A Model-Based Approach to Predict Employee Compensation Components

Andreea L. Erciulescu *† Jean D. Opsomer *†

Abstract

Surveys are often designed with the purpose of producing reliable estimates at upper levels of aggregation, but information at disaggregated levels may also be needed. The demand for official statistics at fine levels and at low cost is motivating researchers to explore alternative estimation methods. For this work, the challenge originated with the U.S. Bureau of Labor Statistics' National Compensation Survey (BLS's NCS), which aims to collect wage and non-wage compensation data of employees in the United States. For this survey program, the BLS is interested in producing hundreds of thousands of employee compensation statistics that serve as key data in producing economic indicators, such as the quarterly Employment Cost Index and the Employer Costs for Employee Compensation, of interest especially to government agencies and institutions. In this paper, a bivariate hierarchical Bayes model is developed using sparse survey data available for fine domains defined by geography, occupations, work levels within occupations, and job characteristics (union/non-union membership, full-time/part-time work status, and incentive-based/time-based pay), from the NCS. Model predictions are then constructed for the small set of domains, for which survey data are available, and for a much larger set of domains, for which survey data are not available. Also, discussed in this paper are methods for addressing challenges in identifying the prediction space, in constructing and selecting the information that serves as model input, and in making use of relationships between variables and domains structure.

Key Words: Bayes, benefits, bivariate models, small area estimation, wage

1. Introduction

With a growing demand for granular statistics, increasing costs of data collection and a need for integrating data from multiple sources, agencies are continuously increasing their interest in small area estimation models. The U.S. Bureau of Labor Statistics publishes monthly small area estimates of employment and unemployment, part of the Local Area Unemployment Statistics program. The U.S. Census Bureau has implemented small area estimation models in the production of official statistics of income and poverty measures for the Small Area Income and Poverty Estimates (SAIPE) program (Bell, Basel, and Maples, 2016). The U.S. National Agricultural Statistics Service is reviewing small area model-based estimates for the production of crops (Erciulescu, Cruze, and Nandram, 2019), cash rents (Erciulescu et al., 2018) and agricultural labor official statistics (Erciulescu, 2018). The Chilean Ministerio de Desarrollo (Casas Cordero Valencia, Encina, and Lahiri, 2016) and the World Bank (Elbers, Lanjouw, and Lanjouw, 2003) use small area estimation models for poverty mapping. These are all examples of important government programs that already employ or are actively considering small area estimation methods. The variety and complexity of each of these practical studies is difficult to generalize to one single strategy. In a recent article, Tzavidis et al. (2018) attempted to present such framework for the production of small area official statistics but limited the presentation to unit-level models, with fairly simple specifications, variance structures that ignored the design effect, and

*Westat, 1600 Research Blvd, Rockville, MD 20850.

†The views presented in this paper are those of the author(s) and do not represent the views of any Government Agency/Department or Westat. The authors thank Jeff Gonzalez and colleagues, from BLS's Office of Compensation and Working Conditions, for access to, and valuable discussions about, the data for the application.

prediction for domains with sample data only.

Small area estimation methods aim to improve on estimates that rely only on the observed survey data within the small domains, by borrowing information across all domains and by using auxiliary data, to allow for estimation in domains with no survey data and provide associated measures of uncertainty for granular statistics in a transparent, reproducible and validated process. There is rich and growing literature on small area estimation models, since the introduction of the major classes, in the pioneering papers by Fay and Herriot (1979) (area-level models) and Battese, Harter, and Fuller (1988) (unit-level models). For more details and extensions, we refer to comprehensive reviews of the field, including Ghosh and Rao (1994), Jiang and Lahiri (2006), Datta and Ghosh (2012), and Pfeffermann (2013), and the monograph Rao and Molina (2015).

In the frequentist framework, likelihood-based or approximate likelihood-based methods are common approaches for estimation of small area models parameters. Other approaches, which have proven to perform not as well as the likelihood-based methods, include method of moments and analysis of variance (ANOVA). For a simulation study comparing different estimation methods, we refer to Datta, Rao, and Smith (2005). In the majority of the literature, empirical best linear unbiased predictions (EBLUPs) are constructed under the framework of linear mixed models with an averaged squared error loss. A major methodological challenge is the derivation of an accurate estimator for the prediction mean squared error (PMSE) of the EBLUP, and an extensive literature on the topics exists, with several competing estimation approaches; for a classical case, we refer to Prasad and Rao (1990), who used an ANOVA-type estimator for the random effects variance component and derived a second-order unbiased estimator for the PMSE of the EBLUP, and for alternative, resampling, methods, we refer to Jiang and Lahiri (2006) and Rao and Molina (2015). Finally, the construction of prediction intervals for the EBLUPs have been of interest to many, but only few publications are available. The motivation for these studies is the fact that Wald-type intervals that use the PMSE estimator in lieu of the variance are often not accurate for areas with small sample sizes. Different corrections to prediction intervals have been studied by Datta et al. (2002) and Diao et al. (2014). Resampling methods have been used to construct bootstrap prediction intervals Hall and Maiti (2006), Chatterjee, Lahiri, and Li (2008), Erciulescu and Fuller (2018).

In a Bayesian framework, both empirical Bayes and hierarchical Bayes approaches have been considered in the literature as estimation approaches of small area models parameters. Under this framework, Markov Chain Monte Carlo (MCMC), often times in combination with other Bayesian sampling algorithms, have been greatly studied. On the other hand, for simple distributional assumptions with practical computational properties, MCMC may not even be needed; see, for example, Molina, Nandram, and Rao, 2014. In an empirical Bayes approach, distributional assumptions are made for the error terms, and the model parameters are treated as fixed, estimated a priori using a classical method. This approach was used in Fay and Herriot (1979) for constructing small area predictions and in Cox (1975) for constructing prediction intervals. In a hierarchical Bayes approach, distributional assumptions are made for the error terms and prior distributions are adopted for the nuisance parameters; when a priori information is not available, common choices are non-informative priors. In either of the two approaches, inference is based on the posterior distribution of the parameters of interest, given the data (survey direct estimates) and nuisance parameters. Arora and Lahiri (1997) proposed one of the initial hierarchical Bayes approaches to small area estimation and showed a case of superiority of the Bayes estimator of a small area over the corresponding BLUP, in terms of reducing the PMSE. One advantage of the Bayesian methods for estimation is that the full conditional distribution is derived for the parameters of interest, without additional computations or approximations; therefore, posterior means,

posterior variances, and posterior credible intervals may be obtained in a straightforward manner. Two additional advantages of the Bayesian methods are that, under appropriate model and priors specification, the posterior mean of a variance component will always be positive, and the implementation and prediction methods of complex functions of parameters can be handled straightforwardly. One drawback of the Bayesian methods may be the computational complexity and time.

Multivariate small area estimation models can be used to incorporate the relationship between two or more variables of interest. The motivation for modeling multiple variables of interest jointly instead of separately is that, when they are correlated, the small area estimates will be more precise under joint modeling. A small number of studies of the multivariate Fay-Herriot (FH) model have appeared in the literature. Fay (1987), Fuller and Harter (1987) and Datta, Day, and Basawa (1999) discussed the use of multivariate regression for small area estimation and constructed empirical Bayes predictions. Datta, Fay, and Ghosh (1991), Datta et al. (1996), and Datta, Day, and Maiti (1998) constructed hierarchical Bayes predictions using a multivariate small area model. Gonzalez-Manteiga et al. (2008) demonstrated the performance of the bivariate FH model using simulation studies; the authors used methods of moments for parameters estimation and compared a closed-form expression with a bootstrap estimator for the prediction mean squared error (a matrix referred to by the mean crossed product error matrix). Fay, Planty, and Diallo (2013) extended the dynamic model in Rao and Yu (1994), and use maximum likelihood estimation methods to estimate the parameters of a multivariate model. Concurrently, Krenzke et al. (2019) developed bivariate hierarchical Bayes area-level models for adult competency proportions.

The goal of this paper is to provide an estimation framework that expands on the classic small area estimation approaches. Whereas direct survey estimation relies only on the observed survey data within the small domains, the classic small area estimation approaches borrow strength from information across all domains and from auxiliary data. Lacking auxiliary data of good predictive power, at the level of interest, we explore the construction of covariates using categorical variables used in the domain definitions. Moreover, our proposed framework borrows strength from the relationship between multiple quantities of interest, in addition to borrowing strength from information across all domains and from covariates. Model predictions and associated measures of uncertainty are constructed for the set of domains with sample data, as well as for a set of domains with no sample data. Unlike classical small area estimation approaches, where predictions are constructed for all the theoretically possible domains, we define the set of domains plausible for prediction a priori, using ancillary information. The methodology developed is a transparent, reproducible and validated process, that relies on hierarchical Bayes methods for model fit and prediction.

In Section 2, we introduce the motivation for this work, describe the data available for the application study, and present a framework for direct survey estimation. Small domain estimation models are described in Section 3, in the form of bivariate Fay-Herriot models, along with a framework for model fit and prediction. Model validation methods are presented in Section 4. Of interest to the motivation study, results for the prediction of employee compensation components are provided throughout the paper. A discussion of practical challenges and concluding remarks are given in Section 5.

2. Application Background

The motivation for this work is the interest in small area estimation methodology at the U.S. Bureau of Labor Statistics' (BLS') Office of Compensation and Working Conditions. The

BLS' National Compensation Survey (NCS) production and publication has experienced a decrease in sample size as a result of the enactment of the 2011 federal budget. See Lettau and Zamora (2013), for more information on the NCS program. However, interest remains in potentially producing hundreds of thousands of compensation statistics by geography, occupation, classification, and work level. The official statistics produced using NCS data serve as key data in producing economic indicators, such as the quarterly Employment Cost Index and the Employer Costs for Employee Compensation, of interest especially to government agencies and institutions.

The NCS is a nationwide establishment-based survey conducted to collect data on labor cost and compensation components, such as wages, salaries, and benefits. NCS data are collected on the pay period that includes the 12th day of the month, using a two-stage sample design: in the first stage, a probability proportional to size sample of establishments is drawn from the list frame, and in the second stage, a sample of occupations (to be denoted by 'quotes') is drawn within the establishments sampled in the first stage, using a complex four-step procedure. The occupations are defined using the six-digit codes from the Standard Occupational Classification (SOC) system. Official statistics produced using NCS data are published four times a year, for the reference months of March, June, September, and December. For more information on the NCS, we refer to Bureau of Labor Statistics (2018).

