given by Doz et al. (2011, 2012), who assume that r can be different from r_1. Additionally, when $r = r_1$, Bai and Ng (2004), Choi (2017) and Corona et al. (2020) also assume possible non-stationarity in the idiosyncratic noises. Barigozzi et al. (2015, 2016) assume possible non-stationarity in F_t, ε_t and $r \neq r_1$. Certainly, the literature to represent DFMs is vast. For example, Bai and Ng

(2002) and Bai (2004) focus on static ones, while Bai and Ng (2007) show the relationship between the static and the dynamic representation, in turn, Stock and Watson (2011) describe the dynamic model showing some estimation methods, including the PC method used in this work, and Forni et al. (2000) show the generalized DFM, where they make the static representation widely applicable. In our work, the dynamics of the common factors is specified as a VAR(p) model to refine the static estimation in a second step, as proposed by Doz et al. (2011), since it is important to assume that the idiosyncratic errors are stationary to obtain factor consistent estimates.

The DFM in Equations (3) to (5) is not identified. As we noted in the Introduction, the factor extraction used in this work is the 2SM; consequently, in the first step, we estimate the common factors by using PC to solve the identification problem and uniquely define the factors; we impose the restrictions $P'P/N = I_r$ and $F'F$ being diagonal, where $F = (F_1, \ldots, F_T)$ is $r \times T$. For a review of restrictions in the context of PC factor extraction, see Bai and Ng (2013).

Note that the transformations considered in this work do not have the goal of achieving stationarity as is frequently done in several empirical applications. Although intrinsically these transformations are stationary transformations regardless of whether $y*$ is stationary; in fact, the transformations m and α tend to be stationary transformations when the time series are $I(1)$, which is frequent in economics; see Corona et al. (2017b). Otherwise, it is necessary that $(f(X_i^*), y^*)$ are cointegrated in the context of Bai (2004). This is so because even though, some common factor, F_t, can be non-stationary, consistent estimations remain even when the idiosyncratic errors are stationary. In this way, we use the PANIC test (Bai and Ng 2004) to verify this assumption. Note that, for example, Bai and Ng (2004) and Barigozzi et al. (2016) obtain consistent estimates for F_t regardless of whether the idiosyncratic errors are non-stationary.

2.2.1. Two-Step Method for Factor Extraction

Giannone et al. (2008) popularized the usage of 2SM factor extraction to estimate the common factors by using monthly information with the goal of generating the nowcasts of quarterly GDP. However, Doz et al. (2011) proved the statistical consistency of the estimated common factor using 2SM. In the first step, PC factor extraction consistently estimates the static common factors without assuming any distributions, allowing weak serial and cross-sectional correlation in the idiosyncratic noises; see, for example, Bai (2003). In the second step, we model the dynamics of the common factors via the Kalman smoother, allowing idiosyncratic heteroskedasticity, a situation that occurs frequently in practice. In a finite sample study, Corona et al. (2020) show that with the 2SM of Doz et al. (2011) based on PC and Kalman smoothing, we can obtain closer estimates of the common factors under several data generating processes that can occur in empirical analysis, such as heteroskedasticity, serial and cross-sectional correlation in idiosyncratic noises. Additionally, following Giannone et al. (2008), this method is useful when the objective is nowcasting given the flexibility to estimate common factors, when all variables are not updated at the same time.

The 2SM procedure is implemented according to the following steps:

1. Set $\hat{P}$ as $\sqrt{N}$ times the eigenvectors corresponding to the r largest eigenvalues of $X'X$, where $X = (X_1, \ldots, X_T)'$ is a $T \times N$ matrix. By regressing X on $\hat{P}$ and using the identifiability restrictions, obtain $\hat{F} = X\hat{P}/N$ and $\hat{\varepsilon} = X - \hat{F}'\,\hat{P}'$. Then, compute the asymptotic confidence intervals for both factor loadings and common factors as proposed by Bai (2003).
2. Set the estimated covariance matrix of the idiosyncratic errors as $\hat{\Psi} = \text{diag}\big(\hat{\sum}_{\varepsilon}\big)$, where the diagonal of $\hat{\Psi}$ includes the variances of each variable of X; hence, $\hat{\sigma}_1^2$ for $i = 1, \ldots, N$.
3. Estimate a VAR(k) model by OLS to the estimated common factors, $\hat{F}$, re express the autoregressive coefficients as the VAR(1) model, denoted by $\hat{\Phi}$. Assuming that $f_0 \sim N(0, \Sigma_f)$, the unconditional covariance matrix of the factors can be estimated as $\text{vec}\,(\hat{\sum}_f) = (I_{r2} - \hat{\Phi} \otimes \hat{\Phi})^{-1}\,\text{vec}\,(\hat{\sum}_{\eta})$, where $\hat{\sum}_{\eta} = \hat{\eta}'\hat{\eta}/T$.
4. Write DFM in Equations (3) to (5) in state-space form, and with the system matrices $\hat{P}$,$\hat{\Psi}$, $\hat{\Phi}$, $\hat{\sum}_{\eta}$ and $\hat{\sum}_f$, use the Kalman smoother to obtain an updated estimation of the factors denoted by $\tilde{F}$.

In practice, X_t are not updated for all t; in these cases, we apply the Kalman smoother, $\tilde{F} = E(F_t|\Omega_T)$, where Ω_T is all the available information in the sample, and we take into account the following two cases:

$$\hat{\Psi}_i = \begin{cases} \hat{\sigma}_i^2 & \text{if } x_{it} \text{ is available,} \\ \infty & \text{if } x_{it} \text{ is not available.} \end{cases}$$

Empirically, when specific data on X_t are not available, Harvey and Phillips (1979) suggests using a diffuse value equal to 10^7; however, we use 10^{32} according to the package nowcast of the R program, see De Valk et al. (2019).

As commented in Stock and Watson (2011), the estimation of common factors based on PC factor extraction is consistent to the presence of breaks or time variation in the factor loadings. Additionally, in our case X_t is seasonally adjusted, that is, we have removed outliers, and the time series selection and transformations is oriented to maximizing the contemporaneous relationship with IGAE. The purpose is to summarize the common movements, including the drastic drops of the economy, thus, all the time series tend to have the same variability. On top of that, the ADF tests tend not to reject the null hypothesis of structural change. In our case, not rejecting the null hypothesis in PANIC tests computed to the idiosyncratic errors could point out the presence of possible structural change.

2.2.2. Determining the Number of Common Factors

To detect the estimated number of common factors, $\hat{r}$, Onatski (2010) proposes a procedure to use when the proportion of the observed variance attributed to the factors is small relative to that attributed to the idiosyncratic term. This method determines a sharp threshold, which consistently separates the bounded and diverging eigenvalues of $X'X/T$. Note that if X is centered and scaled, $X'X/T$ is the covariance matrix.

This algorithm is known as edge distribution, and Onatski (2010) provide the algorithm that generates a consistent estimator of r for any fixed $\delta > 0$, denoted by $\hat{r}$. Corona et al.

(2017b) shows that this method works reasonably well in small samples. Two important features of this method are that the number of factors can be estimated without previously estimating the common components and that the common factors may be integrated.

2.3. Obtaining Nowcasts

The nowcasting models are basically regressions with ARMA errors, in which the dependent variable is the annual percentage variation of the IGAE and the exogenous variable are the estimated common factors. We select the autoregressive and moving average orders in a time series CV period, so that, they minimize the one-step ahead estimation error. In this subsection, we describe in detail the algorithm.

Having estimated the common factors as described in Subsection 2.2.1 by using X_t, we estimate a linear regression model with autoregressive moving average (ARMA) errors to generate the nowcasts

$$y_t^* = a + b\tilde{F}_t + u_t, t = 1, \ldots, T-2, \tag{6}$$

where $u_t = \gamma(L)\phi(L)^{-1}u_t$ with $\gamma(L) = 1 + \sum_{j=1}^{q}\gamma_j L^j$ and $\phi(L) = 1 - \sum_{i=1}^{p}\phi_i L^i$. The parameters are estimated by maximum likelihood. Consequently, the nowcasts are obtained using the following expression:

$$\hat{y}_{T^*+h} = \hat{a} + \hat{b}\tilde{F}_{T^*+h} + \hat{u}_{T^*+h} \text{ for } h = 1, 2. \tag{7}$$

Note that Giannone et al. (2008) propose using the model with $p = q = 0$; hence, the nowcasts are obtained by using expression (7). In our case, we estimate different models by the orders $p = 0, \ldots p_{\max}$ and $q = 0, \ldots q_{\max}$; thus, the case of Giannone et al. (2008) is a particular case of this expression. Our interest is in selecting models with similar performance in a CV period. In this way, we carry out the following procedure:

1. Start with $p = 0$ and $q = 0$.
2. Estimate the nowcasts for $T^* + 1$ and $T^* + 2$, namely, $\hat{y}^{0,0} = \left(\hat{y}_{T^*+1}, \hat{y}_{T^*+2}\right)'$.
3. Split the data for $t = 1, \ldots, T^* - H_t$.
4. For $h = 1$ and for the sample of size $T^* - H_t + h$, estimate Equation (6), generate the nowcasts with expression (7) one step ahead, and calculate the errors and absolute error (AE) as follows:

$$e_1^{0,0} = y_{T^*-H_t+1} - \hat{y}_{T^*-H_t+1}$$

$$AE_1^{0,0} = |e_1^{0,0}|$$

5. Repeat steps 3 and 4 until H_t. Then, estimate $e^{0,0} = \left(e_1^{0,0}, \ldots, e_H^{0,0}\right)'$ and $AE^{0,0} = \left(AE_1^{0,0}, \ldots, AE_{H_t}^{0,0}\right)$. Additionally, we define the weighted AE (WAE) as $WAE^{0,0} = AE^{0,0}\ \Upsilon$ where Υ is a weighted $H_t \times 1$ matrix that penalizes the nowcasting errors such that $\Upsilon 1' = 1$.
6. Repeat steps for all combinations of p and q until p_{max} and q_{max}. Generate the following elements:

$$\hat{y}(p, q) = \left(\hat{y}^{0,0}, \hat{y}^{1,0}, \ldots, \hat{y}^{p_{\max}, q_{\max}}\right),$$

$$e(p,q) = \left(e^{0,0}, e^{1,0}, \ldots, e^{p_{\max},q_{\max}}\right),$$

$$WAE(p,q) = \left(WAE^{0,0}, WAE^{1,0}, \ldots, WAE^{p_{\max},q_{\max}}\right)',$$

where $\hat{y}$ is a $2 \times (p_{\max}+1)(q_{\max}+1)$ matrix of nowcasts, e is an $H_t \times (p_{\max}+1) \times (q_{\max}+1)$ matrix that contains the nowcast errors in the data test, and WAE is an $H_t \times 1$ vector of the weighted errors in data test.

7. We select the best nowcast as a function of p and q, denoted by $\hat{y}(p*, q*)$, where $p*$, $q*$ are obtained as follows:

$$p^*, q^* = \underset{0 \le p,q \le p_{\max} q_{\max}}{\operatorname{argmin}} \; WAE(p,q)$$

8. To use models with similar performance, we combine the nowcasts of $\hat{y}(p*, q*)$ with models with equal forecast errors according to the Diebold and Mariano (1995) tests, by using the $e(p,q)$, carrying out pairs of tests between the model with minimum $AE(p,q)$ and the others. Consequently, from the models with statistically equal performance, we select the median of the nowcasts, namely, $\hat{y}$.

This nowcasting approach allows the generation of nowcasts based on a trained process, taking advantage of the information of similar models. In this case, $\hat{b}$ must be significant to exploit the relationship between the IGAE and the information summarized by the DFM. Note that Υ is a weighted matrix that penalizes the nowcasts errors. The most common form is $\Upsilon = (1/H_t, \ldots, 1/H_t)'$, a $H_t \times 1$ matrix where all nowcasts errors have equal weight named in literature as mean absolute error (MAE). Therefore, we are not considering by default the traditional MAE, but rather a weighted (or equal) average of the individual AE. For example, we could have penalized with more weight the last nowcasts errors, that is, in the COVID-19 period. Also, note that we can obtain $AE(p,q)$ and estimate the median or some specific quantile for each vector of this matrix.

Note that even though root mean squared errors (RMSEs) are often used in the forecast literature, we prefer a weighted function of AEs, although in this work we use equal weights that is, the MAE. The main advantages of MAE over RMSE are two: (1) it is easy to interpret since it represents the average deviation without considering its direction, while the RMSE averages the squared errors and then we apply the root, which tends to inflate the larger errors and (2) RMSE does not necessarily increase with the variance of the errors. Anyway, the two criteria are in the interval [0, ∞) and are indistinct to the sign of errors. In practical terms, in the rest of the work and whenever its appropriate, we use both indistinctly. This allows to compare our model's MAE with models whose error measure is the RMSE.

3. Data and Descriptive Analysis

To illustrate the behaviour of the target variable, Figure 1 shows the historical $y*$ from January 2004 to October 2020.

We can see that IGAE's fall in 2020 has been the sharpest in the recent history. The pandemic brought about -19.7%, -21.4% and -14.4% contractions in total economic

Fig. 1. Evolution of the annual growth rate of IGAE

activity for April, May and June of 2020, respectively, but the economy had already begun to show signs of deterioration since May 2019. Prior to the coronavirus outbreak, the only comparable economic turmoil was the economic contraction derived from the Great Recession of 2008–2009.

Between January 2019 and February 2020, before the COVID-19 outbreak started in Mexico, the annual growth of IGAE already showed signs of slowing and fluctuated around -1.75% and 76%, and since May 2019, the economy exhibited nine consecutive months of negative growth.

The first six months of the COVID-19 pandemic (until September 2020) has had severe impacts on the Mexican economy. The first case of coronavirus in Mexico was documented on February 27, 2020.

Broken down by sector, the economy suffered devastating the consequences in the secondary and tertiary sectors. The industrial sector registered the deepest contractions, reducing its activity in April and May by -29.9% and -29.6%, respectively, in annual terms, mainly driven by the closure of manufacturing and construction operations, which were considered nonessential businesses, following a slight recovery in June, -17.1%, when an important number of activities, including automobile manufacturing, resumed but remained at low activity levels. The services sector also suffered from lockdown measures, falling by -15.9%, -18.7% and -13.7% in the three months of the second quarter, respectively, especially on transportation, retail, lodging and food preparation, mainly due to the decrease in tourist activity, although restaurants and airports were not closed. The primary sector showed signs of resilience and even grew in April and May 2020, by 0.7% and 1.5%, and shrank in June by -2.4% on an annual basis.

On the other hand, note that the variables to estimate the DFM are selected by using the criteria of timely and contemporaneous correlation with respect to $y*$. Thus, we consider a select group of macroeconomic and financial time series that represent contemporaneously and timely, the industrial and service sectors. For example, we take into account

information related with the industrial production from Mexico and US, manufactures, exports, imports, sales, hotel occupation, number of formal workers, and so on. Note that usually the nowcasts models focus on predicting the quarterly GDP. In this sense, there are several monthly variables that anticipate the GDP release. Therefore, for quarterly time series nowcasting, the availability of information is larger.

In this sense, the model differs from those in the traditional literature on large DFMs, which use a large amount of economic and financial variables; see, for example, Corona et al. (2017a) who use 211 time series to estimate the DFM for the Mexican case with the goal of generating forecasts for the levels of IGAE. However, Gálvez-Soriano (2020) uses approximately 30 selected time series to generate nowcasts of Mexican quarterly GDP. Note that Boivin and Ng (2006), in the context of DFM, point out that we can reduce the forecast prediction error with selected variables by estimating the common components. Additionally, Poncela and Ruiz (2016) and Corona et al. (2020) show that with a relatively small sample size, for example, $N = 12$, we can accurately estimate a rotation of the common factors. We acknowledge that recent developments on the inclusion of big data using sparse or dense modeling to generate economic predictions, and have concluded that it is not always possible to identify sparse predictive representations that allow the selection of adequate predictors from a larger amount of macroeconomic variables (Giannone et al. 2021). Thus we keep with the economic variables selected a priori based on our literature review.

Alternatively, when the timeliness of the estimation becomes relevant the set of variables is restricted as well, for example, compared to Corona et al. (2017a) where the key variable is also monthly, although they do not require the timely criterion, thus they exploit the lag correlation between a large amount of time series with respect to IGAE's current value.

Consequently, given the timely and possibly contemporaneous correlation with respect to the $y*$, the variables considered in this work are described in Appendix (Section 6). All variables are seasonally adjusted in the following ways: (1) directly downloadable from their source or (2) by applying the X-13ARIMA-SEATS.

Hence, we initialized with 68 time series divided into traditional and nontraditional time series. The traditional time series is timely macroeconomic and financial information such as the Industrial Production Index values for Mexico and the US, business confidence, and exports, imports, Mexican stock market index, nominal exchange rate, interest rate and the Standard & Poor's 500. We considered the possibility of including other timely time series as inflation, oil prices, consumer confidence, among many others, but were discarded given that they inclusion in the DFM increased the error in the CV. Additionally, for the nontraditional variables, we have daily variables such as the media mobility index obtained from Twitter and the topics extracted from Google Trends. These topics are manually selected according to several phenomena that occur in Mexican society, such as frequencies of politicians' names, natural disasters, economic themes and topics related to COVID-19, such as coronavirus, quarantine, or face mask. The Google Trends variable takes a value of 0 when the topic is not searched in the time span and 100 when the topic has the maximum search in the time span. Although these variables are expressed as monthly variables, for the media mobility index, we average the daily values, and for Google Trends we download the variables by month. While searching for significant Google topics in the past, we normalize each Google topic as $\left(\frac{W_i}{max(W_i)}\right) \times 100$ to remove the effect of the maximum search on recent data.

The media mobility index is calculated based on Twitter information following Graff et al. (2022). We selected around 70,000 daily tweets georeferenced to the Mexican territory, each one is associated with a bounding box. Then, movement data analysis is performed by identifying users and their sequence of daily tweets: a trip is considered for each pair of consecutive geo-tagged tweets found in different bounding boxes. The total number of trips per day is obtained and divided by the average number of users in the month. The number obtained can be interpreted as the average number of trips that tweeters make per day. Twitter allows to estimate the index with a larger sample size than with respect other alternatives as Google or Facebook.

To select the relevant topics, we apply the methodology described in Subsection 2.1.1 by using $H_g = H_t + 1$ and $\alpha = 0.10$ where $H_t = 36$; consequently, we select the topics that are relevant in at least 90% of cases in a sample size that covers the number of months of the CV in the nowcasting approach and also, the period of the current nowcasts. Additionally, we select $H_i^* = 8$.

In this way, the current significant topics (as of January 15, 2021) are "health crisis", "quarantine", "face mask N95" and "recession". Other relevant topics in the past were "coronavirus", "dollar", "oil", "pandemic","unemployment", among many others.

Once X^* is defined, we apply the transformations suggested by Equation (2) to define X. Appendix (Subsection 6.2. Table 1) shows each X_i ordered according to its correlation with $y*$.

In order to highlight how the traditional and nontraditional variables are related with $y*$ in the last $H_g = 37$ months, Figure 2 shows the ordered linear correlations for different sample sizes $T^* - H_g + h$, with $h = 1, \ldots, 37$.

We can see that during the last 37 months the time series are strong or moderately correlated with the IGAE; the exceptions being SP 500 and REM, whose correlations have decreased starting the COVID-19 pandemic. Also, M4 is weakly correlated throughout the period. Although Google topics have become more correlated since COVID-19 started, in the past they were medium correlated, which implies its possible inclusion, not only in the current period but also in the past. Figure 2 also shows the selected topics when the selection is carried out with all sample size T^*, that is, when $h = 37$. Other interesting traditional variables are IRGS, HOTEL and GAS, which show a similar behavior as the Google topics. Intuitively, these variables are more relevant to explain the IGAE's movement since the

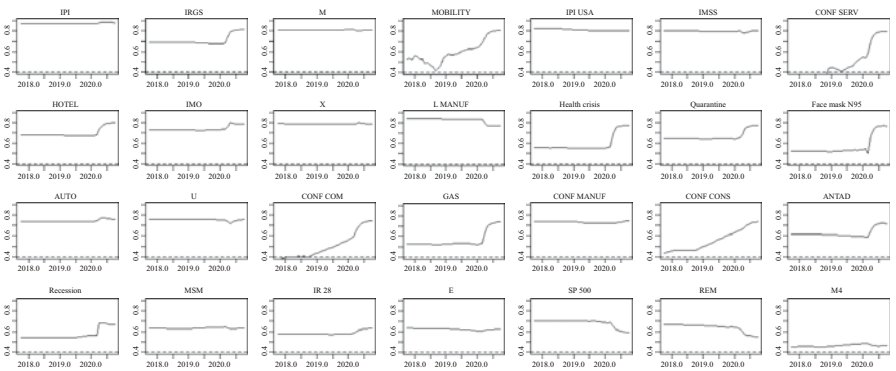


Fig. 2. Linear correlation: between X and y by using* $t = 1, \ldots, T* - H_g + h$ *with* $h = 1, \ldots, 37$.

COVID-19 pandemic started. Additionally, variables that exhibit linear correlations with a positive trend are MOBILITY and the those related to business confidence.

Note that the nowcasts depend on the dates of the information released. Depending on the day of the current month, we can obtain nowcasts with a larger or smaller percentage of updated variables. For example, the financial time series and nontraditional variables are available in real time, but macroeconomic traditional time series, which are timely with respect to the IGAE, are available on different moments according to their official release dates.

To obtain timely estimates, we rely on the fact that nontraditional variables, financial time series, business confidence and fuel demand, can be obtained on the day after the month T^* closes. This indicates that on the first day of month $T^* + 1$, we can generate the nowcasts for $T^* + 2$ with approximately 50% of the updated information and 81% for the current month, $T^* + 1$. Moreover, in the 12th day, the IMSS variable is updated, and on the 16th, the IPI USA is usually updated. Consequently, in official statistics, we recommend conducting the nowcasts on the first day of $T^* + 1$ and 16 days later, updating the nowcasts with the latter two timely traditional and important time series, taking into account the timeliness of the most relevant variables. Note that IPI represents around the 34% of the monthly GDP, and represents more than 97% of the second grand economic activity. Given that the IPI is updated around ten days after the end of the reference month, this information is very valuable to carry out the $T^* + 1$ nowcasts.

Consequently, we carry out the nowcasts around the 15th day of each month. In the results presented herein, the database is current up to January 15, 2021; therefore, we generated the nowcasts ten days before the official result of November 2020 and 41 days before the official value of December 2020 was published, having 90% and 66% of updated variables at $T^* + 1$ and $T^* + 2$, respectively.

Appendix (Subsection 6.2, Table 2) shows the approximate day when the information is released for $T^* + 2$ after the current month T^*. Consequently, the temporal dimension of the panel to carry out the nowcasts for the last two months of 2020 is from January 2004 to December 2020. Whenever missing values exist before October 2020, they are imputed according to the methodology proposed by Giannone et al. (2008). Similarly, missing values for the nowcasts months are operated following the 2SM procedure described in Subsection 2.2.1.

4. Nowcasting Results

4.1. Estimating the Common Factors and the Loading Weights

By applying the Onatski (2010) procedure to X centered, we can conclude that $\hat{r} = 1$ is adequate to define the number of common factors. Alternative criteria as the ratio of eigenvalues of the sample covariance matrix of the observations (Ahn and Horenstein 2013) also conclude $\hat{r} = 1$. Note that the sample effective dependence in the correlation matrix X is 0.79, which can be considered as relatively large. In other words, once the significant Google Trends topics are selected, the 28 time series observed for 204 months that compose the DFM are summarized with one latent variable. The estimated static common factor obtained by PC using the set of variables, X, their asymptotic confidence intervals at 95%, in Figure 3, the asymptotic variance of the estimated common factors is

calculated as $\frac{1}{N}\hat{V}^{-1}\Theta_t\hat{V}^{-1}$ where $\hat{V}$ is the $r \times r$ diagonal matrix of the first r largest eigenvalues of matrix $X'X/(NT)$ and $\Theta_t = \frac{1}{N}\sum_i^N \hat{p}_i'\hat{p}_i'\hat{\varepsilon}_{it}^2$ and the dynamic factor estimates by applying the 2SM procedure with $k = 1$ lags.

We observe the common factors summarizing the previous elements representing the decline in the economy in 2009 and 2020. Note that in the last period, the dynamic common factor shows a slight recovery of the economy because this common factor supplies more timely information than the static common factor. Thus, the static common factor has information up to August 2020, while the dynamic factor has information up to October 2020. Note that the confidence intervals are closed with respect to the static common factor, which implies that the uncertainty attributed to the estimation is well modelled. Note that the dynamic factor can be interpreted as a coincident economic indicator because it measures contemporaneously the global state of the economy.

The correlation between $\tilde{F}_t$ and $y*$ is approximately 0.86 prior to the financial crisis of 2009, increasing from this year to 0.98, showing a slight decrease since 2011, dropping in 2016 to 0.95 and fully reaching levels of 0.96 since 2020. Consequently, we can exploit the contemporaneous relationship between the dynamic factor and the IGAE to generate their nowcasts for the two following months that the common factors have estimated with respect to the IGAE. Appendix (Subsection 6.2, Table 3) shows the correlation coefficient of $\tilde{F}_t$ with $y*$ since 2008.

Having estimated the dynamic factor by the 2SM approach, we show the results of the loading weight estimates that capture the specific contribution of the common factor to each variable, in other words, given the PC restrictions, they can be seen as N times the

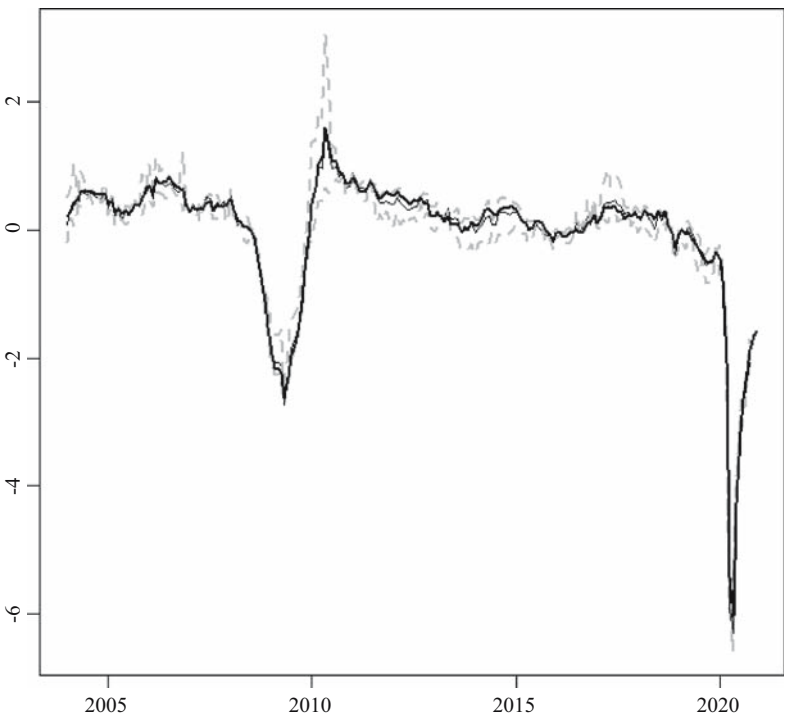


Fig. 3. Factor estimates. The thin line is the static common factor, the gray dotted lines are their asymptotic confidence intervals at 95% and the solid line is the smoothed or dynamic common factor.

contribution of each variable in the common factor. We compute the asymptotic confidence interval at 95% (Bai 2003) denoted by $CI_{\hat{P},0.05}$. Once the dynamic factor is estimated by using the Kalman smoother, it is necessary to reestimate the factor loadings to have $\hat{P} = f(\tilde{F})$, such that $\tilde{F} = g(\tilde{P})$. To do so, we use Monte Carlo estimation iterating 1,000 samples and select the replication that best satisfies the following condition:

$$\tilde{F} \approx X\tilde{P}/N \text{ s.t } \tilde{P} \in CI_{\hat{p},0.05}.$$

The results of the estimated factor loadings are shown in Figure 4. The loadings are ordered from the most positive contribution to the most negative.

We observe several similarities with respect to Figure 2. The more important variables in the factor estimates are (1) the industrial production, both from Mexico and the US, (2) imports, (3) hotel occupancy, (4) insureds by the Mexican Social Security System (IMSS), (5) exports and (6) the Google Trends topics such as "quarantine", "health crisis" and "face mask N95", which makes sense in the COVID-19 period. Obviously, when these variables are updated, it will be more important to update the nowcasts. Other variables, such as IMO, CONF MANUF, GAS, S&P 500, MOBILITY and E, are also relevant. The less important variables are M4 and remittances. Interestingly all variables are significant because their confidence intervals do not include the number zero.

We want to emphasize that the most relevant variables are also the timelier, for example the industrial production index of Mexico and the U.S. are updated around days ten and 16

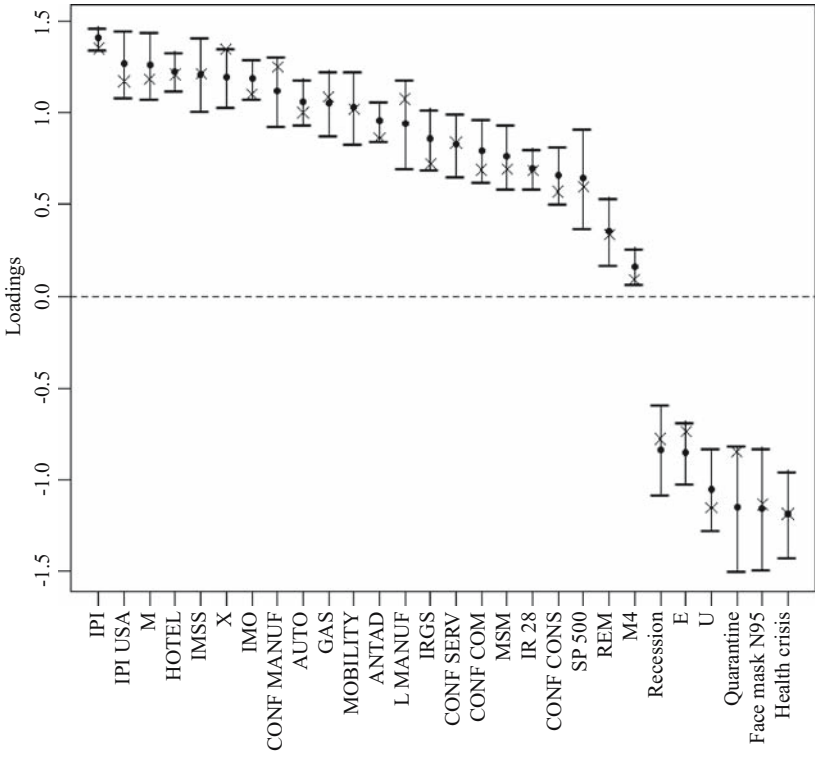


Fig. 4. Factor loadings. The point is each $\hat{p}_i$ with its respective 95% asymptotic confidence interval. Crosses are the $\tilde{p}_i$.

for $T^* + 1$ and $T^* + 2$, respectively, once closed the current month; furthermore, the exports and imports are updated for $T^* + 2$ by the 25th day, while IMO and IMSS are updated the first day and 12th day, respectively for $T^* + 2$. Consequently, this allows to have more accurate and correlated estimates beginning the first day of the current month for both, $T^* + 1$ and $T^* + 2$.

As we have previously noted, to obtain a consistent estimation of $\tilde{F}$ and $\hat{P}$ it is necessary that $\hat{\varepsilon}$ be stationary. We check this point with the PANIC test of Bai and Ng (2004). Table 1 summarizes the results.

We conclude that we have achieved stationarity in the idiosyncratic component, obtaining a statistic of 12.03 that generates a p-value of 0.00; hence, $\hat{\varepsilon}$ does not have a unit root. Additionally, we can verify with the augmented Dickey-Fuller test that $\tilde{F}$ is stationary with a p-value of 0.01; consequently, we have also achieved stationarity in X.

Note that we could have alternatively proposed to estimate the components of the DFMs in blocks, as suggested by Baríbura et al. (2011) by assuming that some common factors can be composed by different groups of variables, for example: macroeconomic, financial and non-traditional time series. However, given that the Onatski (2010) procedure detects one common factor, we considered that the assumption that all factors are global is adequate.

4.2. *Selecting the Nowcasts: Time Series Cross-Validation*

We apply the procedure described in Subsection 2.3 by using a $\Upsilon = (1/H_t, 1/H_t, \ldots, 1/H_t)'$; then, we assume that each AE has equal weight over time. Additionally, we fix $p_{\max} = q_{\max} = 6$. In empirical exercises, when $p_{\max}$ and $q_{\max} > 6$ the results do not change. Consequently, we assume that $p_{\max} = q_{\max} = 6$ is an appropiate window for searching models both from statistical and computational viewpoints. The obtained results indicate that the optimums $p*$ and $q*$ are selected to be equal to 2 with $k = 1$ in the factor equation. Consequently, the best model to carry out the current nowcasts is a regression model with ARMA(2,2) errors.

This specification generates the following nowcasts in CV one-step ahead, during $H_t = 36$ months that are presented in Figure 5.

We can see that the best nowcast model in CV performs well given that in 33 of 36 cases, 91.6%, the observed values are within the confidence interval at 95%. The MAE (equal weights in Υ) is 0.52, and the mean absolute annual growth of IGAE is 3.50%. Regarding the median of the AEs, the estimated value is 0.38. Furthermore, based on Diebold-Mariano tests, 84 of 98 models are statistically equal to the best model using a significance of $\alpha = 0.05$, consequently, our nowcasting models generate consistently accurate results in the CV.

In addition, we compare our results to Corona et al. (2017a), which selects models in a similar way for CV period, we compare the IGAE levels' forecasts one-step ahead. They

Table 1. PANIC test

Test	Value	P-value
ADF (dynamic factor)	-2.70	0.01
Pooled test (idiosyncratic errors)	12.03	0.00

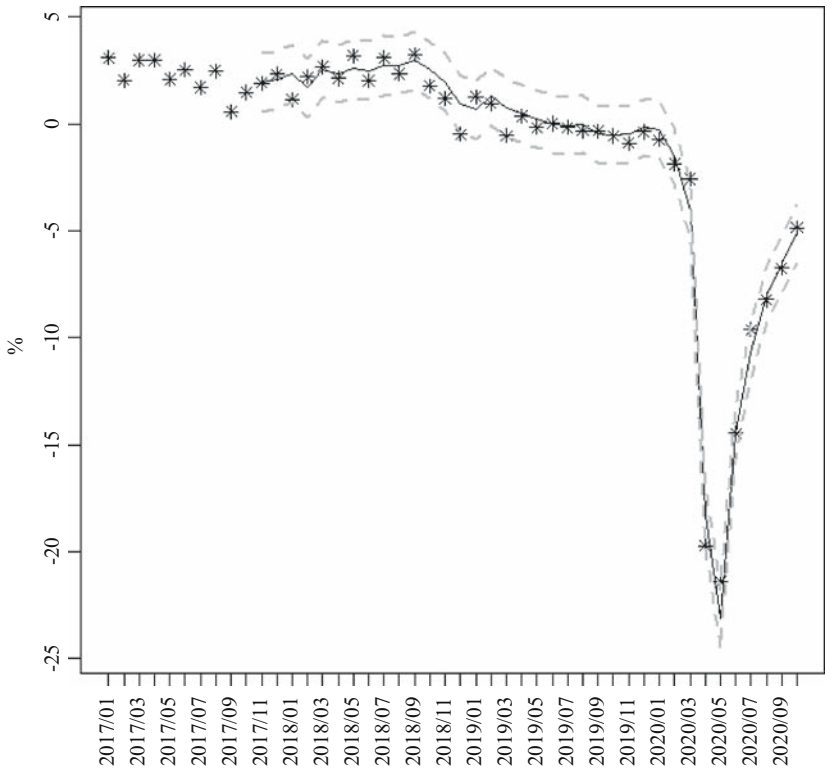


Fig. 5. Nowcasts in CV. Asterisks are the observed values, the line depicts the nowcasts and the gray dotted lines are the confidence intervals at 95%.

obtain a median of the root squared errors of around 0.4 and 0.5 in a period of 13 months (March 2015 to April 2016) while our model obtains a statistic of 0.43 for 36 months. Consequently, our approach tends to have better results.

To contrast the results of our approach with those obtained by other procedures in a certain period of CV, we consider the following two alternative models:

1. Static PC: We apply static PC to X_t^* and by using the first estimated PC in a linear regression, we obtain the nowcasts for the last $H_t = 36$ months. In each nowcasting step we consider their specific relevant topics of Google Trends obtained by applying the procedure described in Subsection 2.1.1.
2. DFM without Google Trends: We estimate a traditional DFM similar to Corona et al. (2017a) or Gálvez-Soriano (2020), but using only economic and financial time series, that is, without considering the relevant topics extracted from Google Trends. Hence, we carry out the last $H_t = 36$ nowcasts.

Figure 6 shows the accumulated MAEs for the out-of-sample period by the previous two models and the obtained by the regression model with ARMA(2,2) errors.

We can see that, in cross-validation period the static PC is the one with the weakest performance, followed by the traditional DFM. Specifically, the MAE is 1.24 for the static PC, 0.74 when using DFM without Google Trends and, as we have commented, 0.52 for

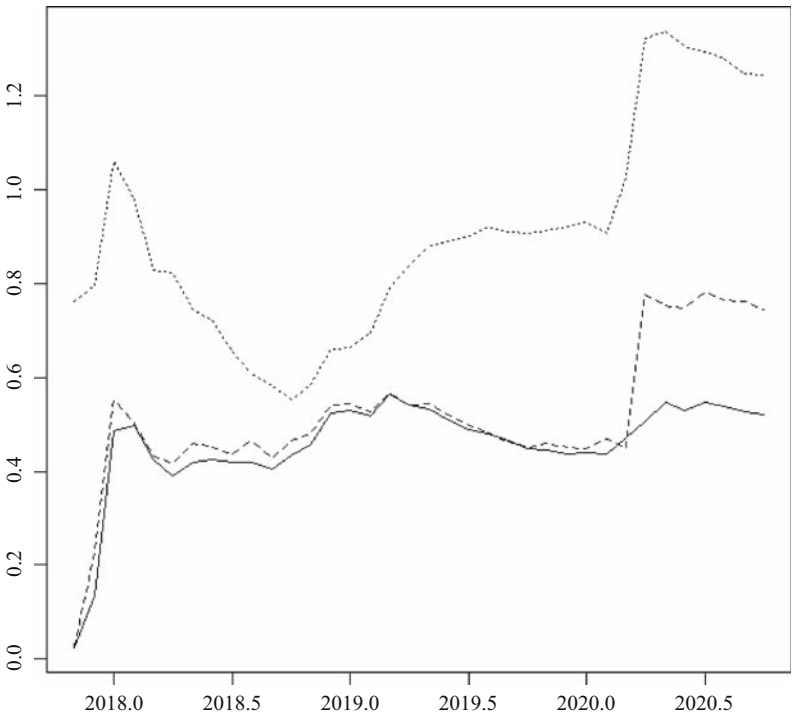


Fig. 6. Cumulative MAEs for models for the period of CV. Solid line is the nowcasting approach suggested in this work, pointed line is the naive model and the dotted line is the traditional DFM (without Google trends topics).

the incumbent model, which indeed includes this type of information. Note that the use of nontraditional information does not affect the behaviour of the MAEs previous to COVID-19 pandemic and reduces the error during this period. The Diebold-Mariano tests indicate that for all cross-validation period, we obtain p values of 0.08 and 0.16 when comparing our approach with static PC and traditional DFM respectively, but since the pandemic period, reflected from April 2020, the MAEs of these models have increased from 1.02 to 1.24 for the static PC, and from 0.45 to 0.74 for the traditional DFM. For this period, the proposed DFM error is slightly increased from 0.47 to 0.52.

Note that the main reason why we cannot reject the null hypothesis can be attributed to the importance of the traditional time series, which are very relevant to timely estimate IGAE movements. Nonetheless, we can further reduce the error estimates during the shock attributed to the COVID-19 pandemic by using nontraditional information and, in the past, the error does not increase.

The conclusion is that the performance of the suggested approach is highly competitive when compared with (1) similar models for nowcasting of GDP, (2) models that estimate the levels of the objective variable and (3) alternative models that can be used in practice.

Although the residuals are not normally distributed, mainly due to the shocks introduced by the pandemic, note that residual normality is not required given that the likelihood estimation of nowcasting equations assumes asymptotic normality, consequently, the t statistics are correctly estimated when no autocorrelation and no ARCH effect are present.

Instead, we do verify and validate the absence of autocorrelation and ARCH effect. Table 2 summarizes the results.

We can see that the residuals of the model with ARMA(2,2) errors does not have autocorrelation nor ARCH effect. Furthermore, the rest of the nowcasting models generate, on average, both non-autocorrelated and non-conditional-heteroskedastic residuals.

4.3. Current Nowcasts and Historical Analysis in Pseudo Real Time

Having selected the highly competitive models in accordance with Diebold and Mariano (1995) tests, the final nowcasts for IGAE's annual percentage variation for November and December 2020 indicate a slight recovery of the economy in November and December 2020, obtaining nowcasts of -4.6% and -4.1%, respectively, with confidence intervals of (-6.3, -3.2) and (-5.6, -2.6) for both months. Considering the observed values for November and December, released on February 25 2021 by INEGI, the annual percentage change for the IGAE was -3.7% for both months; consequently, in this case the model generates nowcasts that falls within the confidence interval with deviations of -0.9% and -0.4% respectively.

The procedure described in the previous subsection allows the generation of nowcasts using databases with different cut off dates. In this way, we carry out the procedure updating the databases once a month during the last 35 periods. We assume similar conditions as in the current nowcasts, that is, we consider the same percentage of time series updated, 90% and 66% for $T^* + 1$ and $T^* + 2$ respectively. Consequently, we assume estimations in "pseudo real time" because they do not consider previous revisions in the databases. Additionally, the seasonal adjustment procedure tends to generate slight changes in the values of the time series in every update. However, note that the methods based on DFMs are sensible to distortions in the variance of the time series, consequently, it is preferable to estimate the components of the DFM with seasonal adjusted time series. Thus, we replicate the exercise with the current database assuming the last 35 months of its observed values. Note that this period encompasses several stages of the Mexican economy cycle: stability, slight drops, drastic drops and recovery. Note that transformations described in Subsection 2.1.2 are assumed fixed during all pseudo real time analysis. However, when generating an historical analysis about the optimal transformations along the last 37 months, only few variables present differences: business confidence (CONF CONS, CONF SERV and CONF COM), M4 and MOBILITY; which, reasonably, validates our assumption.

Figure 7 summarize the nowcasts results for $T^* + 1$ and $T^* + 2$ respectively, comparing them with the observed values.

Table 2. Residual tests about the residuals of nowcasting models. We report the p-values.

Test	Model with ARMA(2,2) errors	Mean: All nowcasting models
Normality (Anderson-Darling test)	0.02	0.05
Autocorrelation (Ljung-Box test)	0.80	0.86
ARCH test (F statistic)	0.29	0.30

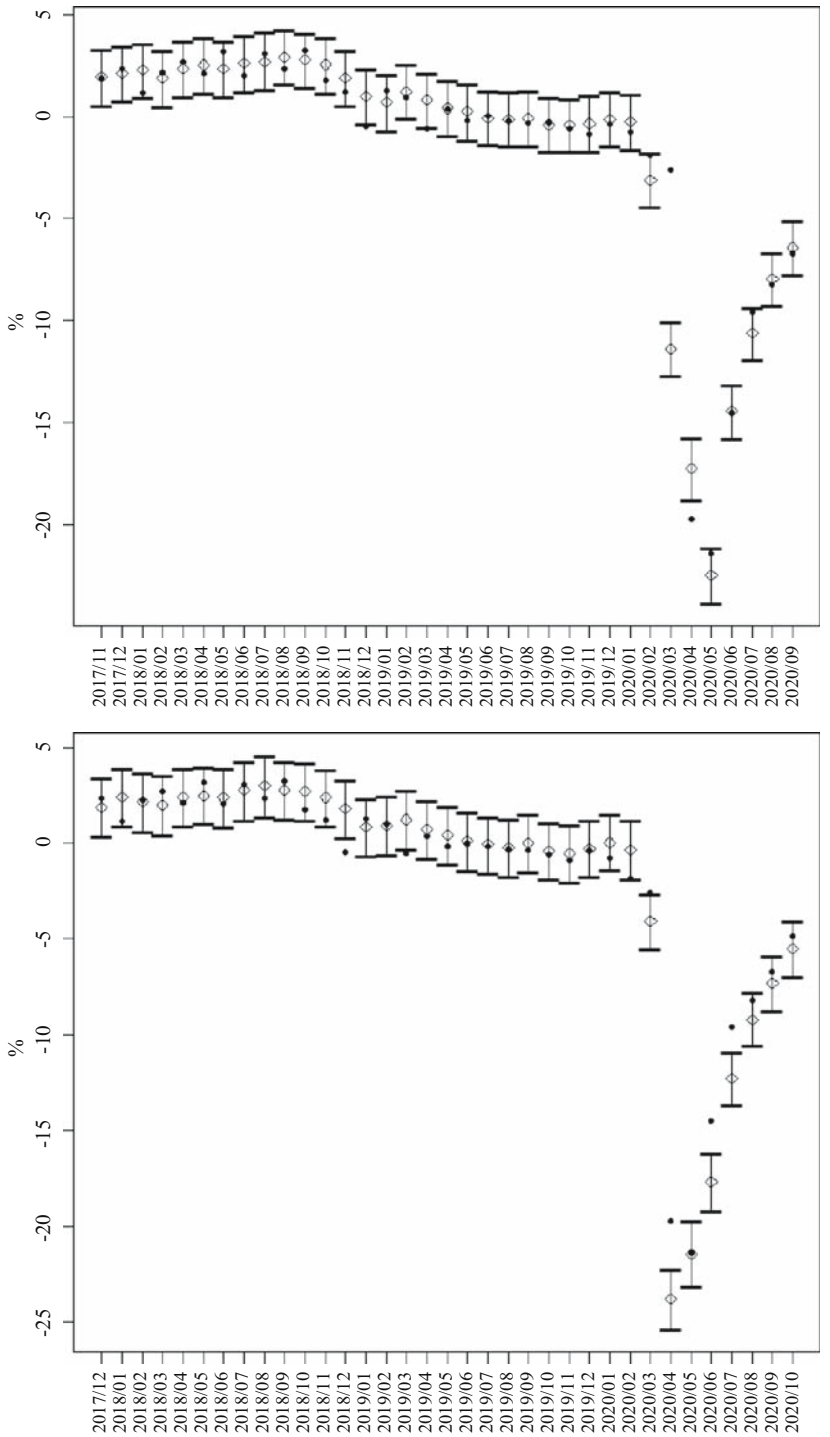


Fig. 7. Nowcasts in pseudo real time for T^*+1 (top panel) and T^*+2 (bottom panel) during November 2017 to September 2020. The rhombus represent the nowcasts with their respective 95% confidence intervals. Points are the observed values.

Focusing on $T^* + 1$, we can see the approach's functioning in the pre COVID-19 crisis period, and how on March 2020 it anticipated a larger fall than the observed value. The reason is that the selected Google Topics during these months overestimate the plunge, which was around -2.6% while the nowcasts for $T^* + 2$ was -4.1% and $T^* + 1$ of -11.4%. Note that in April of the same year, there was a historic fall, larger than 20%, which was outside our confidence interval. Nonetheless in all but these two months, the model has been precise in anticipating the falls and recoveries of the Mexican economy.

Specifically, for the March and April 2020 estimations, the dimension of the panels are $T = 182, N = 31$, and $T = 183, N = 30$, respectively. In turn, for March 2020 the relevant topics are: "coronavirus", "health crisis", "quarantine", "face mask", "face Mask N95", "dollar" and "recession". On the other hand, for April 2020 the selected Google Topics are: "coronavirus", "quarantine", "dollar", "face Mask N95", "pandemic" and "recession. Basic regression models, without assuming ARMA errors, generates in the $H_g^* = 8$ periods, MAEs of 0.96 and 1.05 for March and April 2020, respectively. Alternative procedures to LASSO regression as partial least squares select similar Google topics along the time.

Although Google topics are selected using the procedure described in subsection 2.1.1, they are drawn from a user-predefined list (see Appendix, Subsection 6.1). In this way, the accuracy of our procedure also depends on which topics are decided to start with. This list is used to carry out both, the current nowcasts (November and December 2020) and the previous 35 months. Thus, we acknowledge that this fact represents a limitation of our pseudo real time analysis given that the accuracy depends on which are the predefined topics along the time, above all, when unexpected events can occur.

Considering this fact as given, note that $T^* + 2$ shows consistent results with $T^* + 1$, and help us to conclude that the approach behaved well previous to the pandemic, but also during these phenomena. Note that, the model anticipated the drastic falls of April and May 2020, and subsequently it has showed precise performance, considering that these nowcasts are estimated with a smaller amount of information. In absolute terms, the MAE for $T^* + 1$ *y* $T^* + 2$ are 0.80 and 0.86 respectively, with AE's median of 0.42 and 0.58. Consequently, during several months the performance of the approach tends to generate accurate nowcasts, whether, the Mexican economy is stable or volatile.

A nowcasting model to compare with our approach is INEGI's "Early Monthly Estimation of Mexico's Manufacturing Production Level (INEGI. d.) whose target variable is manufacturing activity, generating the one step ahead nowcasts by using a timely electricity indicator. The average MAE for the annual percentage variation of manufacturing activity in the last 36 months, from November 2017 to October 2020, is 1.43. Consequently, in a similar sample period, we have a smaller average MAE than another nowcasting model where its monthly target variable is specified as annual percentage variation.

To compare the nowcasts/prediction errors with other procedures in the same sample, we estimate $T^* + 1$ and $T^* + 2$ using some benchmark univariate models, based on statistical methods such as AR. The AR models were obtained using the automatic models generated by the `auto.arima` function from `forecast R` package. and machine learning approaches based on neural networks (NN). Additionally, we consider DFMs without Google Trends (DFM-2). With respect to NN, we consider a multilayer perceptron (MLP) model and extreme learning machines (ELM), both adapted for time series

forecasting, see, for example, Huang et al. (2006) and Ord et al. (2017). NN are efficient non-parametric models which have been used in a wide variety of tasks, due to their ability to obtain useful data representations by means of non-linear activation functions. Although there are some NN architectures specialized in time series forecasting with remarkable results, see Gamboa (2017), Sezer et al. (2020), Hewamalage et al. (2021), among many others, they are very expensive in time and computational resources, mainly due to the high number of parameters involved. For this reason, we decided to use an ensemble of 20 single layer MLP and ELM, where the number of hidden nodes and other hyperparameters were selected automatically with a CV criterion, and the output layer weights for ELM are estimated with CV-LASSO. In both cases, we used the median as the combination operator for forecast. See Kourentzes et al. (2014) for details. We use the implementation of the `nnfor R` package. For all benchmark methods, we do not use exogenous regressors.

Table 3 shows the relative MAEs and the Diebold-Mariano p-values of the tests, both with respect to the DFM presented in this work.

We can see that, although the Diebold-Mariano test only rejects the null hypothesis of equal performance for the AR model, the other approaches generate larger MAEs than the DFM, above all, for $T^* + 2$. Consequently, our approach performs by generating lower estimation errors than some alternative procedures.

The inclusion of exogenous variables in NN methods generates more exact comparisons with respect to DFM and DFM-2. Thus, it is expectable that the errors may decrease by including these regressors.

Our approach generates monthly nowcasts of the Mexican economic activity, with results that are very close to the monthly GDP. Consequently, we carry out the quarterly version of GDP nowcasts by assuming that we know the IGAE for the first month of the quarter and then estimating the next two months ($T^* + 1$ and $T^* + 2$) with our pseudo real time analysis. We compare these quarterly nowcasts with the median of the Banco de Mexico Survey of Professional Forecasters (SPF) generated for the expected GDP annual variation of the corresponding quarter made after the first month ends. Table 4 shows the results for 2018:Q1 to 2020:Q3.

We can see that, our approach generated better estimates than the median of the expectations of SPF, which corroborates that our nowcasting approach is frequently much better with respect to the opinion of policy makers and specialized analysts. Specifically, the Diebold-Mariano test allows to conclude that our approach generates better nowcasts by obtaining a p-value of 0.04, that is, our method is more accurate than SPF estimates.

Table 3. Relative MAEs and Diebold-Mariano p-values for $T^* + 1$ and $T^* + 2$ for each alternative method with respect to the DFM.

	MAE		Diebold-Mariano test	
Model	$T^* + 1$	$T^* + 2$	$T^* + 1$	$T^* + 2$
AR	3.127	3.316	0.06	0.02
ELM	1.926	2.842	0.17	0.13
MLP	1.864	2.753	0.17	0.13
DFM-2	1.163	1.768	0.36	0.24

Table 4. Comparison with Banxico's Survey of Proessional Forecasters:

	Percentage annual variation			Absolute errors	
Date	Observed GDP	SPF (median)	Nowcasting (quarterly version)	SPF (median)	Nowcasting (quarterly version)
2018:Q1	2.50	1.80	1.82	0.70	0.68
2018:Q2	2.18	2.50	2.27	0.32	0.09
2018:Q3	2.81	2.45	2.94	0.36	0.13
2018:Q4	1.23	2.10	1.79	0.87	0.56
2019:Q1	0.26	1.50	1.25	1.24	0.99
2019:Q2	0.04	1.10	0.18	1.06	0.14
2019:Q3	-0.12	0.60	-0.13	0.72	0.01
2019:Q4	-0.86	0.18	-0.39	1.04	0.47
2020:Q1	-1.98	0.54	-2.56	2.52	0.58
2020:Q2	-18.80	-14.00	-20.00	4.80	1.20
2020:Q3	-8.65	-12.00	-8.40	3.35	0.25
			Mean:	1.54	0.46

5. Conclusions and Further Research

In this article, we contribute to the nowcasting literature by focusing on the one and two step-ahead of the annual percentage variation of IGAE, the equivalently of the Mexican monthly GDP, during the COVID-19 times. For this purpose, we use statistical and econometric tools to obtain accurate and timely estimates, some around 40 days before the publication of the official data. In summary, the suggested approach consists in (1) using the LASSO regression, along with empirical computational methods, to select the relevant topics that affect the IGAE in the short term that with traditional information allows for building a correlated and timely database to exploit the correlation among the variables and the IGAE, (2) estimate a dynamic factor by using the 2SM approach and (3) selecting linear regressions with ARMA errors in a cross-validation period to select the better models and generate the current nowcasts.

We highlight the following key results. We can see that our approach is highly competitive when compared with models such as static PC, AR, benchmark ELM and MLP; traditional DFM and other models published in the literature. Our procedure frequently captures the observed value, both, in cross-validation period and in real time for the time previous and during the period of the COVID-19 where the usage of Google Trends benefits in the reduction of the error estimation. Another contribution of this article lies in its statistical point of view, given that we compute the confidence interval of the factor loadings and the factor estimates, verifying the significance of the factor on each variable and the uncertainty attributed to the factor estimates. Additionally, we consider some econometric issues to guarantee the consistency of estimates like stationarity in idiosyncratic noises and uncorrelated errors in nowcasting models. Finally, it is of interest to note out-of-sample performance, whether the nowcast error increases when using monthly versus quarterly data.

We highlight the relevance of this method to produce official statistics. Our approach generates preliminary and accurate estimates of the IGAE, a variable that is closely pegged

to GDP. This allows the policy makers and market participants to make informed decisions about the near future of the economy.

Future research topics emerged when doing this research. The most important is to incorporate machine learning techniques to design an algorithm that automatically select the set of possible relevant topics from Google Trends, also explore the inclusion of news and other nontraditional and unstructured information different to predefined lists. Other, is the implementation of an algorithm to estimate non-stationary common factors and make the selection of the number of factors flexible, such as the one developed in Corona et al. (2020), to minimize a measure of nowcasting errors. Also, it would be interesting to incorporate IPI information as restrictions to the nowcasts, by exploring some techniques that incorporate nowcasts restrictions when official countable information is available. We must note that, we can introduce electronic payment information provided by Bank of Mexico in a similar way as Galbraith and Tkacz (2018) and León and Ortega (2018) and compare the results with a large scale DFM. Also, it is interesting the possibility to include time varying intercept as Antolin-Diaz et al. (2017) and González-Astudillo and Baquero (2019) and incorporate dynamic factors as exogenous regressors in machine learning methods is interesting. Furthermore, for future research in this area, it is worth to study further the effects of monthly timely estimate variables versus quarterly time series in nowcasting models; this can be achieved by Monte Carlo analysis with different data generating processes which can occur in practice to compare the increase in the error estimation when distinct frequencies and transformations of time series are used. Finally, it is of interest to generate different transformations for each h during the pseudo real time analysis and to test LASSO or elastic net regression as variable selection method for both traditional and nontraditional blocks of time series.

6. Appendix

6.1. Database

Traditional macroeconomic time series

Short	Variable	Source	Time Span
ANTAD	Total sales of departmental stores	ANTAD	2004/01-2020/11
AUTO	Automobiles production	INEGI	2004/01-2020/12
CONF COM	Right time to invest (Commerce)	INEGI	2011/06-2020/12
CONF CONS	Right time to invest (Construction)	INEGI	2011/06-2020/12
CONF MANU	Right time to invest (Manufacturing)	INEGI	2004/01-2020/12
CONF SERV	Right time to invest (Services)	INEGI	2017/01-2020/12
GAS	Fuel demand	SENER	2004/01-2020/12
HOTEL	Hotel occupancy	Tourism secretariat	2004/01-2020/11
IMO	Index of manufacturing orders	INEGI	2004/01-2020/12
IMSS	Permanent and eventual insureds to the Social Security	IMSS	2004/01-2020/12
IPI	Industrial Production Index	INEGI	2004/01-2020/11
IPI USA	Industrial Production Index (USA)	BEA	2004/01-2020/12
IRGS	Income of retail goods and services	INEGI	2008/01-2020/10
L MANUF	Trend of labor in manufacturing	INEGI	2007/01-2020/10
M	Total imports	INEGI	2004/01-2020/11
M4	Monetary aggregate M4	Banxico	2004/01-2020/11
REM	Total remittances	Banxico	2004/01-2020/11

Traditional macroeconomic time series			
Short	**Variable**	**Source**	**Time Span**
U	Unemployment rate	INEGI	2005/01-2020/11
X	Total exports	INEGI	2004/01-2020/11
Traditional financial time series			
Short	**Variable**	**Source**	**Time Span**
E	Nominal exchange rate	Banxico	2004/01-2020/12
IR 28	Interest rate (28 days)	Banxico	2004/01-2020/12
MSM	Mexican stock market index	Banxico	2004/01-2020/12
SP 500	Standard & Poor's 500	Yahoo! finance	2004/01-2020/12
Nontraditional time series			
Short	**Variable**	**Source**	**Time Span**
AH1N1	AH1N1 online search index	Google	2004/01-2020/12
AMLO	AMLO online search index	Google	2004/01-2020/12
Ayotzinapa	Ayotzinapa online search index	Google	2004/01-2020/12
Calderon	Calderon online search index	Google	2004/01-2020/12
Cartel	Cartel online search index	Google	2004/01-2020/12
Casa Blanca	Casa Blanca online search index	Google	2004/01-2020/12
Chapo	Chapo online search index	Google	2004/01-2020/12
China	China online search index	Google	2004/01-2020/12
Coronavirus	Coronavirus online search index	Google	2004/01-2020/12
Corrupcion	Corrupcion online search index	Google	2004/01-2020/12
Crisis economica	Crisis economica online search index	Google	2004/01-2020/12
Crisis sanitaria	Crisis sanitaria online search index	Google	2004/01-2020/12
Cuarentena	Cuarentena online search index	Google	2004/01-2020/12
Cubrebocas	Cubrebocas online search index	Google	2004/01-2020/12
Desempleo	Desempleo online search index	Google	2004/01-2020/12
Dólar	Dolar online search index	Google	2004/01-2020/12
Elecciones	Elecciones online search index	Google	2004/01-2020/12
EPN	EPN online search index	Google	2004/01-2020/12
Gasolina	Gasolina online search index	Google	2004/01-2020/12
Homicidios	Homicidios online search index	Google	2004/01-2020/12
Huachicol	Huachicol online search index	Google	2004/01-2020/12
Inflacion	Inflacion online search index	Google	2004/01-2020/12
Inseguridad	Inseguridad online search index	Google	2004/01-2020/12
Mascarilla N95	Mascarilla N95 online search index	Google	2004/01-2020/12
Medidas economicas	Medidas econoomicas online search index	Google	2004/01-2020/12
Migracion	Migracioon online search index	Google	2004/01-2020/12
Migrantes	Migrantes online search index	Google	2004/01-2020/12
MOBILITY	Media mobility index	Twitter	2004/01-2020/12
Morena	Morena online search index	Google	2004/01-2020/12
Muertos	Muertos online search index	Google	2004/01-2020/12
Muro	Muro online search index	Google	2004/01-2020/12
Pacto	Pacto online search index	Google	2004/01-2020/12
PAN	PAN online search index	Google	2004/01-2020/12
Pandemia	Pandemia online search index	Google	2004/01-2020/12
PEMEX	PEMEX online search index	Google	2004/01-2020/12
Peso	Peso online search index	Google	2004/01-2020/12
Petroleo	Petrooleo online search index	Google	2004/01-2020/12
PRI	PRI online search index	Google	2004/01-2020/12
Recesion	Recesioon online search index	Google	2004/01-2020/12
Reformas	Reformas online search index	Google	2004/01-2020/12
Salario	Salario online search index	Google	2004/01-2020/12
Sismo	Sismo online search index	Google	2004/01-2020/12
Tipo de cambio	Tipo de cambio online search index	Google	2004/01-2020/12
Trump	Trump online search index	Google	2004/01-2020/12
Violencia	Violencia online search index	Google	2004/01-2020/12

6.2. Tables 1–3

Table 1. Dotted lines represent the specifc Xi and solid lines indicate the specified y. Numbers in parentheses indicate the linear correlation and those between brackets the transformation.*

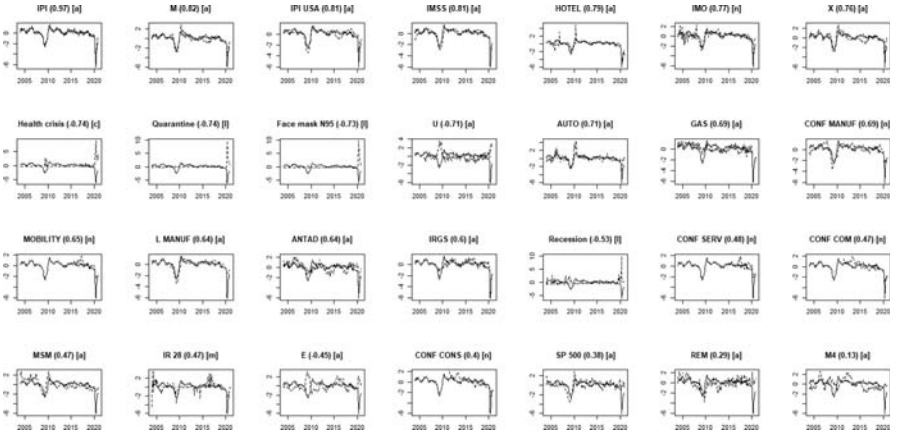


Table 2. Percentage of updated information to carry out the nowcasts T + 2 once the current month T* is closed.*

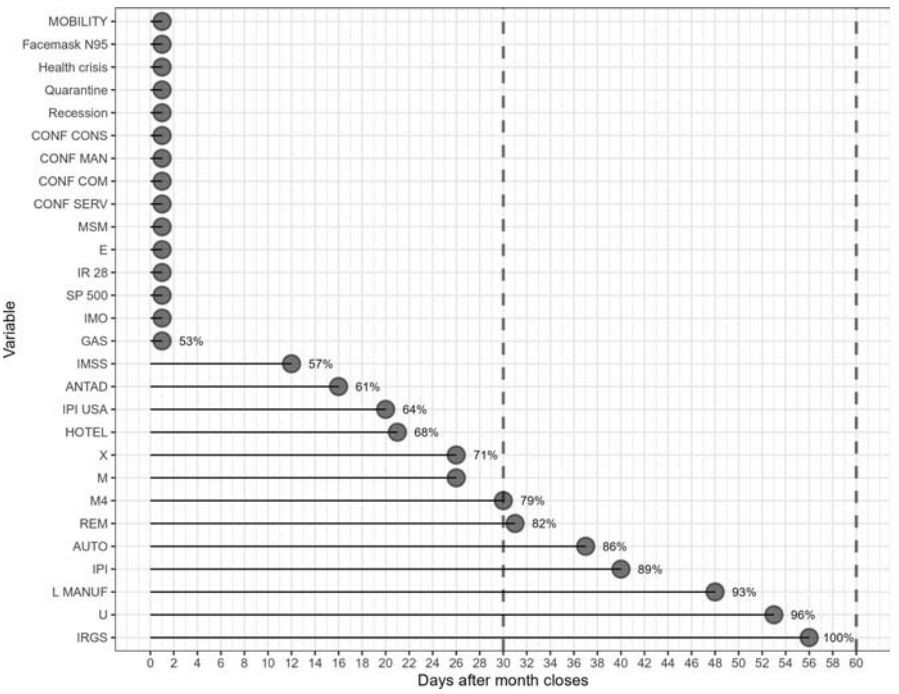
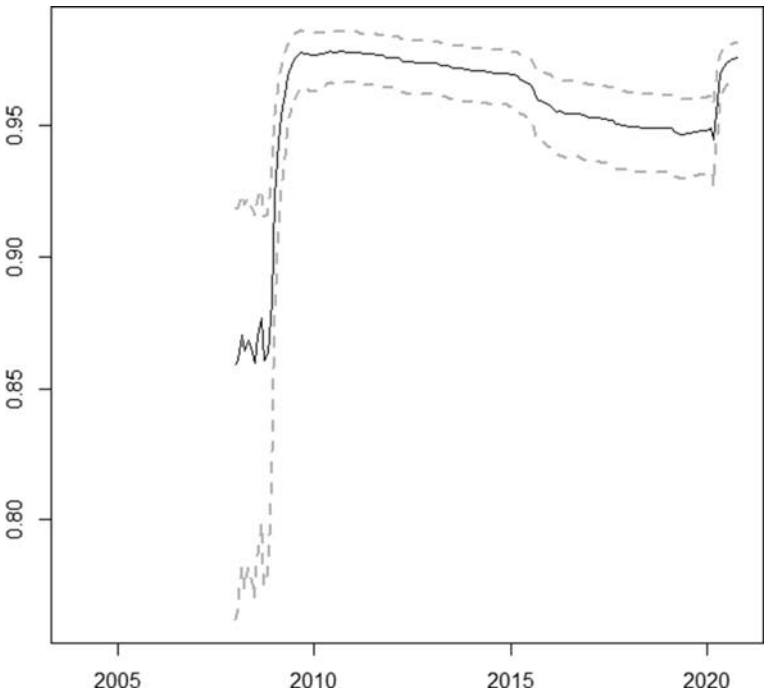


Table 3. Blue line is Corr ($\tilde{F}_t$; y) from January 2008 to October 2020. Lines represent the confidence interval at 95%.*



7. References

Ahn, S. and A. Horenstein, 2013. "Eigenvalue ratio test for the number of factors." *Econometrica* 81(3): 1203–1227. DOI: https://doi.org/10.3982/ECTA8968.

