Bayesian Nonparametric Joint Model for Point Estimates and Variances under Biased Domain Variances November2019

Julie Gershunskaya, Terrance D. Savitsky[1]

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

Abstract: We propose a joint model for point estimates and their variances when observed variances may contain bias. The bias in variances for groups of domains may be induced by an estimation procedure, such the weight smoothing procedure of Beaumont (2008) to compute a domain point estimator. While the weight-smoothed point estimator is more efficient than the original weighted survey estimator, its variance estimation procedure requires truncations that induces bias in the domain variance estimator. The proposed formulation generalizes the joint point estimator and variance models to explicitly parameterize a multiplicative bias in observed variances under a nonparametric formulation that allows the data to discover distinct bias regimes. As a consequence of the better variance estimation, domain point estimates are more robustly estimated under a joint model for the domain point estimates and their associated variances. We compare the performances of alternative models in application to estimates from the Current Employment Statistics survey and in simulations.

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

1. Introduction

Estimates for smaller domains are often made by employing models that allow for "borrowing strength" across multiple domains to produce more efficient domain point estimation. The Fay-Herriot (FH) model provides a relatively simple formulation that produces more efficient domain-level estimation. The success of the FH model depends on the existence of good covariates that allow for a quality description of the underlying process, which customarily is formulated as a regression model; a normally distributed "random effects" term expresses deviations of the signal around the regression line.

In the classical FH formulation, the variances of direct sample estimates are assumed to be fixed and known to researchers. In practice, they are estimates and contain noise. Instead of assuming that variances of direct sample estimates are "fixed and known", Maiti et al (2014) and Sugasawa et al. (2017) proposed models to simultaneously fit point estimates and their variances; a few alternative formulations were considered in Gershunskaya and Savitsky (2018). The joint treatment of point estimates and variances is a more efficient way of modeling, compared to the classical FH, because it exploits the relationship between point estimates and their variances.

In this paper, we focus on the quality of domain-indexed survey sampling variances, which provide the estimation model information on the quality of direct sample inputs. Our goal is to allow more flexibility in their estimation in order that we may more accurately assess the quality of the direct sample inputs, which would be expected to improve the quality of small domain estimations.

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

It is clear that the quality of the resulting model estimates depends on the quality of the inputs; both the domain point estimates and their estimated variances. We begin our assessment of the quality of input variances with a pre-modeling step where we adopt Beaumont's (2008) generalized design-based paradigm and treat survey weight as random variables. This approach permits smoothing survey weights before applying them in the expansion estimator formula. The theory states that the resulting "smooth weight" (SW) estimator is unbiased with respect to a generalized design approach (that includes the distribution over survey samples and the random variable used to generate sampling weights). The SW estimator produces lower domain variance inputs than the traditional estimator that uses the original survey weights (in what follows, we use abbreviation UW to refer to the estimator based on the original "unsmoothed" weights). The approach was applied by Gershunskaya and Sverchkov (2014) for the sample-based domain estimation in the Current Employment Statistics (CES) survey conducted by the Bureau of Labor Statistics. We start with and use smoothed weights for improved CES domain point estimates and variances as inputs in our domain-level modeling.

Beaumont (2008) proposes two ways for estimating the variance of the SW estimator. The first approach uses an additive (negative) adjustment to the traditional design variance of the UW estimator. The adjustment can be estimated using the bootstrap. However, this approach can lead to negative values: in the CES application, we observed that a large portion of domain estimates resulted in negative variance estimates that had to be truncated to zero. Therefore, we used the second approach proposed in the same paper. We describe it in more detail in Section 2. This approach, however, also employs ad hoc truncations to avoid negative variance estimates or squared biases. As a result, variance estimates input to our domain-level modeling for the SW estimator may be of low quality for a subset of domains.

In this paper, we devise a more flexible model for the variances in application to the SW point estimate and variance inputs, both the original UW and smoothed SW variance estimates, with the goal to obtain a more accurate estimator. The paper is organized as follows. In Section 2, we briefly describe the Beaumont's variance estimation approach for the SW estimator. We introduce our joint models for domain point and variance estimation in Section 3. The bias adjustment property is discussed in Section 4. The CES application results are given in Section 5. A simple simulation study in Section 6 is designed as an illustration to help explain the observed CES results.