2.1 Data

For our application, the data are available from the NCS, at the quote level, for June 2017. In particular, sample data on employee compensation components, wage and benefits, measured in $/hr, are available for 36,790 quotes with positive compensation. The area, or the domain, of interest is defined by the combination of geography, occupation, work level, and characteristics (time/incentive, full-time/part-time, union/nonunion). Geography is defined by the nine census divisions (groups of states). Occupation is defined in O*NET-SOC, six-digits SOC system codes being of interest, and there are a total of 759 in the sample data. Work levels range from 1 to 15 and are missing for some of the sample data records. The total number of cells that can be constructed by cross-tabulating these categorical variables, and corresponding to the total number of *theoretically possible* domains, is 874,368. It is important to note that sample data are only available for 16,107 domains, with most domains having very small or zero realized sample size. On the other hand, information on the classification of six-digit SOC system codes and corresponding work levels is available from the 2010 SOC system. From that set, we construct the prediction set of interest to the application, leading to 572,328 *plausible* domains; the range of work levels varies across occupations. Finally, base weights and replicate weights are available for the sample data. In particular, a set of 120 replicate weights was constructed for the balanced repeated replication (BRR) method, with Fay adjustment factor of 0.5 (see Judkins, 1990). These replicates will be used in the construction of sampling variances for the domains of interest.

Most of the small area estimation studies in the literature rely on the availability of auxiliary data, from other surveys, administrative sources or remote sensing sources, to name a few. However, often, difficulties arise in obtaining auxiliary data that align with the survey data in terms of time and granularity, that are free or error or subject to quantifiable error, and that are good covariates for regression models fitted to the survey data. For our application study, such data are not available, and we propose an alternative method for constructing model covariates, as will be described in the next section.

2.2 Survey Summaries

Let $i = 1, \ldots, m$ index the target domains with n_i the sample size in domain i, $j = 1, \ldots, n_i$ the observations (quotes) in the domain, and y_{ij} the value of the target variable (compensation: $y_{1,ij}$ and $y_{2,ij}$ for wage and benefits compensation, respectively) for unit j in domain i. Let Y_i denote the population quantity of interest in domain i, for instance the domain total $\sum_{j=1}^{N_i} y_{ij}$ or domain mean $\sum_{j=1}^{N_i} y_{ij}/N_i$, with N_i the number of population units in the domain. We write $\hat{Y}_i$ as the direct estimator of Y_i, which is constructed using only the survey data and the survey design weights in domain i. Typically, direct estimators are of the Horvitz-Thompson or inverse-probability-weighted type. Note that n_i can be zero, in which case $\hat{Y}_i$ is undefined.

For our application, the estimation targets are the mean hourly wage and benefits compensation in each domain. Their direct estimators are defined as:

$$\hat{Y}_{1i} = \frac{\sum_{j=1}^{n_i}(w_{ij}y_{1,ij})}{\sum_{j=1}^{n_i} w_{ij}},\ \hat{Y}_{2i} = \frac{\sum_{j=1}^{n_i}(w_{ij}y_{2,ij})}{\sum_{j=1}^{n_i} w_{ij}}.$$

Following Guciardo (2019), we construct their associated replicate variance and covariance estimates using the BRR method, with Fay adjustment,

$$\begin{aligned}
\hat{V}(\hat{Y}_{1i}) &= \frac{\sum_{r=1}^{R}(\hat{Y}_{1i}^r - \hat{Y}_{1i})^2}{R(1-k)^2} \\
\hat{V}(\hat{Y}_{2i}) &= \frac{\sum_{r=1}^{R}(\hat{Y}_{2i}^r - \hat{Y}_{2i})^2}{R(1-k)^2} \\
\widehat{Cov}(\hat{Y}_{1i}, \hat{Y}_{2i}) &= \frac{\sum_{r=1}^{R}\left((\hat{Y}_{1i}^r - \hat{Y}_{1i})(\hat{Y}_{2i}^r - \hat{Y}_{2i})\right)}{R(1-k)^2},
\end{aligned}$$

where $R = 120, k = 0.5$, and

$$\hat{Y}_{1i}^r = \frac{\sum_{j=1}^{n_i}(w_{ij}^r y_{1,ij})}{\sum_{j=1}^{n_i} w_{ij}^r},\ \hat{Y}_{2i}^r = \frac{\sum_{j=1}^{n_i}(w_{ij}^r y_{2,ij})}{\sum_{j=1}^{n_i} w_{ij}^r}$$

are the direct estimators constructed using the replicate weights, for $r = 1, ..., R$.

The direct point and variance/covariance estimates just defined are the key inputs into the domain-level models to be considered in the next section. It is not necessary to have estimates in all domains of interest, since the small domain model will be able to handle prediction for unobserved domains. However, observed domains i, for $i = 1, ..., m$, require valid sets of estimates, in particular, non-zero point and variance estimates and non-singular variance-covariance matrices,

$$\begin{bmatrix} \hat{V}(\hat{Y}_{1i}) & \widehat{Cov}(\hat{Y}_{1i}, \hat{Y}_{2i}) \\ \widehat{Cov}(\hat{Y}_{1i}, \hat{Y}_{2i}) & \hat{V}(\hat{Y}_{2i}) \end{bmatrix}.$$

Let the domains of interest be also denoted as fine-level domains (16,107 domains); recall that these domains are defined by the cross-tabulations of census divisions, six-digit SOC system codes, work levels, full-time/part-time, union/nonunion, time/incentive. In contrast, let high-level domains (197 domains) be defined by the cross-tabulations of census divisions and two-digit SOC system code. Then, survey summaries are provided in Table 1: sample sizes, effective sample sizes $n_i^{eff} := (\sum_j w_{ij}^2)^{-1}(\sum_j w_{ij})^2$, effective sample size in domain i, and coefficients of variation (CVs) for wage and benefits survey estimates.

Note that the domain-level sample sizes are very small, having a median of 1, for the fine-level domains, and reasonably large for the high-level domains; similar results for the effective sample sizes. Despite this result, most of the CVs for the fine-level domains estimates are smaller than the CVs for the high-level domains estimates, as a consequence of unreliable survey estimates at fine levels of granularity.

Table 1: Survey Summaries

Statistic	Fine Level				High Level			
	Sample Size	Effective Sample Size	CV(%)		Sample Size	Effective Sample Size	CV(%)	
			Wage	Benefits			Wage	Benefits
Minimum	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
1st Quantile	1.00	1.00	0.00	0.00	60.00	23.90	3.77	6.00
Median	1.00	1.00	0.00	0.00	127.00	48.41	5.63	8.67
Mean	2.28	1.85	2.62	4.69	186.75	68.98	7.40	11.56
3rd Quantile	2.00	1.99	2.80	5.27	234.00	90.85	9.55	14.45
Maximum	74.00	45.86	68.12	92.20	1308.00	367.77	39.12	80.88

Preliminary data investigation and estimation using multiple linear regression models, fitted to the survey estimates of wage and benefits, separately, indicated overall departures from common normality assumptions, which adds an additional complication to the small area estimation modeling. Therefore, we apply a log transformation to the survey estimates,

$$y_{1i} := log(\hat{Y}_{1i}),\ y_{2i} := log(\hat{Y}_{2i}),$$

and transform the associated variance-covariance matrices estimates with entries

$$\sigma^2_{ei,1} := \frac{\hat{V}(\hat{Y}_{1i})}{\hat{Y}^2_{1i}},\ \sigma^2_{ei,2} := \frac{\hat{V}(\hat{Y}_{2i})}{\hat{Y}^2_{2i}},$$

$$\sigma_{ei,12} := \widehat{Cov}(\hat{Y}_{1i}, \hat{Y}_{2i}) \frac{\sigma_{ei,1}\sigma_{ei,2}}{\sqrt{\hat{V}(\hat{Y}_{1i})\hat{V}(\hat{Y}_{2i})}}.$$

Figure 1 illustrates the effect of such transformation on the overall distributions of the domain-level survey estimates.

We note that a possible alternative approach for variance estimation for the transformed estimates consists of applying the replication variance estimation formulas directly to the transformed variables. In our application, the survey summaries for the two sets of variance estimates were similar, so we proceeded with the variance transformation approach. One practical advantage of the variance transformation approach is that it can be applied to the domain-level summary statistics, without the need for record-level data.

To mitigate the zero or very small, fine-level domains variances estimates, we developed a three-step procedure. First, we impute for zero or missing high-level variances (covariance) estimates using the mean (median) of the available, positive, variances (covariances) estimates. Then, we project the high-level variances and covariances estimates onto fine levels, inversely proportional to the effective sample sizes,

$$\Sigma_{ei,1,1} := (n_i^{eff})^{-1}\sigma^{2,H}_{ei,1}n_i^{eff,H},\ \Sigma_{ei,2,2} := (n_i^{eff})^{-1}\sigma^{2,H}_{ei,2}n_i^{eff,H},$$

$$\Sigma_{ei,1,2} := (n_i^{eff})^{-1}\sigma^{H}_{ei,12}n_i^{eff,H},$$

where the superscript H corresponds to high-level domains statistics. Finally, we approximate the variance-covariance matrices by the nearest positive-definite matrices (see

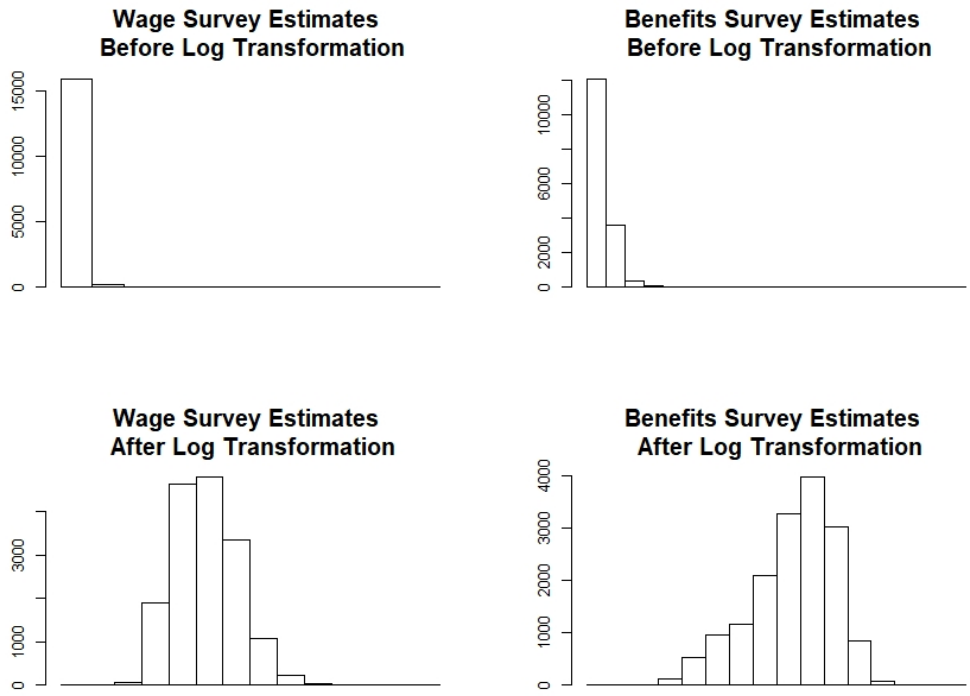

Figure 1: Distributions of the domain-level wage and benefits survey estimates, before and after the transformation and revision, constructed using the sample data. The x-axes are removed due to disclosure limitations.