Ali, U., Herbst, C.M., and C.A. Makridis. 2021. "The impact of COVID-19 on the US child care market: Evidence from stay-at-home orders." *Economics of Education Review* 82: 102094. DOI: https://doi.org/10.1016/j.econedurev.2021.102094.

Antolin-Diaz, J., Drechsel, T., and I. Petrella. 2017. "Tracking the slowdown in long-run GDP growth." *Review of Economics and Statistics* 99(2): 343–356. DOI: https://doi.org/10.1162/REST_a_00646.

Aprigliano, V., and L. Bencivelli. 2013. *Ita-coin: a new coincident indicator for the Italian economy.* Banca D'Italia. Working papers: 935: DOI: https://dx.doi.org/10.2139/ssrn.2405416v.

Bai, J. 2003. "Inferential theory for factor models of large dimensions." *Econometrica* 71(1): 135–171. DOI: https://doi.org/10.1111/1468-0262.00392.

Bai, J. 2004. "Estimating cross-section common stochastic trends in nonstationary panel data." *Journal of Econometrics* 122(1): 137–183. DOI: https://doi.org/10.1016/j.jeconom.2003.10.022.

Bai, J., and S. Ng. 2002. "Determining the number of factors in approximate factor models." *Econometrica* 70(1): 191–221. DOI: https://doi.org/10.1111/1468-0262.00273.

Bai, J., and S. Ng. 2004. "A PANIC attack on unit roots and cointegration." *Econometrica* 72(4): 1127–1177. DOI: https://doi.org/10.1111/j.1468-0262.2004.00528.x.

Bai, J., and S. Ng. 2007. "Determining the number of primitive shocks in factor models." *Journal of Business & Economic Statistics* 25(1): 52–60. DOI: https://doi.org/10.1198/073500106000000413.

Bai, J., and S. Ng. 2013. "Principal components estimation and identification of static factors." *Journal of Econometrics* 176(1): 18–29. DOI: https://doi.org/10.1016Zj.jeconom.2013.03.007.

Baríbura, M., Giannone, D., and L. Reichlin. 2011. "Nowcasting", In *Oxford Handbook of Economic Forecasting* edited by Michael P. Clements and David F. Hendry: 193–224. DOI: https://doi.org/10.1093/oxfordhb/9780195398649.013.0008.

Barigozzi, M., Lippi, M., and M. Luciani. 2015. *Dynamic factor models, cointegration, and error correction mechanisms*. arXiv preprint arXiv:1510.02399. DOI: https://doi.org/10.48550/arXiv.1510.02399.

Barigozzi, M. Lippi, and M. Luciani. 2016. *Non-Stationary Dynamic Factor Models for Large Datasets.* SSRN 2741739. DOI: http://dx.doi.org/10.2139/ssrn.2741739.

Benchimol, J., S. Kazinnik, and Y. Saadon. 2021. "Federal Reserve communication and the COVID-19 pandemic." *Covid Economics* 71: 218. DOI: https://cepr.org/content/-covid-economics-vetted-and-real-time-papers-0 (accessed January 2022).

Boivin, J., and S. Ng. 2006. "Are more data always better for factor analysis?" *Journal of Econometrics* 132(1): 169–194. DOI: https://doi.org/10.1016/j.jeconom.2005.01.027.

Buono, D., Mazzi, G., M. Marcellino, and G. Kapetanios. 2017. "Big data types for macroeconomic nowcasting." *Eurostat Review on National Accounts and Macroeconomic Indicators* 1(2017): 93–145. Available at: https://ec.europa.eu/eurostat/-cros/content/big-data-types-macroeconomic-nowcasting-dario-buono-gian-luigi-mazzi-george-kapetanios_en (accessed October 2021).

Caperna, G., Colagrossi, M., A. Geraci, and G. Mazzarella. 2022. "A Babel of web-searches: Googling unemployment during the pandemic." *Labour Economics* (74). DOI: https://doi.org/10.1016/j.labeco.2021.102097.

Caruso, A. 2018. "Nowcasting with the help of foreign indicators: The case of Mexico." *Economic Modelling* 69: 160–168. DOI: https://doi.org/10.1016Zj.econmod.2017. 09.017.

Choi, H. and Varian. 2012. "Predicting the present with Google Trends." *Economic record* 88: 2–9. DOI: https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Choi, I. 2017. "Efficient estimation of nonstationary factor models." *Journal of Statistical Planning and Inference* 183: 18–43. DOI: https://doi.org/10.1016/j.jspi.2016.10.003.

Corona, F., G. González-Farías, and P. Orraca. 2017a. "A dynamic factor model for the Mexican economy: Are common trends useful when predicting economic activity?" *Latin American Economic Review* 27(1). DOI: https://doi.org/10.1007/s40503-017-0044-7.

Corona, F., P. Poncela, and E. Ruiz. 2017b. "Determining the number of factors after stationary univariate transformations." *Empirical Economics* 53(1): 351–372. DOI: https:// doi.org/10.1007/s00181-016-1158-5.

Corona, F., P. Poncela, and E. Ruiz. 2020. "Estimating Non-stationary Common Factors: Implications for Risk Sharing." *Computational Economics* 55(1): 37–60. DOI: https://doi.org/10.1007/s10614-018-9875-9.

De Valk, S., D. de Mattos, and P. Ferreira. 2019. "Nowcasting: An R Package for Predicting Economic Variables Using Dynamic Factor Models." *The R Journal* 11(1). DOI: https://doi.org/10.32614/RJ-2019-020.

Delajara, M., F.H. Alvarez, and A.R. Tirado. 2016. "Nowcasting Mexico's short-term GDP growth in real-time: A factor model versus professional forecasters." *Economia* 17(1): 167–182. DOI: https://muse.jhu.edu/article/634035.

Diebold, F. and R. Mariano, 1995. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics* 13(3): 253–263. DOI: https://doi.org/10.1080/07350015.1995.10524599.

Doz, C., D. Giannone, and L. Reichlin. 2011. "A two-step estimator for large approximate dynamic factor models based on Kalman filtering." *Journal of Econometrics* 164(1): 188–205. DOI: https://doi.org/10.1016/j.jeconom.2011.02.012.

Doz, C., D. Giannone, and L. Reichlin. 2012. "A quasi maximum likelihood approach for large, approximate dynamic factor models" *The Review of Economics and Statistics* 94(4): 1014–1024. DOI: https://doi.org/10.1162/REST_a_00225.

Forni, M., M. Hallin, M., Lippi, and L. Reichlin. 2000. "The generalized dynamic-factor model: identification and estimation." *Review of Economics and Statistics* 82(4): 540–554. DOI: https://doi.org/10.1162/003465300559037.

Galbraith, J. and G. Tkacz. 2018. "Nowcasting with payments system data." *International Journal of Forecasting* 34(2): 366–376. DOI: https://doi.org/10.1016/j.ijforecast.2016.10.002.

Gálvez-Soriano, O. 2020. "Nowcasting Mexico's quarterly GDP using factor models and bridge equations." *Estudios Economicos* 35(2): 213–265. DOI: https://doi.org/10.24201/ee.v35i2.402.

Gamboa, J.C.B. 2017. *Deep learning for time-series analysis*. arXiv:1701.01887. DOI: https://doi.org/10. 48550/arXiv.1701.01887.

Giannone, D., M. Lenza, and G.E. Primiceri. 2021. *Economic predictions with big data: The illusion of sparsity*. ECB Working Paper: 2021/2542, SSRN. DOI: http://dx.doi.org/10.2139/ssrn.3835164.

Giannone, D., L. Reichlin., and D. Small. 2008. "Nowcasting: The real-time informational content of macroeconomic data". *Journal of Monetary Economics* 55(4): 665–676. DOI: https://doi.org/10.1016/j.jmoneco.2008.05.010.

Goldsmith-Pinkham, P. and A. Sojourner. 2020. "Predicting Initial Unemployment Insurance Claims Using Google Trends." *Technical report, Working Paper*. DOI: https://paulgp.github.io/GoogleTrendsUINowcast/google_trends_UI.html.

González-Astudillo, M. and D. Baquero. 2019. "A nowcasting model for Ecuador: Implementing a time-varying mean output growth." *Economic Modelling* 82: 250–263. DOI: https://doi.org/10.1016/j.econmod.2019.01.010.

Graff, M., D. Moctezuma, S. Miranda-Jiménez, S., E.S. Tellez. 2022. "A Python library for exploratory data analysis and knowledge discovery on Twitter data." *Computers & Geosciences, 159: 105012*. DOI: https://doi.org/10.1016Zj.cageo.2021.105012.

Guerrero, V.M., A.C. García, A. C., and E. Sainz. 2013. "Rapid Estimates of Mexico's Quarterly GDP." *Journal of Official Statistics* 29(3): 397–423. DOI: https://doi.org/10.2478/jos-2013-0033.

Harvey, A., and G. Phillips. 1979. "Maximum Likelihood Estimation of Regression Models With Autoregressive-Moving Averages Disturbances." *Biometrika* 152: 49–58. DOI: https://doi.org/10.1093/biomet/66.1.49.

Hewamalage, H., C. Bergmeir, and K. Bandara. 2021. "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions." *International Journal of Forecasting*: 37(1): 388–427. DOI: https://doi.org/10.1016/j.ijforecast.2020.06.008.

Huang, G.B., Q.Y. Zhu, and C.K. Siew. 2006. "Extreme learning machine: Theory and applications." *Neurocomputing* 70(1–3): 489–501. DOI: https://doi.org/10.1016/j.neucom.2005.12.126.

INEGI. a. Indicador Global de la Actividad Económica. Available at: https://www.inegi.org.mx/ temas/igae/

INEGI. b. Estimación Oportuna del PIB Trimestral. Available at: https://www.inegi.org.mx/temas/pibo/.

INEGI. c. Indicador Oportuno de la Actividad Económica. https://www.inegi.org.mx/investigacion/ioae/

INEGI. d. Indicador Mensual Oportuno de la Actividad Manufacturera. Available at: https://www.inegi.org.mx/investigacion/imoam/

Kourentzes, N., Barrow, D.K., and S.F. Crone. 2014. "Neural network ensemble operators for time series forecasting." *Expert Systems with Applications* 41(9): 4235–4244. DOI: https://doi.org/10.1016/j.eswa.2013.12.011.

León, C., and F. Ortega. 2018. "Nowcasting economic activity with electronic payments data: A predictive modeling approach." *Revista de economia del Rosario* 21(2): 381–407. DOI: https://dialnet.unirioja.es/servlet/articulo?codigo = 7411408.

Onatski, A. 2010. "Determining the number of factors from empirical distribution of eigenvalues." *The Review of Economics and Statistics*: 92(4): 1004–1016. DOI: https://doi.org/10.1162/REST_a_00043.

Ord, J., R. Fildes, and N. Kourentzes. 2017. *Principles of Business Forecasting-2nd Ed.* Wessex, Incorporated.

Poncela, P. and E. Ruiz. 2016. "Small versus big data factor extraction in Dynamic Factor Models: An empirical assessment in dynamic factor models." In *Advances in Econometrics*, edited by E. Hillebrand, and S.J. Koopman. 35: 401–434. DOI: https://doi.org/10.1108/S0731-905320150000035010.

Sezer, O.B., M.U. Gudelek, and A.M. Ozbayoglu. 2020. "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019." *Applied Soft Computing* 90: 106–181. DOI: https://doi.org/10.1016/j.asoc.2020.106181.

Shmueli, G. 2010. "To explain or to predict?" *Statistical science* 25(3): 289–310. DOI: https://doi.org/10.1214/10-STS330.

Stephens-Davidowitz, S. and H. Varian. 2014. *A hands-on guide to Google data*. Technical report, Google Inc. Available at: https://people.ischool.berkeley.edu/ ~ hal/Papers/2015/primer.pdf (accessed October 2021).

Stock, J.H., and M.V. Watson. 2011. "Dynamic factor models." In *Oxford Handbook of Economic Forecasting*. edited by M.P. Clements, and D.F. Hendry, Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398649.013.0003.

Tibshirani, R. 1996. "Regression shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1): 267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Varian, H.R. 2014. "Big data: New tricks for econometrics." *Journal of Economic Perspectives* 28(2): 3–28. DOI: https://doi.org/10.1257/jep.28.2.3.

World Bank, World Bank national accounts data. Available at: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Received October 2021
Revised February 2022
Accepted May 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 767–792, http://dx.doi.org/10.2478/JOS-2022-0034

Small Domain Estimation of Census Coverage – A Case Study in Bayesian Analysis of Complex Survey Data

Joane S. Elleouet[1], *Patrick Graham*[1], *Nikolai Kondratev*[1], *Abby K. Morgan*[1], *and Rebecca M. Green*[1]

Many countries conduct a full census survey to report official population statistics. As no census survey ever achieves 100% response rate, a post-enumeration survey (PES) is usually conducted and analysed to assess census coverage and produce official population estimates by geographic area and demographic attributes. Considering the usually small size of PES, direct estimation at the desired level of disaggregation is not feasible. Design-based estimation with sampling weight adjustment is a commonly used method but is difficult to implement when survey nonresponse patterns cannot be fully documented and population benchmarks are not available. We overcome these limitations with a fully model-based Bayesian approach applied to the New Zealand PES. Although theory for the Bayesian treatment of complex surveys has been described, published applications of individual level Bayesian models for complex survey data remain scarce. We provide such an application through a case study of the 2018 census and PES surveys. We implement a multilevel model that accounts for the complex design of PES. We then illustrate how mixed posterior predictive checking and cross-validation can assist with model building and model selection. Finally, we discuss potential methodological improvements to the model and potential solutions to mitigate dependence between the two surveys.

1. Introduction

In Aotearoa New Zealand a census is conducted every five years. It is a key input to official population estimates and supports a wide range of social and demographic analyses. Although the census would ideally count all people and their attributes of interest in the country at a given time, it inevitably fails to enumerate the full population. Censuses are expensive undertakings and the performance of the census in enumerating the population is therefore a matter of public interest. Consequently, a post-censal survey (the post-enumeration survey, henceforth PES) is conducted to evaluate the population coverage of the census. As well as providing an evaluation of the census, coverage estimates of the New Zealand census are used to adjust census counts in order to produce population estimates in the form of an estimated resident population (ERP) which is highly disaggregated to geographic and demographic groups. The population estimation system run by Statistics New Zealand (Stats NZ), New Zealand's official statistics agency, therefore requires coverage adjustments at a high level of granularity defined by, at least,

[1]Statistical Methods Unit, Stats NZ 8 Willis Street, PO Box 2922, Wellington 6011, New Zealand. Emails: joane.elleouet@gmail.com, patrick.graham.br@gmail.com, nikolai.kondratev@stats.govt.nz, abby.morgan@stats.govt.nz and rebecca.green@stats.govt.nz

Acknowledgments: The views expressed in this article are those of the authors and should not be taken to represent an official view of their affliated organisation. We thank three annonymous referees and an Associate Editor for thoughtful comments.

© Statistics Sweden

combinations of age in single year intervals, sex, ethnicity, 88 local government areas, Māori descent and country of birth (New Zealand or other). As these variables can potentially form hundreds of thousands of observed domains and the PES sample is made of approximately 30,000 individual records, direct estimation meets crucial limitations and the problem is best viewed as a modelling problem in which the objective is to relate the coverage probability to the covariates of interest. We therefore propose a Bayesian multilevel modelling approach to census coverage estimation. Although the official 2018 population estimates were created using a similar method (Stats NZ 2020b), the data and models used here differ from those used for the official published census coverage estimates and should not be regarded as official statistics.

Many countries with traditional census collections run a post-census coverage survey. Published applications include Brown et al. (2019), Chipperfield et al. (2017), Hogan (1993), Mule et al. (2008), with coverage estimation methods ranging from adaptations of dual systems estimation using a variant of the well-known Lincoln-Petersen estimator (Brown et al. 2019) to logistic regression of census coverage (Mule et al. 2008; Chen et al. 2010) followed by inverse coverage probability weighting of the census file to obtain population estimates. Our methods resemble the latter approach, though we use multilevel logistic models to obtain coverage and population estimates at a high level of granularity. Hierarchical Bayes models have also been proposed for estimation of the coverage of the Canadian census (You and Dick 2004). However, these are area-level models, in contrast to the individual level models discussed in this article. Elliott and Little (2000) developed a Bayesian model for census coverage estimation that incorporates information on population sex ratios, in addition to data from the census and a census coverage survey. However, the data structure assumed in that work differs from the one available for the current analysis.

Modelling complex survey data at the individual level requires attending to the impact of the survey design and nonresponse on inclusion in the data. Whereas the design-based approach to survey inference achieves this through the use of survey weights and variance calculations that respect the survey design, the model-based approach accommodates the impact of survey design and nonresponse on inclusion in the observed data within the model structure. The latter approach is often accompanied by the application of model-derived estimates to benchmark population data to obtain small domain estimates that account for differences in covariate structure between the sample and the target population, as illustrated by the so-called MRP (Multilevel Regression and Post-stratification) method (Gelman and Little 1997; Lax and Phillips 2009; Si et al. 2020). We cannot use population benchmarks to aid coverage estimation from PES because one of the purposes of PES is to adjust the census data to produce new population benchmarks. Nevertheless, the application of highly disaggregated model-derived estimates from PES to the census to produce estimates of the usually resident population has some parallels with the MRP approach to estimation.

Although the general Bayesian approach to analysis of complex survey data has been well described (Rubin 1987, chap. 2; Little 2003; Gelman et al. 2014, chap. 8), published applications of individual level Bayesian analyses of complex sample surveys remain relatively rare. Some recent applications, unrelated to census coverage, include small area official statistics (Nandram et al. 2018), political sciences (Ghitza and Gelman 2013;

Shirley and Gelman 2015), and public health (Paige et al. 2020), the latter using simulations to compare design-based to model-based approaches. Bayesian methods, and particularly multilevel Bayesian models, have more commonly been applied to area-level modelling of complex survey data for small domain estimation. In such applications, summary direct estimates with an associated variance estimate are first computed for each area and/or group of interest. Multilevel Bayesian models are then applied to smooth the summary statistics. In the case of complex survey data the direct estimates and variance estimates computed as the first stage of this procedure are usually design-based estimates. Examples include Ghosh et al. (1998), You and Chapman (2006), Molina et al. (2014), and Chen et al. (2014). Reviews of the general approach can be found in Pfeffermann (2013, 45–47) and Rao and Molina (2014, chap. 10). In this approach, design-based estimation is used to deal with the analytical complications of complex sample surveys, freeing the multilevel Bayesian modelling from the requirement to explicitly deal with the survey design.

Application of the area-level approach is problematic in our context where the number of covariate combinations (or domains) exceeds the number of records in the survey data set, so that forming the initial set of domain-level summary statistics is not even possible. Even applying the area-level approach to an aggregated version of the cross-classification of co-variates for which estimates are ultimately required, such that each covariate combination in the aggregated cross-classification occurs in PES, would be difficult unless the degree of coarsening is substantial. In sparse data situations with a binary outcome, conventional design-based variance estimates of proportions can often be zero and this makes subsequent modelling difficult. Consequently, framing the problem as estimation from a model fitted at the level of individual records and from which predictions can then be made seems a logical way forward. However, accounting for a complex survey design complicates the model so that the model fitted to the data is more complex than required for prediction. We illustrate how the model of interest can be, implicitly, recovered from the fitted model by integrating out parameters associated with the survey design but not relevant to the predictions. This article illustrates the potential of Bayesian modelling of complex survey data for challenging small domain estimation problems.

To describe our approach, we first describe the PES design in Section 2. We then present our modelling strategy in Section 3, including model-checking and evaluation. In Section 4, we show results of the model checking procedure. We also include summaries of standardised coverage estimates, by area and by age and ethnic group. The standardisation is achieved by applying the modelled coverage estimates for each group to a common reference population. The reference population used for this estimation is the population estimated by adjusting the census file for under-coverage using the disaggregated coverage estimates obtained from the model. Uncertainty in the estimation of the reference population is automatically incorporated in the posterior distribution for the standardised estimates. Section 5 concludes the article with some discussion of the modelling issues and suggestions for further development.

2. PES and Census Data

The official 2018 census data set comprises census respondents as well as records obtained from administrative sources (Stats NZ 2019). It is subject to under-coverage (eligible

residents missed by the census) and over-coverage (non-eligible individuals mistakenly counted, such as births after the census date and residents temporarily overseas at the time of census). In the 2013 census, over-coverage was approximately 0.7%, in contrast to an under-coverage measure of approximately 3.1% (Stats NZ 2014). The official ERP is corrected for both types of errors estimated on the full census file (Stats NZ 2020b). Estimates presented here differ from previously published estimates of the population coverage of the official census file and should not be regarded as official statistics. The main differences with the methodology used for official statistics is that we focus on under-coverage probability estimation and we perform the estimation on the respondent subset of the census file, which excludes administrative enumerations. However, the estimation challenge we describe is similar to the one faced in constructing the official 2018 ERP. Estimates for under-coverage probabilities hold without having to make assumptions about levels of over-coverage. Stats NZ (2020b) addresses over-coverage estimation in a very similar manner to under-coverage, and we refer the reader to this publication for more details on over-coverage estimation.

The 2018 PES used an area-based, stratified two-stage design. For sampling purposes New Zealand was divided into 23,174 small geographic areas (Primary Sampling Units, PSUs) that were grouped into 101 strata, based on a combination of broad geographic region, major urban status, census delivery mode (whether an access code for the online census form was mailed out or a hard copy census form was delivered) and a measure of deprivation. The PSUs were selected using probability proportional to size (PPS), where the size measure was based on historical estimates and included an adjustment for ethnic group proportions. Sampling fractions varied by strata, with urban strata sampled more intensively than non-urban strata, for fieldwork efficiency reasons. PES operated in all strata and a total of 1,365 PSUs were selected for the PES sample. In most PSUs, 11 dwellings were sampled within each PSU using Stats NZ's standard approach in which dwellings within a PSU are grouped into panels of size 11, and one panel is randomly selected. This resulted in a sample of 15,213 households within 15,015 dwellings in the 1,365 selected PSUs. Dwellings refer to the building in which people live, whereas people residing together and sharing facilities within a dwelling constitute a household, and there can be multiple households per dwelling. All usual residents at selected households were eligible for inclusion in the sample. Henceforth, we refer to households and use the terms household effects and household variables when referring to both household and dwelling characteristics. Within the 15,213 visited households, 37,548 people were interviewed. After filtering for refusal, incomplete responses and ineligibility, the final sample included 12,459 households and 31,600 respondents with responses of sufficient quality to be linked to the census and included in the estimation.

The PES sample was linked to the census file using a conservative probabilistic linking methodology, followed by clerical checking of all non-linked records and a sample of linked records. Details are described in Stats NZ (2020b). Of all eligible PES person records, 30,397 were linked to a census record (1,300 through manual linking), and the remaining 1,203 PES respondents were not linked to any census record. PES respondents linked to a record in the census respondent file were considered covered by census, whereas unlinked PES records constitute instances of under-coverage.

3. Census Coverage Estimation Under a Bayesian Modelling Framework

3.1. The Bayesian Approach

Fully model-based analysis of complex survey data usually requires multilevel models in order to account for the survey design. Such modelling fits neatly into a Bayesian framework. The Bayesian approach to inference permits coherent assessment of uncertainty for all model parameters and provides a flexible framework for propagation of parameter uncertainty to quantities derived from the model. We exploit this flexibility to obtain posterior distributions for highly disaggregated coverage probabilities (see Subsection 3.4) and for useful summaries of these probabilities (see Subsection 4.4). We generally specify prior distributions to be only weakly informative, in the sense of being open-minded as to the range of parameter values, while guaranteeing that inherent range constraints are respected (e.g., positive variances) and discouraging, but not disallowing, extreme values (Gelman et al. 2008).

In our application, we obtain a Monte Carlo approximation to the joint posterior distribution for all model parameters, by generating a sample from the posterior using Markov Chain Monte Carlo (MCMC) methods. Specifically, the sample is obtained using the program Stan (Stan Development Team 2020b) through the R interface (Stan Development Team 2020a; R Core Team 2019). Stan implements Hamiltonian Monte Carlo, a popular type of MCMC algorithm known to reduce the correlation between successive sampled values and, therefore, efficiently converging to the posterior distribution.

3.2. General Assumptions

A critical assumption of Bayesian analysis of survey data is ignorability (Rubin 1987, chap. 2; Little 2003; Gelman et al. 2014, chap. 8), which in the case of PES, requires conditional independence of inclusion in PES and inclusion in census, given the model covariates and *a priori* independence of the parameters of the models for inclusion in PES and in census. The former assumption is similar to the often invoked "independence" assumption of dual systems population estimation (Chandrasekar and Deming 1949; Brown et al. 2019). When ignorability holds, inference for inclusion in census (that is, census coverage) can proceed without specifying and fitting the model for inclusion in PES. In order to justify the assumption of ignorability, it is usually necessary to include the survey design features in the model, along with other covariates associated with nonresponse. We follow this approach in developing the model for census coverage. The nested geographical clustering of the sample design naturally lends itself to multilevel modelling, and, fortunately, in our case, there is overlap between variables of substantive interest and those predictive of nonresponse. We discuss the ignorability assumptions for our analysis in more detail in Section C of Online Supplemental Material, which tailors the general approach to Bayesian analysis of complex surveys given in Gelman et al. (2014, chap. 8) to the specific case of PES.

As well as conditional independence of inclusion in PES and census, we make the other standard assumptions of dual systems population estimation. We assume no errors in the

linkage of PES to census, and we assume the target population is closed over the operating periods of census and PES.

3.3. General Under-coverage Model

We let $\mathbf{X}$ denote demographic covariates and $\mathbf{x}$ a particular covariate combination. We use the notation TA to denote geographic area, and let $t \in \{1, \ldots, 88\}$ indicate a particular TA. To simplify notation in this section we let $\mathbf{V} = (\mathbf{X}, \text{TA})$, so $\mathbf{v} = (\mathbf{x}, t)$ refers to a particular covariate combination $\mathbf{x}$ in TA t. The sample space for $\mathbf{V}$ is the space of all covariate-TA combinations, denoted $\mathcal{V}$.

Introducing the indicators C and Q for inclusion in the census and in the target population respectively, we define the under-coverage probability as

$$p_{under}(\mathbf{v}, \boldsymbol{\xi}) = \Pr(C = 0 | Q = 1, \mathbf{V} = \mathbf{v}, \boldsymbol{\xi}),$$

where $\boldsymbol{\xi}$ is the parameter vector of the under-coverage model.

The purpose of the model presented here is to estimate $p_{under}(\mathbf{V}, \boldsymbol{\xi})$. A coverage-adjusted population estimate based on the census can subsequently be obtained by weighting each census record by the inverse of the under-coverage probability:

$$w_i = \frac{1}{(1 - p_{under}(\mathbf{v}_i, \boldsymbol{\xi}))}, \tag{1}$$

where the subscript refers to the i^{th} census respondent. Using a Bayesian approach enables this adjustment to be applied to each census record for each draw from the posterior distribution for $p_{under}(\mathbf{v}_i, \boldsymbol{\xi})$ in a Monte Carlo procedure which produces as many simulations of the ERP as needed to obtain precise uncertainty measures (e.g., approximate credible intervals). More details on the Monte Carlo methodology of the ERP production can be found in Bryant et al. (2016) and Stats NZ (2020a).

We let N_h^{ind}, denote the number of usual residents within household h, N_p^{hh}, the number of households in PSU p, N_s^{psu}, the number of PSUs in stratum s, N_t^{strat}, the number of strata intersecting TA t, $N_{\text{tot}}^{\text{strat}}$ the total number of strata and N^{ta}, the total number of TAs. After the linking procedure between PES and census, each record j in household h in the PES data set receives an under-coverage indicator Y_{hj} which states whether the record is present in the census file ($Y = 0$) or absent from it ($Y = 1$). Each record is also characterised by a set of demographic covariates $\mathbf{X}_{hj}^{\text{ind}}$, geographic variables related to the survey design, and local government area, TA. We present the model for census under-coverage in two ways: with a directed acyclic graph (DAG) (Figure 1), and with the following equations, followed by a description.

$$\left[Y_{hj} | X_{hj}^{\text{ind}}, p_{under_{hj}}\right] \overset{\text{indep}}{\sim} \text{Bernoulli}\left(p_{under_{hj}}\right); j = 1, \ldots, N_h^{\text{ind}}; h = 1, \ldots, N_{\text{psu}[h]}^{\text{hh}}, \tag{2}$$

$$\text{logit}\left(p_{under_{hj}}\right) = \alpha_h^{\text{hh}} + X_{hj}^{\text{ind}'}\beta; j = 1, \ldots, N_h^{\text{ind}}; h = 1, \ldots, N_{\text{psu}[h]}^{\text{hh}}, \tag{3}$$

$$\left[\alpha_h^{\text{hh}} | \mathbf{X}_h^{\text{hh}}, \alpha_{\text{psu}[h]}^{\text{psu}}, \beta^{\text{hh}}, \sigma_{\text{hh}}^2\right] \overset{\text{indep}}{\sim} \mathcal{N}\left(\mu + \alpha_{\text{psu}[h]}^{\text{psu}} + \mathbf{X}_h^{\text{hh}'}\beta^{\text{hh}}, \sigma_{\text{hh}}^2\right); h : \text{psu}[h] = p;$$
$$p = 1, \ldots, N_s^{\text{psu}}, \tag{4}$$

$$\left[\alpha_p^{\text{psu}}|\mathbf{X}_p^{\text{psu}}, \alpha_{\text{strat}[p]}^{\text{strat}}, \beta^{\text{psu}}, \sigma_{\text{psu}}^2\right] \overset{\text{indep}}{\sim} \mathcal{N}\left(\alpha_{\text{strat}[p]}^{\text{strat}} + \mathbf{X}_p^{\text{psu}\prime}\beta^{\text{psu}}, \sigma_{\text{psu}}^2\right); p : \text{strat}[p] = s; \quad s = 1, \ldots, N_t^{\text{strat}}, \tag{5}$$

$$\left[\alpha_s^{\text{strat}}|\mathbf{W}, \alpha^{\text{ta}}, \sigma_{\text{strat}}^2\right] \overset{\text{indep}}{\sim} \mathcal{N}\left(\mathbf{W}_s\alpha^{\text{ta}}, \sigma_{\text{strat}}^2\right), s = 1, \ldots, N_{\text{tot}}^{\text{strat}}, \tag{6}$$

$$\left[\alpha_t^{\text{ta}}|\mathbf{X}_t^{\text{ta}}, \beta^{\text{ta}}, \sigma_{\text{ta}}^2\right] \overset{\text{indep}}{\sim} t_3\left(\mathbf{X}_t^{\text{ta}\prime}\beta^{\text{ta}}, \sigma_{\text{ta}}^2\right), t = 1, \ldots, N^{\text{ta}}, \tag{7}$$

where the notation psu[h], and strat[p] refer, respectively, to the PSU of the h^{th} household and the stratum of the p^{th} PSU. Note $\alpha^{ta} = \left(\alpha_1^{ta}, \ldots, \alpha_{N^{ta}}^{ta}\right)'$ is a $N^{\text{ta}} \times 1$ vector of TA

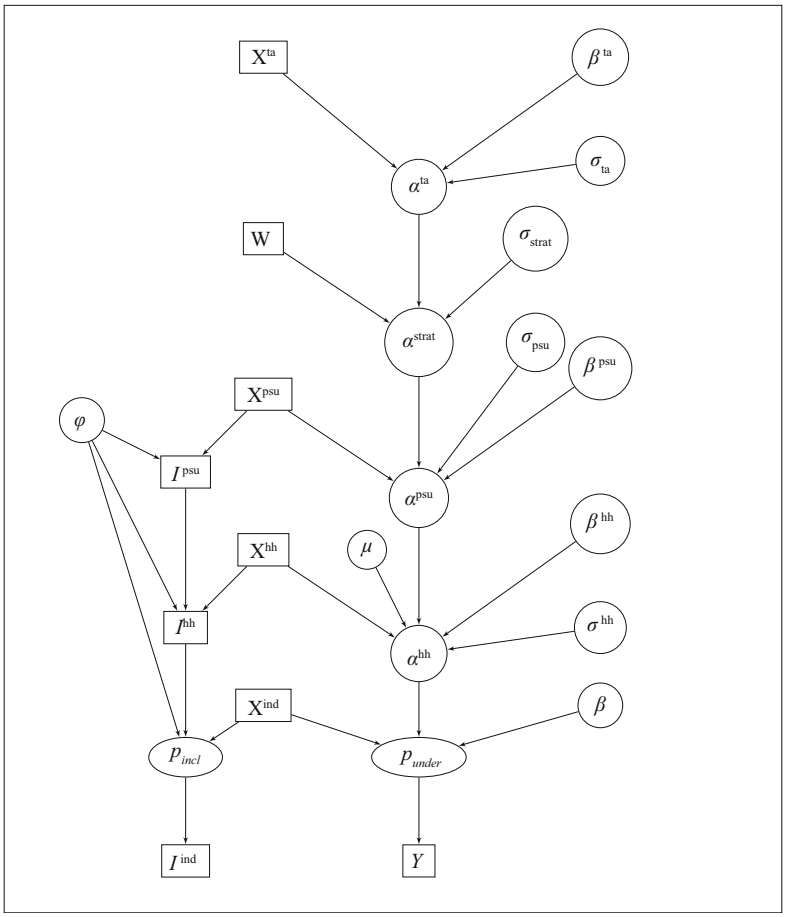


Fig. 1. Directed acyclic graph (DAG) representing the structure of the model for census coverage (Y). Rectangles represent observables and circles represent model parameters. p_{incl} and p_{under} are probabilities obtained as deterministic functions of observables and parameters. I^{psu}, I^{hh}, I^{ind} are indicators for PSU, household and individual inclusion in PES. The inclusion model depends on the parameter vector, φ. The inclusion indicators may depend on the model covariates but, because they are assumed conditionally independent of census coverage and parameters of the census coverage model, given the covariates, and because the inclusion and coverage model parameters are assumed to be a priori independent, inclusion in PES is ignorable. Under these assumptions, modelling of census coverage using PES data can proceed without modelling inclusion in PES.

effects. The notation t_3 in Equation (7) corresponds to a Student-t distribution with three degrees of freedom.

We model the under-coverage indicator Y_{hj} for individual j in household h using a Bernoulli distribution with probability $p_{under_{hj}}$ (2). A logistic regression is specified for $p_{under_{hj}}$ with individual covariates $\mathbf{X}^{\text{ind}}_{hj}$ and household-specific varying intercept α^{hh}_h (3). Equations (4)–(6) show how the varying intercept α^{hh}_h contains an overall average μ and all levels of the hierarchy reflecting the PES sampling design: it is modelled as a normal distribution, and the mean of this distribution is the result of a regression with a PSU-level effect and household covariate effects in Equation (4). These household covariates are a "hard-to-find" binary variable (HTF) which accounts for the variation in dwelling enumeration success between areas (details in Section D of the Online Supplemental Material), and potential individual demographic variables summarised at the household level. The PSU-level effect α^{psu}_p is itself a varying effect that we model with a normal distribution, and the mean of this distribution is the result of a regression with a stratum-level effect term and PSU covariate effects in Equation (5). These covariates are the PSU sampling variables used in the PES sampling design (see Table 1). The stratum-level effect α^{strat}_s is a varying effect modelled with a normal distribution whose mean is a weighted mean of TA-level effects from TAs present in the stratum (6). We add a TA level to the model, as this is the geographic resolution required for the publication of the ERP.

As each of the 101 strata generally spans several TAs, the relationship of strata to TAs is described by an occurrence matrix $\mathbf{W}$ where each row corresponds to a stratum and each column corresponds to a TA. Each matrix cell $\mathbf{W}(s,t)$ therefore represents the proportion of TA t included in stratum s. These cell proportions are estimated based on individual counts within small geographical units in the augmented census file, which includes administrative records in addition to census responses (Stats NZ 2019). We let $\mathbf{W}_s$ denote the s^{th} row of $\mathbf{W}$. Finally the TA effects α^{ta}_t are modelled through covariates $\mathbf{X}^{\text{ta}}$, which correspond to four socio-economic predictors of TA effects that are calculated from NZ Deprivation indices (Atkinson et al. 2019). We choose a t_3 distribution at the TALB level because it has more mass in its tails than the normal distribution, which helps avoid over-shrinkage at higher levels of the hierarchical model. A detailed description of the individual covariates included in $\mathbf{X}^{\text{ind}}$ as well as higher level covariates $\mathbf{X}^{\text{hh}}$, $\mathbf{X}^{\text{psu}}$ and $\mathbf{X}^{\text{ta}}$ is given in Table 1.

In practice, incorporating the group covariates $\mathbf{X}^{\text{psu}}$ and $\mathbf{X}^{\text{hh}}$ at the individual level of the model (3) by allocating all individuals in a group (household or PSU) the covariate values for that group gives an equivalent formulation to Equations (2–7) and makes subsequent predictions easier to compute. We let $\mathbf{X} = (\mathbf{X}^{\text{ind}'}, \mathbf{X}^{\text{hh}'}, \mathbf{X}^{\text{psu}'})'$. A demonstration of the equivalence of the two approaches is detailed in Section B of the Online Supplementary Material.

The choice of individual covariates used in Equation (3) is largely guided by New-Zealand post-enumeration surveys from previous years (Table 1). For instance, age, ethnic group and sex are known to affect census inclusion in distinct ways, so we include these variables and interactions in all models we examine. The four ethnic indicators are Māori, Pacific, Asian and Other. They are mutually non-exclusive, allowing individuals to belong to multiple ethnic groups. We added interaction terms between ethnicities representing two common profiles of people with multiple ethnicities: Māori-Other, and Māori-Pacific.

Table 1. Covariates used in the coverage model. "n. param" shows the number of parameters estimated for each of the covariates and covariate interactions. Covariates only used in the second model are depicted in grey.

Variable	Coding	Description	n. param
Individual covariates			
sex	binary	0 = male, 1 = female	1
age	10 splines	quadratic age splines with knots at ages 10, 20, 30, 40, 51, 61, 71, and 81	10
Māori	binary	Māori ethnicity indicator	1
Pacif	binary	Pacific ethnicity indicator	1
Asian	binary	Asian ethnicity indicator	1
Other	binary	indicator for "other" ethnicities	1
NZ born	binary	0 = born abroad, 1 = born in New Zealand	1
Māori descent	binary	0 = non-Māori descent, 1 = Māori descent	1
Individual covariate interactions			
Māori * Other	binary		1
Māori * Pacif	binary		1
sex * age	binary	sex and all 10 age splines	10
Asian * NZ born	binary		1
ethnicity * age	Binary	5 first age splines with each ethnicity and with Māori* Other (3-way)	25
Household covariates $\mathbf{X}^{hh}$ (model 2 only)			
Māori	binary	presence of Māori	1
Pacif	binary	presence of Pacific	1
Asian	binary	presence of Asian	1
Other	binary	presence of Other	1
Female	binary	presence of females	1
Māori descent	binary	presence of people of Māori descent	1
NZ born	binary	presence of people born in New Zealand	1
HTF	binary	hard-to-enumerate area	1
Household covariate interactions (model 2 only)			
between ethnicity indicators	binary		6
ethnicity * female	binary		1
ethnicity * NZ born	binary		1
PSU covariates $\mathbf{X}^{psu}$			
Pacif prop	continuous	proportion of Pacific adults	1
PSU size	categorical	S($<$ 50 dwellings)/M(50-100)/L(>100)	2
TA covariates $\mathbf{X}^{ta}$			
communication	continuous	Prop. of people with no access to internet at home	1
income	continuous	Prop. of people living in households with income below the poverty threshold	1
qualification	continuous	Prop. of people aged 18-64 without any qualifications	1
internet response	continuous	proportion of online census responses	1

For individual-level variables that are available but whose effect on coverage is less obvious, and for more subtle interactions between covariates, we compute several models differing in their covariates and interactions. Careful examination of resulting parameter

posterior distributions and predictions as well as out-of-sample deviance calculations are the used to guide model selection.

All individual covariates except age are binary variables. The challenge with modelling age is that it is inherently an ordered categorical variable with potentially more than 100 categories. Treating this variable as such creates the challenge of estimating a large number of parameters, and dividing the sample into excessively small categories. One solution is to create broader categories such as five-year age groups, but this solution does not reflect the continuous character of age and its effect on census coverage. It also introduces the additional issue of subjectively selecting categories, and creates breaks among contiguous years that may share extreme values. Another solution, implemented here, is to apply a spline transformation to the original variable. We model age using ten quadratic splines defined by eight internal breakpoints (see Table 1). Figure S1 (Online Supplemental Material) illustrates the transformation by showing the spline values for each age present in the census. One can see that at any given age, a maximum of three splines contribute to describing the underlying age. This stems from our choice of quadratic polynomials rather than higher-degree polynomials, in order to limit the smoothing of patterns that would result from highly overlapping spline curves.

We select *Cauchy* (0, 2.5) as a prior for μ, which is a standard prior recommended in Gelman et al. (2008). Covariate effect parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{\text{hh}}$, $\boldsymbol{\beta}^{\text{psu}}$ and $\boldsymbol{\beta}^{\text{ta}}$ are drawn from independent $\mathcal{N}(0,1)$ distributions. This is not unduly restrictive yet places low prior probability on extreme values. As a reference point, after converting to the odds ratio scale, a $\mathcal{N}(0,1)$ prior for a logistic regression parameter corresponds to the 95% prior interval $\exp(\pm 1.96)$, implying a $\exp(3.92) \approx 50$ – fold range of prior variation for the effect in question. Group-level variances σ^2_{ta}, σ^2_{strat}, σ^2_{psu}, and σ^2_{hh} are drawn from independent $Cauchy^+$ (0, 2.5) distributions, where $Cauchy^+$ refers to the Cauchy distribution truncated to positive values.

We run three HMC chains for each model, with the first half used as warm-up. We set the target average proposal acceptance probability to 0.9 and let all other algorithm parameters be set at their default value. For each model, we determine chain length experimentally by increasing it until convergence is reached. We ensure convergence by using the potential scale reduction factor, $\hat{R}$ (Gelman et al. 2014, 284–285) and by visually assessing chain profiles. We also monitor the effective Monte Carlo sample size to ensure appropriate post-convergence Monte Carlo sample size (Gelman et al. 2014, 286–287).

We explore potential models in two stages. We first focus on individual covariates and their interactions. Group-level covariates at the household, PSU and TA level as well as the stratum level are present in the varying intercept to account for the PES sampling design. The basic model therefore involves all individual covariates as well as the group-level covariates pertaining to the sample design.

However, results associated with this approach (see Section 4) suggest that individual covariates cannot fully account for variation in census coverage. As the census interview process is dwelling-based, households are an important component of the survey design and this is accounted for in the model through the first level of the varying intercept, α^{hh}. It is likely that census response is partly driven by household-level characteristics that are unobserved. It is also possible that an individual's response or nonresponse is influenced by another individual in the household. For instance, it is reasonable to suggest that

children's response to census is dependent on the parents or caregivers they live with. In such cases, we expect nonresponse of the former to depend on nonresponse of the latter, therefore bringing nonresponse at the household level. This is inconsistent with the model structure, which implicitly assumes that the household-level intercept and individual predictors are independent. To allow for correlation between household and individual characteristics, we follow the solution described in Gelman and Hill (2006, 506–507): we create versions of the individual covariates aggregated at the household level. Therefore, in a second stage, we experiment with the creation of many household-level covariates calculated from all individual covariates. The new covariates are included in $\mathbf{X}_h^{\text{hh}}$ in Equation (4), and described in Table 1. The outcomes from including these additional household covariates in the model are addressed in Sections 4 and 5.

3.4. Predicting Under-Coverage Probabilities of Census Records

To produce the ERP, coverage probabilities are required for each combination of covariates occurring in the census file. Household- and PSU-level covariates are included in the ERP production but not individual household or PSU effects. The geographic level for application of coverage probabilities is the TA level. While other choices could have been made, these settings provided a compromise between computational tractability and granularity of estimation. Below we describe how the model can be used to generate coverage probabilities at the desired level of demographic and geographic detail.

After fitting the multilevel logistic model, 1,000 samples are extracted from the posterior distribution. Each of the 1,000 draws from the posterior can be used to predict the undercoverage probability associated with each combination of covariate values that exist in the census. Parameters related to the sampling design (household, PSU, and stratum effects) are integrated to obtain a posterior prediction for each covariate-TA combination. For each combination of TA and individual, household and PSU level covariates, $\mathbf{v} = (\mathbf{x}', t)'$ where $\mathbf{x} = (\mathbf{x}^{\text{ind}'}, \mathbf{x}^{\text{hh}'}, \mathbf{x}^{\text{psu}'})'$, and for each draw from the posterior of $\boldsymbol{\xi}$, we require

$$
\begin{aligned}
&p_{\text{under}}(\mathbf{v}, \boldsymbol{\xi}) = \Pr(Y = 1 | \mathbf{V} = \mathbf{v}, \boldsymbol{\xi}) \\
&= \sum_{s=1}^{N_t^{\text{strat}}} \Pr\left(Y = 1 | \mathbf{X} = \mathbf{x}, \text{strat} = s, \boldsymbol{\xi}\right) \Pr\left(\text{strat} = s | \text{TA} = t\right) \qquad (8) \\
&= \sum_{s=1}^{N_t^{\text{strat}}} \left(\int \left(\Pr(Y = 1 | \alpha^{\text{hh}}, \mathrm{X}^{\text{ind}} = \mathrm{x}^{\text{ind}}, \text{strat} = s, \boldsymbol{\xi}) \times \right. \right. \qquad (9) \\
&\qquad \left. \left. p\left(\alpha^{\text{hh}} | \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \boldsymbol{\xi}\right)\right) \mathrm{d}\alpha^{\text{hh}} \times \Pr\left(\text{strat} = s | \text{TA} = t\right)\right) \\
&= \sum_{s=1}^{N_t^{\text{strat}}} \left(\int \text{expit}\left(\alpha^{\text{hh}} + \mathbf{x}^{\text{ind}'}\beta\right) \mathcal{N}\left(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'}\beta^{\text{hh}} + \mathbf{x}^{\text{psu}'}\beta^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2\right) \mathrm{d}\alpha^{\text{hh}} \right. \\
&\qquad \left. \times \Pr\left(\text{strat} = s | \text{TA} = t\right)\right), \qquad (10)
\end{aligned}
$$

where expit() is the inverse logit function and $\mathcal{N}(.|\mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. Writing $\mathbf{X} = \mathbf{x}$ instead of $\mathbf{V} = \mathbf{v}$ in the first component of Equation (8) follows from the assumptions of the model given by Equations (2–7). TAs affect census under-coverage only via strata, so after conditioning on strata, conditioning

on TAs becomes unnecessary. Similarly, we write $\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}$ instead of $\mathbf{X} = \mathbf{x}$ in the first element in the integral in Equation (9) because conditioning on α^{hh} means we do not need to condition on $\mathbf{X}^{\text{hh}}$ and $\mathbf{X}^{\text{psu}}$. The normal density for the household effects in the integrand in Equation (10) follows from the model Equations (4) and (5) since, by the mixture property of the normal distribution (Gelman et al. 2014, 577), we have

$$
\begin{aligned}
&p\left(\alpha^{\text{hh}} | \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \boldsymbol{\xi}\right) \\
&= \int p\left(\alpha^{\text{hh}} | \alpha^{\text{psu}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \boldsymbol{\xi}\right) p\left(\alpha^{\text{psu}} | \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \alpha_s^{\text{strat}}\right) \mathrm{d}\alpha^{\text{psu}} \\
&= \int \mathcal{N}\left(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}} + \alpha^{\text{psu}}, \sigma_{\text{hh}}^2\right) \mathcal{N}\left(\alpha^{\text{psu}} | \mathbf{x}^{\text{psu}'} \boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{psu}}^2\right) \mathrm{d}\alpha^{\text{psu}} \\
&= \mathcal{N}\left(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}} \boldsymbol{\beta}^{\text{hh}} + \mathbf{x}^{\text{psu}} \boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{hh}}^2\right).
\end{aligned}
$$

$\Pr(\text{strat} = s | \text{TA} = t)$ is estimated using an occurrence matrix constructed using the same data as $\mathbf{W}$, that is the official census file, which augments the census respondent file with administrative records.

The integral in Equation (10) produces predicted coverage probabilities that are marginalised with respect to household and PSU effects. That is, they are not predictions that are relevant to particular households, but are expectations over the distribution of household effects among households with covariate values x^{hh} in PSUs with covariates x^{psu}. An alternative, conditional prediction, could be obtained by setting the household and PSU effects to zero (or some other value) but such predictions are tied to households and PSUs with the specified effect and would not be appropriate for application to the census file for which the desired notion is that of an unknown household with particular household covariate values in an unspecified PSU with particular PSU covariate values. Further discussion on the marginal and conditional predictions can be found in Skrondal and Rabe-Hesketh (2009) and Pavlou et al. (2015) and some more details on the derivation of Equation (10) are given in Section B of the Online Supplemental Material. In our application we use Monte Carlo integration to approximate the integral in Equation (10).

4. Results

4.1. *Using Mixed Predictive Checks to Assist With Model Assessment*

Models were run with three HMC chains of sufficient length (11,000–12,000 iterations) to ensure convergence. Stan run times with parallel chains were 8.0 hours for the initial model (model 1), and 17.9 hours for the model with household covariates (model 2, see Subsection 4.2). After discarding the first half of each chain as warm-up period, the $\hat{\mathrm{R}}$ convergence diagnostic (Gelman et al. 2014, 284–285) was less than 1.01 for all monitored parameters for both models. We assess the quality of the models using posterior predictive checking focusing on marginal predictions for two different groupings: demographic categories formed by all binary demographic covariates, and TAs. Results for all checks performed on two models are presented in Figure 2 for predictions on demographic groupings, and in Figure 3 for TA-level predictions. Note that only the first 42 TAs of the North Island are shown in Figure 3 (See Online Supplementary Material Figure S2 for all other TAs). For the

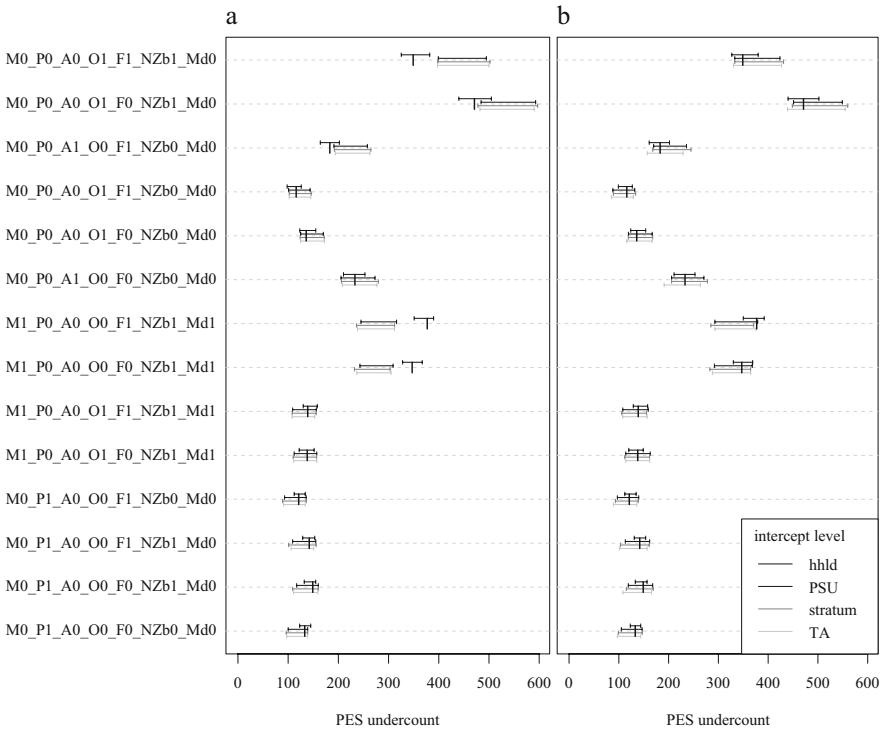


Fig. 2. Posterior predictive checks (PPCs) for model 1 (without household covariates, left) and model 2 (with household covariates and their interactions, right). Only the 14 most common demographic categories are represented, out of 81 demographic categories, all geographic areas pooled. The vertical black bars represent the PES observed counts for the categories displayed on the left axis. PPC results for different levels of integration are represented by horizontal 90% credible intervals. For each category, the top interval corresponds to PPCs performed on the raw output from the logistic regression, and lower intervals (different shades of grey) correspond to PPC results after Monte Carlo integration of household parameters, household and PSU parameters, and household, PSU, stratum parameters, respectively. The labels on the left are to be interpreted as a combination of demographic variables (letters) with whether or not the category comprises individuals corresponding to the demographic variable (0 = no, 1 = yes). M = Māori, P = Pacific, A = Asian, O = Other, F = Female, NZb = New-Zealand born, Md = Māori descent. For instance the top category corresponds to females of Other ethnicity only, who were born in New Zealand and are not of Māori descent.

first predictive check, we use each sample from the joint posterior distribution of parameters to replicate the PES data under the logistic model described in Equations (2)–(3). We compare the 1,000 simulated data sets with the observed data. We summarise the aggregated undercount distributions from simulated data sets using 90% posterior predictive intervals and assess whether observed undercounts fall within these intervals (Figures 2a and 3a, top intervals). This self-consistency check allows us to confirm that the model fits the data: all observed aggregated undercounts fall within the 90% posterior predictive intervals from simulated data, for both demographic and geographical groupings.

The PES model is designed to predict under-coverage for census records. Census records can be considered as "new observations" that we need the model to output predictions for. We therefore need to assess not only the fit but also the predictive ability of the model when confronted with new observations. This is especially important as these observations do not fit into the hierarchy of households, PSUs, and strata that was solely defined to account for

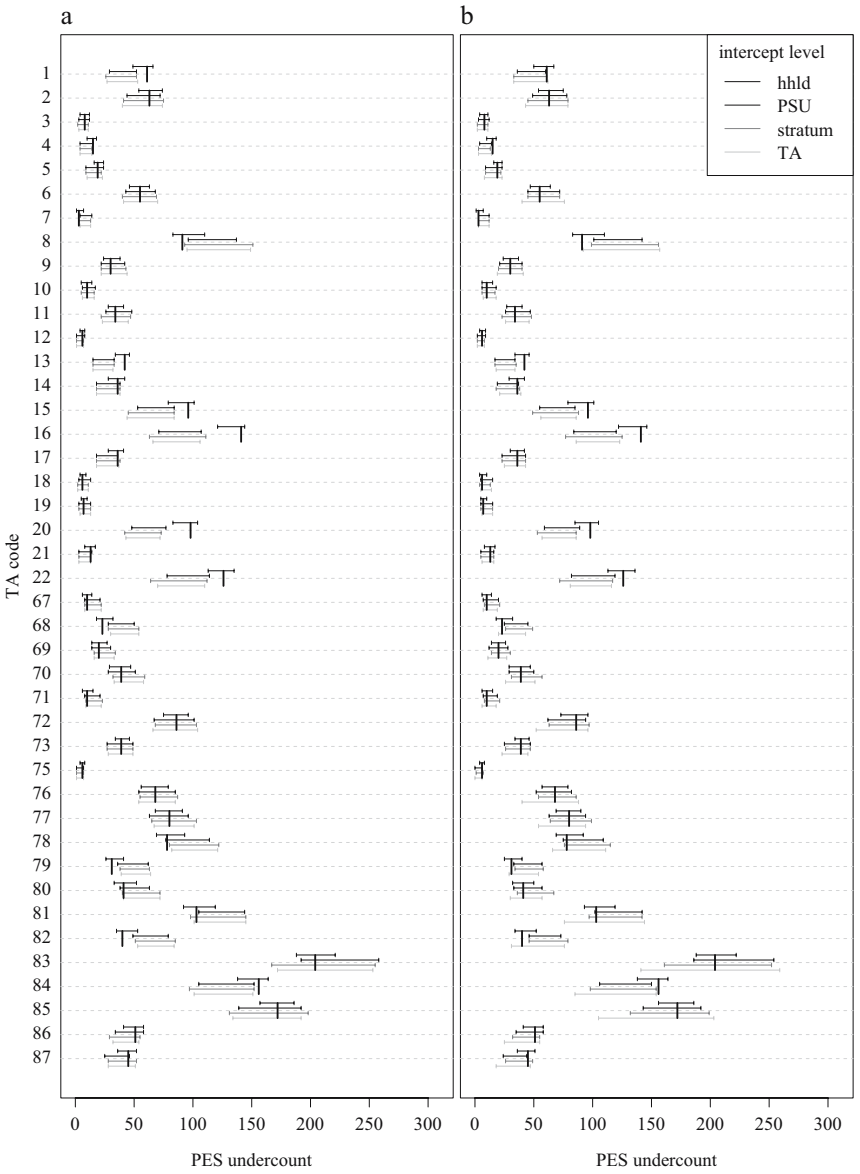


Fig. 3. Posterior predictive checks (PPCs) for model 1 (without household covariates, left) and model 2 (with household covariates and their interactions, right). Undercounts by TAs are represented for the first 42 TAs of the North Island, all demographic categories pooled. See Figure 2 caption for further details.

the PES sampling design. Therefore, we need to determine how good the model is at estimating under-coverage probabilities for census records, which are characterised by the same TA and demographic information as PES records but are for the most part not included in the households and PSUs selected for PES. This can be done using mixed predictive checking, whereby predictions are performed for new individuals (outside of the PES sampling frame) with exactly the same demographic and group-level predictors as PES individuals (Gelman et al. 1996). In our case, it amounts to applying Equation (10) to all PES

records, drawing from the Bernoulli process to simulate the under-coverage indicator, and aggregating the results to the same groupings as previous posterior predictive checks (TA and demographic categories). The results are displayed as light grey 90% credible intervals (bottom interval) on Figures 2a and 3a. With results aggregated to demographic groups, the model shows a substantial misfit for five out of the 14 most common demographic groups, which is a higher proportion than the 10% roughly expected under the assumption that the model is adequate. The TA grouping also shows widespread misfit, with almost a third of all TA under-coverage counts lying outside of the predicted 90% credible intervals.

When a misfit is observed after integration of the sampling design parameters, it is useful to investigate what level of the hierarchy causes the problem, especially in models with more than two levels. In our case, unaccounted for variation could be present at the individual, household, PSU, or stratum level. To assess the problematic level, we compute mixed predictive checks where some but not all of the grouping levels in the hierarchy are integrated. We first perform predictions from the PES data considering that the stratum and PSU of each individual are known but the household is new, therefore sampling the varying intercept from the population distribution for households with their given covariates. For an individual in PSU p, in a household with household covariate values x^{hh} and with individual covariate value $\mathrm{x}^{\mathrm{ind}}$, this means calculating the following under-coverage probability:

$$q^{\mathrm{psu}}(\mathbf{x}^{\mathrm{ind}}, \mathbf{x}^{\mathrm{hh}}, \boldsymbol{\xi}, \alpha_p^{\mathrm{psu}}) = \int \mathrm{expit}(\alpha^{\mathrm{hh}} + \mathbf{x}^{\mathrm{ind}'}\boldsymbol{\beta})\mathcal{N}(\alpha^{\mathrm{hh}} | \mu + \mathbf{x}^{\mathrm{hh}'}\boldsymbol{\beta}^{\mathrm{hh}} + \alpha_p^{\mathrm{psu}}, \sigma_{\mathrm{hh}}^2)\mathrm{d}\alpha^{\mathrm{hh}}.$$

The results are displayed as dark grey 90% credible intervals (second from the top) on Figures 2a and 3a. We repeat the procedure where both households and PSUs are new, which in practice consists in sampling the intercept from the population distribution for households, integrated over PSUs. In this case, the under-coverage probability in stratum s is calculated as follows:

$$q^{\mathrm{strat}}(\mathbf{x}^{\mathrm{ind}}, \mathbf{x}^{\mathrm{hh}}, \mathbf{x}^{\mathrm{psu}}, \boldsymbol{\xi}) = \int \mathrm{expit}(\alpha^{\mathrm{hh}} + \mathbf{x}^{\mathrm{ind}'}\boldsymbol{\beta})\mathcal{N}(\alpha^{\mathrm{hh}} | \mu + \mathbf{x}^{\mathrm{hh}'}\beta^{\mathrm{hh}} + \mathbf{x}^{\mathrm{psu}'}\beta^{\mathrm{psu}} + \alpha_s^{\mathrm{strat}}, \sigma_{\mathrm{hh}}^2 + \sigma_{\mathrm{psu}}^2)\mathrm{d}\alpha^{\mathrm{hh}}.$$

The results of these predictions are displayed as grey 90% credible intervals (third from the top) on Figures 2a and 3a.

4.2. Adding Household Covariates Improves the Model for Demographic Groups

The four performed predictive checks, with integration occuring at different grouping levels, clearly show that a major misfit arises when predicting under-coverage of individuals in new households. This misfit is visible for both demographic groupings (Figure 2a) and TA groupings (Figure 3a), but does not grow larger when predictions are calculated with unknown PSUs and/or strata. This suggests that the model did not estimate an adequate population distribution for households. Further graphical investigations (not shown) suggest some dependence between the varying intercept and individual covariates. As noted in Subsection 3.3, the logistic model described in Equation (2)–(7) assumes independence between intercept and individual-level predictors. Following Gelman and Hill (2006, 506), we address

this inconsistency by aggregating the individual covariates suspected to cause the dependency to the hierarchical level in question, and introduce the new variables as group-level covariates. We first test different ways of aggregating demographic covariates (ethnicity, age, sex, and New-Zealand born) at the household level. For continuous variables, it makes sense to average values across individuals in a group. However, several ways to aggregate categorical variables at the group level can be considered. For each individual categorical covariate, we test the following aggregation methods: (1) binary variable indicating presence/absence of household occupants with the demographic characteristic, (2) continuous variable of proportion of household occupants with the demographic characteristic, and (3) binary variable indicating a majority of household occupants with the demographic characteristic. We find that overall model performance is best when using (1) for all group-level covariates. We therefore only present results with these covariates (see Table 1 for the final list of household covariates and their interactions). We apply the same four predictive checks as we applied to the original model, and present the results in Figures 2b and 3b.

Introducing household covariates and their interactions considerably improves the fit of predicted undercounts of demographic groups to the observed data (Figure 2), with only one observed value sitting just outside the 90% credible intervals from the TA-level prediction. However, the modification only partially improves the fit to TA counts (Figure 3). While the model including household covariates fits most TAs well, seven TAs on Figure 3 still show predictive credible intervals that contain the true value when simple posterior predictive checks are performed but do not encompass it when performing any of the mixed predictive checks. For these TAs, the estimated household population distribution seems wrong, and we hypothesise that some unknown household characteristics cause unaccounted-for heterogeneity. Consistent with this hypothesis, between household variation is the largest component of unexplained variation and did not shrink after the inclusion of household level covariates (Table 2).

4.3. *Complementary Cross-validation Tests*

Posterior predictive checks give insight into the fit of the model to the data. They are a first "sense-check" of an analysis. The three additional mixed predictive checks, simulating data with the same covariates as the data but different hierarchical groups, constitute a step further in assessing not only the fitness but also the predictive power of the model, and where its limitations lie. However, these checks still use the exact same covariate values in the simulated data sets as in the observed data sets. One way to further determine the

Table 2. Marginal posterior quantiles of variance parameters for model 1 (without household covariates) and model 2 (with household covariates)

	Model 1			Model 2		
	2.5%	50%	97.5%	2.5%	50%	97.5%
σ^{hh}	4.10	4.35	4.62	4.18	4.45	4.74
σ^{psu}	0.29	0.76	1.05	0.08	0.63	0.98
σ^{strata}	0.01	0.18	0.45	0.01	0.15	0.42
σ^{ta}	0.01	0.13	0.43	0.01	0.12	0.43

predictive power of the models is cross-validation. For all tested models, and to assist with model selection, we calculated approximate leave-one-out-cross-validation scores using Pareto-smoothed importance sampling (PSIS). We show the results for the two main models presented in Subsections 4.1 and 4.2 in Table 3. Lower leave-one-out cross-validation scores (or their importance sampling approximation, LOOIC) indicate a lower out-of-sample deviance, and therefore more accurate predictions to new data. Model 2, with additional household covariates, has a lower LOOIC value than model 1, although the difference is of the order of one standard error. While LOOIC is the most appropriate measure of predictive accuracy for complex hierarchical models, the Pareto-k diagnostic values for both models but especially model 2 suggest the error in the LOO approximations might be high and the LOOIC values might understate predictive accuracy (Vehtari et al. 2017). This is typical of flexible hierarchical models where some groups have very few observations. Both model 1 and 2 are like this: the lowest level of the hierarchy, households, often contains only one or two observations (individuals).