2. Sample estimation with the weight smoothing

We refer the reader to Gershunskaya and Sverchkov (2014) for the description of the CES estimator and the weight smoothing procedure.

We next describe the variance estimation approach used to compute the SW variances that we propose to use for domain modeling in the sequel. The method for SW variances is based on Beaumont (2008). In Beaumont's notation, variance of the SW estimate $\hat{T}^{SW}$ of finite population target T is

$$V\left(\hat{T}^{SW} \mid Y\right) = E\left\{\left(\hat{T}^{SW} - T\right)^2 \mid Y\right\} = E\left\{V\left(\hat{T}^{SW} \mid Z,Y\right) + B^2\left(\hat{T}^{SW} \mid Z,Y\right) \mid Y\right\},$$

where $B\left(\hat{T}^{SW} \mid Z,Y\right) = E\left(\hat{T}^{SW} - T \mid Z,Y\right)$, Y is a response variable and Z is a variable used to determine the inclusion probabilities (i.e., conditioning on both Z and Y denotes the classical design distribution, conditioning on Y only denotes the generalized Beaumont's approach where weights are treated as random variables.)

Thus, the variance is estimated by

$$\hat{V}\left(\hat{T}^{SW} \mid Y\right) = \hat{V}\left(\hat{T}^{SW} \mid Z,Y\right) + \hat{B}^2\left(\hat{T}^{SW} \mid Z,Y\right). \tag{1}$$

Note that both parts on the right hand side of (1) are expressed with respect to the traditional design approach, i.e. where the weights are treated as fixed. The first term in (1), $\hat{V}\left(\hat{T}^{SW} \mid Z,Y\right)$, is the estimate of variance where the smooth weights are fixed (through Z) and treated as if they were the design weights. This quantity may be obtained with the usual methods for survey variance estimation. In the CES application, we used Balanced Repeated Replication (BRR) resampling method. The squared bias term, $\hat{B}^2\left(\hat{T}^{SW} \mid Z,Y\right)$, may be estimated as $\left(\hat{T}^{SW} - \hat{T}^{UW}\right)^2$, where $\hat{T}^{UW}$ is an estimator based on the original survey weights. However, this would be a biased estimate of the squared bias, since

$$E\left[\left(\hat{T}^{SW} - \hat{T}^{UW}\right)^2 \mid Z,Y\right] = \left[E\left(\hat{T}^{SW} - \hat{T}^{UW} \mid Z,Y\right)\right]^2 + V\left(\hat{T}^{SW} - \hat{T}^{UW} \mid Z,Y\right).$$

Thus, the unbiased estimate of the squared bias is

$$\hat{B}^2\left(\hat{T}^{SW} \mid Z,Y\right) = \left(\hat{T}^{SW} - \hat{T}^{UW}\right)^2 - \hat{V}\left(\hat{T}^{SW} - \hat{T}^{UW} \mid Z,Y\right). \tag{2}$$

Note that the second term in (2), $\hat{V}\left(\hat{T}^{SW} - \hat{T}^{UW} \mid Z,Y\right)$, may be obtained from the same BRR procedure as $\hat{V}\left(\hat{T}^{SW} \mid Z,Y\right)$. Since (2) can produce negative values, Beaumont (2008) recommended to replace negative values by zeros, so that the estimate of the squared bias is

$$\hat{B}^2 = \max\left\{0, \left(\hat{T}^{SW} - \hat{T}^{UW}\right)^2 - \hat{V}\left(\hat{T}^{SW} - \hat{T}^{UW} \mid Z,Y\right)\right\}.$$

Finally, according to the theory, the variance of the SW is no greater than the variance of the UW estimator. Thus, another truncation sets the variance of SW to be equal to the variance of UW in cases where it happens to go above the latter:

$$\hat{V}\left(\hat{T}^{SW} \mid Y\right) = \min\left\{\hat{V}\left(\hat{T}^{UW} \mid Z,Y\right), \hat{V}\left(\hat{T}^{SW} \mid Z,Y\right) + \hat{B}^2\right\}. \tag{3}$$

