

Small area co-modeling of point estimates and their variances for domains in the Current Employment Statistics Survey

Julie Gershunskaya and Terrance D. Savitsky¹

U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Washington, DC 20212

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract: Every month, the Bureau of Labor Statistics publishes estimates of employment from the Current Employment Statistics survey at the state and national total levels, as well as at various detailed levels by industry and geography. For smaller domains, where the direct sample-based estimates are not reliable, estimates are produced using models. We adopt a Bayesian approach and consider the area level Fay-Herriot model along with several alternatives that: (i) co-model the variances of the direct estimators instead of adhering to the traditional assumption of the “fixed and known” variances; (ii) accounts for possible deviations from the normality assumption of the random effects by assuming a mixture of the normal distributions. Models are compared based on the direct estimates and variances from the Current Employment Statistics survey, as well as using a simulation study. We further propose a model-based method of screening that could become a useful tool for analyst’s review of the estimates before they are released for publication.

Key Words: Bayesian Hierarchical Modeling, Dirichlet process, Fay-Herriot, Variational Bayes, Stan

1. Introduction

Large government surveys, such as the Current Employment Statistics survey considered in this paper, are designed to produce high quality sample-based estimates for a number of state and national industrial levels. More detailed geographical and industrial domains often contain a small number of sample units (e.g., business establishments). Direct sample-based estimates at these detailed levels are not reliable, and models are used to improve the quality of the estimates. One of the most popular models is the classical Fay-Herriot model by (Fay and Herriot 1979). The Fay-Herriot (FH) model yields an estimator that can be conveniently presented in the form of a weighted average of the direct sample-based estimator and a so-called “synthetic” component. Both the synthetic component and the mixture weights depend on specific distributional assumptions. Direct sample-based estimates are used as the data input in the FH modeling. In the classical FH model, variances of the direct sample estimates are assumed to be fixed and known. In reality, these variances are not known and some estimated variances are plugged in as if they were true variances; for example, direct sample based estimates of variances could be used for this purpose. However, such sample-based estimates of variances contain noise; and so, the usual practice, is to smooth the noise by using model-based estimates of variances extracted from a generalized variance function (GVF). Such GVF-based variances are implemented in a separate model from that for the direct point estimates.

Maiti et al. (2014) showed that co-modeling of direct point estimates and their variances in the same model may improve estimates of both quantities as it would exploit the relationship between the point estimates and their variances. Maiti et al. (2014) proposed a solution within the frequentist paradigm, and Sugasawa et al. (2017) considered a Bayesian approach. In this paper, we extend Sugasawa et al. (2017) to include nonparametric probabilistic clustering and apply it to estimates from the Current Employment Statistics (CES) survey. Our clustering formulation relaxes the assumption of normality of the random effects in the models for both the direct point estimates and the variances as a means of addressing deviations in employment from linearity assumption among industry domains. Employment may grow or decrease faster in some groups of domains included in the model. This phenomena may be described by imposing the mixture of normal distributions assumption on the random effects.

¹ Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

The models considered may still fail to describe true population target in domains having large deviations from the model linearity assumption. We adapt a posterior predictive checking approach to uncover domains that are not well described by a model. We identify such domains using a Bayesian multiple hypotheses testing approach. Each domain’s probability of not being generated by the target model is considered in conjunction with the overall false discovery rate (FDR) (Benjamini and Hochberg 1995), to identify a relatively small number of “suspected” domains whose estimates are posited as not having been generated by our joint model. The list of these domains may be sent to analysts for review. Analysts may conclude that the deviation is due to a few outlying units used in deriving the domain level estimates; otherwise, analysts may decide that a particular domain’s deviation from the linear model expresses real economic movement. The approach provides a potentially useful tool for analysts if applied for screening the estimates before publication.

We compare the performances of alternative models using the CES data; we also study the model behavior for several scenarios using Monte Carlo simulations. Our simulation results confirm that co-modeling of the direct point estimates and their variances leads to improved estimates. In addition, the model estimates of variances for the direct estimator can be considered as a useful by-product of the modeling efforts.

We adopt the hierarchical Bayesian paradigm for development of the models. The code is implemented in the Stan modeling language (Gelman et al. 2015) using a Variational Bayes algorithm (Kucukelbir et al. 2017) implemented in RStan V2.15.1 package, which is the R interface for the Stan modeling language (Gelman et al. 2015, Stan Development Team 2017), to implement our models.

The paper is organized as follows. In Section 2 we provide brief introduction of estimation procedures and the form of the sample-based estimator used in CES. The models considered in this paper are stated in Section 3. In Section 4, we discuss the results of application of the models to the real CES data and to the synthetic data generated by adding noise to the true historical series. In Section 5, we conduct a robustness study of our candidate model formulations to assess their performances under deviations from linearity. Section 6 introduces additional uses for our models with large-sized domains where modeling is not traditionally performed because the direct estimates are published; in particular, we introduce a model-based screening procedure to identify a set of domains whose direct sample-based estimates are not adequately described by the model. We also assess whether raw variances may be replaced with modeled estimates to provide improved measures of uncertainty, even for larger-sized domains. We conclude with a summary discussion in Section 7.

2. CES Data Construction

Estimates of employment from the CES survey are published every month for various industries (defined by the North American Industry Classification System (NAICS)) at the national level, as well as for State and Metropolitan Statistical Area levels.

The focus of this paper is construction of an area-level model for CES domains defined by intersections of industry and geography. (In our application discussed in Section 4, our geographic resolution is the State level). Since the direct survey estimates are used as input data in the proposed area-level models, we start by briefly describing relevant details pertaining to construction of the CES estimator.

For a given month, t , the target of the CES estimation is the change in employment from the previous to current month. Consider a set of (geography-by-industry) domains, $i = 1, \dots, N$. The population ratio, $R_{i,t}$, is the target employment change, defined as

$$R_{i,t} = \frac{Y_{i,t}}{Y_{i,t-1}}, \tag{1}$$

where $Y_{i,t}$ is the employment level in domain i at month t .

The estimated relative change in employment level $\hat{R}_{i,t}$ can be described as an adjusted sample based estimator of the relative change

$$\hat{r}_{i,t} = \frac{\sum_{j \in S_t^{(i)}} w_j y_{jt}}{\sum_{j \in S_t^{(i)}} w_j y_{j,t-1}}, \quad (2)$$

where y_{jt} is the employment of business j at time t , w_j is the sampling weight of unit j , and $S_t^{(i)}$ is a set of units sampled in domain i that provide non-zero employment inputs in both previous and current months as a “matched” set of respondents. The presence of matched sets of sampled units is typically high from one month to another but there are also unmatched units; thus, there is an adjustment to $\hat{r}_{i,t}$, yielding estimator $\hat{R}_{i,t}$ of $R_{i,t}$. The adjustment is described in some detail, for example, in Gershunskaya and Savitsky (2017) and is omitted here for brevity. In what follows, we assume $\hat{R}_{i,t}$ to be an unbiased estimator of target, $R_{i,t}$.

Every year, the estimation cycle starts at month 0 from a known employment level $Y_{i,0}$ and after twelve months the CES estimated employment level $\hat{Y}_{i,12}$ is compared to the census data, maintained by BLS’ Quarterly Census of Employment and Wages (QCEW) program. The QCEW data become available with a lag of about 6 to 9 months, while the CES estimates provide timely snapshot of the economy on a monthly basis. Once a year, the CES estimated levels are revised to reflect newly available QCEW levels (in a procedure commonly known as the annual revision), and a new cycle of estimation starts with the new true census level as the new month 0. Employment seasonal patterns in the QCEW are affected by the quarterly submission of administrative data provided by units (business establishments). CES estimates are unaffected by this quarterly seasonal influence due to a monthly submission cycle. So we may not compare monthly QCEW and CES estimates. Nevertheless, the annual levels from QCEW are considered a “gold standard” and the quality of the CES employment estimates of levels are judged based on the size of the annual revision that benchmark to the QCEW.

To summarize, monthly ratios $\hat{R}_{i,t}$, along with their respective sampling variances $v_{i,t}$, constitute the domain-level data supplied for the modeling. In order to compare estimates with the QCEW gold standard, we produce estimated employment level after 12 months of estimation as

$$\hat{Y}_{i,12} = Y_{i,0} \prod_{\tau=1}^{12} \hat{R}_{i,\tau}, \quad (3)$$

where $Y_{i,0}$ is a known “benchmark” employment level at month 0, available from QCEW. We compute the analogous levels using model fitted ratios $\tilde{R}_{i,t}$ and use them to compare the models.

Figure 1 presents a plot of the estimation cycle. It shows monthly estimated levels for one of the CES domains. The lines on the plot correspond to alternative (model-based) estimates considered in the paper. The black line with solid circles is the target QCEW line. The goal is to be closer to the QCEW line at the 12th month of the cycle. Direct sample-based estimates in small domains may be appreciably volatile. Model-based estimates usually present various degree of smoothness compared to the direct estimates, as exemplified in Figure 1.