4.4. Standardised Estimates

The PES model output gives estimates of $p_{under}(\mathbf{v}, \boldsymbol{\xi})$ and gives us insight into how different geo-demographic groups respond to census. From a demographer's point of view and for the sake of planning future censuses, it is also valuable to know what factors are actually driving nonresponse patterns. For instance, in a TA with a high estimated census undercount, it can be of interest to know if nonresponse is due to the demographic composition of the TA, or if there are there intrinsic difficulties associated with operating a large-scale survey in this area. If demographic effects are predominant, then we can assume nonresponse is driven by behavioural patterns in the respondents, whereas area effects would suggest potential issues with incomplete address registers or other operational pitfalls. Insight into the relative impact of the different covariates on census coverage can be gained by calculating under-coverage probabilities across the categories of the variable of interest for a standardised distribution of all other covariates. For instance, one can obtain area-level estimates where differences due to their demographic composition are statistically removed, leaving only differences pertaining to intrinsic area characteristics. The same standardisation logic can be applied to other variables, for instance one can obtain estimates by ethnicity, standardising

Table 3. Out-of-sample deviance diagnostics for the model without household covariates (w/o hh covar.) and with household covariates (w/ hh covar.). $elpd_{loo}$: expected log pointwise predictive density. p_{loo}: effective number of parameters. LOOIC: Pareto-smoothed importance sampling leave-one-out cross-validation approximation. Values in brackets correspond to standard error estimates. The Pareto-k distribution section bins estimates of importance for all data records into categories ordered by decreasing quality.

model	$elpd_{loo}$	p_{loo}	LOOIC	Pareto-k distribution			
				$(-\infty,0.5]$	(0.5,0.7]	(0.7,1]	(1,1)
W/o hh covar.	−5,747.1 (100.6)	2,858.6 (60.3)	11,494.3 (201.2)	65.1%	24.0%	9.9%	0.9%
W/ hh covar.	−5,639.8 (99.4)	2,814.0 (59.7)	11,279.7 (198.8)	62.1%	25.5%	11.3%	1.1%

areas and all other demographic covariates. Following the example of TA-level standardised estimates, we can define, for a given TA t:

$$p_{under}^{\text{std}}(t, \boldsymbol{\xi}) = \sum_{\mathbf{x}} p_{under}(\mathbf{x}, t, \boldsymbol{\xi}) \Pr^{\text{std}}(\mathbf{X} = \mathbf{x}|\boldsymbol{\xi}), \tag{11}$$

where $\Pr^{\text{std}}()$ refers to the covariate probabilities from some standard distribution. Note that the standard distribution is allowed to depend on the model parameters. This is not usual but suits our situation because a natural choice of standard population is the corrected version of the census file, based on the under-coverage probabilities, estimated from the model. Thus, if the inverse under-coverage probability for the i^{th} census record corresponding to a particular setting of parameter values $\boldsymbol{\xi}$ is $w_i(\boldsymbol{\xi})$ we can define $\Pr^{\text{std}}(\mathbf{X} = \mathbf{x}|\boldsymbol{\xi})$ as

$$\Pr^{\text{std}}(\mathbf{X} = \mathbf{x}|\boldsymbol{\xi}) = \frac{\sum_{i:\mathbf{X}_i=\mathbf{x}} w_i(\boldsymbol{\xi})}{\sum_i w_i(\boldsymbol{\xi})}, \tag{12}$$

where the summations are over records in the census file. With the standard probabilities defined as in Equation (12), standardised under-coverage probabilities can be obtained for each TA and repeating this for each draw from the posterior for ξ will produce a sample from the joint posterior for the standardised under-coverage probabilities by TA. Credible intervals and other summaries, including for contrasts between TAs, can be computed from the posterior sample. The standardised coverage probability, given by Equation (11) can be contrasted with the marginal TA under-coverage probability which is

$$\Pr(Y = 1|TA = t, \boldsymbol{\xi}) = \sum_{\mathbf{x}} p_{under}(\mathbf{x}, t, \boldsymbol{\xi}) \Pr(\mathbf{X} = \mathbf{x}|TA = t, \boldsymbol{\xi}). \tag{13}$$

Comparing Equations (13) and (11) it can be seen that they differ only in the covariate distribution, with the standardised probabilities using the covariate distribution of the chosen standard population in place of the TA-specific covariate distribution used to obtain the marginal coverage probability. By definition, the standardised probabilities are all based on the same covariate distribution, so differences in standardised TA under-coverage probabilities reflect genuine geographic differences in census under-coverage.

We focus on standardised under-coverage probability estimates for TAs and for age-sex profiles, using the output from the model with household covariates (model 2). Figure 4 shows results across TAs. Although uncertainty bounds are wide and overlap among most TAs, some TAs seem to have higher nonresponse probabilities than the majority, all else being equal. As our mixed predictive checks have identified some inaccuracies in the undercoverage predictions for some TAs (Subsection 4.2), we are cautious about drawing conclusions on our TA-level standardised estimates. Figure 5 shows that Māori and Pacific people generally have higher nonresponse levels than other ethnic groups. People in their twenties have the highest nonresponse levels across all ethnic groups. A secondary under-coverage peak around age 50 is also visible, although it is more pronounced in Pacific people than in other ethnic groups.

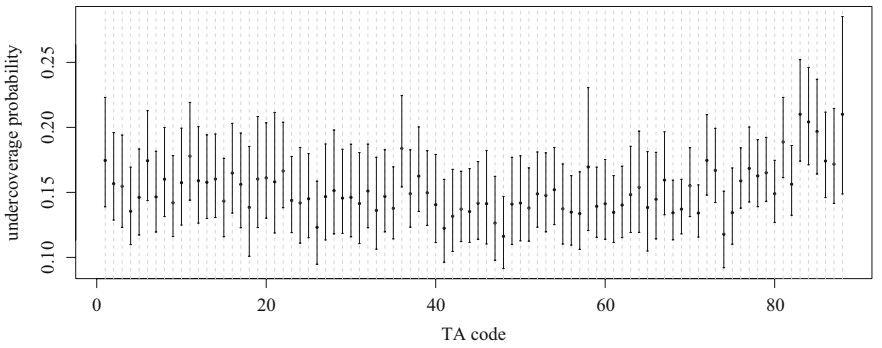


Fig. 4. Standardised estimates of under-coverage probabilities for each TA. Points correspond to posterior medians and error bars correspond to 90% credible intervals.

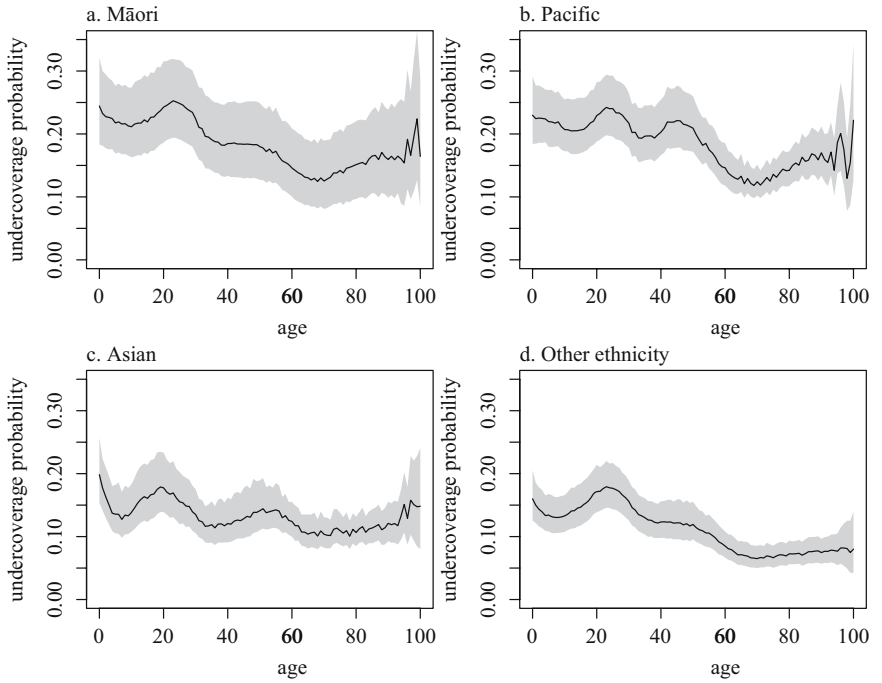


Fig. 5. Standardised estimates of under-coverage probabilities by age, for each ethnicity. Black lines correspond to posterior medians and grey shading corresponds to 90% credible intervals.

4.5. Effects of Coverage Adjustment on Census Counts

As stated in Subsection 3.3, census counts by demographic and geographic categories are corrected using the posterior values for p_{under} ($\mathbf{v}_i$, $\boldsymbol{\xi}$), following Equation (1). Although producing adjusted counts is out of scope for this article, it is useful to provide a sense of the scale of the correction to the census counts implied by the estimated under-coverage probabilities. Assuming model 2 is chosen, we find that overall average undercoverage of census responses is 10.9%. For a hypothetical census data set comprising four million responses this would mean adding about 490,000 individuals to the census respondent population. However, some demographic groups are better represented in census than

others. We find that census undercoverage in young male adults of Pacific and Māori populations can reach 35%. This means that the population for such categories is about 1.5 times larger than in the census responses, though in the official census file this effect is offset by the inclusion of adminstrative records.

5. Discussion

5.1. Posterior Predictive Checks Help Understand and Select the Multilevel Model

Through the present analysis, we have illustrated how census coverage can be quantified using a modelling approach to the analysis of complex survey data, thereby allowing insights at high levels of granularity in demographic and geographic attributes. To estimate undercoverage of the New Zealand 2018 census, we have fitted a binomial model with four nested geographic levels reflecting the complex sampling design of the post-censal survey. Especially, we have shown that multilevel models are a promising approach to analyse survey data with complex sampling designs. We reiterate that our results are experimental results that reflect a different analysis to the one used to output official 2018 census coverage estimates (Stats NZ 2020b).

We have also illustrated how performing extensive posterior predictive checking can assist in model selection in the case of a complex multilevel model structure. Combining posterior predictive checking, mixed predictive checking and cross-validation allowed us to assess the fit and predictive limitations of competing models as well as identifying aspects of the models that require modifications. Especially, performing mixed predictive checking at all levels of the hierarchy allowed us to identify the lack of fit at a specific level (here, the household level). We could assess the improvement associated with the subsequent addition of household-level covariates using additional mixed predictive checks and comparing cross-validation results across models.

Even after the addition of household-level covariates, posterior predictions did not always fit under-coverage data at the TA level. Some TA under-coverage counts were over-estimated, with 10 (of 88) TAs having PES-observed census under-coverage counts below the predicted 90% credible intervals. Other TA counts were under-estimated, with 12 TAs having observed under-coverage counts over the predicted 90% credible intervals (Figures 3 and S2). Mixed posterior predictive checks at each level of geographic parameter integration allowed us to attribute most of this misfit to the household level. The lack of fit is unlikely to be related to the demographic attributes of household occupants, as most of these attributes have been accounted for as household-aggregated level variables, and posterior predictive checks for demographic groups show no apparent bias. Subsequent investigations have failed to identify commonalities between TAs with similar misfit patterns.

The only noticeable pattern in this result is the relationship between observed undercoverage proportion of a TA and the direction of the estimation error: TAs which tend to be under-estimated are the ones with high observed under-coverage, whereas over-estimated TAs tend to have a low observed under-coverage proportion. Taken at face value this result suggests the model may be over-regularising more extreme estimates. In multilevel modelling, one expects small hierarchical groups with extreme observations to have predictions shrunk towards their expected value under the model. As the number of

levels increases, we can expect shrinkage to increase too for predictions made at higher levels. However, if overshrinkage, *per se*, was the main reason for the TA-level misfit, we would expect the issue to primarily affect smaller areas, whereas several of the TAs that the model fits poorly are large urban areas with a relatively large PES sample size. Further, if over-shrinkage was the primary reason for the lack of model fit in some areas, we would expect predictions to gradually show more shrinkage as we move from household-level predictions towards TA-level predictions, instead of a single jump from adequate fit at the household level to misfit for prediction at all other levels. We experimented with replacing the normal distributions for household, PSU and stratum effects by the heavier tailed t distribution with three degrees of freedom, but this had no impact on results. If the lack of fit in some TAs was due to the normal models tending to over-shrink extreme values we would have expected to see some improvement in fit when t_3 priors were adopted. The most plausible explanation for the lack of fit in some TAs is that one or several important factors related to geography were not included in the model. If this explanation holds, it follows that estimates are, in some cases, being shrunk towards expectations that do not exhibit the appropriate amount of geographic variation because a geographically varying covariate has been omitted from the model. In this specific sense, the estimates for some TAs may be exhibiting the effects of over-shrinkage. As noted above, given our posterior predictive checking results, the omitted covariate(s) seems likely to be a household-level variable, such as an aspect of dwelling construction (e.g freestanding versus in an apartment block) that varies by area and is related to census coverage (e.g., census enumeration may be more difficult in apartment blocks). A natural next step for future model improvement would involve attempting to identify the missing covariate(s). If the missing covariates cannot be identified or sourced, a potential alternative is to specify the problematic group distribution as a mixture of several distributions. This may allow the model to recover unobserved categories within groups and improve model fit. However moving to mixture distributions at one or more of the model levels introduces additional computational complexity.

5.2. Individual Demographic Characteristics Drive Census Coverage Patterns

Standardised estimates give insight into the role played by different demographic and geographic attributes in driving coverage differences between groups. Though common practice in epidemiology and demography, the use of standardisation to adjust for differences in co-variate distributions has been less common in official statistics. In our case it provided a simple way to present comparative results from a complex model. Comparing Figures 4 and 5 suggests high census under-coverage patterns are in general driven more by individual demographic attributes than by geographic ones. For instance, Māori and Pacific people as well as people in their early twenties tend to respond to census less than other demographic groups, independently of where they live. Although most TAs do not seem to intrinsically drive census under-coverage, clusters of TAs with higher under-coverage propensity can be identified. In this case standardised estimates can be used in the planning of future census operations. For instance, incentivisation and follow-up efforts could be allocated more heavily in TAs where under-coverage propensity has historically been high.

5.3. Design-Based versus Model-Based Approach

The most common approach to analysing complex surveys has traditionally been through a design-based method, where individual sampling weights are calculated from the sampling frame and subsequently adjusted for nonresponse. This approach has limitations when survey nonresponse is difficult to track. For instance, we do not know the number of occupants in nonresponding households nor the number of nonrespondents in a responding household. Sampling weights are often adjusted for nonresponse using benchmark population data to ensure that weighted sample distributions are close to known population distributions. However, for PES, such benchmark population data is not available, because PES is used in conjunction with census to estimate a new benchmark population. A further challenge to the application of design-based methods in PES is the absence of an accurate count of dwelling numbers by PSU at the time of the PES fieldwork, which complicates the computation of selection probabilities and hence sampling weights. Moreover, a sample size of about 15,000 households does not allow precise design-based estimates at the required level of geographic and demographic disaggregation. In this regard, the modelling approach seems natural, especially when geographical attributes are treated in a hierarchical fashion. Multilevel modelling facilitates pooling of information across areas and is desirable for small area and small domain estimation problems.

Modelling of survey data is, of course, possible from a design-based perspective, though there appear to be efficiency gains through explicitly modelling the survey design structure rather than dealing with the impact of survey design through sampling weights (Lumley and Scott 2017). Design-based multilevel modelling is challenging, because the pseudo-likelihood methods commonly used for design-based fitting of single-level models are more difficult to apply in the case of multilevel models. Pseudo-likelihood estimates of multilevel models are potentially sensitive to the scaling of survey weights, even when design and analysis clusters are identical (Rabe-Hesketh and Skrondal 2006). Methods based on pairwise composite likelihood are a promising alternative to pseudo-likelihood methods for fitting design-based multilevel models but require knowledge of joint selection probabilities (Rao et al. 2013; Yi et al. 2016). In the PES analysis the geographic clusters of analytical interest are the TAs, which were not part of the sample design and this further complicates the application of design-based methods to multilevel modelling (Lumley and Scott 2017).

5.4. Mitigating the Ignorable Inclusion Assumption

One of the fundamental assumptions of the PES model is the independence between inclusion in PES and inclusion in census, conditional on design features and covariates included in the model. This means that the list of dwelling addresses used for census and the PES sampling frame need to be built independently, a requirement sometimes difficult to satisfy. Another challenge to the conditional independence assumption is respondent behaviour. For instance, a respondent's negative experience with census might influence whether they open the door to PES interviewers. Such behaviour would lead to non-ignorable nonresponse and complicate the analysis by requiring that the model for inclusion in PES be explicitly formulated and included in the model fitting. Pfeffermann et al. (2006) develops a conditional likelihood approach to incorporating non-ignorable

nonresponse in multilevel modelling of survey data. Extending the PES model to deal with non-ignorable nonresponse may be a worthwhile direction for future development of the model. In Figure 1, this would result in additional edges between one or more of the inclusion indicators and the coverage indicator, illustrating the need to explicitly specify the model for inclusion and to estimate the inclusion model jointly with the coverage model. Alternatively, it may be possible to incorporate external information that allows the assumption of conditional independence between PES and census inclusion to be weakened (Elliott and Little 2000; Brown et al. 2019).

6. References

Atkinson, J., C. Salmond, and P. Crampton. 2019. *NZ Dep 2013 index of deprivation interim research report.* Technical report, Department of Public Health, University of Otago,Wellington. Available at: https://www.otago.ac.nz/wellington/otago730394.pdf. (accessed February 2022).

Brown, J.J., C. Sexton, O. Abbott, and P.A. Smith. 2019. "The framework for estimating coverage in the 2011 census of England and Wales: Combining dual-system estimation with ratio estimation." *Statistical Journal of the International Association of Official Statistics* 35: 481–499. DOI: https://doi.org/10.3233/SJI-180426.

Bryant, J., K. Dunstan, P. Graham, N. Matheson-Dunning, E. Shrosbree, and R. Speirs. 2016. *Measuring uncertainty in the 2013-base estimated resident population* (Statistics New Zealand Working paper: 16–04). Wellington, New Zealand: Statistics New Zealand. Available at: https://www.stats.govt.nz/ (accessed February 2022).

Chandrasekar, C. and W.E. Deming. 1949. "On a method of estimating birth and death rates and the extent of registration." *Journal of the American Statistical Association* 44: 101–115. DOI: https://doi.org/10.1080/01621459.1949.10483294.

Chen, C., J. Wakefield, and T. Lumley. 2014. "The use of sampling weights in Bayesian hierarchical models for small area estimation." *Spatial and Spatio-Temporal Epidemiology* 11: 33–43. DOI: https://doi.org/10.1016/j.sste.2014.07.002.

Chen, S.X., C.Y. Tang, and V.T. Mule Jr. 2010. "Local post-stratification in dual system accuracy and coverage evaluation for the US census." *Journal of the American Statistical Association* 105: 105–119. DOI: https://doi.org/10.1198/jasa.2009.ap08404.

Chipperfield, J., J. Brown, and P. Bell. 2017. "Estimating the count error in the Australian census." *Journal of Official Statistics* 33: 43–59. DOI: https://doi.org/10.1515/jos-2017-0003.

Elliott, M.R. and R.J. Little. 2000. "A Bayesian approach to combining information from a census, a coverage measurement survey, and demographic analysis." *Journal of the American Statistical Association* 95 (450): 351–362. DOI: https://doi.org/10.1080/ 016 21459.2000.10474205.

Gelman, A., J. Carlin, H. Stern, D. Dunson, D. Vehtari, and A. Rubin. 2014. *Bayesian Data Analysis*. Boca Raton, FL.: CRC Press.

Gelman, A., and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical models*. Cambridge: Cambridge university press.

Gelman, A., A. Jakulin, M.G. Pittau, and Y.-S. Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics* 2: 1360–1383. DOI: https://doi.org/10.1214/08-A0AS191.

Gelman, A. and T.C. Little. 1997. "Poststratification into many categories using hierarchical logistic regression." *Survey Methodology* 23: 127–135. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf?st=76F1g34m (accessed July 2022).

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

Ghitza, Y., and A. Gelman. 2013. "Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups." *American Journal of Political Science* 57: 762–776. DOI: https://doi.org/10.1111/ajps.12004.

Ghosh, M., K. Natarajan, T. Stroud, and B.P. Carlin. 1998. "Generalized linear models for small-area estimation." *Journal of the American Statistical Association* 93: 273–282. DOI: https://doi.org/10.1080/01621459.1998.10474108.

Hogan, H.P. 1993. "The 1990 post-enumeration survey: Operations and results." *Journal of the American Statistical Association* 88: 1047–1060. DOI: https://doi.org/10.1080/01621459.1993.10476374.

Lax, J.R., and J.H. Phillips. 2009. "How should we estimate public opinion in the States?" *American Journal of Political Science* 53: 107–121. DOI: https/doi.org/10.1111/j.1540-5907.2008.00360.x.

Little, R.J. 2003. "The Bayesian approach to sample survey inference." In *Analysis of Complex Surveys*, edited by R. Chambers and C. Skinner: 49–57. John Wiley and Sons.

Lumley, T., and A. Scott. 2017. "Fitting regression models to survey data." *Statistical Science* 32: 265–278. DOI: https//doi.org/10.1214/16-STS605.

Molina, I., B. Nandram, and J. Rao. 2014. "Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach." *Annals of Applied Statistics* 8: 852–885. DOI: https/doi.org/10.1214/13-A0AS702.

Mule, T., T. Schellhamer, D. Malec, and J. Maples. 2008. "Using continuous variables as modeling covariates for net coverage estimation." In *JSM Proceedings: Section on Survey Research Methods*: 1941–1948. Denver. Available at: http://www.asasrms.org/Proceedings/y2008/Files/301279.pdf (accessed February 2022).

Nandram, B., L. Chen, and B. Manandhar. 2018. "Bayesian analysis of multinomial counts from small areas and sub-areas." In *JSM proceedings: Section on Survey Research Methods:* 1140–1162. Vancouver. Available at: http://www.asasrms.org/Proceedings/y2018/files/867100.pdf (accessed February 2022).

Paige, J., G.-A. Fuglstad, A. Riebler, and J. Wakefield. 2020. "Design-and model-based approaches to small-area estimation in a low and middle income country context: comparisons and recommendations." *Journal of Survey Statistics and Methodology.* DOI: https://doi.org/10.1093/jssam/smaa011.

Pavlou, M., G. Ambler, S. Seaman, and R.Z. Omar. 2015. "A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes." *BMC Medical Research Methodology* 15: 1–6. DOI: http://doi.org/10.1186/s12874-015-0046-6.

Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28: 40–68. DOI: https://doi.org/10.1214/12-STS395.

Pfeffermann, D., F.A.D.S. Moura, and P.L.D.N. Silva. 2006. "Multilevel modelling under informative sampling." *Biometrika* 93: 943–959. DOI: https://doi.org/10.1093/biomet/93.4.943.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (accessed February 2022).

Rabe-Hesketh, S., and A. Skrondal. 2006. "Multilevel modelling of complex survey data." *Journal of the Royal Statistical Society:* 169: 805–827. DOI: https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J. and I. Molina. 2014. *Small-area estimation.* Hoboken, NJ: John Wiley & Sons, Inc.

Rao, J., F. Verret, and M.A. Hidiroglou. 2013. "A weighted composite likelihood approach to inference for two-level models from survey data." *Survey Methodology* 39(2): 263–282. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/12-001-x2013002-eng.pdf?st=LsDJmvSV (accessed July 2022).

Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys.* Hoboken, NJ: John Wiley & Sons.

Shirley, K.E., and A. Gelman. 2015. "Hierarchical models for estimating state and demographic trends in US death penalty public opinion." *Journal of the Royal Statistical Society* 178: 1–28. DOI: https://doi.org/10. 1111/rssa.12052.

Si, Y., R. Trangucci, J.S. Gabry, and A. Gelman. 2020. "Bayesian hierarchical weighting adjustment and survey inference." *Survey Methodology* 46: 181–214. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2020002/article/00003-eng.htm. (accessed July 2022).

Skrondal, A., and S. Rabe-Hesketh. 2009. "Prediction in multilevel generalized linear models." *Journal of the Royal Statistical Society:* 172: 659–687. DOI: https://doi.org/10.1111/j.1467-985X.2009.00587.x.

Stan Development Team. 2020a. *R Stan: the R interface to Stan. R package version 2.21.2*. Available at: http://mc-stan.org/ (accessed February 2022).

Stan Development Team. 2020b. *Stan Modeling Language Users Guide and Reference Manual, version 2.25.* Available at: http://mc-stan.org/ (accessed February 2022).

Stats NZ. 2014. *Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey.* Wellington: Statistics New Zealand. Available at: https://www.stats.govt.nz/. (accessed February 2022).

Stats NZ. 2019. *Overview of statistical methods for adding admin records to the 2018 Census dataset.* Wellington, NZ: Statistics New Zealand. Available at: https://www.stats.govt.nz/ (accessed February 2022).

Stats NZ 2020a. *Estimated resident population 2018: Data sources and methods.* Wellington, NZ: Statistics New Zealand. Available at: https://www.stats.govt.nz/ (accessed February 2022).

Stats NZ. 2020b. *Post-enumeration survey 2018: Methods and Results.* Wellington, NZ: Statistics New Zealand. Available at: https://www.stats.govt.nz/ (accessed February 2022).

Vehtari, A., A. Gelman, and J. Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and Computing* 27: 1413–1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

Yi, G.Y., J. Rao, and H. Li. 2016. "A weighted composite likelihood approach for analysis of survey data under two-level models." *Statistica Sinica* 26: 569–587. DOI: https://doi.org/10.5705/ss.2013.383.

You, Y., and B. Chapman. 2006. "Small area estimation using area level models and estimated sampling variances." *Survey Methodology* 32: 97–104. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf?st=a4rH5VTf (accessed July 2022).

You, Y. and P. Dick. 2004. "Hierarchical Bayes small area inference to the 2001 census undercoverage estimation." In *JSM Proceedings: Section on Government Statistics:* 1836–1840. Available at: http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000377.pdf (accessed February 2022).

Received July 2021
Revised December 2021
Accepted March 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 793–822, http://dx.doi.org/10.2478/JOS-2022-0035

Identifying Data Quality Challenges in Online Opt-In Panels Using Cognitive Interviews in English and Spanish

Yazmín García Trejo[1], *Mikelyn Meyers*[1], *Mandi Martinez*[1], *Angela O'Brien*[1], *Patricia Goerman*[1], *and Betsarí Otero Class*[1]

In this article, we evaluate how the analysis of open-ended probes in an online cognitive interview can serve as a metric to identify cases that should be excluded due to disingenuous responses by ineligible respondents. We analyze data collected in 2019 via an online opt-in panel in English and Spanish to pretest a public opinion questionnaire (n = 265 in English and 199 in Spanish). We find that analyzing open-ended probes allowed us to flag cases completed by respondents who demonstrated problematic behaviors (e.g., answering many probes with repetitive textual patterns, by typing random characters, etc.), as well as to identify cases completed by ineligible respondents posing as eligible respondents (i.e., non-Spanish-speakers posing as Spanish-speakers). These findings indicate that data collected for multilingual pretesting research using online opt-in panels likely require additional evaluations of data quality. We find that open-ended probes can help determine which cases should be replaced when conducting pretesting using opt-in panels. We argue that open-ended probes in online cognitive interviews, while more time consuming and expensive to analyze than close-ended questions, serve as a valuable method of verifying response quality and respondent eligibility, particularly for researchers conducting multilingual surveys with online opt-in panels.

Key words: Data quality; cognitive interviews, open-ended probes; online opt-in panels.

1. Introduction

Online opt-in panels are becoming increasingly prevalent in survey research as a method for quick data collection (Chmielewski and Kucker 2019; Edgar et al. 2016; Kennedy et al. 2021). In recent years researchers have even used opt-in panels with convenience samples to pretest early versions of survey instruments in place of or in addition to traditional pretesting methods such as in-person cognitive interviews (Murphy et al. 2014; Lenzner and Neuert 2017; Neuert and Lenzner 2019). Having an easily accessible group of respondents who have signed up to fill out multiple surveys sent to them over time can be a

[1]U.S. Census Bureau, 4600 Silver Hill Road, Suitland-Silver Hill, MD 20746, U.S.A. Emails: yazmin.argen.garcia.trejo@census.gov, mikelyn.v.meyers@census.gov, mandi.martinez@census.gov, angela.c.obrien@census.gov, patricia.l.goerman@census.gov and betsari.otero.class@census.gov

Acknowledgments: The authors would like to thank both U.S. Census Bureau reviewers as well as this journal's reviewers for their time and thoughtful feedback. We also thank our collaborators at U.S. Census Bureau and our sponsors for making this research possible. Finally, we are thankful for our families, who encourage us and keep us motivated.

This article is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau. The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release [Approved for release under CBDRB-FY20-189 and CBDRB-FY19-546].

© Statistics Sweden

quick and efficient way for researchers to collect data on a whole host of topics (Matthijsse et al. 2015; Porter et al. 2019). However, questions remain about the quality and comparability of data that can be obtained from online opt-in panels in general and across languages in particular. It can be challenging to put together an online opt-in panel of members of hard-to-count populations, such as non-English speaking immigrants in the U.S., who have demographic characteristics that are representative of their population in the U.S. (Barreto et al. 2018; Ortega-Santos 2019). Moreover, researchers who purchase non-English convenience samples from online opt-in panel vendors can encounter challenges evaluating data quality, such as ensuring respondents' proficiency in non-English languages. Researchers need to be able to identify problematic answers and response patterns in order to differentiate between legitimate and non-legitimate responses across languages, which may need to be discarded from analysis.

In this article, we examine data from cognitive interview probes administered through online opt-in panels in English and Spanish to pretest survey questions from the 2020 U.S. Census Attitudes Survey (Team Y&R 2020). We were particularly interested in understanding how the inclusion of open-ended probes in online opt-in panels might help us identify data quality issues that indicate that a panel respondent was providing low quality responses. We refer to this problem as "response quality." An example of an indicator of a response quality problem is respondent behavior like copying and pasting text from the question stem rather than providing their own answer. Notably, this type of behavior would be highly unusual in a face-to-face cognitive interview or survey, and seems to be unique to the online, asynchronous format of this project.

Open-ended probes, such as those used in online cognitive interviews, require respondents to think about and write in their own response to a survey question, rather than choosing from a pre-set list of response options (Behr et al. 2017; Meitinger and Behr 2016). Such probes are common in face-to-face survey pretesting methods like cognitive interviewing, which typically involves the use of probing questions to gauge respondent understanding of survey questions and concepts (Neuert and Lenzner 2019; Willis 2005). Open-ended probes allow respondents to share what they are thinking without imposing the restrictions of categorical questions with predetermined response options (Willis 2005). Probing questions can detect cognitive issues with comprehension, retrieval, and the decision-making process when answering survey questions (Behr et al. 2017; Edgar 2013). For example, a typical probe may ask a participant to explain with their own words what a concept like "confidentiality" means to them.

In recent years, researchers have increasingly experimented with online pretesting using opt-in panels as a supplement to or a replacement for traditional in-person cognitive interviewing (Yu et al. 2019; Lenzner and Neuert 2017; Meitinger and Behr 2016; Edgar et al. 2016; Behr et al. 2012). Although scholars have studied the efficacy of many online quality checks in detecting bogus respondents using methods like evaluating open-ended responses (Buchanan and Scofield 2018; Chandler and Paolacci 2017; C. Kennedy et al. 2020; R. Kennedy et al. 2020; Kennedy et al. 2021), little is known about the role that open-ended probes in English and Spanish can play in data quality checks for online cognitive interviews. We argue that open-ended probes in online pretesting can help to detect potential response quality problems when using opt-in panels to pretest survey questions in English and Spanish. Notably, online cognitive testing has become even more

common in recent times due to safety restrictions on face-to-face interactions during the COVID-19 pandemic (Shepperd et al. 2021; Upadhyay and Lipkovich 2020). Thus, scholarship needs to identify challenges associated with data quality issues using online opt-in panels. This type of analysis could also help vendors to think of new ways to improve non-English online opt-in panel data quality as the demand for this type of panels increases. For this article, we focus on three research questions:

1. How can open-ended probes serve to identify response quality issues in online cognitive interviews using online opt-in panels in English and Spanish?
2. Are response quality issues identified in open-ended probes different (or similar) across English and Spanish responses in online cognitive interviews using a large U.S. based online opt-in panel?
3. If poor quality cases (i.e., disingenuous respondents) are identified, are open-ended probes helpful in determining which responses should be eliminated and/or replaced?

2. Literature Review

Online opt-in panels provide a relatively affordable and efficient way to pretest surveys in comparison to the funding and time needed for in person cognitive interviewing (Callegaro et al. 2014; Edgar et al. 2016). For example, researchers do not need to worry about in person scheduling logistics when using online opt-in panels for survey pretesting (Barreto et al. 2018; Edgar et al. 2016). Additionally, online pretesting provides for greater geographic dispersion and increased diversity among research participants as compared with traditional cognitive testing (Meitinger et al. 2019; Meitinger and Behr 2016), which typically takes place in fewer geographic locations due to time and resource constraints (Chandler et al. 2019; Edgar et al. 2016; Yarrish et al. 2019).

At the same time, there are drawbacks to pretesting surveys online via opt-in panels. Due to the lack of an interviewer in self-administered online pretesting, it is impossible to include spontaneous, unscripted probes that might arise based on topics introduced by respondents (Behr et al. 2012; Behr et al. 2017). Self-response data collected via online opt-in panels lacks interviewer notes, which can often include insights about respondent affect or difficulty exhibited during in-person pretesting (Edgar et al. 2016). Consequently, online pretesting results may be more difficult to interpret and may result in less detailed findings (Behr et al. 2012; Behr et al. 2017; Edgar et al. 2016). An additional drawback is the presence of bots and/or humans purposely misrepresenting themselves, which pose a threat to data quality in online research (Chmielewski and Kucker 2019; Hillygus et al. 2014). Given these challenges, it is important to understand the data quality issues that might arise when using online opt-in panels for cognitive interviews.

While research on deceptive responses to pretesting surveys is still in its infancy, research to date indicates that disingenuous respondents and bots can have a negative impact on data quality and results (Chmielewski and Kucker 2019; Kennedy et al. 2021; Simone 2019a; Yarrish et al. 2019). Kennedy et al. (2021, 3) defines bots in online surveys as "computer algorithms designed to complete online surveys automatically." Additionally, researchers have pointed out that people can misrepresent themselves in order to qualify for studies with financial incentives for participants (Chandler and Paolacci 2017; Chmielewski and Kucker 2019; Kennedy et al. 2021). This type of misrepresentation can be particularly rampant in

online surveys that seek to recruit vulnerable populations (Chandler and Paolacci 2017). For example, Chandler and Paolacci (2017) determined that their study on the LGB community had a fraud rate of 62% when no measures were taken to curtail cheating. In other words, 62% of respondents falsely claimed to qualify for the study based on their sexual orientation when researchers did not try to prevent respondent misrepresentation of their sexuality. Kennedy et al. (2021) has found that bots tend to be particularly effective with multiple-choice, close-ended survey questions, but they can be programmed to complete other types of questions as well (Kennedy et al. 2021; Simone 2019a).

There are several ways to spot misrepresentation (both electronic and human) within a survey (Kennedy et al. 2021; Prince et al. 2012; Yarrish et al. 2019). First, researchers should look to see if the prevalence of a required specific characteristic is realistic within a sample (Chandler and Paolacci, 2017). For example, Ortega-Santos (2019) finds that although M Turk provides a large pool of Spanish-speakers, there are limitations on the diversity of bilingual and non-bilingual respondents. Ortega-Santos (2019) notes that in her M Turk study, most bilingual Spanish respondents resided in the United States, while the majority of monolingual Spanish-speaking respondents came from countries outside of the United States. Another potential problem with online surveys is the presence of bogus responses. Kennedy et al. (2021, 11–12) notes that when asked open-ended questions such as "When you were growing up, what was the big city nearest where you lived?", many participants responded with text appropriate for reviewing a product, but not answering that question, such as "ALL SERVICES SOUND VERY GOOD" or responding with the name of a state or other non sequitur responses. Moreover, some of the responses were word-for-word identical, such as a multitude of respondents who provided the identical response that their favorite thing to do in a new city is to, "Explore by Foot. VirtualTourist members agree that the best way to see a new destination is to experience it by foot" (C. Kennedy et al. 2020, 67).

This finding raised questions about the integrity of the participants and their subsequent responses to other questions in the survey. Another potential indicator of poor-quality responses is quick completion time. Although quick completion can be an indication of a satisficer, or a respondent who chooses the easiest possible response rather than putting much thought into their answer, it may also be an indication of a bot, since bots often complete surveys with inhuman quickness (C. Kennedy et al. 2020; Kennedy et al. 2021).

There are several steps researchers can take to prevent poor quality in online opt-in panel data. Simone (2019b) recommends including at least two or three open-ended questions as a quality check. Although electronic bots are becoming increasingly adept, open-ended questions remain a reliable way of detecting and flagging unusual data. Kennedy et al. (2021) recommends adding homonyms, or words that have identical spellings and/or sounds, but different definitions, to open-ended questions to make problematic answers more detectable. For example, Kennedy et al. (2021) asked respondents, "What would you like to see elected leaders in Washington get done during the next few years? Please give as much detail as you can." Although some of her respondents responded appropriately with political issues, many respondents replied by detailing the life of George Washington or talking about Washington state. Particularly suspicious answers featured reordered text clearly taken from somewhere else on the Internet.

Nearly all research on detecting data quality issues in online opt-in panels has focused on English-speakers. To our knowledge, the only research focusing on detecting data quality issues with Spanish-speakers comes from studies where researchers were trying to

crowdsource translations. Crowdsourcing is a method by which researchers acquire services by asking for contributions from a large number of people, often from an online community (Chandler et al. 2019; Chmielewski and Kucker 2019; Ortega-Santos 2019). Some crowdsourced language studies have detected professional respondents in all languages, including Spanish (Aranberri et al. 2017; Ortega-Santos 2019). Our article contributes to a gap in the literature relating to Spanish-language responses to online opt-in panels as compared to English-speaking respondents. We hope that this article will start a methodological and ethics conversation on data quality issues for English- and Spanish-language online opt-in panels, particularly for those that aim to include vulnerable populations. Advancing this research topic will benefit both researchers and vendors as both need to understand the challenges in data quality that their research efforts may encounter.

3. Data

The data we discuss here were collected via an online opt-in panel in both English and Spanish. These data were collected from a convenience sample as part of a round of online cognitive interviews to pretest questions for the 2020 Census Attitudes Survey, a public opinion survey about people's attitudes and planned participation in the 2020 Census. The survey was also pretested via two rounds of traditional in-person cognitive interviews in English and Spanish prior to the online testing.

The online pretesting occurred in July of 2019 and focused on unresolved issues with the question wording from the previous in-person testing. In the survey introduction we explained that participants needed to first answer a survey questionnaire and then respond to follow-up items about the questions previously asked. Eighteen survey questions and thirteen probes were included in a forced-choice design (see Appendix, Subsection 8.1), and there were 463 completed cases (265 in English and 198 in Spanish). These numbers included unedited cases with signs of problematic behavior; for example, "speeders" (i.e., people who answered the survey below the 25th percentile respondent completion time, which is less than four minutes in both languages) and respondents in the Spanish-language sample who did not self-identify as Hispanic or Latino. While the online survey platform we used typically delivers cleaned data (e.g., with "speeders" deleted), we had a unique opportunity to examine the initial un-scrubbed data because our agency has regulations that only permit individuals with U.S. Census Bureau Special Sworn Status to review unedited responses in order to protect respondents' confidentiality. Thus, the typical vendor data scrub procedure was not used before we received the data. Rather, we received unedited responses and had the opportunity to review them ourselves.

For the pretesting study, respondents first answered a series of survey questions. Next, they answered a series of retrospective probes about a subset of the survey questions (see Appendix, Subsection 8.2). This retrospective probing approach was consistent with the prior in-person cognitive interview approach, with minor adaptations to the probe wording for an online audience. Half of the sample in both languages was randomly assigned to receive one of two alternate versions of two of the survey questions.

The survey platform we used typically includes a language toggle that allows respondents to switch between Spanish and English at any point while completing a survey. We asked the vendor to disable this toggle for our study to better separate our sample of English-speakers

from Spanish-speakers. We wanted to discourage ineligible English-speakers from posing as Spanish-speakers and filling out the survey more than once. In addition, a separate link to access the survey in each language was created to distribute invitations to Spanish- and English-speaking respondents directly in the appropriate language.

When designing respondent samples, researchers typically set goals or quotas for how many participants they would like to include with specific characteristics. In face-to-face cognitive interviews, it is common to seek a small, diverse sample of respondents with demographic characteristics of interest for a given study. In most online opt-in panels, researchers can also set quotas for a specific number of completed cases of any given type. For instance, researchers can request that half the sample be comprised of men and half of women, or that a certain percentage of respondents have an advanced educational degree.

We wanted to include Spanish-speaking respondents in this study who reflected the variation in educational level, sex, birthplace, and region of the general population of Hispanics in the U.S. Monolingual Spanish-speaking respondents in the U.S. are more likely to have completed high school or less than the general population. Therefore, we requested overrepresentation of Spanish-dominant speakers who had completed high school or less (see Table 1). Of note, for this study, it took longer to fill the quotas for Spanish-dominant speakers than it did for English-speakers. This may be due to the challenges that researchers face in recruiting vulnerable or hard-to-reach populations across survey modes.

Table 2 shows the demographic characteristics of the respondents who completed the survey, which in some cases, differed from the quotas that we had requested in Table 1. For example, we ended up with a larger number of English-speakers (265) than we had originally requested (250). The Spanish-speaking sample fell short of the originally requested 250 responses, with only 198 completed cases. Educational attainment was another variable for which quotas were not fulfilled in both languages. Spanish-speakers with less than high school education were difficult to find among the online opt-in panel respondents, and they only comprised about 20% of the respondents. Spanish-speaking women were over-represented compared with men. We did not impose quotas by race and ethnicity, so any respondent in the vendor's contact frame who was identified as a Spanish-speaker was eligible to answer the survey. One unexpected finding, shown in Table 2, was that about 48% of the sample of Spanish-speaking participants did not identify as Hispanic.

Table 1. Sample quotas requested for the study.

Variable	English	Spanish
Sex	Men (120) Women (130)	Men (120) Women (130)
Education	High school education or less (100) More than High School (150)	High school education or less (150) More than High School (100)
Birthplace	U.S. Born (250)	U.S. Born (150) Foreign born (100)
Region	Census regions (125 from the West and 125 from Northeast and South)	Census regions (150 from the West and 100 from Northeast and South)

Source: 2020 U.S. Census Attitudes Survey.

Table 2. Characteristics of the participants in the pretesting project.

		English	Spanish
Sex	Male	49.8%	36.2%
	Female	50.2%	63.8%
Ethnicity	Hispanic	18.1%	52.3%
	Non-Hispanic	81.9%	47.7%
Race	White	56.3%	58.5%
	Black or African American	27.1%	10.6%
	American Indian or Alaska Native	3.2%	3.4%
	Asian	2.8%	6.8%
	Native Hawaiian or Pacific Islander	1.8%	2.4%
	Some other race	8.8%	18.4%
Age	18-30	54.7%	44.2%
	31-45	22.3%	34.2%
	46-55	9.8%	11.6%
	56-65	6.0%	6.5%
	Over 65	7.2%	3.5%
Education	Less than High School	40%	19.6%
	Completed High School	29.8%	17.6%
	Some College, no degree	9.4%	15.6%
	Associate's degree (AA/AS)	6.0%	11.6%
	Bachelor's degree (BA/BS)	9.1%	27.6%
	Post Bachelor's degree	5.7%	8.0%
Total N		265	198

Source 2020 U.S. Census Attitudes Survey.

We requested some information from the vendor such as the number of invitations sent, language recruitment measurements used and fraud control mechanisms the vendor uses for their online opt-in panel participants. Disclosure of this type of information may vary from vendor to vendor for online opt-in panels. There were 326,272 invitations sent for this study, and a total of 8,439 respondents accessed the survey but did not complete it. Given our 463 completed cases, the completion rate for this study was 0.14%, with an additional 2.6% of the sample providing incomplete responses. Since we have not sought to compare data from other vendors, we do not know how common these extremely low response rates are for online opt-in panels, and we are unaware of any scholarly publications that compare this type of information or any other type of methodological information across online opt-in panels. The design of guidelines for participation rates and data quality in online opt-in panels is an area ripe for future research.

At our request, the vendor shared that they used the following question about language spoken at home for the recruitment of Spanish-speaking participants:

What best describes the combination of English and Spanish that you speak within your home?

1. Spanish only
2. Spanish more than English
3. Spanish/English equally,
4. English more than Spanish
5. English only

6. None of the above
7. Prefer not to answer.

We requested that the vendor included only people who responded "Spanish only" or "Spanish more than English" for our study. The vendor further told us that all demographic information collected during the recruitment of respondents for the opt-in panel was self-reported and no additional language or reading proficiency items were used. The vendor also shared that they typically use the following mechanisms to prevent fraud:

- Diverse recruitment partnerships with blue chip companies in retail, travel, entertainment, and business,
- Proprietary members who are known and trusted,
- Consistent source blending that is constantly monitored,
- Double opt-in "member join process" measuring thought consistency, unlikely correlations, low-probability answers, speeding and more,
- Close panelist monitoring for behavioral anomalies (e.g., abrupt location changes or abnormal qualification rates), and
- Phone/SMS validation for reward redemption 3rd party data validation, network of checks for browser, system and BIOS configurations.

The Spanish translation of the survey instrument was produced by certified translators with extensive experience translating survey materials in the U.S. Census Bureau's Decennial Translation Branch. The translations were also expert reviewed to verify that the phrasing of probes was built on findings from prior multilingual pretesting research using a process outlined in [Citation removed].

4. Methods

Through our analysis we sought to provide answers to the following hypotheses:

- *Hypotheses 1*(H1): Open-ended probes can serve to identify response quality issues in multilingual online cognitive interviews using online opt-in panels in English and Spanish,
- *Hypotheses 2* (H2): Open-ended probes in Spanish online opt-in panel pretesting will have more response quality issues compared to respondents answering in the language predominant for online opt-in panels (i.e., English in the U.S.), and
- *Hypotheses 3* (H3): Open-ended probes can help identify and remove disingenuous responses from online cognitive interviews when used in tandem with other data quality measures (e.g., removing "speeders").

To test whether open-ended probes can uncover data quality issues in online cognitive interviews, a team of researchers developed a coding scheme based on recommendations from the 2019 Washington Statistical Society annual address (Kennedy et al. 2021) to document seven types of problematic responses (i.e., responses provided by respondents who were not making a good faith effort to fill out the survey). We did not include valid survey responses in our analysis. Table 3 provides a definition for each of these types of problematic responses. The study included three different types of open-ended probes: (1) probes that asked about the meaning of concepts or words, (2) probes about a specific part

Table 3. Coding categories for valid and problematic responses.

Coding categories	Definition
Valid response.	Valid answers/No issues found
Problematic responses	Responses that do not answer the question adequately
1. Incomplete responses	Respondent uses language related to the topic of the question but provides an incomplete answer (e.g., "About United States")
2. Non sequitur	The answer has nothing to do with the topic of the question (e.g.,"Pink, "yes" or "no" as the response to a question asking for a definition or a preference)
3. Copied survey text	Text copied from the question stem
4. Random typing	Set of letters typed at random (e.g., "asdf;asdf;ljasf")
5. Profanity	Respondent uses inappropriate or offensive language (e.g., "stop asking the same fxxxxxx questions")
6. Volunteers that respondent is not a Spanish-speaker	Respondent writes that they do not know or speak Spanish (e.g., "I don't speak Spanish)
7. Repetitive textual pattern	Uses the same text across three or more responses (e.g., "no idea," "DK," "Yes," etc.)

of the question stem, and (3) probes about participants' question wording preference. In total, we analyzed data from thirteen open-ended probes, each of which respondents had to answer in order to advance to the next screen.

The codes enable us to measure the extent of response quality problems (to address H1), and to compare their prevalence across languages and probe types (to address H2). For our final hypothesis (H3), we conduct an exploratory analysis on flagging cases that might need to be removed due to data quality issues. In particular, we focus on how response quality issues identified in open-ended probes can be used in combination with other data quality measures, such as identifying "speeders," to make decisions about which cases should be removed.

As mentioned previously, when conducting in-person cognitive interviews, researchers rarely encounter these types of data quality problems. Typical data quality issues encountered during cognitive interviewing might include discovering a respondent does not have the expected demographic characteristics or interviewing a respondent who provides minimal responses to probing questions. However, response quality issues comparable to a respondent providing "random answers" to complete the survey more quickly and receive an incentive would not typically occur in face-to-face cognitive interviews. Moreover, for in-person cognitive interviews, which are considered a qualitative method, researchers do not typically need to interview a large number of respondents in order to obtain useful findings. Thus, even if researchers need to discard a relatively large number of the online cases when using an opt-in panel, the results could still be useful for pretesting purposes. This consideration distinguishes online pretesting efforts from both in person cognitive testing and regular production surveys.

Qualitative analysis of the open-ended responses was done using NVivo. Researchers first collaborated on creation of the coding scheme. Next, two bilingual researchers independently coded a portion of the open-ended responses in English and Spanish. After they had coded

four of the probes, we met as a team to come to a consensus on how to code the more complex cases. We conducted an evaluation of intercoder reliability based on a third of the probes coded, which resulted in a 92% intercoder reliability rate for Spanish and 94% for English. Researchers determined that the number of potential codes and the frequency that multiple codes were assigned to the same response made the probability of agreement relatively low. As a result, this intercoder reliability evaluation was completed using percent agreement. As a last data quality step, the coding was reviewed by a senior researcher. Once all coding was completed, we produced frequency tables of the seven indicators of poor-quality. Given that multiple codes could be applied to each response, the denominator used to calculate percentages corresponds to the number of references (or codes) for a given question, rather than the total number of respondents who answered the question.

5. Findings (Hypotheses 1 and Hypotheses 2)

5.1. *H1: Open-ended Probes Can Serve to Identify Response Quality Issues in Multilingual Online Cognitive Interviews Using Online Opt-in Panels in English and Spanish*

After coding the responses to the thirteen open-ended probes, we assessed our first hypothesis by evaluating how responses to the open-ended probes could serve as a metric to identify response quality issues in the pretesting. Table 4 summarizes the number of responses coded as valid (appropriate/helpful) responses and problematic responses for both English and Spanish. Results are presented by type of open-ended probe and by language. The rows add up to a hundred percent in each language because the denominator is the number of references coded for each row, showing the prevalence of the data quality issue.

Table 4. Prevalence of valid responses and problematic responses by probe and by language.

Probe types	Response Types (English)			Response Types (Spanish)		
	Valid response	Problematic responses	Total	Valid response	Problematic responses	Total
Probes about meaning of concepts or words	%	%		%	%	
Meaning of "questions on the census form"	49	51	267	32	68	184
Meaning of "confidential"	71	29	268	30	70	188
Meaning of "The Census Bureau will not keep answers to the 2020 Census confidential"	51	49	267	25	75	183
Meaning of "share individuals answers with other government agencies"	48	52	268	19	81	193
Meaning of the phrase "used against you"	55	45	266	27	73	185
Meaning of the "2020 Census" (Version 2 asked to half of the sample)	62	38	133	31	69	93
Conditional probe about the content of the question						
Concerns respondents had in mind when asked, "How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?"	49	51	267	25	75	184
Concerns respondents had in mind when asked, "How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census?"	52	48	267	29	71	184
Government agencies respondent had in mind	58	42	267	25	75	185
Probes related to preference of the wording of a question						
Comparing two versions of the intention to participate in the census questions. Version A vs. Version B1 (Asked to half of the sample)	69	31	67	30	70	47
Comparing two versions of the intention to participate in the census questions. Version A vs. Version B2 (Asked to half of the sample)	52	48	73	40	60	43
Reasons for preference of Census awareness questions (Form 1)	69	31	77	43	57	46
Reasons for preference of Census awareness questions (Form 2)	74	26	76	31	69	55
Average across all questions	**58**	**42**		**30**	**70**	

Note: "Valid responses" refers to all the responses that did not present indications of response quality issues as measured in this paper. The columns labeled as "Problematic Responses" aggregated all quality issues defined in Table 3. The average across all questions was calculated adding the percentages reported in each row and divided by the number of rows. Source: 2020 U.S. Census Attitudes Survey.

Interestingly, although responses in both languages were coded as problematic, the Spanish-language responses had a higher incidence of problematic responses across all probe types. Table 4 shows that valid English-language responses range from 48 to 74% depending on the type of probe asked, while valid responses occurred far less often in Spanish, ranging from 19 to 43%. For example, when respondents were asked about the meaning of the term "confidential," about 71% of responses were coded as valid in English compared to only 30% of the Spanish-language responses.

5.2. *H2: Open-ended Probes in Spanish Online Opt-in Panel Pretesting Will Have More Response Quality Issues Compared to Respondents Answering in the Predominant Language for Opt-in Panels (i.e., English in the U.S.)*

For our second hypothesis, we focus only on the problematic responses shown in Table 4 above. We expected more response quality issues with Spanish-language responses than English-language responses based on past findings from in-person cognitive interview studies. For example, research using face-to-face pretesting among Spanish-dominant Hispanics indicates that internet access does not necessarily correlate to the type of digital affinity necessary to complete online surveys (e.g., experience typing in URLs, familiarity with establishing accounts with a username and password, etc.) (Brown et al. 2016; Trejo and Schoua-Glusberg 2017). Moreover, research on survey pretesting also shows reliability measurement challenges for the Spanish-speaking population in the U.S. For instance, educational attainment is a well-established item used in many surveys in English. However, when pretested in Spanish, researchers have found that Spanish-speakers can interpret educational attainment response options differentially based on conventions for naming levels of education in their country of origin, which implies that further improvements to the translation of certain questions may be necessary to ensure reliability (Goerman et al. 2018).

Table 5 shows the frequency of six response quality issues for each probe type in English and seven response quality issues for Spanish. The code for respondents who volunteered that they were not Spanish-speakers was only applicable to Spanish-language cases, which results in one additional type of data quality issue possible in Spanish-language cases. One of the most salient themes that emerged was the prevalence of repetitive textual patterns. In our initial coding, we focused on coding one probe at a time, with each probe in its own column. As we worked, we noticed that contiguous responses by the same respondent across columns or different questions sometimes contained identical responses, indicating that a respondent was answering multiple open-ended probes the same way. Once we noticed this potential copying and pasting behavior, we began to code this as a repetitive textual pattern. This problem occurred on average in 42% of English-language responses, but the frequency of repetitive textual patterns on average was higher in Spanish (54%). These repetitive textual patterns have important implications for response quality and may be the result of respondents attempting to complete the survey as quickly as possible, or they may indicate the presence of a bot.

The second common code in both languages and across all types of probes was non sequiturs (26% on average in both English and Spanish), which is similar to what other researchers have discussed as a quality issue in the open-ended question literature (Kennedy et al. 2021). For instance, researchers have hypothesized that respondents may

Table 5. Prevalence of problematic responses and probe and by language.

Probe Types	Type of Problematic Responses																
	Incomplete responses	Incomplete responses	Non sequitur	Non sequitur	Copied survey text	Copied survey text	Random typing	Random typing	Profanity	Profanity	Volunteers that Respondent is not English speaker	Volunteers that Respondent is not Spanish speaker	Repetitive Textual Pattern	Repetitive Textual Pattern	Total	Total
Probes' language (EN= English, SP= Spanish)	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP
Probes about meaning of concepts or words	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	
Meaning of "questions on the census form"	4	5	30	20	10	0	13	6	1	1	0	3	41	66	135	126
Meaning of "confidential"	1	17	17	17	0	0	27	8	4	2	0	4	51	54	78	132
Meaning of "The Census Bureau will not keep answers to the 2020 Census confidential"	19	1	22	30	0	0	17	7	2	2	0	5	39	54	132	138
Meaning of "share individuals' answers with other government agencies"	17	9	26	20	0	10	17	7	3	1	0	4	38	48	139	157
Meaning of the phrase "used against you"	5	10	24	19	4	1	26	6	2	1	0	1	39	61	120	135
Meaning of the "2020 Census" Version 2 (asked to half of the sample)	16	0	16	27	0	9	10	6	0	2	0	8	58	48	50	64
Conditional probe on the content of the question																
Concerns respondents had in mind when asked "How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?"	6	1	33	28	1	0	19	7	1	1	0	3	40	59	136	138
Concerns respondents had in mind when asked "How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census?"	3	5	30	26	1	0	20	9	2	2	0	2	45	57	128	130
Government agencies respondent had in mind	5	2	21	32	0	1	24	7	2	1	0	3	47	53	112	139
Probes related to preference of the wording of a question																
Comparing two versions of the intention to participate in the census questions. Version A vs. Version B1 (Asked to half of the sample)	5	12	33	18	0	0	29	6	5	9	0	0	29	55	21	33
Comparing two versions of the intention to participate in the census questions. Version A vs. Version B2 (Asked to half of the sample)	3	4	37	31	0	0	29	15	0	0	0	4	31	46	35	26
Reasons for preference of Census awareness questions (Form 1)	4	4	17	42	0	0	33	4	4	0	0	0	42	50	24	26
Reasons for preference of Census awareness questions (Form 2)	5	11	30	21	0	0	10	8	5	3	0	3	50	55	20	38
Average across all questions	7	6	26	26	1	2	21	7	2	2	0	3	42	54		

Note: The rows add up to one hundred percent in each language. The total for each language in each row shows the number of problematic references coded for each probe (the sum of all seven quality issues defined in Table 3). The average across all questions was calculated adding the percentages reported in each row and divided by the number of rows. Source: 2020 U.S. Census Attitudes Survey.

be tired of answering too many open-ended probes, or they may type nonsensical responses in order to advance to the next question more quickly and complete the survey sooner (Neuert and Lenzner, 2019).

While many findings were common across languages, some codes were more commonly applied to responses in one language than another. Random typing, where respondents entered gibberish such as "adjkipsdjklfuid," was on average more common in English (21%) than in Spanish (7%) across all probes. One behavior only seen in the Spanish-language data was respondents volunteering that they were not Spanish-speakers (3% on average). While data quality issues like non sequiturs and random typing can be attributed to different causes and may not always be indicative of a deceptive respondent (e.g., respondents may have made a good faith effort to answer most of the survey but may demonstrate fatigue as the survey goes on), this is not the case for respondents who volunteer that they are not Spanish-speakers when filling out a Spanish-language survey. This behavior is much more clearly indicative of a deceptive respondent. In both languages, there were relatively few instances of text copied directly from the survey, with two percent on average or fewer of responses coded as "copied text" in both languages. Use of profanity was also relatively rarely observed (2% on average across languages). When this behavior occurred, it was a red flag in terms of respondents not making a good faith effort to fill out the survey.

To our knowledge, there are no widely accepted standards in the survey industry for ensuring high quality data when including Spanish-dominant Spanish-speakers in online opt-in panels based in the U.S. However, as our experience shows, the onus is often on researchers to request the inclusion of project-specific language screening questions in order to ensure that the correct respondents are included in the final sample. As we found in this study, even when Spanish proficiency is a requirement for the study, the sample of respondents may erroneously include respondents who are not fluent in the language in question.

5.3. *Findings: Further Exploratory Analysis for Hypothesis 3 (H3): Open-Ended Probes Can Help Identify and Remove Disingenuous Responses from Online Cognitive Interviews when used in Tandem with other Data Quality Measures (i.e., Removing "Speeders")*

The occurrence of repetitive textual patterns in open-ended probes, as shown in the previous section in Table 5, allowed us to identify cases that might need to be eliminated from or replaced in the analysis. As an exploratory analysis for this article, we decided to compare these findings on problematic open-ended responses with a data quality metric used to evaluate close-ended responses: respondents who showed signs of being "speeders." Speeding is one of the behaviors commonly examined when evaluating the data quality of survey responses (Kennedy et al. 2021; Storozuk et al. 2020).

We defined "speeders" for the purposes of this exploratory research as participants whose time to complete the main survey and the probing questions fell on or below the 25th percentile (227 seconds for English and 215.5 seconds for Spanish). Respondents whose response time was above the 75th percentile (557 seconds for English and 566 seconds for Spanish) were considered to be respondents who took a substantial amount of time to respond to the online cognitive interview. There are two reasons we selected the 25th percentile to define speeders in our exploratory analysis. First, in analysis not shown in this article, we calculated the number of data quality issues on the 10th fastest percentile based on time of completion of the online cognitive interview. When we did this calculation, we found that there were no differences in the average number of data quality problems when comparing the 10th and 25th percentiles. For example, for the cases in Spanish, the 10th percentile had an average of ten data quality issues and the 25th percentile had 9.8 problems on average. Our second reason for defining speeders using the 25th percentile is that the 25th percentile has more than twice the number of cases, which improves comparability with other quartiles and the sample overall. In particular, the decile size is about 20 cases but the quartile size is about 50 cases. Once we identified "speeders," we calculated the average number of problematic codes for respondents in each quartile. The results for the average number of problems by topic and time to answer are shown in Table 6. This table shows that "speeders" in the survey also accounted for most, but not all, data quality problems identified in the analysis of the open-ended questions in English and Spanish. These results signal that the analysis of open-ended items may identify additional participants displaying problematic behaviors that are missed when only using more traditional methods for verifying data quality like identifying "speeders."

In an effort to better understand the data quality issues found in the Spanish-language sample, we completed an analysis not shown in this article in which we evaluated how different sociodemographic groups varied in their responses to open-ended questions. When analyzing gender and age, we did not find differences by age across response quality issues. However, the results by gender show that women had fewer response quality issues.

6. Limitations

Some limitations should be kept in mind regarding our study results. Online opt-in panel vendors typically use data cleaning procedures to weed out poor quality responses prior to sending data to clients. As mentioned before, due to our agency's data security

Table 6. Average Number of Problems found by quartile of the duration of the stuvey.

Quartile of the duration of the survey	Average number of problems	Type of Problematic Responses: Incomplete responses	Non sequitur	Copied survey text	Random typing	Profanity	Volunteers that respondent is not a Spanish sneaker	Repetitive textual pattern
English								
Q1 (0-25)	8.2	0.2	1.5	0.1	2.6	0.2	0.0	3.7
Q2 (25-50)	3.9	0.5	1.1	0.1	0.4	0.1	0.0	1.9
Q3 (50-75)	2.5	0.3	0.9	0.1	0.4	0.0	0.0	0.8
Q4 (75-100)	2.4	0.4	0.9	0.1	0.1	0.1	0.0	0.9
Spanish								
Q1 (0-25)	9.8	0.1	1.5	0.1	1.3	0.2	0.4	6.2
Q2 (25-50)	8.3	0.6	2.1	0.1	0.3	0.0	0.4	4.7
Q3 (50-75)	4.5	0.4	1.5	0.1	0.2	0.0	0.0	2.2
Q4 (75-100)	3.3	0.4	1.2	0.2	0.0	0.1	0.0	1.2

Note: Q = Quartile.
Source: 2020 U.S.Census Attitudes Survey.

requirements, the vendor's customary data cleaning procedures could not be implemented, and as such, the raw data that we started with was likely of poorer quality than what consumers might typically receive from online opt-in panels. However, we suspect that the ability of many online opt-in panel vendors in the U.S. to provide in depth data cleaning in non-English languages is probably somewhat limited. Some of the problems with the Spanish-language online opt-in panel may have persisted even if the typical data cleaning procedures had been followed prior to our receiving the dataset. For example, Kennedy et al. (2021) found similar data quality concerns to those we identified when analyzing responses from Hispanic respondents in edited datasets. They also found that bots often select the first response option of close-ended questions, which can lead to an overrepresentation of deceptive respondents reporting themselves as Hispanic, since this option appears first in most ethnicity questions (Kennedy et al. 2021). In addition, some language or cultural groups may be less likely to be "looped in" to participating in this type of online opt-in panel than others. This type of finding can help online opt-in panel vendors to refine their recruitment and data cleaning methods, particularly for non-English-speaking, U.S. based respondents.

We encourage online opt-in panel vendors and researchers to revise the guidelines on the recruitment of vulnerable populations (University of Maryland 2014), in particular for non-English-speaking populations, to encourage greater participation by these types of respondents. We recommend that the recruitment of participants who speak languages other than English be a two-part process, with quality control implemented not just by sample vendors but also by researchers who evaluate data quality to ensure that respondents meet eligibility criteria and provide useful responses.

Another limitation of this study is the increased respondent burden when answering open-ended probes as compared to closed-ended items. It seems likely that open-ended probes will impose some burden since they deal with comprehension issues and require in-depth thinking, and this burden may be compounded for participants who are responding to these questions on their smartphones. This leads to a potential dilemma: the burden of answering probing questions may motivate some of the respondent "shortcuts" that we aim to identify in this manuscript. This dilemma is an important limitation to consider for future analyses and may be more relevant in pretesting efforts that include many open-ended probes than in production surveys that contain fewer open-ended questions.

A further limitation of our study is the lack of quotas imposed for race and Hispanic ethnicity. We focused our analysis on participants who reported being Spanish-dominant Spanish-speakers, but we did not screen out respondents who selected that they did not self-identify as Hispanic. Interestingly, we found that 48% of the respondents who self-identified as Spanish-dominant speakers did not identify as Hispanics. This discrepancy may be an artifact of participants interpreting the recruitment question about language spoken at home in different ways, or it may be related to the issue of participants who posed as Spanish-speakers to receive the incentive but reported not speaking Spanish in open-ended questions. In their article about best practices for collecting data on racially diverse online panels, Barreto et al. (2018) stress the importance of including quotas relevant to race and ethnic groups that help researchers better represent the groups under study. These recommendations should be considered in future designs of quotas for online cognitive interviews using opt-in panels.