As a result of the above mentioned truncations, the quality of the resulting variance estimate may suffer. Therefore, for the model inputs, we consider a case that uses the BRR-based variance of UW in place of the variance of SW to avoid the truncations inherent in the SW variances. We will next see how we parameterize the model for our variances to remove the bias in use of the UW variance for the SW point estimate inputs. It turns out that the model fit based on such variance "bias-corrected" version performs better than when the input variances (3) are used without such correction term.

3. Description of the models

Our Model 1 is 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

$$y_i \mid \theta_i \overset{ind}{\sim} N\left(\theta_i, v_i\right), \tag{4}$$

$$\theta_i \mid \mu, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right). \tag{5}$$

Equation (4) says that survey estimates are unbiased and normally distributed with variance v_i that is assumed to be fixed and known. Equation (5) states the assumption about the underlying process. In particular, we assume that target population values θ_i are normally distributed around regression line $\mu + \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$ is a vector of covariate values for domain i; model parameters $\mu, \boldsymbol{\beta}, \tau_u^2$ are to be estimated from the model.

In practice, true variances of the survey estimates are not known, and v_i represent some estimate of the variance. Since direct survey estimates of the variances contain noise, the traditional approach is to use some sort of smoothing of the variances (for example, based on a generalized variance function, or GVF) before they are used in the modeling.

A more efficient approach to use available direct variance estimates is to fit them simultaneously with the point estimates in a single model. This co-modeling approach was proposed by Maiti et al. (2014) who also presented an EM algorithm for their model. Sugasawa et al. (2017) considered the co-modeling approach within the Bayesian paradigm. We next employ as our Model 2 the revised version of Sugasawa et al. (2017).

In Model 2, we present a “simple” co-modeling formulation (the model is referred to as FHS). Assume the following holds for pairs of direct survey estimates (y_i, v_i) for each domain i:

$$y_i \mid \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \sigma_i^2\right), \tag{6}$$

$$\theta_i \mid \mu, \boldsymbol{\beta}, \tau_u^2 \overset{ind}{\sim} N\left(\mu + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right). \tag{7}$$

$$v_i \mid a, \sigma_i^2 \overset{ind}{\sim} G\left(\frac{a n_i^*}{2}, \frac{a n_i^*}{2\sigma_i^2}\right), \tag{8}$$

$$\sigma_i^2 \mid \boldsymbol{\gamma} \overset{ind}{\sim} IG\left(2, \exp\left(z_i^T \boldsymbol{\gamma}\right)\right). \tag{9}$$

Conditions (6) and (7) are the same as FH’s (4) and (5), except that σ_i^2 signifies an unknown variance parameter. Condition (8) states that direct variance estimates v_i are unbiased and have gamma distribution. The shape and scale parameters of the gamma distribution also depend on an unknown parameter a and domain sample size, n_i (we use standardized values for sample sizes, $n_i^* = \left(n_i - \left\{\min_i n_i - 1\right\}\right) \Big/ \left(\max_i n_i - \min_i n_i\right) \in [0,1]$.) Condition (9) is a model for true variance σ_i^2 : the variance has the inverse gamma distribution and the mean depends on vector of covariates z_i through $\exp\left(z_i^T \boldsymbol{\gamma}\right)$, where $\boldsymbol{\gamma}$ are unknown parameters.

In this paper, we construct a new formulation for a nonparametric model for the variances (not included in Gershunskaya and Savitsky (2018)) designed to capture bias and noise in either the case we use SW variances as inputs or UW variances in lieu of the SW variances as inputs to our models. The former express high rates of truncation that induce bias and noise in subgroups of domains while the latter is positively biased under our use of SW point estimates.