The domain-level auxiliary information used in the models is the employment ratio, obtained as a forecast from the historical QCEW series. For this paper, we used five-year averages of historic QCEW ratios $R_{i,t}^H = \sum_{m=1}^5 R_{i,(t-12m)} / 5$ as auxiliary information for the point estimate part of the models.

It is possible to extend the cross-sectional models by including multiple months and thus using both cross-sectional and time information (Gershunskaya and Savitsky 2017.) However, in this paper, we concentrate on the one-month-at-a-time estimation of monthly ratios, which historically express little month-over-month correlation after transformation to ratios.

3. Description of the models

3.1 Models for point estimates with known variances

We start with the classical Fay-Herriot (FH) model (Fay and Herriot 1979.) Let y_i be a survey estimate of target parameter θ_i for domain i . For each domain, $i = 1, \dots, N$, assume

$$y_i | \theta_i \stackrel{ind}{\sim} N(\theta_i, v_i), \quad (4)$$

$$\theta_i | \mu, \boldsymbol{\beta}, \tau_u^2 \stackrel{ind}{\sim} N(\mu + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2). \quad (5)$$

Sample estimated y_i 's are assumed to be normally distributed and unbiased for target parameter θ_i , with variances v_i that are treated as known (equation (4)). Equation (5) links true signal θ_i to a vector of covariates \mathbf{x}_i via the linear regression by assuming the normally distributed deviation of the true signal from “synthetic” part $\mu + \mathbf{x}_i^T \boldsymbol{\beta}$ (to facilitate the ensuing description, we explicitly write the intercept term μ .)

The normality assumption used in (5) may not be realistic. For example, if a single or a handful of domains deviate significantly from $\mathbf{x}_i^T \boldsymbol{\beta}$, this assumption of the Fay-Herriot model would result in the under-shrinkage of the bulk of the observations.

We introduce a new model, referred to as CFH, by relaxing the normality assumption in (5). We replace the normal distribution with a finite mixture normal distributions. Specifically, we assume the existence of K latent clusters having cluster specific intercepts μ_k , for $k = 1, \dots, K$, and common variance τ_u^2 :

$$\theta_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k N(\mu_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2) \quad (6)$$

The FH and CFH models are summarized in Table 1 (formulated for a single covariate x_i , for simplicity.) The CFH model is designed to allow for deviations from linearity assumption $\mu + \mathbf{x}_i^T \boldsymbol{\beta}$ for some subsets of domains. The form for the Dirichlet prior, with hyperparameters set to α/K , induces a Dirichlet process (DP) mixture formulation in the limit of the maximum number of allowable mixture components, K (see Neal 2000). The larger

is α , the more of the K possible mixture components (also referred to as clusters) will have $\pi_k \neq 0$, so a further gamma prior is imposed to allow the data to learn the number of mixture components.

Table 1. FH and CFH models

FH	CFH
$y_i \theta_i \overset{ind}{\sim} N(\theta_i, v_i)$	$y_i \theta_i \overset{ind}{\sim} N(\theta_i, v_i)$
$\theta_i \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N(\boldsymbol{\mu} + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2)$	$\theta_i \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \overset{iid}{\sim} \sum_{k=1}^K \pi_k N(\boldsymbol{\mu}_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2)$
$\boldsymbol{\beta} \lambda_\beta \sim N(\mathbf{0}, \lambda_\beta^{-1})$	$\boldsymbol{\pi} \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha}/K, \dots, \boldsymbol{\alpha}/K)$
$\boldsymbol{\mu} \lambda_\mu \sim N(\mathbf{0}, \lambda_\mu^{-1})$	$\boldsymbol{\beta} \lambda_\beta \sim N(\mathbf{0}, \lambda_\beta^{-1})$
$\tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1, 1)$	$\boldsymbol{\mu}_k \lambda_\mu \overset{iid}{\sim} N(\mathbf{0}, \lambda_\mu^{-1})$
	$\boldsymbol{\alpha}, \tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1, 1)$

As noted, sampling variances v_i in models FH and CFH of Table 1 are considered fixed and known. In practice, estimates of true variances are used. In our survey application and in the simulation study, we consider two possibilities for the treatment of v_i in models FH and CFH: 1) using direct sample based estimates of true variances (FH-BRR and CFH-BRR), which treats the variances as fixed and known and 2) using a smoothed estimator of variances. For these models (referred to as FH-V and CFH-V), the estimation of the variances are performed, separately, in a first step and then used as plug-in estimators for v_i in estimations of FH and CFH. The first step of the variance estimation is based on the same set of covariates as used in the models described below. Note that this approach ignores any uncertainty in the estimation of the variances, and so, is not a fully Bayesian approach (though we estimate the variance portion of FH-V and CFH-V under a Bayesian construction).

3.2 Co-modeling of sampling variances and point estimates (FHS)

Rather than fixing the variances at the estimated value, v_i , we view direct sample-based estimates of variances as data and model them together with the vector of point estimates y_i in a fully Bayesian model specification.

Table 2 contains a summary of the three models that co-model point estimates and estimates of their variances.

The first model (referred to as FHS) is slight modification of a model considered by Sugawara et al. (2017). Assume the following model holds for pair of direct survey estimates (y_i, v_i) for each domain i :

$$y_i | \theta_i, \sigma_i^2 \overset{ind}{\sim} N(\theta_i, \sigma_i^2), \quad (7)$$

$$\theta_i | \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N(\boldsymbol{\mu} + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2). \quad (8)$$

$$v_i | a, \sigma_i^2 \stackrel{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right), \quad (9)$$

$$\sigma_i^2 | b, \boldsymbol{\gamma} \stackrel{ind}{\sim} IG(2, b \exp(\mathbf{z}_i^T \boldsymbol{\gamma})). \quad (10)$$

Table 2. FHS, CFHS, and CFHSc models: joint modeling point estimates and estimates of their variances

FHS	CFHS	CFHSc
$y_i \theta_i, \sigma_i^2 \stackrel{ind}{\sim} N(\theta_i, \sigma_i^2)$	$y_i \theta_i, \sigma_i^2 \stackrel{ind}{\sim} N(\theta_i, \sigma_i^2)$	$y_i \theta_i, \sigma_i^2 \stackrel{ind}{\sim} N(\theta_i, \sigma_i^2)$
$\theta_i \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \stackrel{ind}{\sim} N(\boldsymbol{\mu} + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2)$	$\theta_i \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k N(\boldsymbol{\mu}_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2)$	$\theta_i \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k N(\boldsymbol{\mu}_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2)$
$v_i a, \sigma_i^2 \stackrel{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right)$	$v_i a, \sigma_i^2 \stackrel{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right)$	$v_i a, \sigma_i^2 \stackrel{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right)$
$\sigma_i^2 b, \boldsymbol{\gamma} \stackrel{ind}{\sim} IG(2, b e^{\mathbf{z}_i^T \boldsymbol{\gamma}})$	$\sigma_i^2 b, \boldsymbol{\gamma} \stackrel{ind}{\sim} IG(2, b e^{\mathbf{z}_i^T \boldsymbol{\gamma}})$	$\sigma_i^2 \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\pi} \stackrel{ind}{\sim} \sum_{k=1}^K \pi_k IG(2, b_k e^{\mathbf{z}_i^T \boldsymbol{\gamma}})$
	$\boldsymbol{\pi} \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha}/K, \dots, \boldsymbol{\alpha}/K)$	$\boldsymbol{\pi} \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha}/K, \dots, \boldsymbol{\alpha}/K)$
$\boldsymbol{\gamma} \boldsymbol{\Sigma} \sim \mathbf{N}_p(0, \boldsymbol{\Sigma})$	$\boldsymbol{\gamma} \boldsymbol{\Sigma} \sim \mathbf{N}_p(0, \boldsymbol{\Sigma})$	$\boldsymbol{\gamma} \boldsymbol{\Sigma} \sim \mathbf{N}_p(0, \boldsymbol{\Sigma})$
$\boldsymbol{\beta} \lambda_\beta \sim N(0, \lambda_\beta^{-1})$	$\boldsymbol{\beta} \lambda_\beta \sim N(0, \lambda_\beta^{-1})$	$\boldsymbol{\beta} \lambda_\beta \sim N(0, \lambda_\beta^{-1})$
$\boldsymbol{\mu} \lambda_\mu \sim N(0, \lambda_\mu^{-1})$	$\boldsymbol{\mu}_k \lambda_\mu \stackrel{iid}{\sim} N(0, \lambda_\mu^{-1})$	$\boldsymbol{\mu}_k \lambda_\mu \stackrel{iid}{\sim} N(0, \lambda_\mu^{-1})$
$\tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1, 1),$	$\boldsymbol{\alpha}, \tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1, 1),$	$\boldsymbol{\alpha}, \tau_u^{-2}, \lambda_\beta, \lambda_\mu, b_k \sim G(1, 1),$
$\log(a) \sim t_3(0, 1),$	$\log(a) \sim t_3(0, 1),$	$\log(a) \sim t_3(0, 1),$
$\log(b) \sim t_3(0, 1),$	$\log(b) \sim t_3(0, 1),$	$\log(b) \sim t_3(0, 1),$
$prior(\boldsymbol{\Sigma})$	$prior(\boldsymbol{\Sigma})$	$prior(\boldsymbol{\Sigma})$

Lines (7)-(8) are the usual FH assumptions on the point estimates y_i and lines (9)-(10) describe the variance model, where parameter σ_i^2 is the true sampling variance; \mathbf{z}_i is a vector of covariates for the variance model for area i ; $a, b, \boldsymbol{\gamma}$ are the model parameters. Note that in equation (9), estimated variances depend on the sample size n_i , where for a set of domains with unequal number of respondents, we use standardized response size,

$$n_i^* = \left(n_i - \left\{ \min_i n_i - 1 \right\} \right) / \left(\max_i n_i - \min_i n_i \right) \in [0, 1].$$

Our assumption is slightly different from Maiti et al. (2014) and Sugawara et al. (2017) as we allow additional (unknown) parameter, a , to regulate the scale and shape of the distribution. In our application, we found that for moderate sample sizes, using sample size alone would result in predicted variances that are similar to direct estimates of variances.