A final limitation we would like to acknowledge is that the survey that we administered was programmed with a "forced choice" design, in which respondents were required to enter a response to the open-ended questions before they could continue with the survey. This practice is often implemented to determine participant eligibility for subsequent questions, or to ensure that researchers have enough data to analyze for each question. However, a drawback to forced choice question designs is that some participants may be forced to answer a question they do not know the answer to, or to choose between inadequate response choices in order to complete the survey and get their incentive. Therefore, some respondents may have been forced to answer a question they would have left blank under other circumstances.

7. Summary and Discussion

This article provides evidence that open-ended probes used in online cognitive interviews can root out response quality issues in the emerging domain of online opt-in panels. While respondents taking part in live cognitive interviews may provide low quality responses by engaging in behaviors like giving vague or brief responses, in our experience, most respondents would be reluctant to completely flout conversational norms in a live interview (e.g., by stringing together nonsense syllables, or only repeating the question verbatim instead of providing a response to the probe). Thus, a unique artifact of conducting online cognitive interviews is that the lack of a live interviewer may make respondents feel more comfortable engaging in behaviors that result in low quality responses such as typing gibberish or copying and pasting the question text into the response field.

While this research highlights the benefits of using open-ended probes as a data quality validation for cognitive interviews conducted with online opt-in panels in English and Spanish, our findings also signal that analysis of open-ended items is an important data quality check for multilingual surveys beyond pretesting efforts. We argue that research about data quality and best practices for multilingual online opt-in panels should be carried out in earnest as online opt-in panels in non-English languages become more common. Online opt-in panels are an innovative medium that have great promise for survey research in general and for the inclusion of diverse multilingual and multicultural populations as well. Moreover, due to the COVID-19 pandemic, remote data collection methods are becoming increasingly common as researchers around the world adapt face-to-face data collection to other modes appropriate to the current context.

An interesting issue raised by this analysis is the quality of non-English opt-in panels in countries such as the United States. The analysis of Spanish-language open-ended probe responses uncovered response quality issues, including the use of repetitive textual patterns and participants in the Spanish-language survey who explicitly said that they were not Spanish-speakers. We also found that open-ended probes can help determine which cases should be replaced when conducting a survey using an online opt-in panel. These findings aligned with the exploratory analysis we conducted on respondents who were speeding, since speeders often provided low-quality open-ended responses.

Based on our findings, there are several areas that we identify here for future research. It would be valuable to compare responses to open-ended probes across other online opt-in panels to evaluate the type and frequency of problematic behaviors when vendors use

different approaches to recruitment. Further research could expand on our exploratory analysis of respondents who provide low quality data for open-ended items, and how they correspond (or not) to respondents identified as providing problematic responses using traditional measures for detecting disingenuous responses (e.g., speeders, straightlining, etc.). As researchers adapt to the COVID-19 pandemic, conducting online cognitive interviews may become increasingly common. While cognitive interviews are traditionally a qualitative method, the usage of online opt-in panels could result in larger sample sizes with potentially low response rates such as those that we encountered in this study. As researchers tailor cognitive interviews to an online format with larger samples, the study of non-response bias in online pretesting studies may become a necessary topic of discussion.

Additionally, further research to compare data from different vendors or in other languages beyond English and Spanish is needed. For example, experiments comparing online opt-in panels could help to evaluate the variables needed to recruit monolingual, bilingual and multilingual participants in Spanish and other languages, as well as the incidence of poor data quality. Another potential area of research could investigate whether some participants would feel more comfortable recording audio of their responses in non-English languages rather than typing their responses, especially if they are fluent speakers of the language but rarely use it to respond in writing to surveys.

While some of our findings regarding data quality for U.S. Spanish-language cases are troubling, we believe that efforts to improve the representation of non-English-speakers in survey research using emerging methods like online opt-in panels are important given both the increasing linguistic diversity of the U.S. as well as declining survey response rates (De Heer and De Leeuw 2002; Hanson and Espinosa 2016; Luiten et al. 2020; Rumbaut and Massey 2013). We conclude that the inclusion of open-ended probes and survey questions, while more time-consuming and expensive to analyze, is an important method of verifying data quality, particularly in multilingual research. Much research is still to be conducted on this timely topic.

8. Appendix

8.1. Wording of questions and probes

Survey question	Retrospective probe	Sample
Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?	The first question you were asked was, "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" This question mentions the "2020 Census." Based on your understanding, what is the "2020 Census"?	All

Continued

Survey question	Retrospective probe	Sample
Now we're going to show you an alternative version of the question. Version 1 (question you answered): "The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" Version 2 (Alternative): "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much haveyou seen or heard recently – within the last week or so – about the 2020 Census?	Can you say more about why you prefer that version?	Half
Now we're going to show you an alternative version of the question. Version 1 (question you answered): "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" Version 2 (Alternative): "The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?"	Can you say more about why you prefer that version?	All
How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form. • Definitely will (1) • Probably will (2) • Might or might not (3) • Probably will not (4) • Definitely will not (5)	Another survey question was, "How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form." What comes to mind when you see the phrase "questions on the census form" in this question?	All
Now we're going to show you an alternative version of the question. Version 1 (question you answered):	Can you say more about why you prefer that version?	All

Continued

Survey question	Retrospective probe	Sample
"How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form." Version 2 (alternative): "How likely do you think it is that you or someone in your household will answer the questions on the census form?"		
Now we're going to show you an alternative version of the question. Version 1 (question you answered): "How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form." Version 2 (alternative): "In 2020, the Census Bureau will mail nearly every household in the U.S. a questionnaire and an invitation to participate online. How likely do you think it is that you or someone in your household will mail back the questionnaire or submit it online?"	Can you say more about why you prefer that version?	Half
How concerned are you, if at all, that the Census Bureau will not keep answers to the 2020 Census confidential? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3) • Not too concerned (4) • Not at all concerned (5)	What does the word "confidential" mean to you in this question?	Half
How concerned are you, if at all, that the Census Bureau will not keep answers to the 2020 Census confidential? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3) • Not too concerned (4) • Not at all concerned (5)	In your own words, what does the phrase "the Census Bureau will not keep answers to the 2020 Census confidential" mean to you in this question?	All
How concerned are you, if at all, that the Census Bureau will share individuals' answers to the 2020 Census with other government agencies? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3)	What government agencies were you thinking of when you were answering this question?	All

Continued

Survey question	Retrospective probe	Sample
• Not too concerned (4) • Not at all concerned (5)		
How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3) • Not too concerned (4) • Not at all concerned (5)	Another survey question was, "How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?" What concerns came to mind when you answered the question that appears above, if any	All
How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3) • Not too concerned (4) • Not at all concerned (5)	What does the phrase "used against you" mean to you in this question?	All
How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census? • Extremely concerned (1) • Very concerned (2) • Somewhat concerned (3) • Not too concerned (4) • Not at all concerned (5)	The next survey question was, "How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census?" What concerned came to mind when you answered the question that appears above, if any?	All

8.2. 2020 Census Attitudes Survey (Instrument Used for Pretesting Purposes)

Thank you for agreeing to participate in this research study. On the next screen, you will be asked to begin filling out a questionnaire. After you have filled out some of the questions, you will be asked to answer some follow-up questions about the previous question items. Your responses will help us make sure that everyone understand the questions the same way. The survey will take 10 minutes or less to complete. By law, the Census Bureau is required to keep your answers confidential. This survey is voluntary and you may refuse to answer any question. The results of this study will be used to inform future surveys and censuses. This collection has been approved by the Office of Management and Budget (OMB). This eight-digit OMB approval number, 0607-0725, confirms this approval. If this number was not displayed, we could not conduct this survey. We estimate that this survey will take about 10 minutes. By proceeding with this study, you give your consent to participate in this study. Your privacy is protected by the Privacy Act. There are a limited number of uses of your data that are permitted under the Privacy Act. You can find a list of these uses in the System of Records Notice (SORN). For this survey, the SORN is named "COMMERCE/CENSUS-5, Decennial Census Program." For more information, please

visit our website at census.gov and click on “Data Protection and Privacy Policy” at the bottom of the home page.

What is your age?

_____years

[PROGRAMMING NOTE: Terminate under 18 yrs]

Are you ....

1. Male
2. Female

What state do you live in?

1. Alabama
2. Alaska
3. Arizona
4. Arkansas
5. California
6. Colorado
7. Connecticut
8. Delaware
9. Florida
10. Georgia
11. Hawaii
12. Idaho
13. Illinois
14. Indiana
15. Iowa
16. Kansas
17. Kentucky
18. Louisiana
19. Maine
20. Maryland
21. Massachusetts
22. Michigan
23. Minnesota
24. Mississippi
25. Missouri
26. Montana
27. Nebraska
28. Nevada
29. New Hampshire
30. New Jersey
31. New Mexico
32. New York
33. North Carolina
34. North Dakota

35. Ohio
36. Oklahoma
37. Oregon
38. Pennsylvania
39. Rhode Island
40. South Carolina
41. South Dakota
42. Tennessee
43. Texas
44. Utah
45. Vermont
46. Virginia
47. Washington
48. West Virginia
49. Wisconsin
50. Wyoming
51. I do not reside in the United States

Were you born in the United States?
1. Yes
2. No

Are you of Hispanic, Latino, or Spanish origin?
1. Yes, Hispanic, Latino, Spanish origin
2. No, not of Hispanic, Latino, or Spanish origin

What is your race? Please select all that apply.
1. White
2. Black or African American
3. American Indian or Alaska Native
4. Asian
5. Native Hawaiian or Pacific Islander
6. Some other race

What is the highest degree or level of school you have completed?
1. No high school
2. Some high school
3. High school graduate or equivalent (for example GED)
4. Some college, but degree not received or is in progress
5. Associate degree (for example AA, AS)
6. Bachelor's degree (for example BA, BS, AB)
7. Graduate degree (for example master's, professional, doctorate)

Lang_Eng. Do you speak a language other than English at home?
1. Yes
2. No

ASK IF Lang_Eng = 1:

What is this language?

ASK IF Lang_Eng = 1:
How well do you speak English?

1. Very well
2. Well
3. Not well
4. Not at all

The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census? Would you say a great deal, some, a little, or nothing at all?

1. A great deal
2. Some
3. A little
4. Nothing at all

Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?

1. A great deal
2. Some
3. A little
4. Nothing at all

How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form.

1. Definitely will
2. Probably will
3. Might or might not
4. Probably will not
5. Definitely will not

How concerned are you, if at all, that the Census Bureau will not keep answers to the 2020 Census confidential?

1. Extremely concerned
2. Very concerned
3. Somewhat concerned
4. Not too concerned
5. Not at all concerned

How concerned are you, if at all, that the Census Bureau will share individuals' answers to the 2020 Census with other government agencies?

1. Extremely concerned
2. Very concerned
3. Somewhat concerned

4. Not too concerned
5. Not at all concerned

How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?

1. Extremely concerned
2. Very concerned
3. Somewhat concerned
4. Not too concerned
5. Not at all concerned

How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census?

1. Extremely concerned
2. Very concerned
3. Somewhat concerned
4. Not too concerned
5. Not at all concerned

/* RANDOM ROTATE SERIES */

Which one of the following is the most important concern, if any concern, about filling out the census form??Mark ONE box.

1. The Census Bureau will not keep answers to the 2020 Census confidential
2. The Census Bureau will share individuals' answers to the 2020 Census with other government agencies
3. The answers you provide to the 2020 Census will be used against you
4. An unauthorized person from outside the Census Bureau will access your answers to the census
5. No concerns
6. Other (Please specify)

_____

Now, we would like to ask you a few questions about some of the questions you just answered.

The first question you were asked was, "The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?"

_____

Now we're going to show you an alternative version of the question. Version 1 (question you answered): "The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" Version 2 (Alternative): "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?

Are these questions asking the same thing or something different?
1. Same thing
2. Something different

Q67. Which version of the question do you prefer?
1. Version 1 (question you answered)
2. Version 2 (alternative)
3. No preference

ASK IF Q67 = 1 OR 2

Can you say more about why you prefer that version?

The first question you were asked was, "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" This question mentions the "2020 Census." Based on your understanding, what is the "2020 Census"?

Now we're going to show you an alternative version of the question. Version 1 (question you answered): "Every ten years, the United States Census Bureau conducts a nationwide census, or a count of all individuals living in the United States. The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?" Version 2 (Alternative): "The next few questions are about the 2020 Census, which will occur on April 1st, 2020. How much have you seen or heard recently – within the last week or so – about the 2020 Census?"

Are these questions asking the same thing or something different?
1. Same thing
2. Something different

Q48. Which version of the question do you prefer?
1. Version 1 (question you answered)
2. Version 2 (alternative)
3. No preference

ASK IF Q48 = 1 OR 2

Can you say more about why you prefer that version?

Another survey question was, "How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form." What comes to mind when you see the phrase "questions on the census form" in this question?

Now we're going to show you an alternative version of the question.

Version 1 (question you answered):
"How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form."

Version 2 (alternative):
"How likely do you think it is that you or someone in your household will answer the questions on the census form?"

Are these questions asking the same thing or something different?
1. Same
2. Something different

Q60. Which version of the question do you prefer?
1. Version 1 (question you answered)
2. Version 2 (alternative)
3. No preference

ASK IF Q60 = 1 OR 2

Can you say more about why you prefer that version?

Now we're going to show you an alternative version of the question.

Version 1 (question you answered):
"How likely are you to participate in the 2020 Census? By participate we mean answer the questions on the census form."

Version 2 (alternative):
"In 2020, the Census Bureau will mail nearly every household in the U.S. a questionnaire and an invitation to participate online. How likely do you think it is that you or someone in your household will mail back the questionnaire or submit it online?"

Are these questions asking the same thing or something different?
1. Same
2. Something different

Q86. Which version of the question do you prefer?
1. Version 1 (question you answered)
2. Version 2 (alternative)
3. No preference

ASK IF Q86 = 1 OR 2

Can you say more about why you prefer that version?

Another survey question was, "How concerned are you, if at all, that the Census Bureau will not keep answers to the 2020 Census confidential?"

What does the word "confidential" mean to you in this question?

In your own words, what does the phrase "the Census Bureau will not keep answers to the 2020 Census confidential" mean to you in this question?

_____

Another question was, "How concerned are you, if at all, that the Census Bureau will share individuals' answers to the 2020 Census with other government agencies?"

In your own words, what does the phrase "share individuals' answers with other government agencies" mean to you in the question above?

_____

What government agencies were you thinking of when you were answering this question?

_____

Another survey question was, "How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?"

What concerns came to mind when you answered the question that appears above, if any

_____

What does the phrase "used against you" mean to you in this question?

_____

The next survey question was, "How concerned are you, if at all, that an unauthorized person from outside the Census Bureau will access your answers to the census?"

What concerned came to mind when you answered the question that appears above, if any?

_____

Thank you for your participation! If you have any additional thoughts on this survey, please provide them in the space below.

_____

8. References

Aranberri, N., G. Labaka, A. Díaz de Ilarraza, and K. Sarasola. 2017. "Ebaluatoia: Crowd Evaluation for English–Basque Machine Translation." *Language Resources and Evaluation* 51(4): 1053–1084. DOI: https://doi.org/10.1007/s10579-016-9335-x.

Barreto, M., L. Frasure-Yokley, E. Vargas, and J. Wong. 2018. "Best Practices in Collecting Online Data with Asian, Black, Latino, and White Respondents: Evidence from the 2016 Collaborative Multiracial Post-Election Survey." *Politics, Groups, and Identities* 6(1): 171–180. DOI: https://doi.org/10.1080/21565503.2017.1419433.

Behr, D., L. Kaczmirek, W. Bandilla, and M. Braun. 2012. "Asking Probing Questions in Web Surveys: Which Factors Have an Impact on the Quality of Responses?" *Social Science Computer Review* 30(4): 487–498. DOI: https://doi.org/10.1177/0894439311435305.

Behr, D., K. Meitinger, M. Braun, and L. Kaczmirek. 2017. "Web Probing-Implementing Probing Techniques from Cognitive Interviewing in Web Surveys with the Goal to Assess the Validity of Survey Question." *Social Science Open Access Repository*: 18. DOI: https://doi.org/10.15465/gesis-sg_en_023

Brown, A., G. López, and M. Hugo Lopez. 2016. *Digital Divide Narrows for Latinos as More Spanish Speakers and Immigrants Go Online*. Washington, DC: Pew Research Center. Available at: https://www.pewresearch.org/hispanic/2016/07/20/digital-divide-narrows-for-latinos-as-more-spanish-speakers-and-immigrants-go-online/ (accessed July 2021).

Buchanan, E. and J. Scofield. 2018. "Methods to Detect Low Quality Data and Its Implication for Psychological Research." *Behavior Research Methods* 50(6): 2586–2596. DOI: https://doi.org/10.3758/s13428-018-1035-6.

Callegaro, M., A. Villar, D.S. Yeager, and J.A. Krosnick. 2014. "A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels Based on Probability and Nonprobability Samples." In *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavrakas: 23–53. Sussex, UK: Wiley.

Chandler, J. and G. Paolacci. 2017. "Lie for a Dime: When Most Prescreening Responses are Honest but Most Study Participants are Impostors." *Social Psychological and Personality Science* 8(5): 500–508. DOI: https://doi.org/10.1177/194855061769 8203.

Chandler, J., C. Rosenzweig, A. Moss, J. Robinson, and L. Litman. 2019. "Online Panels in Social Science Research: Expanding Sampling Methods Beyond Mechanical Turk." *Behavior Research Methods* 51(5): 2022–2038. DOI: https://doi.org/10.3758/s13428-019-01273-7.

Chmielewski, M., and S. Kucker. 2019. "An MTurk Crisis? Shifts in Data Quality and The Impact on Study Results." *Social Psychological and Personality Science* 11(4): 464–473. DOI: https://doi.org/10.1177/1948550619875149.

De Heer, W. and E. de Leeuw. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little: 41–54. New York: Wiley.

Edgar, J. 2013. "Self-Administered Cognitive Interviewing." In 68th American Association for Public Opinion Research Conference, May 2013. Boston, MA. Available at: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2013/Session_A-1-3-Edgar.pdf (accessed July 2021).

Edgar, J., J. Murphy, and M. Keating. 2016. "Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions." *Sage Open* 6(4): 1–14. DOI: https://doi.org/10.1177/2158244016671770.

Goerman, P., L. Fernandez, and R. Quiroz. 2018. "Translation of Country-Specific Programs and Survey Error: Measuring the Education Level of Immigrants." *Translation & Interpreting* 10(2): 21–33. DOI: https://doi.org/10.1177/2158244016671770.

Hanson, M. and L. Espinosa. 2016. "Culture, Ethnicity, and Linguistic Diversity: Implications for Early Childhood Special Education." In *Handbook of Early Childhood Special Education*, edited by B. Reichow, Boyd, E. Barton, and S. Odom: 455–471. Springer, Cham.

Hillygus, S., N. Jackson, and M. Young. 2014. "Professional respondents in nonprobability online panels." In *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavrakas: 219–237. Sussex, UK: Wiley.

Kennedy, C., N. Hatley, A. Lau, A. Mercer, S. Keeter, J. Ferno, and D. Asare-Marfo. 2020. *Assessing the Risks to Online Polls from Bogus Respondents*. Washington, DC: Pew Research Center. Available at: https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2020/02/PM_02.18.20_dataquality_FULL.REPORT.pdf (accessed July 2021).

Kennedy, C., N. Hatley, A. Lau, A. Mercer, S. Keeter, J. Ferno, and D. Asare-Marfo. 2021, "Strategies for Detecting Insincere Respondents in Online Polling." *Public Opinion Quarterly* 85(4): 1050–1075. DOI: https://doi.org/10.1093/poq/nfab057.

Kennedy, R., S. Clifford, T. Burleigh, P. Waggoner, R. Jewell, and N. Winter. 2020. "The Shape of and Solutions to the MTurk Quality Crisis." *Political Science Research and Methods* 8(4): 614–629. DOI: https://doi.org/10.1093/poq/nfab057

Lenzner, T. and C. Neuert. 2017. "Pretesting Survey Questions Via Web Probing–Does it Produce Similar Results to Face-to-Face Cognitive Interviewing?" *Survey Practice* 10(4): 2768. DOI: https://doi.org/10.29115/SP-2017-0020.

Luiten, A., J. Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study Across Countries and Surveys." *Journal of Official Statistics* 36(3): 469–487. DOI: https://doi.org/10.2478/jos-2020-0025.

Matthijsse, S., E. de Leeuw, and J. Hox. 2015. "Internet Panels, Professional Respondents, and Data Quality." *Methodology* 11(3). DOI: https://doi.org/10.1027/1614-2241/a000094.

Meitinger, K. and D. Behr. 2016. "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?" *Field Methods* 28(4): 363–380. DOI: https://doi.org/10.1177/1525822X15625866.

Meitinger, K., D. Behr, and M. Braun. 2019. "Using Apples and Oranges to Judge Quality?: Selection of Appropriate Cross-National Indicators of Response Quality in Open-Ended Questions." Social Science Computer Review 39(3): 1–22. DOI: https://doi.org/10.1177/0894439319859848.

Murphy, J., Edgar, J., and Keating, M. 2014. "Crowdsourcing in the Cognitive Interviewing Process." In the Annual Meeting of the American Association for Public Opinion Research, May, 2014. Anaheim, CA. Available at: http://dc-aapor.org/2014%20conference%20slides/EdgarMurphyKeating.pdf (accessed January 2021).

Neuert, C., and T. Lenzner. 2019. "Effects of the Number of Open-Ended Probing Questions on Response Quality in Cognitive Online Pretests." *Social Science Computer Review* 39(3): 456–468. DOI: https://doi.org/10.1177/0894439319866397.

Ortega-Santos, Iván. 2019. "Crowdsourcing for Hispanic Linguistics: Amazon's Mechanical Turk as a source of Spanish data." Borealis – An International Journal of Hispanic Linguistics 8(1): 187–215. DOI: https://doi.org/10.7557/1.8.1.4670.

Porter, C.O., R. Outlaw, J.P. Gale, and T.S. Cho. 2019. "The Use of Online Panel Data in Management Research: A Review and Recommendations". *Journal of Management* 45(1): 319–344. DOI: https://doi.org/10.1177/0149206318811569.

Prince, K.R., A.R. Litovsky, and D.G. Friedman-Wheeler. 2012. "Internet-Mediated Research: Beware of Bots." *The Behavior Therapist*, 35(5): 85–88. Available at: https://services.abct.org/i4a/doclibrary/index.cfm?category_id = 15

Rumbaut, R.G., and D.S. Massey. 2013. "Immigration & Language Diversity in the United States." *Daedalus* 142(3): 141–154. DOI: https://doi.org/10.1162/DAED_a_ 00224.

Shepperd, J.A., G. Pogge, J.M. Hunleth, S. Ruiz, and E.A. Waters. 2021. "Guidelines for Conducting Virtual Cognitive Interviews During a Pandemic." *Journal of Medical Internet Research* 23(3): 25173e. DOI: https://doi.org/10.2196/25173.

Simone, M. 2019a. "How to Battle the Bots Wrecking Your Online Study," *Behavioral Scientist*. Available at: https://behavioralscientist.org/how-to-battle-the-bots-wrecking-your-online-study/. (accessed January 2021).

Simone, M. 2019b. "Bots started sabotaging my online research. I fought back," STAT. Available at: https://www.statnews.com/2019/11/21/bots-started-sabotaging-my-online-research-i-fought-back/. (accessed January 2021).

Storozuk, A., M. Ashley, V. Delage, and E. Maloney. 2020. "Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks." *Quantitative Methods for* Psychology 16(5): 472–481. DOI: https://doi.org/10.20982/tqmp. 16.5.p472.

Team Y&R. 2020. *Census Tracking Survey*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Available at: https://doi.org/10.3886/E125761V3.

Trejo, Y.A.G., and A. Schoua-Glusberg. 2017. "Device and Internet Use among Spanish-Dominant Hispanics: Implications for Web Survey Design and Testing." *Survey Practice* 10(3). DOI: https://doi.org/10.29115/SP-2017-0016.

University of Maryland. 2014. "Institutional Review Board Investigator Handbook." College Park, MD: University of Maryland. Available at: https://research.umd.edu/-sites/default/files/documents/irb-forms/Investigator%20Handbook%20FINAL%20112114.pdf (accessed July 2021).

Upadhyay, U.D., and H. Lipkovich. 2020. "Using Online Technologies to Improve Diversity and Inclusion in Cognitive Interviews with Young People." *BMC Medical Research Methodology* 20(1): 1–10. DOI: https://doi.org/10.1186/s12874-020-01024-9.

Willis, G.B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications. DOI: https://dx.doi.org/10.4135/9781412983655.

Yarrish, C., L. Groshon, J. Mitchell, A. Appelbaum, S. Klock, T. Winternitz, and D. Friedman-Wheeler. 2019. "Finding the Signal in the Noise: Minimizing Responses from Bots and Inattentive Humans in Online Research." *The Behavior Therapist* 42(7): 235–242. Available at: https://services.abct.org/i4a/doclibrary/index.cfm?category_id = 22

Yu, E., A. Fobia, J. Graber, J. Holzberg, R. Kaplan, B. Kopp, K. Kubzdela, B. Mockovak, R. Morrison, P. Scanlon. 2019. *Experiences Using Online Testing to Support Survey-Methods Research and Pre-Testing in the Federal Government.* Research and Methodology Directorate, Center for Behavioral Science Methods Research Report Series (Survey Methodology 2019-06). Washington, DC: U.S. Census Bureau. Available at: http://www.census.gov/content/dam/Census/library/working-papers/2019/adrm/rsm2019-06.pdf (accessed January 2021).

Received January 2021
Revised July 2021
Accepted March 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 823–845, http://dx.doi.org/10.2478/JOS-2022-0036

Measuring and Mapping Micro Level Earning Inequality towards Addressing the Sustainable Development Goals – A Multivariate Small Area Modelling Approach

Saurav Guha[1] *and Hukum Chandra*[1]

The earning inequality in India has unfavorably obstructed underprivileged in accessing elementary needs like health and education. Periodic labour force survey conducted by National Statistical Office of India generates estimates on earning status at national and state level for both rural and urban sectors separately. However, due to small sample size problem, these surveys cannot generate reliable estimates at micro-level viz. district or block. Thus, owing to unavailability of district-level estimates, analysis of earning inequality is restricted to the national and the state level. Therefore, the existing variability in disaggregate-level earning distribution often goes unnoticed. This article describes multivariate small area estimation method to generate precise and representative district-wise estimate of earning distribution in rural and urban areas of the Indian State of Bihar by linking Periodic labour force survey data of 2018–2019 and 2011 Population Census data of India. These disaggregate-level estimates and spatial mapping of earning distribution are essential for measuring and monitoring the goal of reduced inequalities related to the sustainable development of 2030 agenda. They expected to offer insightful information to decision-makers and policy experts for identifying the areas demanding more attention.

Key words: Multivariate small area estimation; earning inequality; SDG; periodic labour force survey; NSO; census.

1. Introduction

The economy of India has developed at verifiably notable degrees and is currently one of the fastest developing nation in terms of economy in the world. India has progressed substantially in reforming its economy, reducing the hardship and fulfilling opportunities for day to day comfort of its vast population. This unceasing economic growth has also led to significant reduction in poverty level and has big impact on the overall wellbeing of the people in this country. However, in spite of these remarkable achievement, the earning distribution in India remains obstinately uneven.

The movement of economy also directs the behavior of the labour market. The volatility in the economy, both in its inter and intra sectored linkages as well as in the framework of economic integration with rest of the world, is reflected in the domestic labour market (MoSPI 2020a). Global economic slowdown creates extreme volatility which can enormously influence the contemporary economic atmosphere. Thus, it is hugely

[1]ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi, India. Emails: saurav.iasri@gmail.com and hchandra12@gmail.com
Acknowledgments: The authors would like to acknowledge the valuable comments and suggestions of the Editor, the Associate Editor and three anonymous referees.

© Statistics Sweden

imperative to quantify its short term impact on labour market which involves the collection of labour force data at regular interval. In India, labour force participation in unorganized sector is much higher as compared to the organized sector. The frequent availability of labour force data was the need of the hour and that led to the launch of Periodic labour force survey (PLFS) in 2017 by National Statistical Office (NSO), Ministry of Statistics and Program Implementation (MoSPI), Government of India. In India, NSO is the primary body to collect PLFS data for generating estimates at the state and the national level for both rural and urban areas. The PLFS data provides estimates for a range of employment and unemployment indicators such as unemployment rate, worker population ratio, labour force participation rate, earning of different working groups. Even though being exceptionally crucial, the estimates of earning distribution are unattainable further down the state level in India for example, district, block or further level of disaggregation.

The rising earning inequality in the world indicates that the richest 10% have up to 40% of global earning whereas the poorest 10% earn only between 2% to 7% and the population growth inequality in developing countries has increased by 11% (UNDP 2015). Regardless of the achievement of Millennium Development Goals (MDGs), several objectives including equality have not been unified. As the primary focus of MDGs was on national and global averages, the growing disparities at the local level remained unchecked. The Post-2015 Development Agenda reckon with the Rio + 20 promises to "strive for a world that is just, equitable and inclusive". The Sustainable Development Goals (SDGs), also known as the Global Goals, were adopted by all United Nations Member States in 2015 as a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity by 2030 (UNDP 2015). Through the pledge to "Leave No One Behind", the SDGs ensure the addition of marginalized and nonempowered groups, and reduction of inequalities together with eradication of poverty. Extreme inequality persists in multiple extents within and between nations and it often creeps into the decision-making processes as well. According to UNCDP (2018), segments of the population that are typically excluded from meaningful participation in decision-making are unlikely to see their interests safeguarded in policy and investment decisions, as is the case for countries that are marginalized in international decision-making structures. The SDG10 among the 17 goals that constitute the 2030 Agenda for SDG aims to "reduce inequality within and among countries". The comprehensive targets to reduce inequality "within" and "among" countries emphasize the significance of inequality reduction by all means in all countries.

Inequality creates barrier to growth and development when it denies people of opportunities which in turn, lead to the state of extreme poverty. There is a growing consensus that economic growth is not sufficient to reduce poverty if it is not inclusive and if it does not involve the three dimensions of sustainable development – economic, social and environmental, (UNDP 2015). Goal 10 aims that the income growth of the bottom 40% of country's population is higher than the national average by the year 2030. The Gini coefficient of income inequality for India fell from 36.8% in 2010 to 33.6% in 2015. The Government of India prioritizes the policy of inclusion, financial empowerment and social security via major initiatives like Jan Dhan Yojona, Aadhaar etc. These comprehensive steps are in line with the SDG targets intended for achieving better equality and

encouraging the socio-economic, and political inclusion of all by 2030. In this present situation, the budding interests of the policy makers, public agencies, scientists, government organizations are focused in attaining the local (or micro) level estimates. The emphasis on disaggregate level SDG indicators by various national and international agencies has further emphasized the inevitable need of such local level estimates. These local level areas or domains, better known as small areas or small domains are formed by cross-classification of several demographic and topographic variables that includes small topographic areas (e.g., districts) or small demographic groups (e.g., land category, social groups, religion, age-sex groups) or cross classifying both (Guha and Chandra 2021a). Besides, in the existing PLFS data of NSO, the small areas or districts may have very small or even zero sample sizes which may lead to large sampling error in case of direct estimation. The small area estimation (SAE) methodology provides a viable and cost effective solution to this problem of small sample sizes (Rao and Molina 2015). The SAE techniques borrow strength from various external sources viz. time periods, areas and so on to obtain precise and reliable estimates.

The key idea behind the SAE methodology is to link the variable under study with the auxiliary information through different statistical models. This defines the model-based small area estimates corresponding to these small areas. Based on the availability of the auxiliary information, the unit level or the area level models are mostly used in SAE. The Fay–Herriot (FH) model (Fay and Herriot 1979) is a widely accepted area level model in SAE as it is better suited to account for complex survey design for point and variance estimates compared to the unit-level model and when the model covariates are available only in aggregate form. The FH model assumes the availability of area specific survey estimates and these estimates follow an area level linear mixed model with area as random effects. Application of the FH model are readily available in literature in multiple dimensions. The uncertainty of the SAE estimates was deliberated by Prasad and Rao (1990), Datta et al. (2011), Fay (1987) and Datta et al. (1991) introduced the multivariate version of the FH model while Benavent and Morales (2016) extended it further. Often, there is a need of estimating correlated processes viz. poverty indicators, unemployment, and so on. Multivariate models often allow for the correlation of several variables and usually suitable in these circumstances. Unlike the FH model, more than one variable of interest is modelled via multivariate Fay–Herriot model (MFH) by allowing for different covariance structure between the vector of the variable of interest and the random effects, see Guha and Chandra (2021b). A number of small area applications for estimating socio-economic indicators have been described in literature based on univariate FH model that ignores the correlation between the target variables, see, for example, Chandra et al. (2011), and references therein. Furthermore, surveys are generally planned to collect information on more than a few variables. In SAE problem, when the target areas comprise of insufficient sample size, taking into account of correlation between the target variables can provide an added advantage in obtaining precise and reliable small area estimates (Rao and Molina 2015). Franco and Bell (2021) also pointed out that precision in bivariate area-level models is only improved if one of the outcomes has very low variance and the correlation between the two outcomes is very strong.

According to the quarterly report of MoSPI (2020b), the unemployment rate in the age group of 15 years and above has sharply increased from 9.1% in January–March 2020 to

20.8% in April–June 2020 with the working population ratio decreased from 43.7% to 36.4%. These figures indicate the severity of the job losses and sufferings faced by the majority of the working population in the country during the first phase of the COVID-19 pandemic. Given the severe economic hardships faced by a large section of the populations during this pandemic, having precise knowledge of district-level estimates of pre-pandemic earning distribution is critical for evaluating the true impact of the disaster. India reported the second highest number of COVID-19 cases for any country in the world (>32 million) by mid-2021, with Bihar contributing to almost 8 million cases (MoHFW 2021). Bihar, the third most populous state in the country, accounts for about 8.6% of India's population with an area of 94163 square km that equals to 2.9% of India's total geographical area. About 33.74% population of the state lives below the poverty line which is higher than the national average of 21.92% (NITI Aayog 2019). According to the Global Hunger Index 2020 (GHI 2020), out of 132 countries India stands on 94th position with an overall score of 27.2. The state of Bihar ranks 27th out of 28 states in "zero hunger" with a score of 31 which is much lower than the national average of 47 (NITI Aayog 2019). In addition, 75.2% of the population is in the lowest two wealth quintiles in Bihar which is highest in India and Bihar ranks 25th out of 28 states in "reduced inequality" parameter with a score of 48 which is much lower than the national average of 67 (NITI Aayog 2019). Therefore, it seems rational to consider Bihar to generate the district level estimates of earning inequalities at rural and urban sectors using SAE techniques. To the best of our knowledge, no prior study has been done to estimate the micro level or disaggregate level earning inequalities in India.

The article is organized as follows. Section 2 describes the data from the 2018–2019 Periodic labour force survey of the NSO of India and the 2011 Population Census of India that will be used to estimate the district level earning distribution in rural and urban sector of the Indian State of Bihar. In this section, we also set out the theoretical background of the area level MFH model, and then discuss the different variant of this model used in estimating small area means under this model. The results obtained from the application of district-level inequalities in earning distribution along with various diagnostic measures are reported in Section 3. We also provide spatial mapping of earning distribution in this section that serves to demonstrate the degree of district-level inequalities in the distribution of earning between rural and urban sector in Bihar. Finally, Section 4 summarizes the article and provides concluding remarks.

2. Data and Model Specification

In this section, we introduce the major data sources utilized in multivariate SAE application. The 2018–2019 PLFS data of the NSO for rural and urban districts of Bihar and the data from 2011 Population Census of India are used for estimating the district level inequality in the earning distribution between rural and urban sector of the state. Afterward, we briefly describe the multivariate SAE methodology applied in the estimation of district level inequality in distribution of average monthly earning.

2.1. Description of PLFS and Population Census Data

The PLFS survey data is freely downloadable from the MoSPI , Government of India (http://mospi.nic.in/). Since 2017, NSO carries out the PLFS every year. In PLFS a

rotational panel sampling design for first visit both in rural and urban areas and three periodic revisit in urban areas has been used while there was no revisit in rural areas. A stratified multistage survey design was adopted, with the final units being households. The 2018–2019 PLFS of NSO is intended to produce precise and reliable estimates at the state and the national level for both rural and urban areas in the country. However at district level, this PLFS data cannot directly be used to generate precise and reliable estimates, since sample size within each district is not adequate to offer district-level estimates with acceptable reliability and precision. Although, district is always being a very crucial part of the planning process in the country, there are no surveys conducted to produce district level estimates in India and this leads to limit the policy interventions at the district or even further lower level (Guha and Chandra 2021b).

The 2018–2019 PLFS data of the NSO comprised 15,441 persons in 3,179 households from the rural and urban areas in 38 districts of Bihar. For all the districts, the sample size ranges from 07 to 205 with an average of 92 for rural areas while for urban areas, it is 06 to 362 with an average of 45. This survey provides information on earning estimate of every person separately for rural and urban areas in the state of Bihar. Districts included moderately small sample sizes with an average sampling fraction of 0.000046 for rural and 0.00027 for urban areas. Due to the problem of small sample size, it is barely possible to obtain precise and reliable direct estimates at district level which eventually leads to producing large standard errors from this survey, see Chandra et al. (2011) and Rao and Molina (2015). In this paper, we made an attempt to address the problem of small sample size in attaining district-level estimates from the 2018–2019 PLFS data. The multivariate small area method has been implemented to tackle this problem by including related auxiliary variables from the Population Census 2011 of India (Census 2011).

We have considered the following information on earning from employment from PLFS 2018–2019 viz. (a) self-employed persons, (b) salaried employees and regular wage earner, and (c) person working as casual labour. For salaried employees and regular wage earner in current weekly status (CWS), information on earnings in the previous calendar month was collected. For self-employed persons in CWS, information on earnings in the last 30 days from the self-employment was collected. It is important to note that average gross earnings from the self-employment activity have been calculated by excluding those self-employed persons who had reported earning as zero or not reported. For the person working as casual labour (except public works), information on earnings was collected for the casual labour work in which the person was engaged for each and every day of the reference week that is, last seven days prior to the date of the survey. For the sake of the analysis, we have transformed the daily data into monthly data for the casual labour work. The estimates in this section are derived using the data collected in the first visit schedules in the rural areas (since there was no revisit in rural areas) and for the urban areas using the data collected in the schedules of first visit and the corresponding revisits conducted during the four quarters of the survey period, viz., during July to September followed by October to December in 2018 and January to March followed by April to June in 2019. For more detailed information on the method of the data collection, readers may refer to the annual report of the PLFS 2018–2019 (MoSPI 2020a). The target variables in the 2018–2019 PLFS data are Y_1: average monthly earning (in Rs.) of a person from employment in rural areas (hereafter denoted by Rural-Earning), Y_2: average monthly

earning (in Rs.) of a person from employment in urban areas (hereafter denoted by Urban-Earning). This paper targets to estimate the inequality in average monthly earning of a person in rural and urban districts of Bihar at small area level through joint modelling of the target variables that is, Rural-Earning and Urban-Earning.

2.2. *Multivariate Small Area Modelling*

Let us consider that the population consists of M domains and let there are R target variables in this study. All the way through, a subscript $m(m = 1, \ldots, M)$ is used to denote the quantities possess by domain m and a subscript $r(r = 1, \ldots, R)$ is used to index the variable r under study. Assume a finite population Ω of size N comprises M non-overlapping domains $\Omega_m; m = 1, \ldots, M$ and a sample s of size n is drawn from Ω by any probability sampling design. We also assume that the domain size N_m is known for each domain and n_m units are selected in the sample from N_m units of m^{th} domain (hereafter denoted by small area). The population total is given by $N = \sum_{m=1}^{M} N_m$ and the corresponding sample size is $n = \sum_{m=1}^{M} n_m$. Let y_{mrj} be the value corresponding to j^{th} unit of the r^{th} target variable in m^{th} area, $r = 1, \ldots, R, j = 1, \ldots, N_m$ and $m = 1, \ldots, M$. The aim is to estimate small area mean $\bar{Y}_{mr} = N_m^{-1} \sum_{j \in \Omega_m} y_{mrj}$, $r = 1, \ldots, R$ and $m = 1, \ldots, M$. The traditional direct survey estimator (hereafter denoted by Direct) for $\bar{Y}_{mr}$ is given by $\bar{y}_{mr} = \sum_{j=1}^{n_m} \tilde{w}_{mrj} y_{mrj}$ with $\tilde{w}_{mrj} = w_{mrj} / \sum_{j=1}^{n_m} w_{mrj}$ where $\tilde{w}_{mrj}$ is the normalized survey weight for j^{th} unit of the r^{th} variable in m^{th} area. In addition, $\tilde{w}_{mrj}$ satisfies $\sum_{j=1}^{n_m} \tilde{w}_{mrj} = 1$ with w_{mrj} being the survey weight for j^{th} unit of the r^{th} variable in m^{th} area. Following Särndal et al. (1992), the estimated variance of Direct estimator is approximated by $v(\bar{y}_{mr}) = \sum_{j=1}^{n_m} \tilde{w}_{mrj}(\tilde{w}_{mrj} - 1)(y_{mrj} - \bar{y}_{mr})^2$. Under simple random sampling without replacement (SRSWOR), $w_{mrj} = 1/\pi_{mrj}$ where $\pi_{mrj} = n_m/N_m$ is the inclusion probability for j'$^{\text{th}}$ unit of the r^{th} variable in the m^{th} area.

Let us further assume that $y_{mr}(r = 1, \ldots, R)$ be an unbiased direct survey estimates of an unknown population parameter (e.g., the population mean) Y_{mr} of the target variable r for small area m. Let x_{mr} be a p_r -vector of available auxiliary variables corresponding to area m that are associated to the population mean Y_{mr} for the variable r under study. Usually, area-specific auxiliary information are acquired from some available secondary sources, for example, administrative registers, the population census, etc. We denote $\mathbf{y}_m = (y_{m1}, \ldots, y_{mr})^T$, a vector of direct survey estimates of $\boldsymbol{Y}_m$ where $\boldsymbol{Y}_m$ is the m-vector population mean of target variables. In line with Benavent and Morales (2016), an area-level FH model (Fay and Herriot 1979) used for more than one target variables is given by

$$\mathbf{y}_m = \boldsymbol{Y}_m + \boldsymbol{\varepsilon}_m \text{ and } \boldsymbol{Y}_m = \mathbf{X}_m \boldsymbol{\beta} + \mathbf{u}_m. \tag{1}$$

In literature, the model in Equation (1) is time and again referred to the multivariate form of the FH model. The MFH model in Equation (1) consists of two stages, the first one takes care of the sampling variability of the direct survey estimates $\mathbf{y}_m$ of true area means of the target variable $\boldsymbol{Y}_m$ while the second stage accounts for linking of the true area means of the target variable $\boldsymbol{Y}_m$ to $\mathbf{X}_m = diag\,(\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mR})_{R \times p}$, a matrix of available auxiliary variables where $p = \sum_{r=1}^{R} p_r$. This model in Equation (1) can be denoted as an area level random effect model as

$$\mathbf{y}_m = \mathbf{X}_m\boldsymbol{\beta} + \mathbf{u}_m + \boldsymbol{\varepsilon}_m, \quad m = 1, \ldots, M. \tag{2}$$

Here $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_r')_{p\times 1}'$ and $\boldsymbol{\beta}_r$ is a p_r - vector of unknown fixed effect parameters. The vector of ind, random area effects $\mathbf{u}_m$ are independent and identically distributed with $\mathbf{u}_m$ *ind* $N\left(0,\ \mathbf{V}_{u_m}\right)$ while vectors of sampling errors ε_m are independent and normally distributed with $\boldsymbol{\varepsilon}_m \sim N(0, \mathbf{V}_{\varepsilon_m})$. Moreover, these two vector of errors $\mathbf{u}_m$ and $\boldsymbol{\varepsilon}_m$ are independent of each other within and between areas with $\mathbf{V}_{\varepsilon_m}$, the covariance matrices of ε_m are known while the covariance matrices of $\mathbf{u}_m$ denoted by $\mathbf{V}_{u_m}$ depend on unobservable parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R)$. Combining M-area-level models, the model in Equation (2) can be denoted in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{y} = \text{col}\ (\mathbf{y}_m; 1 \le m \le M)$ is the vector of direct estimates of order $MR \times 1$, $\mathbf{X} = \text{col}\ (\mathbf{X}_m; 1 \le m \le M)$ is the matrix of known covariates of dimension $MR \times p$, $\mathbf{u} = \text{col}\ (\mathbf{u}_m; 1 \le m \le M)$ is the vector of random area effects of dimension $MR \times 1$ and $\boldsymbol{\varepsilon} = \text{col}(\boldsymbol{\varepsilon}_m; 1 \le m \le M)$ is the vector of sampling errors of dimension $MR \times 1$ with $\mathbf{u} \sim N(0, \mathbf{V}_u)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V}_\varepsilon)$. $\mathbf{Z} = \text{col}'(\mathbf{Z}_m; 1 \le m \le M)$ is a matrix of order $MR \times MR$ whose mth column $z_m, m = 1, \ldots, M$, is an indicator variable which takes the value 1 if a unit belongs to an area m and zero otherwise. Especially, in model (3) $\mathbf{Z}$ is an $MR \times MR$ identity matrix. Furthermore, it is supposed that the random area effects u are independently distributed of the sampling errors ε where $\mathbf{u} \sim N(0, \mathbf{V}_u)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V}_\varepsilon)$. The random effects covariance matrix is denoted by $\mathbf{V}_u = \text{diag}\ (\mathbf{V}_{um}; 1 \le m \le M)$ and $\mathbf{V}_\varepsilon = \text{diag}\ (\mathbf{V}_{\varepsilon m}; 1 \le m \le M)$ is the matrix of design variances.

Next we consider three types of the model (3) to obtain model-based small area estimates. First, we take $\mathbf{V}_{u_m} = \text{diag}\ (\sigma_{ur}^2;\ 1 \le r \le R)$, $\mathbf{V}_{\varepsilon_m} = \text{diag}\ (\sigma_{\varepsilon mr}^2;\ 1 \le r \le R)$, $m = 1, \ldots, M$ and $\sigma_{\varepsilon mr}^2$'s are known for the estimator based on univariate FH model (UFH). Second estimator, denoted by MFH-1, is based on MFH model with $\mathbf{V}_{u_m} = \text{diag}$ $(\sigma_{ur}^2; 1 \le r \le \text{R})$, $m = 1, \ldots, M$, and a known but not necessarily diagonal matrix $\mathbf{V}_\varepsilon$. The third estimator, denoted by MFH-2, is also based on MFH model where the random effects $\mathbf{u}_m = (\mathbf{u}_{m1}, \ldots, \mathbf{u}_{mR})'$ is generated via a first order heteroscedastic autoregressive HAR(1) process $\mathbf{u}_{mr} = \rho\mathbf{u}_{mr-1} + \boldsymbol{\tau}_{mr}$ with $\mathbf{u}_{m0}$–$N(0, \sigma_0^2)$, $\boldsymbol{\tau}_{mr}$–$N(0, \sigma_2^2)$, $r = 1, \ldots, R$ and σ_2^2, $\mathbf{u}_{m0}$, $\boldsymbol{\tau}_{mr}$ are independent. The components of $\mathbf{V}_{u_m}$ are given by $\sigma_{umii} = \sum_{k=0}^{i} \rho^{2k}\sigma_{1-k}^2$ and $\sigma_{umij} = \sum_{k=0}^{|j-i|} \rho^{2k+|j-i|}\sigma_{|j-i|-k}^2, i \ne j$. It is further assumed that sampling errors are not independent with each other that is, $\mathbf{V}_\varepsilon$ is known but not essentially a diagonal matrix. For UFH and MFH-1 estimators, the number of unknown variance component parameters to be estimated is equal to R with $\theta_r = \sigma_{ur}^2$, $r = 1, \ldots, R$ and for MFH-2, it is $R + 1$ with $\theta_r = \sigma_{ur}^2$, $r = 1, \ldots, R$ and $\theta_{R+1} = \rho$. Under the model (3), $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Var(\mathbf{y}) = \mathbf{V}_y = \mathbf{V}_u + \mathbf{V}_\varepsilon = diag(\mathbf{V}_{ym;} 1 \le m \le M)$, with $\mathbf{v}_u = \mathbf{z}'\mathrm{v}_u\mathbf{z}$ and $\mathbf{V}_{ym} = \mathbf{V}_{um} + \mathbf{V}_{sm}$, $m = 1, \ldots, M$. Here, the covariance matrix $\mathbf{V}_y$ depends on R and $R + 1$ unknown variance component parameters given by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R)$ for UFH and MFH-1 and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{R+1})$ for MFH-2 model respectively. The restricted maximum likelihood (REML) method is applied to estimate θ. Replacing the estimated values $\hat{\theta}$ of parameters $\boldsymbol{\theta}$ in $\mathbf{V}_u$ to obtain $\hat{\mathbf{V}}_u = \mathbf{V}_u(\hat{\theta})$ and $\hat{\mathbf{V}}_y = \hat{\mathbf{V}}_u + \mathbf{V}_\varepsilon$, the multivariate version of empirical best linear unbiased predictors (EBLUP) of $\boldsymbol{Y}$ is defined as

$$\hat{\boldsymbol{Y}}_{\mathrm{MFH}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}. \tag{4}$$

Here, the empirical best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the EBLUP of $\mathbf{u}$ are obtained $\hat{\boldsymbol{\beta}} = (\mathbf{X}', \hat{\mathbf{V}}_y^{-1}\mathbf{X})^{-1}\ \mathbf{X}'\hat{\mathbf{V}}_y^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \hat{\mathbf{V}}_u\mathbf{Z}'\hat{\mathbf{V}}_y^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ respectively. In small area applications, the mean squared error (MSE) estimates are desirable to measure the precision of estimates and also to construct the confidence interval for the estimates (Guha and Chandra 2021b). The analytical MSE estimate of EBLUP of MFH in Equation (4) is obtained following Benavent and Morales (2016).

3. Results and Discussions

3.1. Variable Selection and Model Fitting

We used Population Census 2011 (Census 2011) data of India for selection of suitable covariates for small area modelling. As these covariates are available as counts at district level, area-level multivariate small area models were used in this analysis to obtain the small area estimates. There are almost 30 auxiliary variables accessible from the census data and we did some exploratory analysis prior to selection of appropriate covariates for multivariate small area model fitting. Principal Component Analysis (PCA) was also applied to obtain combined scores for some particular sets of auxiliary variables. The PCA variables (i.e., combined scores obtained using PCA) in place of Population Census variables are frequently used in SAE method since maximum variation is explained with dimension reduction of auxiliary variables, see, for example Johnson et al. (2010) and Chandra et al. (2011). In particular, we did PCA independently on three sets of auxiliary variables for both Rural-Earning and Urban-Earning and all measured at district level. For Rural-Earning and Urban-Earning, these three sets of auxiliary variables are noted as PC1, PC2 and PC3 below. The first set (PC1) included the literacy rates by gender and proportions of worker population by gender, the second set (PC2) included the proportions of main worker by gender, proportions of main cultivator by gender and proportions of main agricultural labourers by gender and the third set (PC3) included the proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. Table 1 describes the proportion of explained variability by these principal components in all the three sets of auxiliary variables for both the target variables. This followed by a stepwise regression for choosing significant auxiliary variables based on Akaike information criterion value.

Initially, the direct estimates of two target variables that is, Rural-Earning and Urban-Earning are plotted to get an impression about the correlation between them. From Figure 1, it seems that these two target variables Rural-Earning and Urban-Earning are loosely correlated. Here it should be noted that, for Urban-Earning, there was 05 non-sample districts and for Rural-Earning, there was one non-sample district. Consequently, we fit a FH model separately to both the target variables with the sample areas and then a synthetic estimation (see Chandra et al. 2011) is carried out to estimate the non-sample areas. MSE of the synthetic estimates are obtained following Chandra et al. (2011). In most of the cases, the estimates of error variance are extremely noisy and need to be smoothed before being used

Table 1. Principal components and explained variations for the target variable rural-earning and urban-earning.

Variable	Set	Principal component	Proportion of variance	Cumulative proportion
	PC1	R-PC11	0.958	0.958
		R-PC12	0.041	1.000
Rural-earning	PC2	R-PC21	0.953	0.953
		R-PC22	0.027	0.980
	PC3	R-PC31	0.963	0.963
		R-PC32	0.037	1.000
	PC1	U-PC11	0.915	0.915
		U-PC12	0.085	1.000
Urban-earning	PC2	U-PC21	0.714	0.714
		U-PC22	0.217	0.931
	PC3	U-PC31	0.807	0.807
		U-PC32	0.193	1.000

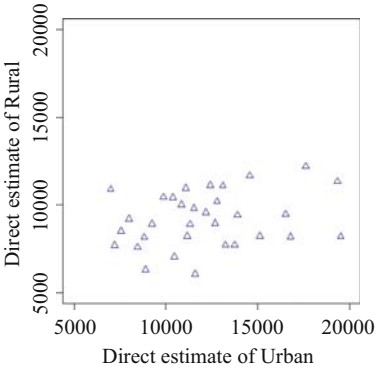


Fig. 1. Scatter plot of the direct estimates of rural-earning and urban-earning.

in the analysis. In line with Schall (1991), we applied a generalized variance function (GVF) method for smoothing the estimated error variance of the direct estimates. The details of the GVF method is described in the supplementary materials. Next, we proceed with the MFH-2 model described in the previous section using a combination of the direct estimates and the synthetic estimates corresponding to the sample and the non-sample areas as the input of the two target variables and some selected covariates from the census data as suitable auxiliary variables. Finally, three significant covariates viz. proportions of marginal cultivator population (MCP), proportions of marginal agriculture labour population (MALP) and R-PC31 corresponding to the target variable Rural-Earning and for Urban-Earning, four significant covariates viz. proportions of main worker population (MWP), proportions of marginal cultivator population (MCP), U-PC21 and U-PC31 are included in the model. The regression parameter estimates are reported in Table 2 for the two dependent variable Rural-Earning and Urban-Earning. Observing the signs of the regression parameters estimates, it can be concluded that rural districts having lesser proportion in all the three covariates have more earning while urban districts having greater proportion U-PC21 and U-PC31 and lesser proportion in MWP and MCP have more earning.

Table 2. Regression parameters, standard error and p-values for the target variables rural-earning and urban-earing.

Variable	Parameter	Estimate	Standard error	P-value
Rural-Earning	Intercept	21,906	4,835	< 0.001
	MALP	-14,987	5,553	< 0.01
	R-PC31	-5,429	1,907	< 0.01
Urban-Earning	Intercept	45,139	12,169	< 0.001
	MWP	-38,588	14,644	< 0.01
	MCP	-45,798	19,477	0.018
	U-PC31	7,699	3,499	0.021

The values of the estimates from fitting the multivariate small area model using 2018–2019 PLFS data are described as follows. The estimate of variance component parameters for the MFH-2 model are $\hat{\sigma}_{u1} = 1317$, $\hat{\sigma}_{u2} = 2448$ and $\hat{\rho} = 0.478$. We also test the null hypothesis $H_0 : \sigma^2_{ui} = \sigma^2_{uj},\ i,j = 1,2,\ i \neq j$ against the alternative hypothesis $H_1 : \sigma^2_{ui} \neq \sigma^2_{uj}$. The test statistic is given by $t_{ij} = \hat{\sigma}^2_{ui} - \hat{\sigma}^2_{uj}\ /\sqrt{v_{11} + v_{22} - 2v_{12}};\ i,j = 1,2,\ i \neq j$, where $v_{rs}, r,s = 1,2$ are the elements of the inverse of the matrix of Fisher information corresponding to MFH-2 model calculated at $\hat{\theta} = \left(\hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}, \hat{\rho}\right)$. The value of the test statistic is given by $t_{12} = -2.056$ (0.039) with p-value given in parenthesis. As the value of the test statistic is significant at 5% level, this leads to the conclusion that variance of random area effects for Rural-Earning and Urban-Earning are significantly different. This followed by testing $H_0 : \rho = 0$ with the test statistic $t_\rho = \hat{\rho}/\sqrt{v_{33}}$ and the value of $t_p = 1.071(0.284)$, p-value is in parenthesis. This reveals that the correlation between the two target variables is not significantly different from zero and we proceed with MFH-1 model with diagonal covariance matrix of random effects instead of MFH-2 model. It is important to note that, although the variance of random area effects for Rural-Earning and Urban-Earning are significant at 5% level, they are not significant at 1% level and also the correlation between the two target variables is not significantly different from zero. This leads to the almost identical results in univariate and multivariate estimates which established the idea reported in Franco and Bell (2021) that that precision in multivariate area-level models is only improved if one of the outcomes has very low variance and the correlation between the two outcomes is very strong. Finally, the MFH-1 model is applied with all the significant auxiliary variables to obtain the earnings estimates that is, Rural-Earning and Urban-Earning for all the districts in Bihar.

3.2. Diagnostic Measures

In what follows, we described some standard diagnostic measures to examine the model assumptions and inspect the reliability and validity of the generated estimates through MFH method. In line with Brown et al. (2001), two forms of diagnostics viz. (a) the model diagnostics, and (b) the multivariate SAE diagnostics are employed to endorse the model assumptions. The reliability of the model-based estimates of Rural-Earning and Urban-Earning attained by SAE method under MFH-1 model is validated by some additional diagnostics. Corresponding to the target variable Rural-Earning and Urban-Earning, the random effects in MFH-1 model are supposed to follow a normal distribution with zero

mean and constant variance σ_{ur}^2, $r = 1, 2$. If the underlying model assumptions hold, the residuals are supposed to be distributed randomly around zero. We used the normal probability (Q-Q) plots to examine the normality assumption. Q-Q plots of district level residuals corresponding to the two target variables Rural-Earning and Urban- Earning are given in Figure 2. In addition, we also examined the normality assumption of the random area effects via Shapiro-Wilk test and the p-values of the test are 0.321 and 0.457 for Rural-Earning and Urban-Earning respectively. Furthermore, it is evident from the Q-Q plot in Figure 2 that the normality assumption holds while p-values of the Shapiro-Wilk test are greater than 0.05 and both of these evidences taken together indicate that the district wise random area effects are likely to be distributed normally. Next we evaluated the validity and the reliability of the small area estimates by some frequently used diagnostics. In line with Brown et al. (2001) and Chandra et al. (2011), model-based small area estimates should be (1) consistent with unbiased direct survey estimates and (2) more efficient than direct estimates in terms of MSE. The subsequent measures e.g., the bias diagnostic, goodness of fit (GoF) diagnostic, the percentage coefficient of variation (CV) and the 95 percentage confidence interval (CI) are selected. Later, we classified the measurements of CV and CI as internal diagnostic measures as these indicate the efficiency of the small area estimates. Moreover, for benchmarking, a calibration diagnostic is also applied in which the model-based small area estimates are combined to an upper level so that these estimates can be compared with direct estimates at that higher level and we classified this as an external diagnostic measure. It is important to note that in this case, the direct estimates are in survey weighted form.

3.2.1. Bias Diagnostic

The bias diagnostic measure tests the validity while the precision of the model-based estimates is examined by the CI and CV. Following Chandra et al. (2011), the bias diagnostic is performed. Being unbiased of the population values, the regression of the direct estimates on the true population values likely to be linear with the identity line. If the model-based estimates are close to these true values of the population, the regression of direct estimates on model-based small area estimates expected to be similar. Consequently, we plotted the direct estimates and model-based estimates in the *y* and *x-axis* respectively and examined the departure of the small area estimates from the regression line fitted values. The plot given in Figure 3 demonstrates that small area

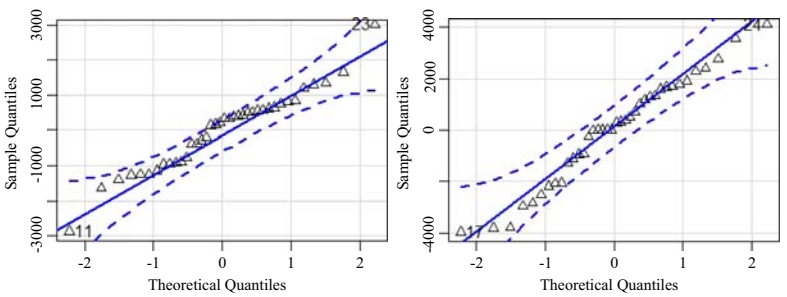


Fig. 2. Normal Q-Q plot of district-level residuals for rural (on the left) and urban (on the right).

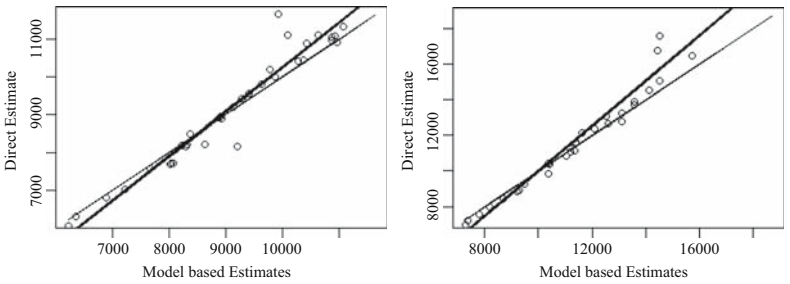


Fig. 3. Bias diagnostic plot with $y = x$ line (thin line) and regression line (solid black line) for rural (on the left) and urban (on the right).

estimates are not as extreme as the direct estimates signifying the usual SAE result of diminishing greater extreme values to the average values. Largely, this diagnostic specify that the small area estimates are expected to be consistent and approximately design unbiased when compared with direct estimates. This is expected as the MFH estimates are realization of random variables and so the regression of the direct estimates on the MFH estimates is unbiased for a test of common expected values. The Goodness of Fit (GoF) diagnostic provides such test. The statistic for GoF diagnostic is obtained using the Wald statistic for MFH estimate,

$$W = \sum_m \left[\frac{(direct\ estimate_m - MFH\ estimate_m)^2}{\hat{\text{v}}\text{ar}(direct\ estimate_m) + \hat{MSE}(MFH\ estimate_m)} \right]$$

The test statistic value is compared against the value from a chi-square distribution with M degrees of freedom. The value of chi square statistic with $M = 37$ degrees of freedom for rural districts is 52.19 and for urban districts is 47.40 with $M = 33$ degrees of freedom at 5% level of significance. We obtained the value of Wald statistic for rural districts as $W = 3.69$ and for urban districts it is $W = 5.10$. A smaller value than the table value indicates no significant difference between the direct and the model-based MFH estimates. This diagnostic results clearly indicate that the MFH estimates are consistent with the direct survey estimates.

3.2.2. Internal Diagnostic

Afterward, the degree of improvement in precision of model-based multivariate small area (i.e., district level) estimates of Rural-Earning and Urban-Earning are examined against the FH and direct survey estimates. Typically small area estimates having smaller CVs are likely to be reliable. The summary of %CVs of the Direct, FH and MFH estimates of Rural-Earning and Urban-Earning are given in Table 3. District specific %CV is demonstrated in Figure 4. The direct survey estimates possess greater CV compared to the FH and MFH estimates of Rural-Earning and Urban-Earning. It is obvious from Table 3 and Figure 4 that direct survey estimates of Rural-Earning and Urban- Earning seem to be highly unstable. It is important to note that, for the target variable Urban- Earning, there was 05 non-sample districts and for Rural-Earning, there was one non-sample district. So for Urban-Earning we first compare the performance with the sampled districts and comparison of non-sample districts are given separately. For Rural-Earning, the CV of

Table 3. Distribution of % CV for the direct and model-based small area estimates of rural-earning and urban-earning.

Values	Rural-earning (37 sample districts)			Urban-earning (33 sample districts)			Urban-earning (05 non-sample districts)		
	Direct	FH	MFH	Direct	FH	MFH	Direct	FH	MFH
Minimum	3.22	3.18	3.18	5.10	5.06	5.06	-	19.46	14.70
Q1	4.74	4.61	4.61	8.64	8.23	8.18	-	23.62	16.95
Median	6.86	6.34	6.34	9.57	9.02	8.93	-	23.68	18.09
Mean	7.77	6.57	6.56	10.37	9.48	9.38	-	23.36	17.23
Q3	8.60	7.93	7.92	10.69	10.22	10.16	-	24.61	18.19
Maximum	21.60	11.81	11.77	24.22	17.08	16.46	-	25.41	18.24

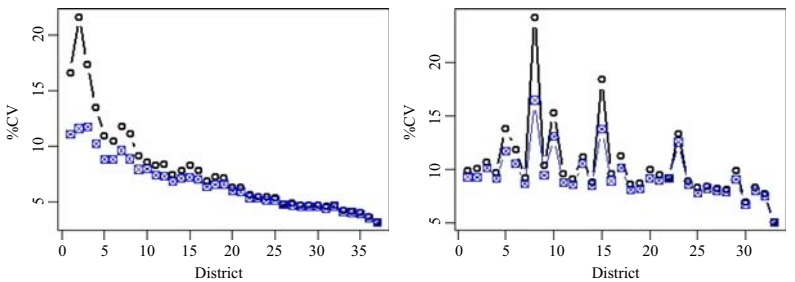


Fig. 4. District specific percentage coefficient of variation (CV) of direct (black, °) and MFH (blue, □) estimate for rural-earning (on the left) and urban-earning (on the right). Districts are arranged in increasing order of sample size.

