Robust estimation in the presence of deviations from linearity in small domain models

Julie Gershunskaya, Terrance D. Savitsky[1]

U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212

Abstract:

Small domain estimation models, like the Fay-Herriot, often assume a normally distributed latent process centered on a linear mean function. The linearity assumption may be violated for domains that express idiosyncratic phenomena not captured by the predictors. Under a single component normal distribution prior for the random effects, direct sample estimates for those domains would be viewed as if they were outliers with respect to the model, when in fact they may reflect the underlying true population value. The model interpretation is also confounded by the variances of direct sample estimates because, while typically treated as fixed and known, they are estimates and thus contain noise. In this paper, we construct a joint model for the direct estimates and their variances where we replace the normal distribution for the latent process with a nonparametric mixtures of normal distributions with the goal to improve robustness in estimation quality for these idiosyncratic domains. We devise a model-based screening tool that leverages the posterior predictive distribution under the model to nominate domains where the model may not accurately account for deviations from the linearity assumption. Our screening tool nominates a few domains to allow for a focused investigation to determine whether a deviation from linearity is real. The U.S. Bureau of Labor Statistics' Current Employment Statistics (CES) survey publishes monthly employment estimates for domains defined by industry and geography. Model estimation is performed for smaller domains to improve the reliability of the direct estimator. We compare fit performances for our candidate models under data constructed to be similar to the CES and conduct a simulation study to assess the robustness of our candidate models in the presence of deviations from linearity. We apply our model-based screening method and quantify its ability to improve the quality of published estimates.

Key Words: Bayesian Hierarchical Modeling, Posterior predictive distribution, False discovery rate, Dirichlet process, Fay-Herriot, Variational Bayes, Stan

1. Introduction

Large government surveys, such as the Current Employment Statistics (CES) survey considered in this paper, are designed to produce high quality sample-based estimates for a number of state and national 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

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

improve the quality of the estimates. One of the most popular models is the classical Fay-Herriot (FH) model (Fay and Herriot 1979). The 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 point estimates are assumed to be fixed and known. In reality, these variances are not known and sample-based variances can be plugged in as if they were true variances. However, such sample-based estimates of variances contain noise.

The usual practice to address noise in sampling-based variances is to smooth the noise by using model-based estimates 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 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 implied in Sugasawa et al. (2017). Employment may grow or decrease faster in some groups of domains included in the model due to idiosyncratic effects in a particular domain that is not shared among other domains. This phenomenon may be captured by imposing a mixture of normal distributions assumption on the random effects.

The models considered may still fail to describe true population target in domains having large deviations from the linearity, even under a prior formulation on random effects that induces clustering of domains. So we devise a posterior predictive checking approach to uncover domains that are not well described (or generated) by a particular 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 estimates, in which case the modeled estimates are accepted; otherwise, analysts may decide that a particular domain's deviation from the linearity expresses real economic movement. In the case that the deviation from linearity is deemed real, the modeled estimate for that particular domain would be replaced by the direct estimate. We show in the sequel that making such replacements of selected model values would be expected to notably improve estimation performance. Our testing approach is also expected to be useful when applied to non-modeled domains.

We compare the performances of alternative models under synthetic data generated from the Quarterly Census of Employment and Wages (QCEW), which is considered the gold standard (because it lacks sampling error) for evaluation of CES estimates. Our simulation results confirm that co-modeling of the direct point estimates and their variances leads to

improved estimates. We follow by conducting a simulation study to assess the robustness of our models for capturing deviations from linearity among some domains.

We adopt the hierarchical Bayesian paradigm for development of the models. The code is written 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).

The paper is organized as follows. The models considered in this paper are stated in Section 2. In Section 3, after a brief introduction of estimation procedures and the form of the sample-based estimator used in CES, we discuss the results of application of the models to the synthetic data generated by adding noise to the true historical series. In Section 4, we conduct a robustness study of our candidate model formulations to assess their performances under deviations from linearity. Section 5 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 conclude with a summary discussion in Section 6.

2. Description of the models

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, $i = 1,...,N$, assume

$$y_i \mid \theta_i \overset{ind}{\sim} N\left(\theta_i, v_i\right), \tag{1}$$

$$\theta_i \mid \mu, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right). \tag{2}$$

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 (1)). Equation (2) 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 as μ.)

As noted, sampling variances v_i in the FH model are considered fixed and known. In practice, estimates of true variances are used. We consider two possibilities for the treatment of v_i : 1) using direct sample based estimates of true variances, which treats the variances as fixed and known in a Fay Herriot model that we refer to as FH; 2) using a smoothed estimator of variances that is plugged into the Fay-Herriot model. For this model (referred to as FH-V), the estimation of the variances is performed, separately, in a first step and then used as plug-in estimators for v_i in the Fay-Herriot model. 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 under a Bayesian construction).

In the next two models, 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.