Model CFHS, in the second column of Table 2, is the analogue of the clustering model CFH. It is described by replacing (8) of the FHS model with the finite mixture (that contracts on a DP mixture), as in (6).

Table 3. Summary of models FH, CFH, FHS, CFHS, CFHSc

	Level	FH	CFH	FHS	CFHS	CFHSc
Point Estimate	<i>Sampling model</i>	normal	normal	normal	normal	normal
	<i>Linking model</i>	normal	normal mixture	normal	normal mixture	normal mixture
Sampling Variance	<i>Sampling model</i>			gamma	gamma	gamma
	<i>Linking model</i>			inverse gamma	inverse gamma	inverse gamma mixture

Finally, the inverse gamma assumption in (10) is relaxed in the CFHSc construction by specifying a mixture of the inverse gamma distributions with the cluster-specific shape parameter b_k :

$$\sigma_i^2 | \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\pi} \stackrel{ind}{\sim} \sum_{k=1}^K \pi_k IG\left(2, b_k \exp\left(z_i^T \boldsymbol{\gamma}\right)\right). \quad (11)$$

It is reasonable to suppose that point estimates and estimates of their variances are related, and we parameterize this assumption by assuming a common cluster structure for pairs, (μ_k, b_k) . That is, each mixture / cluster component in the joint distribution for (θ_i, σ_i^2) share the same π_k .

Table 3 provides a brief summary for the models introduced in this section for quick comparison and reference.

4. Model fit comparison

We next compare the model performances based on the CES data. The first part of this analysis is based on the actual historical data. We subsequently generate a synthetic dataset by adding noise to the census data, which allows us to compare additional properties of the models relevant to the specifics of CES.

4.1 Analysis based on the real CES data

The CES data used in this paper are defined for a set of $N = 2233$ sub-industry-by-State domains; the data series are based on September 2008 as the starting point. We chose this particular year estimation cycle because of the non-trivial employment pattern that occurred during the period of the “great recession”, which induced a marked shift in employment trends from previous years.

We fit separate models for sets of domains belonging to 17 major industries. Each set consists of different number of domains defined by subindustries and States.

The sample-based estimates of variances $\hat{V}_{i,t}$ of the point estimator $\hat{R}_{i,t}$ are computed using the balanced repeated replication (BRR) methodology.

Before fitting the models, we standardized the input values as follows: $y_{i,t} = (\hat{R}_{i,t} - \bar{R}) / \sqrt{\bar{V}}$,

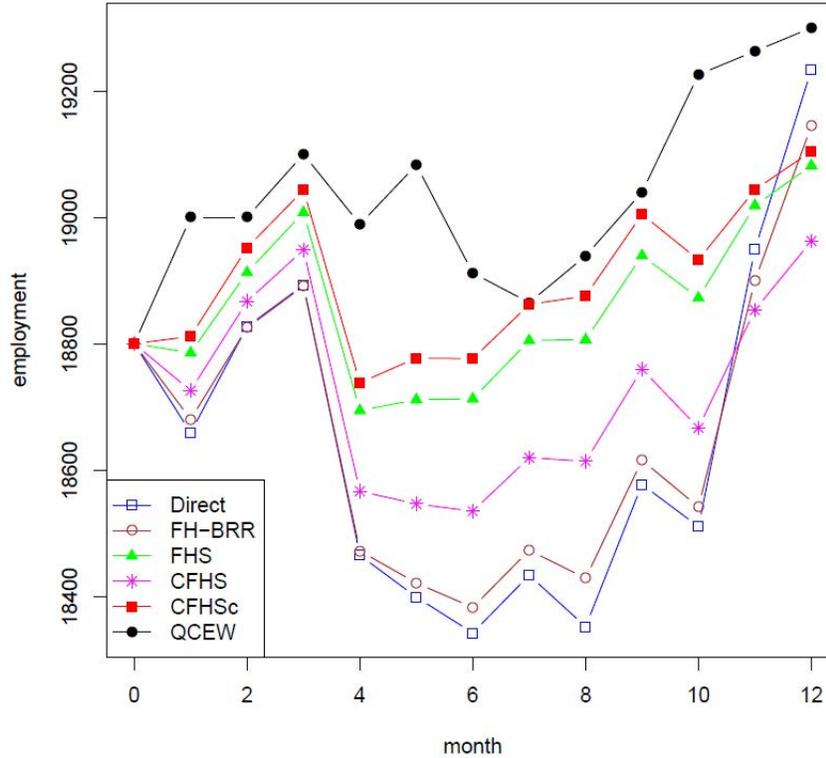
$x_{i,t} = (R_{i,t}^H - \bar{R}^H) / \sqrt{\bar{V}}$, and $v_{i,t} = \hat{V}_{i,t} / \bar{V}$, where \bar{R} , \bar{R}^H , and \bar{V} are the average values of respective original data $\hat{R}_{i,t}$, $R_{i,t}^H$, and $\hat{V}_{i,t}$. To obtain the model estimator, $\tilde{R}_{i,t}$, we perform a back transformation.

For the variance part of the model, the vector of covariates $z_{i,t}$ contains the following three components: the number of sample respondents $n_{i,t}$, the domain employment level $Y_{i,0}$ at month 0, and the estimated domain-specific fraction of employment not covered by the responding sample, $f_{i,t}$. Each covariate was log transformed and standardized before fitting the model, i.e., $z_{i,t}^p = (l_{i,t}^p - \bar{l}_t^p) / s_t^p$, where $l_{i,t}^p$ is the log transformed component $p = 1, 2, 3$; \bar{l}_t^p is the average and s_t^p is the standard error of $l_{i,t}^p$ computed over domains.

We obtain model estimates $\tilde{R}_{i,t}$ for relative monthly changes for each month over the 12-month estimation period. The estimates of employment levels at month t are obtained from the set of $\tilde{R}_{i,\tau}$, $\tau = 1, \dots, t$, as

$$\tilde{Y}_{i,t} = Y_{i,0} \prod_{\tau=1}^t \tilde{R}_{i,\tau}. \quad (12)$$

Figure 1: Domain #60 in Health Care and Social Assistance industry (average number of responding units in the domain is 16.6)



Due to different seasonality patterns between the employment series derived from the administrative QCEW data and CES, the most meaningful comparison of the two series is after 12 months of estimation. Results for each major industry and overall, presented in Table 4, are based on the mean absolute deviation (MAD):

$$MAD = N^{-1} \sum_{i=1}^N \left| \tilde{Y}_{i,12} - Y_{i,12} \right|, \quad (13)$$

where $Y_{i,12}$ comes from the (QCEW) census data and is used as “the gold standard” for the estimates.

The real data results (Table 4) show that the co-modeling of point estimates and variances leads to the estimates with smaller MAD than the estimates based on a variance that is plugged in the model as fixed and known parameter. This is true for the case of raw sample-based variances (FH-BRR and CFH-BRR) as well as for a smoothed version derived from a separate model for variances (FH-V and CFH-V).

Table 4. Real data results

Ind	N	Direct	FH-BRR	CFH-BRR	FH-V	CFH-V	FHS	CFHS	CFHSc
1000	50	792	774	757	789	784	840	777	817
2000	141	2152	1782	1783	1801	1825	1770	1760	1779
3100	234	1112	1072	1058	968	970	1075	1053	1081
3200	140	955	918	894	862	844	949	898	960
4100	124	1485	1234	1230	1215	1212	1201	1164	1179
4200	286	1439	1364	1365	1366	1340	1321	1305	1311
4300	194	1310	1023	1022	1044	1045	903	928	929
5000	83	1204	768	781	718	750	728	711	736
5500	149	1473	1031	1041	1047	1072	1051	1042	1060
6054	150	1450	1185	1193	1152	1177	1160	1145	1165
6055	45	1066	992	1005	917	950	892	975	937
6056	115	2344	1862	1876	2001	2040	2034	1853	1955
6561	59	1901	1779	1768	1538	1581	1537	1614	1593
6562	214	1551	1226	1244	1242	1247	1178	1225	1192
7071	59	2047	1431	1421	1208	1352	1136	1243	257
7072	80	1912	1819	1708	1641	1665	1736	1689	1733
8000	110	1773	1211	1250	1175	1219	1097	1186	1128
Overall	2233	1502	1252	1250	1224	1236	1215	1207	1221

4.2 Analysis based on synthetic data

In order to more fully compare model performances than is possible on the CES data, we created synthetic data by adding noise to the QCEW series, thus preserving the existing structure of the target. Our synthetic response expresses the same seasonality as the QCEW series, facilitating comparison and, unlike for the real data where the QCEW will contain some unknown measurement error (which we ignore), we know that the QCEW are the true values for these synthetically generated data. We added Student's t distributed noise to QCEW-based ratios $R_{i,t}$. To specify true variances, we fitted the real data using the CFHS model and obtained the fitted variances for each month of the 2008 benchmark year. We used these fitted variances in generating of the noise, as described below.