Direct distributed form 3.22% to 21.60% with a median value of 6.86% whereas it is 3.18% to 11.77% with a median value of 6.34% for MFH which indicate that the MFH estimates are more strongly distributed compared to the direct estimates as the distribution of CVs from MFH estimates are much narrower than that of the direct estimates. Similarly for Urban-Earning, when we compare the Direct with MFH for the 33 sampled districts, the CV of Direct is distributed form 5.10% to 24.22% with a median value of 9.57% whereas it is 5.06% to 16.46% with a median value of 8.93% for MFH. It is important to note that, for sample districts the performance of FH and MFH are almost identical which in turn validate the point about the usefulness of MFH model discussed earlier in the model fitting section. Furthermore, when the performance of the FH and MFH estimates are compared for the non-sampled areas, there is fair improvement in both the rural and urban areas. For the single non-sample district of Rural-Earning, the CV of FH is 15.27% while it is 11.03% for MFH estimate and for Urban-Earning with 05 non-sampled district, the CV of FH is distributed from 19.46% to 25.41% with a median value of 23.68% whereas it is 14.70% to 18.24% with a median value of 18.09% for MFH estimates. As we have mentioned that, the input for the target variables in the multivariate model is a combination of the direct estimate and the synthetic estimate, the MFH model is utilized the extra information from the synthetic estimate apart from the information used by the FH model. It seems clear From Figure 4, with decreasing sample sizes of the districts the relative performance of the MFH estimates for Rural-Earning and Urban-Earning has improved. Accordingly, these precise and reliable MFH estimates generate the district level earning

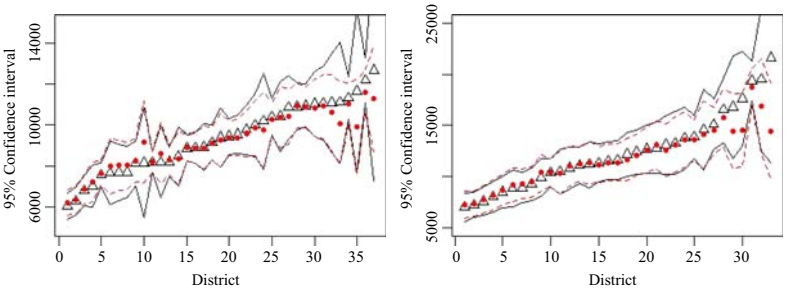


Fig. 5. District-wise 95% nominal confidence interval for direct (black) and MFH (red) estimates for rural-earning (on the left) and urban-earning (on the right). Districts are arranged in increasing order of direct estimates.

estimate much better than direct and FH estimates. The 95% confidence intervals (CIs) are given in Figure 5. The Figure 5 indicates that the CI of MFH estimate is much tighter than the direct survey estimates.

3.2.3. External Diagnostic

The aggregation property of the MFH based district-level SAE estimates at higher aggregation level (viz. state and regional level) are examined. The divisional and state-level estimates of Rural-Earning and Urban-Earning is obtained by

$$\hat{Y}_i = \sum_{j=1}^{M} N_{ij}\hat{Y}_{ij} / \sum_{j=1}^{M} N_{ij},\ i = 1, 2 \text{ and } j = 1, \ldots, M,$$

where $\hat{Y}_{ij}$ denote the MFH estimate of Rural-Earning and Urban-Earning for $i = 1, 2$ and district j and the population size, corresponding to the variable i and district j is N_{ij}. The districts of Bihar are classified in nine divisions and we studied the aggregation property. The state and the divisional-level estimates of Rural-Earning and Urban-Earning are reported in Table 4. While comparing the small area estimates against the direct estimates, it seems that in both the state and the regional level these small area estimates are close enough to the direct estimates.

Table 4. Aggregated estimates of rural-earning and urban-earning obtained from Direct and MFH. Estimates are aggregated over 37 districts for rural areas and 33 districts for urban areas at the state and divisional levels.

Division	Rural-earning		Urban-earning	
	Direct	MFH	Direct	MFH
State	9,149	9,201	14,478	13,971
Patna	9,709	9,841	11,247	11,567
Tirhut	8,754	9,147	11,413	11,365
Saran	8,080	8,180	12,643	11,970
Darbhanga	8,283	8,803	17,342	13,784
Kosi	8,148	8,289	13,740	13,397
Purnia	8,570	8,837	12,562	12,269
Bhagalpur	12,150	10,790	17,552	14,527
Munger	10,710	9,820	17,745	17,188
Magadh	10,722	10,745	11,174	11,369

3.3. Spatial Distribution of Earning Inequality

Table 5 and 6 report the distribution of CV and earning range across all the districts respectively. The district specific direct survey estimates and MFH estimates together with the 95% CI and CV for Rural-Earning and Urban-Earning are given in Table 7 and 8. The spatial maps of earning distribution (in Rs.) by districts for both rural and urban areas are produced for the district level estimates generated by MFH method. Figure 6 displays spatial maps of the MFH estimates of earning for rural and urban areas of Bihar. These spatial mapping assist in describing the magnitude of inequality in earning distribution between the district of rural and urban areas of the state. In case of rural areas, central part of Bihar has lower earning followed by the eastern and western part. For urban areas, the lower earning level exists in eastern part followed by the western and central part. The average monthly earning is ranging from Rs. 6,231 to Rs. 11,634 in rural areas whereas it is Rs. 7,279 to Rs. 18,727 in urban areas. This clearly indicates that there is a huge difference in average monthly earning between rural and urban areas in Bihar. From Table 5, the distribution of CVs indicate the precision of the MFH estimates compared to the direct estimates. In rural areas of Bihar, 08 districts have more than 10% CV of direct estimates while it is only 04 districts for FH and MFH estimates and in case of urban sector, 12 districts have more than 10% CV whereas it is only 09 districts for FH and MFH estimates. Based on the results obtained in Table 5, we further described the distribution of earning range based on precise MFH estimates in Table 6. From Table 6 and Figure 6, MFH estimates also reveal that number of districts having average monthly earning more than Rs. 10,000 is only 12 in rural areas while it is 30 for urban areas. In case of lower earning level, 26 districts in rural areas have average monthly earning of Rs. 10,000 or less whereas it is only 08 districts for urban areas. Table 6 further described that all of the rural areas possess an average monthly earning of Rs. 12,500 or less. However, nearly 40% of

Table 5. Distribution of CV of 37 districts in rural areas and 33 districts in urban areas.

Group	%CV	Rural-earning			Urban-earning		
		Direct	FH	MFH	Direct	FH	MFH
1	$< = 5$	12	12	12	0	0	0
2	5.01–10	17	21	21	21	24	24
3	10.01–15	05	04	04	09	08	08
4	15.01–20	02	0	0	02	01	01
5	$> = 20$	01	0	0	01	0	0

Table 6. Distribution of earning range of 38 districts in rural and urban areas based on MFH estimates.

Group	Earning range (in Rs.)	Rural-earning	Urban-earning
1	$< = 7{,}500$	04	02
2	7,501–10,000	22	06
3	10,001–12,500	12	15
4	12,501–15,000	0	12
5	$> = 15{,}000$	0	03

Table 7. Direct and MFH estimates along with 95% confidence interval (95 % CI) and percentage coefficient of variation (CV) of the target variable rural-earning by district in Bihar.

		District				MFH			
			95% CI				95% CI		
District	Sample size	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Pashchim Champaran	156	8,912	8,188	9,636	4.15	8,899	8,192	9,606	4.05
Purba Champaran	154	10,423	9,565	11,281	4.20	10,275	9,451	11,099	4.09
Sheohar	67	7,038	6,012	8,064	7.44	7,234	6,262	8,207	6.86
Sitamarhi	163	7,588	6,986	8,190	4.05	7,676	7,086	8,266	3.92
Madhubani	56	7,720	6,463	8,977	8.31	8,050	6,882	9,218	7.40
Supaul	34	10,198	7,843	12,553	11.78	9,777	7,932	11,623	9.63
Araria	140	8,223	7,486	8,960	4.57	8,326	7,607	9,044	4.40
Kishanganj	136	9,192	8,347	10,037	4.69	9,124	8,312	9,936	4.54
Purnia	80	8,160	7,063	9,257	6.86	8,298	7,266	9,329	6.34
Katihar	55	8,499	7,066	9,932	8.60	8,382	7,075	9,688	7.95
Madhepura	114	6,053	5,409	6,697	5.43	6,231	5,599	6,863	5.17
Saharsa	135	9,459	8,583	10,335	4.72	9,389	8,548	10,231	4.57
Darbhanga	107	6,806	6,057	7,555	5.62	6,902	6,174	7,630	5.38
Muzaffarpur	92	8,952	7,838	10,066	6.35	8,915	7,869	9,960	5.98
Gopalganj	131	9,559	8,650	10,468	4.85	9,411	8,535	10,287	4.75
Siwan	205	8,207	7,689	8,725	3.22	8,223	7,709	8,736	3.19
Saran	112	6,303	5,627	6,979	5.47	6,368	5,704	7,032	5.32
Vaishali	68	9,414	7,965	10,863	7.85	9,288	7,990	10,585	7.13
Samastipur	83	10,881	9,349	12,413	7.18	10,427	9,077	11,777	6.61
Begusarai	7	8,175	5,511	10,839	16.63	9,198	7,194	11,202	11.12
Khagaria	28	11,121	8,171	14,071	13.53	10,086	8,063	12,109	10.24
Bhagalpur	16	11,673	7,698	15,648	17.37	9,922	7,633	12,210	11.77
Banka	140	12,201	11,078	13,324	4.70	11,634	10,582	12,686	4.61
Munger	137	11,340	10,286	12,394	4.74	11,065	10,068	12,062	4.60
Lakhisarai	95	9,805	8,592	11,018	6.31	9,623	8,503	10,743	5.94
Sheikhpura	65	10,432	8,720	12,144	8.37	10,374	8,889	11,859	7.30
Nalanda	68	11,089	9,279	12,899	8.33	10,931	9,378	12,484	7.25
Patna	133	10,958	9,915	12,001	4.86	10,869	9,880	11,858	4.64
Bhojpur	46	7,715	6,333	9,097	9.14	8,072	6,819	9,326	7.92
Buxar	178	8,897	8,259	9,535	3.66	8,921	8,297	9,545	3.57
Kaimur (Bhabua)	33	7,687	6,103	9,271	10.52	8,021	6,636	9,405	8.81
Rohtas	71	10,009	8,481	11,537	7.79	9,865	8,513	11,216	6.99
Aurangabad	32	8,212	6,456	9,968	10.91	8,636	7,139	10,134	8.85
Gaya	10	12,699	7,322	18,076	21.60	11,283	8,716	13,850	11.61
Nawada	82	11,051	9,473	12,629	7.29	10,862	9,459	12,265	6.59
Jamui	118	10,906	9,750	12,062	5.41	10,962	9,864	12,060	5.11
Jehanabad	42	11,112	8,688	13,536	11.13	10,622	8,777	12,467	8.86
			Non-sample District						
Arwal	0	-	-	-	-	8,981	7,039	10,924	11.03

Table 8. Direct and MFH estimates along with 95% confidence interval (95 % CI) and percentage coefficient of variation (CV) of the target variable urban-earning by district in Bihar.

		District				MFH			
			95% CI				95% CI		
District	Sample size	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Pashchim Champaran	29	9,251	7,679	10,823	8.67	9,458	7,941	10,975	8.18
Purba Champaran	89	10,384	8,971	11,797	6.94	10,404	9,031	11,777	6.73
Sitamarhi	8	10,463	8,271	12,655	10.69	10,367	8,302	12,432	10.16
Madhubani	18	8,468	6,957	9,979	9.10	8,702	7,240	10,165	8.57
Araria	24	13,744	10,708	16,780	11.27	13,598	10,888	16,308	10.17
Kishanganj	32	12,790	10,276	15,304	10.03	13,119	10,761	15,477	9.17
Purnia	102	15,108	12,654	17,562	8.29	14,518	12,249	16,787	7.97
Katihar	9	7,988	6,472	9,504	9.68	8,192	6,722	9,661	9.15
Madhepura	7	8,831	7,082	10,580	10.10	9,199	7,529	10,870	9.26
Saharsa	15	7,573	6,212	8,934	9.17	7,788	6,460	9,115	8.70
Darbhanga	52	11,591	9,676	13,506	8.43	11,392	9,555	13,229	8.23
Muzaffarpur	83	16,513	13,325	19,701	9.85	15,723	12,934	18,512	9.05
Siwan	15	21,609	11,351	31,867	24.22	14,446	9,785	19,107	16.46
Saran	39	12,670	10,300	15,040	9.55	12,605	10,398	14,812	8.93
Vaishali	41	12,167	9,978	14,356	9.18	11,627	9,549	13,705	9.12
Samastipur	22	19,518	12,485	26,551	18.38	16,926	12,366	21,486	13.74
Begusarai	28	8,904	7,396	10,412	8.64	9,272	7,807	10,737	8.06
Bhagalpur	102	13,891	11,791	15,991	7.71	13,572	11,581	15,563	7.48
Banka	6	6,995	5,641	8,349	9.88	7,279	5,952	8,605	9.30
Munger	16	16,782	11,753	21,811	15.29	14,440	10,746	18,134	13.05
Lakhisarai	50	12,407	10,250	14,564	8.87	12,107	10,082	14,132	8.53
Sheikhpura	9	14,565	10,622	18,508	13.81	14,120	10,878	17,362	11.72
Nalanda	42	17,610	13,010	22,210	13.33	14,535	10,964	18,106	12.53
Patna	362	19,354	17,418	21,290	5.10	18,727	16,868	20,586	5.07
Bhojpur	59	11,530	9,690	13,370	8.14	11,390	9,624	13,156	7.91
Buxar	17	9,859	8,009	11,709	9.58	10,389	8,612	12,166	8.73
Kaimur (Bhabua)	20	13,101	10,241	15,961	11.14	12,540	9,960	15,120	10.50
Rohtas	51	11,070	9,275	12,865	8.27	11,195	9,480	12,910	7.82
Aurangabad	14	13,251	10,169	16,333	11.87	13,128	10,420	15,836	10.52
Gaya	54	11,317	9,488	13,146	8.25	11,159	9,412	12,906	7.99
Nawada	21	7,214	5,968	8,460	8.81	7,346	6,119	8,572	8.52
Jamui	22	10,847	8,797	12,897	9.64	11,037	9,109	12,965	8.91
Jehanabad	15	11,174	8,907	13,441	10.35	11,369	9,257	13,481	9.48
			Non-sample Districts						
Sheohar	0	-	-	-	-	13,923	9,912	17,934	14.70
Supaul	0	-	-	-	-	11,443	7,641	15,245	16.95
Gopalganj	0	-	-	-	-	11,487	7,379	15,595	18.24
Khagaria	0	-	-	-	-	10,665	6,862	14,468	18.19
Arwal	0	-	-	-	-	11,011	7,107	14,915	18.09

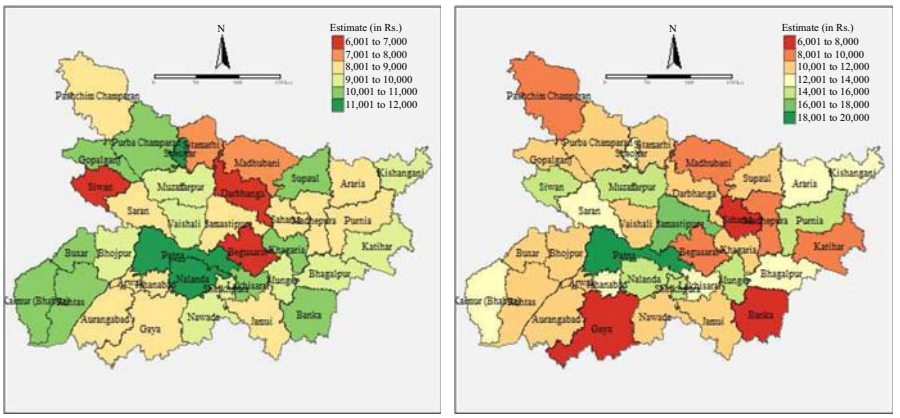


Fig. 6. Model-based MFH estimates displaying the spatial distribution of earning inequality by district between rural (on the left) and urban (on the right) districts in Bihar.

urban areas hold an average monthly earning of more than Rs. 12,500. Taken together, it is evident from these results that the degree of inequality in earning distribution between rural and urban districts is extremely severe. The difference in average monthly earning between rural and urban areas of Bihar can be obtained from Table 7–Table 8 and we may conclude that out of 38 districts in Bihar, 04 districts in rural areas are having earning higher than urban areas. But this seems not be the case as the direct estimates for urban areas in these 04 districts are truly unstable with higher CV percentage. Districts viz. Katihar, Saharsa, Banka and Gaya indicate higher level of earning in rural areas compared to urban. Sample sizes for rural areas in these districts also indicate that these particular districts covered more rural parts than the urban areas. These spatial maps and results provide useful information to policymakers in effective policy formulation and financial planning.

3.4. Distribution of Average Monthly Earning and Earning Inequality

In what follows, we tried to get a closer look into the degree of inequality between rural and urban areas. Figure 7 demonstrates two bivariate plots showing the spatial distribution of average monthly earning and Theil index of inequality by district in rural and urban areas in Bihar. We have used the Theil index to measure the degree of earning inequality as this index is perhaps best suited to measure the economic inequality. The U.S. Census Bureau (2016) mentioned that "the Theil index measures an entropic "distance" the population is away from the "ideal" egalitarian state of everyone having the same income." We have calculated the Theil index separately for Rural-Earning and Urban-Earning as

$$T_i = \sum_{j=1}^{M} \frac{\hat{Y}_{ij}}{\hat{Y}_i} T_{ij} + \sum_{j=1}^{M} \frac{\hat{Y}_{ij}}{\hat{Y}_i} \ln\left(\frac{\hat{Y}_{ij}/\hat{Y}_i}{N_{ij}/N_i}\right),$$

where $\hat{Y}_{ij}$ denote the MFH estimate of Rural-Earning and Urban-Earning for $i = 1, 2$ and district j and the population size corresponding to the variable i and district j is N_{ij} with $N_i = \sum_{j=1}^{M} N_{ij}$. Here $\hat{Y}_i$ is defined in Subsection 3.2.3 and T_{ij} is the Theil index

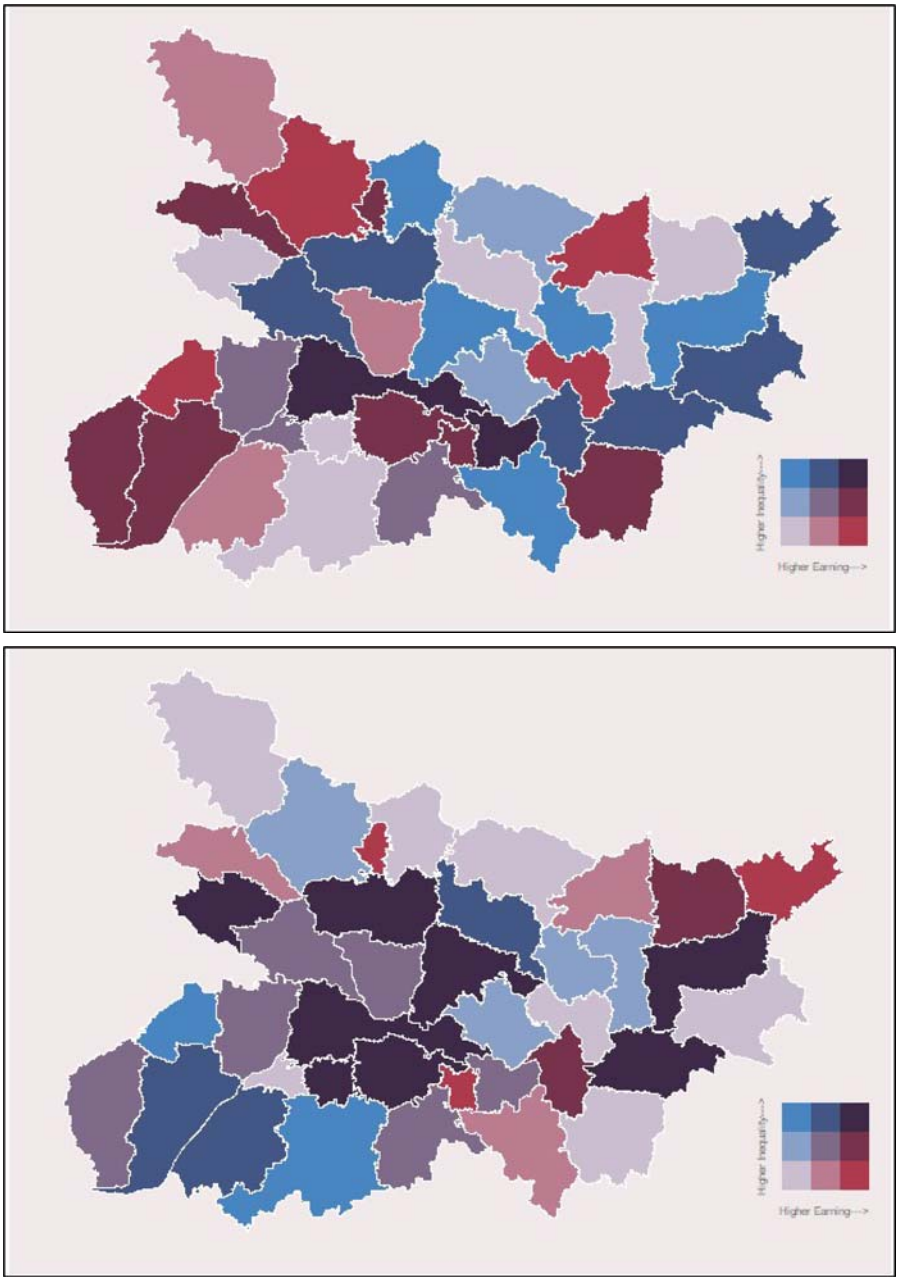


Fig. 7. Bivariate plot showing the spatial distribution of average monthly earning and Theil index of earning inequality by district in rural (at the top) and urban (at the bottom) districts in Bihar.

corresponding to the variable *i* and district *j* obtained from the original survey data. From Figure 7, the areas in grey indicate low earning and low inequality, blue areas show high earning and low inequality, areas in violate indicate high earning and high inequality and lastly the red areas specify low earning and high inequality. These bivariate spatial maps in Figure 7, also reveal that degree of inequality is increased with earning level and

consequently urban areas indicate greater degree of inequality than its counterpart. These bivariate maps offer worthy evidence to administrators and decision makers in prioritizing the areas more attention.

4. Conclusion

According to UNDP (2015), during 1990–2010 earning inequality has increased by 11% in developing countries and in India, 27.5% of the population are multidimensionally poor. World Bank (2020) mentioned that India accounts for 139 million of the total 689 million people (20.17%) living in extreme poverty in 2017 while its population being 17.8% of the world population (World Bank 2019) which indicates that India's share of the world's extreme poor population is higher than its share of the world population. Similar to the major developed economies around the globe, the COVID-19 pandemic has also hit the India's economy very hard with the loss of millions of jobs, causing in considerably reduced household incomes and extreme poverty. Due to the socioeconomic and health crisis in this pandemic, India's economy has experienced the biggest annual contraction of 7.3% in its gross domestic product (MoSPI 2021) since independence.

As part of the 2030 Agenda for sustainable development, the target 1 of the 10th goal aims that "By 2030, progressively achieve and sustain income growth of the bottom 40% of the population at a rate higher than the national average". From Table S2 in the online supplementary material, we can dig deeper to see the past pattern of earning inequality in India. The share in total earning of the top 1% declined gradually from 13% (1961) to 6.9% (1981) but after the major economic reforms in 1990s, it is continuously increasing from 10.4% (1991) to 21.7% (2019). The share in total earning of the bottom 50% remained fairly constant from 19.3% in 1961 to 21.4% in 1981. Afterward, it started declining from 19.8% in 1991 to 13.1% in 2019. To implement the agenda of sustainable development, India currently lacks the critically essential disaggregate level measures and maps of localized earning inequality.

In this study, the MFH model and its corresponding empirical best linear unbiased predictor for estimating small area means along with MSE estimation are summarized. Then, we applied this method in the 2018–2019 PLFS data of NSO, Govt. of India to produce the district level estimates and spatial mapping of earning inequalities in rural and urban areas of the state of Bihar in India. For the selection of suitable covariates for MFH model, data from 2011 Population Census of India are used and we applied PCA followed by a stepwise regression technique for selecting significant covariates. Guha and Chandra (2021b) described that multivariate modelling approach incorporating the correlation between the target variables often leads to efficient estimation of small area parameters of correlated measures. However, following Franco and Bell (2021) it should also be noted that the gain in precision for multivariate models can only be realized if one of the variables under study has low variance and the target variables are strongly correlated. Otherwise, both FH and MFH models may produce identical results. In this analysis, both the target variables viz. Rural-Earning and Urban-Earning are jointly modelled via MFH model and the gain is achieved in terms of efficiency (i.e., MSE) and CV in estimating the district level estimates of rural and urban earning inequalities. These estimates related to earning inequalities across the state of Bihar can assist in initiating the dialogue about the

drivers of earning inequalities in this state. Various diagnostic methods were used to assess the model-based MFH estimates and it also reveals significant gains in efficiency in producing district level estimates of earning which consequently measures the distribution of inequality that persist between rural and urban areas in the state. Moreover, the spatial maps so obtained show the evidence of unequal earning distribution and uneven distribution of inequality across the districts of rural and urban areas in Bihar. In rural areas, districts such as Patna, Lakhisarai exhibit higher level of earning and very high level of inequality whereas in urban areas, districts like Patna, Nalanda, Muzaffarpur, Samastipur, Purnia, Bhagalpur revealed higher level of earning as well as very high level of inequality in the state. When we tried to get a closer look on the degree of inequality between the rural and urban areas, we found that 08 districts in urban areas indicate very high level of inequality compared to only 02 districts in rural areas which eventually shows that though the earning level is much higher in urban areas, the degree of inequality is also far greater than the rural areas in Bihar.

This analysis undoubtedly established the advantages of SAE approach to deal with the problem of small sample sizes in obtaining precise and cost effective disaggregate or local level estimates along with the confidence intervals from existing survey data, (Guha and Chandra 2021b). In addition, this study also reveals the advantages of MFH over FH model in case of nonsample districts. This analysis established the fact that many areas in rural sector of Bihar possess very low level of earning compared to the urban sector while the degree of inequality is much higher in urban areas. The NSO surveys of Government of India are intended for obtaining state and national level estimates and these surveys do not reveal the real situation at the micro level (for example block or district level). However, substantial importance is given on micro level planning by the Government of India for realizing a stable economic development together with earning generation. For definite planning and development in a country, district is always an important unit and thus availability of district-level data and statistics are much vital for planning and monitoring of policy action plans. These cost effective and precise model-based estimates together with spatial maps may be useful for various Ministries and Departments of Government of India along with International organizations for effective policy planning and monitoring related to sustainable development goal ten – reduced inequalities. This study can be useful in obtaining the district level estimates and examine the inequality in earning distribution in the remaining parts of the country. Moreover, as earning data are generally skewed in nature, authors are working on this problem in multivariate SAE framework.

5. References

Benavent, R., and D. Morales. 2016. "Multivariate Fay-Herriot models for small area estimation". *Computational Statistics and Data Analysis* 94: 372–390. DOI: https://doi.org/10.1016/j.csda.2015.07.013.

Brown, G., R. Chambers, P. Heady, and D. Heasman. 2001. "Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS". In Proceedings of Statistics Canada Symposium 2001. Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Ottawa: Statistics Canada. Available

at: https:// www150.statcan.gc.ca/n1/pub/11-522-x/2001001/session6/6247-eng.pdf (accessed April 2021).

Census. 2011. Primary Census Abstracts, Registrar General of India, Ministry of Home Affairs, Government of India. Available at: https://www.censusindia.gov.in/2011-Common/CensusData2011.html (accessed April 2021).

Chandra, H., N. Salvati, and U.C. Sud. 2011. "Disaggregate-level Estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique". *Journal of Applied Statistics* 38(11): 2413–2432. DOI: https://doi.org/10.1080/02664763.2011.559202.

Datta, G.S., R.E., Fay, and M. Ghosh. 1991. "Hierarchical and empirical Bayes multivariate analysis in small area estimation". In Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, January, Washington, D.C., U.S.A.: 63–79 Available at: https://play.google.com/books/reader?id = Lr3rAAAAMAAJ&pg = GBS.PA64&hl = en (accessed April 2021).

Datta, G., T. Kubokawa, I. Molina, and J.N.K. Rao. 2011. "Estimation of mean squared error of model-based small area estimators". *TEST* 20(2): 367–388. DOI: https://doi.org/10.1007/s11749-010-0206-2.

Fay, R.E., and R. Herriot. 1979. "Estimates of income for small places: an application of James stein procedures to census data". *Journal of the American Statistical Association* 74: 269–277.DOI: https://doi.org/10.2307/2286322.

Fay, R.E. 1987. "Application of multivariate regression of small domain estimation". In *Small Area Statistics,* edited by R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh: 91–102. New York: Wiley. DOI: https://doi.org/10.1002/bimj.4710300737.

Franco, C., and W.R. Bell. 2021. "Using American community survey data to improve estimates from smaller U.S. surveys through bivariate small area estimation models". *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smaa040.

GHI. 2020. *Global Hunger Index 2020*. Available at: https://www.globalhungerindex.org/results.html (accessed September 2021).

Guha, S., and H. Chandra. 2021a. "Measuring and Mapping Disaggregate Level Disparities in Food Consumption and Nutritional Status via Multivariate Small Area Modelling". *Social Indicators Research* 154(2): 623–646. DOI: https://doi.org/10.1007/s11205-020-02573-8.

Guha, S., and H. Chandra. 2021b. "Measuring disaggregate level food insecurity via multivariate small area modelling: evidence from rural districts of Uttar Pradesh, India". *Food Security* 13(3): 597–615. DOI: https://doi.org/10.1007/s12571-021-01143-1.

Johnson, F.A., H. Chandra, J. Brown, and S. Padmadas. 2010. "Estimating district-level births attended by skilled attendants in ghana using demographic health survey and census data: an application of small area estimation technique". *Journal of Official Statistics* 26(2): 341–359. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/district-level-estimates-of-institutional-births-in-ghana-application-of-small-area-estimation-technique-using-census-and-dhs-data.pdf

MoHFW. 2021. *COVID-19 State wise Status*. Available at: https://www.mohfw.gov.in (accessed September 2021).

MoSPI. 2020a. *Annual Report: PLFS, 2018–19*. Government of India, Ministry of Statistics and ProgrammeImplementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/publicationreports/Annual_Report_PLFS_2018_19_HL.pdf (accessed April 2021).

MoSPI. 2020b. *Quarterly Bulletin, PLFS: April–June* 2020. Government of India, Ministry of Statistics and Programme Implementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/publication_reports/PLFS_Quarterly_Bulletin_April-June_2020.pdf (accessed September 2021).

MoSPI. 2021. Press note: *On estimates of gross domestic product for the first quarter (april − june) 2021–2022*. Government of India, Ministry of Statistics and Programme Implementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/press-release/PRESS_NOTE-Q1_2021-22.pdf (accessed September 2021).

NITI Aayog. 2019. *SDG India Index*. Available at: https://sdgindiaindex.niti.gov.in (accessed September 2021).

Prasad, N.G.N., and J.N.K. Rao. 1990. "The estimation of the mean squared error of small-area estimators". *Journal of the American Statistical Association* 85: 163–171. DOI: https://doi.org/10.2307/2289539.

Rao, J.N.K., and I. Molina. 2015. Small area estimation. New York: Wiley. DOI: https://doi.org/10.1002/9781118735855.

Särndal, C.E., B. Swensson, and J.H. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer-Verlag.

Schall, R. 1991. "Estimation in generalized linear models with random effects". *Biometrika* 78(4): 719–727. DOI: https://doi.org/10.2307/2336923.

UNCDP. 2018. *Official Records of the Economic and Social Council, 2018*. Available at: https://undocs.org/pdf?symbol = en/E/2018/33 (accessed September 2021).

UNDP. 2015. *Sustainable Development Goals*. Available at: https://www.undp.org (accessed April 2021).

U.S. Census Bureau. 2016. *Theil Index*. Available at: https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html (accessed September 2021).

World Bank. 2019. Open Data. Available at: https://www.worldbank.org/en/home (accessed September 2021).

World Bank. 2020. Open Data. Available at: https://www.worldbank.org/en/home (accessed September 2021).

Received March 2021
Revised September 2021
Accepted January 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 847–873, http://dx.doi.org/10.2478/JOS-2022-0037

Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM)

Daniel Hopp[1]

Artificial neural networks (ANNs) have been the catalyst to numerous advances in a variety of fields and disciplines in recent years. Their impact on economics, however, has been comparatively muted. One type of ANN, the long short-term memory network (LSTM), is particularly well-suited to deal with economic time-series. Here, the architecture's performance and characteristics are evaluated in comparison with the dynamic factor model (DFM), currently a popular choice in the field of economic nowcasting. LSTMs are found to produce superior results to DFMs in the nowcasting of three separate variables; global merchandise export values and volumes, and global services exports. Further advantages include their ability to handle large numbers of input features in a variety of time frequencies. A disadvantage is the stochastic nature of outputs, common to all ANNs. In order to facilitate continued applied research of the methodology by avoiding the need for any knowledge of deep-learning libraries, an accompanying Python (Hopp 2021a) library was developed using PyTorch. The library is also available in R, MATLAB, and Julia.

Key words: Forecasting; machine learning; python.

1. Introduction

A defining feature of the 21st century so far has been the explosion in both the volumes and varieties of data generated and stored (Domo 2017). Almost every industry and aspect of life has been affected by this "data revolution" (Einav and Levin 2014; MacFeely 2020). Simultaneously, rapid advancements in machine learning methods have been made, spurred on in part by the need for novel methods to analyze these new data quantities. Perhaps no methodology has gained greater prominence than the artificial neural network (ANN). ANNs are the engine behind tremendous leaps in fields as disparate as machine translation, image recognition, recommendation engines, and even self-driving vehicles. Yet to date, their impact in the field of economic policy has been largely muted or exploratory in nature (Falat and Pancikova 2015).

This is not to suggest that economic data have been immune to the transformative forces of the data revolution. Quite the opposite in fact, as classical economic data series from national statistical offices (NSO) and other organizations can now be fortified by alternative data sources like never before, helping to provide glimpses into the developments of the global economy with unparalleled granularity and timeliness. The

[1]UNCTAD E-9042, Palais des Nations, 1211 Geneva 10, Switzerland. Email: daniel.hopp@un.org
Acknowledgments: I would like to thank Fernando Cantú, Anu Peltola, Onno Hoffmeister, Henri Luomaranta and Stephen MacFeely for their valuable comments and support.

© Statistics Sweden

COVID-19 pandemic and ensuing economic crisis showcased this, with analysts and policy-makers gaining insight to the rapidly evolving economic situation from such alternative data sources as Google mobility data (Yilmazkuday 2021), booking information from dining apps (OpenTable 2021), and transaction data from e-commerce sites (Statista 2021), among many others.

The availability of a broad range of novel, timely indicators should ostensibly have led to significant advances in the field of economic nowcasting, where real-time macroeconomic variables that may be published with a significant lag are estimated based on an array of more timely indicators (Banbura et al. 2010; Giannone et al. 2008). In reality, the field has not experienced the degree of progress seen in other fields, such as image recognition, in the past ten years. A large factor in this relative stagnation is the fact that many of the issues facing nowcasting are not addressed by more data alone. Issues such as multicollinearity, missing data, mixed-frequency data, and varying publication dates are sometimes even exacerbated by the addition of variables (Porshakov et al. 2016). As such, advancements in the field come from a combination of both new data and methodological developments. Dynamic factor models (DFM) in particular have been found to address many of the data issues inherent in nowcasting (Stock and Watson 2002), and have been applied successfully in applications such as nowcasting economic growth in 32 countries (Matheson 2011), nowcasting German economic activity (Marcellino and Schumacher 2010), and nowcasting Canadian GDP growth (Chernis and Sekkel 2017). The basic premise of DFMs is that one or more latent factors dictates the movement of many different variables, each with an idiosyncratic component in relation to the factor(s). With historical data, the factor(s) can be estimated from the variables. Subsequently, even in future periods where not all data are complete, the factor(s) can still be estimated and used to generate forecasts for variables that are not yet published, as each variable's relation to the factor(s) has already been estimated.

Despite DFMs' strengths in addressing a wide swath of nowcasting's data issues, the impressive performance of ANNs in other domains raises the question of their performance in nowcasting. ANNs have been applied to economic nowcasting in the past (Loermann and Maas 2019). However, due to the time-series nature of many economic nowcasting applications, the long short-term memory (LSTM) architecture is better suited to the problem than the traditional feedforward architecture explored in Loermann and Maas (2019). LSTMs are an extension of recurrent neural network (RNN) architecture, which introduces a temporal component to ANNs. LSTMs have been used to nowcast meteorological events (Shi et al. 2015) as well as GDP (Kurihara and Fukushima 2019).

However, use of LSTMs in nowcasting economic variables remains in its infancy, perhaps partly due to high barriers to their implementation. Many common deep learning frameworks, including Keras and PyTorch, include provisions for LSTMs. However, the implementations are general and require knowledge of the frameworks to successfully implement. As such, a Python library focused on economic nowcasting has been published alongside this article, available for install on PyPi (Hopp 2021a) along with wrappers for R, MATLAB, and Julia. Hopefully, a more accessible library will help stimulate interest and expand the applications of these models. More information on the libraries is available at:

- Python (Hopp 2021a)
- R (Hopp 2021b)
- MATLAB (Hopp 2021c)
- Julia (Hopp 2021d)

The remainder of this article is structured as follows: the next section will further explain nowcasting and its challenges; Section 3 will explore ANNs and LSTMs in more detail; Section 4 will examine the LSTM's empirical performance compared with DFMs in nowcasting three series: global merchandise trade exports expressed in both values and volumes and global services exports; the final section will conclude and examine areas of future research.

2. Exposition of Nowcasting Problem

Nowcasting, a portmanteau of "now" and "forecast", is the estimation of the current, or near to it either forwards or backwards in time, state of a target variable using information that is available in a timelier manner. Keith Browning coined the term in 1981 (WMO 2017) to describe forecasting the weather in the very near future based on its current state. The concept and term remained in the meteorological domain for years before being adopted into the economic literature in the 2000s. The concept of real-time estimates of the macroeconomic situation predates the adoption of the nowcasting terminology, as evidenced by Mariano and Murasawa (2003). However, Giannone et al. (2005) explicitly referenced the term "nowcasting" in its title and the term became commonplace in subsequent years, being applied for example to Portuguese GDP in 2007 (Morgado et al. 2007) and to Euro area economic activity in 2009 (Giannone et al. 2009). The 2010s saw a wealth of papers examining the topic both for a range of target variables as well as with a range of methodologies. Targets most often included GDP (Rossiter 2010; Bok et al. 2018), and trade (Cantú 2018; Guichard and Rusticelli 2011). Common methodologies include dynamic factor models (DFM) (Guichard and Rusticelli 2011; Antolin-Diaz et al. 2020), mixed data sampling (MIDAS) (Kuzin et al. 2009; Marcellino and Schumacher 2010) and mixed-frequency vector autoregression (VAR) (Kuzin et al. 2009), among others. Nowcasting also has relevance in the context of the 2030 Agenda for Sustainable Development (UN 2015). Many indicators face issues in terms of data quality, availability, timeliness, or all three. As such, nowcasting is being discussed as a possible method of ensuring maximum coverage in terms of indicators (UNSD 2020).

Economic nowcasting is generally confronted with three main issues regarding data. The first is mixed frequency data, or when all independent variables and the dependent variable are not recorded with the same periodicity. This occurs frequently in economic data, for instance when trying to nowcast a quarterly target variable, such as GDP growth, using monthly indicators, or estimating a yearly target variable with a mixture of monthly and quarterly variables. The second is the heterogeneous publication schedules of independent variables, frequently referred to as "ragged edges". Any nowcasting methodology must provide provisions for incomplete or partially complete data, as varying availability of latest data is the reality of most data sets of economic series. Finally, there is the issue of the "curse of dimensionality", which renders many classical econometric methods less effective in the nowcasting context and hinders the application

of "big data" to the field (Buono et al. 2017). The problem stems from the nature of many economic variables, where they may have few observations relative to the potential pool of explanatory variables or features. The quarterly target series for the United Nations Conference on Trade and Development's (UNCTAD) nowcasts for global merchandise trade, for instance, only began in 2005 (Cantú 2018). That leaves only 60 observations for training a model at the end of 2020. Meanwhile, many more than 60 potential independent variables can be conceived of to estimate a model of global merchandise trade.

The nowcasting methodologies previously mentioned address these problems in varying ways to achieve better predictions, and LSTMs are no different. The following section will provide background information on their network architecture as well as the characteristics that allow them to address the aforementioned nowcasting data problems. Those interested in an even more comprehensive examination of neural networks should see Gurney (1997).

3. Artificial Neural Networks and Long Short-Term Memory Networks

3.1. Artificial Neural Networks and Recurrent Neural Networks

3.1.1. Layers, Nodes, and Weights

ANNs are made up of various inter-connected layers composed of groups of nodes or neurons. Figure 1 shows one of the simplest forms of neural network, a dense, single layer

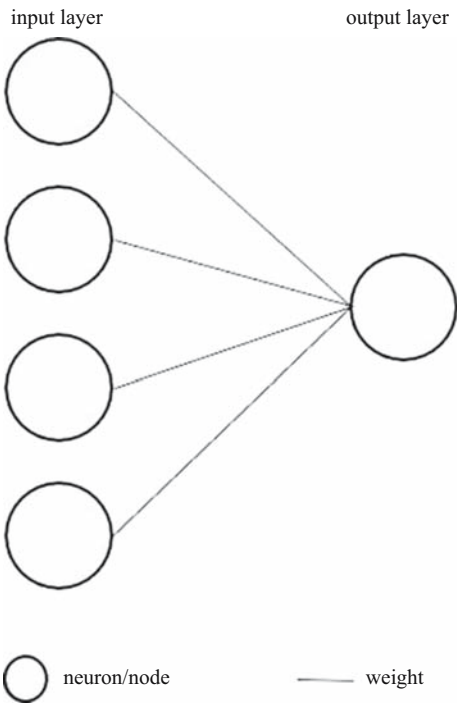


Fig. 1. A single layer feedforward neural network.

feedforward network. Feedforward means that signal or information flows only one way through the network, from input, then through any intermediate layers, finally to output. Dense, or fully connected, means that each node of every layer is connected to each node of the next layer. This distinction is superfluous in the single layer case, but relevant in more complex network architectures.

This simple network can be interpreted in the following manner: four input variables, represented by the four nodes in the input layer, are multiplied by four coefficients or weights, represented by the solid lines connecting each input node to the output node, then summed to obtain an output or prediction. This sounds similar to linear regression because, in this simplified case and in the absence of an activation function, which will be discussed below, it essentially is. The output layer can have more than one node, for instance in cases of categorical classification, but only the single output layer node case is relevant for the regression application in this article.

Figure 2 shows a more complex network architecture, where one hidden layer is added with two additional nodes. Now, coefficients exist between both the initial input layer and the hidden layer as well as between the hidden layer and the final output layer. Usually, no good semantic interpretation of hidden layers exists as they become an abstracted amalgamation of previous layers. They are best thought of as intermediate processing layers which help the network approximate the target function. However, post analysis on a trained network can sometimes lead to human-interpretable meanings, identifying

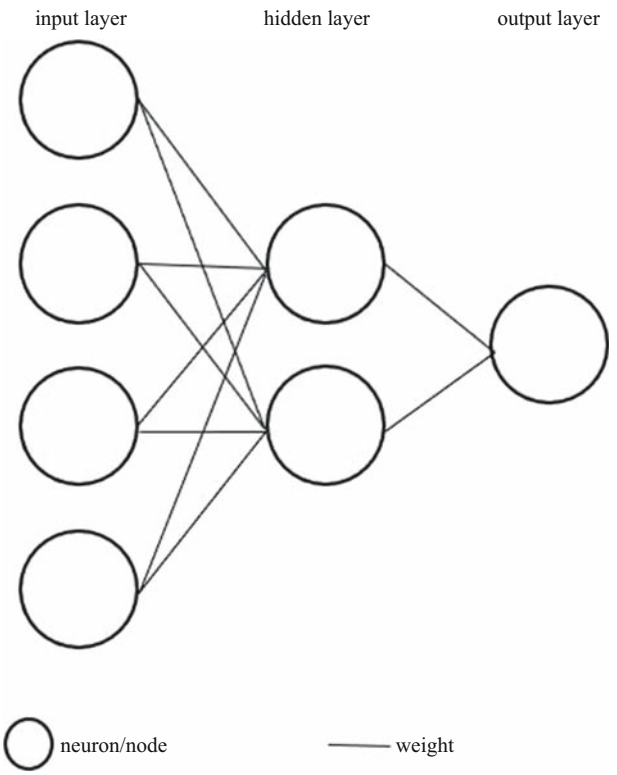


Fig. 2. A multi-layer feedforward neural network with one hidden layer

hidden layers or neurons associated with the identification of say eyes or textures in a photograph, or with macro-concepts like investment or developing economy performance in an economic context. Those interested in learning more about the interpretation of how and what neural networks learn should see Subsection 10.1 of Molnar (2019). The formula for an individual node on the hidden layer is below.

$$H_j = \sum_{i=1}^{n} w_i v_i \tag{1}$$

where:

H_j = the value of hidden node j
n = the number of nodes in the previous layer connected to the hidden node
w = the weight or coefficient between the previous layer's node and the hidden node
v = the value or output of the previous layer's node

3.1.2. Activation Functions

The simple networks described above contain a rather large drawback, the fact that they can only represent linear relationships. The ability to approximate complex, non-linear relationships is one of the defining characteristics of ANNs and an essential component of their predictive power. This component is introduced by means of a non-linear activation function. While there are many different types of activation function, see Sharma et al. (2020) for an in depth explanation of several, a commonly used one is the Rectified Linear Unit, ReLU for short. It will be used here as the illustrative case of an activation function. The formula for ReLU is below.

$$f(x) = \max(0, x) \tag{2}$$

In words, given an input, if the value is greater than or equal to zero, leave it unchanged, if the value is less than zero, output zero.

To introduce the activation function to our network, we run the result of our weighted sums at each node through the activation function before passing the result on to the next layer.

$$H_j = K(\sum_{i=1}^{n} w_i v_i) \tag{3}$$

Equation (3) is the same as Equation (1), with the addition of K, the activation function. A different activation function is typically used for the final output layer, depending on the application. For instance, a sigmoid function may be used for a binary outcome problem, or none at all for a regression problem. The intuition is the following, if we employed the ReLU activation function in the output layer of our network, we would never be able to predict values less than zero with our network.

3.1.3. Backpropagation and Gradient Descent

With a basic understanding of neural network architecture, the next question becomes how this network can learn or be trained to improve the qualities of its predictions. We have

enough information to understand the first step in this process. For the first training epoch, or run of data through the network, all weights in the network are randomly initialized, the input data is fed through, and a prediction is obtained. This initial randomization step is in fact a primary source of ANNs' stochasticity. Ten identical ANNs trained on the same data will output ten slightly different predictions because their starting points, or initial weights, were all different. This initial prediction will most likely be of poor quality, since weights were chosen randomly. For the model to learn, that is, update its weights, and increase the quality of its predictions, two more steps are necessary: backpropagation and gradient descent.

Although other methods of training a neural network exist, the combination of backpropagation and gradient descent has been by far the most common since the methodology's introduction for use in neural networks in 1986 (Rumelhart et al. 1986). Before a network can be trained to minimize error or loss, a loss function must be chosen to determine that error. The type of loss function chosen depends on the application. For instance, whether the network is used for classification or regression. Common loss functions for regression problems include mean absolute error (MAE) and mean squared error (MSE). See PyTorch (2021b) for more examples of loss functions. The formula for the MAE loss function is below.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{4}$$

where:

MAE = mean absolute error
n = the number of training observations
y = the actual value of an observation
x = the network's predicted value of an observation

Backpropagation is in turn an algorithm that computes the gradients, or partial derivatives, of this loss function with respect to the weights of each layer in the network using the chain rule. Figure 3 helps illustrate the intuition of this process.

Figure 3 shows the plot of a loss function in a network with a single weight. The goal is to minimize loss. In this simple case, we could set our derivative equal to zero and solve for weight. However, in more complicated networks, the loss function does not have a closed form derivative, so gradient descent is employed instead to reduce loss. At our initial weight, the gradient is calculated via the backpropagation algorithm, this information is then used to determine which direction to move the weight to reduce loss, represented by the cluster of arrows pointing towards the function minimum in Figure 3. The process is then repeated. Backpropagation is a rich mathematical field in its own right, interested readers should see Chapter 2 of Nielsen (2015) for a deeper examination.

Backpropagation is only the step that calculates gradients. To actually update weights and thus train the model, gradient descent is then employed. A representation of gradient descent is detailed below.

$$w_i = w_{i-1} - \delta \nabla C(w_{i-1}) \tag{5}$$

where:

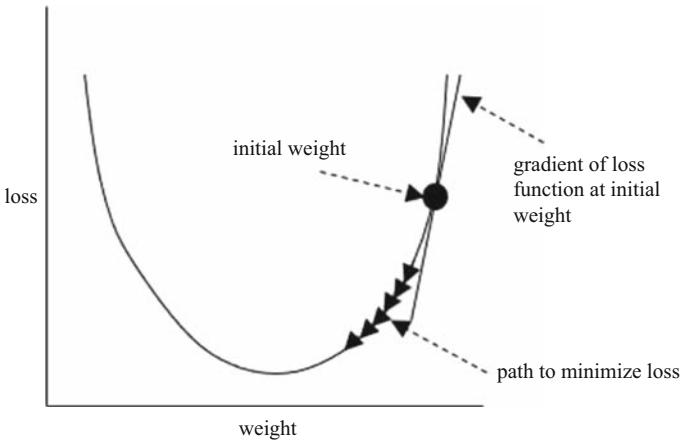


Fig. 3. Graphical representation of backpropagation and gradient descent with a single weight.

$w =$ the vector of weights in the network
$i =$ the current training epoch of the network
$\delta =$ the step size or learning rate
$C =$ the cost function

In words, a particular weight in the network is equal to its weight in the previous epoch minus the gradient of the cost function at the previous weight times the learning rate. Learning rate, or step size, is an important hyperparameter in ANNs. It dictates how much to update weights by in each training epoch. A very large learning rate risks unstable weights, as the cost function minimum may be continually jumped over. A small learning rate risks moving towards the cost function minimum very slowly. Taking the example of Figure 4, a high learning rate risks jumping to the left and right side of the parabola, never actually reducing loss, while a low learning rate risks moving the weight only slightly down to the left, even after many training epochs. In practice, calculating gradients for all observations, called gradient descent or batch gradient descent, can be computationally expensive, time consuming, and lead to overfitting (Keskar et al. 2017). Accordingly, stochastic gradient descent or mini batch gradient descent is often employed to speed up this process. In stochastic gradient descent, rather than calculating gradients for all observations, they are only calculated for one observation. Mini batch splits the difference between these two approaches, calculating gradients for a subset of observations. It should also be mentioned that networks may not always converge, or reduce loss, during training. This may be due to poor-quality input data, ill-suited hyperparameters, network architecture, choice of activation function, or other factors.

With that, we now understand the basics of how feedforward ANNs are structured and trained. A network architecture is specified with input data connected to an output via intermediary hidden layer(s). The weights between these layers and nodes are initially randomized. Data is passed through these weights and layers, run through a non-linear activation function, and an output is obtained. This output is then run through a cost function, which is then used to calculate gradients via backpropagation. Gradient descent is then employed to update the network's weights in the direction of reducing loss. The

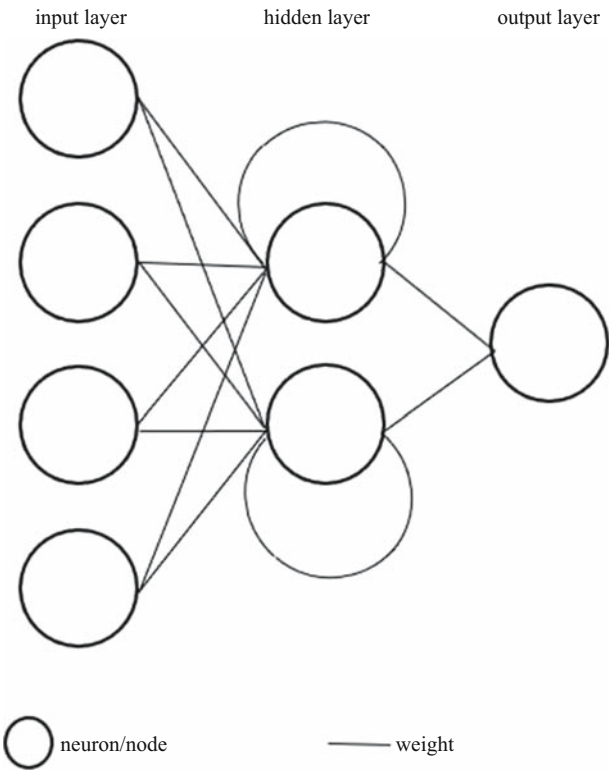


Fig. 4. A recurrent neural network.

input data is then run through the network again, and the process is repeated until loss no longer decreases or after a predetermined number of training epochs has been reached. Weights are then fixed at final values and the network is considered trained. New data can then be fed through to obtain new outputs or predictions

3.1.4. Recurrent Neural Networks

Feedforward networks are extremely powerful and have proven very effective in a variety of applications. However, as their name suggests, the flow of information through the network is unidirectional. This limits their usefulness in applications with a temporal aspect. They can, however, still be used with time series by flattening the data, that is, converting each lag of a variable into a separate column. In some cases, this approach can produce good predictions and can be considered before turning to the more explicitly time-based architectures detailed below (Brownlee 2018). Recurrent neural networks address feedforward networks' temporal deficiency by introducing feedback loops and converting the network into a directed acyclic graph (Amidi and Amidi 2019; Stratos 2020; Dematos et al. 1996). Figure 4 illustrates a simple RNN.

The network is similar to the feedforward network in Figure 2, but the outputs of the hidden layer are fed back into the network rather than directly to the output layer. The introduction of the temporal component means that the entirety of the network can no

longer be fully expressed in a diagram like Figure 4. Rather, each hidden node now conceals more folded in layers. To better illustrate this concept, Figure 5 displays one of the hidden nodes from Figure 4 unfolded.

where:

$x =$ input from the previous layer to the hidden node
$s =$ state of the hidden node
$o =$ output of the hidden node
$u =$ weight between the input layer and the hidden layer
$w =$ weight between the previous hidden state of the node and the current state
$v =$ weight between the hidden layer and the output layer

The mathematics for the cost function, backpropagation, and gradient descent remain the same, with the additional dimension of time. This temporal component makes RNNs well-suited for applications such as natural language processing or speech processing. However, due to vanishing or exploding gradients, which give RNNs a short memory, their usefulness in nowcasting is limited (Grosse 2017).

Vanishing and exploding gradients arise from the mechanics of backpropagation through time. With the introduction of time, the cost function is now calculated at each point in time, so weights need to be updated not only for each node, but for each node at each point in time. Additionally, later nodes are dependent on the input from earlier nodes, so weights w from Figure 5 are multiplied many times throughout the network and can thus tend towards zero or extremely large numbers. Because of temporal dependencies where later weights are based on early weights, this is an issue for the entire network. Long short-term memory networks, discussed in the next section, are one way of addressing this issue.

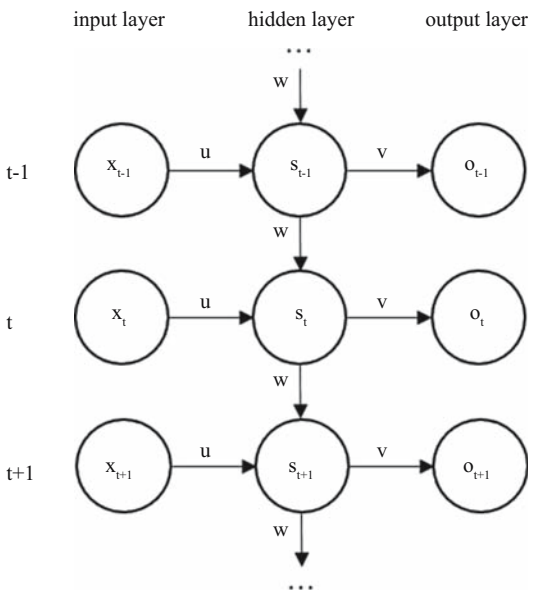


Fig. 5. Unfolded recurrent neural network node.

3.2. Long Short-Term Memory Networks

3.2.1. LSTM Architecture

Long short-term memory networks (LSTM) introduce three gates, a forget gate, an input gate, and an output gate, to RNN nodes (Chung et al. 2014). Crucially, these gates allow gradients to flow unchanged through the network, mitigating the exploding and vanishing gradients problem. Figure 6 displays an LSTM node with time *t* hidden node enlarged for detail. The graphical representation of an LSTM node is rather complicated, if readers still have trouble grasping the concepts, see Olah (2015).

where:

$x =$ input from the previous layer to the hidden node
$s =$ state of the hidden node
$o =$ output of the hidden node
$u =$ weight between the input layer and the hidden layer
$w =$ weight between the previous hidden state of the node and the current state
$v =$ weight between the hidden layer and the output layer
$c =$ memory state
$\sigma =$ sigmoid layer
$tanh =$ tanh layer

In contrast with the RNN diagram of Figure 5, we see that the LSTM node now has three inputs: data from the previous layer, *x,* the previous state or output of the node, s_{t-1}, and *c*, the memory state of the previous node. *c* is what allows gradients to flow through the network and gives LSTMs a longer memory. There are essentially three things happening

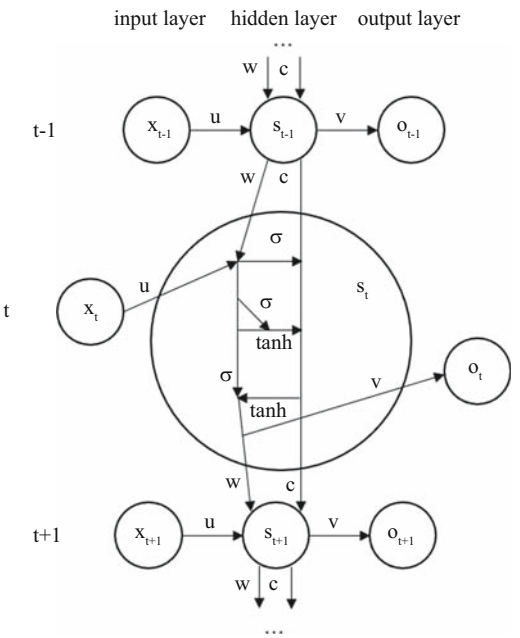


Fig. 6. Detail of an LSTM node.

in an LSTM node: first, deciding what information to keep or discard in the memory state; second, deciding what new information to introduce into the memory state; third, deciding what to output to the next layer.

These three steps are represented in Figure 6 by the three horizontal lines leading into and out of c in the hidden node at time t. Moving sequentially from top to bottom, the first horizontal line is the forget gate. The σ on this line represents a layer with the sigmoid activation function, which outputs a value between zero and one. This represents how much information from the previous memory state to allow to pass onwards. The next step, choosing what to introduce to the memory state, has two parts. First, simultaneously another sigmoid layer, the input gate, determines which values to update while a tanh layer scales those potential values to add to the memory state. The output of these two layers is then combined to update the memory state. At this point, the memory state is finished updating and is passed on to the next node. However, one last step is carried out to determine the output of the node. That is, the output of the node as we understand it from feedforward networks and RNNs, for instance to the final output layer. In a process similar to the second step, a sigmoid layer determines which parts of the memory state to output, while a tanh layer transforms the values of the memory state. These two are then combined to determine the output of the node.

3.2.2. LSTM Suitability for Nowcasting

LSTMs' ability to address the first common nowcasting data issue, mixed frequency data, stems from ANNs' ability to learn complex, non-linear relationships in data, a product of multiple neuron layers coupled with non-linear activation functions. As such, mixed frequency data can be fed to the network in the highest frequency available, with lower frequency data having missings at time periods where data are not published. These missing data can then be filled using a variety of approaches, including with the mean, the median, with values sampled from a distribution (Ennett et al. 2001), or with other more complex methods (Smieja et al. 2019). In the analysis performed in this article, mean replacement was chosen and implemented in the accompanying Python library.

LSTMs are able to address the ragged edges problem through no special mechanism other than standard missing-filling methods. These include using ARMA or VAR models to fill in ragged-edges (Kozlov et al. 2018), as well as using the mean or Kalman filters (Doz et al. 2011). The method chosen in the context of LSTM nowcasting can be considered a hyperparameter to be tuned and tested empirically. At the time of writing, the Python library supports ARMA filling and any n-to-1 series transformation, for example, mean, median, etc. Both ARMA and mean filling were used in the analysis performed in this article, depending on the results of hyperparameter tuning for each individual model.

The last major problem of nowcasting, the curse of dimensionality, is partially addressed by LSTMs' efficiency compared with other methods, i.e., their computation time scales very slowly with the number of variables (Hochreiter and Schmidhuber 1997). In empirical testing, the DFM's computation time scales exponentially with the number of features, while the LSTM's remains constant. Both methodologies' computation time scales linearly with the number of observations.

As a result of this efficiency, a functional model can be trained with many more features than a DFM. Computational feasibility of models with a large number of input variables is

however only part of how LSTMs and ANNs address the curse of dimensionality. Neural networks are remarkable for their ability to extract relevant features from higher dimensional spaces and project them onto lower dimensional spaces, which is already a form of feature reduction (Hodas and Stinis 2018), and for their robustness to multicollinearity (De Veaux and Ungar 1994). This is why incredibly complex networks with millions or more coefficients can still generate effective predictions with many fewer training observations. They are also compatible with standard dimensionality and overfitting reduction techniques, such as regularization. For instance, L2 regularization can be introduced to the *nowcast_lstm* library via the PyTorch optimizer function's *weight_decay* argument. PyTorch's *dropout* parameter is another effective means of regularization implemented in the *nowcast_lstm* library (PyTorch 2021a). Within the LSTM architecture, as in any ANN, there are many choices to be made regarding network architecture and hyperparameters. Some examples include the number of hidden states, the number of layers, the loss function and the activation function, among many others.

4. Empirical Analysis

4.1. Description of Data and Models

In order to assess the relative performance of LSTMs vs DFMs, three target variables were used: global merchandise exports in both value (WTO 2020) and volume (UNCTAD 2020a), and global services trade (UNCTAD 2020a). These are the same series UNCTAD currently produces nowcasts for using DFMs (UNCTAD 2020b) and which were examined in an UNCTAD research paper (Cantú 2018). The target series are all quarterly. A large pool of 116 mixed-frequency monthly and quarterly independent series was used to estimate each of the target series. These series are listed in Online Supplemental Material while more information on any individual series is available upon request. All series were converted to seasonally adjusted growth rates using the US Census Bureau's X13-ARIMA-SEATS methodology (USCB 2017). Modelling on seasonally adjusted data is standard practice in nowcasting and has been shown to produce empirically better results with DFMs than using non-seasonally adjusted data (Camacho et al. 2015).

The DFM model used was the same examined in Cantú (2018). In this model, the DFM is modeled in a state-space representation where it is assumed that the target and independent variables share a common factor as well as individual idiosyncratic components. The Kalman filter is then applied, and maximum likelihood estimates of the parameters obtained. This is a common method of estimating DFMs and is explained in further detail in Bańbura and Rünstler (2011). The LSTM model used was that present in the *nowcast_lstm* Python library using the average of 10 networks' output with basic hyperparameter tuning. Table 1 lists all hyperparameters and values used for tuning. The logic of averaging the output of more than one network to obtain predictions is discussed further in Subsection 4.4, but see Stock and Watson (2004) for a discussion of forecast combination. There exist different variants of LSTMs, but PyTorch's LSTM class, which is a classic or general LSTM architecture, served as the base for the library (PyTorch 2021c), with a general structure of a variable number of LSTM layers plus one linear densely connected layer for generating final predictions.

Table 1. Hyperparameters and values used for tuning.

Hyperparameter	Values
Batch Size	15, 30, 60
Loss function	L1, MSE
Number of hidden layers	10, 20, 40
Number of LSTM layers	2, 4
Number of training epochs	50, 100, 200
Ragged edge filling method	mean, ARMA

4.2. Modelling Steps

Hyperparameter tuning of the LSTM and model performance was evaluated using a training set dating from the second quarter of 2005 to the third quarter of 2016 and a test set dating from the fourth quarter of 2016 to the fourth quarter of 2019. Time series are generally not suitable for cross-fold validation, a model validation technique often used to avoid overfitting. Cross validation involves the random selection of subsamples from the data to train and assess models on different data. For more information on the topic, see Scikit-learn (2021). Time series, however, are often not independent and it makes little intuitive sense to predict the past using information from the future. Though there do exist methods of using cross validation with time series, for instance using rolling points in time in the past as the folds, they were not used in this analysis due to the already relatively short time span of the data.

A pool of independent variables was used to ensure the robustness of results, as either model could perform better on a single set of features due to chance. As such, the models' performance was evaluated by taking random samples of between five and 20 features, then fitting both an LSTM and DFM model on this same sample. Both methods' performance was then evaluated on the test set via mean absolute error (MAE) and root-mean-square error (RMSE) on five different data vintages, repeating the process 100 times for each of the three target variables. In this manner, a distribution of relative performance over a wide breadth of independent variables could be obtained. The number of features was restricted to a maximum of 20 due to the high computational time of estimating DFMs with more than this number, though performance of the LSTM may have been improved further if additional features were used.

Data vintages in this case refer to the artificial withholding of data to simulate what the availability of data would have looked like at different points in the past. This is important in evaluating model performance in the nowcasting context, as in real life series have varying publication schedules which nowcasting models must be robust to. The five vintages simulated were: two months before the target period, for example, if the target was the second quarter of 2019, the data as it would have appeared in April 2019; one month before; the month of; a month afterwards; and two months afterwards. The model continues to be evaluated even after the target period has theoretically "passed" as data continue to be published for a given month well after it has passed, depending on the series' individual publication schedule. For example, two months after the second quarter of 2019 simulates being in August 2019, when much more data on the second quarter is

available. The variables' publication lags were obtained based on empirical observations from the period from April to November 2020.