Higham, 1988). Figure 2 illustrates the effect of such revisions on the overall distributions of the domain-level survey variances and covariances estimates.

3. Models

Model-based estimation methods are useful tools for improving survey-based estimation for domains with small amounts of data because they provide a principled way to incorporate the survey estimates, the relationships between variables and external (auxiliary) information, while also accounting for sources of uncertainty in these components. In the context of small domain estimation, the models are fit to the set of domains with sample data, but predictions are constructed for all the domains in the prediction space.

In this section, we introduce the small domain estimation models as hierarchical models, with a sampling level corresponding to the model for the domain direct estimates and a linking level corresponding to the model for the domain target quantities and auxiliary data. To incorporate the relationship between two or more variables of interest, we consider multivariate small domain estimation models. In particular for the application, wage and benefits compensation components are jointly modeled. The motivation for modeling multiple variables of interest jointly instead of separately is that, when they are correlated, the small domain estimates can be expected to be more precise under joint modeling. In the application using the NCS data under consideration, the correlation between the domain-level wage rate estimates and the corresponding benefits estimates, for fine-level domains is approximately 0.8. Therefore, in the application study, we expect some reductions in the

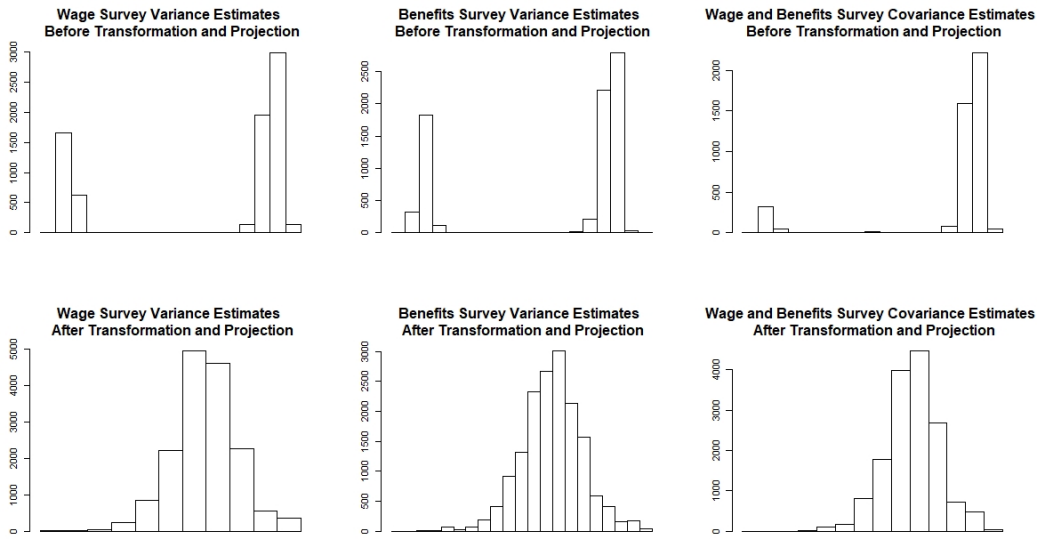

Figure 2: Distributions of the domain-level wage and benefits survey variances and covariances estimates, before and after the transformation and revision, constructed using the sample data. The x-axes are removed due to disclosure limitations.

variances of the model predictions based on a bivariate model, compared to variances of model predictions based on separate univariate models.

3.1 Model Matrix

As mentioned in the previous section, auxiliary data available from sources other than the NCS are not available at the level of granularity of interest for this study. Therefore, we explore the possibility of constructing covariates from the survey data available; in particular, the categorical variables that are used to define the domains of interest, as well as their interactions. In this subsection, we describe a method for constructing the model matrix.

First, a full model matrix is constructed by cross-tabulating the categorical variables defined by the census divisions, six-digit SOC system codes, work levels, full-time/part-time, union/nonunion, time/incentive, and their two-way interactions. The number of rows in the resulting matrix equals to the number of domains in the sample data. Because all of the variables considered are categorical, with different number of levels each (9, 759, 16, 2, 2, 2, respectively), the resulting model matrix is a binary matrix, with multiple columns corresponding to each of the six variables levels, and additional columns corresponding to the interaction terms. Therefore, the number of columns in the full model matrix equal to 20,684 $(= (9-1)+(759-1)+(16-1)+(2-1)+(2-1)+(2-1)+((9-1)+(759-1)+(16-1))\times((2-1)+(2-1)+(2-1))+(9-1)\times(759-1)+(9-1)\times(16-1)+(759-1)\times(16-1)+(2-1)\times(2-1)+(2-1)\times(2-1)+(2-1)\times(2-1))$.

To reduce the number of dimensions of the model matrix, we implemented three classic selection procedures: a multiple or multi-task learning method, a decision trees method, and a likelihood-based method. For these, each column in the model matrix is treated as a covariate. Note that we violate the interaction criterion by selecting levels of the original variables or their interactions, without constraining on selecting all their main effects lev-

els; this is not a concern because our interest is in prediction, and not in determining and explaining the effects of each of the variables on the final predictions.

Aligning with the goal of modeling jointly wage and benefits compensation components, we start the selection of columns in the model matrix with a 10-fold cross-validated fit of a multivariate Lasso, assuming multivariate normal distribution. Applying a multivariate Lasso penalty results in a set of selected covariates across both responses (wage and benefits survey estimators), determined using the tuning parameter that gives the minimum mean cross-validated error. This effort is validated by previously fitted multiple linear regression models to wages and benefits, separately, using the same set of covariates defined using the six categorical variables under discussion, that resulted in comparable coefficients of determination. Finally, we impose a constraint that all the estimated regression coefficients be greater than 0.03, in absolute value. The set of selected covariates using multivariate Lasso consists of 612 variables.

To further reduce the dimensions of the model matrix, we apply two additional variable selections methods, separately to wage survey estimates and benefits survey estimates, starting with the set of covariates selected by the multivariate Lasso-type method described above. A subset of covariates is selected using an algorithm of decision trees models that recursively partitions the data until the outcomes in the final disaggregated group are as nearly homogeneous. A second subset of covariates is selected using stepwise selection using Akaike Information Criterion with both backward and forward selection. The final set of covariates is defined as the intersection of the these two sets, and consists of 18 variables; see Appendix A for a descriptive list of the final set of covariates selected for modeling. The final model matrix includes these 18 variables and an intercept.

Multiple linear regression models are again fit to the wage survey estimates and benefits survey estimates, to asses the predictive power of the final set of covariates selected for modeling. Plots of the survey estimates against the fitted values, as well as the coefficients of determination, for these two models, are illustrated in Figure 3. Assuming homogeneous sampling variances across the domains (which is not the case), the results in Figure 3 suggest a resonable percentage of the sampling variability is being explained by the set of selected covariates. Next, we allow for heterogeneous sampling variances across the domains, for joint modeling of wage and benefits survey estimates, and for use of information across the domains; expecting additional variability to be explained by the last two.

3.2 Small Domain Hierarchical Bayes Bivariate Model

We previously described the construction of the model input: survey estimates of wage and benefits, with associated variance-covariances matrices and a set of covariates with reasonable predictive power. Now, the proposed bivariate model is developed as a hierarchical Bayes model of the form:

$$
\begin{array}{rrcl}
\text{Sampling Model:} & y_i|(\theta_i, \Sigma_{ei}, \beta, \Sigma_v) & \sim & \text{N}(\theta_i, \Sigma_{ei}) \\
\text{Linking Model:} & \theta_i|(\beta, \Sigma_v) & \sim & \text{N}(x_i\beta, \Sigma_v) \\
 & & & \\
\text{Priors:} & \beta & \sim & \text{N}(0, 10^4 I), \text{ component-wise} \\
 & \Sigma_v & \sim & \text{Inverse-Wishart}(3, I)
\end{array}
$$

where $i = 1, \cdots, m = 16,107$ is an index for the small domains (cross-tabulations of census divisions, six-digit SOC system codes, work levels, binary characteristics), y_i is the vector of domain-level direct survey estimates for wage and benefits, θ_i is a vector of the two quantities of interest (true wage and true benefits), x_i is the final model matrix defined in the previous section, of dimensions $m \times p = 16,107 \times 19$, β is a vector of $p = 19$

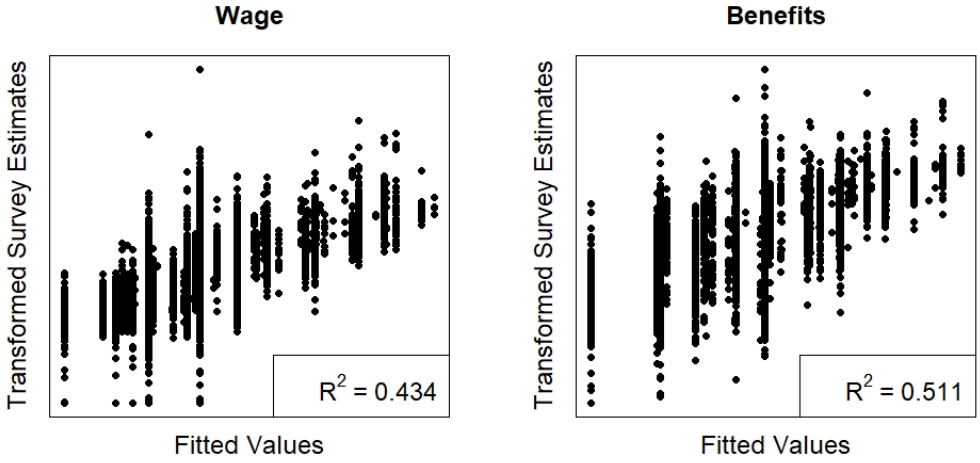

Figure 3: Survey estimates against the fitted values from multiple linear regression models, fitted separately to wage survey estimates and benefits survey estimates

unknown regression coefficients, Σ_v is the linking model variance-covariance matrix, of dimension 2×2, and Σ_{ei} is a known survey variance-covariance matrix, of dimensions 2×2. Independent, noninformative, proper priors are adopted for the parameters β and Σ_v.