We summarize our synthetic data generating process with the following steps:

- 1) For each domain $i = 1, \dots, N$, generate multivariate normal vector $(\varepsilon_{i,1}, \dots, \varepsilon_{i,12})$, using fitted variances from CFHS as true $\sigma_{i,1}^2, \dots, \sigma_{i,2}^2$ sampling variances;
- 2) To obtain the Student's t distribution with 6 degrees of freedom, generate parameter $\delta_{i,t}$, independently for each domain i and month t ,

$$\delta_{i,t} \stackrel{iid}{\sim} IG(3, 3).$$

- 3) Let $\hat{R}_{i,t} = R_{i,t} + e_{i,t}$, where $e_{i,t} = \varepsilon_{i,t} \sqrt{\delta_{i,t}}$.

On the simulated data (see Table 5), we see the pattern that is similar to what we observed in the real data: the co-modeling versions work better than where the point estimates only are modeled or where smoothed variances are used to fit the model.

With the simulated data, we can also look at the monthly results. In Table 6, we present MAD averaged over domains and months, as follows:

$$MAD = N^{-1} 12^{-1} \sum_{i=1}^N \sum_{t=1}^{12} |\Delta \tilde{Y}_{i,t} - \Delta Y_{i,t}|, \text{ where } \Delta \tilde{Y}_{i,t} = \tilde{Y}_{i,t} - \tilde{Y}_{i,t-1} \text{ and } \Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}.$$

Table 5. Simulated data results

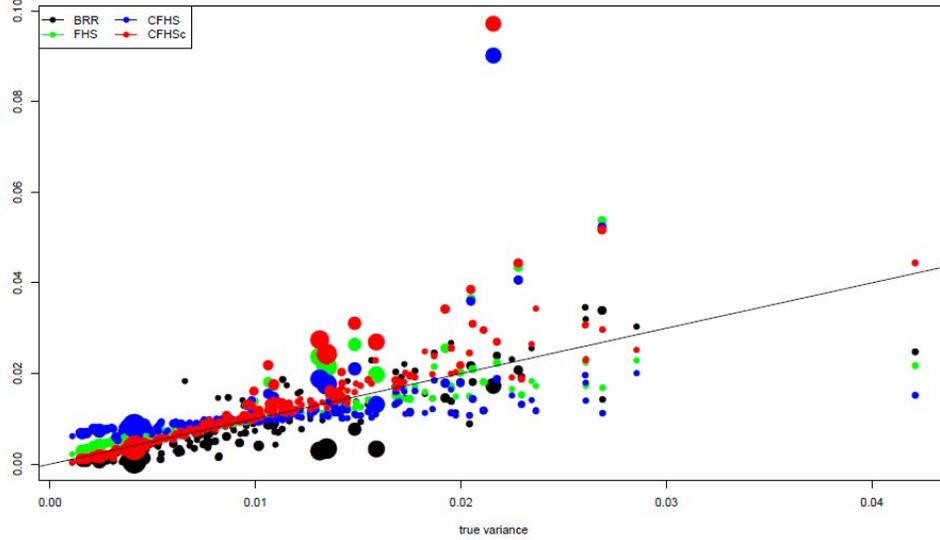
Ind	N	Direct	FH-BRR	CFH-BRR	FH-V	CFH-V	FHS	CFHS	CFHSc
1000	50	796	801	789	873	848	811	777	772
2000	141	2128	1477	1468	1584	1609	1484	1473	1448
3100	234	816	684	662	753	754	715	678	687
3200	140	683	541	550	609	600	600	568	566
4100	124	1073	757	750	812	823	760	715	720
4200	286	1798	1002	995	919	920	858	849	872
4300	194	1192	846	853	903	891	773	778	762
5000	83	995	662	670	674	637	552	514	551
5500	149	952	679	690	708	703	639	664	662
6054	150	1241	783	794	936	954	804	825	806
6055	45	521	532	524	512	502	537	525	533
6056	115	2283	1691	1665	1893	1897	1812	1715	1814
6561	59	2666	2079	2081	1987	2094	1771	1907	1710
6562	214	1213	853	866	864	891	786	805	798
7071	59	2126	1281	1305	1393	1485	1278	1241	1223
7072	80	2338	1722	1745	1713	1780	1565	1668	1563
8000	110	1628	1087	1094	998	1058	936	1025	968
Overall	2233	1393	967	967	1004	1017	923	921	915

Table 6. Simulated data, over-the-months results

Ind	N	Direct	FH-BRR	CFH-BRR	FH-V	CFH-V	FHS	CFHS	CFHSc
1000	600	285	220	226	205	213	188	202	204
2000	1692	795	503	519	507	535	444	480	448
3100	2808	328	275	274	279	280	273	281	277
3200	1680	275	213	216	185	191	187	193	192
4100	1488	396	252	262	234	247	213	234	221
4200	3432	613	350	351	315	324	278	289	276
4300	2328	456	292	297	300	307	265	279	266
5000	996	282	206	211	187	189	188	188	194
5500	1788	391	267	272	232	246	221	230	216
6054	1800	481	314	323	314	327	290	304	292
6055	540	213	176	181	162	170	158	169	167
6056	1380	923	689	687	642	661	601	643	616
6561	708	751	564	578	528	557	512	543	519
6562	2568	444	297	306	289	301	250	271	251
7071	708	771	470	500	480	516	435	480	459
7072	960	770	578	593	551	581	506	557	510
8000	1320	639	377	395	355	383	331	370	340
Overall	26796	511	344	351	328	342	302	321	306

We conclude that joint point estimates and variance models FHS and CFHSc perform better, based on the monthly results as well.

**Figure 2. Estimated vs true variances of the direct estimator
(Health Care and Social Assistance industry, month #1)**



Sampling variances fitted using different models can be compared to the true variances used to generate the synthetic data. Figure 2 presents an example of a scatter plot, for all domains in one month in Health Care and Social Assistance industry. Dots correspond to domains and show estimated variances versus true variances; black dots represent the direct estimates of variances; green, blue, and red colors show estimated variances from, respectively, models FHS, CFHS, and CFHSc. The closer the dots to the 45-degree line, the more accurate (less biased) are the estimates of the variances. We observe that for the bulk of domains the CFHSc model variance estimates lie along the 45-degree line.

The sizes of the dots are proportional to standardized distances between the direct point estimates and respective true values, $d_i = |y_i - \theta_i| / \sqrt{v_i}$. We can see a couple of larger blue and red dots on the upper edge of the plot. The dots correspond to different estimators for the same domain. The size of the dots suggests that the domain has an outlying value of the direct point estimate. The location of the dots indicate that the variance is overestimated by all three models for this outlying domain (the variance estimates from the FHS and CFHS models are very close and thus they came out on the plot as a single blue dot.) This would have the effect of over-shrinkage of the estimated value to the mean (the “synthetic part”) of the model.

While we might accept over-shrinkage of an outlier, in practice we do not observe the true value. This same over-shrinkage phenomenon would be also expected to occur in the case where the true (but unobserved) generating value for a domain deviates from the assumption of linearity. The three joint models provide various degrees of smoothing based on the input data. Whenever the joint models encounter large residuals, i.e., deviations of the observed input data from the linearity assumptions stipulated by formula (8), they may enlarge the estimated variance, particularly for models that impose a single component normal distribution as the prior for random effects. Therefore, it is important to study robustness of the models to deviations to the linearity assumption. We approach this in the next section by introducing a Monte Carlo simulation study where the true domain values are generated such that the global linearity assumption does not hold for some of the domains.

5. Model robustness study

The purpose of the simulation exercise described in this section is to study how each of the proposed joint models behaves in the case when there are domains with large deviations from the model’s linearity assumption. To this end, we generate data using several scenarios, as described below.

For a set of $i = 1, \dots, 100$ domains, we generate estimation targets θ_i as

$$\theta_i = \mu + \beta x_i + u_i, \quad (14)$$

where auxiliary data $x_i \sim U(5, 10)$, $\beta = 1$ and random effects are $u_i \sim N(0, 1)$.

We set $\mu = 0$ for the first 95 domains and $\mu = 3$ for the last 5 domains. Thus, the last 5 domains induce a deviation from the (overall) linearity assumption of the models.