Our first model that co-estimates the direct estimates and their variances is referred to as FHS and is a modification of an approach considered by Sugasawa et al. (2017). Assume the following hold for pair of direct survey estimates $\left(y_i, v_i\right)$ for each domain i:

$$y_i \mid \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \sigma_i^2\right), \tag{3}$$

$$\theta_i \mid \mu, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right). \tag{4}$$

$$v_i \mid a, \sigma_i^2 \overset{ind}{\sim} G\left(\frac{a n_i^*}{2}, \frac{a n_i^*}{2\sigma_i^2}\right), \tag{5}$$

$$\sigma_i^2 \mid b, \boldsymbol{\gamma} \overset{ind}{\sim} IG\left(2, b \exp\left(z_i^T \boldsymbol{\gamma}\right)\right). \tag{6}$$

Lines (3)-(4) are the usual FH assumptions on the point estimates y_i and lines (5)-(6) describe the variance model, where parameter σ_i^2 is the true sampling variance; 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 (5), estimated variances depend on the sample size n_i, where for a set of domains with unequal number of respondents, we use the standardized response size, $n_i^* = \left(n_i - \left\{\min_i n_i - 1\right\}\right) \Big/ \left(\max_i n_i - \min_i n_i\right) \in [0,1]$. Our assumption is somewhat different from Maiti et al. (2014) and Sugasawa et al. (2017) as we include an additional (unknown) parameter, a, to regulate the scale and shape of the distribution. In our application, we found that for moderate sample sizes, using the sample size alone would result in predicted variances that are overly similar to direct estimates of variances.

The normality assumption used in (2) and (4) may not be realistic. For example, if a single or a handful of domains deviate significantly from $\mu + \mathbf{x}_i^T \boldsymbol{\beta}$, assumption (2) of the FH model would result in the under-shrinkage of the bulk of the observations. In the FHS model, violation of assumption (4) would result in overestimation of sampling variances in the domains where the signal deviates from the assumed linearity and, hence, in over-shrinkage of estimates for these domains.

Our capstone model is termed FHSC and is designed to allow for deviations from linearity assumption $\mu + \mathbf{x}_i^T \boldsymbol{\beta}$ for some subsets of domains. Namely, we replace the normal distribution of assumption (4) with a finite mixture normal distributions. Specifically, we assume the existence of K latent clusters having cluster specific intercepts μ_k, for $k = 1, ..., K$, and common variance τ_u^2:

$$\theta_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k N\left(\mu_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right) \tag{7}$$

In addition, the inverse gamma assumption in (6) can be relaxed by specifying a mixture of the inverse gamma distributions with the cluster-specific shape parameter b_k:

$$\sigma_i^2 \mid \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\pi} \overset{ind}{\sim} \sum_{k=1}^{K} \pi_k IG\left(2, b_k \exp\left(z_i^T \boldsymbol{\gamma}\right)\right). \tag{8}$$

Table 1. FH, FHS, and FHSC models

FH	FHS	FHSC
$y_i \mid \theta_i \overset{ind}{\sim} N\left(\theta_i, v_i\right)$	$y_i \mid \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \sigma_i^2\right)$	$y_i \mid \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \sigma_i^2\right)$
$\theta_i \mid \mu, \beta, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \beta x_i, \tau_u^2\right)$	$\theta_i \mid \mu, \beta, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \beta x_i, \tau_u^2\right)$	$\theta_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \beta, \tau_u^2 \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k N\left(\mu_k + \beta x_i, \tau_u^2\right)$
	$v_i \mid a, \sigma_i^2 \overset{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right)$	$v_i \mid a, \sigma_i^2 \overset{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2\sigma_i^2}\right)$
	$\sigma_i^2 \mid b, \boldsymbol{\gamma} \overset{ind}{\sim} IG\left(2, be^{z_i^T \gamma}\right)$	$\sigma_i^2 \mid \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\pi} \overset{ind}{\sim} \sum_{k=1}^{K} \pi_k IG\left(2, b_k e^{z_i^T \gamma}\right)$
$\beta \mid \lambda_\beta \sim N\left(0, \lambda_\beta^{-1}\right)$	$\beta \mid \lambda_\beta \sim N\left(0, \lambda_\beta^{-1}\right)$	$\beta \mid \lambda_\beta \sim N\left(0, \lambda_\beta^{-1}\right)$
$\mu \mid \lambda_\mu \sim N\left(0, \lambda_\mu^{-1}\right)$	$\mu \mid \lambda_\mu \sim N\left(0, \lambda_\mu^{-1}\right)$	$\mu_k \mid \lambda_\mu \overset{iid}{\sim} N\left(0, \lambda_\mu^{-1}\right)$
	$\boldsymbol{\gamma} \mid \Sigma \sim \mathbf{N}_p\left(0, \Sigma\right)$	$\boldsymbol{\gamma} \mid \Sigma \sim \mathbf{N}_p\left(0, \Sigma\right)$
		$\boldsymbol{\pi} \mid \alpha \sim Dir\left(\alpha/K, ..., \alpha/K\right)$
	$\tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1,1),$	$\alpha, \tau_u^{-2}, \lambda_\beta, \lambda_\mu . b_k \sim G(1,1),$
$\tau_u^{-2}, \lambda_\beta, \lambda_\mu \sim G(1,1)$	$\log(a) \sim t_3(0,1),$	$\log(a) \sim t_3(0,1),$
	$\log(b) \sim t_3(0,1)$, $prior(\Sigma)$	$prior(\Sigma)$

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, $\left(\mu_k, b_k\right)$. That is, each mixture / cluster component in the joint distribution for (θ_i, σ_i^2) share the same π_k.