In Model 3 (CFHG), we assume that (6) holds and relax assumptions (7),(8), and (9) of the FHS model by replacing them with the finite mixtures of respective distributions:

$$\theta_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \tau_u^2 \overset{iid}{\sim} \sum\nolimits_{k=1}^{K} \pi_k N\left(\mu_k + \mathbf{x}_i^T \boldsymbol{\beta}, \tau_u^2\right), \tag{10}$$

$$v_i \mid a, \sigma_i^2, \mathbf{b}, \boldsymbol{\pi} \overset{ind}{\sim} \sum\nolimits_{k=1}^{K} \pi_k G\left(\frac{a n_i^*}{2}, \frac{a n_i^*}{2 b_k \sigma_i^2}\right), \tag{11}$$

$$\sigma_i^2 \mid \boldsymbol{\gamma}, \boldsymbol{\pi} \overset{ind}{\sim} \sum\nolimits_{k=1}^{K} \pi_k IG\left(2, \exp\left(z_i^T \boldsymbol{\gamma}_k\right)\right). \tag{12}$$

In (10)-(12), we assume the existence of several underlying processes, each characterized by its own cluster-specific parameters $\mu_k, b_k, \boldsymbol{\gamma}_k$. Parameters b_k may be used to detect bias; for example, the use of UW variances are expected to be positively biased for estimation of the true variances, σ_i^2. Domains with (non-truncated) SW variances that are lower than UW variance would be allocated to cluster, k, with $b_k > 1$. We also expect that a subset of domains will express noisy and inefficient variance estimates due to small sample sizes. In this case the b_k term helps to stabilize estimation of σ_i^2 in a similar manner as a ridge regression term (on the logarithm scale). The $\boldsymbol{\gamma}_k$'s represent regression coefficients for the true variances and allow for different clusters of domains – induced by bias or noise – to express distinct sensitivities to predictor inputs.

We, next, allow the prior distributions to be influenced by predictors, $\mathbf{g}_i$, so that two domains, i_1 and i_2, who have similar predictor values, $\mathbf{g}_{i_1}$ and $\mathbf{g}_{i_2}$ would be assigned a higher probability to cluster together, a priori. These predictors specifically influence how domains co-cluster. Using predictors to influence the prior probability of co-clustering is another way to use predictors beyond a regression relationship with the response variable.

In CES, for example, information on average earnings in domains may inform on the structure of the clusters. Let $\mathbf{g}_i$ denote an l-dimensional vector of such covariates. Assume

$$\mathbf{g}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\pi} \overset{ind}{\sim} \sum\nolimits_{k=1}^{K} \pi_k \mathbf{N}_l\left(\boldsymbol{\mu}_{gk}, \boldsymbol{\Sigma}_g\right). \tag{13}$$

Thus covariates $\mathbf{g}_i$ are used as additional input data in the model. We are not interested in the fitted values of parameters $\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$, but fitting $\mathbf{g}_i$ and including the $\boldsymbol{\mu}_{gk}$ and $\boldsymbol{\Sigma}_g$ in the clusters along with $\mu_k, b_k, \boldsymbol{\gamma}_k$ helps better identify the clustering structure to give sharper estimates.

4. The bias correction property of the multiplicative variance model

We formulate the bias protection property for a particular case of a scalar multiplicative bias term, when all domains are assumed to have similar bias. A multi-cluster bias regime is viewed as a generalization of this single-cluster case.

Suppose the following model holds for domains $i = 1,...,N$:

$$y_i \mid \theta_i, \sigma_i^2 \overset{ind}{\sim} N\left(\theta_i, \sigma_i^2\right), \tag{14}$$

$$\theta_i \mid \tau_u^2 \overset{ind}{\sim} N\left(0, \tau_u^2\right), \tag{15}$$

$$v_i \mid a, b, \sigma_i^2 \overset{ind}{\sim} G\left(\frac{an_i^*}{2}, \frac{an_i^*}{2}\frac{1}{b\sigma_i^2}\right), \tag{16}$$

$$\sigma_i^{-2} \overset{ind}{\sim} G\left(\frac{\beta_\sigma}{2}+1, \frac{\beta_\sigma}{2}\gamma_i\right), \tag{17}$$

and let

$$b^{-1} \overset{ind}{\sim} G\left(\frac{\beta_b}{2}+1`, \frac{\beta_b}{2}\right). \tag{18}$$