The “observed point estimates” are

$$y_i = \theta_i + e_i, \quad (15)$$

where $e_i \sim N(0, v_i)$ and “true” variances are

$$\sigma_i^2 \sim IG(\lambda_g + 1, \lambda_g b). \quad (16)$$

Variances σ_i^2 are not observed directly. Instead, their estimates are available to a modeler. We simulate these “observed estimates of variances” as

$$v_i \sim G\left(3, 3 \frac{1}{\sigma_i^2}\right). \quad (17)$$

We consider several scenarios by varying the values of parameters $[b, \lambda_g]$, thus reflecting various schemes for the noise in the data:

- 1) Low average variance $b = 0.5$;
- 2) Medium average variance $b = 1$;
- 3) High average variance $b = 1.5$.

For each level of b , consider three levels of variability of the true variance. The value of $\lambda_g = 1$ induces the highest degree variability (of the variances, σ_i^2), while $\lambda_g = 4$ and $\lambda_g = 8$ induce gradually lower variability in the generated variances. The higher variability scenarios (inversely proportional to λ_g) are expected to generate a heavier tailed distribution for σ_i^2 that will induce outlying values of y_i for some domains.

After $S = 100$ simulations, we compute MSE for each of the above scenario $[b, \lambda_g]$ for domain i as

$$MSE_i(\hat{\theta}) = \frac{1}{S} \sum_{s=1}^S (\hat{\theta}_i - \theta_i)^2.$$

Average MSE over all 100 domains is $MSE(\hat{\theta}) = \frac{1}{100} \sum_{i=1}^{100} MSE_i(\hat{\theta})$. We also compute average MSE

separately over a set of domains with $\mu = 0$, $MSE_{\mu=0}(\hat{\theta}) = \frac{1}{95} \sum_{i=1}^{95} MSE_i(\hat{\theta})$, and over a set of domains with

$\mu = 3$, $MSE_{\mu=3}(\hat{\theta}) = \frac{1}{5} \sum_{i=96}^{100} MSE_i(\hat{\theta})$.

In Table 7, we present MSE for models each of the models and scenarios $[b, \lambda_g]$, separately for the 95 domains that were generated with $\mu = 0$ and for the 5 domains with $\mu = 3$.

Let us first discuss results for the $\mu = 0$ case, presented in the upper half of the table. We observe that all the models perform better than the direct estimator. In most of the scenarios considered, the joint models have lower MSE than the FH-based models. The exceptions are scenarios $[b = 0.5, \lambda_g]$, where the generating variance is small, such that treating it as fixed induces little distortion in the resulting model-based point estimates. In these scenarios, the co-clustering model CFHSc outperforms the other models. When the variances are relatively high, the FHS model outperforms all other models, including the clustering models CFHS and CFHSc.

Table 7: MSE, separately for domains with $\mu = 0$ and $\mu = 3$

$[b, \lambda_g]$	Y	FH	CFH	FH-V	CFH-V	FHS	CFHS	CFHSc
over 95 domains with $\mu = 0$								
[0.5, 8]	0.508	0.358	0.359	0.342	0.358	0.374	0.362	0.352
[0.5, 4]	0.513	0.351	0.353	0.348	0.362	0.373	0.363	0.347
[0.5, 1]	0.482	0.291	0.290	0.333	0.340	0.341	0.329	0.286
[1, 8]	1.016	0.568	0.574	0.526	0.560	0.515	0.519	0.559
[1, 4]	1.025	0.556	0.564	0.533	0.571	0.513	0.512	0.549
[1, 1]	0.963	0.456	0.459	0.509	0.533	0.441	0.439	0.443
[1.5, 8]	1.524	0.736	0.752	0.669	0.710	0.642	0.659	0.714
[1.5, 4]	1.538	0.717	0.733	0.672	0.720	0.633	0.644	0.694
[1.5, 1]	1.445	0.585	0.591	0.637	0.674	0.531	0.540	0.556
over 5 domains with $\mu = 3$								
[0.5, 8]	0.511	1.209	1.122	1.269	0.858	3.719	3.317	1.331
[0.5, 4]	0.556	1.204	1.140	1.311	0.889	3.851	3.534	1.419
[0.5, 1]	0.361	0.943	0.898	1.386	0.890	3.892	3.511	1.589
[1, 8]	1.023	1.955	1.815	2.089	1.652	3.350	2.948	1.952
[1, 4]	1.112	1.946	1.823	2.168	1.707	3.469	3.124	2.032
[1, 1]	0.721	1.540	1.453	2.235	1.758	3.375	3.010	1.901
[1.5, 8]	1.534	2.423	2.268	2.554	2.242	3.392	3.016	2.478
[1.5, 4]	1.668	2.392	2.250	2.629	2.287	3.484	3.121	2.538
[1.5, 1]	1.082	1.951	1.840	2.672	2.344	3.316	2.933	2.303

We next turn to the lower part of Table 7 that presents results for the 5 domains with $\mu = 3$, i.e. the cases where the true domain target values deviate from the models' linearity assumption. As expected, all the models perform worse

than the direct estimator for these domains. The joint models (FHS, CFHS, CFHSc) tend to inflate the estimated variances for those domains deviating from linearity such that the resulting model-based point estimate are overly shrunk. Interestingly, the FH and CFHSc results are best among the models, although for different reasons. The FH is more likely to return the data value when the data value deviates from the normality assumption on the random effects. So the FH performs relatively well here because the data provides a better estimator than an overly smoothed model value that fails to capture the deviation from linearity. The CFHSc possesses the flexibility to construct a separate cluster (or component normal distribution) for a domain whose point estimate deviates from linearity.

Table 8: Properties of the credible intervals, over 95 domains with $\mu = 0$

$[b, \lambda_g]$	FH	CFH	FH-V	CFH-V	FHS	CFHS	CFHSc
Coverage (0.95 nominal)							
[0.5, 8]	0.914	0.916	0.957	0.959	0.933	0.943	0.933
[0.5, 4]	0.915	0.915	0.952	0.955	0.934	0.943	0.935
[0.5, 1]	0.921	0.923	0.957	0.958	0.943	0.951	0.951
[1, 8]	0.922	0.924	0.963	0.964	0.951	0.958	0.938
[1, 4]	0.923	0.924	0.960	0.961	0.951	0.956	0.942
[1, 1]	0.927	0.930	0.963	0.965	0.957	0.966	0.952
[1.5, 8]	0.928	0.931	0.971	0.969	0.960	0.964	0.944
[1.5, 4]	0.930	0.929	0.968	0.966	0.959	0.963	0.946
[1.5, 1]	0.933	0.933	0.970	0.967	0.965	0.970	0.956
Length							
[0.5, 8]	2.259	2.281	2.392	2.492	2.259	2.322	2.293
[0.5, 4]	2.208	2.230	2.384	2.489	2.234	2.288	2.244
[0.5, 1]	2.004	2.019	2.371	2.488	2.186	2.255	2.069
[1, 8]	2.933	2.985	3.102	3.226	2.881	2.983	2.966
[1, 4]	2.869	2.917	3.080	3.222	2.846	2.940	2.897
[1, 1]	2.599	2.633	3.060	3.200	2.725	2.834	2.642
[1.5, 8]	3.445	3.523	3.659	3.741	3.366	3.504	3.435
[1.5, 4]	3.364	3.434	3.625	3.738	3.316	3.439	3.353
[1.5, 1]	3.027	3.082	3.595	3.701	3.135	3.271	3.036

Coverage probabilities and interval lengths for 95% nominal credible intervals for the fitted values based on all the models are presented in Table 8 (for $\mu = 0$ domains) and Table 9 (for $\mu = 3$ domains.) Coverages are derived for each domain over 100 simulations. After that, the domain results are averaged over respective groups of domains:

$$\bar{c}_{\mu=0} = \frac{1}{95} \sum_{i=1}^{95} \hat{c}_i \quad \text{and} \quad \bar{c}_{\mu=3} = \frac{1}{5} \sum_{i=96}^{100} \hat{c}_i, \quad \text{where} \quad \hat{c}_i = \frac{1}{S} \sum_{s=1}^S I\{\hat{q}_{i,0.025} \leq \theta_i \leq \hat{q}_{i,0.975}\}$$

and $\hat{q}_{i,0.025}$, $\hat{q}_{i,0.975}$ are quantiles of the posterior distribution of the fitted values for domain i . The average length of the intervals are obtained as $\bar{l}_{\mu=0} = \frac{1}{95} \sum_{i=1}^{95} \hat{l}_i$ and $\bar{l}_{\mu=3} = \frac{1}{5} \sum_{i=96}^{100} \hat{l}_i$, where $\hat{l}_i = \hat{q}_{i,0.975} - \hat{q}_{i,0.025}$.

For $\mu = 0$ cases, coverages for all models, except the FH, are close to nominal. The FH coverage is somewhat low, especially for lower variances scenarios of $[b = 0.5, \lambda_g]$ and $[b = 1, \lambda_g]$. The coverages for the FH-V and CFH-V models are slightly higher than the nominal; their average interval lengths are longer than in the other models. This result is consistent with the previously observed higher MSE results for these models.