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 contains a summary of the three models considered in this paper (formulated for a single covariate x_i, for simplicity.)

3. Model fit comparison

We applied the models introduced in Section 2 to CES estimates of employment for the period from October 2008 through September 2009. The quality of the employment estimates can be assessed several months after their CES-based publication by comparing the estimates 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.

CES domains are defined by intersections of industry and geography: industries in the CES survey are defined by the North American Industry Classification System (NAICS); the geographic resolution considered in this paper is the State level. Since the direct CES 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. A more detailed description of the CES estimation procedures can be found in chapter 2 of Bureau of Labor Statistics (2004).

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, $i = 1, ..., N$. The population ratio, $R_{i,t}$, is the target employment change, defined as

$$R_{i,t} = \frac{Y_{i,t}}{Y_{i,t-1}}, \tag{9}$$

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

$$\hat{r}_{i,t} = \frac{\sum_{j \in s_t^{(i)}} w_j y_{jt}}{\sum_{j \in s_t^{(i)}} w_j y_{j,t-1}}, \tag{10}$$

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, $R_{i,t}$. Monthly ratios $\hat{R}_{i,t}$, along with their respective sampling variances $v_{i,t}$, constitute the domain-level data supplied for the modeling.

Estimates of employment levels for month t are obtained by multiplying estimated previous month level, $\hat{Y}_{i,t-1}$, by the estimate of the relative employment change:

$$\hat{Y}_{i,t} = \hat{Y}_{i,t-1}\hat{R}_{i,t}. \tag{11}$$

The corresponding estimate of the over-the-month employment change is

$$\Delta\hat{Y}_{i,t} = \hat{Y}_{i,t} - \hat{Y}_{i,t-1}. \tag{12}$$

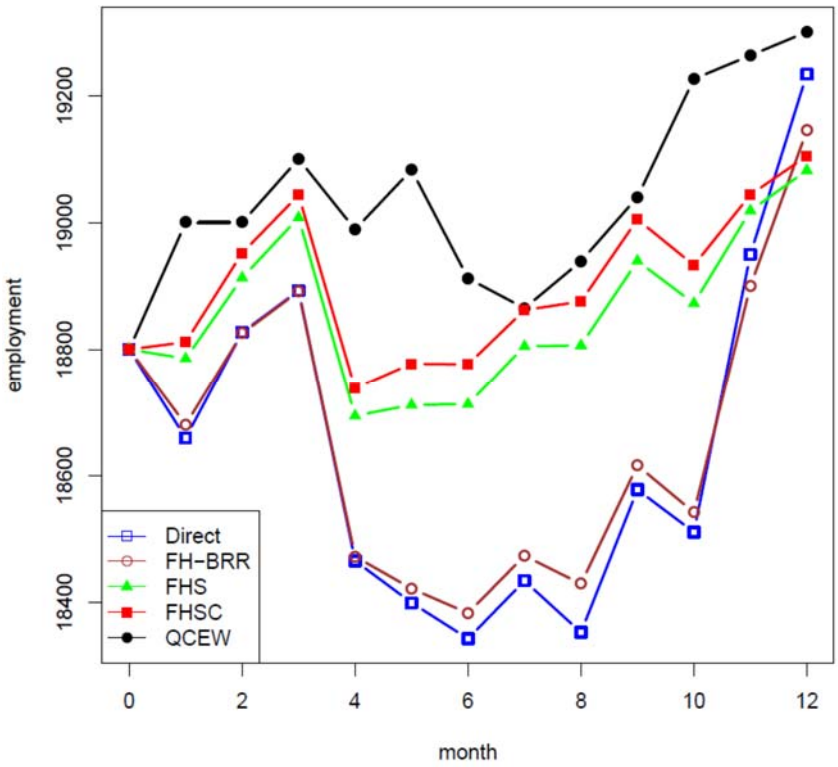
Every year, the estimation cycle starts at month 0 from a known QCEW-based employment level $Y_{i,0}$ and after twelve months the CES estimated employment level $\hat{Y}_{i,12}$ is compared to the QCEW employment levels. Once a year, the CES estimated levels are revised to reflect newly available QCEW levels (in a procedure commonly known as the annual revision).