Consider the following condition:

$$\frac{\left(y_i-\theta_i\right)^2+\beta_\sigma\gamma_i}{1+\beta_\sigma} \Bigg/ \frac{\left(y_i-\theta_i\right)^2+an_i^*\frac{v_i}{b}+\beta_\sigma\gamma_i}{3+an_i^*+\beta_\sigma} \approx 1 \tag{19}$$

Condition (19) states that observed variances that are assumed to be having bias b do not deviate “too much” from $b\sigma_i^2$.

The following statement holds (see the derivation in Apendix).

Proposition 1. Assume model (14)-(18) and condition (19) hold. Then the conditional distribution of b, after integrating out σ_i^2, given data and other parameters is:

$$b \mid ... \sim IG\left(\frac{N}{2}\left(a\bar{n}^*+\beta_b\right)+1, \frac{N}{2}\left(\beta_b+\left(1+\beta_\sigma\right)\frac{1}{N}\sum_{i=1}^{N}\frac{an_i^*v_i}{\left(\left(y_i-\theta_i\right)^2+\beta_\sigma\gamma_i\right)}\right)\right). \tag{20}$$

For instance, the conditional expected value of b, given data and other parameters, after integrating out σ_i^2, is

$$E(b|\ldots)=w+(1-w)\tilde{b}, \tag{21}$$

where $w=\dfrac{\beta_b}{a\bar{n}^*+\beta_b}$ and $\tilde{b}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\dfrac{\dfrac{n_i^*}{\bar{n}^*}v_i}{\left(\dfrac{(y_i-\theta_i)^2+\beta_\sigma\gamma_i}{1+\beta_\sigma}\right)}$,

$\tilde{b}$ represents a multiplicative bias correction. Since y_i is unbiased, the MSE term in y_i corrects for the bias in v_i by multiplying the ratio into σ_i^2. If "average reliability of estimated variances" $a\bar{n}^*$ is large relative to β_b (which is typical), $a\bar{n}^* >> \beta_b$, then weight w is small, allocating more weight to $\tilde{b}$. If, however, estimated variances are less reliable under small domain sample sizes, such that $a\bar{n}^*$ becomes small, then b would still be between 1 and $\tilde{b}$ but closer to 1 (which shrinks the estimate towards the prior).

5. CES application results

We fit our three models for ten estimation cycles for domains defined by area and industry: for example, the 2008 year cycle starts with October 2008 estimates and ends in September 2009, the last cycle considered, 2017 year cycle, starts in October 2017 and ends in September 2018. Domains are grouped together by major industries and the modeling is performed independently by month and each major industry. In our application, there were from 54 to 725 domain counts in different industries. Some of the domains have sample that is considered large enough and estimates for these domains are published without the use of a modeling. The majority of domains, however, have small samples; estimates for such domains are published using a model. In our research, we use the set of all domains to fit the models. The summary results are reported based on the "domains of interest", i.e. those domains that are designated to be published using a model (see Table 1, column "N" for the "model domain" counts.)

The evaluation is based on comparison of the CES estimates to the "true" employment levels that become available to researchers on a lagged basis from the administrative Quarterly Census of Employment and Wages (QCEW) file. Due to different seasonality patterns between the employment series derived from QCEW data and CES, the most meaningful comparison of the two series is after 12 months of estimation. Mimicking the production setup, we obtain level estimates after 12 months of estimation from monthly ratio estimates $\hat{R}_{i,\tau}$ (based on various models, as well as on the sample): respective monthly ratio estimates are multiplied together and by respective September's starting level, $Y_{i,0}$, that is available to CES at the start of the estimation period:

$$\hat{Y}_{i,12}=Y_{i,0}\prod_{\tau=1}^{12}\hat{R}_{i,\tau}. \tag{22}$$

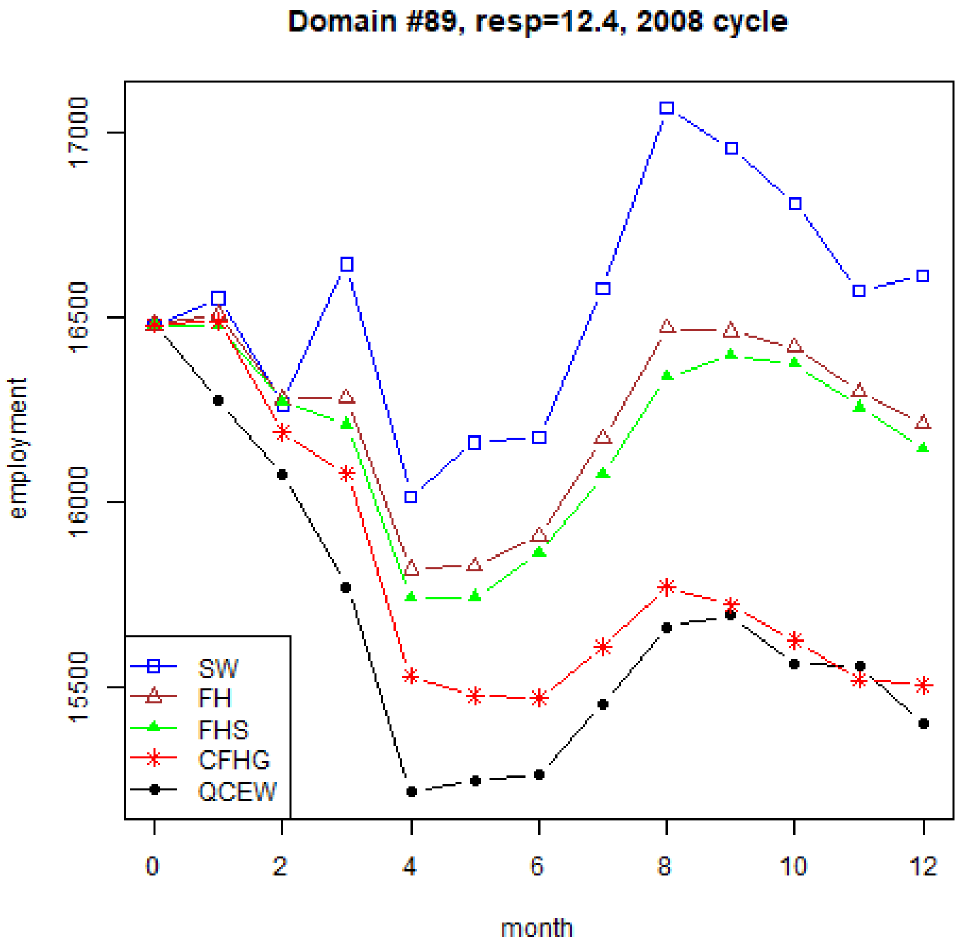

Figure 1: Example of a 12-month CES estimation cycle

For illustration of alternative estimates in conjunction with 12 months of the CES production cycle, in Figure 1, we plot estimates of employment levels for the period from September 2008 to September 2009 for one of the domains in industry 7072 (Hospitality). Black line (with solid dots) represent true population levels from the QCEW source. The other lines show alternative estimators: direct SW and based on FH, FHS, and CFHG models (we present here only the versions where we used UW variances in lieu of the SW variances as inputs.)

In Table 1 we report industry level and overall results for the 2008 cycle. The resulting September 2009 level estimates, $\hat{Y}_{i,12}$, are then compared with the true levels that are available from QCEW after a lag of 6 to 9 months after the reference period. Results for each major industry and overall, presented in Table 1, are based on the mean absolute deviation (MAD):

$$MAD = N^{-1}\sum_{i=1}^{N}\left|\tilde{Y}_{i,12} - Y_{i,12}\right|, \qquad (23)$$

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

Table 1: Real data results for 2008 estimation cycle.