Table 9: Properties of the credible intervals, over 5 domains with $\mu = 3$

$[b, \lambda_g]$	FH	CFH	FH-V	CFH-V	FHS	CFHS	CFHSc
Coverage (0.95 nominal)							
[0.5, 8]	0.678	0.718	0.658	0.814	0.330	0.368	0.706
[0.5, 4]	0.702	0.732	0.654	0.810	0.342	0.354	0.694
[0.5, 1]	0.756	0.766	0.614	0.820	0.332	0.394	0.680
[1, 8]	0.676	0.720	0.654	0.778	0.482	0.572	0.706
[1, 4]	0.690	0.738	0.642	0.770	0.474	0.540	0.710
[1, 1]	0.734	0.752	0.632	0.756	0.468	0.564	0.692
[1.5, 8]	0.708	0.740	0.706	0.782	0.566	0.642	0.718
[1.5, 4]	0.726	0.750	0.704	0.776	0.566	0.666	0.702
[1.5, 1]	0.728	0.766	0.694	0.748	0.556	0.652	0.698
Length							
[0.5, 8]	2.241	2.269	2.390	2.475	2.564	2.647	2.443
[0.5, 4]	2.231	2.243	2.388	2.474	2.575	2.646	2.448
[0.5, 1]	2.006	2.020	2.371	2.490	2.551	2.637	2.377
[1, 8]	2.927	2.973	3.118	3.221	3.094	3.197	3.071
[1, 4]	2.880	2.943	3.072	3.204	3.081	3.179	3.040
[1, 1]	2.608	2.635	3.062	3.196	2.977	3.089	2.819
[1.5, 8]	3.428	3.511	3.650	3.738	3.515	3.639	3.508
[1.5, 4]	3.378	3.455	3.611	3.723	3.489	3.641	3.467
[1.5, 1]	3.038	3.076	3.603	3.689	3.342	3.466	3.194

The model coverages for $\mu = 3$ domains are low under all of the models. Among the three joint models, the co-clustering model CFHSc has the highest coverage (for even shorter, interval lengths), which is consistent with the previously noted result of the lower MSE for the co-clustering model.

These results show that none of the models considered provide satisfactory estimates for the domains where there are significant deviations from the model linearity assumption. Therefore, it is important to develop a procedure that would identify domains that do not fit the model well. In the next Section, we propose such a procedure to create a list of ‘suspect’ domains that are not well described by the model.

6. Improved handling of non-modeled domains

Although the CES survey uses models for a number of its small domains, the direct sample-based estimator is used for publication of moderately and larger sized domains. Before these estimates are published, they have to be reviewed. In this section, we propose a screening procedure that can be used to facilitate the analyst’s review of the direct estimates before they go to production.

The proposed screening creates a list of domains that are not well described by the assumed model. For the larger, direct sample-based domains, analysts may find influential reports (that may need to be downweighted) or submission errors (that would be subsequently repaired) among establishments that would induce outliers in the sample estimates. So, even though models would not be used to provide estimators for large-sized domains, they may be used to check for outliers in an efficient way.

Our screening procedure would also be expected to flag deviations from linearity among all domains – including those which are modeled – for analyst checking. To the extent that data submission errors and low quality data (due

to small domain sizes) are ruled out, the nominated domain may be assumed to represent a deviation from linearity, in which case the direct estimator for that domain would replace the modeled estimate.

6.1 Efficient identification of outliers and deviations from linearity

We earlier showed that our models may poorly fit domains expressing deviations from the linearity assumption due to over-smoothing. Ideally, we want to flag these domains as not generated from our model, in this case, and just use the direct estimator. Similarly, our models may be useful to flag outliers with respect to the model due to unreliable estimators or establishment input errors. We would like to flag and correct these points. It is time prohibitive to have a survey analyst perform manual checking of all domains due to the tightly scheduled CES production environment. In what follows, we formulate a hypothesis test from the posterior predictive distribution under the model to assess whether the direct estimator for each domain was generated from our chosen model. We nominate a few domains out of many under this procedure that allows focused, efficient investigation by the survey analyst of whether any of the few identified domains are outliers. If so and the survey analyst concludes that there are input errors, they will be corrected. If not, the large difference between modeled estimators and the direct estimators for these domains are assumed to represent deviations from linearity.

The usual strategy for introducing of a model in the CES production is to consider a set of candidate models $\mathbf{M}_1, \dots, \mathbf{M}_W$ and thoroughly test them on a number of historical series over several years. Suppose researchers are satisfied with the results of such a multi-year study, and one of the models, \mathbf{M}_w , is accepted for production. The question remains, what if the selected model \mathbf{M}_w works well in general but fails for some domains in some months?

As we earlier noted, the analysis may suggest that model-based estimates for some of these domains are unreliable in the case of deviations from linearity; in such a case, the direct sample estimates would be used for publication. Alternatively, the direct estimates may be considered not trustworthy (for example, due to small sample size or extreme sample reports). In the latter case, model estimates could be used even though they are seemingly inconsistent with the data.

We now proceed to describe the method of creating the list of suspect domains. The method is based on the Bayesian multiple hypotheses testing and posterior predictive checking.

For a given model \mathbf{M}_w over the space of candidate models indexed by $w = 1, \dots, W$, let $y_i^l, l = 1, \dots, L$ be replicate data draws from posterior predictive distribution $p(y_i^l | y_i, \mathbf{M}_w)$ for domain i (after marginalizing out the model parameters).

For each domain i , consider hypothesis H_{i0} that the domain response is generated from the model, which means that y_i follows $p(y_i^l | y_i, \mathbf{M}_w)$.

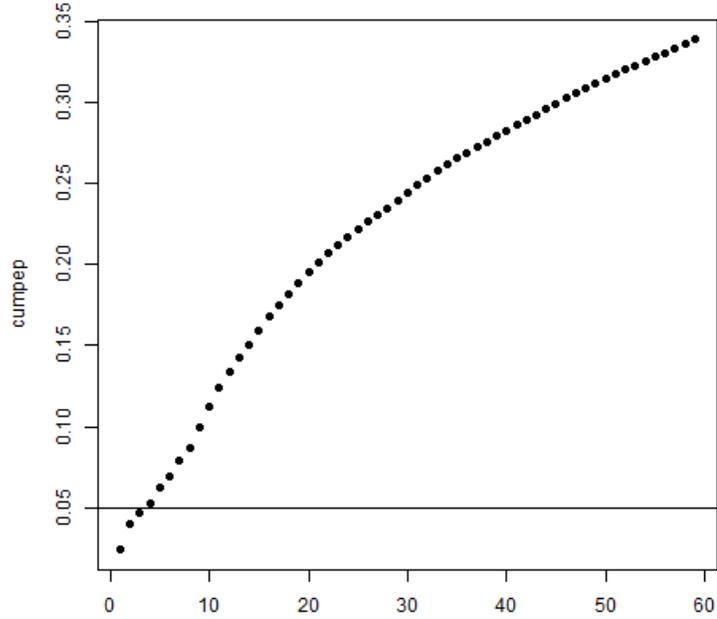
Define $p_i = \min\left(P\{y_i^l \leq y_i | y_i, \mathbf{M}_w\}, P\{y_i^l \geq y_i | y_i, \mathbf{M}_w\}\right)$, where P denotes the probability of the event that y_i is generated from model \mathbf{M}_w . In this sense, p_i denotes the probability of erroneously rejecting H_{i0} that domain i is generated from the model.

Let set D be the set of “discoveries” (i.e., the domains that are deemed not generated from the model according to the definition of H_{i0} .) Then the expected number of “false discoveries” is $F = E[p_i | i \in D, \mathbf{M}_w]$ and the estimated F is computed from the average,

$$\hat{F} = \frac{1}{|D|} \sum_D p_i. \quad (18)$$

Next, we set threshold, q , a hyperparameter setting that denotes the maximum percent of allowable “falsely discovered domains” (Storey, 2003). The size of the list of “discoveries” will depend on q : set D will contain the maximum number of domains such that \hat{F} does not exceed q .