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 quality of CES estimates could be judged by their proximity to employment from QCEW (which is considered a gold standard due to the absence of sampling error). However, 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. To overcome this difficulty related to monthly comparisons of CES and QCEW and to facilitate focused comparisons to true population figures, we provide results from the synthetic data. We created synthetic data by adding Student's t distributed 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 month-by-month comparisons. The

relative fit performances of our candidate models are the same on both the real and synthetic data sets because we use the QCEW to compose the synthetic data.

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



In Table 2, we present mean absolute deviation (MAD) averaged over domains and months, as follows:

$$MAD = N^{-1}12^{-1}\sum_{i=1}^{N}\sum_{t=1}^{12}\left|\Delta\tilde{Y}_{i,t} - \Delta Y_{i,t}\right|,$$

where $\Delta\tilde{Y}_{i,t} = \tilde{Y}_{i,t} - \tilde{Y}_{i,t-1}$, $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$, and $\tilde{Y}_{i,t}$, $\tilde{Y}_{i,t-1}$ signify estimates based on the sample or on a model.

Results in Table 2 suggest that all model-based estimates are more efficient than the direct sample estimates. The models that jointly model the point estimates and the variances, FHS and FHSC, perform similarly to one another, but notably better than FH-V, which separately models the point estimates and the variances because the point estimates and variances are dependent.

Sampling variances fitted using our two 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. Symbols (empty and filled circles and stars) correspond to domains and

show estimated variances versus true variances; stars represent the direct estimates of variances; empty circles show estimated variances from the FHS model; filled circles correspond to variance estimates from FHSC. The closer the symbols are 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 FHSC model variance estimates lie along the 45-degree line.

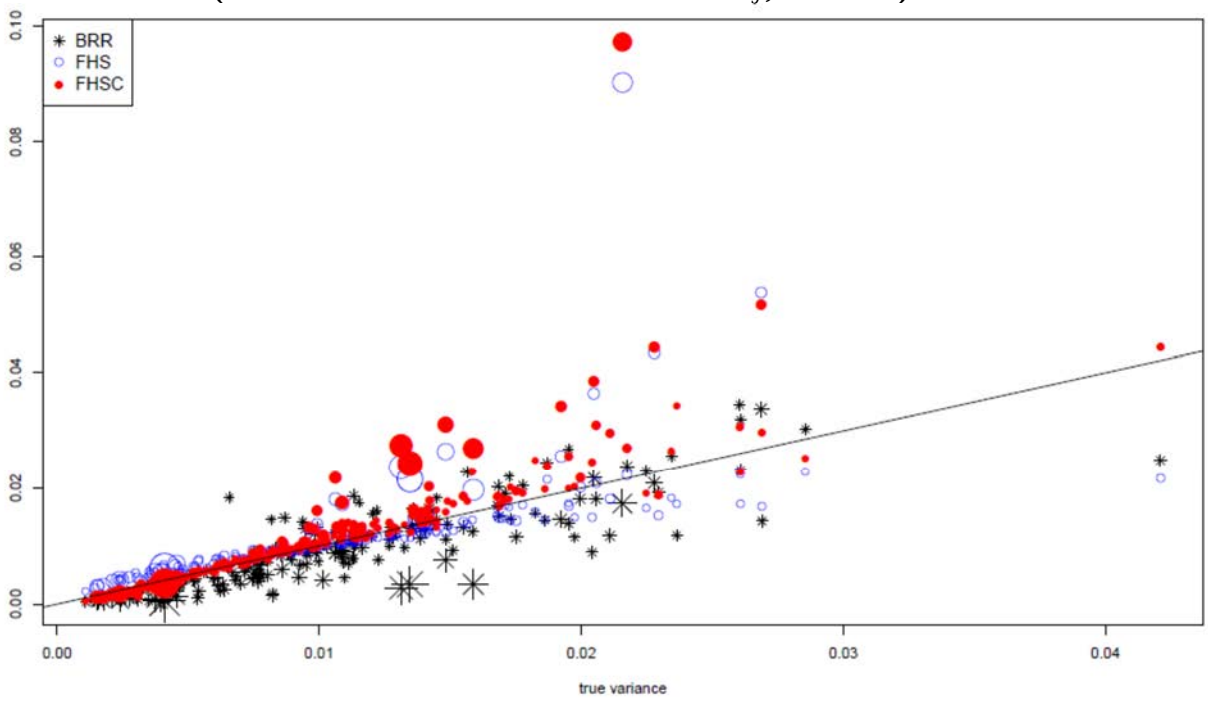
Table 2. Simulated data, over-the-months results