4.3. Relative Performance

Figure 7 shows the distribution of the LSTM's error as a proportion of the DFM's for each target variable. A value less than one for an individual model indicates better performance on the test set for the LSTM, while a value greater than one indicates worse performance. Consequently, a distribution centered around one, that is, the vertical line, indicates

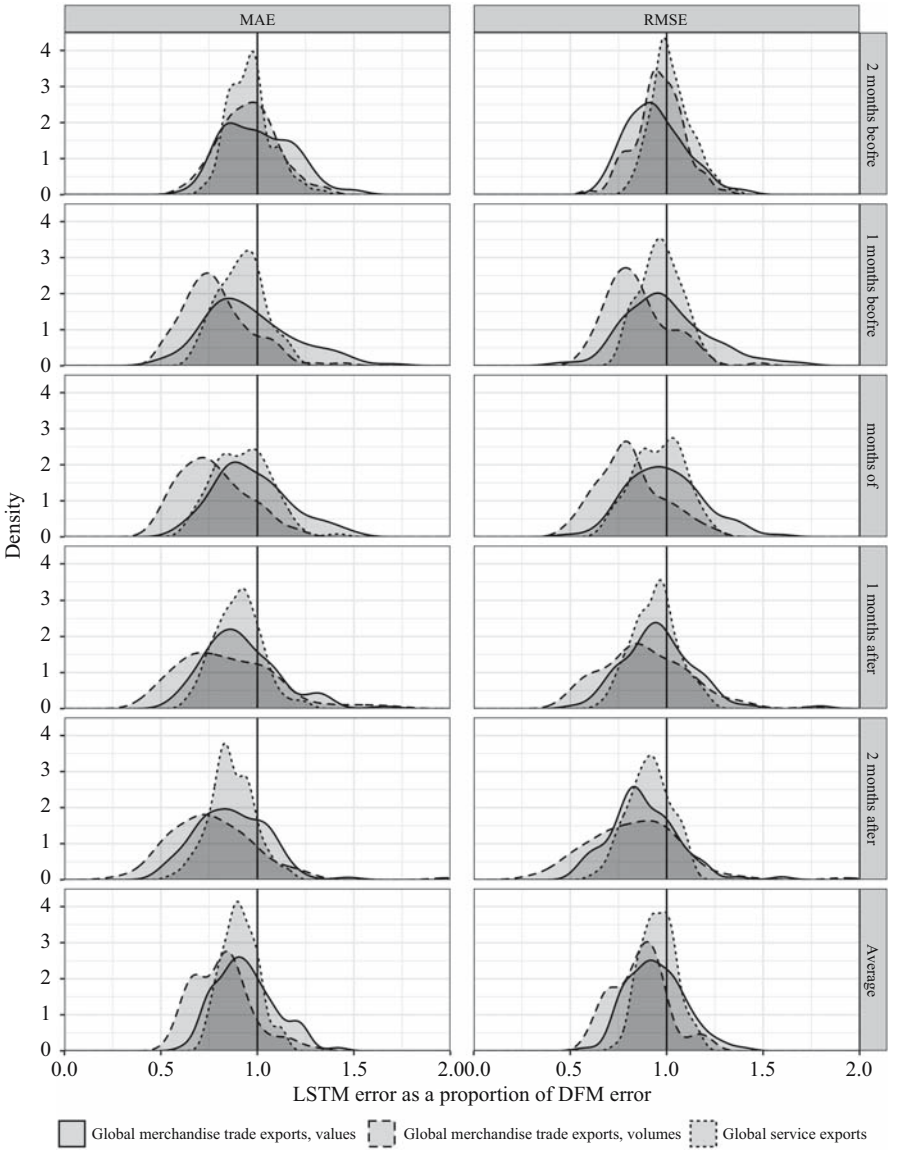


Fig. 7. LSTM error as a proportion of DFM error.

comparable performance between the two models, while one to the left of the vertical line indicates better performance on average for the LSTM model.

The results clearly favor the LSTM model, obtaining better average performance for both performance metrics across all data vintages and target variables, with the sole exception of RMSE for the two months before services exports vintage. Tables 2, 3 and 4 display the average performance metrics for the two models over the sample of 100 different feature combinations, as well as the results using a simple autoregressive model as a benchmark. A one-tailed t-test was performed on the LSTM and DFM errors to ascertain the significance of these differences in performance, with the alternative hypothesis that the LSTM errors were lower. Results are displayed in the LSTM columns.

Table 2. Average performance metrics, global merchandise trade exports, values.

Vintage	ARMA MAE	LSTM MAE	DFM MAE	ARMA RMSE	LSTM RMSE	DFM RMSE
2 months before	0.0177	0.0149	0.0150	0.0233	0.0176**	0.0185
l month before	0.0177	0.0112*	0.0117	0.0233	0.014	0.0141
month of	0.0177	0.0115	0.0118	0.0233	0.0142	0.0142
1 month after	0.0168	0.0108***	0.0117	0.0217	0.0138	0.0142
2 months after	0.0168	0.0094***	0.0109	0.0217	0.0119***	0.0135
Average	0.0173	0.0115**	0.0122	0.0227	0.0143	0.0149

Note: *p < .05 **p < .01 ***p < .001.

Table 3. Average performance metrics, global merchandise trade exports, volumes.

Vintage	ARMA MAE	LSTM MAE	DFM MAE	ARMA RMSE	LSTM RMSE	DFM RMSE
2 months before	0.0085	0.006**	0.0064	0.0097	0.0075**	0.0078
1 month before	0.0085	0.0051***	0.0066	0.0097	0.0066***	0.0079
month of	0.0085	0.0049***	0.0065	0.0097	0.0063***	0.0079
1 month after	0.0084	0.0045***	0.0057	0.0108	0.0059***	0.0069
2 months after	0.0084	0.0042***	0.0056	0.0108	0.0054***	0.0067
Average	0.0085	0.0049***	0.0062	0.0101	0.0063***	0.0074

Note: *p < .05 **p < .01 ***p < .001.

Table 4. Average performance metrics, global services exports.

Vintage	ARMA MAE	LSTM MAE	DFM MAE	ARMA RMSE	LSTM RMSE	DFM RMSE
2 months before	0.0119	0.0123***	0.0129	0.0151	0.0154	0.0152
1 month before	0.0119	0.0103***	0.0113	0.0151	0.0135**	0.0140
month of	0.0119	0.0103***	0.0111	0.0151	0.0135**	0.0141
l month after	0.0119	0.0103***	0.0115	0.0151	0.0137***	0.0146
2 months after	0.0119	0.0101***	0.0117	0.0151	0.0134***	0.0147
Average	0.0119	0.0107***	0.0117	0.0151	0.0139***	0.0145

Note: *p < .05 **p < .01 ***p < .001.

An additional metric for comparing the performance specifically of two forecasts is the Diebold-Mariano test. Table 5 displays the proportion of models where a one-sided Diebold-Mariano test was significant at the 5% level. The "DFM" column displays results where the alternative hypothesis was that the DFM was a more accurate forecast than the LSTM, while the "LSTM" column displays the reverse. The results closely mirror the findings in Figure 7 and Tables 2–4. In all target-vintage combinations but three the LSTM had a higher proportion of significance.

Tables 6 and 7 display additional information on the two methodologies' errors on the test set, namely average skewness, kurtosis, and various quantiles of absolute errors.

Table 5. Proportion of models with forecasts on the test set better than those of the other methodology, according to one-sided Diebold-Mariano test at 5% significance level.

Target	Vintage	DFM	LSTM
Values	2 months before	4%	10%
Values	1 month before	9%	6%
Values	month of	2%	7%
Values	1 month after	2%	9%
Values	2 months after	1%	17%
Volumes	2 months before	0%	3%
Volumes	1 month before	0%	9%
Volumes	month of	0%	15%
Volumes	1 month after	1%	15%
Volumes	2 months after	1%	26%
Services	2 months before	6%	2%
Services	1 month after	1%	0%
Services	month of	0%	4%
Services	1 month after	0%	10%
Services	2 months after	0%	21%

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

Table 6. DFM, average skewness, kurtosis, and quantiles of absolute errors on the test set.

Target	Vintage	Skewness	Kurtosis	50th quantile	75th quantile	90th quantile	95th quantile	99th quantile
Values	2 month before	−0.198	2.437	0.013	0.021	0.028	0.033	0.037
Values	1 month before	0.078	2.223	0.011	0.017	0.021	0.024	0.026
Values	month of	0.113	2.268	0.011	0.016	0.022	0.025	0.027
Vatues	1 month after	−0.013	2.401	0.010	0.016	0.022	0.026	0.028
Values	2 months after	0.168	2.644	0.010	0.015	0.021	0.024	0.026
Volumes	2 months before	−0.310	2.544	0.006	0.009	0.012	0.014	0.015
Volumes	1 month before	−0.205	2.391	0.006	0.009	0.012	0.014	0.015
Volumes	month of	−0.123	2.318	0.006	0.009	0.012	0.014	0.015
Volumes	1 month after	0.077	2.392	0.005	0.008	0.011	0.012	0.013
Volumes	2 months after	0.118	2.114	0.005	0.008	0.010	0.011	0.012
Services	2 months before	0.634	2.476	0.012	0.017	0.022	0.025	0.029
Services	1 month before	0.594	2.992	0.010	0.015	0.022	0.025	0.029
Services	month of	0.715	3.222	0.009	0.015	0.023	0.026	0.030
Services	1 month after	0.676	3.314	0.009	0.015	0.024	0.028	0.031
Services	2 months after	0.713	3.170	0.009	0.016	0.023	0.027	0.031

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

Table 7. LSTM, average skewness, kurtosis, and quantiles of absolute errors on the test set.

Target	Vintage	Skewness	Kurtosis	50th quantile	75th quantile	50th quantile	95th quantile	99th quantile
Values	2 months before	−0.247	1.820	0.014	0.021	0.026	0.028	0.030
Values	1 month before	−0.083	2.481	0.009	0.016	0.022	0.024	0.027
Values	month of	−0.011	2.435	0.009	0.016	0.021	0.024	0.027
Values	1 month after	−0.224	2.555	0.008	0.016	0.021	0.024	0.027
Values	2 months. after	−0.296	2.668	0.008	0.013	0.018	0.021	0.024
Volumes	2 months before	−1.036	3.635	0.005	0.008	0.010	0.013	0.016
Volumes	1 month before	−1.024	4.052	0.004	0.007	0.009	0.012	0.015
Volumes	month of	−0.850	3.633	0.004	0.007	0.009	0.011	0.014
Volumes	1 month after	−0.763	3.638	0.004	0.006	0.008	0.011	0.013
Volumes	2 months after	−0.728	3.611	0.003	0.005	0.007	0.010	0.012
Services	2 months before	0.280	2.387	0.011	0.018	0.024	0.028	0.031
Services	1 month before	0.383	2.768	0.008	0.014	0.023	0.026	0.028
Services	month of	0.508	2.792	0.007	0.014	0.023	0.026	0.028
Services	1 month after	0.564	2.938	0.007	0.013	0.024	0.027	0.029
Services	2 months after	0.777	3.104	0.008	0.013	0.022	0.026	0.029

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

To test for autocorrelation of errors, a portmanteau test was performed on the test predictions for all models (Johansen 1995). Average p-values are presented in Table 8, with the null hypothesis that autocorrelation was not present in the residuals.

At the individual model level, only 11 were able to reject the null hypothesis at a 5% significance level, all DFMs predicting global merchandise exports values at various time vintages. To test for heteroskedasticity in the residuals, a Lagrange Multiplier test was additionally performed, with the null hypothesis that the residuals were homoskedastic (Engle 1982). Average p-values are presented in Table 9.

No individual models were able to reject the null hypothesis at a 5% significance level. The results of these tests imply that neither the DFM nor the LSTM are likely to suffer from systematic issues of autocorrelation or heteroskedasticity in their errors. Table 8 displays average bias and variance of the two models at different vintages. Bias and variance were calculated individually for each model in the sample then averaged by target variable-data vintage combination (Table 10).

In terms of bias, the two methods are comparable, with relative performance varying depending on the target series. Broadly, the LSTM had a lower bias on average for

Table 8. Average p-values of portmanteau test for autocorrelation.

Vintage	DFM, values	LSTM, values	DFM, volumes	LSTM, volumes	DFM, services	LSTM, services
2 months before	0.72	0.69	0.70	0.92	0.82	0.71
1 month before	0.65	0.73	0.73	0.95	0.83	0.72
month of	0.49	0.68	0.72	0.92	0.87	0.74
1 month after	0.57	0.74	0.73	0.88	0.87	0.84
2 months after	0.84	0.76	0.75	0.86	0.83	0.84

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

Table 9. Average p-values of Lagrange Multiplier test for heteroskedasticity.

Vintage	DFM, values	LSTM, values	DFM, volumes	LSTM, volumes	DFM, services	LSTM, services
2 months before	0.94	1.00	0.98	0.99	0.95	0.90
1 month before	0.97	0.94	0.98	0.96	0.99	0.82
month of	0.98	0.97	0.98	0.98	0.97	0.83
1 month after	0.97	0.97	0.97	0.95	0.94	0.83
2 months after	0.97	0.96	0.98	0.96	0.96	0.91

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

Table 10. Average bias and variance of predictions on the test set at different vintages.

Target	Vintage	DFM bias	LSTM bias	DFM variance	LSTM variance
Values	2 months before	−0.004833	−0.001999	0.000265	0.000096
Values	1 month before	−0.003976	−0.002097	0.000412	0.000174
Values	month of	−0.004144	−0.002134	0.000444	0.000238
Values	1 month after	−0.004955	−0.003545	0.000352	0.000260
Values	2 months after	−0.003793	−0.004156	0.000415	0.000330
Volumes	2 months before	0.000195	0.000854	0.000060	0.000013
Volumes	1 month before	0.000543	0.000879	0.000096	0.000024
Volumes	month or	0.000511	0.000684	0.000104	0.000034
Volumes	1 month after	−0.000018	0.000191	0.000085	0.000039
Volumes	2 months after	−0.000043	0.000137	0.000075	0.000044
Services	2 months before	−0.001474	0.001746	0.000070	0.000021
Services	1 month before	−0.001184	0.001913	0.000112	0.000047
Services	month of	−0.001261	0.002161	0.000116	0.000064
Services	1 month after	−0.001729	0.001250	0.000097	0.000073
Services	2 months after	−0.001616	0.001066	0.000105	0.000105

Note: For brevity, "Values" refers to global merchandise exports in values, "Volumes" refers to global merchandise exports in volumes, and "Services" refers to global services exports.

merchandise exports in terms of values, while the DFM had a lower bias on average for merchandise exports in terms of volumes. For services, which methodology had a lower bias depended on the vintage. Variance in the LSTM was lower than that of the DFM in all target-vintage combinations, though the degree to which this was the case varied. The fact that the LSTM is able to combine higher accuracy with lower volatility suggests the DFM may be overly reactive to signals in the data, predicting large changes in the target variable that are either inaccurate or excessive.

4.4. Comparison With the Dynamic Factor Model

The fact that the LSTM performed better than the DFM on average for all three target variables across almost all vintages and both performance metrics is strong evidence for their relevance in the economic nowcasting space. The LSTM's ability to handle long-term temporal dependencies due to its architecture may be a source of some of this improved

predictive performance. Of course, the results do not indicate that LSTMs are superior to this type of DFM in every instance. They rather provide some evidence that they can be a competitive alternative to DFMs and have the potential to become a more commonly used methodology in nowcasting. There are, however, characteristics of the methodology with pros and cons relative to DFMs that are independent of predictive performance. One of the pros relative to the DFM was discussed in Section 3. Namely, LSTMs' ability to handle many more features than the DFM before coming up against computational bottlenecks. This could be beneficial by lessening the need for variable selection in the early stages of an analysis, easing the obtainment of initial results. Additionally, a model is able to be reliably trained on any given set of features and values, which is not the case for the DFM, the training of which may fail if input matrices are non-invertible.

A third advantage is the ability to easily use mixed frequency variables with no corresponding change in the underlying modeling and formulas. Annual, quarterly, monthly, and even theoretically daily data can be combined in a single model just by changing the structure or frequency of the input data, as explained in Subsection 3.2.

Computational speed is more difficult to ascribe to either method as an advantage. There are many factors affecting the computation time of the two models. For DFMs, this includes the number of features, the number of observations, and especially the time taken for maximum likelihood convergence. For LSTMs, this includes the number of observations, as well as nearly all of the hyperparameters. As such, there are cases where either method can be faster. Even still, training a single LSTM network with any choice of hyperparameters is usually faster than estimating a DFM on the same data. For instance, in the 300 model runs of this analysis, this was the case 96% of the time, with the LSTM taking on average just 25% of the time needed to estimate the DFM. However, the results in Figure 7 were obtained by fitting ten LSTM models and averaging the result, in which case the LSTM was faster just 42% of the time, taking on average 2.5 times as long to estimate compared with the DFM. These numbers are slightly skewed in favor of the DFM however, as the number of features was restricted to a maximum of 20. Models with a number of features above this would favor the LSTM in computation time. Because the LSTM needs to be hyperparameter tuned for every new target series, likely more LSTM models need to be estimated than DFM models in the model selection phase of an exercise, leading to comparatively longer run times initially. However, once variables and hyperparameters are selected and models just need to be retrained periodically with the latest data, this disadvantage disappears.

The fact that results were evaluated using ten networks for the LSTM has to do with one of their disadvantages relative to DFMs, namely, the stochastic nature of ANNs. Ten LSTM networks trained on the same data will output ten different predictions due to the randomization of initial weights, which is not the case for DFMs. Training many networks and taking their average predictions is a way to mitigate this characteristic. Figure 8 illustrates how the distribution of predictions develops as more networks are used.

The distributions were obtained by taking a single set of variables predicting global merchandise exports values and training an LSTM model averaging the outputs of between one and 20 networks on the data, then generating a prediction for a single time period. This was repeated 100 times by retraining the same model, thus generating 100 predictions for a single target period, creating the distributions. The predictions were then

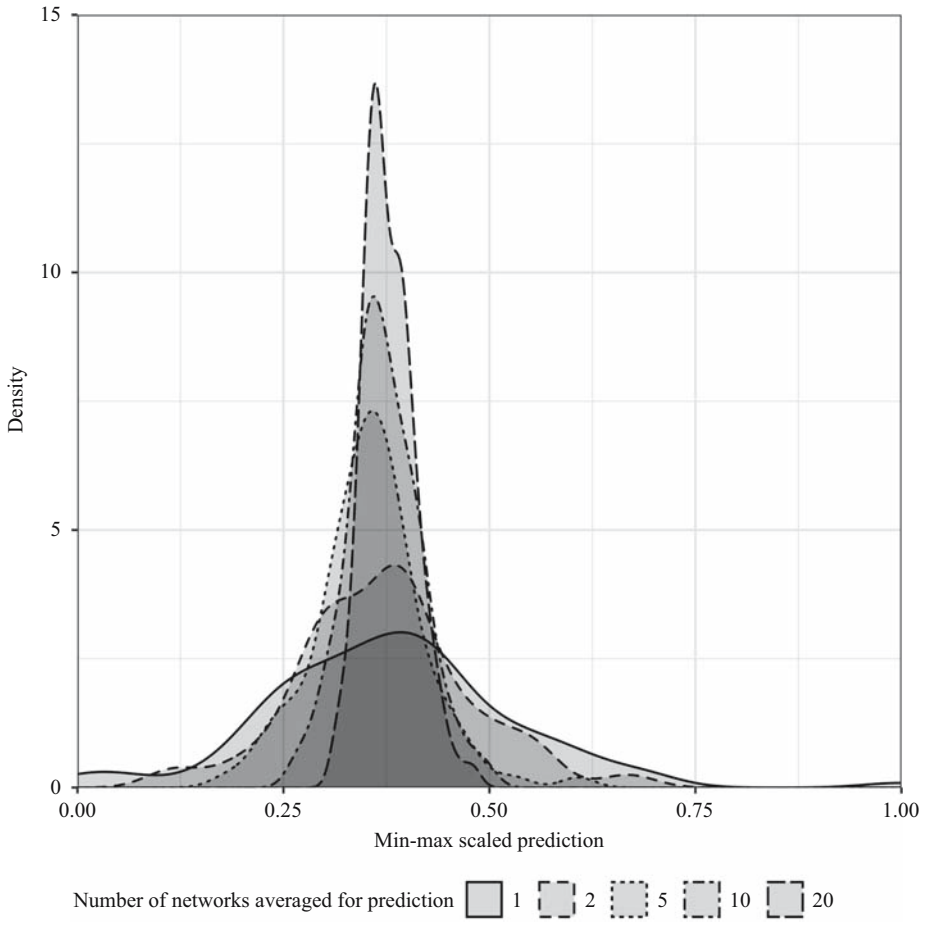


Fig. 8. Distribution of predictions of the same target observation.

scaled to be between 0 and 1 for clarity. Variance decreases as more networks are added. While adding more networks can mitigate the stochastic nature of LSTMs' predictions, adding very many can substantially increase computation time while never achieving perfectly consistent outputs.

A final disadvantage LSTMs have compared with DFMs is the lack of interpretability in their parameters, and the consequent lack of inference as to what is driving changes in the model. DFMs have the advantage of being able to offer precise insights to various features' impact on predictions, as illustrated by the New York Fed's nowcasts (Federal Reserve Bank of New York 2021). This is a well-known characteristic of ANNs in general (Fan et al. 2020). In this regard, there is opportunity for further research into applying existing ANN interpretability methods, such as activation maximization or sensitivity analysis (Montavon et al. 2018), to the nowcasting LSTM framework, though one has already been implemented in the *nowcast_lstm* library. The methodology employed is similar to a simplified implementation of calculating Shapley values, whose use in adding interpretability to machine learning methodologies is explained in further detail in

Subsection 5.9 of Molnar (2019). For more information on usage and the methodology, see the repository's README and example file.

5. Conclusion

Timely, accurate estimates of macroeconomic series can be valuable in helping inform policy decisions. This article provides evidence for stronger consideration of LSTMs for this purpose, as well as introduces Python, R, MATLAB, and Julia libraries to facilitate future research. LSTMs were shown to produce superior predictions compared with DFMs on three different target series: global merchandise trade exports expressed in both values and volumes and global services exports, and over five different data vintages.

In addition to better empirical performance for the three target series, LSTMs provide advantages over DFMs by being able to handle large numbers of features without computational bottlenecks, not relying on the invertibility of any matrices, thus being able to be fit on any dataset, and the ability to use any mixture of frequencies in features or target. Disadvantages relative to DFMs include LSTMs' stochastic nature and opacity regarding feature contribution to predictions.

The *nowcast_lstm* library can facilitate the use of LSTMs in economic nowcasting by lowering the barrier to experimentation. LSTMs' ability to reliably generate predictions on a large number of input features makes it easier to quickly verify whether or not a given series has the potential to be nowcast, a characteristic that could help expand the variety and quantity of economic variables monitored via nowcasting.

There remains much scope for future research and development on this topic. Further testing should be performed to verify LSTMs' performance on a wider variety of series and frequency mixtures. More hyperparameter tuning could be performed to see if tweaking other aspects of model architecture could result in better results. There is also much scope for exploring different methods of filling missing values beyond ARMA or mean-filling. Finally, methods in addition to the one already implemented for interpreting LSTMs and ascertaining feature contribution to predictions would increase the method's viability as a policy-informing instrument. The library could be extended in the future to incorporate any improvements to performance or functionality deriving from future research, continuing to facilitate the adoption and development of the methodology in the nowcasting domain.

6. References

Amidi, A., and S. Amidi. 2019. "Recurrent Neural Networks Cheatsheet". Recurrent Neural Networks Cheatsheet. 2019. Available at: https://stanford.edu/ ~ shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks. (accessed December 2020).

Antolin-Diaz, J., T. Drechsel, and I. Petrella. 2020. *Advances in Nowcasting Economic Activity: Secular Trends, Large Shocks and New Data*. DOI: http://dx.doi.org/10.2139/ssrn.3669854.

Banbura, M., D. Giannone, and L. Reichlin. 2010. *Nowcasting*. ECB Working Paper 1275. DOI: http://dx.doi.org/10.2139/ssrn.1717887.

Bańbura, M, and G. Rünstler. 2011. "A Look into the Factor Model Black Box: Publication Lags and the Role of Hard and Soft Data in Forecasting GDP". *International Journal of Forecasting* 27 (2): 333–346. DOI: https://doi.org/10.1016/j.ijforecast.2010.01.011.

Bok, B., D. Caratelli, D. Giannone, A.M. Sbordone, and A. Tambalotti. 2018. "Macroeconomic Nowcasting and Forecasting with Big Data". *Annual Review of Economics* 10 (1): 615–643. DOI: https://doi.org/10.1146/annurev-economics-080217-053214.

Brownlee, J. 2018. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery. Available at: https://books.google.ch/books?id = o5qnDwAAQBAJ.

Buono, D., G. Mazzi, M. Marcellino, and Kapetanios. 2017. "Big Data Types for Macroeconomic Nowcasting". *Eurona*.

Camacho, M., Y. Lovcha, and G. Perez Quiros. 2015. "Can We Use Seasonally Adjusted Variables in Dynamic Factor Models?" *Studies in Nonlinear Dynamics & Econometrics* 19 (3): 377–391. DOI: https://doi.org/doi:10.1515/snde-2013-0096.

Cantú, F. 2018. *Estimation of a Coincident Indicator for International Trade and Global Economic Activity*. 27. UNCTAD Research Paper. UNCTAD. Available at: https://unctad.org/system/files/official-document/ser-rp-2018d9_en.pdf.

Chernis, T., and R. Sekkel. 2017. "A Dynamic Factor Model for Nowcasting Canadian GDP Growth". *Empirical Economics* 53(1): 217–234. DOI: https://doi.org/10.1007/s00181-017-1254-1.

Chung, J., Ç. Gülçehre, K. Cho, and Y. Bengio. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". *CoRR* abs/1412.3555. Available at: http://arxiv.org/abs/1412.3555.

De Veaux, R.D., and L.H. Ungar. 1994. "Multicollinearity: A Tale of Two Nonparametric Regressions". In *Selecting Models from Data*, edited by P. Cheeseman and R.W. Oldford, 393–402. New York, NY: Springer New York.

Dematos, G., M.S. Boyd, B. Kermanshahi, N. Kohzadi, and I. Kaastra. 1996. "Feedforward versus Recurrent Neural Networks for Forecasting Monthly Japanese Yen Exchange Rates". *Financial Engineering and the Japanese Markets* 3(1): 59–75. DOI: https://doi.org/10.1007/BF00868008.

Domo. 2017. "Data Never Sleeps 5.0". *Data Never Sleeps 5.0*. 2017. Available: https://www.domo.com/learn/data-never-sleeps-5. (accessed September 2021).

Doz, C., D. Giannone, and L. Reichlin. 2011. "A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering". *Journal of Econometrics* 164(1): 188–205. DOI: https://doi.org/10.1016/j.jeconom.2011.02.012.

Einav, L., and J. Levin. 2014. "The Data Revolution and Economic Analysis". *Innovation Policy and the Economy* 14: 1–24. DOI: https://doi.org/10.1086/674019.

Engle, R.F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". *Econometrica* 50(4): 987–1007. DOI: https://doi.org/10.2307/1912773

Ennett, C.M., M. Frize, and C.R. Walker. 2001. "Influence of Missing Values on Artificial Neural Network Performance." *Studies in Health Technology and Informatics* 84 (1): 449–53.

Falat, L., and L. Pancikova. 2015. "Quantitative Modelling in Economics with Advanced Artificial Neural Networks". *Procedia Economics and Finance* 34: 194–201. DOI: https://doi.org/10.1016/S2212-5671(15)01619-6.

Fan, F., J. Xiong, and G. Wang. 2020. "On Interpretability of Artificial Neural Networks". *CoRR* abs/2001.02522. Available at: http://arxiv.org/abs/2001.02522.

Federal Reserve Bank of New York. 2021. "Nowcasting Report: Methodology". *Nowcasting Report: Methodology*. 2021. Available at: https://www.newyorkfed.org/research/policy/nowcast/methodology.html. (accessed November 2020)

Giannone, D., L. Reichlin, and S. Simonelli. 2009. "Nowcasting Euro Area Economic Activity in Real Time: The Role of Confidence Indicators". *National Institute Economic Review* 210 (1): 90–97. DOI: https://doi.org/10.1177/0027950109354413.

Giannone, D., L. Reichlin, and D. Small. 2005. *Nowcasting GDP and Inflation: The Real Time Informational Content of Macroeconomic Data Releases*. Centre for Economic Policy Research. Available et: https://cepr.org/active/publications/discussion_papers/dp.php?dpno = 5178.

Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data". *Journal of Monetary Economics* 55(4): 665–666. DOI: https://doi.org/10.1016/j.jmoneco.2008.05.010.

Grosse, R. 2017. *Lecture 15: Exploding and Vanishing Gradients*. Available at: http://www.cs.toronto.edu/ ~ rgrosse/courses/csc321_2017/readings/L15%20Exploding%20and%20Vanishing%20Gradients.pdf. (accessed December 2020)

Guichard, S., and E. Rusticelli. 2011.*A Dynamic Factor Model for World Trade Growth:* 874.OECD Economics Department Working Papers. DOI: https://doi.org/10.1787/5kg9zbvvwqq2-en.

Gurney, K. 1997. *An Introduction to Neural Networks*. USA: Taylor & Francis, Inc.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory". *Neural Computation* 9: 1735–80. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.

Hodas, N.O., and P. Stinis. 2018. "Doing the Impossible: Why Neural Networks Can Be Trained at All". *Frontiers in Psychology* 9: 1185. DOI: https://doi.org/10.3389/fpsyg.2018.01185.

Hopp, D. 2021a. "nowcast_lstm". Available at: https://github.com/dhopp1/nowcast_lstm/

Hopp, D. 2021b. "nowcastLSTM". Available at: https://github.com/dhopp1/nowcastLSTM/

Hopp, D. 2021c. "nowcast_lstm_matlab". Available at: https://github.com/dhopp1/nowcast_lstm_matlab/

Hopp, D. 2021d. "NowcastLSTM.jl". Available at: https://github.com/dhopp1/NowcastLSTM.jl/

Johansen, S. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. New York.

Keskar, N.S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang. 2017. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima.* Available at: https://arxiv.org/abs/1609.04836.

Kozlov, M., S. Karaivanov, D. Tsonev, and R. Valkov. 2018. "The News on Norcasting". *The News on Nowcasting*. Available at: https://www.weareworldquant.com/en/thought-leadership/the-news-on-nowcasting/. (accessed December 2020)

Kurihara, Y., and A. Fukushima. 2019. "AR Model or Machine Learning for Forecasting GDP and Consumer Price for G7 Countries". *Applied Economics and Finance* 6: 1. DOI: https://doi.org/10.11114/aef.v6i3.4126.

Kuzin, V.N., M. Marcellino, and C. Schumacher. 2009. *MIDAS versus Mixed-Frequency VAR: Nowcasting GDP in the Euro Area*. Discussion Paper Series 1: Economic Studies 2009,07. Deutsche Bundesbank. Available at: https://ideas.repec.org/p/zbw/bubdp1/7576.html.

Loermann, J., and B. Maas. 2019. Nowcasting US GDP with Artificial Neural Networks. MPRA Paper 95459. University Library of Munich, Germany. Available at: https://ideas.repec.org/p/pra/mprapa/95459.html.

MacFeely, S. 2020. "In Search of the Data Revolution: Has the Official Statistics Paradigm Shifted?" *Statistical Journal of the IAOS* 36(4): 1075–1094. DOI: https://doi.org/10.3233/SJI-200662.

Marcellino, M., and C. Schumacher. 2010. "Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP". *Oxford Bulletin of Economics and Statistics* 72(4): 518–550.

Mariano, R.S., and Y. Murasawa. 2003. "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series". *Journal of Applied Econometrics* 18(4): 427–443. DOI: https://doi.org/10.1002/jae.695.

Matheson, T. 2011. "New Indicators for Tracking Growth in Real Time". *IMF Working Paper* 11 (43): 1–22.

Molnar, C. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Montavon, G., W. Samek, and K.-R. Müller. 2018. "Methods for Interpreting and Understanding Deep Neural Networks". *Digital Signal Processing* 73: 1–15. DOI: https://doi.org/10.1016/j.dsp. 2017.10.011.

Morgado, A.J., L. Catela Nunes, and S. Salvado. 2007. *Nowcasting an Economic Aggregate with Disaggregate Dynamic Factors: An Application to Portuguese GDP*. GEE Papers 0002. Gabinete de Estratégia e Estudos, Ministério da Economia. Available at: https://ideas.repec.org/p/mde/wpaper/0002.html.

Nielsen, M.A. 2015. "Neural Networks and Deep Learning." *Determination Press*.

Olah, C. 2015. *Understanding LSTM Networks.* Available at: https://colah.github.io/-posts/2015-08-Understanding-LSTMs/.

OpenTable. 2021. "The State of the Restaurant Industry". *The State of the Restaurant Industry*. Available at: https://www.opentable.com/state-of-industry. (accessed September 2021).

Porshakov, A., A. Ponomarenko, and A. Sinyakov. 2016. "Nowcasting and Short-Term Forecasting of Russian GDP with a Dynamic Factor Model". *Journal of the New Economic Association* 30(2): 60–76.

PyTorch. 2021a. "Dropout". Available at: https://pytorch.org/docs/stable/generated/-torch.nn.Dropout.html. (accessed October 2021)

PyTorch. 2021b. "Loss Functions". Available at: https://pytorch.org/docs/stable/n-n.html#loss-functions. (accessed August 2021).

PyTorch. 2021c. "LSTM". Available at: https://pytorch.org/docs/stable/generated/-torch.nn.LSTM.html. (accessed October 2021).

Rossiter, J. 2010. *Nowcasting the Global Economy. 2010–2012*. Bank of Canada. Available at: https://ssrn.com/abstract = 1674952. DOI: http://dx.doi.org/10.2139/ssrn.1674952.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams. 1986. "Learning Representations by Back-Propagating Errors". *Nature* 323 (6088): 533–36. DOI: https://doi.org/10.1038/323533a0.

Scikit-learn. 2021. "3.1. Cross-Validation: Evaluating Estimator Performance". *3.1. Cross-Validation: Evaluating Estimator Performance*. 2021. Availavle at: https://Scikit-learn.org/stable/modules/cross_validation.html. (accessed October 2021)

Sharma, S., S. Sharma, and A. Athaiya. 2020. "Activation Functions in Neural Networks". *International Journal of Engineering Applied Sciences and Technology,* 4(12): 310–316.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo. 2015. *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*.

Smieja, M.Ł. Struski, J. Tabor, B. Zieliński, and P. Spurek. 2019. *Processing of Missing Data by Neural Networks*.

Statista. 2021. "Coronavirus Impact on Retail E-Commerce Website Traffic Worldwide as of June 2020, by Average Monthly Visits". *Coronavirus Impact on Retail E-Commerce Website Traffic Worldwide as of June 2020, by Average Monthly Visits*. 2021. Available at: https://www.statista.com/statistics/1112595/covid-19-impact-retail-e-commerce-site-traffic-global/. (accessed March 2021).

Stock, J.H., and M.W. Watson. 2002. "Forecasting Using Principal Components From a Large Number of Predictors". *Journal of the American Statistical Association* 97(460): 1167–1179. DOI: https://doi.org/10.1 198101 621450238861 8960.

Stock, J.H., and M.W. Watson. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set". *Journal of Forecasting* 23(6): 405–430. DOI: https://doi.org/10.1002/for.928.

Stratos, K. 2020. *Feedforward and Recurrent Neural Networks*. Available at: http://www1.cs.columbia.edu/ ~ stratos/research/neural.pdf. (accessed December 2020)

"Transforming Our World: The 2030 Agenda for Sustainable Development". *Transforming Our World: The 2030 Agenda for Sustainable Development*. Available at:https://sdgs.un.org/2030agenda.

UN. 2015. "Transforming Our World: The 2030 Agenda for Sustainable Development". *Transforming Our World: The 2030 Agenda for Sustainable Development*. Available at: https://sdgs.un.org/2030agenda. (accessed September 2020).

UNCTAD. 2020a. "UNCTADStat". *UNCTADStat*. 2020. Available at: https://unctadstat.unctad.org/EN/Index.html. (accessed March 2021).

UNCTAD. 2020b. "Global Merchandise Trade Nowcast December 2020". *Global Merchandise Trade Nowcast December 2020*. Available at: https://unctad.org/system/files/official-document/gdsdsimisc2020d8_en.pdf. (accessed August 2020).

UNSD. 2020. "Nowcasting and Forecasting for SDG Monitoring". Presented at the Nowcasting and Forecasting for SDG Monitoring, February 3, Geneva, Switzerland. Available at: https://unstats.un.org/unsd/statcom/51st-session/side-events/20200302-2L-Nowcasting-and-Forecasting-for-SDG-Monitoring/. (accessed February 2021).

"The X-13ARIMA-SEATS Seasonal Adjustment Program". *The X-13ARIMA-SEATS Seasonal Adjustment Program*. 2017. Available at:https://www.census.gov/srd/www/-x13as/.

USCB. 2017. "The X-13ARIMA-SEATS Seasonal Adjustment Program". *The X-13ARIMA-SEATS Seasonal Adjustment Program*. 2017. Available at: https://www.census.gov/srd/www/x13as/. (accessed March 2021).

WMO. 2017. "Guidelines for Nowcasting Techniques". 1198. WMO. WMO. Available at: https://library.wmo.int/doc_num.php?explnum_id = 3795. (accessed March 2021).

WTO. 2020. "Statistics on Merchandise Trade". *Statistics on Merchandise Trade*. 2020. Available at: https://www.wto.org/english/res_e/statis_e/merch_trade_stat_e.htm. (accessed November 2020).

Yilmazkuday, Hakan. 2021. "Stay-at-Home Works to Fight against COVID-19: International Evidence from Google Mobility Data". *Journal of Human Behavior in the Social Environment* 31 (1–4): 210–220. DOI: https://doi.org/10.1080/10911359.2020.1845903.

Received July 2021
Revised August 2021
Accepted February 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 875–900, http://dx.doi.org/10.2478/JOS-2022-0038

Variable Inclusion Strategies through Directed Acyclic Graphs to adjust Health Surveys subject to Selection Bias for Producing National Estimates

Yan Li[1], *Katherine E. Irimata*[2], *Yulei He*[2], *and Jennifer Parker*[2]

Along with the rapid emergence of web surveys to address time-sensitive priority topics, various propensity score (PS)-based adjustment methods have been developed to improve population representativeness for nonprobability- or probability-sampled web surveys subject to selection bias. Conventional PS-based methods construct pseudo-weights for web samples using a higher-quality reference probability sample. The bias reduction, however, depends on the outcome and variables collected in both web and reference samples. A central issue is identifying variables for inclusion in PS-adjustment. In this article, directed acyclic graph (DAG), a common graphical tool for causal studies but largely under-utilized in survey research, is used to examine and elucidate how different types of variables in the causal pathways impact the performance of PS-adjustment. While past literature generally recommends including all variables, our research demonstrates that only certain types of variables are needed in PS-adjustment. Our research is illustrated by NCHS' Research and Development Survey, a probability-sampled web survey with potential selection bias, PS-adjusted to the National Health Interview Survey, to estimate U.S. asthma prevalence. Findings in this article can be used by National Statistics Offices to design questionnaires with variables that improve web-samples' population representativeness and to release more timely and accurate estimates for priority topics.

Key words: Kernel weighting; logistic regression; propensity score model; survey inference.

1. Introduction

Producing timely data is a priority of National Statistics Offices (NSOs). However, some of the more timely data collections, including web-based surveys, may be subject to biases relative to large nationally representative surveys conducted by NSOs due to lower coverage and response rates. Adjusting these timelier sources with less timely but higher quality reference surveys may decrease their biases.

Selection bias has been acknowledged in different areas (Hernán 2004) and is becoming more critical in the big data era with the rapid emergence of various web surveys to address time-sensitive priority topics, referred to here as target samples. Data collected in target samples, such as web panels, can result in attrition and response rates are often found to be 10% or lower (Baker et al. 2013). Although low response is not necessarily

[1] Joint Program in Survey Methodology, Department of Epidemiology and Biostatistics, 1218 Lefrak Hall, University of Maryland College Park, College Park, MD 20742, U.S.A. Email: yli6@umd.edu
[2] Division of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782, U.S.A. Emails: oui4@cdc.gov, wdq7@cdc.gov and jdp3@cdc.gov
Acknowledgments: The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

© Statistics Sweden

indicative of response bias (Groves and Peytcheva 2008; Brick and Tourangeau 2017), selection bias has been of great concern because the composition of web panels often differs from that of the underlying population. Panel members tend to be more educated and to have higher socioeconomic status than non-panel-members (Craig et al. 2013). Epidemiologic target samples often recruit "healthy volunteers" with lower estimates of disease incidence and mortality that a general population (Pinsky et al. 2007). To reduce the selection bias of the target samples, various propensity score (PS)-based adjustment methods have been developed which use an existing high-quality probability sample (e.g., national representative surveys) as a reference, where high quality refers to probability-sampled surveys with relatively low sampling and non-sampling errors that lead to confidence in their ability to produce representative estimates (Groves 1989). Recent PS adjustment methods include, but are not limited to, PS weighting (Valliant 2020; Chen et al. 2019) and PS matching (Kern et al. 2021) methods.

The amount of bias reduction, however, varies depending on the outcome and variables that are collected in both the target and reference data sources. Wang et al. (2020a) studied the bias reduction of different PS adjustment methods using the non-representative U.S. National Institutes of Health – American Association of Retired Persons cohort (NIH-AARP 2006), with the National Health Interview Survey (NHIS) as the reference survey. Among the ten selected diseases examined, they found the amount of relative bias reduction ranged from 8% to 30% using their proposed PS-based kernel weighting (KW) method. There is still a large amount of bias that is not removable by PS adjustment methods alone due to the uncollected information in the reference probability sample. The effectiveness of PS adjustment methods depends on the identification of the proper set of covariates, their availability, and their quality (Baker et al. 2013). Some references (e.g., Mercer et al. 2018) have even argued that choosing the correct variables can be more important than choosing the correct adjustment models, including PS methods.

High-quality probability samples surveyed through well-designed questionnaires are in great demand as reference surveys for at least two reasons: (1) Different PS adjustment methods, including PS-based weighting and matching methods, require a high-quality probability sample as the reference in order to create a set of pseudo-weights for the target sample to better represent the underlying target population; (2) Different target samples may use a common high-quality probability sample as the reference for cost efficiency by using the same questions with exact wordings to avoid potential reporting/measurement error. Given a high-quality population representative reference survey, we are interested in identifying the types of variables that are critical for collection in the target sample to improve its external validity in estimating population quantities. The findings can be used in turn to plan for future surveys.

The target sample motivating this research is collected through the National Center for Health Statistics' (NCHS) Research and Development Survey (RANDS), a probability-based panel survey that has been conducted using online and phone administration (www.cdc.gov/nchs/rands). Although the RANDS data are more structured than nonprobability samples, RANDS is subject to potential selection bias as RANDS has lower response rates, as well as potential measurement and coverage errors compared to traditional interviewer-administered national population health surveys (Parker et al. 2020). On the other hand, probability survey panels such as RANDS have the potential to

produce more timely information than national population surveys. To reduce the potential selection bias in RANDS estimates, NCHS' NHIS has been used as a reference sample to construct pseudo-weights using PS-based weighting methods (Parker et al. 2020; Irimata et al. 2020) and raking. These adjustments have been applied to the estimation of several population health outcomes, including diagnosed asthma, diagnosed hypertension, diagnosed diabetes, health insurance, as well as for health outcomes related to the coronavirus disease 2019 (COVID-19) pandemic such as access to health care. In these studies, adjustment to the NHIS (PS-based weighting or raking) has typically been performed using the main effects for all common covariates between RANDS and the NHIS, including sociodemographic, health, and internet use variables; the adjustment for RANDS during COVID-19 used a limited subset of variables for the public release of COVID-19-related estimates (www.cdc.gov/nchs/covid19/rands.htm). While PS weighting and raking adjustments have been shown to improve RANDS estimates relative to those without any adjustment to the NHIS (Parker et al. 2020; Irimata et al. 2020), stability of the estimated propensity scores and how the inclusion of different variables in the propensity model or calibration affects bias and efficiency of the estimated population mean for various outcomes have been a major concern.

Propensity model variable inclusion has been widely studied in different areas, including clinical trial or medical research and survey research. In clinical trial research, participants are included for clinical and experimental purposes (mainly for treatment effect estimation) and are not necessarily representative of the U.S. population. Simulations (Brookhart et al. 2006; Leyrat et al. 2013) were performed to examine the effect of the choice of variables that are included in a propensity model has on the bias, variance, and mean squared error of estimated treatment effects. It was concluded that omitting confounding factors increases bias and the inclusion of variables that are independent of the exposure but related to the outcome in the propensity model gains efficiency without increasing bias of estimated treatment effects. However, covariate inclusion for propensity score models in clinical trial research has been limited. Ali et al. (2015) provided a systematic review of covariate inclusion in the PS model for medical studies and concluded that the quality of reporting variable inclusion is far from optimal in the medical literature. Similarly, Grose et al. (2020) found 90% out of 303 systematically reviewed studies did not provide justification for covariates included in their PS models.

In survey research, propensity analyses have been conducted to estimate response propensity (Groves 2006; Iannacchione et al. 1991) and to adjust sampling weights in representative surveys to reduce the estimation bias due to unit nonresponse. The best auxiliary variables to be included for nonresponse adjustment are those simultaneously correlated with response propensity and the key survey outcomes (Lessler and Kalsbeek 1992). Little and Vartivarian (2005) further suggested that most important feature of variables for inclusion is that the variables are predictive of survey outcomes; prediction of response propensity is a secondary, though useful, goal.

This article, in contrast to the interest of estimating treatment effects in clinical research, aims to estimate population quantities such as the population mean. We are interested in identifying key auxiliary information in a reference probability survey to improve the external validity of inferences from a target data set. This is an important obligation for survey designers because the choice and inclusion of these variables has a tremendous

effect on both the bias and the precision of the estimates of population quantities. This differs from the goal of nonresponse adjustment which uses chosen covariates for predicting response propensities as, in nonresponse adjustment research, respondents are nested within the sampled units, and respondents and nonrespondents share common sampling design variables. As a result, unweighted analysis of response propensity can be performed conditional on the design and response predictive variables (Little and Vartivarian 2003). However, this is not true for estimating the propensity of target sample inclusion because the reference survey and the target samples are often independent without sharing design variables (Wang et al. 2021). Variable choice for the propensity model used to predict the target sample inclusion propensity should be performed with additional care.

This article aims to examine how different types of variables included in a propensity model impact the performance of population mean estimation using target samples through the directed acyclic graph (DAG), a common graphical tool in causal studies but largely under-utilized in survey research. The DAG is used to identify certain types of variables in the causal pathway to be included in the PS model which results in the lowest bias and highest precision under various scenarios. Estimated population means and their variances are evaluated analytically and numerically under various mis-specified propensity models, including with and without interactive effects. Different levels of variable correlations in the finite population are considered to mimic real data scenarios. The findings are applied to RANDS, with NHIS as the reference, to estimate the prevalence of asthma in the U.S. The RANDS evaluation demonstrates the advantage of this approach compared to the approach when the propensity model includes all available variables.

The results from this research provide insight for data analysts on propensity model construction to improve the population representativeness of target samples. It also provides insight for questionnaire designers on the critical auxiliary information to collect from the reference survey. NSOs, using the article results, can design the questionnaires for both the target and reference surveys and release accurate estimates for priority topics from more timely data sources.

2. Methods

We first introduce some notation. Suppose Y is a binary outcome of interest (e.g., for estimating the prevalence of a disease or health condition: $Y = 1$ if event and 0 otherwise). In the context of survey sampling, suppose A is the binary selection indicator variable (i.e., $A = 1$ if a population unit participates in the target sample and 0 otherwise). Note A indicates the target sample participation with value of one representing population units who are recruited and respond to the survey.

We adapt the framework of Brookhart et al. (2006) of employing a directed acyclic graph (DAG) to study potential selection bias induced by three types of covariates (see Figure 1a):

1. variables related to both the outcome Y and the selection indictor A of the target sample − confounders (X_1),
2. variables related to Y but not related to A – outcome predictors (X_2), and
3. variables related to A but not related to Y – selection variables (X_3).

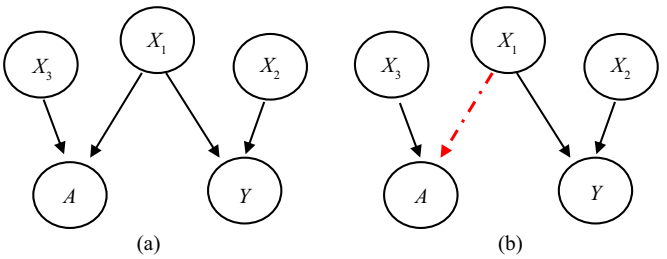


Fig. 1. DAG for three types of covariates with the selection indicator (A) and the outcome (Y).

We now present some background about PS adjustment methods. For estimation of the finite population (FP) mean of a binary outcome $E(Y) = p(Y = 1)$, the naïve unweighted estimate using the selected target sample ($A = 1$) has bias relative to the FP, given by $Bias = p\big(Y = 1|A = 1\big) - p(Y = 1) = (R - 1)p(Y = 1)$ with $R = \frac{p\big(A=1|Y=1\big)}{p(A=1)}$. In order to remove the bias, it requires the conditional distribution of A given Y is the same as the distribution of A, denoted by $A \perp Y$ and adjustment methods based on PS are often employed (Valliant 2020).

More specifically in PS-based adjustment methods, the population mean μ of the outcome Y, is estimated by

$$\hat{\mu}(x) = \frac{1}{\sum_{j\in S_c} w_j(x)} \sum_{j\in S_c} w_j(x) y_j, \tag{1}$$

where S_c denotes the set of sample units in the target sample of size n_c; y_j for $j \in S_c$ and x are, respectively, realized values of the outcome Y and X; the pseudo-weights $w_j(x)$ for $j \in S_c$ is constructed to balance the distribution in covariates between the target sample and the reference survey. Note X can be a single covariate or a vector of covariates from X_1- X_3, and are available in both the target sample and the reference survey, while the outcome variable Y is available in the target sample S_c only.

Various PS-based adjustment methods, including PS weighting and PS matching methods, have been developed under the following assumptions. First, the reference survey sample (in our real data example, the NHIS), through weighting, properly represents the target population of interest. Second, all finite population units have a positive participation rate (i.e., each individual in the population has a positive propensity to volunteer to participate in RANDS panel). Third, conditional exchangeability holds with no unmeasured confounders, that is, the probability for each individual in the FP to participate in the target sample is not related to his/her outcome, after adjusting for all measured variables. It is a common practice that the variables in the target sample are measured using same question wordings as in the reference survey to avoid potential reporting or measurement error.

While PS weighting and PS matching methods have similar assumptions, PS *weighting* methods construct the pseudo-weight by the inverse of the inclusion probability conditional on x, that is, $w_j = \frac{1}{e(x_j)}$ for $j \in S_c$, with $e(x) = p\big(A = 1|x\big)$, the target sample inclusion probability conditional on x. It can be verified that $A \perp x|e(x)$. In contrast, PS

matching methods distribute the survey sample weights to target sample units that have similar predicted propensity scores. For example, the KW method (Wang et al. 2021) first assigns the sample weight of each survey unit, say unit i, to cohort members proportionally according to kernel distances, defined by propensity scores $K\left(\frac{e(x_i)-e(x_j)}{h}\right)$ for $j \in S_c$, where $K(.)$ a kernel function such as the standard normal density function, and h is the band width selected by Silverman's rule of thumb method (Silverman 1986). As such, most of the sample weight for survey unit i is assigned to those cohort members with similar propensity scores. The assigned portions from survey members to cohort member j are then summed up to form the pseudo-weight w_j.

In Subsections 2.1–2.2, we assume that X_1, X_2, and X_3 are mutually independent in the FP and study how the PS-based adjustment methods reduce the bias and variance through the incorporation of different types of variables in the propensity models. We further consider real situations in Subsection 2.3 when different types of variables are correlated in the FP using DAG. Although DAG is a graphical tool developed for causal interpretation, we used it to rule out possible confounding and identify a conditioning covariate set for $Y \perp A$. The actual causation is not important in this context. Subsection 2.4 summarizes some practical guidelines for identifying the variable types in real data and choosing between PS-based methods to construct pseudo-weights when covariates interactively affect the target sample participation and the outcome.

2.1. *Bias of $\hat{\mu}(x)$ By Various Types of Covariates*

It is readily shown in Figure 1a that the confounders X_1 induce the bias when we use the simple sample mean to estimate the population mean $p(Y=1)$. Intuitively, the information can be exchanged between the two nodes of A and Y through X_1, but not X_2 or X_3. This result is consistent with the bias calculation below. For selection variables $(X = X_3)$ or predictors $(X = X_2)$, we have

$$R = \frac{\sum_x \left[p(Y=1|X=x)p(A=1|X=x)p(X=x)\right]}{p(Y=1)\sum_x \left[p(A=1|X=x)p(X=x)\right]} = 1 \text{ and hence Bias} = 0.$$

For confounders, however, $Bias \neq 0$. To correct for the bias induced by X_1, PS-based adjustment methods create pseudo-weights and reweight the target sample such that the weighted sample distribution of the confounder X_1 is same as that in the FP, that is, $X_1 \perp A$ as shown in Figure 1b. The dotted line denotes the path X_1-A is blocked (i.e., there is no information exchange between the two nodes) by reweighting the target sample and hence $A \perp Y$.

As a result, the estimator $\hat{\mu}(x_1)$ withthe set of pseudo-weights $w(x_1)$, where x_1 is the realized value of the confounder X_1, is approximately unbiased. Analogously, it is readily shown that the estimator $\hat{\mu}$ with pseudo-weights defined by the inverse of sample inclusion probabilities that balance the x_1 distribution between the target sample and the FP, including $e(x_1, x_2)$, $e(x_1, x_3)$, or $e(x_1, x_2, x_3)$, is also unbiased. Note that the three sets of pseudo-weights of $w(x_1, x_2)$, $w(x_1, x_3)$, or $w(x_1, x_2, x_3)$ balance the x_1 distribution and also the distribution of x_2, x_3, or x_2 and x_3, respectively, between the target sample S_c and the FP (Rosenbaum and Rubin 1983).

In contrast, pseudo-weights of $w(x_2)$, $w(x_3)$ or $w(x_2, x_3)$ do not balance the X_1 distribution and therefore the corresponding weighted estimators in Equation (1) are biased.

2.2. *Variance of $\hat{\mu}(x)$ By Various Types of Covariates*

Among the four unbiased estimators based on $e(x_1)$, $e(x_1, x_2)$, $e(x_1, x_3)$, and $e(x_1, x_2, x_3)$, we compare their efficiencies. We first compare the variance of $\hat{\mu}(x)$ with x the realized value of $X = X_1$ versus $X = (X_1, X_3)$, denoted by $V(\hat{\mu}(x_1))$ and $V(\hat{\mu}(x_1, x_3))$, respectively. We write

$$V(\hat{\mu}(x_1)) = V\left(\frac{1}{\sum_{j\in S_c} w_j(x_1)} \sum_{j\in S_c} w_j(x_1) y_j\right), \text{ and}$$

$$V(\hat{\mu}(x_1, x_3)) = V\left(\frac{1}{\sum_{j\in S_c} w_j(x_1, x_3)} \sum_{j\in S_c} w_j(x_1, x_3) y_j\right)$$

Note the selection variable is independent of the outcome and thus the pseudo-weights based on x_3 are non-informativeof the outcome Y. The corresponding pseudo-weighted mean, although adding no bias, loses efficiency due to the differential non-informative pseudo-weights. Taking the adjusted logistic propensity pseudo-weights (denoted by ALP in Wang et al. 2021) as an example, under the logistic regression propensity model

$$\log \frac{p_j}{1 - p_j} = \log \pi_j = \beta_x^T x_j, \tag{2}$$

where $p_j = p(j \in S_c^* | FP^*)$ and $FP^* = S_c^* \cup FP$ denotes the pseudopopulation by combining S_c^* (i.e., a copy of the target sample S_c) and the FP, and β_x is the regression coefficient associated with x. The ALP pseudo-weight $w_j(x)$ is constructed as $w_j(x) = \exp^{-1}(\beta_x^T x_j)$ for $j \in S_c$.

For simple illustration, assume model (2) includes main effects of covariates, so $w_j(x_1, x_3) = w_j(x_1) w_j(x_3)$ and $w_j(x_3)$ are noninformative weights since $x_3 \perp y$. Under the assumption that the variance of the observations is approximately constant (Kish 1992), the proportional increase in variance from weighting, denoted by L, is approximated to be

$$L = \frac{V(\hat{\mu}(x_1, x_3))}{V(\hat{\mu}(x_1))} - 1 = CV^2(w_j(x_3)) > 0$$

where CV is the coefficient of variation of the $w_j(x_3)$ weights. Thus, $V(\hat{\mu}(x_1, x_3)) > V(\hat{\mu}(x_1))$.

Note that the model parameter $\beta_x = 0$ in Equation (2) if x is an outcome predictor, which does not predict the target sample membership A, such as x_2 in Figure 1a. As a result, $w(x_1, x_2) = \exp^{-1}\left(\beta_{x_1}^T x_1 + 0\right) = w(x_1)$ and thus

$$V\left(\hat{\mu}(x_1,x_2)=\frac{1}{\sum_{j\in S_c}w_j(x_1,x_2)}\sum_{j\in S_c}w_j(x_1,x_2)y_j\right)$$

$$=V\left(\hat{\mu}(x_1)=\frac{1}{\sum_{j\in S_c}w_j(x_1)}\sum_{j\in S_c}w_j(x_1)y_j\right).$$

Along the same line of justification, $V\left(\hat{\mu}(x_1,x_3)\right)=V\left(\hat{\mu}(x_1,x_2,x_3)\right)$. In conclusion,

$$V\left(\hat{\mu}(x_1)\right)=V\left(\hat{\mu}(x_1,x_2)\right)<V\left(\hat{\mu}(x_1,x_3)\right)=V\left(\hat{\mu}(x_1,x_2,x_3)\right)$$

In summary, to achieve unbiasedness and efficiency of pseudo-weighted mean estimators, the propensity model that considers confounders (X_1) alone, or together with outcome predictors (X_2), should be used to construct the pseudo-weights in Equation (1). The resulting mean estimates are unbiased and most efficient. The inclusion of selection variables in the propensity model, in addition to all confounders, adds no more bias, however, will inflate the variances of the estimates. In contrast, the inclusion of outcome predictors does not inflate the variance while retaining the unbiasedness of the FP mean estimates. A short version of the recommendation for PS-based pseudo-weights construction: include all confounders but avoid selection variables in the propensity model.

The above justification assumesthe logistic regression model (2) with main effects is true. More rigorous justification is needed when different types of covariates are correlated; propensity models are mis-specified (that is, the logistic regression model (2) is not the true model); and the pseudo-weights (i.e., propensity model coefficients) are unknown and have to be estimated.

2.3. *Correlation Between Covariates*

We now consider more realistic scenarios in which the confounders, the outcome predictors, and the selection variables can be correlated to each other. Figure 2 shows cases where correlation exists between the pairs X_1 and X_3, X_1 and X_2, and X_2 and X_3, respectively. In addition, any two or all three pairs can be correlated simultaneously in the FP.

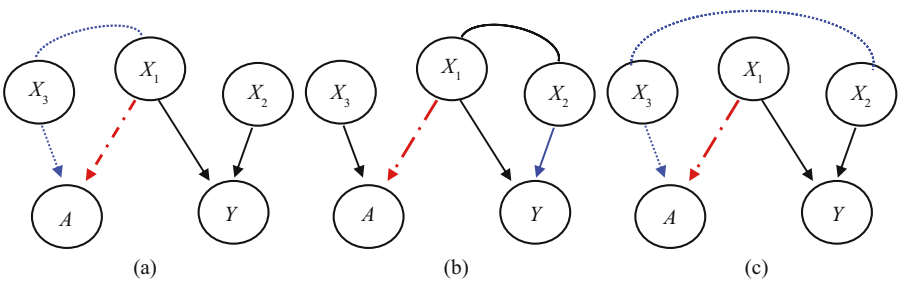


Fig. 2. DAG for $Y \perp A$ when correlation exists between $X_1 - X_3$ (a), $X_1 - X_2$ (b) and $X_2 - X_3$ (c). Blocking dotted path(s) to have $Y \perp A$.

For unbiased estimation of the FP mean of Y using the target sample (A = 1), we need to block all paths connecting A and Y such that $Y \perp A$. We focus on paths that point to A since in the propensity model we construct weights for the target sample units (with A = 1) so that the weighted target sample and the FP have same distributions in certain covariates X, that is, A ⊥ X|e(X).

As shown by the dotted lines in Figure 2, two paths (indicated by two different types of dotted line) in Figure 2a and 2c, and one path in Figure 2b are identified and need to be blocked, that is, prevent information flow between A and Y, in order to achieve $Y \perp A$. The backdoor criteria (Pearl 2009) is a way to rule out confounding via conditioning, and allows identifying the causal effect from A to Y (equal to zero in Figure 2, i.e., $Y \perp A$) after conditioning a set of covariates that block the backdoor paths between A and Y. Here, the identified dotted paths in Figure 2 are backdoor paths because the arrows point into A (not the opposite direction if arrows point from A towards X_1-X_3). By the backdoor criteria, X_1 blocks the identified paths in Figure 2a-b. As follows, we construct PS-based pseudo-weights $w(x_1)$ so that the X_1 distribution in the weighted target sample is same as that in the FP in Figure 2a-b when X_1 and X_3 or X_1 and X_2 are correlated (i.e., $\rho_{x_1x_3} \neq 0$ or $\rho_{x_1x_2} \neq 0$). Thus, the $w(x_1)$ -weighted target sample mean of Y is an unbiased estimator of the FP mean. In Figure 2c, X_2 or X_3, in addition to X_1, block the two identified paths (Pearl 2009). Following the same logic, pseudo-weights that balance the distributions in X_2 or X_3, in addition to X_1, denoted by $w(x_1, x_2)$ or $w(x_1, x_3)$, should be constructed for the target sample units when the pair of X_2 and X_3 are correlated in Figure 2c ($\rho_{x_2x_3} \neq 0$). This result also applies to cases when any two pairs or all three pairs of covariates are simultaneously correlated in the FP, and the $w(x_1,x_2)-$ or $w(x_1,x_3)$-pseudo-weighted target sample means are approximately unbiased.

In summary, similar to the scenario shown in Figure 2, in order to block the dotted paths when covariates interactively affect the outcome, PS-based adjustment methods can be applied to construct pseudo-weights that balance the distributions of X_1 (Figure 2a-b) and (X_1, X_2) or (X_1, X_3) (Figure 2c) in the pseudo-weighted sample and the FP.

2.4. Practical Guidelines

In practice, the variable types (confounder, predictor, selection variable) need to be identified for propensity model construction. Since we are not concerned about model interpretation, parametric models with complex functional forms or nonparametric models can be fitted. In our RANDS example (Section 4), both the outcome and propensity models were selected by automatic backward selection methods. Starting from the full model containing all factors and their pairwise interaction terms, we removed the interaction term with the largest p-value and re-fit the model. We continued the iterations until all p-values of the interaction terms were less than 0.05. For each interaction, complex survey designs were accounted for in the logistic regression analysis using the svyglm() function in the R survey package (Lumley 2020). The main effects with p-values greater than 0.05 were removed only if they were not involved in any of significant interaction terms. As results, each type of variables is identified: confounders are common terms in both the selected propensity and the outcome models; the selection variables (or predictors) are those selected in the propensity (or outcome) model only. Note that each type of variables may

contain multiple variables as well as their nonlinear or nonadditive functions (e.g., pairwise interactions) in the final outcome and the propensity models.

Alternative model selection criteria can be employed, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Lumley and Scott 2015) to identify variable types. Yang et al. (2020) and Chen et al. (2018) have also proposed variable selection methods, including penalized estimating equations or LASSO regression, which can be used to identify variable types for inclusion in the PS model. In the case where the outcome of interest is not available in the reference probability sample or the outcome has not yet been collected in the target sample, subject matter literature and knowledge may have to be used to assign the covariate types. Variable type identification is critical in practice and comparing different model selection methods to create the final models is of future research interest.

The true propensity model of the underlying selection mechanism of the target sample (A = 1) is often unknown but complicated, which may involve covariate terms of higher orders of nonlinearity and/or nonadditivity. For example, X_1 and X_2 (or X_1 and X_3) can interactively affect the outcome Y (or selection indicator A), the scenario considered in simulation study 3 (to be shown in Section 3). In order to estimate the propensity scores accurately to achieve the covariate balance so that the condition of $Y \perp A$ holds, data analysts need to be careful in choosing the PS-based adjustment methods among PS weighting or matching methods based on parametric models such as logistic regression and nonparametric methods such as machine learning.

For example, PS weighting methods (such as the ALP) can be sensitive to model misspecification (Wang et al. 2020a). ALP-weighted target sample distributions match the FP distributions when the assumed propensity model is true. For instance, ALP pseudo-weights that are constructed based on the propensity model of A on X_1, X_3, and their interaction $X_1 * X_3$, produce unbiased estimators. The estimators, however, are biased if the model is misspecified, for example, the interaction term is omitted from the propensity model.

In contrast, PS *matching* methods construct weights by matching target sample units with reference sample units using the estimated propensity scores, followed by distributing reference sample weights to target sample units with similar propensities. It has been well recognized that PS matching, compared to PS weighting methods, is more robust to model misspecification (Wang et al. 2020a). In our limited simulation studies, it is shown that the balance in covariate distributions between the KW-weighted sample and the FP can be achieved as long as the blocking variables, that is, X_1 in Figure 2a-b; X_1 and X_3 or X_1 and X_2 in Figure 2c, are included (with or without $X_1 * X_3$ interaction) in the propensity model (as shown in Figure 3). For complicated propensity models with higher orders of interactions and/or nonlinearity, including only the main effects of the blocking variables in the propensity model by KW methods might not be sufficient. Nonparametric modeling such as machine learning methods maybe promising (Kern et al. 2021) in identifying nonlinear or nonadditive relationships of covariates with the target sample selection.

3. Simulation

Simulation studies were conducted to evaluate the performance of the mean estimator from Equation (1) with the pseudo-weights constructed by the ALP and the KW methods based on propensity models that consider different types of covariates.