Figure 3: Cumulative mean, by domain for model FHS (industry 6561, month #3)



The algorithm follows:

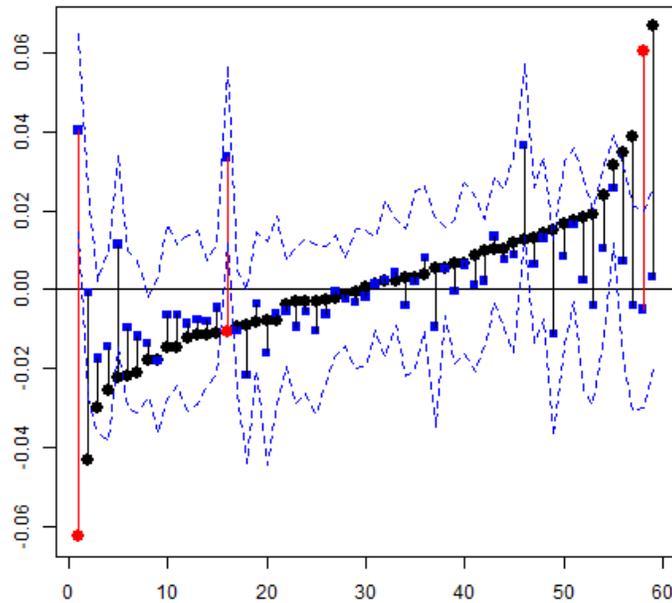
1. Sort p_i 's in the ascending order, $p_{(1)} \leq \dots \leq p_{(N)}$ and compute the cumulative mean.
2. An example of the plot of the cumulative mean is presented in Figure 3. We may review the plot to think of what the reasonable q -value could be. Or we may just set the q -value once in advance.
3. Suppose we choose $q = 0.05$. Then D will consist of the first d domains with smallest p_i 's:

$p_{(1)} \leq \dots \leq p_{(d)}$, such that

$$\frac{1}{d} \sum_{i=1}^d p_{(i)} \leq q.$$

In other words, $p_{(d)}$ is the p -value that guarantees that the false discovery rate does not exceed $q = 0.05$.

Figure 4: Sorted residuals for model FHS (industry 6561, month #3)



In Figure 4, we show results of the screening procedure using one of the industries and months as an example. The design of Figure 4 is aimed to visualize possible cases of “discoveries”. Since in our simulation study, we know the true value of the response for each domain, we can compute the residual of the direct estimate relative to the true value. We sort the residuals of the direct estimates on the plot in Figure 4: each black or red dot corresponds to a domain indexed residual value. The farther the dot is from the horizontal reference line, the farther from the truth is the direct point estimate. The blue squares correspond to residuals formed based on the model fitted values. Blue dotted lines show confidence interval for the fitted values. The segments connecting blue squares with the corresponding dots show the distance between the direct sample-based estimates and respective model-based estimates. Our screening procedure identified three domains (displayed as red dots) as not likely to have been generated from our model. Our intent would be to forward these three domain values for further analyst review. The two red dots on the edges of the plot express large residual values and are likely to be regarded as having noisy direct estimates, such that we suspect these points as being outliers; notice that the respective fitted values are closer to the horizontal line, thus providing a better estimate than the direct estimate. By contrast, the red dot closer to the center of the plot is also on the list for review, but in this case we suspect the direct estimates are better than the model based because the residual of the modeled point estimate is large. This outcome is a case of a large deviation from linearity in the signal that is over-smoothed by the model. It bears mention that these nominated points resulted from applying our screening procedure, which does not rely on knowledge of the true values. This is important because the true values are not available to analysts and they will not see this plot, but they will have the small number of domains screened out by our procedure. To make their judgment, analysts may look for possible outliers at the unit (e.g., business establishment) level data that could have affected the direct estimates or by using subject matter knowledge outside the sample.

We applied this test to the data from the simulation study considered in Section 5 and created a list of “discoveries” to be sent for the review by analysts. In our study, we set the threshold of $q = 0.10$. Since this review is not available in our simulations, we make a favorable assumption that analysts make right decisions of whether an estimate on the list is an outlier or a true phenomenon. Namely, assume that all “discoveries” from the set of the 95 domains generated with $\mu = 0$ were attributed to “bad sample” and analysts’ decision was to use model estimates for these domains; all “discoveries” from the set of the 5 domains generated with $\mu = 3$ were attributed to failure of

the model linearity assumption and the direct sample estimates were used for such domains, instead of the model estimates. The result of replacement is given in Table 10.

Table 10: MSE for domains with $\mu = 3$, after “analysts review.”

$[b, \lambda_g]$	Y	FH	FHS	CFHS	CFHSc	FHS*	CFHS*	CFHSc*
[0.5, 8]	0.511	1.209	3.719	3.317	1.331	0.849	0.930	0.911
[0.5, 4]	0.556	1.204	3.851	3.534	1.419	0.842	0.953	0.948
[0.5, 1]	0.361	0.943	3.892	3.511	1.589	0.618	0.720	0.741
[1, 8]	1.023	1.955	3.350	2.948	1.952	1.741	1.820	1.742
[1, 4]	1.112	1.946	3.469	3.124	2.032	1.893	1.909	1.771
[1, 1]	0.721	1.540	3.375	3.010	1.901	1.367	1.488	1.350
[1.5, 8]	1.534	2.423	3.392	3.016	2.478	2.479	2.493	2.407
[1.5, 4]	1.668	2.392	3.484	3.121	2.538	2.675	2.721	2.464
[1.5, 1]	1.082	1.951	3.316	2.933	2.303	2.023	2.112	1.854

The first columns of Table 10 are given for comparison: these are MSEs for the original estimates, previously reported in Table 7. The last three columns, labeled FHS*, CFHS*, CFHSc*, are MSE results after the replacement of domains in the list by the direct sample estimates. As we can see, this replacement lead to visible reduction in MSE; however, the results are still higher than the respective MSE’s for the estimates given in column labeled Y. It is possible to increase the threshold q , which would increase the number of domains in the list for analysts.

However, this would also increase the workload for analysts and may result in more domains to be mislabeled as “deviations from linearity assumptions”, while in fact their appearance on the list could be due to poor direct estimates. In practice, certain tuning will be required to set a threshold, taking in consideration the workload and timeline restrictions.

6.2 Replacing Variances with Modeled Estimates

A useful outcome of the joint models is the fitted values for variances of the direct sample estimators. As noted earlier, estimates for the medium-to-larger sized domains in the CES survey are published using the direct survey estimates. However, the corresponding variance estimates are not stable even for larger domains. The model-based estimates of variances potentially provide a more stable alternative. We next examine the coverage properties of the 95% confidence intervals for the direct sample estimator, constructed under the assumption of normality of the

direct estimates and using alternative estimates of variances of the direct estimates: $\bar{d}_{\mu=0} = \frac{1}{95} \sum_{i=1}^{95} \hat{d}_i$ and

$\bar{d}_{\mu=3} = \frac{1}{5} \sum_{i=96}^{100} \hat{d}_i$, where $\hat{d}_i = \frac{1}{S} \sum_{s=1}^S I \left\{ \hat{\theta}_i - z_{0.975} \sqrt{\hat{\sigma}_i^2} \leq \theta_i \leq \hat{\theta}_i + z_{0.975} \sqrt{\hat{\sigma}_i^2} \right\}$, $\hat{\sigma}_i^2$ is the estimated variance of

$\hat{\theta}_i$ derived from respective models, and $z_{0.975}$ is the quantile of the standard normal distribution. Since the true variances σ_i^2 are available in the simulation study, we also included in the analysis intervals based on the true

variances. The interval lengths are $\bar{l}_{\mu=0}^d = \frac{1}{95} \sum_{i=1}^{95} \hat{l}_i^d$ and $\bar{l}_{\mu=3}^d = \frac{1}{5} \sum_{i=96}^{100} \hat{l}_i^d$, where $\hat{l}_i^d = 2z_{0.975} \sqrt{\hat{\sigma}_i^2}$.

Coverage properties and lengths of the intervals are given in Table 11 (for the $\mu = 0$ domains) and Table 12 (for $\mu = 3$ domains). The column titled “True” corresponds to intervals constructed using true variances σ_i^2 generated by (16). The column titled “direct” is based on the observed variance values v_i containing noise, as given by (17). The FHS, CFHS, and CFHSc columns are based on the respective model fitted variances.

Table 11: Properties of the 95% confidence intervals for direct estimator based on alternative variance estimates, over 95 domains with $\mu = 0$

$[b, \lambda_g]$	True	“direct”	FHS	CFHS	CFHSc
Coverage (0.95 nominal)					
[0.5, 8]	0.946	0.897	0.973	0.974	0.934
[0.5, 4]	0.944	0.896	0.973	0.974	0.937
[0.5, 1]	0.949	0.907	0.984	0.983	0.957
[1, 8]	0.946	0.897	0.966	0.964	0.932
[1, 4]	0.944	0.896	0.966	0.964	0.935
[1, 1]	0.949	0.907	0.977	0.976	0.954
[1.5, 8]	0.946	0.897	0.962	0.960	0.932
[1.5, 4]	0.944	0.896	0.962	0.961	0.935
[1.5, 1]	0.949	0.907	0.975	0.973	0.953
Length					
[0.5, 8]	2.724	2.612	3.131	3.126	2.772
[0.5, 4]	2.682	2.569	3.116	3.115	2.738
[0.5, 1]	2.449	2.355	3.122	3.134	2.600
[1, 8]	3.853	3.695	4.182	4.182	3.889
[1, 4]	3.793	3.633	4.150	4.159	3.837
[1, 1]	3.464	3.330	4.042	4.061	3.591
[1.5, 8]	4.719	4.525	5.039	5.037	4.752
[1.5, 4]	4.646	4.449	4.987	4.999	4.686
[1.5, 1]	4.242	4.079	4.795	4.809	4.367

The interval lengths reported in Tables 11 and 12 are longer than the intervals in Tables 8 and 9, which is expected and consistent with the fact that model-based estimates are more efficient, overall.