Ind	N	Direct	FH	FH-V	FHS	FHSC
1000	600	285	220	205	188	204
2000	1692	795	503	507	444	448
3100	2808	328	275	279	273	277
3200	1680	275	213	185	187	192
4100	1488	396	252	234	213	221
4200	3432	613	350	315	278	276
4300	2328	456	292	300	265	266
5000	996	282	206	187	188	194
5500	1788	391	267	232	221	216
6054	1800	481	314	314	290	292
6055	540	213	176	162	158	167
6056	1380	923	689	642	601	616
6561	708	751	564	528	512	519
6562	2568	444	297	289	250	251
7071	708	771	470	480	435	459
7072	960	770	578	551	506	510
8000	1320	639	377	355	331	340
Overall	**26796**	**511**	**344**	**328**	**302**	**306**

The sizes of the symbols 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 circles on the upper edge of the plot. These circles correspond to different estimators for the same domain. The sizes of the circles suggests that the domain has an outlying value of the direct point estimate. The location of the circles indicate that the variance is overestimated by both models for this outlying domain. 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 two 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 (4), they may enlarge the estimated variance, particularly for the FHS model that imposes a single component normal distribution as the prior for random effects. Therefore, it is important to study robustness of the models to deviations from 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.

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



4. Model robustness study for deviations from linearity

The purpose of the simulation exercise described in this section is to study how the proposed joint models behave 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 $i = 1,...,100$ domains, we generate estimation targets θ_i as

$$\theta_i = \mu + \beta x_i + u_i, \tag{13}$$

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

$$y_i = \theta_i + e_i, \tag{14}$$

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

$$\sigma_i^2 \sim IG(\lambda_g + 1, \lambda_g b). \tag{15}$$

Generating true variances σ_i^2 are not observed directly (or available for subsequent modeling). We simulate observed estimates of variances as

$$v_i \sim G\left(3, 3\frac{1}{\sigma_i^2}\right). \quad (16)$$

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

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

Table 3: Properties of the credible intervals, over 95 domains with $\mu = 0$

$\left[b, \lambda_g\right]$	**FH**	**FH-V**	**FHS**	**FHSC**
Coverage (0.95 nominal)				
[0.5, 8]	0.914	0.957	0.933	0.933
[0.5, 4]	0.915	0.952	0.934	0.935
[0.5, 1]	0.921	0.957	0.943	0.951
[1, 8]	0.922	0.963	0.951	0.938
[1, 4]	0.923	0.960	0.951	0.942
[1, 1]	0.927	0.963	0.957	0.952
[1.5, 8]	0.928	0.971	0.960	0.944
[1.5, 4]	0.930	0.968	0.959	0.946
[1.5, 1]	0.933	0.970	0.965	0.956
Length				
[0.5, 8]	2.259	2.392	2.259	2.293
[0.5, 4]	2.208	2.384	2.234	2.244
[0.5, 1]	2.004	2.371	2.186	2.069
[1, 8]	2.933	3.102	2.881	2.966
[1, 4]	2.869	3.080	2.846	2.897
[1, 1]	2.599	3.060	2.725	2.642
[1.5, 8]	3.445	3.659	3.366	3.435
[1.5, 4]	3.364	3.625	3.316	3.353
[1.5, 1]	3.027	3.595	3.135	3.036

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

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

Average MSE over all 100 domains is $MSE\left(\hat{\theta}\right)=\frac{1}{100}\sum_{i=1}^{100}MSE_i\left(\hat{\theta}\right)$. We also compute average MSE separately over a set of domains with $\mu=0$, $MSE_{\mu=0}\left(\hat{\theta}\right)=\frac{1}{95}\sum_{i=1}^{95}MSE_i\left(\hat{\theta}\right)$, and over a set of domains with $\mu=3$, $MSE_{\mu=3}\left(\hat{\theta}\right)=\frac{1}{5}\sum_{i=96}^{100}MSE_i\left(\hat{\theta}\right)$.

Table 4: Properties of the credible intervals, over 5 domains with $\mu=3$

$\left[b,\lambda_g\right]$	FH	FH-V	FHS	FHSC
Coverage (0.95 nominal)				
[0.5, 8]	0.678	0.658	0.330	0.706
[0.5, 4]	0.702	0.654	0.342	0.694
[0.5, 1]	0.756	0.614	0.332	0.680
[1, 8]	0.676	0.654	0.482	0.706
[1, 4]	0.690	0.642	0.474	0.710
[1, 1]	0.734	0.632	0.468	0.692
[1.5, 8]	0.708	0.706	0.566	0.718
[1.5, 4]	0.726	0.704	0.566	0.702
[1.5, 1]	0.728	0.694	0.556	0.698
Length				
[0.5, 8]	2.241	2.390	2.564	2.443
[0.5, 4]	2.231	2.388	2.575	2.448
[0.5, 1]	2.006	2.371	2.551	2.377
[1, 8]	2.927	3.118	3.094	3.071
[1, 4]	2.880	3.072	3.081	3.040
[1, 1]	2.608	3.062	2.977	2.819
[1.5, 8]	3.428	3.650	3.515	3.508
[1.5, 4]	3.378	3.611	3.489	3.467
[1.5, 1]	3.038	3.603	3.342	3.194