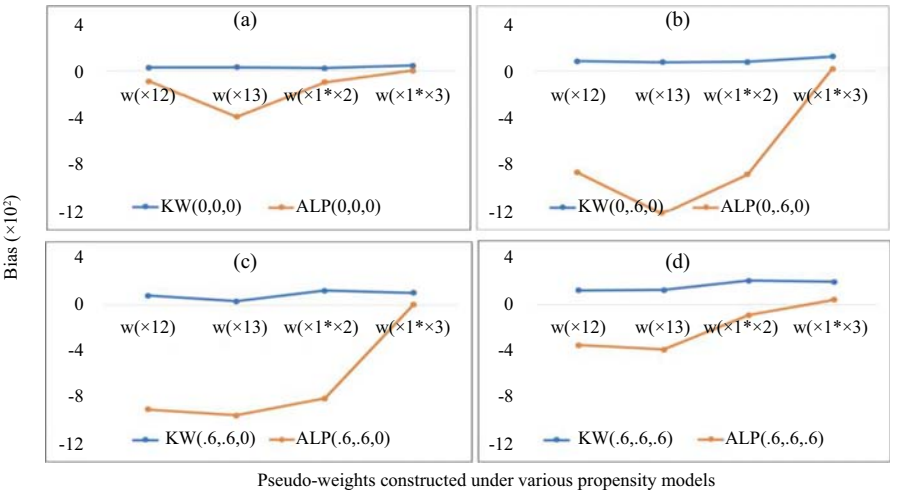


Fig. 3. Bias of Kernel Weighting (KW) vs. Adjusted Logistic Propensity (ALP) Estimated under Various Propensity Models with $w(x_{12})$, $w(x_{13})$, $w(x_1{*}x_2)$, and $w(x_1{*}x_3)$ including, respectively, main effects of X_1 and X_2, main effects of X_1 and X_3, main and interactive effects of X_1 and X_2, and main and interactive effects of X_1 and X_3, with pairwise covariate correlations $(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0,0,0)$ (a), (0,.6,0) (b), (.6,.6,0) (c), and (.6,.6,.6) (d), in the FP. Propensity model with $w(x_1{*}x_3)$ is the true model.

3.1. Population Generation

We generate a finite population $FP = \{X_{1i}, X_{2i}, X_{3i}, Y_i \text{ for } i = 1, \ldots N\}$ with population size $N = 20{,}000$. Three covariates X_1, X_2, and X_3 follow standard trivariate normal distributions with pairwise correlations $\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$. A binary outcome Y is generated following the Bernoulli distribution with a mean of

$$p(Y=1) = \frac{\exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{12} X_1 X_2)}{1 + \exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{12} X_1 X_2)}.$$

We specify $(\alpha_0, \alpha_1, \alpha_2) = (-1, .5, .5)$ so that x_1 and x_2 are associated with Y as in Figure 1, but vary $\alpha_{12} =$ 0.5 or 0 with and without the interaction term. As a result, the FP mean $\bar{Y} \approx 0.29$.

3.2. Selection of the Target Sample (with A = 1)

A sample of size $n_c = 1{,}000$, denoted by S_c, is selected from the FP, using the design of probability proportional to size (PPS) sampling with measure of size (*mos*): $mos = \exp(\beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{13} X_1 X_3)$ so that the inclusion probability is

$$p(j \in S_c | FP) = \frac{n_c \times mos_j}{\sum_{i \in FP} (mos)_i} \propto \exp(\beta_1 X_{1j} + \beta_3 X_{3j} + \beta_{13} X_{1j} X_{3j}).$$

We specify $\beta = (\beta_0, \beta_1, \beta_3) = (-1, .5, .5)$ so that x_1 and x_3 are associated with A as in Figure 1. In addition, we vary $\beta_{13} = .5$ or 0 with or without the interactive effect in the propensity model. We have the target sample participation rate of $E(A) = .05$.

The inclusion probabilities (i.e., sample weights) are masked in the analysis and treated as unknown (i.e., equal sample weights of 1 used). Note that the target sample without weights is not representative of the population.

3.3. Selection of a Probability Sample

An independent probability sample of size $n_s = 500$, denoted by S_s, is selected using the same sampling design as the target sample selection. The selected probability sample has known selection probabilities. The weighted probability sample is used as the reference survey, representing the underlying FP in the propensity analysis.

Pseudo-weighted means, that is, (1), with estimated pseudo-weights constructed under different propensity models, including the confounders (X_1), outcome predictors (X_2), the selection variables (X_3), and/or their interactions, were compared. Three simulation studies are conducted with results presented in Tables 1–2 and Figure 3. Simulation 1 considers a simple scenario of independent covariates in the FP (with $\rho_{x_1x_2} = \rho_{x_1x_3} = \rho_{x_2x_3} = 0$) without interaction effects of covariates on the outcome or the target sample inclusion (i.e., $\alpha_{12} = \beta_{13} = 0$). Simulation 2 varies the covariate correlation in the FP by

$$\left(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}\right) = \left(.6, 0, 0\right), \left(0, .6, 0\right), \left(0, 0, .6\right),$$

$$\left(.6, .6, 0\right), \left(.6, 0, .6\right), \left(0, .6, .6\right), \text{ or } \left(.6, .6, .6\right),$$

while keeping $\alpha_{12} = \beta_{13} = 0$. Simulation 3 further complicates the underlying outcome model and the propensity model by including the interaction terms with $\alpha_{12} = \beta_{13} = 0.5$.

Tables 1–2 show the bias, empirical variance (EmpVar), and MSE of the KW estimate, over B = 500 iterations, from simulations 1–2, respectively, and

$$bias = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^{(b)} - \bar{Y};\ EmpVar = (B-1)^{-1}\sum_{b=1}^{B}\left\{\hat{\mu}^{(b)} - B^{-1}\sum_{b=1}^{B}\hat{\mu}^{(b)}\right\}^2;\ \text{and}$$

$$\text{MSE} = \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\mu}^{(b)} - \bar{Y}\right\}^2,$$

where $\hat{\mu}^{(b)}$ is the KW estimate of the population mean using the b$^{\text{th}}$ simulated target sample under various analytical propensity models. The w(x_1), w(x_{12}), and w(x_{13}) denote the propensity models including main effects of, respectively, x_1, x_1 and x_2, x_1 and x_3.

Table 1. Results from population mean estimation[1] under various propensity score models[2] with covariate correlations $\left(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}\right) = (0, 0, 0)$ *and interaction effects* $\alpha_{12} = \beta_{13} = 0$.

	Sample[3]	w(x_1)	w(x_2)	w(x_3)	w(x_{12})	w(x_{13})	w(x_{23})
Bias ($\times 10^2$)	4.61	0.26	4.50	4.83	0.26	0.41	4.77
EmpVar ($\times 10^4$)	2.20	2.68	2.62	2.96	2.92	3.43	3.32
MSE ($\times 10^4$)	23.48	2.75	22.85	26.31	2.99	3.60	26.04

[1]Kernel weighting estimator (Wang et al. 2020b) is applied for population mean estimation.
[2]w(x_1), w(x_2), w(x_3), w(x_{12}), w(x_{13}), and w(x_{23}) denote pseudo-weighted means with pseudo-weights constructed under the propensity model with main effect(s) of x_1, x_2, x_3, x_1 and x_2, x_1 and x_3, and x_2 and x_3, respectively.
[3]Sample denotes the unweighted mean

Table 2. Results from population mean estimation[1] under various propensity score models[2] by covariate correlations with interaction effects $\alpha_{12} = \beta_{13} = 0$.

	Sample[3]	w(x_1)	w(x_2)	w(x_3)	w(x_{12})	w(x_{13})	w(x_{23})
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6, 0, 0)$					
Bias ($\times 10^2$)	7.35	0.37	2.98	7.60	0.37	0.59	3.25
EmpVar ($\times 10^4$)	2.15	2.59	2.64	2.77	2.66	2.88	2.84
MSE ($\times 10^4$)	56.14	2.72	11.52	60.57	2.79	3.23	13.42
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, .6, 0)$					
Bias ($\times 10^2$)	7.27	0.32	7.16	3.21	0.30	0.41	3.12
EmpVar ($\times 10^4$)	2.17	3.60	2.39	3.68	3.53	4.05	3.66
MSE ($\times 10^4$)	54.98	3.70	53.67	13.97	3.62	4.22	13.39
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, 0, .6)$					
Bias ($\times 10^2$)	7.55	2.98	4.65	4.87	0.26	0.37	4.83
EmpVar ($\times 10^4$)	2.01	2.52	2.38	2.57	2.66	2.75	2.66
MSE ($\times 10^4$)	59.00	11.38	24.00	26.30	2.73	2.89	26.03
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6, .6, 0)$					
Bias ($\times 10^2$)	9.81	-1.04	5.36	6.09	0.54	0.54	1.70
EmpVar ($\times 10^4$)	2.16	3.45	2.94	3.79	3.94	3.98	4.19
MSE ($\times 10^4$)	98.39	4.52	31.61	40.93	4.23	4.27	7.09
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6, 0, .6)$					
Bias ($\times 10^2$)	10.27	3.11	1.67	7.60	0.50	0.60	2.36
EmpVar ($\times 10^4$)	2.33	2.84	2.96	2.94	2.93	3.11	3.00
MSE ($\times 10^4$)	107.86	12.51	5.76	60.72	3.18	3.46	8.58
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, .6, .6)$					
Bias ($\times 10^2$)	10.09	3.11	7.24	1.56	0.33	0.44	2.34
EmpVar ($\times 10^4$)	1.98	3.65	2.63	3.28	3.45	3.61	3.48
MSE ($\times 10^4$)	103.83	13.33	54.98	5.71	3.55	3.80	8.97
		$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6, .6, .6)$					
Bias ($\times 10^2$)	13.28	1.73	4.52	4.81	0.77	0.89	3.30
EmpVar ($\times 10^4$)	1.98	3.93	3.39	3.88	3.91	4.29	3.97
MSE ($\times 10^4$)	178.29	6.91	23.83	26.99	4.50	5.07	14.88

[1]Kernel weighting estimator (Wang et al. 2020b) is applied for population mean estimation.
[2]w(x_1), w(x_2), w(x_3), w(x_{12}), w(x_{13}), and w(x_{23}) denote pseudo-weighted means with pseudo-weights constructed under the propensity model with main effect(s) of x_1, x_2, x_3, x_1 and x_2, x_1 and x_3, and x_2 and x_3, respectively.
[3]Sample denotes the unweighted mean.

Models including x_2 only, x_3 only, and x_2 and x_3 are denoted as w(x_2), w(x_3), and w(x_{23}), respectively.

Three observations are made in Table 1. *Firstly*, consistent with our expectations, all propensity models that include the confounder x_1, that is, w(x_1), w(x_{12}), w(x_{13}), produce approximately unbiased estimates of the FP mean of Y; the estimates are badly biased under the propensity models which include x_2 only, x_3 only, or x_2 and x_3. *Secondly*, the propensity model w(x_3) yields inflated variance estimates compared to w(x_1) or w(x_2), and w(x_2) has the smallest empirical variances. *Thirdly*, among the three approximately unbiased estimators, w(x_1) yields the most efficient estimates relative to w(x_{12}) or w(x_{13}).

Table 2 presents results from simulation 2 with varying covariate correlations. Three observations are made. *Firstly*, pseudo-weights that balance the distributions in x_2 or x_3, in addition to x_1, produced approximately unbiased estimates across various correlations; see the shaded two columns of w(x_{12}) and w(x_{13}). *Secondly*, among the two, w(x_{12}) and w(x_{13}), the empirical variance estimates and MSEs under w(x_{12}) tend to be smaller than those under w(x_{13}). *Thirdly*, the inclusion of only the confounder x_1 in the propensity model, that

is, $w(x_1)$, although efficient, may induce bias, especially when correlation exists between x_2 and x_3.

Simulation 3 compares biases of estimated population means by the KW matching method and the ALP weighting method when the underlying outcome and propensity models include the interaction terms, i.e., $\alpha_{12} = \beta_{13} = 0.5$ (see Figure 3). Four analytic propensity models, including X_2 or X_3 in addition to the confounder x_1, are considered and they are (1) w(x_{12}), X_1 and X_2 main effects only, (2) w(x_{13}), X_1 and X_3 main effects only, (3) w(x_1*x_2), X_1 and X_2 main effects and their interaction, and (4) w(x_1*x_3), including X_1 and X_3 main effects and their interaction. Recall KW is a type of PS matching method and expected to be more robust to model misspecification compared to the ALP method. As expected, the KW method consistently yields approximately unbiased estimates across four propensity models with or without interaction terms. In the contrast, the ALP approach directly uses the inverse of the participation rates estimated from the assumed propensity model as pseudo-weights, and the ALP estimates are approximately unbiased only under the true propensity model w(x_1*x_3). Furthermore, it can be observed that biases of the ALP estimates are consistently closer to zero than the KW under the true model. Results with covariate correlations $\left(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}\right) = \left(.6, 0, 0\right)$, $\left(0, 0, .6\right)$, $\left(.6, 0, .6\right)$ and $\left(0, .6, .6\right)$ showed a similar pattern and hence are not shown.

4. Real Data Analysis

RANDS, a series of web-based probability panel surveys conducted at NCHS (https://www.cdc.gov/nchs/rands), has been used for methodological research and, more recently, for providing early experimental estimates on the COVID-19 pandemic. RANDS has the capability to collect data quickly and is less costly than traditional national household surveys, but is subject to potential selection biasdue to low response rates. Adjustment methods to construct pseudo-weights, including propensity-score based methods, are applied to balance the covariate distributions in the target sample and the FP, and are an important component of the RANDS program. We consider the simulation findings from this article for selecting PS-model covariates to estimate the national prevalence of asthma compared to NCHS' NHIS.

Data from the third round of RANDS (RANDS 3) is evaluated. RANDS 3 was collected in 2019 using NORC's AmeriSpeak® Panel (https://amerispeak.norc.org) and included responses from 2,646 panelists aged 18 years and older. RANDS 3 panelists were surveyed via web and were asked questions related to general and mental health, medical conditions, opioid use, and pain. The RANDS 3 cumulative response rate was 18.1%. The RANDS 3 original panel weights were developed by the inverse of the probability of inclusion in the AmeriSpeak® Panel, subject to nonresponse adjustment and poststratification adjustment to external population totals of age, sex, education, race/ethnicity, housing tenure, telephone status, and Census Division (National Center for Health Statistics 2020). The original panel-weighted estimate of diagnosed asthma (ever been told you had asthma) in RANDS 3 was 16.86% (standard error = 0.98%). For comparison, the unweighted estimate of diagnosed asthma in RANDS 3 was 16.40% (standard error = 0.72%). The 2019 NHIS (n = 31,997) is evaluated as the gold standard. The NHIS (https://www.cdc.gov/nchs/nhis) is a cross-sectional household interview

survey that collects information on a broad range of health topics, primarily through face-to-face interviews. The NHIS sample adult file, which is a collection of responses from one randomly selected adult per selected household, was used to evaluate the prevalence of ever having asthma among adults. The percentage of adults who ever had asthma based on the 2019 NHIS (n = 31,997) was 13.46% (standard error = 0.25%).

Common covariates available in RANDS 3 and the 2019 NHIS that were potentially related to diagnosed asthma or the selection indicator were considered (see Table 3). All percent estimates in Table 3 (when expressed as proportions) meet the NCHS data presentation standards for proportions (Parker et al. 2017). As observed, demographic variables of age, sex and race/ethnicity have similar weighted distributionsin the RANDS and NHIS. This result is as expected, since these variables are poststratification variables used to construct the sample (or original panel) weights in both NHIS and RANDS. As observed, persons with higher levels of education, with selected health conditions (i.e., diagnosed high cholesterol, diagnosed chronic obstructive pulmonary disease (COPD), emphysema, or chronic bronchitis, diagnosed diabetes, and diagnosed hypertension), who are current or former smokers, or who are not married (with the exception of those who are widowed) participated in RANDS at a higher rate compared to the NHIS. Since the percent of missing values across all considered variables was relatively low for both data sources, ranging from 0%-0.68% for RANDS and 0%-2.64% for NHIS (unweighted), missing values were excluded for evaluation.

To check for correlation between covariates, bivariate correlations were assessed on the weighted NHIS data. Bivariate correlations for all selected covariates were statistically significant. Prior to evaluating the propensity models, the survey weights for both data sets were normalized to their respective sample sizes (n = 2,646 for RANDS, n = 31,997 for NHIS) as suggested by Li et al. (2011) and Wang et al. (2021). The KW method was implemented for demonstration to construct RANDS pseudo-weights that adjust for potential selection bias due to differential non-response and under-coverage of some groups on the sample frame using the NHIS data as the reference dataset.

A full propensity model (denoted by model. all) that includes all covariates and their pairwise interactions was used to create pseudo-weights. Due to the large number of parameters in the full model, estimated propensity scores can be unstable. As a result, some form of stepwise propensity model selection methods have been conducted in different studies (Weitzen et al. 2004; Austin 2008; Wang et al. 2020a), using the combined target sample and the reference sample to identify significant terms out of the full propensity model. In the framework of our article, the combination of the confounders and selection predictors (i.e., model.x13 which contains X_1 and X_3), which can be main effects of covariates or their nonlinear/nonadditive combinations such as pairwise interactions, are recommended as terms for inclusion. Based on the simulation results, we expect that the pseudo-weighted mean under model.x13 would be unbiased but with higher variability when compared withthe estimates under model.x12 that includes the confounders and outcome predictors.

Accordingly, we conducted the outcome model selection using backward selection on the reference survey (e.g., the 2019 NHIS), to identify terms which were confounders or outcome predictors. We defined the selected model as model.x12.n (contains X_1 and X_2) with "n" indicating that the outcome predictors were identified using the NHIS. However,

Table 3. Distribution of selected covariates and asthma in the Research and Development Survey (RANDS) 3 and the 2019 National Health Interview Survey (NHIS).

Variable	Subgroup	RANDS (n = 2,646)			NHIS (n = 31,997)	
		N	%	Wt %	n	Wt %
Outcome						
Ever diagnosed with asthma	Yes	431	16.4	16.9	4,229	13.5
	No	2,197	83.6	83.1	27,718	86.5
Covariates						
Age group (years)	18–34	721	27.2	29.9	7,058	29.7
	35–49	652	24.6	24.1	7,250	24.3
	50–64	687	26.0	25.1	8,313	24.9
	65+	586	22.1	20.9	9,376	21.1
Sex	Male	1,318	49.8	48.3	14,733	48.3
	Female	1,328	50.2	51.7	17,261	51.7
Race/ethnicity	Non-Hispanic white	1,729	65.3	63.1	21,915	63.2
	Non-Hispanic black	273	10.3	11.9	3,483	11.8
	Non-Hispanic other	227	8.6	8.5	2,447	8.5
	Hispanic	417	15.8	16.5	4,152	16.5
Marital status	Married	1,282	48.5	47.7	14,759	52.4
	Widowed	134	5.1	4.5	3,115	6.0
	Divorced	350	13.2	12.4	4,317	9.0
	Separated	50	1.9	1.8	456	1.2
	Never married	618	23.4	24.3	6,368	22.5
	Living with partner	212	8.0	9.3	2,136	8.9
Education level	High school diploma or less	577	21.8	38.8	11,155	39.9
	Some college	1,222	46.2	27.7	9,386	31.1
	Bachelor's degree or more	847	32.0	33.5	11,277	29.0
Smoking status[1]	Current	409	15.5	17.2	4,296	14.0
	Former	811	30.8	28.9	7,973	22.5
	Never	1,411	53.6	53.9	18,931	63.5
Diagnosed with high cholesterol	Yes	976	37.1	36.4	9,179	24.9
	No	1,657	62.9	63.6	22,697	75.1
Diagnosed with COPD, emphysema, or chronic bronchitis	Yes	213	8.1	8.4	1,787	4.6
	No	2,420	91.9	91.6	30,158	95.4
Diagnosed with diabetes[2]	Yes	279	10.6	10.5	3,355	9.3
	No	2,352	89.4	89.5	28,594	90.7
Diagnosed with hypertension	Yes	989	37.5	37.0	11,480	31.7
	No	1,648	62.5	63.0	20,458	68.3
Employment status	Paid employee	1,630	61.6	58.6	18,810	64.6
	Looking for work	166	6.3	7.2	485	2.0
	Not looking for work	850	32.1	34.2	11,919	33.4

Notes: n = unweighted sample size, % = unweighted percent, Wt % = weighted percent

[1]Smoking status: Current smoker is defined as someone who has smoked at least 100 cigarettes in their lifetime and now smokes every day or some days. Former smoker is defined as someone who has smoked at least 100 cigarettes in their lifetime and now does not smoke. Never smokers are defined as persons who have smoked less than 100 cigarettes in their lifetime.

[2]Diagnosed diabetes excludes pre − diabetes and gestational diabetes.

it is often the case that the reference probability surveys have no collected information on the outcome variable. In this case, we have only the target sample (e.g., RANDS) available for outcome model selection. With the assumption of the conditional noninformative sampling of the target sample, it is expected the unweighted regression of the outcome would produce unbiased estimates of regression coefficients (Korn and Graubard 1999). As follows, outcome model variable selection was conducted based on the unweighted outcome regression of the RANDS data, and the selected model included both confounders and outcome predictors, denoted by model.x12.r (contains X_1 and X_2) indicating that the outcome predictors were identified using RANDS. The common terms in model.x13 and model.x12 (denoted by either x12.n or x12.r based on the information available) are confounders, and the corresponding propensity model is denoted by model.x1. The identified covariate types under each model are reported in the Appendix (Section 6). Due to the correlation between x1, x2 and x3, we expect estimates under model.x1, albeit efficient, may not remove as much bias as under the model.x12 or model.x13.

The outcome models utilized the observations in the NHIS or the RANDS, whereas the propensity model utilized the observations in the combined NHIS and RANDS data, from which the estimated propensities were obtained and used for construction of the KW pseudo-weight for each individual in RANDS. Note that RANDS has panel weights, which were computed as an overall sampling weight for the selection of each panel member from the sampling frame and the selection of the panel member into RANDS. We considered two scenarios of (1) panel weights or (2) no panel weights for the propensity analysis.

Various propensity models that included different types of covariates were evaluated by the coefficient of variation (CV) of the KW pseudo-weights ($CV = sd(KW)/mean(KW)$ with *sd* denoting standard deviation), relative bias ($\text{relBias} = \frac{\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}}{\hat{\mu}_{NHIS}} \times 100\%$), standard error (se), and mean squared error ($\text{MSE} = \left(\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}\right)^2 + se^2\left(\hat{\mu}_{RANDS}\right)$). The relative bias was calculated as the estimated asthma prevalence in RANDS relative to the NHIS estimate where the RANDS estimate, $\hat{\mu}_{RANDS}$, was calculated using the KW pseudo-weights produced from the various propensity models. The standard error $se\left(\hat{\mu}_{RANDS}\right)$ considered the variability due to estimating the propensity scores, sampling, kernel weighting, as well as differential pseudo-weights by the Taylor linearization method (Wang et al. 2020b). For comparison purposes, we also report the relative bias, standard error, and MSE of the original panel-weighted and unweighted estimates of asthma prevalence in RANDS 3. Results are presented in Table 4.

Four observations can be made from Table 4. *Firstly*, all 12 (panel-weighted or unweighted) propensity-adjusted estimates perform better, with a smaller MSE (or relative bias), compared to the original panel-weighted RANDS estimate of asthma prevalence without PS adjustment. When the RANDS panel weights are considered in the propensity analysis, the standard errors tend to be inflated, relative to those in the lower pane, due to more variable KW pseudo-weights with their CVs ranging from 1.07-1.13 (see the upper pane) versus 0.69–0.83 (see the lower pane). Accordingly, observations 2–4 focus on the results in the lower panel when the panel weights are not used to construct KW pseudo-weights. *Secondly*, consistent with our expectations, the propensity models that contain confounders and selection variables, that is, Model.x13, produce larger estimated

Table 4. Analysis results for estimation of the prevalence of diagnosed asthmafor adults from RANDS 3 under various propensity models and RANDS 3 weights.

Propensity model[1]	CV[2](KW)	relBias[3] (%)	se[4] ($\times 10^2$)	MSE[5] ($\times 10^4$)
Original panel-weighted	0.91	25.31	0.98	12.56
Unweighted	0	21.89	0.72	9.19
	panel weights			
Model.all	1.13	17.55	1.21	7.04
Model.x13	1.10	13.35	1.04	4.31
Model.x12.n	1.07	11.41	0.93	3.23
Model.x1.n	1.07	12.38	0.95	3.67
Model.x12.r	1.08	12.85	0.97	3.94
Model.x1.r	1.08	12.85	0.97	3.93
	no weights			
Model.all	0.83	14.02	1.07	4.70
Model.x13	0.80	13.51	0.97	4.24
Model.x12.n	0.70	11.38	0.81	3.01
Model.x1.n	0.69	13.67	0.82	4.06
Model.x12.r	0.73	11.37	0.84	3.04
Model.x1.r	0.71	13.44	0.84	3.98

[1]Original panel-weighted denotes the RANDS 3 estimate using the original panel weights without PS adjustment; model.all: the full propensity model with all main and pairwise interaction terms; Model.x12.n: propensity model including terms of the confounders and outcome predictors selected using the National Health Interview Survey (NHIS); Model.x12.r: propensity model including terms of the confounders and outcome predictors selected using the Research and Development Survey (RANDS). Panel weights (no weights) indicates that the RANDS 3 original panel weights were used as the base weight for the PS adjustment.

[2]$CV = \frac{sd(KW)}{mean(KW)}$, for standard deviation *sd*

[3]relBias (%) $= \frac{\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}}{\hat{\mu}_{NHIS}} \times 100\%$

[4]se = standard error of estimated mean

[5]MSE $= \left(\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}\right)^2 + se^2\left(\hat{\mu}_{RANDS}\right)$

variances compared with Model.x12 irrespective of the outcome predictors being selected from RANDS or the NHIS (e.g., se = 0.97 vs. 0.81-0.84). *Thirdly*, comparing the estimates under the propensity models containing the confounders and outcome predictors that are selected using NHIS data (i.e., Model.x12.n) vs. the RANDS data (i.e., Model.x12.r), similar relative bias, se and MSE are observed (relBias = 11.38 versus 11.37; se = 0.81 versus 0.84, MSE = 3.01 versus 3.04). *Lastly*, the relative biases under Model.x1 are somewhat larger than that under Model.x12 (relBias = 13.67 versus 11.38 or 13.44 versus 11.37). This result could be due to the existing correlation between the confounders and the outcome predictors. Adjusting for confounders only may not be sufficient for maximum bias reduction.

In brief, for the evaluation of diagnosed asthma using the RANDS data, we would recommend the pseudo-weights constructed under Model.x12.n with the confounders and predictors selected from the reference survey (e.g., NHIS). In situations where outcome variables are not collected in the reference survey but available only in the target sample (e.g., RANDS), Model.x12.r can be the alternative model to construct the KW pseudo-weights, assuming conditional noninformative sampling holds for the target sample.

5. Discussion

Identifying and collecting the best information on more timely target sample and on higher quality reference surveys can increase the ability of NSOs to produce timely estimates with lower bias from target samples. This article examined how different types of variables that are included in a propensity model impact the performance of PS-based pseudo-weighted estimators for population mean estimation from a target sample. Means and variances of estimated population means under various mis-specified propensity models, including different types of variables with and without interactive effects, were evaluated analytically and numerically. Different levels of variable correlations in the finite population were also considered to reflect real data scenarios. We have the following findings: (1) confounders, the variables related to both the selection indicator and the outcome of interest, are important variables to include in the propensity model; (2) pseudo-weights that balance the distributions in the outcome predictor x_2 or the selection variable x_3, in addition to the confounder x_1, denoted by $w(x_1, x_2)$ or $w(x_1, x_3)$, should be constructed for the target sample units so that the corresponding pseudo-weighted target sample mean is approximately unbiased; (3) compared to $w(x_1, x_3)$, the pseudo-weights $w(x_1, x_2)$ gain more efficiency in estimating population means. In contrast, the inclusion of selection variables, compared to the outcome predictors, in the propensity model tended to inflate the estimated variances. Intuitively, the outcome predictor is related to the outcome variable; including outcome predictors in the propensity score model distinguishes differences between the outcome in the reference and target samples, which results in weights related to outcome and therefore yields estimates with smaller variance estimates. Finally, findings are applied to real target data from RANDS, a survey that uses commercial probability panels, which has potential selection bias. Under the model with confounders and outcome predictors (Model.x12) or model with confounders and selection variables (Model.x13), the RANDS estimate of U.S. asthma prevalence had the greatest bias reduction (relative bias ranging from $11.37\% - 13.51\%$ compared to the NHIS) when the panel weights are not used to construct KW pseudo-weights, compared to the original panel-weighted RANDS estimates (relative bias of 25.31%).

Results from this articlehave several important applications in practice for NSOs that collect data from both target surveys and high-quality reference surveys. First, this study provides a principled approach to select covariates for the PS model. Rather than including all variables or selecting certain demographic variables, covariates are assessed based on their variable type (confounder, outcome predictor, selection variable) to be included in the PS model for population mean estimation. Second, guidance on how to design the questionnaire for a target survey with specific research questions (e.g., SARS-CoV2 seropositivity web survey by Kalish et al. 2021) is provided to survey practitioners. The attributes that are most effective in reducing bias/variances of estimates can be collected and used to reduce potential selection bias for the purpose of timely data collection and minimum response burden. Third, the findings from this study can be used for future development of a high-quality probability survey, including the planning of covariates to collect through paradata or the survey questionnaire with minimized measurement/reporting error, to be used as a high-quality reference survey by various nonprobability or web-based probability surveys with selection bias.

The proposed variable inclusion strategies have limitations that can be of interest for future research. First, the strategy is developed for single-outcome studies with research questions related to one outcome of interest, for example, SARS-CoV2 seropositivity study (Kalish et al. 2021). The target sample was collected in a web survey with questions related to COVID-19 infection only. For studies with multiple key outcome variables, it would be of interest to study how the correlation among outcome variables, the overlap for each variable type across outcomes, and their interplays affect population mean estimation by different variable inclusion strategies. Second, in our simulation, we demonstrated the use of a PS matching method (KW) and a PS weighting method numerically. It showed that KW produced approximately unbiased estimates when the analytic propensity model is slightly mis-specified (without the interaction term) while the PS weighting methods require the true propensity model to obtain unbiased estimates. In practice, the underlying selection mechanism of the target sample is often complicated, involving higher orders of nonlinearity and/or nonadditivity. For complicated propensity models, only including main or interactive effects of blocking variables in the logistic model by KW methods may not be sufficient. Nonparametric modeling such as machine learning methods may be promising (Kern et al. 2021).

Third, the focus of this articlewas one valuating the bias and variance reduction of Horvitz-Thompson estimators of FP mean by the types of covariates in the propensity model and thus we did not study how the pseudo-weights, when combined with different estimators, affect the FP mean estimation. Alternative analysis methods, such as doubly robust estimators (Chen et al. 2018) or augmented estimation equations in the missing data imputation context (Robins et al. 1994) can be employed, after identifying the appropriate type of variables to include in the propensity score model. Fourth, selection bias in target samples, compared to more rigorous probability samples, can be induced by low response, different question wording/ordering, topic salience for different types of questions (for example health and health conditions can have large selection bias as shown in Table 2). It would be interesting to study how the proposed variable inclusion strategy can be adapted to reduce the selection bias in target samples with different response rates, question order/wording and salience effects. Lastly, in our data example the backward selection is employed for identifying the type of variables. It would be interesting to employ and compare alternative variable selection methods such as AIC or BIC (Lumley and Scott 2015) that incorporate complex sample designs for the model selection.

6. Appendix: Real Data Analysis Covariate Types

Covariate types (X_1, confounder; X_2, predictor; X_3, selection indicator) reported for each model covariate used in the real data analysis (Section 4). Covariate interactions are denoted by $*$. Eleven predictors (age group; sex; race/ethnicity; marital status; education level; smoking status; diagnosed with high cholesterol; diagnosed with COPD, emphysema, or chronic bronchitis; diagnosed with diabetes; diagnosed with hypertension; and employment status) and their pairwise interactions were included in the initial propensity score models for all adjustments. The covariate types are reported by model set up including the inclusion of RANDS panel weights (panel weights) and exclusion of RANDS panel weights (no weights). Model.n indicates that the outcome predictors were identified using the NHIS; Model.r indicates that the outcome predictors were identified using RANDS.

Table 5. Description of Notation

		Covariate type			
		Panel weights		No weights	
	Variable	Model.n	Model.r	Model.n	Model.r
1	Age group (years)	X_1	X_1	X_1	X_1
2	Sex	X_1	X_1	X_1	X_1
3	Race/ethnicity	X_1	X_1	X_1	X_1
4	Marital status	X_1	X_1	X_1	X_1
5	Education level	X_1	X_1	X_1	X_1
6	Smoking status	X_1	X_1	X_1	X_1
7	Diagnosed with high cholesterol	X_1	X_1	X_1	X_1
8	Diagnosed with COPD, emphysema, or chronic bronchitis	X_1	X_1	X_1	X_1
9	Diagnosed with diabetes	X_1	X_1	X_1	X_1
10	Diagnosed with hypertension	X_1	X_1	X_1	X_1
11	Employment status	X_1	X_1	X_1	X_1
12	Age group (years) * Sex				
13	Age group (years) * Race/ethnicity	X_1	X_1	X_1	X_1
14	Age group (years) * Marital status	X_3	X_1	X_3	X_1
15	Age group (years) * Education level	X_3	X_3	X_3	X_3
16	Age group (years) * Smoking status	X_2	X_2	X_2	
17	Age group (years) * Diagnosed with high cholesterol				
18	Age group (years) * Diagnosed with COPD, emphysema, or chronic bronchitis	X_3	X_3	X_3	
19	Age group (years) * Diagnosed with diabetes	X_3	X_1	X_3	
20	Age group (years) * Diagnosed with hypertension			X_3	X_3
21	Age group (years) * Employment status			X_3	X_3
22	Sex * Race/ethnicity				
23	Sex * Marital status	X_2	X_2	X_2	X_2
24	Sex * Education level	X_1	X_3	X_1	X_3
25	Sex * Smoking status	X_2		X_1	X_3
26	Sex * Diagnosed with high cholesterol	X_2	X_2	X_2	X_2
27	Sex * Diagnosed with COPD, emphysema, or chronic bronchitis		X_2		
28	Sex * Diagnosed with diabetes				X_2
29	Sex * Diagnosed with hypertension	X_3	X_3		
30	Sex * Employment status			X_3	X_3
31	Race/ethnicity * Marital status	X_2	X_2	X_1	X_1
32	Race/ethnicity * Education level	X_3	X_3	X_3	X_3
33	Race/ethnicity * Smoking status	X_3	X_3	X_3	X_3
34	Race/ethnicity * Diagnosed with high cholesterol				

Table 5. Continued

		Covariate type			
		Panel weights		No weights	
	Variable	Model.n	Model.r	Model.n	Model.r
35	Race/ethnicity * Diagnosed with COPD, emphysema, or chronic bronchitis				
36	Race/ethnicity * Diagnosed with diabetes				
37	Race/ethnicity * Diagnosed with hypertension	X_2		X_2	
38	Race/ethnicity * Employment status				
39	Marital status * Education level		X_2		X_2
40	Marital status * Smoking status		X_2		X_2
41	Marital status * Diagnosed with high cholesterol		X_2		X_2
42	Marital status * Diagnosed with COPD, emphysema, or chronic bronchitis		X_2		X_2
43	Marital status * Diagnosed with diabetes	X_2	X_2	X_2	X_2
44	Marital status * Diagnosed with hypertension		X_2		X_2
45	Marital status * Employment status		X_2		X_2
46	Education level * Smoking status				X_2
47	Education level * Diagnosed with high cholesterol	X_3	X_3	X_3	X_3
48	Education level * Diagnosed with COPD, emphysema, or chronic bronchitis				
49	Education level * Diagnosed with diabetes				
50	Education level * Diagnosed with hypertension				
51	Education level * Employment status				
52	Smoking status * Diagnosed with high cholesterol				
53	Smoking status * Diagnosed with COPD, emphysema, or chronic bronchitis	X_2		X_2	
54	Smoking status * Diagnosed with diabetes		X_2		X_2
55	Smoking status * Diagnosed with hypertension				
56	Smoking status * Employment status				

Table 5. Continued

		Covariate type			
		Panel weights		No weights	
	Variable	Model.n	Model.r	Model.n	Model.r
57	Diagnosed with high cholesterol * Diagnosed with COPD, emphysema, or chronic bronchitis	X_2		X_2	
58	Diagnosed with high cholesterol * Diagnosed with diabetes			X_2	
59	Diagnosed with high cholesterol * Diagnosed with hypertension	X_3	X_3	X_3	X_3
60	Diagnosed with high cholesterol * Employment status			X_3	X_3
61	Diagnosed with COPD, emphysema, or chronic bronchitis * Diagnosed with diabetes	X_2			X_2
62	Diagnosed with COPD, emphysema, or chronic bronchitis * Diagnosed with hypertension				
63	Diagnosed with COPD, emphysema, or chronic bronchitis * Employment status				
64	Diagnosed with diabetes * Diagnosed with hypertension				
65	Diagnosed with diabetes * Employment status				
66	Diagnosed with hypertension * Employment status				

7. References

Ali, M.S., R.H.H. Groenwold, S. Belister, W.R. Pestman, AW. Hoes, K.C.B. Roes, A. de Boer, and O.H. Klungel. 2015. "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review." *Journal of Clinical Epidemiology* 68: 122–131. DOI: https://doi.org/10.1016/j.jclinepi.2014.08.011.

Austin, P. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Statistics in Medicine* 27: 2037–2049. DOI: https://doi.org/10.1002/sim.3150.

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 on the APPOR task force on non-probability sampling." *Journal of Survey Statistics and Methodology* 1(2): 90–143. DOI: https://doi.org/10.1093/jssam/smt008.

Brookhart, M.A., S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, and T. Stürmer. 2006. "Variable selection for propensity score models." *American Journal of Epidemiology* 163(12): 1149–1156. DOI: https://doi.org/10.1093/aje/kwj149.

Brick, J.M. and R. Tourangeau. 2017. "Responsive Survey Designs for Reducing." *Journal of Official Statistics* 33(3):735–752. DOI: https://doi.org/10.1515/jos-2017-0034.

Chen, Y., P. Li, and C. Wu. 2019. "Doubly Robust Inference with Nonprobability Survey Samples." *Journal of the American Statistical Association*: 1(11): 2011–2021. DOI: https://doi.org/10.1080/01621459.2019.1677241.

Chen, J.K.T., R. Valliant, and M.R. Elliott. 2018. "Model-assisted calibration of non-probability sample survey data using adaptive LASSO." *Survey Methodology*, 44(1): 117–144. Available at: http://link.gale.com/apps/doc/A663832408/AONE?u = anon ~ 9db3f809&sid = googleScholar&xid = 411f8f5e (accessed March 2022).

Craig B.M, R.D. Hays, A.S. Pickard, D. Cella, D.A. Revicki, and B.B. Reeve. 2013. "Comparison of US panel vendors for online surveys." *Journal of the Medical Internet* 15 (11): e260. DOI: https://doi.org/10.2196/jmir.2903.

Grose, E., S. Wilson, J. Barkun, K. Bertens, G. Martel, F. Balaa, and J. Abou Khalil. 2020. "Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature." *Journal of the American College of Surgeons,* 230(1): 101–112.e2. DOI: https://doi.org/10.1016/j.jamcollsurg.2019.10.003.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves R.M. and E. Peytcheva. 2008. "The impact of nonresponse rates on nonresponse bias." *Public Opinion Quarterly* 72 (2):167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Hernán, M.A. 2004. "A definition of causal effect for epidemiological research." *Journal of Epidemiology & Community Health,* 58(265): 271. DOI: https://doi.org/10.1136/-jech. 2002.006361.

Iannacchione, V.G., J.G. Milne, and R.E. Folsom. 1991. "Response probability weight adjustments using logistic regression." *Proceedings of the section on survey research methods, American Statistical Association* :637–642. Available at: http://www.asasrms.org/Proceedings/papers/1991_109.pdf (accessed March 2022).

Irimata, K.E., Y. He, B. Cai, H.C. Shin, V.L. Parsons, and J.D. Parker. 2020. "Comparison of Quarterly and Yearly Calibration Data for Propensity Score Adjusted Web Survey Estimates." *Survey Methods: Insights from the Field, Special issue "Advancements in Online and Mobile Survey Methods."* DOI: https://doi.org/10.13094/SMIF-2020-00018.

Kalish, H., C. Klumpp-Thomas, S. Hunsberger, H.A. Baus, M.P. Fay, N. Siriponget et al. 2021. "Mapping a Pandemic: SARS-CoV-2 Seropositivity in the United States." Available at: https://www.medrxiv.org/content/10.1101/2021.01.27.21250570v1 (accessed June 2022).

Kern, C., Y. Li, and L. Wang. 2021. "Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples." *Journal of Survey Statistics and Methodology* 9(5): 1088–1113. DOI: https://doi.org/10.1093/jssam/s-maa028.

Kish, L. 1992. "Weighting for unequal Pi." *Journal of Official Statistics* 8(2): 183–200. https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/weighting-for-unequal-empemsubemiemsub.pdf (accessed March 2022).

Korn, E. and B.I. Graubard 1999. *Analysis of Health Surveys*. John Wiley & Sons.

Lessler, J.T. and W.D. Kalsbeek. 1992. *Nonsampling Errors in Surveys*. New York: John Wiley.

Leyrat, C., A. Caille, A. Donner, and B. Giraudeau. 2013. "Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study." *Statistics in Medicine* 32(19): 3357–3372. DOI: https://doi.org/10.1002/sim.5795.

Li, Y., B. Graubard, and R. DiGaetano. 2011. "Weighting methods for population-based case-control study." *Journal of Royal Statistical Society C* 60: 165–185. DOI: https://doi.org/10.1111/j.1467-9876.2010.00731.x.

Little, R.J. and S. Vartivarian. 2003. "On weighting the rates in non-response weights." *Statistics in Medicine* 22(9):1589–1599. DOI: https://doi.org/10.1002/sim.1513.

Little, R.J. and S. Vartivarian. 2005. "Does weighting for nonresponse increase the variance of survey means?" *Survey Methodology* 31(2): 161–168. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf?st = VelcJijp (accessed March 2022).

Lumley, T. and A. Scott. 2015. "AIC and BIC for modeling with complex survey data." Journal of Survey Statistics and Methodology 3 (1): 1–18. DOI: https://doi.org/10.1093/jssam/smu021.

Lumley, T. 2020. Survey: *analysis of complex* surveysamples.R packageversion 4.4.0.

NIH-AARP (National Institutes of Health and AARP Diet and Health Study). 2006. "Data Dictionary. August 2006." Available at: http://dietandhealth.cancer.gov/docs/DataDictionary_Aug2006.pdf (accessed June 2022).

Mercer, A., C. Kennedy, and H. Klein. 2018. *For weighting online opt-in samples, what matters most*. Pew Research Center. Available at: https://www.pewresearch.org/-methods/wp-content/uploads/sites/10/2018/01/Weighting-Online-Opt-In-Samples.pdf (accessed March 2022).

National Center for Health Statistics. 2020. *RANDS 3 Technical Documentation*. Hyattsville, Maryland. Available at: https://www.cdc.gov/nchs/rands/files/RAND-S3_technical_documentation.pdf (accessed June 2022).

Parker, J., K. Miller, Y. He, P. Scanlon, B. Cai, H.-C. Shin, V. Parsons, and K. Irimata. 2020. "Overview and Initial Results of the National Center for Health Statistics' Research and Development Survey." *Statistical Journal of the International Association for Official Statistics* 36(4): 1199–1211. DOI: https://doi.org/10.3233/SJI-200678.

Parker, J.D., M. Talih, D.J. Malec, V. Beresovsky, M. Carroll, and J.F. Gonzalez. 2017. "National Center for Health Statistics Data Presentation Standards for Proportions. National Center for Health Statistics." *Vital Health Stat*, 2(175). Available at: https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf (accessed June 2022)

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Pinsky, P.F., A. Miller, B.S. Kramer, T. Church, D. Reding, P. Prorok, E. Gelmann, R.E. Schoen, S. Buys, R.B. Hayes, and C.D. Berg. 2007. "Evidence of a healthy volunteer

effect in the prostate, lung, colorectal, and ovarian cancer screening trial." American Journal of Epidemiology165(8): 874–881. DOI: https://doi.org/10.1093/aje/kwk075.

Robins, J., A. Rotnitzky, and L. Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American Statistical Association* 89(427): 846–886. https://doi.org/10.1080/01621459.1994.10476818.

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

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Valliant, R. 2020. "Comparing alternatives for estimation from nonprobability samples." *Journal of Survey Statistics and Methodology* 8(2):231–263. DOI: https://doi.org/10.1093/jssam/smz003.

Wang, L., B.I. Graubard, A.K. Hormuzd, and Y. Li. 2020a. "Improving External Validity of Epidemiologic Cohort Analyses: a Kernel Weighting Approach." *Journal of the Royal Statistical Society* 183(3): 1293–1311. DOI: https://doi.org/10.1111/rssa.12564.

Wang, L., B.I. Graubard, A.K. Hormuzd, and Y. Li. 2020b. "Efficient and Robust Propensity-Score-Based Methods for Population Inference using Epidemiologic Cohorts." *International Statistical Review* 90(1):146–164. DOI: https://doi.org/10.1111/insr.12470.

Wang, L., R. Valliant, and Y. Li. 2021. "Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts." *Statistics in Medicine* 40(24): 5237–5250. DOI: https://doi.org/10.1002/sim.9122.

Weitzen, S., K.L. Lapane, A.Y. Toledano, A.L. Hume, and V. Mor. 2004. "Principles for modeling propensity scores in medical research: a systematic literature review." *Phamacoepidemiologyand Drug Safety* 13(12): 841–853. DOI: https://doi.org/10.1002/PDS.969.

Yang, S., J.K. Kim, and R. Song. 2020. "Doubly robust inference when combining probability and non-probability samples with high dimensional data." *Journal of the Royal Statistical Society* 82: 445–465. DOI: https://doi.org/10.1111/rssb.12354.

Received July 2021
Revised January 2022
Accepted April 2022

sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 901–928, http://dx.doi.org/10.2478/JOS-2022-0039

Pseudo Bayesian Mixed Models under Informative Sampling

Terrance D. Savitsky[1] *and Matthew R. Williams*[2]

When random effects are correlated with survey sample design variables, the usual approach of employing individual survey weights (constructed to be inversely proportional to the unit survey inclusion probabilities) to form a pseudo-likelihood no longer produces asymptotically unbiased inference. We construct a weight-exponentiated formulation for the random effects distribution that achieves approximately unbiased inference for generating hyperparameters of the random effects. We contrast our approach with frequentist methods that rely on numerical integration to reveal that the pseudo Bayesian method achieves both unbiased estimation with respect to the sampling design distribution and consistency with respect to the population generating distribution. Our simulations and real data example for a survey of business establishments demonstrate the utility of our approach across different modeling formulations and sampling designs. This work serves as a capstone for recent developmental efforts that combine traditional survey estimation approaches with the Bayesian modeling paradigm and provides a bridge across the two rich but disparate sub-fields.

Key words: Labor force dynamics; Markov chain Monte Carlo; pseudo-posterior distribution; survey sampling; weighted likelihood.

1. Introduction

Hierarchical Bayesian models provide a flexible and powerful framework for social science and economic data, which often include nested units of analysis such as industry, geography, and individual. Yet, social science and economic data are commonly acquired from a survey sampling procedure. It is typically the case that the underlying survey sampling design distribution governing the procedure induces a correlation between the response variable(s) of interest and the survey sampling inclusion probabilities assigned to units in the finite population from which the sample was taken. Survey sampling designs where there is a correlation between the response variable and the sampling inclusion probabilities are referred to as informative and will result in the distribution of the response variable in observed samples being different from that from the underlying population about which we seek to perform model-based inference. Sample designs may also be informative when the inclusion probabilities for groups are correlated with the corresponding latent random effects. The resulting distribution of random effects in the sample is different from that of the finite population of random effects.

The current literature for Bayesian methods has partially addressed population model estimation of survey data under informative sampling designs through the use of survey

[1] U.S. Bureau of Labor Statistics, Office of Survey Methods Research, 1669 Gales St NE, Washington D.C., 20002, U.S.A. Email: Savitsky.Terrance@bls.gov
[2] RTI International, 3040 East Cornwallis Road, Research Triangle Park, North Carolina, 27709-2194, U.S.A. Email: matthew.dunn.williams@gmail.com

© Statistics Sweden

sampling weights to obtain consistent estimates of fixed effects or top level global parameters. Yet the survey statistics literature (Pfeffermann et al. 1998; Rabe-Hesketh and Skrondal 2006) has demonstrated that parameters related to random effects, or local parameters are still potentially estimated with bias. The possibility for survey-induced bias in estimation of random effects severely limits the applicability of the full suite of Bayesian models to complex social and economic data.

This article proposes a Bayesian survey sample-weighted, plug-in framework for the simultaneous estimation of fixed effects and generating hyperparameters (e.g., variance) of random effects that is unbiased with the respect to the distribution over samples and asymptotically consistent with respect to the population distribution which generated the finite population.

1.1. Informative Sampling Designs

Survey sampling designs that induce a correlation between the response variable of interest, on the one hand, and the survey sample inclusion probabilities, on the other hand, are deemed "informative" and produce samples that express a different balance of information than that of the underlying population, thus estimation methods that do not incorporate sample design information lead to incorrect inferences. For example, the U.S. Bureau of Labor Statistics (BLS) administers the Job Openings and Labor Turnover Survey (JOLTS) to business establishments for the purpose of estimating labor force dynamics, such as the total number of hires, separations and job openings for area-indexed domains. The units are business establishments and their inclusion probabilities are set to be proportional to their total employment (as obtained on a lagged basis from a census instrument). Since the number of hires, separations and job openings for establishments are expected to be correlated to the number of employees, this survey sampling design induces informativeness, so that hiring, separations and openings would be expected to be larger in the samples than in the underlying population.

1.2. Bayesian Models for Survey Data

There is growing and rich literature on employment of survey sampling weights (constructed to be inversely proportional to unit inclusion probabilities) for correction of the population model estimated on the observed survey sample to produce asymptotically unbiased estimation. Some recent articles focus on the use of Bayesian modeling for the specific purpose of producing mean and total statistics under either empirical or nonparametric likelihoods, but these methods don't allow the data analyst to specify their own population models for the purpose of parameter estimation and prediction (Kunihama et al. 2016; Rao and Wu 2010; Si et al. 2015). In particular, the set-up for our article is one where the data analyst has specified a particular Bayesian hierarchical model for the population (from which the sample was taken) under which they wish to perform inference from data taken from a complex sampling design. So, having to specify a model that is specific to the realized sample, but unrelated to the population model constructed by the data analyst does not allow them to conduct the inference they seek. It may be possible to first generate a large synthetic population and then perform the intended inference (Dong et al. 2014). Savitsky and Toth (2016) and Williams and Savitsky (2020) complement

these efforts by employing a pseudo-posterior to allow the data analyst to estimate a population model of their choice on an informative sample taken from that population. The pseudo-likelihood exponentiates each unit likelihood contribution by its sampling weight, which re-balances information in the observed sample to approximate that in the population. Maximizing the pseudo-likelihood can lead to consistent estimation of global population generating parameters (Isaki and Fuller 1982). Combining with a proper prior distribution, we achieve a pseudo-posterior. The use of the pseudo-posterior may be situated in the more general class of approximate or composite likelihoods used for Bayesian inference (Ribatet et al. 2012). All of the above Bayesian approaches that allow analyst specification of the underlying population generating model to be estimated on the observed informative sample *only* address models with fixed or global effects, not random effects. Yet, it is routine in Bayesian modeling to employ one or more sets of random effects under prior formulations designed to capture complex covariance structures. Hierarchical specifications make such population models readily estimable. This property may also be exploited to achieve maximum likelihood estimates with techniques such as data cloning (Lele et al. 2010).

1.3. Extending the Pseudo-Posterior to Mixed Effects Models

There are two survey designs considered in this article: 1. Clusters or groups of units are sampled in a first stage, followed by the sampling of nested units in a second stage. Nested units sampled together in the first stage each share the same group level random effect from the population. We refer to this procedure as the "direct" sampling of clusters or groups; 2. Units are sampled in a single stage without directly sampling the clusters or groups in which they naturally nest (e.g., geography). We refer to the case where groups used in the population model are not included in the sampling design as "indirect" sampling of groups, since a group is included in the sample to the extent that a nested unit is directly sampled.

This article extends the approaches of Savitsky and Toth (2016) and Williams and Savitsky (2020) from global-only parameters to mixed effects (global and local parameter) models by exponentiating both the data likelihood contributions *and* the group-indexed random effects prior distributions by sampling weights – an approach that we label, "double-weighting" – that is multiplied, in turn, by the joint prior distribution for the population model parameters to form a joint pseudo-posterior distribution with respect to the observed data and random effects for the sampled groups.

Our augmented (by sample-weighting the prior for random effects) pseudo-posterior method introduced in the next section is motivated by a data analyst who specifies a population generating (super population) model that includes group-indexed random effects for which they desire to perform inference. The observed data are generated under an informative sampling design such that simply estimating the population model on the observed sample will produce biased parameter estimates. Our augmented pseudo-posterior model estimator uses survey sampling weights to perform a relatively minor adjustment to the model augmented likelihood such that parameter draws taken on the observed informative sample approximates inference with respect to the population generating distribution.

We demonstrate that our pseudo-posterior formulation achieves both unbiasedness with respect to the two-stage survey sampling design distribution *and* consistency with respect to the population generating (super population) model for the observed response variable and the latent cluster/group random effects under both direct and indirect sampling of groups. Our weighted pseudo-posterior is relatively simple in construction because we jointly sample the random effects and model parameters in our hierarchical modelling scheme and marginalize out the random effects *after* estimation. For simple models analytic integration may also be possible (Savitsky and Williams 2020). However jointly estimating the random effects also allows for subsequent inference on group level statistics either using the posterior mean or another summary measure that better preserves the population distribution of random effects. This is beyond the scope of this work, but see Shen and Louis (1998) for more details.

The case of indirect sampling of groups is particularly important in Bayesian estimation as it is common to specify multiple random effects terms that parameterize a complex covariance structure because the random effects terms are readily estimable in a hierarchical modelling scheme.

The remainder of the article proceeds to develop our proposed double-weighting methods for estimation of mixed effects models under both direct and indirect sampling of groups on data acquired under informative sampling in Section 2. Simulation studies are conducted in Section 3 that compare our proposed method to the usual case of likelihood weighting under direct sampling of groups. Section 4 applies our doubleweighting method to estimation of the number of hires for business establishments under employment of industry-indexed random effects in the population model where we reveal that our double-weighting approach produces more heterogeneous random effects estimates to better approximate the population from the observed sample than does the usual practice. We offer a concluding discussion in Section 5.

2. Mixed Effects Pseudo Posterior Distribution

The focus of this article addresses sampling units naturally collected into a population of groups; for example, defined by geography. There is typically a dependence among the response values for units within each group such that units are more similar within than between groups. Sampling designs are typically constructed as multi-stage where the collection of groups in the population are first sampled, followed by the sequential taking of a sub-sample of units from the population of units within *selected* or sampled groups.

By contrast, an alternative set of sampling designs may proceed to draw a sample from the population of units in a *single* stage such that the groups are included in the sample, indirectly, when one or more of their member units are selected. These two sampling designs – sampling groups, followed by sampling units within groups in a multi-stage sampling design, on the one hand, as compared to a single-stage sampling of units without directly sampling their member groups, on the other hand – will lead us to design two formulations for extending the pseudo-posterior distribution of Williams and Savitsky (2020).

The pseudo-posterior exponentiates the likelihood contributions of the *single level* fixed effects model (that does not utilize random effects) by the survey weights for the observed

sample units $i \in \{1, \ldots, n\}$, which are inversely proportional to their probabilities of being selected into the sample $w_i \propto 1/\pi_i$.

$$f^{\pi}(\theta|\boldsymbol{y}, \tilde{\boldsymbol{w}}) \propto \left[\prod_{i=1}^{n} f(y_i|\theta)^{\tilde{w}_i}\right] f(\theta) \tag{1}$$

where the normalized weights $\tilde{w}_i = w_i / \frac{\sum w_i}{n}$ to the sample size n. The sum of the weights directly affects the amount of posterior uncertainty estimated in the posterior distribution for θ, so normalizing it to sum to the sample size regulates that uncertainty. Equation (1) is a noisy approximation to the true (but not fully known) joint distribution of the population model $P_{\theta_0}(\boldsymbol{y})$ and the sampling process $P^{\pi}(\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ denotes a vector of sample design inclusion indicators for units and groups (that are governed by P^{π}) and formally defined under the 2-stage class of sampling designs considered in this article in the upcoming section.

The noisy approximation for the population likelihood obtained by constructing the sample-weighted pseudo-posterior estimator for the observed (informative) sample leads to consistent estimation of population generating (super-population) parameters θ (as the sample size, n, grows) for essentially all survey sampling designs used in practice, including multistage designs (Williams and Savitsky 2020). A symptotically correct uncertainty quantification can be achieved through post-processing adjustments (Williams and Savitsky 2021).

The pseudo-posterior construction requires only a minor change to the population model specified by the data analyst on which they wish to perform inference (by weighting each unit-indexed likelihood contribution by its associated marginal sampling weight). In particular, the data analyst may specify population distributions for $f(y_i|\theta)$ and priors $f(\theta)$; for example, if the data are count data that we work with in the sequel, they may specify a Poisson likelihood with mean, μ, for which they define a latent regression model formulation. The data analyst is interested to perform inference for the generating parameters under the population generation and not the distribution of the observed sample. Under informative sampling the two distributions are different and the pseudo-posterior corrects the distribution of the observed sample back to the population of interest.

We demonstrate in the sequel, that the formulation in Equation (1) can be extended to multi-level models by exponentiating *both* the likelihood (conditioned on the random effects) *and* the prior distribution for random effects by sampling weights.

2.1. Mixed Effects Posterior Under Direct Sampling of Population Groups

Assign units, $i \in (1, \ldots, N)$, that index a population, U, to groups, $h \in (1, \ldots, G_U)$, where each population group, h, nests $U_h = 1, \ldots, N_h$ units, such that $N = |U| = \sum_{h=1}^{G_U} N_h$, with $N_h = |U_h|$. Construct a 2-stage informative sampling design whose first stage takes a direct sample of the G_U groups, where $\pi_h \in (0, 1]$ denotes the marginal sample inclusion probability for group, $h \in (1, \ldots, G_U)$. Let $g \in (1, \ldots, G_S)$, index the *sampled* groups, where G_S denotes the number of observed groups from the population of groups, $G_U \supset G_S$.

Our first result defines a pseudo-posterior estimator on the observed sample for our population model that includes group-indexed random effects in the case where we

directly sample groups, followed by the sampling of units nested within selected groups, in a multistage survey sampling design. Our goal is to achieve unbiased inference for (θ,ϕ) (where θ denotes fixed effects for generating population responses, y, and ϕ denotes the generating parameters of random effects, u, for the population), estimated on our observed sample taken under an informative survey sampling design. We assume that random effects u are indexed by group and are independent conditional on the generating parameter ϕ. Multistage designs that sample groups or clusters, followed by the further sampling of nested units, are commonly used for convenience to mitigate costs of administration where in-person interviews are required and also in the case where a sampling frame of end-stage units is constructed after sampling groups in the first stage. The second stage of the survey sampling design takes a sample from the N_g (second stage) units $\forall g \in S_g$, where $S_g \subset U_g$. The second stage units are sampled with conditional inclusion probabilities, $\pi_{\ell|g} \in (0,1]$ for $\ell = 1, \ldots, N_g$, conditioned on inclusion of group, $g \in (1, \ldots, G_S)$. Let $j \in (1, \ldots, n_g)$ index the sampled or observed second stage units linked to or nested within sampled group, $g \in (1, \ldots, G_S)$. Denote the marginal group survey sampling weight, $w_g \propto 1/\pi_g$ for $\pi_g \in (0,1]$. Denote the marginal unit survey sampling weight, $w_{gj} \propto 1/\pi_{gj}$, for $\pi_{gj} \in (0,1]$, the joint inclusion probability for unit, j, nested in group, g, both selected into the sample. The group marginal inclusion probabilities and conditional unit inclusion probabilities under our 2-stage survey sampling design are governed by distribution, P^{π}.

Theorem 1. Under a proper prior specification, $f(\theta)f(\phi)$, the following pseudoposterior estimator achieves approximately unbiased inference for super-population (population generating) model, $f(\theta,\phi|y)$, with respect to the distribution governing the taking of samples from an underlying finite population, P^{π},

$$f^{\pi}(\theta,\phi|\mathbf{y}) \propto \left[\int_{\mathbf{u}\in\mathscr{U}} \left\{ \prod_{g\in S} \left(\prod_{j\in S_g} f(y_{gj}|u_g,\theta)^{w_{gj}} \right) f(u_g|\phi)^{w_g} \right\} d\mathbf{u} \right] f(\theta)f(\phi). \tag{2}$$

where $f^{\pi}(\cdot)$ denotes a sampling-weighted pseudo-distribution, $j \in S_g$ denotes the subset of units, $j \in (1, \ldots, n_g = |S_g|)$, linked to group, $g \in (1, \ldots, G_S)$. Parameters, (θ,ϕ), index the super-population model posterior distribution, $f(\theta,\phi|y)$, that is the target for estimation. The integral for the vector $\mathbf{u} = (u_1, \ldots, u_{n_g})$ is taken over its support, $\mathscr{U}$, for each component, $u_g \in \mathbf{u}$.

We employ a pseudo-likelihood for the first level of the model for sampled observations y_{gj} within sampled clusters g by exponentiating by the sample weight w_{gj}. This provides a noisy approximation to the first stage likelihood. For the second level model (or prior) for the random effects u_g, we exponentiate this distribution by its corresponding sampling weights w_g. This provides a noisy approximation to the population distribution of random effects. Both approximations are needed because the distributions of both the responses and the random effects in the sample can differ substantially from those in the corresponding population due to the informative sampling design at both the cluster g and the within cluster $j|g$ stages, where the latter notation denotes the sampling of unit j conditioned on / within sampled group g.

Under our augmented (by weighting the prior of the group-indexed random effects) pseudo-likelihood of Equation (15), $f(y_{gj}|u_g,\theta)$ and $f(u_g|\phi)$ are not restricted; for example, we select a Poisson distribution for the observed data likelihood, $f(y_{gj}|u_g,\theta)$, for our simulation study and application in the sequel. Similarly, the form of the distribution for the random effects prior distribution is not restricted under our construction, though it is most commonly defined as Gaussian under a GLM specification. Replacing the single Gaussian with a mixture of Gaussian distributions would also fit our set-up. Our approach also readily incorporates additional levels of random effects with no conceptual changes.

Proof. We first construct the complete joint model for the finite population, U, as if the random effects, (u_h), were directly observed,

$$f_U(\theta, \phi|\mathbf{y}, \mathbf{u}) \propto \left[\prod_{h=1}^{G_U}\left(\prod_{\ell=1}^{N_h} f(y_{h\ell}|u_h, \theta)\right) f(u_h|\phi)\right] f(\theta) f(\phi). \tag{3}$$

Under a complex sampling design, we specify random sample inclusion indicators for groups, δ_h, with marginal probabilities $\pi_h = P(\delta_h = 1)$ for $h \in (1, \ldots, G_U)$, governed by P^π. We further specify random sample inclusion indicator, $\delta_{\ell|h} = (\delta_\ell|\delta_h = 1) \in \{0, 1\}$, with probability $\pi_{\ell|h} = P(\delta_{\ell|h} = 1)$, for unit $\ell \in (1, \ldots, N_h)$, conditioned on the inclusion of group, h, such that the indicator for the joint sampling of unit ℓ nested within group h is denoted as $\delta_{h\ell} = \delta_{\ell|h} \times \delta_h$, with the associated marginal inclusion probability, $\pi_{h\ell} = P(\delta_{h\ell} = 1)$.