The relationship between the small domain population value θ_i and the auxiliary data x_i is described in the linking model. Note that the population value θ_i, which is usually treated as fixed in the survey context, is a random variable under this model specification. Inference is based on the predictive distribution, $[\theta_i|y_i, x_i, \Sigma_{ei}, \beta, \Sigma_v]$.

The model is fit using MCMC, with three chains, each chain starting with 3,000 samples, of which 1,000 samples are burned in. To speed up the convergence, the least squares estimates from the multiple regression models, fit separately to wage and benefits survey estimates, are used as initial values for the regression parameters; the least squares estimates are deviated by $(-0.1, 0, 0.1)$ for the three chains, respectively. To reduce the autocorrelation in the chains to make inference that assumes nearly independent samples, we thin each of the chains so that every 10^{th} iteration is kept. The resulting set of 600 samples is used for inference.

3.3 Prediction

The model predictions are composites of direct survey estimates and synthetic predictions based on the hierarchical Bayes model described above. Direct estimates with large sampling variances, relative to all the sampling variances, would be smoothed more than the others, toward a common trend prediction that depends on the linear relationship assumed between the direct estimates and the covariates, as well as the bivariate relationship assumed for the wage and benefits compensation components. In this section, we will discuss prediction for two types of domains: in-sample domains, defined as the domains where survey sample data are available, and not-in-sample domains, defined by the domains where survey sample data are not available, but classification data from the SOC suggests they are plausible and, hence, are part of the prediction space.

Using the R samples $\theta_{i\zeta}, \zeta = 1, \cdots, R$, from the predictive distribution $[\theta_i|y_i, x_i, \Sigma_{ei}, \beta, \Sigma_v]$, posterior summaries such as means, variances, credible intervals, may be constructed for the wage and benefits, on the log scale. Since the interest is in these quantities

on the original scale, we also consider back-transforming the samples using an exponential transformation, $\theta^*_{i\zeta} := \exp(\theta_{i\zeta})$. Then, final model predictions are defined as

- point estimates: $R^{-1}\sum_{\zeta=1}^{R}\tilde{\theta}_{i\zeta}$,
- variance estimates: $R^{-1}\sum_{\zeta=1}^{R}\left(\tilde{\theta}_{i\zeta} - R^{-1}\sum_{\zeta=1}^{R}\tilde{\theta}_{i\zeta},\right)^2$
- p quantile estimates: $\tilde{\boldsymbol{\theta}}_{i(p)}, \tilde{\boldsymbol{\theta}}_i = (\tilde{\theta}_{i1}, \cdots, \tilde{\theta}_{iR})$,

where $\tilde{\theta}_i, \tilde{\theta}_i \in \{\theta_i, \theta^*_i\}$.

Prediction for not-in-sample domains relies on the model output and on the covariates available for these additional domains that were not part of the set of domains to which the model was fit. For the set of domains $i^{'}$, with no sample data, the covariates matrix with rows $x_{i^{'}}$ is constructed similarly to the model matrix described previously.

Recall that the additional information on the not-in-sample domains is limited to the cross-tabulation of six-digit SOC system codes and the plausible work levels within. Therefore, we assume that all the other job characteristics apply to the set of plausible cross-tabulations of six-digit SOC system codes and work levels. In the application, there are 743 six-digit SOC system codes with available information on work levels. The final model matrix is constructed after stacking 743 submodel matrices, each submatrix being specific to a six-digit SOC system code.

To construct the submodel matrices, we propose the following algorithm. For a given six-digits SOC system code, the full model matrix is built using the cross-tabulation of census division categories, *plausible* work level categories, binary characteristics, and their two-way interactions, with the rows that correspond to not-in-sample domains only. Then, the dimension of the sub-model matrix is reduced by selecting main effects and two-way interactions that use common variables to the ones used to define the columns in the model matrix used to fit the model. Finally, an intercept term is included in the submodel matrix. The number of not-in-sample domains ranges from 353 to 864, across the 743 six-digit SOC system codes considered, leading to a not-in-sample model matrix of dimensions 556,197 $\times$ 19.

Finally, for a domain $i^{'}$, with no sample data, we generate R samples $\theta_{i^{'}\zeta}, \zeta = 1, \cdots, R$, from $N(x_{i^{'}}\beta_\zeta, \Sigma_{v\zeta})$, and transform them back to the original scale $\theta^*_{i^{'}\zeta} := \exp(\theta_{i^{'}\zeta})$. Similarly to the prediction for in-sample domains, posterior summaries for domain $i^{'}$ are constructed using the set of samples $\tilde{\theta}_{i^{'}}, \tilde{\theta}_{i^{'}} \in \left\{\theta_{i^{'}}, \theta^*_{i^{'}}\right\}$, for a not-in-sample domain $i^{'}$.

4. Model Validation

As is the case with any modeling approach, validation is a critical piece in the development. We discuss two types of model validation checks, that were conducted iteratively with the model fit and prediction steps, and led to the final model specification presented in the previous section. For example, the final model matrix, the revised sampling variances, the normal distribution in the sampling model assumed for the log transformed direct survey estimates, the noninformative proper priors adopted for the model parameters, the number of MCMC samples, including burn-in and thinning, are consequences of the iterative steps implemented between model fit and prediction, and model validation.

4.1 Internal Model Validation

In this subsection, we discuss internal model validation. For this, we consider mixing and convergence diagnostics for the MCMC sampler, residual diagnostics for the normality

assumptions, and posterior predictive checks for the normality and linearity assumptions, and for other characteristics of the sample data (such as order statistics for each of the response variables, or correlation between the two response variables).

4.1.1 Convergence and Mixing Diagnostics

Convergence and mixing diagnostics for the MCMC sampler are presented in Table 2, using summaries of the Gelman-Rubin $\hat{R}$ statistic and the MC effective sample size, across all of the monitored parameters (regression coefficients, variance-covariance matrices, log-likelihood). The ranges of these two statistics are acceptable; see Gelman et al. (2013). While it is preferred that the $\hat{R}$ be below 1.1, we do report one value of 1.13, for one of the variance-covariance parameters for benefits but do not consider it a major concern. Also, the effective sample sizes are above 5% of the number of samples used for inference, which is again acceptable. The diagnostics presented in Table 2 may be improved by running longer MC chains, but we believe the improvement in convergence and mixing would not outgain the cost in computational time, while the changes in model predictions and their variances would be negligible.

Table 2: Convergence and Mixing Diagnostics

Statistic	$\hat{R}$	MC Effective Sample Size
Minimum	1.00	55.00
1st Quantile	1.00	430.00
Median	1.00	600.00
Mean	1.00	511.10
3rd Quantile	1.00	600.00
Maximum	1.13	600.00

4.1.2 Residuals

Inspired by the transformed residuals computed in Battese, Harter, and Fuller (1988), we define unconditional residuals as $\frac{y_{i,1} - x_i\hat{\beta}}{\sqrt{\hat{\sigma}^2_{v,1} + (\sigma_{ei,1})^2)}}$ and $\frac{y_{i,2} - x_i\hat{\beta}}{\sqrt{\hat{\sigma}^2_{v} + (\sigma_{ei,2})^2)}}$, corresponding to wages and benefits, respectively; the authors constructed unit-level transformed residuals that are approximately independent and identically distributed with mean zero and variance equal to the common constant across all the units, while we constructed domain-level transformed residuals that are approximately independent and identically distributed with mean zero and variance equal to one. Alternative residuals may be constructed as conditional residuals, $\frac{y_{i,1} - \hat{\theta}_{i,1}}{\sigma_{ei,1}}$ and $\frac{y_{i,2} - \hat{\theta}_{i,2}}{\sigma_{ei,2}}$, for wage and benefits, respectively. The subscripts 1 and 2 correspond to the two quantities modeled, wage and benefits, for the vectors being the first or second entry, and for the matrices being the first or second diagonal elements.

Residuals plots against the fitted values $x_i\hat{\beta}$ are presented in the first columns of Figures 4 and 5, for the unconditional and conditional residuals, respectively. Despite the observed decrease in the residuals values with an increase in the fitted values, the scatterplots indicate no significant departure from the homogeneous variance assumption for residuals. Normal quantile-quantile plots are presented in the second columns of Figures 4 and 5. The plots in Figure 5 indicate slightly heavier tails because the conditional residuals are standardized by

the sampling errors only. Overall, the quantile-quantile plots for benefits look better than the ones for wages, but no major concerns are indicated; we would expect all the points to fall close to the 45 degrees line in the case of independent and identically distributed residuals; however, that is not the case here (wages and benefits were modeled jointly and the predictions depend on the estimated variance parameters).