Coverage probabilities and interval lengths for 95% nominal credible intervals for the fitted values based on all the models are presented in Table 3 (for $\mu=0$ domains) and Table 4 (for $\mu=3$ domains.) Coverages are derived for each domain over 100 simulations. The domain results are averaged over respective groups of domains: $\bar{c}_{\mu=0}=\frac{1}{95}\sum_{i=1}^{95}\hat{c}_i$ and $\bar{c}_{\mu=3}=\frac{1}{5}\sum_{i=96}^{100}\hat{c}_i$, where $\hat{c}_i=\frac{1}{S}\sum_{s=1}^{S}I\left\{\hat{q}_{i,0.025}\le\theta_i\le\hat{q}_{i,0.975}\right\}$ 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 $\hat{l}_i=\hat{q}_{i,0.975}-\hat{q}_{i,0.025}$ and S denotes the number of Monte Carlo simulation iterations.

For the $\mu = 0$ domains, coverages for joint models are close to nominal. The FH coverages are somewhat low, especially for lower variances scenarios of $\left[b = 0.5, \lambda_g\right]$ and $\left[b = 1, \lambda_g\right]$. The coverages for the FH-V model are slightly higher than the nominal; their average interval lengths are longer than in the other models.

The model coverages for $\mu = 3$ domains are low under all of the models, with the lowest coverages for the FHS 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.

5. Improved handling of domains

Although the CES survey uses models to produce estimates for 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.

We earlier demonstrated 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 the survey analyst concludes that there are input errors in the nominated or flagged domains, the errors will be corrected. If there are not input errors, 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 $\mathcal{M}_1, ..., \mathcal{M}_W$ 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, $\mathcal{M}_w$, is accepted for production. The question remains, what if the selected model $\mathcal{M}_w$ 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 nominated suspect domains. The method is based on the Bayesian multiple hypotheses testing and posterior predictive checking.

For a given model $\mathcal{M}_w$ over the space of candidate models indexed by $w = 1,...,W$, let $y_i^l, l = 1,...,L$ be replicate data draws from posterior predictive distribution $p\left(y_i^l \mid y_i, \mathcal{M}_w\right)$ 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 $p\left(y_i^l \mid y_i, \mathcal{M}_w\right)$.

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

Let set D denote 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 $F = E\left[p_i \mid i \in D, \mathcal{M}_w\right]$ and the estimated F is computed from the average,

$$\hat{F} = \frac{1}{|D|}\sum_D p_i \,. \tag{17}$$

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.
The algorithm follows:

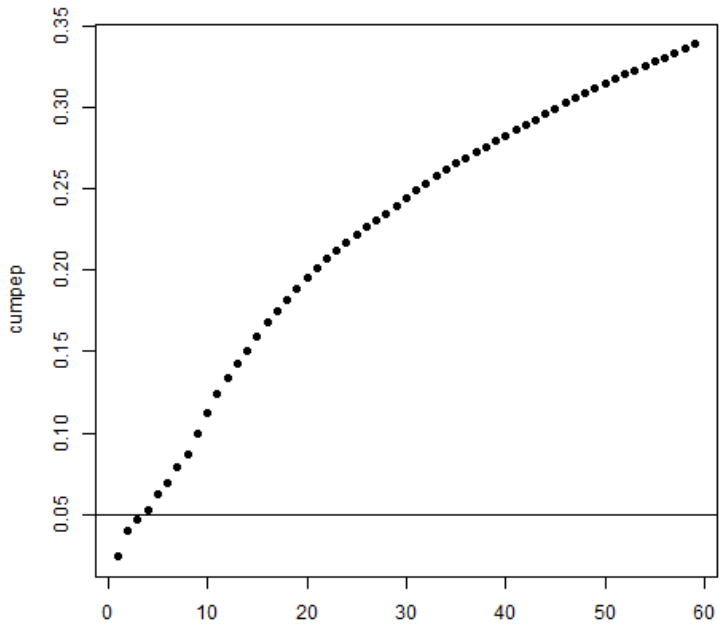
1. Sort p_i's in the ascending order, $p_{(1)} \le ... \le p_{(N)}$ and compute the cumulative mean.
2. An example of the plot of the cumulative mean is presented in Figure 3. We may review the plot to think of what the reasonable q-value could be. Or we may just set the q-value once in advance.

3. Suppose we choose $q = 0.05$. Then D will consist of the first d domains with smallest p_i's: $p_{(1)} \le ... \le p_{(d)}$, such that

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

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

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



We applied this test to the data from the simulation study considered in Section 4 and created a list of "discoveries" to be sent for the review by analysts. Since this review is not available in our simulations, we make a favorable assumption that analysts make correct decisions of whether an estimate on the list is an outlier or a true (deviation from linearity) phenomenon. Namely, we assume that all "discoveries" from the set of the 95 domains generated with $\mu = 0$ were attributed to a "bad sample" cause and that the analyst's decision was to use modeled estimates for these domains; all "discoveries" from the set of the domains generated with $\mu = 3$ were attributed to the failure of model's linearity assumption and the direct sample estimates were used for such domains, instead of the model estimates.