Note that confidence intervals based on the “direct” variance estimator result in significant undercoverage (in the range of 88-91% for 95% nominal). This is the result of the normality-based interval construction. Given (17), it is easy to see that correct intervals based on v_i should be constructed using quantiles $t_{6,0.975}$ of the Student’s t distribution with 6 degrees of freedom, rather than normal quantiles $z_{0.975}$. This construction can be easily corrected in this simulation to achieve the nominal level; however, in practice, such an adjustment to the degrees of freedom is not always considered or is easy to make. Thus, results in column “direct” demonstrate the effect of such a misspecification.

For the model-fitted variances, note the over-coverage for intervals based on the FHS and CFHS variances, which is more pronounced for the $\mu = 3$ cases. The average interval length is also larger for FHS and CFHS, compared to the interval length based on the true variance, as well as to the other models. This result is consistent with the previously noted tendency of the FHS and CFHS models to overestimate variances of direct estimates and with a higher degree of overestimation for the $\mu = 3$ domains. The latter is the result of the fact that FHS and CFHS tend to “confuse” the $\mu = 3$ domains with outliers.

The coverage of the CFHSc-based intervals is close to nominal. Their average length is somewhat larger than the true intervals’ length, yet it is substantially smaller than for the FHS and CFHS cases. This observation suggests that the CFHSc-fitted variances are less biased than the estimates based on FHS or CFHS. Thus, even though the model failed to provide satisfactory point estimates for the $\mu = 3$ domains, the estimated variance provides nominal

confidence intervals for the direct sample estimates. This is an encouraging outcome suggesting that the CFHSc-based variance estimates can be used with the publication of the direct sample-based estimates for those domains that are not model based.

Table 12: Properties of the 95% confidence intervals for direct estimator based on alternative variance estimates, over 5 domains with $\mu = 3$

$[b, \lambda_g]$	True	“direct”	FHS	CFHS	CFHSc
Coverage (0.95 nominal)					
[0.5, 8]	0.930	0.886	0.978	0.980	0.950
[0.5, 4]	0.926	0.878	0.980	0.982	0.952
[0.5, 1]	0.966	0.908	0.996	0.996	0.980
[1, 8]	0.930	0.886	0.968	0.970	0.948
[1, 4]	0.926	0.878	0.974	0.974	0.950
[1, 1]	0.966	0.908	0.990	0.986	0.970
[1.5, 8]	0.930	0.886	0.964	0.968	0.944
[1.5, 4]	0.926	0.878	0.972	0.974	0.946
[1.5, 1]	0.966	0.908	0.984	0.982	0.966
Length					
[0.5, 8]	2.685	2.592	4.116	4.053	3.017
[0.5, 4]	2.658	2.593	4.220	4.174	3.077
[0.5, 1]	2.452	2.387	4.389	4.315	3.132
[1, 8]	3.797	3.666	4.807	4.767	4.086
[1, 4]	3.759	3.667	4.891	4.864	4.137
[1, 1]	3.468	3.376	4.906	4.851	3.994
[1.5, 8]	4.650	4.490	5.530	5.491	4.941
[1.5, 4]	4.604	4.491	5.596	5.572	4.987
[1.5, 1]	4.247	4.135	5.483	5.435	4.752

Although the simulation study of Section 5 was constructed by choosing the parameters of the gamma distribution in formula (17) that immitate a “poor quality” sample, there is indication that similar results for variance estimates also hold for “larger” samples. We completed a similar simulation study by choosing the shape and scale parameters of the gamma distribution equal to 10, thus simulating the situation for larger samples. In Table 13, we present the coverage properties for all domains for the case of “larger sample”, corresponding to 20 degrees of freedom in generating variances v_i .

Raw sampling variances in Table 13 provide about 93% coverage, which is only slightly below the nominal. Correspondingly, the average interval length is slightly lower than the length based on the true variances. The MSE of the interval length, however, is larger than the MSE of the CFHSc model fitted variances. The exception are scenarios with $\lambda_g = 1$, that correspond to cases of frequent outliers in the estimates. The latter, however, is not expected to happen in the larger domains that are slated for the sample based estimation.

Table 13: Properties of the 95% confidence intervals for direct estimator based on alternative variance estimates

$[b, \lambda_g]$	True	“direct”	FHS	CFHS	CFHSc
Coverage					
[0.5, 8]	0.946	0.933	0.972	0.972	0.947
[0.5, 4]	0.944	0.930	0.973	0.973	0.950
[0.5, 1]	0.950	0.939	0.986	0.985	0.962
[1, 8]	0.946	0.933	0.966	0.966	0.947
[1, 4]	0.944	0.930	0.968	0.967	0.949
[1, 1]	0.950	0.939	0.980	0.978	0.960
[1.5, 8]	0.946	0.933	0.963	0.964	0.947
[1.5, 4]	0.944	0.930	0.966	0.965	0.949
[1.5, 1]	0.950	0.939	0.977	0.976	0.960
Average of CI lengths					
[0.5, 8]	2.722	2.689	3.059	3.053	2.802
[0.5, 4]	2.681	2.646	3.059	3.061	2.769
[0.5, 1]	2.449	2.427	3.123	3.099	2.604
[1, 8]	3.850	3.803	4.152	4.154	3.944
[1, 4]	3.792	3.742	4.122	4.127	3.892
[1, 1]	3.464	3.433	4.000	4.001	3.619
[1.5, 8]	4.715	4.658	5.027	5.033	4.823
[1.5, 4]	4.644	4.583	4.977	4.986	4.757
[1.5, 1]	4.242	4.204	4.763	4.765	4.414
MSE of CI lengths					
[0.5, 8]	0	0.188	0.314	0.302	0.155
[0.5, 4]	0	0.189	0.382	0.381	0.161
[0.5, 1]	0	0.189	1.031	0.971	0.233
[1, 8]	0	0.376	0.371	0.366	0.307
[1, 4]	0	0.379	0.426	0.428	0.314
[1, 1]	0	0.379	0.950	0.946	0.389
[1.5, 8]	0	0.564	0.494	0.491	0.457
[1.5, 4]	0	0.568	0.550	0.554	0.467
[1.5, 1]	0	0.568	1.113	1.119	0.562

7. Summary

In this paper, we applied joint modeling of the point estimates and their variances to CES data and obtained more efficient results than in the case of the plugged in “fixed and known” variances. We extended the models of Maiti et al. (2014) and Sugawara et al. (2017) by allowing the data to estimate a clustering structure on random effects and variances to account for deviations from linearity and outlyingness. For the bulk of domains, the co-clustering model provides better estimates of direct survey variances. Our simulations show that co-clustering model is more robust to deviations from linearity assumptions in terms of coverage. In the presence of large deviations from linearity, we observed that although the resulting estimates from the co-clustering model are better than with the alternatives, they are still not “good enough”: in the presence of large deviations from the linearity assumption, model-based estimates may be worse than direct survey estimates.

It is a good practice to perform careful model checks before choosing a model. However, thorough model evaluation can be an unrealistic task in a tightly scheduled production environment. The checking task is so important, however, that estimates are thoroughly tested based on a number of historical series before a model is accepted for implementation in production. Therefore, we devised an automated, fast computing testing procedure based on the Bayesian FDR to nominate a small subset of domains for analysts review on a timely basis. Our procedure evaluates the probability that the direct estimate for a domain was generated from our candidate model. This procedure could become a useful tool for analysts to mark unusual estimates before they are published.

Lastly, there is indication that model fitted variances for direct survey estimates provide a more stable alternative to the raw sample-based estimates of variances. This is a potentially useful by-product from the joint modeling of direct estimates of point estimates and variances.

References:

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

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*, 269–277.

Gelman, A., Lee, D., and Guo, J. (2015) Stan: A probabilistic programming language for Bayesian inference and optimization. In press, *Journal of Educational and Behavior Science*.
http://www.stat.columbia.edu/~gelman/research/published/stan_jebs_2.pdf

Gershunskaya, J. and Savitsky, T.D. (2017) Dependent Latent Effects Modeling for Survey Estimation with Application to the Current Employment Statistics Survey. *Journal of Survey Statistics and Methodology*, Volume 5, Issue 4, 433–453, <https://doi.org/10.1093/jssam/smx021>

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D.M. (2017), Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(14):1–45.

Maiti, T., Ren, H. and Sinha, A. (2014). Prediction error of small area predictors shrinking both means and variances, *Scandinavian Journal of Statistics*, *41*, 775-790.

Neal, R. M. (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Stan Development Team (2017), Stan modeling Language User’s Guide and Reference Manual, Version 2.17.0 [Computer Software Manual], available at <http://mc-stan.org/>. Last accessed 04/23/2018

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q -value. *Annals of Statistics* **31**, 2013–2035.

Sugasawa, S., Tamae, H., and Kubokawa, T. (2017) Bayesian Estimators for Small Area Models Shrinking Both Means and Variances. *Scand J Statist*, *44*: 150–167. doi: [10.1111/sjos.12246](https://doi.org/10.1111/sjos.12246).