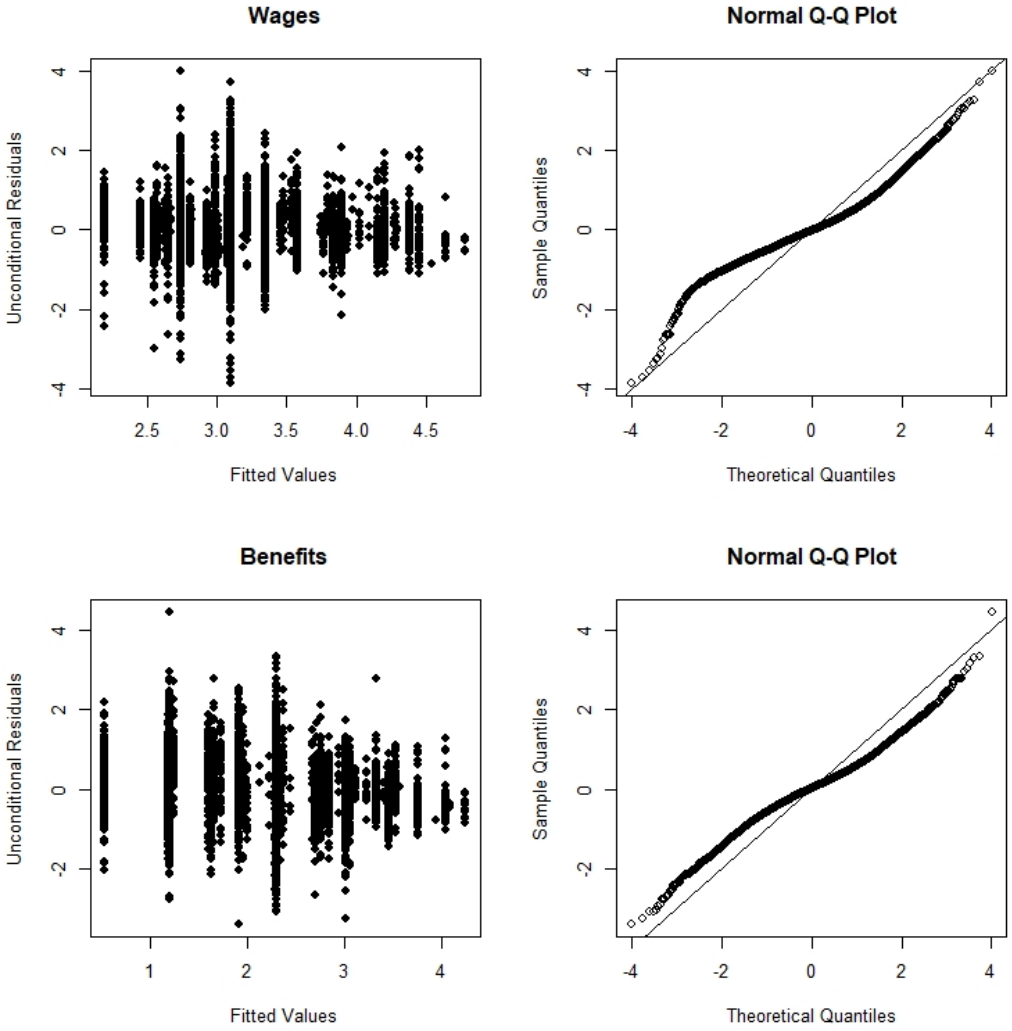

Figure 4: Unconditional residuals. Scatterplots, versus fitted values $x_i\hat{\beta}$, and normal quantile-quantile plots, for wage and benefits

4.1.3 Posterior Predictive Checks

To further assess the adequacy of the model fit, we conducted posterior predictive checks. In particular, for a set of pre-defined statistics, we compare their posterior predictive distribution to their corresponding values obtained using the original sample. The procedure to generate data from the posterior predictive distribution is as follows. Consider the posterior samples for β and Σ_v, denoted by β^t and Σ_v^t, respectively, for $t = 1, ..., T = 600$. Construct θ_i^t and draw replicates y_i^t following the model:

$$\begin{aligned} \theta_i^t &\sim N(x_i'\beta^t, \Sigma_v^t), \\ y_i^t &\sim N(\theta_i^t, \Sigma_{ei}), \end{aligned}$$

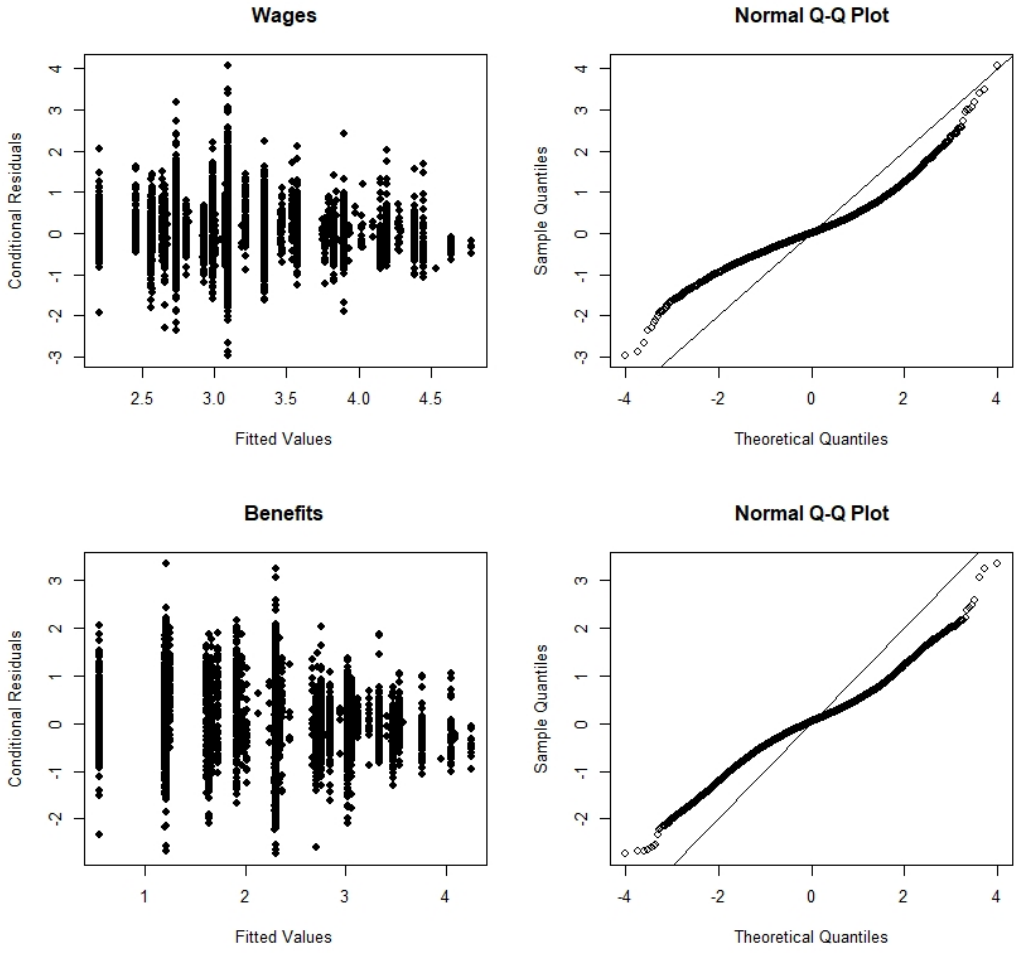

Figure 5: Conditional residuals. Scatterplots, versus fitted values $x_i\hat{\beta}$, and normal quantile-quantile plots, for wage and benefits

For this application, we consider three test statistics, with the corresponding posterior predictive quantities as:

- Deviance, $T^{-1}\sum_t(y_i^t - \theta_i^t)$, $i = 1, ..., m$,
- Unscaled Residual, $T^{-1}\sum_t y_i^t - y_i$, $i = 1, ..., m$,
- Scaled Residual, $\frac{T^{-1}\sum_t y_i^t - y_i}{(T-1)^{-1}\sum_t(y_i^t - T^{-1}\sum_t y_i^t)^2}$, $i = 1, ..., m$,
- Identity, y_i^t, $i = 1, ..., m$, $t = 1, ..., T$,
- Correlation, $Cor((y_{i,1})^t, (y_{i,2})^t)$, $i = 1, ..., m$, $t = 1, ..., T$.

Posterior predictive p-values are constructed for the last two test statistics as:

- Indicator, $\sum_i I\left\{y_i^t \geq y_i\right\}$, $t = 1, ..., T$,
- Correlation, $\sum_i I\left\{Cor((y_{i,1})^t, (y_{i,2})^t) \geq Cor((y_{i,1}), (y_{i,2}))\right\}$, $t = 1, ..., T$.

By definition, the posterior predictive p-value is the proportion of summary statistics calculated with samples generated from the posterior predictive distribution that exceed the

corresponding value based on the original sample. A p-value close to 0.5 indicates that the model provides a reasonable fit to the sample data.

Based on the summary results in Tables 3, 4, and 5, there is a tendancy to generate more smaller values than the survey estimates (distributions for unscaled and scaled residuals have slightly longer lower tails than upper tails), and joint estimates with slightly larger correlation than the sample correlation. Nevertheless, overall, there is no substantial indication of model lack of fit.

Table 3: Posterior Predictive Checks, Wages

Statistic	Deviance (statistics)	Unscaled Residual (statistics)	Scaled Residual (statistics)	Indicator (p-values)
Minimum	-0.005	-3.349	-3.606	0.480
1st Quantile	-0.004	-0.113	-0.262	0.489
Median	-0.003	-0.001	-0.002	0.492
Mean	-0.003	-0.020	-0.028	0.492
3rd Quantile	-0.002	0.099	0.242	0.495
Maximum	-0.001	1.245	2.403	0.506

Table 4: Posterior Predictive Checks, Benefits

Statistic	Deviance (statistics)	Unscaled Residual (statistics)	Scaled Residual (statistics)	Indicator (p-values)
Minimum	-0.002	-3.357	-2.993	0.479
1st Quantile	-0.001	-0.196	-0.283	0.491
Median	-0.001	-0.020	-0.034	0.494
Mean	-0.001	-0.005	-0.014	0.494
3rd Quantile	-0.001	0.173	0.247	0.497
Maximum	-0.000	2.358	2.416	0.505

Table 5: Posterior Predictive Checks, Wages and Benefits

Statistic	Correlation (p-values)
Minimum	0.759
1st Quantile	0.814
Median	0.816
Mean	0.816
3rd Quantile	0.819
Maximum	0.825

4.2 External Model Validation

External model validation checks are based on comparisons of survey estimates to model predictions. For domains where sample data are available, direct comparisons of model and survey estimates may be conducted. The two plots in the first row in Figure 6 illustrate the relationship between the model predictions and the survey direct estimates, for wages and benefits, respectively. Note that there are deviations from the 45 degrees line, corresponding mostly to the domains with large sampling variances (after transformation, and revision). The two plots in the second row in Figure 6 illustrate the relationship between the model predictions variances and the survey direct estimates variances, for wages and benefits, respectively. Note that the model prediction variances are smaller than the sampling variances (after transformation, and revision), for most of the domains. These observations demonstrate the model performance with respect to improving survey-only estimation for the domains where sample data are available but only domain-specific data are used in the estimation at the domain level.