The taking of an observed sample is governed by the survey sampling distribution, P^π (as contrasted with $P_{\theta,\phi}$, the population generation distribution for (y,u)). The pseudo-likelihood with respect to the joint distribution, $(P^\pi, P_{\theta,\phi})$, is then constructed by exponentiating components of the likelihood in the population such that the *expected value* of the survey sample pseudo log-likelihood function with respect to P^π equals that of the log-likelihood for the entire population (and thus the score functions also match in expectation). Let $\ell_U(\mathbf{y}, \mathbf{u}|\theta, \phi) \equiv \log f_U(\theta, \phi|\mathbf{y}, \mathbf{u})$ denote the population model log-likelihood. Applying this approach to the log-likelihood of the joint model, above, leads to the following pseudo-likelihood formulation:

$$\ell_U^\pi(\mathbf{y}, \mathbf{u}|\theta, \phi) \propto \sum_{h=1}^{G_U}\left(\sum_{\ell=1}^{N_h}\left(\frac{\delta_{\ell|h}}{\pi_{\ell|h}}\right)\left(\frac{\delta_h}{\pi_h}\right)\ell(y_{h\ell}|u_h, \theta)\right) + \left(\frac{\delta_h}{\pi_h}\right)\ell(u_h|\phi) \tag{4}$$

$$= \sum_{h=1}^{G_U}\left(\sum_{\ell=1}^{N_h}\left(\frac{\delta_{h\ell}}{\pi_{h\ell}}\right)\ell(y_{h\ell}|u_h, \theta)\right) + \left(\frac{\delta_h}{\pi_h}\right)\ell(u_h|\phi) \tag{5}$$

where P^π governs all possible samples, $(\delta_h, \delta_{\ell|h})_{\ell \in U_h, h=1,\ldots,G_U}$, taken from population, U. Let joint group-unit inclusion indicator, $\delta_{h\ell} = \delta_h \times \delta_{\ell|h}$ with $\pi_{h\ell} = P(\delta_{h\ell} = 1) = P(\delta_h = 1,\ \delta_{\ell|h} = 1)$. For each *observed* sample $\ell_U^\pi(\mathbf{y}, \mathbf{u}|\theta, \phi) = \ell_S^\pi(\mathbf{y}, \mathbf{u}|\theta, \phi)$ where

$$\ell_S^\pi(\mathbf{y}, \mathbf{u}|\theta, \phi) = \sum_{g=1}^{G_S}\left(\sum_{j \in S_g} w_{gj}\ell(y_{gj}|u_g, \theta)\right) + w_g\ell(u_g|\phi) \tag{6}$$

and $w_{gj} \propto \pi_{g,j}^{-1}$ and $w_g \propto \pi_g^{-1}$. The expectation of our estimator in Equation (5) is unbiased with respect to P^π,

$$\mathbb{E}^{\pi}\left[\ell_U^{\pi}(\mathbf{y}, \mathbf{u}|\theta, \phi)|P_{\theta,\phi}\right] \equiv \tag{7}$$

$$\mathbb{E}^{\pi}\left[\ell_U^{\pi}(\mathbf{y}, \mathbf{u}|\theta, \phi)\right] = \ell_U(\mathbf{y}, \mathbf{u}|\theta, \phi), \tag{8}$$

where the expectation, $\mathbb{E}^{\pi}(\cdot)$, is taken with respect to the survey sampling distribution, P^{π}, that governs the survey sampling inclusion indicators, $\{\delta_{h\ell}, \delta_h\}$, conditional on the data $\{\mathbf{y}, \mathbf{u}\}$ generated by $P_{\theta,\phi}$. The final equality in Equation (8) is achieved since $\mathbb{E}^{\pi}(\delta_{h\ell}) = \pi_{h\ell}$ and $\mathbb{E}^{\pi}(\delta_h) = \pi_h$.

Thus, we use the following sampling-weighted model approximation to the complete population model of Equation (3):

$$f^{\pi}(\theta, \phi|\mathbf{y}, \mathbf{u}) \propto \left[\prod_{g\in S}\left(\prod_{j\in S_g} f(y_{gj}|u_g, \theta)^{w_{gj}}\right) f(u_g|\phi)^{w_g}\right] f(\theta) f(\phi). \tag{9}$$

We can then construct a sampling-weighted version of the observed model:

$$f^{\pi}(\theta, \phi|\mathbf{y}) \propto \left[\int_{\boldsymbol{u}\in\mathscr{U}}\left\{\prod_{g\in S}\left(\prod_{j\in S_g} f(y_{gj}|u_g, \theta)^{w_{gj}}\right) f(u_g|\phi)^{w_g}\right\} d\mathbf{u}\right] f(\theta) f(\phi). \tag{10}$$

The walk from Equation (9) to Equation (10) is possible because we co-estimate the $(\mathbf{u})$ with (θ, ϕ) and then perform the integration step to marginalize over the $(\mathbf{u})$ *after* estimation.

Theorem 1 requires the exponentiation of the prior contributions for the sampled random effects, (u_g), by a sampling weight, $w_g \propto 1/\pi_g$ in order to achieve approximately unbiased inference for ϕ; it is not enough to exponentiate each data likelihood contribution, $f(y_{gj}|u_g, \theta)$, by a unit (marginal) sampling weight, w_{gj}. This formulation is generally specified for any population generating model, $P_{\theta,\phi}$. Our result may be readily generalized to survey sampling designs of more than two stages where each collection of later stage groups are nested in earlier stage groups (such as households of units nested within geographic PSUs).

The proposed method under direct sampling of Equation (2) is categorized as a plug-in estimator that exponentiates the likelihood contributions for nested units by the unit-level marginal sampling weights *and*, in turn, exponentiates the prior distribution for cluster-indexed random effects by the cluster (or PSU) marginal sampling weights. Samples from the joint pseudo-posterior distribution over parameters are interpreted as samples from the underlying (latent) population generating model since the augmented (by the weighted random effects prior distribution) pseudo-likelihood estimated on the observed sample provides a noisy approximation for the population generating likelihood.

Although the augmented pseudo-likelihood is unbiased with respect to the distribution over samples, it will not, generally, produce correct uncertainty quantification. In particular, the credibility intervals will be too optimistic relative to valid frequentist confidence intervals because the plug-in method is not fully Bayesian in that it doesn't model uncertainty in the sampling weights (i.e., the dependence of the weights on the responses $\{y_{gj}, u_g\}$ from the population). Although the employment of random effects

captures dependence among nested units, the warping and scaling induced by the sampling weights will result in failure of Bartlett's second identity such that the asymptotic hyperparameter covariance matrix for our plug-in mixed effects model will not be equal to the sandwich form of the asymptotic covariance matrix for the MLE. The result of the lack of equality is that the model credibility intervals, without adjustment, will not contract on valid frequentist confidence intervals.

A recent work of Leon-Novelo and Savitsky (2021) jointly models the unit level marginal sampling weights and the response variable and includes group-indexed random effects parameters in their joint model. They demonstrate correct uncertainty quantification because the asymptotic covariance matrix of their fully Bayesian model (that also co-models the sampling weights) is equal to that for the MLE. Their method specifies an exact likelihood for the observed sample that is complicated and requires a closed-form solution for an integral that restricts the class of models that may be considered. This approach requires a different model formulation than that specified for the population and of interest to the data analyst.

By contrast, our plug-in augmented pseudo-posterior distribution requires only minor change to the underlying population model specified by the data analyst and may be easily adapted to complicated population models.

Correct uncertainty quantification for (θ, ϕ) may be achieved by using the method of Williams and Savitsky (2021) to perform a post-processing of the posterior parameter draws that replaces the pseudo-posterior covariance with the sandwich form of the MLE. This article, by contrast, focuses on providing unbiased point estimation for mixed effects models as an extension of Savitsky and Toth (2016) because Williams and Savitsky (2021) may be readily applied, post sampling.

Our pseudo-likelihood in Equation (5) is jointly conditioned on $(\mathbf{y}, \mathbf{u})$, such that the random weights, $\left(\delta_{h\ell}/\pi_{h\ell}\right)$, are specified in linear summations. This linear combination of weights times log-likelihoods ensures (design) unbiasedness with respect to P^{π} because the weight term is separable from the population likelihood term. We may jointly condition on $(\mathbf{y}, \mathbf{u})$ in our Bayesian set-up because we jointly sample the posteior $(\mathbf{u}, \theta, \phi)$, such that the integration step over $\mathbf{u}$ is applied *after* joint estimation. In other words, we accomplish estimation by sampling $\mathbf{u}$ jointly with (θ, ϕ) on each posterior draw and then ignoring $\mathbf{u}$ to perform marginal inferences on θ and ϕ, which is a common approach with Bayesian hierarchical models. By contrast, Pfeffermann et al. (1998), Rabe-Hesketh and Skrondal (2006), and others (for example see Kim et al. 2017) specify the following integrated likelihood under frequentist estimation for an observed sample where units are nested within groups,

$$\ell^{\pi}(\theta, \phi) = \sum_{g=1}^{G_S} w_g \ell_i^{\pi}(\theta, \phi), \tag{11}$$

for $\ell_i^{\pi}(\theta, \phi) = \log L_i^{\pi}(\theta, \phi)$ for,

$$L_i^{\pi}(\theta, \phi) = \int_{u_g \in \mathscr{U}} \exp\left[\sum_{j \in S_g} w_{j|g} \ell(y_{gj}|u_g, \theta)\right] f(u_g|\phi) du_g, \tag{12}$$

which will *not*, generally, be unbiased with respect to the distribution governing the taking of samples for the population likelihood because the unit level conditional weights, $(w_{j|g})_j$, are nested inside an exponential function (such that replacing $w_{j|g}$ with $\delta_{\ell|h}/\pi_{\ell|h}$ inside the exponential and summing over the population groups and nested units will not produce separable sampling design terms that each integrate to 1 with respect to P^{π}, conditioned on the generated population) (Yi et al. 2016). The non-linear specification in Equation (12) results from an estimation procedure that integrates out $\mathbf{u}$ *before* pseudo-maximum likelihood point estimation of (θ, ϕ).

This design biasedness (with respect to P^{π}) is remedied for pseudo-maximum likelihood estimation by Yi et al. (2016) with their alternative formulation,

$$\ell^{\pi}(\theta, \phi) = \sum_{g=1}^{G_S} w_g \sum_{j<k; j,k \in S_g} w_{j,k|g} \ell_{gj,k}(\theta, \phi) \tag{13}$$

$$\ell_{gj,k}(\theta, \phi) = \log \left\{ \int_{u_g \in U} f(y_{gj}|u_g, \theta) f(y_{gk}|u_g, \theta) f(u_g|\phi) du_g \right\}, \tag{14}$$

where $w_{j,k|g} \propto 1/\pi_{j,k|g}$ denotes the joint inclusion probability for units (j, k), both nested in group, g, conditioned on the inclusion of group, g, in the observed sample. Equation (13) specifies an integration over u_g for each $f(y_{gj}|u_g, \theta) f(y_{gk}|u_g, \theta)$ pair, which allows the design weights to enter in a linear construction outside of each integral. This set-up establishes linearity for inclusion of design weights, resulting in unbiasedness with respect to the distribution governing the taking of samples for computation of the pseudo-maximum likelihood estimate, though under the requirement that pairwise unit sampling weights be published to the data analyst or estimated by them.

Yet, the marginalization of the random effects *before* applying the group weight, w_g, fails to fully correct for the prior distribution for u_g. We show in the sequel that ϕ is estimated with bias by Yi et al. (2016) due to this integration of the random effects being performed on the unweighted prior of u_g. Our method, by contrast, weights the prior for u_g and performs the integration of u_g indirectly by first jointly estimating u_g and θ and then using only the marginal distribution for θ.

Using either Equations (12) and (10) will lead to consistent estimation (collapsing of the estimators around the true population generating values) when both the number and size of clusters are large. But for finite clusters neither will lead to consistent estimation for arbitrary within cluster sampling designs. Slud (2020) proves this for the maximum likelihood estimation of the canonical case of a one-way ANOVA model. Savitsky and Williams (2020) demonstrate constructively that using Equation (10) and the psuedo-posterior can achieve consistent results for the one-way ANOVA model for finite population cluster sizes with an additional restriction on the within cluster balance of the sample.

2.2. *Mixed Effects Posterior Under Indirect Sampling of Population Groups*

Bayesian model specifications commonly employ group-level random effects (often for multiple simultaneous groupings) to parameterize a complex marginal covariance

structure. Those groups are often not *directly* sampled by the survey sampling design. We, next, demonstrate that weighting the prior contributions for the group-indexed random effects is *still* required, even when the groups are not directly sampled, in order to achieve unbiased inference for the generating parameters of the random effects, ϕ. Again, as throughout, we assume the group-indexed random effects are conditionally independent given generating parameter ϕ. We focus our result on a simple, single-stage sampling design, that may be readily generalized, where we reveal that the group-indexed survey sampling weights are constructed from unit marginal inclusion probabilities. Constructing sampled group weights from those of member units appeals to intuition because groups are included in the observed sample only if any member unit is selected under our single-stage survey sampling design.

Suppose the same population set-up as for Theorem 1, with population units, $\ell \in U_h$, linked to groups, $h \in (1, \ldots, G_U)$, where each unit, (h, ℓ), maps to $i \in (1, \ldots, N)$. We now construct a *single* stage sampling design that directly samples each (h, ℓ) unit with marginal inclusion probability, $\pi_{h\ell}$, governed by P^{π}. Group, $g \in G_S$, is *indirectly* sampled based on whether there is *any* linked unit, (gj), observed in the sample.

Theorem 2. The following pseudo-posterior estimator achieves approximately unbiased inference with respect to P^n,

$$f^{\pi}(\theta, \phi|\mathbf{y}) \propto \left[\int_{\mathbf{u} \in \mathscr{U}} \left\{ \prod_{g \in S} \left(\prod_{j \in S_g} f(y_{gj}|u_g, \theta)^{w_{gj}} \right) f(u_g|\phi)^{w_g = \frac{1}{N_g} \sum_{j \in S_g} w_{gj}} \right\} d\mathbf{u} \right] f(\theta) f(\phi), \tag{15}$$

where $w_{gj} \propto 1/\pi_{gj}$.

Proof. We proceed as in Theorem 1 by supposing the population U of units and associated group-indexed random effects, (u_h), were fully observed. We first construct the likelihood for the fully observed population.

$$f_U(\theta, \phi|\mathbf{y}, \mathbf{u}) \propto \left[\prod_{h=1}^{G_U} \left(\prod_{\ell=1}^{N_h} f(y_{h\ell}|u_h, \theta) \right) f(u_h|\phi) \right] f(\theta) f(\phi) \tag{16}$$

$$= \left[\prod_{h=1}^{G_U} \prod_{\ell=1}^{N_h} \left\{ f(y_{h\ell}|u_h, \theta) f(u_h|\phi)^{\frac{1}{N_h}} \right\} \right] f(\theta) f(\phi). \tag{17}$$

We proceed to formulate the pseudo-likelihood for all possible random samples taken from $U, f_U^{\pi}(\cdot)$, governed jointly by $(P^{\pi}, P_{\theta,\phi})$, from which we render the pseudo-likelihood for any sample, $f^{\pi}(\cdot)$, which is constructed to be unbiased with respect to the distribution governing the taking of samples for the population model of Equation (17) under P^{π},

$$f_U^{\pi}(\theta, \phi|\mathbf{y}, \mathbf{u}) \propto \left[\prod_{h=1}^{G_U} \prod_{\ell \in U_h} \left\{ f(y_{h\ell}|u_h, \theta) f(u_h|\phi)^{\frac{1}{N_h}} \right\}^{\frac{\delta_{h\ell}}{\pi_{h\ell}}} \right] f(\theta) f(\phi) \tag{18}$$

$$= \left[\prod_{h=1}^{G_U} f(u_h|\phi)^{\frac{1}{N_h} \sum_{\ell U_h} \frac{\delta_{h\ell}}{\pi_{h\ell}}} \prod_{\ell \in U_h} f(y_{h\ell}|u_\ell, \theta)^{\frac{\delta_{h\ell}}{\pi_{h\ell}}} \right] f(\theta) f(\phi). \tag{19}$$

This pseudo-posterior reduces to the following expression for the observed sample,

$$f^{\pi}(\theta, \phi|\mathbf{y}, \mathbf{u}) \propto \left[\prod_{g=1}^{G_S} f(u_g|\phi)^{\frac{1}{N_g} \sum_{j \in S_g} w_{gj}} \prod_{j \in S_g} f(y_{gj}|u_j, \theta)^{w_{gj}} \right] f(\theta) f(\phi), \tag{20}$$

where $\pi_{gj} = P(\delta_{gj} = 1)$ (under P^{π}), $w_{gj} \propto 1/\pi_{g,j}$ and N_g denotes the number of units in the population linked to observed group, $g \in (1, \ldots, G_S)$ observed in the sample. We set $w_g := 1/N_g \times \sum_{j \in S_g} w_{gj}$ and the result is achieved.

This result derives from eliciting group-indexed weights from unit inclusion probabilities for units linked to the groups. While the resulting pseudo-posterior estimators look very similar across the two theorems, the sampling designs are very different from one another in that groups are not directly sampled in this latter case, which is revealed in their differing formulations for w_g.

The averaging of unit weights formulation for w_g naturally arises under the derivation of Equation (17) when sampling units, rather than groups under a model that utilizes group-indexed random effects to capture within group dependence that naturally arises among units in the population. Exponentiating the augmented pseudo-likelihood of Equation (16) by survey variables anticipates the integration of the random effects to produce an observed data pseudo-likelihood. We may intuit this result by interpreting this form for w_g proportional to the average importance of units nested in group each group, g. It bears mention that in the indirect sampling case, there is no probability of group selection defined for a single stage design.

In practice, it is not common for the data analyst to know the population group sizes, (N_g), for the groups, $g \in (1, \ldots, G_S)$ observed in the sample, so one estimates an $\hat{N}_g$ to replace N_g in Equation (15). Under a single-stage sampling design where the groups are indirectly sampled through inclusions of nested units into the observed sample, we assume that we only have availability of the marginal unit inclusion sampling weights, (w_{gj}). The group population size, N_g, needed for the sum-weights method of Equation (15), may be estimated by $\hat{N}_g = \sum_{j=1}^{n_g} w_{j|g}$. To approximate $w_{j|g}$, we first utilize the sum-probabilities result to estimate, $\hat{w}_g = 1/\hat{\tilde{\pi}}_g$, and proceed to extract $(w_{j|g})$ from $w_{gj} \approx w_g w_{j|g}$. If we invert the resultant group-indexed weight, $w_g = 1/N_g \times \sum_{j \in S_g} w_{gj}$, for the case where groups are not directly sampled, we may view the inverse of the group g weight, $\tilde{\pi}_g = 1/w_g$, as a "pseudo" group inclusion probability, since we don't directly sample groups. One may

envision other formulations for the pseudo group inclusion probabilities, $\tilde{\pi}_g$, that we may, in turn, invert to formulate alternative group-indexed survey sampling weights, (w_g). Please see Appendix (Subsection 6.1) where we develop other methods, in addition to sum-weights, for computing $\tilde{\pi}_g$.

In application, we normalize the by-group, survey sampling weights, $(w_g)_g = 1, \ldots, G_S$, to sum to the number of observed groups in the sample, G_S, and normalize unit weights, $(w_{gj})_{j=1}, \ldots, n_g$ to sum to the overall sample size, n. These normalizations regulate uncertainty quantification for posterior estimation of (u_g) and global parameters, (ϕ, θ) by encoding an effective number of observed groups and units. So, we normalize them to sum to the number of groups and units observed in the sample to regulate the estimated pseudo-posterior variance of (ϕ, θ). (In practice, these normalizations often produce somewhat optimistic credibility intervals due to dependence induced by the survey sampling design. Williams and Savitsky (2021) provide an algorithm that adjusts pseudo-posterior draws to incorporate this dependence). For the pseudo-likelihood of Equation (12), many have investigated different scaling factors for weights (Asparouhov 2006). Some of these approach may improve efficiency for the pseudo-posterior approach.

We refer to our proposed procedure for weight exponentiating both the data likelihood contributions and the prior distributions of the (u_g) as "double-weighting", as mentioned in the introduction, to be contrasted with the usual approach of "single-weighting" of Williams and Savitsky (2020) developed for models with global effects parameters.

3. Simulation Study

Our simulation study in the sequel focuses on a count data response rather than the usual continuous response, both because count data are the most common data type for the employment data collected by BLS and because our Bayesian construction is readily estimable under any response data type.

We generate a count data response variable, y, for a population of size, $N = 5,000$ units, where the logarithm of the generating mean parameter, μ is constructed to depend on a size predictor, x_2, in both fixed and random effects terms; in this way, we construct both fixed and random effects to be informative, since our proportion-to-size survey sampling design sets unit inclusion probabilities to be proportional to x_2. We generate a population of responses using,

$$\begin{aligned} y_i &\sim P(\mu_i) \\ \log\, \mu_i &= \alpha_0 + x_{1i}\alpha_1 + x_{2i}\alpha_2 + [1, x_{2i}]\boldsymbol{\gamma}_{h\{i\}}, \end{aligned} \tag{21}$$

where $\mathscr{P}(\cdot)$ denotes the Poisson distribution, $x_{1i} \sim \mathcal{N}(0,1)$ is the inferential predictor of interest to the data analyst and $x_{2i} \sim \mathscr{E}(1/2.5)$ (where $\mathscr{E}(\cdot)$ denotes the Exponential distribution) is the size variable, which is generated from a skewed distribution to reflect real-world survey data, particularly for employment counts. The expression, $h\{i\}$, denotes the group $h \in (1, \ldots, G_U)$ linked to unit $i \in (1, \ldots, N)$. We generate $\underset{2\times 2}{\boldsymbol{\gamma}_h^{2\times 1}} \sim \mathcal{N}_2(0, \mathrm{diag}(\boldsymbol{\sigma}) \times R \times \mathrm{diag}(\boldsymbol{\sigma}))$, where $\boldsymbol{\sigma} = (1.0, 0.5)'$. We set $R = \mathbb{I}_2$, where $\mathbb{I}_2$ denotes the identity matrix of size 2. Finally, we set $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)' = (0, 1.0, 0.5)$, where we

choose the coefficient of x_2 to be lower than that for x_1 to be moderately informative, which is conservative.

The allocation of units, $i = 1, \ldots, N$ to groups, $h = 1, \ldots, G_U$ is performed by sorting the units, i, based on size variable, x_2. This allocation procedure constructs sized-based groups that accord well with survey designs that define groups as geographic regions, for convenience, where there is expected more homogeneity within groups than between groups.

The population size for each group, N_h, is fixed under direct sampling of groups; for example, $N_h = 4$ in the case of $G_U = 1,250$, which produces $N = 5,000$ units, so the number of population units per group is constructed as $(4, 10, 25, 50, 100)$ for population group sizes, $G_U = (1,250, 500, 200, 100, 50)$, respectively.

Although the population response y is generated with $\mu = f(x_1, x_2)$, we estimate the marginal model $\mu = f(x_1)$ for the population *generating* distribution (for large population sizes there is little difference between these distribution parameters and the finite population parameters). We will compare estimated results on samples taken from the population to assess bias and mean-squared error (MSE). We use x_2 in the generation of the population values for y because the survey sampling inclusion probabilities are set proportionally to x_2, which instantiates the informativeness of the sampling design. In practice, however, the analyst does not have access to x_2 for the population units or, more generally, to all the information used to construct the survey sampling distribution that sets inclusion probabilities for all units in the population. The marginal estimation model under exclusion of size variable, x_2, is specified as

$$\begin{aligned} y_i &\sim \mathscr{P}(\mu_i) \\ \log\, \mu_i &= \beta_0 + x_{1i}\beta_1 + u_{h\{i\}} \\ u_h &\sim \mathscr{N}(0, \sigma_u^2) \end{aligned} \tag{22}$$

where now u_h is an intercept random effect, $h = 1, \ldots, \mathrm{G}_U$.

Our goal is to estimate the global parameters, $(\beta_0, \beta_1, \sigma_u^2)$, from informative samples of size, $n = 500$, taken from the population (of size, $N = 5,000$). We utilize the following simulation algorithm:

1. Each Monte Carlo iteration of our simulator (that we run for $B = 300$ iterations) generates the population $(y_i, x_{1i}, x_{2i})_{i=1}^N$ from Equation (21) on which we estimate the marginal population model of Equation (22) to determine the population true values for $(\mu_{\mathrm{i}}, \sigma_u^2)$.
2. Our simulation study focuses on the direct sampling of groups, followed by a sub-sampling of units within the selected groups. We use a proportion-to-size design to directly sample from the G_U groups in the first stage, where the group inclusion probabilities, $\pi_h \propto \frac{1}{N_h}\sum_{i \in U_h} x_{2i}$.
   We draw a sample of groups in the first stage and observe $G_S < G_U$ groups. In particular, fixed sample of total size $n = 500$ is taken where the number of groups sampled, $G_S = n/(Nf) \times \mathrm{G}_U$.
3. The second stage size-based sampling of units is accomplished with inclusion probabilities, $\pi_{\ell|g} \propto x_{2\ell}$ for $\ell \in (1, \ldots, N_g)$.

We perform a further sub-sampling of $f\%$ of population units in the selected G_S groups.

4. Estimation is performed for $(\beta_0, \beta_1, \sigma_u^2)$ from the observed sample of $n = 500$ under three alternatives:
   (a) Single-weighting, where we solely exponentiate the likelihood contributions for (y_{gj}) by sampling weights, $(w_{gj} \propto 1/\pi_{gj})$ (and don't weight the prior for the random effects, (u_g));
   (b) Double-weighting, where we exponentiate *both* the likelihood for the (y_{gj}) by sampling weights, (w_{gj}), and also exponentiate the prior distribution for u_g by weight, $w_g \propto 1/\pi_g$ (for each of $g = 1, \ldots, G_S$). We compute the marginal unit weights used in both single- and double-weighting as $w_{gj} \propto 1/\pi_{gj}$, where π_{gj} is the marginal inclusion probability, formulated as, $\pi_{gj} = \pi_g \pi_{j\,g}$ for $j = 1, \ldots, n_g$ for each group, $g \in 1, \ldots, G_S$ in the case of direct sampling of groups.
   (c) SRS, where in the case of direct sampling of groups, we take a simple random (equal probability) sample of groups in a first stage, followed by a simple random sample of units within selected groups. We take the SRS sample from the same population as is used to take the two-stage, probability proportional to size (PPS) informative sample. The inclusion of model estimation under (a non-informative) SRS is intended to serve as a gold standard against which we may judge the bias and MSE performance of single- and double-weighting under informative sampling.

We use Stan (Carpenter et al. 2017) to estimate the double-weighted mixed effects model of Equation (9) for the specific case of the Poisson likelihood that we use in our simulations and application that next follows. We fully specify our Stan probability model for the Poisson likelihood under double-weighting in the online supplement. In particular, we specify a multivariate Gaussian joint prior distribution for the $K \times 1, \beta$ coefficients with a vector of standard deviation parameters, σ_β drawn from a truncated Cauchy prior. The associated correlation matrix for the multivariate Gaussian prior for β is drawn from a prior distribution that is uniform over the space of $K \times K$ correlation matrices. The prior for the standard deviation parameter of the random effects, σ_u, is also specified as a truncated Cauchy distribution. These prior distributions are designed as weakly informative by placing large probability mass over a large support region, while expressing a mode to promote regularity and a stable posterior geometry that is readily explored under Stan's Hamiltonian Monte Carlo scheme. The single-weighting case is achieved as a special / simplified case of the double-weighting model.

Please see Appendix (Subsection 6.2) for results of a second simulation study under the indirect sampling of groups.

3.1. Informative Random Effects Under Direct Sampling of Groups

To make concrete the notion of informative random effects, we generate a single population and subsequently take a single, informative sample of groups from that population of groups under a proportion-to-size design, using the procedures for population generation and the direct sampling of groups, described above. The size for

each population is $N = 5,000$ and the sample size is $n = 500$. We, next, average the item responses, y in each group after centering by removing the fixed effects observed in the sample (excluding x_2). For illustration, the computed $\overline{((y - \exp(x\beta))_g)}$ will be used as a naive indicator of the distribution of the random effects, $\exp(u_g)$. Each plot panel in Figure 1 compares the distributions of this group-indexed centered mean statistic between the generated population and resulting informative sample for a population. A collection of plot panels for a sequence of populations with $G_U = (1,250, 500, 100)$ number of population groups is presented, from left-to-right. Fixing a plot panel, each violin distribution plot includes horizontal lines for the $(0.25, 0.5, 0.75)$ quantiles. We see that under a proportion-to-size design that the distributions for the centered, group mean statistic in the sample are different from the underlying populations and skew larger than those for the populations. This upward skewness in each sample indicates that performing population estimation on the observed sample will induce bias for random effects variance, ϕ, without correction of the group indexed random effects distribution in the sample, which we accomplish by weighting the distribution over random effects back to that for the population.

3.2. *Varying Number of Population Groups, G_U*

We assess bias for a population model constructed using group-indexed random effects, where each group links to multiple units. Our results presented in Figure 2 compare our double-weighting method to single-weighting in the case we conduct a proportion-to-size direct sampling of groups and, subsequently, sub-sample $f = 50\%$ of member units within groups. We include an SRS sample of groups and units within selected groups taken from the same population. The results reveal that bias is most pronounced in the case of a relatively larger number of groups e.g., $G_U = (1,250, 500)$ for $N = 5,000$ where each group links relatively few units. By contrast, as the number of groups decreases, fixing the

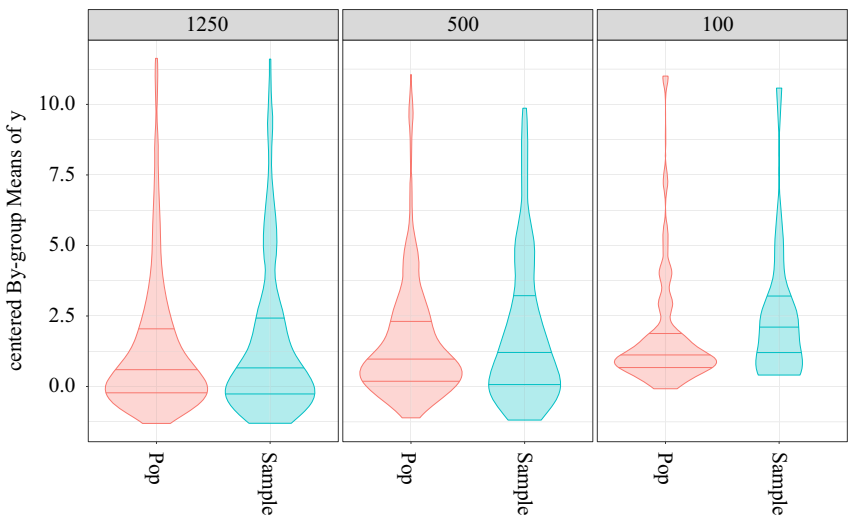


Fig. 1. Each plot panel compares distributions of $\overline{(y - exp(x\beta))_g}$ for each of a synthetic population and a single sample from that population, faceted by a sequence for the number of population groups, G_U for fixed population size $N = 5,000$. The resulting violin plots present each distribution within 95% quantiles.

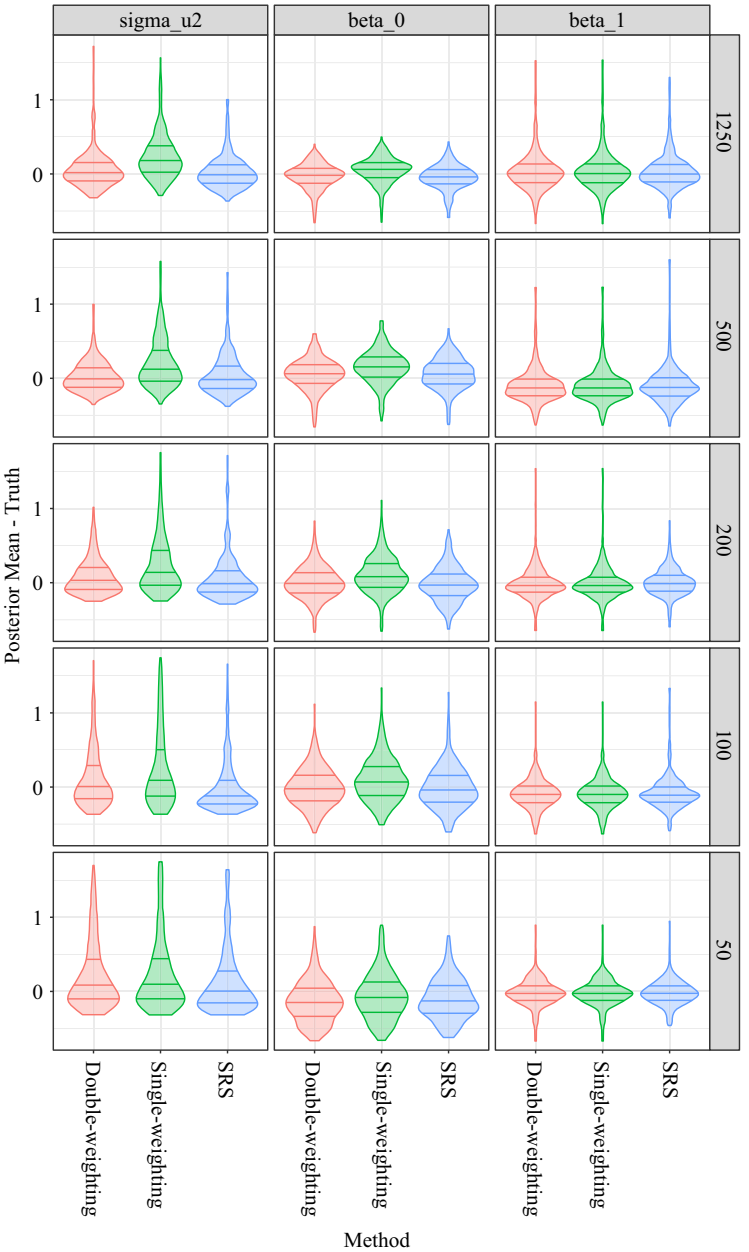


Fig. 2. Direct sampling of groups: Each panel displays the Monte Carlo distributions and quantiles (0.5%, 25%, 50%, 75%, 99.5%) of differences between Posterior Means and truth under Single- and Double-weighting schema as compared to SRS for varying number of random effect groups, G_U, under $x_2 \sim \mathcal{E}(1/(2.5))$ for $N = 5{,}000$ and $n = 500$ across $B = 300$ generated data sets. Parameter estimates $(\beta_0, \beta_1, \sigma_u^2)$ are displayed along columns and number of population groups, G_U, is varied in descending order along the rows.

population size, N, the number of units linking to each group increases, which will have the effect of reducing variation in resulting sampling weights among the groups until, in the limit, there is a single group (with $\pi_g = 1$). The relative bias of single-weighting,

therefore, declines as the number of groups declines (and units per group increases), such that residual bias in σ_u^2 for $G_U = 100$ is dominated by increasing variability (because we sample fewer groups) for all methods. We, nevertheless, detect a small decrease in bias when we use double-weighting. We include Table 1 that presents the bias in the estimation for the posterior mean values of (β_0, β_1, σ_u^2) that confirms the reduction in bias for σ_u^2 under double-weighting for $G_U = 100$. Our set-up may be viewed as more likely to induce bias because we assign units to groups by sorting units on the values of the size variable, $x_2 \sim \mathscr{E}(1/(2.5))$ for allocation to groups. Our proportion-to-size sampling design selects groups based on the mean size variable for each group, $\bar{x}_2$. This set-up will tend to accentuate the variance in the resulting group-indexed size variable (and, hence, the resulting survey sample inclusion probabilities) as compared to a random allocation of units to groups. Our simulation set-up is, nevertheless, realistic because many surveys are characterized by relatively homogeneous clusters; for example, the geographically-indexed metropolitan statistical areas (MSAs) (which may be viewed as clusters) used by the Current Employment Statistics survey (administered by the BLS) tends to express larger (higher number of employees) establishments in more highly populated MSAs.

We next compare our double-weighting approach to the best available method in the literature, the pairwise composite likelihood method of Yi et al. (2016), specified in Equation (13), which we refer to as "pair-integrated". We compare both methods in the case of relatively few units linked to each group (e.g., $G = 500, 1,250$) because Yi et al. (2016) demonstrate superior bias removal properties as compared to Rabe-Hesketh and Skrondal (2006) in this setup. We exclude smaller values of G because as the number of individuals within each group grows, the number of pairwise terms to include in the pair-integrated method grows quadratically. Our simulation set-up conducts a first-stage proportion to size sampling of groups in exactly the same manner as the previous simulation study. We additionally include an SRS of groups and, in turn, units within

Table 1. Estimated bias of posterior mean from single- and double-weighting and SRS across varying number of groups, G_U $x_2 \sim \mathscr{E}(1/2.5)$. Population of $N = 5,000$ and sample of $n = 500$ across $B = 300$ generated data sets.

β_0	β_1	σ_u^2	Statistic	Method	G_U
0.03	0.03	0.25	bias	Single-weighting	1250
-0.06	0.03	0.06	bias	Double-weighting	1250
-0.05	0.01	0.02	bias	SRS	1250
0.15	-0.13	0.19	bias	Single-weighting	500
0.04	-0.13	0.02	bias	Double-weighting	500
0.04	-0.12	0.04	bias	SRS	500
0.10	-0.02	0.26	bias	Single-weighting	200
-0.01	-0.02	0.09	bias	Double-weighting	200
-0.02	0.00	0.05	bias	SRS	200
0.09	-0.09	0.31	bias	Single-weighting	100
-0.01	-0.10	0.11	bias	Double-weighting	100
-0.03	-0.10	0.07	bias	SRS	100
-0.06	-0.03	0.44	bias	Single-weighting	50
-0.14	-0.03	0.29	bias	Double-weighting	50
-0.13	-0.02	0.27	bias	SRS	50

groups, as a benchmark. The custom R code to implement the "pair-integrated" point estimation can be found in the online supplement.

Figure 3 presents the Monte Carlo distributions for parameter estimates, where the columns denote parameters, $(\beta_0, \beta_1, \sigma_u^2)$, and the rows denote number of population groups, G_U. The results demonstrate that double-weighting leads to unbiased estimation of both the fixed effects parameters and the random effects variance relative to using a two-stage SRS sample. By contrast, the pair-integrated method demonstrates both bias and variability for the random effects variance, of σ_u^2, which is exactly the set-up where it is hoped to perform relatively well. This bias for pair-integrated in the random effects variance also induces bias for the fixed effects intercept, β_0. As mentioned in Section 2.1 the pair-integrated method integrates out the random effects (from the unweighted prior distribution) *before* applying the group weights, which fails to fully correct for the informative sampling of groups. Our method, by contrast, weights the prior for the random effects and integrates them out after estimation. It bears mention that Yi et al. (2016) only evaluate informative sampling of units within groups, but not the informative sampling of the groups themselves, which may be why the estimation bias for σ_u was not discovered.

We briefly comment on the simulation study for the indirect sampling of groups detailed in Appendix (Subsection 6.2). The results accord with the direct sampling of groups where double-weighting outperforms single-weighting. When the number of population groups, G_U, is small, however, noise induced by sampling error results in double-weighting underperforming compared to SRS. Yet, as the number of units per group increases with

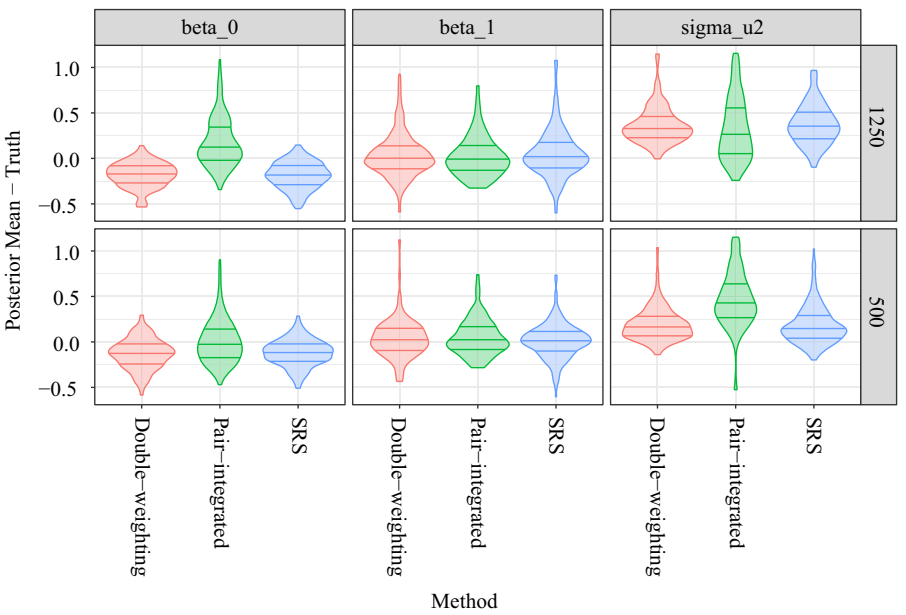


Fig. 3. Direct sampling of groups: Each panel displays the Monte Carlo distributions and quantiles (0.5%, 25%, 50%, 75%, 99.5%) of differences between Posterior Means and truth under Double-weighting, Pair-integrated estimation as compared to Simple Random Sampling (SRS) for varying number of random effect groups, G_U, under $x_2 \sim \mathcal{E}(1/2.5)$ for $N = 5{,}000$ and $n = 500$ across $B = 300$ generated data sets. Parameter estimates $(\beta_0, \beta_1, \sigma_u^2)$ are displayed along columns and number of population groups, G_U, is varied in descending order along the rows.

$G_U = 500$, the sum-weights approach outperforms SRS, which is expected because the PPS design is generally more efficient such that the contraction rate of the estimator on the truth will be faster for PPS (occur at a lower sample size).

Lastly, we note that while we have focused on a simple Poisson random effects formulation, our survey-weighted pseudo Bayesian posterior method readily extends to any number of levels and simultaneous employment of multiple sets of random effects without any modification to the approach. Competitor methods, by contrast, are not readily estimable. While simple models such as in Savitsky and Williams (2020) may allow for analytic integration to obtain the observed likelihood in Equation (10), the augmented approach applies to these broader classes of models.

4. Application

We compare single- and double-weighting under a linear mixed effects model estimated on a dataset published by the Job Openings and Labor Turnover Survey (JOLTS), which is administered by BLS on a monthly basis to a randomly-selected sample from a frame composed of non-agricultural U.S. private (business) and public establishments. JOLTS focuses on the demand side of U.S. labor force dynamics and measures job hires, separations (e.g., quits, layoffs and discharges) and openings. We construct a univariate count data population estimation model with our response, y, defined to be the number of hires. We formulate the associated log mean with,

$$\log\, \mu_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + u_{g\{i\}}, \tag{23}$$

where groups, $g = 1, \ldots, (G = 892)$, denote industry groupings (defined as 6- digit North American Industry Classification (NAICS) codes) that collect the participating business establishments. We expect a within-industry dependence among the hiring levels for business establishments since there are common, industry-driven economic factors that impact member establishments. We construct the fixed effects predictors, $\mathbf{x} = [1,$ ownership status, region], which are categorical predictors where ownership status holds four levels; 1. Private; 2. Federal government; 3. State government; 4. Local government. The region predictor holds four levels, 1. Northeast; 2. South; 3. Midwest; 4 West. Private and Northeast are designated as the reference levels.

The JOLTS sampling design assigns inclusion probabilities (under sampling without replacement) to establishments to be proportional to the number of employees for each establishment (as obtained from the Quarterly Census of Employment and Wages (QCEW)). This design is informative in that the number of employees for an establishment will generally be correlated with the number of hires, separations and openings. We perform our modeling analysis on a May, 2012 data set of $n = 9,743$ responding establishments. We *a priori* expect the random effects, (u_g), to be informative since larger-sized establishments would be expected to express larger variances in their hiring levels. We choose the sum-weights method for inducing industry-level weights (from Equation 15) to construct our double-weighted estimation model on the observed sample.

The more diffuse distribution over the $G = 892$ posterior mean values for random effects, (u_g), under double-weighting than single-weighting shown in Figure 4 demonstrates that co-weighting the likelihood and random effects distribution produces notably different

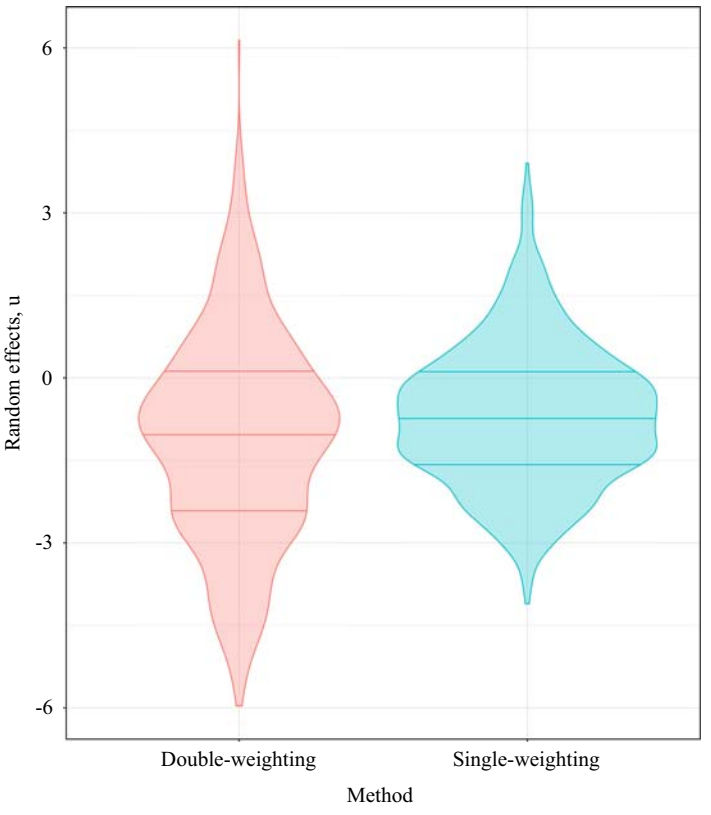


Fig. 4. Distribution and quantiles (25%, 50%, 75%) of estimated posterior mean values for random effects, $u_g, g = 1, \ldots, (G = 892)$, for the JOLTS application under single- and double-weighting.

inference for the group-indexed random effects; in particular, the observed sample is more homogeneous in the number of hires by setting inclusion probabilities to concentrate or over-sample large-sized establishments relative to the more size-diverse population of establishments. So the weighting of the random effects distributions in the observed sample produces a distribution over the posterior mean values for the random effects that better reflects the size-diversity of establishments in the underlying population from which the sample was taken. Figure 5 presents the estimated pseudo-posterior distributions for the generating random effects variance, σ_u^2 and also a single random effect parameter, u_i, under both single- and double-weighting. This figure reinforces the observed result for the random effects where the observed hiring levels in the survey data are more homogenous than those in the underlying population, which induces a larger posterior variation in the estimated random effects parameters for double-weighting.

5. Discussion

In this work, we demonstrate the existence of biased estimation of both fixed *and* random effects parameters when performing inference on data generated from a complex survey sample. This risk is largely unrecognized in the Bayesian literature. The current remedies come from the survey literature and are motivated from a frequentist perspective. They

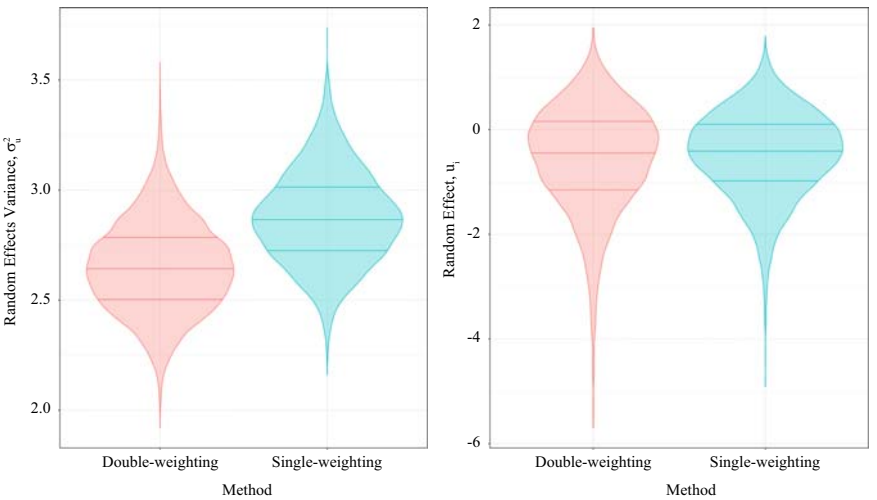


Fig. 5. Distributions and quantiles (25%, 50%, 75%) of posterior samples for σ_u^2, the generating variance for random effects, and a single random effect parameter, u_i, for the JOLTS application, under single- and double-weighting.

provide an incomplete and somewhat ad-hoc approach to the solution. We present a principled development of the "double-weighting" approach based on the joint distribution of the population generating model of inferential interest and the complex sampling design represented by sample inclusion indicators. We exploit the latent variable formulation of mixed models and their related posterior sampling techniques to avoid awkward numerical integration required for frequentist solutions. We show that this simplicity also leads to reductions in bias.

This work culminates recent developmental work in combining traditional survey estimation approaches with the Bayesian modeling paradigm. The pseudo-posterior framework simultaneously offers complex survey data analysis to Bayesian modelers and the full suite of hierarchical Bayesian methods to those well-versed in traditional fixed effect analysis of survey data.

6. Appendix

6.1. Alternative Pseudo Group Inclusion Probabilities Under Indirect Sampling

If we invert the resultant group-indexed weight, $w_g = 1/N_g \times \sum_{j \in S_g} w_{gj}$, from Equation (15), where groups are not directly sampled, we may view the inverse of the group g weight, $\tilde{\pi}_g = 1/w_g$, as a "pseudo" group inclusion probability, since we don't directly sample groups. The construction for one form of $\tilde{\pi}_g$ motivates our consideration of other formulations for the pseudo group inclusion probabilities that we may, in turn, invert to formulate alternative group-indexed survey sampling weights, (w_g).

The resulting w_g of Equation (15) requires either knowledge of N_g or a method for its approximation. The sum of nested unit weights is further composed as a harmonic sum of inverse inclusion probabilities of member units in each group, which may be overly

dominated by units with small unit inclusion probabilities. Our first alternative more directly constructs a pseudo group inclusion probability as the union of probabilities for inclusion of any member unit in the observed sample (in which case the group will be represented in the sample) and does not require estimation of population quantities, such as N_g. Under a weak assumption of nearly independent sampling within groups, this alternative is constructed as,

$$\tilde{\pi}_g = \sum_{\ell=1}^{N_g} \pi_\ell \tag{24}$$

$$\hat{\tilde{\pi}}_g = \sum_{\ell=1}^{N_g} \frac{\delta_\ell}{\pi_\ell} \times \pi_\ell \tag{25}$$

$$= \sum_{j=1}^{n_g} w_j \times \pi_j \tag{26}$$

where π_ℓ denotes the marginal inclusion probability for unit, $\ell \in (1, \ldots, N_g)$, where we recall that N_g denotes the number of units linked to group, $g \in (1, \ldots, G_U)$, in the population of groups. We may estimate the pseudo group inclusion probabilities in the observed sample by making the same walk from population-to-observed-sample as is done in Equation (5) to Equation (6); by including unit sampling weights, $(w_j)_{j \in S_g}$ ($S_g = \{1, \ldots, n_g\}$). We normalize the $(w_j)_{j \in S_g}$ to sum to 1 as our intent is to re-balance the information (among sampled units) within a group to approximate that of the population of units within the group. While this estimator has the undesirable property of computing $\tilde{\pi}_g > 1$, we utilize this quantity to weight the random effects prior density contributions with, $w_g \propto 1/\tilde{\pi}_g$, so we focus on the effectiveness of estimation bias removal for generating hyperparameters of the $(u_g)_{g \in G_U}$. We label this method as the "sum-probabilities" method in contrast to the "sum-weights" methods with which we label the result of Equation (15).

Our second alternative for estimation of a pseudo group inclusion probability is designed to ensure $\tilde{\pi}_g \leq 1$ by using a product complement approach that computes the union of member unit probabilities for a group, indirectly, by first computing its complement and subtracting that from 1. To construct this estimator, we assume that units, $j \in s$ are sampled independently *with* replacement, which is a tenable assumption when drawing a small sample from a large population of units. Let $\pi_j^{(1)}$ denote the probability of selecting unit, j, in a sample of size 1 (e.g., a single draw). Then we may construct the marginal inclusion probability of unit, π_j, for a sample of size, $n = |S|$, as the complement that unit j does not appear in any of the n draws,

$$\pi_j = 1 - \left(1 - \pi_j^{(1)}\right)^n, \tag{27}$$

where $\sum_{j \in U} \pi_j^{(1)} = 1$. By extension, $0 < \tilde{\pi}_g^{(1)} = \sum_{j \in U_g} \pi_j^{(1)} \leq 1$, where $\tilde{\pi}_g^{(1)}$ denotes the pseudo group, $g \in (1, \ldots, G_U)$ inclusion probability for a sample of size 1 and is composed as the union of size 1 probabilities for member units. The expression for the pseudo group inclusion probability derives from the underlying sampling of members with replacement,

$$\tilde{\pi}_g = 1 - (1 - \tilde{\pi}_g^{(1)})^n = 1 - \left(1 - \sum_{j=1}^{N_g} \pi_j^{(1)}\right)^n, \tag{28}$$

where we exponentiate the complement term, $\left(1 - \tilde{\pi}_g^{(1)}\right)$, by the number of draws of units, n (rather than G_S, the number of groups represented in the observed sample) because we don't directly sample groups. We solve for $\pi_j^{(1)}$ using Equation (27), $\pi_j^{(1)} = 1 - (1 - \pi_j)^{(1/n)}$, and plug into Equation (28) to achieve,

$$\tilde{\pi}_g = 1 - \left(1 - \sum_{j=1}^{N_g} \left(1 - (1 - \pi_j)^{(1/n)}\right)\right)^n \tag{29}$$

$$\hat{\tilde{\pi}}_g = 1 - \left(1 - \sum_{j=1}^{N_g} \frac{\delta_j}{\pi_j} \left(1 - (1 - \pi_j)^{(1/n)}\right)\right)^n \tag{30}$$

$$= 1 - \left(1 - \sum_{\ell=1}^{n_g} w_\ell \left(1 - (1 - \pi_\ell)^{(1/n)}\right)\right)^n, \tag{31}$$

where, as with the sum-probabilities formulation, we normalize the unit weights within each group, $(w_\ell)_{\ell \in S_g}$, to sum to 1. We label this method as "product-complement".

6.2. *Simulation Study Results for Alternative Pseudo Group Inclusion Probabilities*

We present the results for the simulation study that samples units, rather than groups, for the expanded set of methods developed in Appendix (Subsection 6.1) for the pseudo group inclusion probabilities. We recall that under this single stage sampling of units, groups are not directly sampled under the survey sampling and are included to the extent that one or more member units are sampled.

The synthetic population (for each Monte Carlo iteration) utilizes group-indexed random effects under size-based assignment of population units to groups under each alternative for total number of groups, G_U. In this study, we randomly vary the number of population units assigned to each group with the mean values for each G_U set to be equal to the fixed number of units per group. We allocate a relatively higher number of units to those groups with smaller-sized units under each group size, G_U, to mimic our application. The number of population units per group, N_h, is set to randomly vary among the G_U population groups using a log-normal distribution centered on the (4,10,25,50,100) units per group used in the case of direct sampling, with a variance of 0.5. In the case of $G_U = 1{,}250$, this produces a right-skewed distribution of the number of units in each group, ranging from approximately 1 to 40 units per group and the total number of population units per group is restricted to sum to $N = 5{,}000$.

We sort the groups such that groups with larger-sized units are assigned relatively fewer units and groups with smaller-sized units are assigned relatively more units. This set-up of assigning more units to smaller-sized groups mimics the estimation of employment counts among business establishments analyzed in our application in the sequel, where there are relatively few establishments with a large number of employees (e.g., >50) (which is the

size variable), while there are, by contrast, many more establishments (small businesses) that have a small number of employees (e.g., < 10).

The survey sampling design employed here is a *single-stage*, proportion-to-size design that directly samples the units (not the groups) with unit inclusion probabilities proportional to the size variable, $x_2 \sim \mathscr{E}(1/(m_2 = 3.5))$. Each Monte Carlo iteration of our simulator (that we run for $B = 300$ iterations) generates the population $(y_i, x_{1i}, x_{2i})_{i=1}^N$, assigns group and unit inclusion probabilities for the population in the case of direct sampling of groups or assigns unit inclusion probabilities in the case of indirect sampling. A sample of $n = 500$ is then taken and estimation is performed for $(\beta_0, \beta_1, \sigma_u^2)$ from the observed sample under three alternatives:

1. Single-weighting, where we solely exponentiate the likelihood contributions for (y_{gj}) by sampling weights, $(w_{gj} \propto 1/\pi_{gj})$ (and don't weight the prior for the random effects, (u_g));
2. Double-weighting, where we exponentiate *both* the likelihood for the (y_{gj}) by sampling weights, (w_{gj}), and also exponentiate the prior distribution for u_g by weight, $w_g \propto 1/\pi_g$ (for each of $g = 1, \ldots, G_S$). We estimate π_g using each of the three methods presented in Subsection 6.1: "sum-weights", "sum-probabilities", and "product-complement".
3. SRS, where we take a single-stage simple random sample of units.

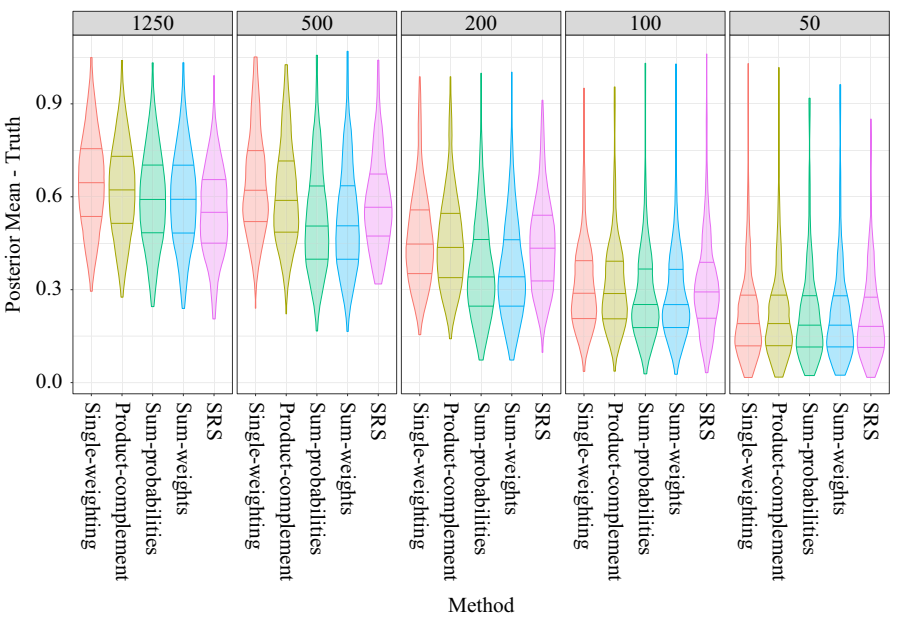


Fig. 6. Indirect sampling of groups: Monte Carlo distributions and quantiles (0.5%, 25%, 50%, 75%, 99.5%) for $B = 300$ iterations for σ_u^2 for difference of posterior means and truth under alternative weighting schema for varying number of groups, G_U. Population of $N = 5{,}000$ and sample of $n = 500$. In each plot panel, from left-to-right is the Single-weighting method, Product Complement double-weighting method of Equation (31), Sum-probabilities double-weighting method of Equation (26), Sum-weights double-weighting method of Equation (15) and Simple-random sampling (SRS).

The inclusion of model estimation under (a non-informative) SRS is intended to serve as a gold standard against which we may judge the bias and MSE performance of single- and double-weighting under informative sampling.

Each plot panel in Figure 6 shows the distributions over Monte Carlo simulations for estimates of the generating variance, σ_u^2, of the random effects, (u_g), under each of the following weighting methods: single-weighting, product-complement doubleweighting (Equation 31), sum-probabilities double-weighting (Equation 26), sum-weights double-weighting (Equation 15), and SRS (no weighting under simple random sampling of the same population from which the PPS sample was taken). The panels are ordered from left-to-right for a sequence of $G_U = (1,250, 500, 200, 100, 50)$. As earlier mentioned, the number of population units per group, N_h, is set to randomly vary under a lognormal distribution, though there will on average be more units sampled per group from synthetic populations with a smaller number of population groups, G_U, than there will be units per group sampled under a larger number of population groups. The sum-probabilities and sum-weights methods for accomplishing double-weighting generally perform nearly identically to one another and better than single-weighting for all group sizes. Since sum-

Table 2. Normalized Bias and Root Mean Squared error (RMSE) for Double-weighting methods as compared to Single-weighting; and Simple Random Sampling (SRS) for Increasing Units Per Random Effect Under Indirect Sampling of Groups for B = 300 generated data sets with population of N = 5,000 and sample of n = 500.

		Relative Bias		Normalized RMSE	
Model	G	beta_0	beta_1	beta_0	beta_1
Product-complement	1250	-0.54	-0.01	0.58	0.11
Single-weighting	1250	-0.55	0.00	0.58	0.11
SRS	1250	-0.58	0.00	0.61	0.11
Sum-probabilities	1250	-0.52	-0.01	0.55	0.11
Sum-weights	1250	-0.52	-0.01	0.55	0.11
Product-complement	500	-0.45	0.01	0.48	0.13
Single-weighting	500	-0.45	0.01	0.48	0.14
SRS	500	-0.46	0.00	0.49	0.11
Sum-probabilities	500	-0.41	0.00	0.45	0.13
Sum-weights	500	-0.41	0.00	0.45	0.13
Product-complement	200	-0.26	0.00	0.34	0.12
Single-weighting	200	-0.26	0.00	0.34	0.12
SRS	200	-0.29	0.00	0.36	0.11
Sum-probabilities	200	-0.23	0.00	0.33	0.12
Sum-weights	200	-0.23	0.00	0.33	0.12
Product-complement	100	-0.13	-0.02	0.30	0.13
Single-weighting	100	-0.13	-0.02	0.30	0.13
SRS	100	-0.19	0.00	0.33	0.12
Sum-probabilities	100	-0.13	-0.02	0.30	0.13
Sum-weights	100	-0.13	-0.02	0.30	0.13
Product-complement	50	-0.07	0.00	0.39	0.11
Single-weighting	50	-0.07	0.00	0.39	0.11
SRS	50	-0.08	-0.01	0.39	0.11
Sum-probabilities	50	-0.07	0.00	0.39	0.11
Sum-weights	50	-0.07	0.00	0.39	0.11

probabilities and sum-weights perform nearly identically, one may choose to prefer use of the former because it does not require our estimation of $\hat{N}_g$, as does the latter.

Table 2 presents the relative bias, defined as the bias divided by the true value, and the normalized root MSE, defined as the square root of MSE divided by the true value, for the regression coefficients, (β_0, β_1), to accompany Figure 6. We show the relative bias and normalized RMSE quantities in this study because the true values of the marginal model, $\sigma_u^2 = (0.578, 0.349, 0.216, 0.169, 0.136)$, varies over the sequence of sizes for G_U. As in the case of direct sampling of groups, there is an association between the amount of bias in estimation of σ_u^2 and in the intercept coefficient, β_0.

7. References

Asprouhov, T. 2006. "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics – Theory and Methods* 35: 439–460. DOI: https://doi.org/10.1080/03610920500476598.

Carpenter, B., A. Gelman, M.D. Hoffman, D. LEE, B. Goodrich, M. Betnacourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76: 1–32. DOI: https://doi.org/10.18637/jss.v076.i01.

Dong, Q., M.R. Elliott, T.E. Ragunathan. 2014. "A nonparametric method to generate synthetic populations to adjust for complex sampling design features." *Survey methodology* 40: 29. Available at: PMC5708580 (accessed February 2022).

Isaki, C.T., and W.A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96. DOI: https://doi.org/10.1080/01621459.1982.10477770.

Kim, J.K., S. Park, and Y. Lee. 2017. "Statistical inference using generalized linear mixed models under informative cluster sampling." *Canadian Journal of Statistics* 45: 479–497. DOI: https://doi.org/10.1002/cjs.11339.

Kunihama, T., A. Herring, C. Halpern, and D. Dunson. 2016. "Nonparametric Bayes modeling with sample survey weights." *Statistics & probability letters* 113: 41–48. DOI: https://doi.org/10.1016Zj.spl.2016.02.009.

Lele, S.R., K. Nadeem, and B. Schmuland. 2010. "Estimability and likelihood inference for generalized linear mixed models using data cloning." *Journal of the American Statistical Association* 105: 1617–1625. DOI: 10.1198/jasa.2010.tm09757.

Leon-Novelo, L.G., and T.D. Savitsky. 2021. "Fully Bayesian Estimation Under Dependent and Informative Cluster Sampling." *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smab037.

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society* 60: 23–40. DOI: https://doi.org/10.1111/1467-9868. 00106.

Rabe-Hesketh, S., and A. Skrondal. 2006. "Multilevel modelling of complex survey data." *Journal of the Royal Statistical Society* 169: 805–827. DOI: https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J.N.K., and C. Wu. 2010. "Bayesian pseudo-empirical-likelihood intervals for complex surveys." *Journal of the Royal Statistical Society* 72: 533–544. DOI: https://doi.org/10.1111/j.1467-9868.2010.00747.x.

Ribatet, M., D. Cooley, and A.C. Davison. 2012. "Bayesian inference from composite likelihoods, with an application to spatial extremes." *Statistica Sinica* 22: 813–845. Available at: JST0R:24310036 (accessed February 2022).

Savitsky, T.D., and D. Toth. 2016. "Bayesian estimation under informative sampling." *Electronic Journal of Statistics* 10: 1677–1708. DOI: https://doi.org/10.1214/16-EJS1153.

Savitsky, T.D., and M.R. Williams. 2020. "Pseudo Bayesian Estimation of Oneway ANOVA Model in Complex Surveys." *arXiv preprint.* Available at: arXiv:2004.06191 (accessed February 2022).

Shen, W., and T.A. Louis. 1998. "Triple-Goal Estimates in Two-Stage Hierarchical Models." *Journal of the Royal Statistical Society* 60: 455–471. DOI: https://10.1111/1467-9868.00135.

Si, Y., N.S. Pillai, and A. Gelman. 2015. "Bayesian Nonparametric Weighted Sampling Inference." *Bayesian Analysis* 10: 605–625. DOI: https://doi.org/10.1214/14-BA924.

Slud, E.V. 2020. "Nonidentifiability of Mixed-Model Parameters under Informative Sampling using only Single-Inclusion Weights." In *Proceedings of the Survey Research Methods Section, ASA*. Available at: http://www.asasrms.org/Proceedings/y2020/files/1505312.pdf (accessed November 2021).

Williams, M.R., and T.D. Savitsky. 2020. "Bayesian Estimation Under Informative Sampling with Unattenuated Dependence." *Bayesian Analysis* 15: 57–77. DOI: https://doi.org/10.1214/18-BA1143.

Williams, M.R., and T.D. Savitsky. 2021. "Uncertainty Estimation for Pseudo-Bayesian Inference Under Complex Sampling." *International Statistical Review* 89: 72–107. DOI: https://doi.org/10.1111/insr.12376.

Yi, G.Y., J.N.K. Rao, and H. Li. 2016. "A Weighted Composite Likelihood Approach for Analysis of Survey Data under Two-level Models." *Statistica Sinica* 26: 569–587. DOI: https://doi.org/10.5705/ss.2013.383.

Received March 2021
Revised August 2022
Accepted March 2022