We first focus on comparing the relative effectiveness of our models to discover deviations form linearity on domains generated with $\mu = 3$. The resulting MSEs after the replacement, for different levels of thresholds, are given in Table 5.

Table 5: MSE for domains with $\mu = 3$

				$q = 0.05$		$q = 0.10$		$q = 0.15$	
$[b, \lambda_g]$	**Y**	**FHS**	**FHSC**	**FHS***	**FHSC***	**FHS***	**FHSC***	**FHS***	**FHSC***
[0.5, 8]	0.511	3.719	1.331	1.449	1.070	0.849	0.911	0.646	0.771
[0.5, 4]	0.556	3.851	1.419	1.603	1.135	0.842	0.948	0.681	0.807
[0.5, 1]	0.361	3.892	1.589	1.636	1.096	0.618	0.741	0.441	0.526
[1, 8]	1.023	3.350	1.952	2.442	1.891	1.741	1.742	1.374	1.506
[1, 4]	1.112	3.469	2.032	2.615	1.940	1.893	1.771	1.442	1.558
[1, 1]	0.721	3.375	1.901	2.467	1.678	1.367	1.350	0.920	1.006
[1.5, 8]	1.534	3.392	2.478	3.052	2.525	2.479	2.407	2.002	2.098
[1.5, 4]	1.668	3.484	2.538	3.228	2.572	2.675	2.464	2.157	2.243
[1.5, 1]	1.082	3.316	2.303	2.815	2.168	2.023	1.854	1.388	1.437

The first columns of Table 5 show MSEs of the "direct" estimates Y and the estimates based on the FHS and FHSC models. Columns labeled FHS* and FHSC* show MSE results after analysts correctly replace domains from group $\mu = 3$ nominated under each model by the direct sample estimates. The set of nominated or discovered domains for the FHS* and FHSC* columns were created using, respectively, the FHS- and FHSC-based screening procedures with the same respective threshold levels. We show results for threshold choices of 0.05, 0.10, and 0.15. As can be seen in Table 5, the correct replacement of discoveries with the direct estimates leads to visible reductions in MSE; however, the values of MSE are still higher than the respective MSEs for the "direct" estimates given in the column labeled Y. The reason is that not all the deviations from linearity were captured by the test under the chosen threshold levels, i.e., there remained domains with $\mu = 3$ that were not captured by the screening and thus their respective (overly shrunken) model-based estimates were not replaced by the direct estimates.

Table 6 displays the fraction of "true discoveries", defined as the number of correct discoveries divided by the total number of the $\mu = 3$ domains, i.e. of true deviations from linearity, at increasing threshold levels. Table 6 shows that higher threshold levels increase the number of domains correctly discovered. This increase in discoveries, in turn, would reduce MSEs for those correctly replaced domains (Table 5); however, investigating the added discoveries would also increase the workload for analysts. Table 7 shows the fraction of "false discoveries", defined as the number of discoveries among the domains where there is no deviation from linearity, e.g., $\mu = 0$, divided by the total number of such domains. By construction, the false discovery rate increases with an increasing threshold, which could result (after analyst investigation) in more domains to be mislabeled as "deviations from linearity assumptions", when in fact their appearance on the list could be due to poor direct estimates (that induce outlyingness). In practice, tuning will be required to set the threshold, taking into consideration the workload and timeline restrictions.

We note that the FHSC produces a lower discovery rate. This lower discovery rate is expected, since FHSC is a more flexible model than FHS in the sense of adapting to deviations from linearity. It is able to better model some of these domains by allocating them to a cluster, which reduces the shrinking of these domain estimates. The flexible estimation property of the FHSC model is also evidenced by the MSE results for the $\mu = 3$ domains shown in Table 5, where MSE values are lower for the FHSC model compared to

the FHS model; in particular, results for the $q = 0.05$ threshold are better for FHSC*, as compared to FHS*, even though the discovery rate is lower for the FHSC model; similarly, results for the $q = 0.10$ and $q = 0.15$ threshold levels are close, even though FHSC has much lower discovery rate. Some of the $\mu = 3$ domains are less shrunken under FHSC (than FHS) and are, therefore, better predicted by the model and, hence, not "discovered". That our testing procedure produces fewer discoveries under FHSC for modeled domains, but yet FHSC produces relatively lower errors is a feature of this model.