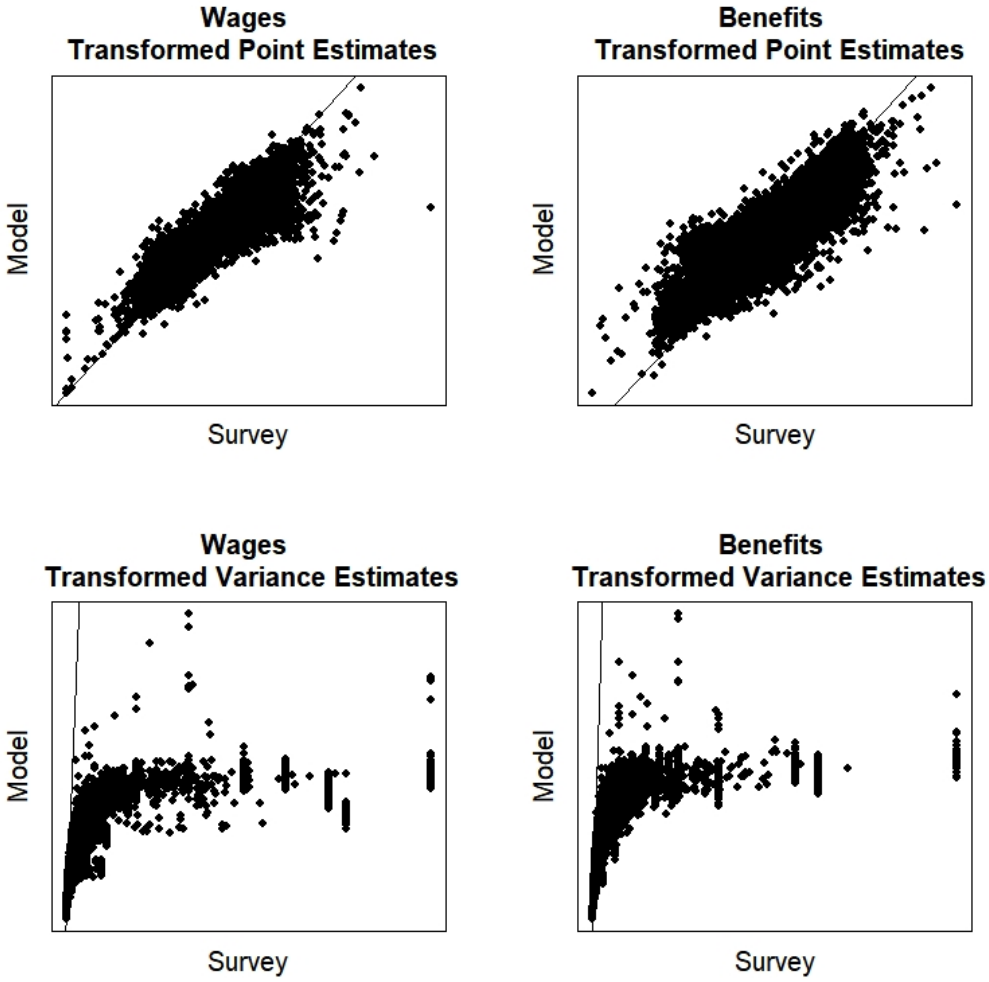

Figure 6: Model predictions versus survey direct estimates, for wages and benefits; point estimates

A different way of visualizing the results in Figure 6 is to illustrate relative measures of model predictions versus survey direct estimates. For this, we consider two relative

measures: the difference between the model predictions and the survey direct estimates, and the ratio of the model predictions variances to the survey direct estimates variances. The estimated relative measures for the domains with sample data are plotted against the domain effective sample sizes, in Figure 7. In agreement with the conclusions above, the largest differences between the model predictions and the survey direct estimates are for domains with small effective sample sizes, and the ratios between the model variances and the sampling variances (after transformation, and revision) increase (to 1) with an increase in the effective sample size.

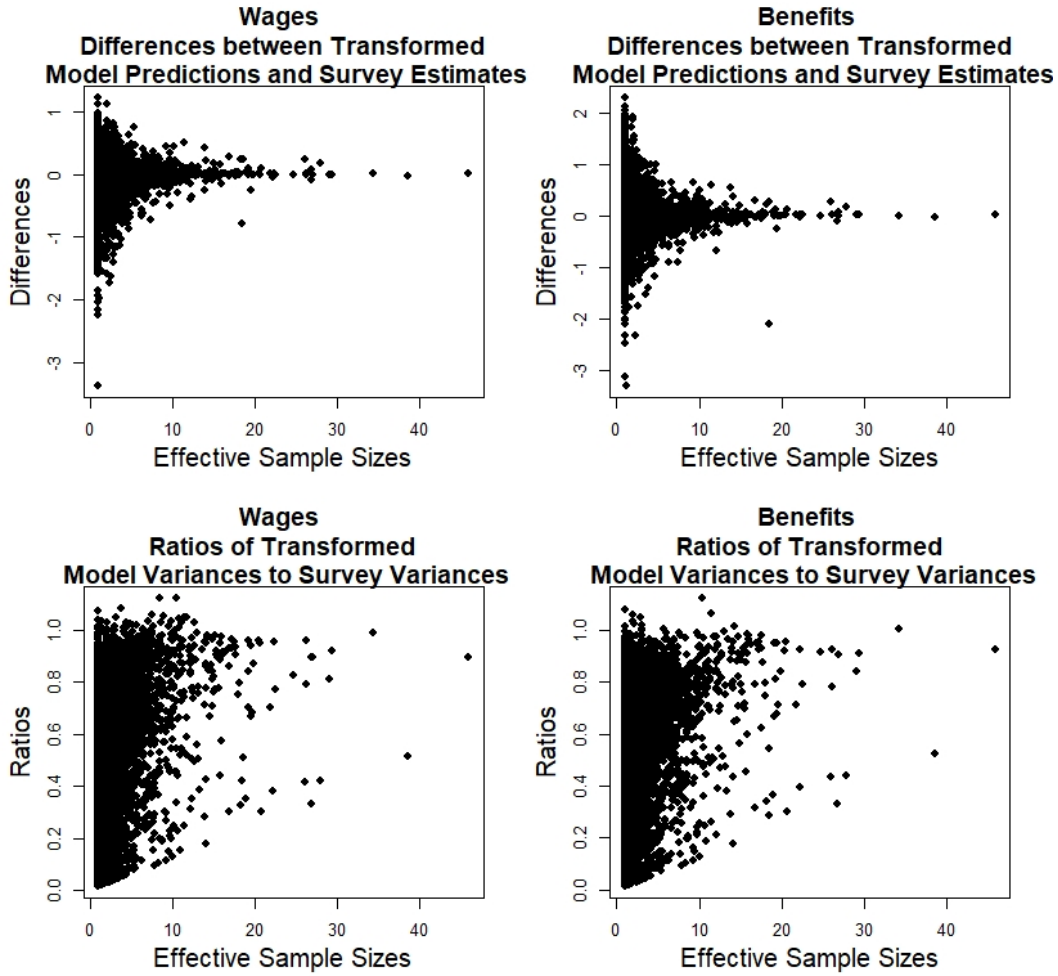

Figure 7: Relative measures of model predictions versus survey direct estimates, for wages and benefits; variance estimates

For domains where sample data are not available, direct comparisons of model and survey estimates can not be conducted because survey estimates can not be constructed. Therefore, we consider a novel visual approach for comparing model predictions for the full prediction space of domains to the survey estimates for the set of domains with sample data. Figures 8, 9, 10, and 11 depict the distributions of the model predictions and survey estimates of wage, point estimates and standard errors, benefits, point estimates and standard errors, respectively: one curve (in blue) corresponds to the distribution of the survey estimates for the set of domains with sample data; one curve (in red) corresponds to the distribution of the model predictions for the set of domains with sample data; and one curve (in black) corresponds to the distribution of the model predictions for the set of do-

mains without sample data. The range of the estimates is truncated to improve visualization (longer right tails result in difficult to see curves).

To reiterate, the results in Figures 8, 9, 10, and 11 are not meant for one-to-one comparisons across the different sets of estimates/predictions but rather for comparison of ranges of estimates/predictions. For example, it is noticeable that the mode of the in-sample model predictions is shifted to the right, compared to the model of the survey estimates. This is an effect of shrinkage of the estimates toward the common trend, estimated by the model, especially shrinkage of estimates for domains with sample sizes or large sampling variances (after transformation, and revision). However, the one-to-one comparison between these two sets of estimates is better quantified in Figures 6 and 7. The multiple modes in the distribution of the model predictions for not-in-sample domains correspond to the same mean structure, the $x_i\hat{\beta}$ component, due to similar job characteristics encountered in the set of covariates selected for modeling and prediction. Moreover, the range of the model predictions, for both in-sample and not-in-sample data, are comparable to the range of the survey estimates.

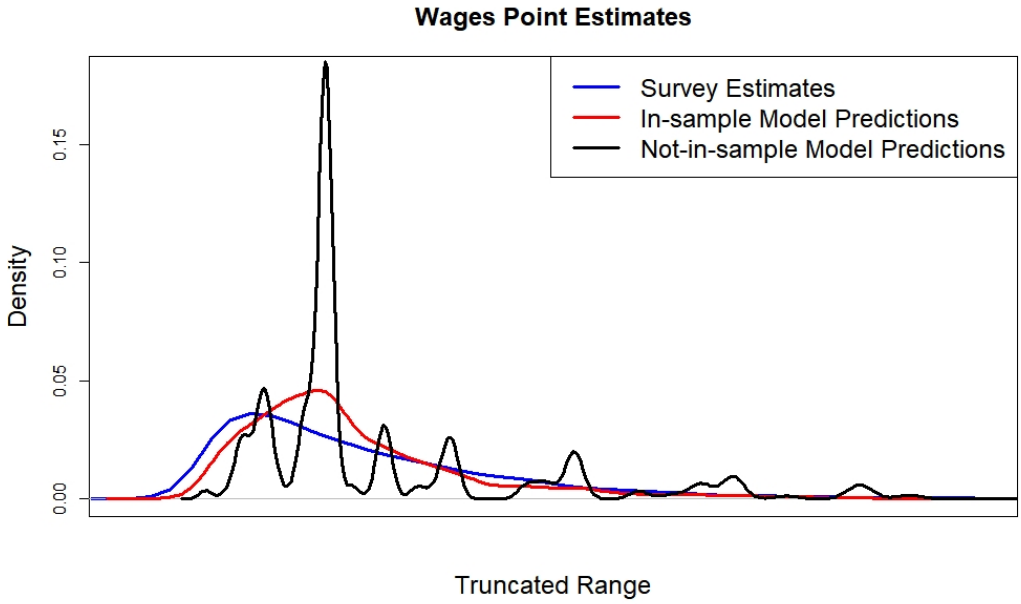

Figure 8: Distributions of the model predictions and survey estimates of wage; point estimates

5. Discussion

Various practical challenges were encounter when applying the methodology developed to the application data. First, preparing the model input data was an intensive process, including data manipulation, construction of direct survey estimates, data investigation to assess the quality of the direct survey estimates and development of methods to revise such survey summaries and assess relationships between quantities of interest and necessary transformations. Another important challenge was in defining the prediction space, effort that is still ongoing. We note that the prediction space is specific to the application, for example, certain job characteristics do not apply for some of the six-digit SOC system codes, such as unleveled jobs.

The next set of practical challenges relate to variable selection. For the application we

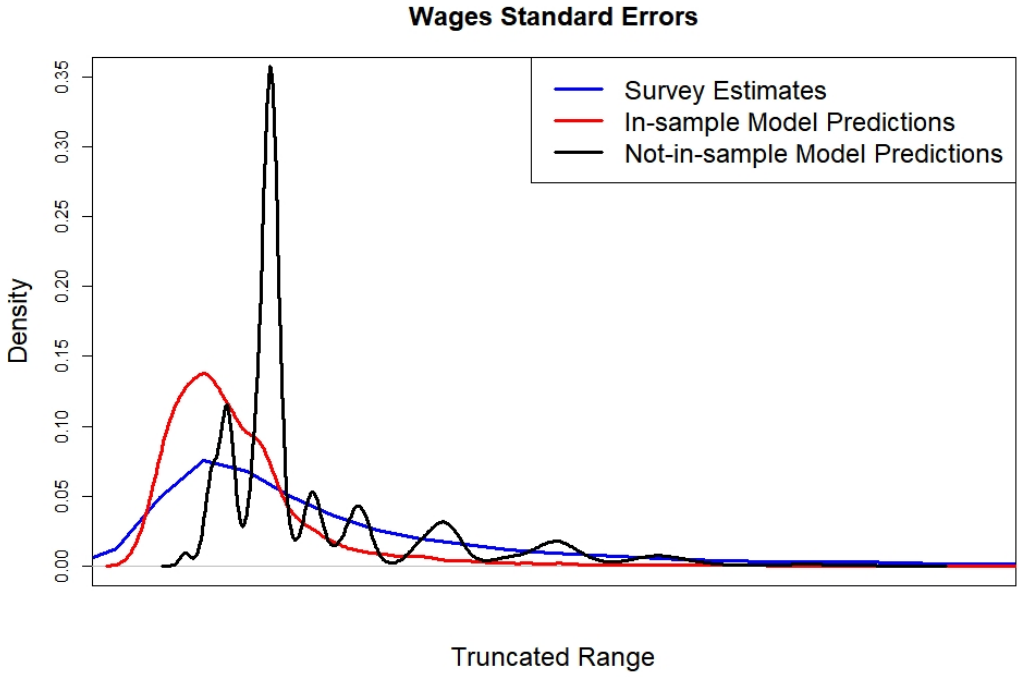

Figure 9: Distributions of the model predictions and survey estimates of wage; standard errors

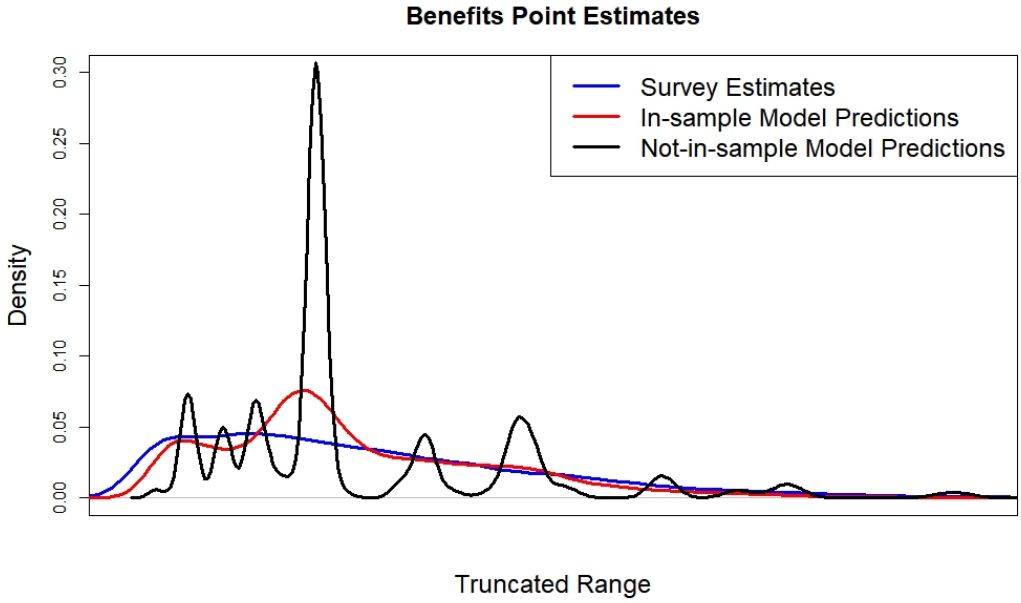

Figure 10: Distributions of the model predictions and survey estimates of benefits; point estimates

considered, there was a set of six categorical variables, each with a different number of categories. Considering all these variables and their multiple interactions was not possible due to computational difficulties. Therefore, we started the variable selection effort with the full set of variables and their two-way interactions only. The final set of variables we

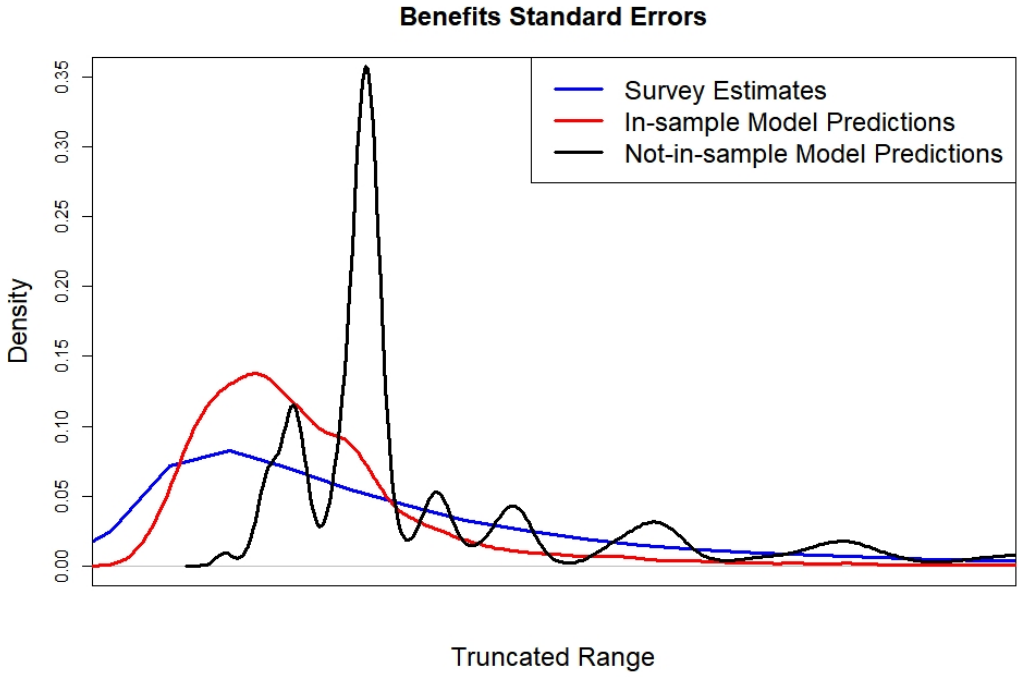

Figure 11: Distributions of the model predictions and survey estimates of benefits; standard errors

selected to construct the model matrix was very small, relative with the full initial set, and the addition of any other variable resulted in an increase in computational time or a limitation in computation capability (e.g., it was not possible to keep union of the sets of variables selected by different methods).

Other software and computational time challenges relate to the use of R STAN and R JAGS for model fit. There are limitations in model specifications for R JAGS. For example, the default priors for the linking model variance-covariance matrix in R STAN are LKJ-type priors (Lewandowski, Kurowicka, and Joe, 2009), specification not straightforwardly available in R JAGS. On the other hand, for these types of models, the computation time for R STAN is significantly longer than for R JAGS (see Erciulescu, 2019). The models are expensive to fit because of the multivariate nature (modeling multiple response variables), the multiple nature (using a large number of covariates), the large number of domains considered for model fit (over 16,000), the inference method considered (Bayes and MCMC). The results in this paper are based on models fit in R JAGS.

Moreover, the prediction for not-in-sample domains has limitations due to storage (multi-dimensional array objects corresponding to posterior samples for model parameters) and is computationally expensive due to the time necessary for predictions (new samples are generated to construct the posterior predictive distribution of $\theta_{i'}$). Therefore, increasing the number of MCMC samples, considering other model diagnostics, such as autocorrelation plots, or adopting alternative priors for the model variance-covariance matrix, would result in a computational cost.

We developed alternative methodology for fit and prediction using other model specifications, such as LKJ-type prior distributions for the linking model variance-covariance matrix or univariate models fit independently to wage and benefits. However, due to computing resources limitations, we were only able to work with small subsets of the in-sample prediction space (few hundreds of domains) and different subsets of covariates. Also, we

compared the posterior means for the model parameters to the initial values provided to the MCMC and decided to further investigate an empirical Bayes estimation approach, treating the regression parameters except the intercept as fixed. As a result, the computational time decreased, at the cost of lower quality convergence and mixing diagnostics.

In summary, we modeled employee compensation components for domains defined by the crossing of census division, six-digit SOC system codes, work levels and characteristics. In future investigations, we would like to explore even finer domains (e.g., with geography defined by states, counties, or metropolitan areas). In addition, we developed bivariate small domain models for wage and benefits compensation but would like to explore multivariate models for wage and non-wage (more than just benefits) compensation subcategories. Alternative computation methods are of interest and exploration of other sources that help refine the prediction space remains of interest, too.