Table 6: Percent of "true discoveries": for the $\mu = 3$ domains

	$q = 0.05$		$q = 0.10$		$q = 0.15$	
$[b, \lambda_g]$	**FHS***	**FHSC***	**FHS***	**FHSC***	**FHS***	**FHSC***
[0.5, 8]	37%	8%	66%	20%	81%	40%
[0.5, 4]	35%	10%	68%	24%	82%	43%
[0.5, 1]	32%	8%	69%	26%	84%	47%
[1, 8]	18%	5%	48%	19%	66%	40%
[1, 4]	17%	7%	48%	23%	70%	44%
[1, 1]	15%	5%	49%	21%	72%	45%
[1.5, 8]	11%	4%	37%	21%	59%	42%
[1.5, 4]	11%	6%	37%	24%	61%	45%
[1.5, 1]	9%	4%	37%	21%	65%	47%

Table 7: Percent of "false discoveries": for the $\mu = 0$ domains

	$q = 0.05$		$q = 0.10$		$q = 0.15$	
$[b, \lambda_g]$	**FHS***	**FHSC***	**FHS***	**FHSC***	**FHS***	**FHSC***
[0.5, 8]	1%	0%	8%	0%	18%	2%
[0.5, 4]	1%	0%	7%	1%	18%	3%
[0.5, 1]	1%	0%	7%	1%	18%	3%
[1, 8]	1%	0%	5%	1%	14%	4%
[1, 4]	1%	0%	5%	1%	14%	5%
[1, 1]	1%	0%	5%	2%	13%	5%
[1.5, 8]	0%	0%	4%	1%	13%	6%
[1.5, 4]	0%	0%	4%	2%	13%	7%
[1.5, 1]	1%	0%	4%	2%	13%	7%

We, next, assess the effectiveness of our test procedure when it is applied to *non-modeled* domains as a screening tool of sample-based, direct estimates before they are released for publication. Here, the goal is to detect the domains where direct estimates are impacted by poor sample or extreme sample measurements. For this purpose, we use the same simulation and test results as in Tables 5-7 but change the focus of the evaluation to compare the original "direct" estimates with the "corrected" estimates following analysts' review of the list. In the actual production, we expect analysts to identify extreme measurements and errors and update domain sample-based estimates after making appropriate corrections. In this simulation study, we do not have the analysts review stage; in place of the review, we use the following assumptions and approximations: assume analysts correctly identify the flagged domains in the $\mu = 0$ group as being affected by

sample outliers; for those flagged domains in the $\mu = 0$ group, we use respective model-based estimates to replace the original "direct" estimates as proxy of what the estimates will look like after analysts' treatment of outliers. We compute the MSE of the original direct-based estimates and the revised estimates after replacements. The results are reported in Table 8.

Table 8: MSE for domains with $\mu = 0$

		$q = 0.05$		$q = 0.10$		$q = 0.15$	
$[b, \lambda_g]$	**Y**	**Y_FHS***	**Y_FHSC***	**Y_FHS***	**Y_FHSC***	**Y_FHS***	**Y_FHSC***
[0.5, 8]	0.508	0.500	0.506	0.461	0.501	0.419	0.484
[0.5, 4]	0.513	0.502	0.510	0.463	0.495	0.420	0.472
[0.5, 1]	0.482	0.449	0.467	0.404	0.429	0.369	0.391
[1, 8]	1.016	0.984	1.013	0.864	0.976	0.724	0.896
[1, 4]	1.025	0.981	1.014	0.848	0.956	0.710	0.863
[1, 1]	0.963	0.873	0.910	0.711	0.800	0.585	0.690
[1.5, 8]	1.524	1.487	1.513	1.276	1.421	1.038	1.251
[1.5, 4]	1.538	1.472	1.511	1.241	1.369	1.016	1.193
[1.5, 1]	1.445	1.301	1.339	1.023	1.121	0.808	0.923

Column Y shows MSE for the original "direct" estimates. Columns Y_FHS* and Y_FHSC* show MSEs after the original estimates for the domains on the list where replaced by the respective model-based estimates, used as approximation of estimates smoothed over the presumed outliers. We observe that the increasing of the threshold levels results in more flagged domains; the replacement of respective original estimates by more smooth model-based values results in improved estimation. We also observe that FHS-based procedure gives slightly better results, compared with the FHSC-based counterpart. Since FHS does not cluster extreme values – be they deviations from linearity or extreme sample measurements – this model will tend to produce more discoveries. This greater sensitivity of FHS relative to FHSC could also lead to false discoveries. The discoveries may also include a more equal mix deviations from linearity and extreme sample measurements than under FHSC (which may better estimate some domains expressing deviations from linearity). The more equal mix produces a greater reliance on the analyst to correctly differentiate the two phenomena.

6. Summary

In this paper, we applied joint modeling of the point estimates and their variances to the synthetic 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 Sugasawa 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 without clusters, 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.

Bureau of Labor Statistics (2004), Employment, Hours, and Earnings from the Establishment Survey, BLS Handbook of Methods, chap. 2, Washington, DC: US Department of Labor. Available at http://www.bls.gov/opub/hom/pdf/homch2.pdf, Last accessed on May 1, 2018.

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.