REFERENCES

Arora, V., and Lahiri, P. (1997). "On the superiority of the Bayesian method over the BLUP in small area estimation problems." *Statistica Sinica*, 7, 1053-63.

Battese, G. E., Harter R. M. , and Fuller W. A. (1988). "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association*, 83(401). [American Statistical Association, Taylor Francis, Ltd.], 28-36. http://www.jstor.org/stable/2288915.

Bell, W. R., Basel W. W. , and Maples J. J. (2016). "An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program." *Analysis of Poverty Data by Small Area Estimation*, M. Pratesi (Ed.). West Sussex: Wiley Sons, Inc., 349-77.

Bureau of Labor Statistics (2018). "National Compensation Measures Handbook." Accessed December 7. https://www.bls.gov/opub/hom/ncs/home.htm.

Casas Cordero Valencia, C., Encina J., and Lahiri P. (2016). "Poverty Mapping in Chilean Comunas." Analysis of Poverty Data by Small Area Estimation, M. Pratesi (Ed.). West Sussex: Wiley Sons, Inc., 379-403.

Chatterjee, S., Lahiri P., and Li, H. (2008). "On Small Area Prediction Interval Problems." *Annals of Statistics*, 36, 1221-45.

Cox, D. R. (1975). "Prediction Intervals and Empirical Bayes Confidence Intervals." *Journal of Applied Probability*, 12(S1). Cambridge University Press, 47-55. doi:10.1017/S0021900200047550.

Datta, G. S., Fay, R. E., and Ghosh, M. (1991). "Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation." *Proceeding of the US Census Bureau 1991 Annual Research Conference*, U.S. Census Bureau, Washington, DC, 63-79.

Datta, G. S., Ghosh M., Nangia N., and Natarajan, K. (1996). "Estimation of Median Income of Four-Person Families: A Bayesian Approach." In *Bayesian Analysis in Statistics and Econometrics*, Eds.: D.A. Berry, K.M. Chaloner and J.K. Geweke. New York: Wiley, 6129-6140.

Datta, G. S., Day B., and Basawa, I.V. (1999). "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation." *Journal of Statistical Planning and Inference*, 75, 269-79.

Datta, G. S., Day B., and Maiti, T. (1998). "Multivariate Bayesian Small Area Estimation: An Application to Survey and Satellite Data." *Sankhya A*, 60, 344-62.

Datta, G. S., Ghosh M., Smith D. D., and Lahiri, P. (2002). "On an Asymptotic Theory of Conditional and Unconditional Coverage Probabilities of Empirical Bayes Confidence Intervals." *Scandinavian Journal of Statistics*, 29, 139-52.

Datta, G. S., Rao, J. K., and Smith, D. D. (2005). "On Measuring the Variability of Small Area Estimators Under a Basic Area Level Model." *Biometrika*, 92, 1, 183-96.

Datta, G., and Ghosh, M. (2012). "Small Area Shrinkage Estimation." *Statistical Science*, 27(1), 95-114. doi:10.1214/11-STS374.

Diao, L., Smith, D. D., Datta, G. S., Maiti, T., and Opsomer, J. D. (2014). "Small Area Shrinkage Estimation." *Scandinavian Journal of Statistics*, 41, 497-515. doi:10.1111/sjos.12045.

Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). "Micro-Level Estimation of Poverty and Inequality." *Econometrica*, 71(1), 355-64. http://www.jstor.org/stable/3082050.

Erciulescu, A. L. (2018). "Transparent and Reproducible Research in Agricultural Official Statistics." Government Advances in Statistical Programming Workshop. http://washstat.org/presentations/20181024/ Erciulescu.pdf.

Erciulescu, A. L. (2019). "Hierarchical models in the production of official statistics: a discussion of some practical aspects." Government Advances in Statistical Programming Workshop. To appear.

Erciulescu, A. L., and Fuller, W. A. (2018). "Bootstrap Confidence Intervals for Small Area Proportions."

Journal of Survey Statistics and Methodology. doi:10.1093/jssam/smy014.

Erciulescu, A. L., Berg E. J., Cecere W., and Ghosh, M. (2018). "A Bivariate Hierarchical Bayesian Model for Estimating Cropland Cash Rental Rates at the County Level." *Survey Methodology, Statistics Canada, Catalogue No. 12-001-X*, 45(2). https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00001-eng.htm.

Erciulescu, A. L., Cruze N. B., and Nandram, B. (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:10.1111/rssa.12390.

Fay, R. (1987). "Application of Multivariate Regression to Small Domain Estimation." R. Platek, J.N.K. Rao, C.-E. Sarndall, and M.P. Singh (Eds.), *Small Area Statistics*, New York: John Wiley Sons, Inc., 91-201.

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). Taylor Francis, 269-77. doi:10.1080/01621459.1979.10482505.

Fay, R.E., Planty M., and Diallo, M.S. (2013). "Small Area Estimates from the National Crime Victimization Survey." JSM Proceedings, Survey Research Methods Section, 1544-57.

Fuller, W. A., and Harter, R. (1987). "The multivariate components of variance model for small area estimation." R. Platek, J.N.K. Rao, C.-E. Sarndall, and M.P. Singh (Eds.), *Small Area Statistics*, New York: John Wiley Sons, Inc., 103-123.

Gelman, A., J.N. Carlin, H. S. Stern, D. B. Dunson, A. Vehatri, and D. B. Rubin. 2013. *Bayesian Data Analysis*. 3rd edition. Chapman Hall/CRC.

Ghosh, M., and Rao, J. N. K. (1994). "Small Area Estimation: An Appraisal (with Discussion)." *Statistical Science*, 9, 55-93.

Gonzalez, M. E. (1973). "Use and Evaluation of Synthetic Estimators." Proceedings of the Social Statistics Section, American Statistical Association, Washington DC, 33-36.

Gonzalez-Manteiga, W., Lobardia, M. J., Molina, I., Morales, D., and Santamaria, L. (2008). "Analytic and Bootstrap Approximations of Prediction Errors Under a Multivariate Fay-Herriot Model." *Computational Statistics and Data Analysis*, 52, 5242-52.

Guciardo, C. J. 2019. "Estimating Variance in the National Compensation Survey, Using Balanced Repeated Replication." Accessed April 26. https://www.bls.gov/osmr/pdf/st010110.pdf.

Hall, P., and Maiti, T. (2006). "On Parametric Bootstrap Methods for Small-Area Prediction." *Journal of Royal Statistical Society, Series B*, 68, 221-38.

Higham, N. J. (1988). "Computing a Nearest Symmetric Positive Semidefinite Matrix." *Linear Algebra and Its Applications*, 103, 103-118.

Jiang, J., and Lahiri, P. (2006). "Mixed Model Prediction and Small Area Estimation." *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 15(1), 1-96. https://EconPapers.repec.org/RePEc:spr:testjl:v:15:y:2006:i:1:p:1-96.

Judkins, D. R. (1990). "Fay's method for variance estimation." *Journal of Official Statistics*, 6(3), 223-239.

Krenzke T., Mohadjer L., Li J., Erciulescu A. L., Fay R., Ren W., Van de Kreckhove W., Li L., Rao J. N. K. (forthcoming) "Program for the International Assessment of Adult Competencies. State and County Indirect Estimation Methodology." *United States Department of Education*. NCES2019012.

Lettau, M. K., and Zamora, D. A. (2013). "Wage estimates by job characteristic: NCS and OES program data." *Monthly Labor Review*, U.S. Bureau of Labor Statistics, August. https://doi.org/10.21916/mlr.2013.27.

Lewandowski, D., Kurowicka, D., Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method." *Journal of Multivariate Analysis*, 100, 9, 1989-2001. doi: 10.1016/j.jmva.2009.04.008.

Molina, I., Nandram B., and Rao, J. N. K. (2014). "Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach." *Annals of Applied Statistics*, 8(2), 852-85. doi:10.1214/13-AOAS702.

Pfeffermann, D. (2013). "New Important Developments in Small Area Estimation." *Statistical Science*, 28(1). The Institute of Mathematical Statistics, 40-68. doi:10.1214/12-STS395.

Prasad, N. G. N., and Rao, J. N. K. (1990). "The Estimation of the Mean Squared Error of Small Area Estimators." *Journal of the American Statistical Association*, 85, 163-71.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, 2nd ed. Hoboken, NJ: John Wiley Sons.

Rao, J.N.K., and Yu, M. (1994). "Small Area Estimation Combining Time Series and Cross-Sectional Data." *Canadian Journal of Statistics*, 22, 511-28.

Tzavidis, N., Zhang, L., Hernandez, A. L., Schmid, T., and Rojas-Perilla, N. (2018). "From Start to Finish: A Framework for the Production of Small Area Official Statistics." *Journal of the Royal Statistical Society A*, 181(4), 927-79. doi:10.1111/rssa.12364.

6. Appendix

The variables selection methods described in Section 3 resulted into a final set of 18 covariates. A descriptive list of these 18 covariates is provided in Table 6.

Table 6: Selected Covariates

Intersection of full-time/part-time variable, part-time category, with time/incentive variable, time category
Intersection of full-time/part-time variable, part-time category, with work level variable, level 10
Intersection of full-time/part-time variable, part-time category, with work level variable, level 4
Intersection of six-digit SOC system code variable, code 112022 (Sales Managers), with work level variable, level 13
Main effect six-digit SOC system code variable, code 291151 (Registered Nurses)
Intersection of six-digit SOC system code variable, code 411012 (First-Line Supervisors/Managers of Non-Retail Sales Workers), with work level variable, level 9
Intersection of six-digit SOC system code variable, code 413021 (Insurance Sales Agents), with work level variable, level 9
Intersection of six-digit SOC system code variable, code 413031 (Securities, Commodities, and Financial Services Sales Agents), with work level variable, level 11
Intersection of six-digit SOC system code variable, code 413031 (Securities, Commodities, and Financial Services Sales Agents), with work level variable, level 9
Intersection of six-digit SOC system code variable, code 472111 (Electricians), with union/nonunion variable, union category
Intersection of union/nonunion variable, union category, with time/incentive variable, time category
Intersection of union/nonunion variable, union category, with work level variable, level 3
Intersection of union/nonunion variable, union category, with work level variable, level 4
Main effect work level variable, level 11
Main effect work level variable, level 12
Main effect work level variable, level 13
Main effect work level variable, level 2
Main effect work level variable, level 9