		Direct		*FH*			*FHS*			*CFHG*	
				UW	*SW*	*SW*	*UW*	*SW*	*SW*	*SW*	*SW*
Ind	*N*	*UW*	*SW*	*vrnc UW*	*vrnc SW*	*vrnc UW*	*vrnc UW*	*vrnc SW*	*vrnc UW*	*vrnc SW*	*vrnc UW*
1000	55	725	563	603	519	515	473	448	487	504	529
1500	252	944	683	834	654	657	490	424	405	419	420
2000	91	1555	1036	1035	821	818	905	803	819	826	820
3000	232	930	698	792	671	674	564	524	520	523	516
3100	262	1065	922	1039	920	945	929	914	946	885	892
3200	157	775	627	618	502	500	521	466	466	455	449
4000	43	430	333	402	270	269	313	248	236	250	244
4100	297	655	455	493	398	398	384	354	348	336	342
4200	433	620	520	532	486	485	447	448	441	430	429
4300	381	732	491	637	537	501	391	361	365	357	355
5000	364	494	366	397	343	335	323	300	293	286	288
5500	509	691	529	541	483	480	474	422	418	411	410
6000	200	877	660	638	593	591	485	483	470	473	471
6054	98	977	785	707	646	673	612	623	639	645	650
6055	58	868	550	573	502	487	453	423	421	418	431
6056	180	1558	1395	1118	1270	1233	1099	1320	1292	1354	1309
6500	229	676	531	722	595	621	327	327	319	323	320
6561	50	1881	1477	1982	1700	1653	1529	1267	1272	1071	1132
6562	266	922	781	702	726	745	601	656	653	655	658
7000	235	798	445	395	344	332	356	317	304	311	304
7071	51	1681	1056	873	680	653	747	582	568	596	591
7072	131	849	647	488	491	468	449	503	501	465	463
8000	290	711	550	700	625	658	329	318	314	303	304
Overall	4864	825	629	665	591	590	511	495	491	485	484

In Table 1, we observe that, for every major industry, SW direct estimates have smaller MAD than the estimates based on the original survey weight. The models based on the SW estimates as inputs also have smaller MAD compared to those based on the UW (to save space, we omitted from the Table results for the CFHG based on UW as point estimates.)

FHS and CFHG perform better than the FH model, and CFHG is better than FHS.

Using the UW variance as input, even when SW is used as input point estimate, gives better MAD results.

As noted earlier, we ran the models for 10 estimating cycles. In Table 2 we present the overall results for each year from 2008 to 2017.

In Table 2, we observe that overall results are consistent over the 10 years considered, in that the CFHG model performs slightly better than the FHS models. The last two columns, showing the CFHG results based on SW and the UW variances indicate that the respective MAD results are close.

Table 2: Overall MAD, relative to MAD of SW (after 12 months of each estimation cycle).

Year	*N*	*Direct*		*FH*			*FHS*			*CFHG*	
		UW	*SW*	*UW*	*SW*	*SW*	*UW*	*SW*	*SW*	*SW*	*SW*
				vrnc UW	*vrnc SW*	*vrnc UW*	*vrnc UW*	*vrnc SW*	*vrnc UW*	*vrnc SW*	*vrnc UW*
2008	4,864	825	629	665	591	590	511	495	491	485	484
2009	4,609	784	610	598	543	540	436	412	403	397	398
2010	4,570	776	665	538	519	515	411	410	404	396	399
2011	4,609	664	496	515	457	447	380	354	349	341	341
2012	4,497	641	506	457	423	420	370	351	347	341	339
2013	4,537	618	488	460	412	405	363	339	334	332	332
2014	4,624	598	434	411	362	362	359	333	332	329	330
2015	4,559	589	445	413	374	369	356	337	334	333	333
2016	4,496	610	475	446	409	405	343	328	324	322	324
2017	4,566	616	461	427	379	375	330	307	303	303	302

In Figure 2, we give an example of the distribution of estimated variances (on the log scale) under alternative models and the direct variance estimates (based on industry 7072 at month 6.) In this example, the fitted variance based on the clustering model CFHG is smaller, on average, than the fitted variance based on the FHS model, indicating that the bias/ridge term, b_k, in CFHG has corrected for over-estimation of variances.

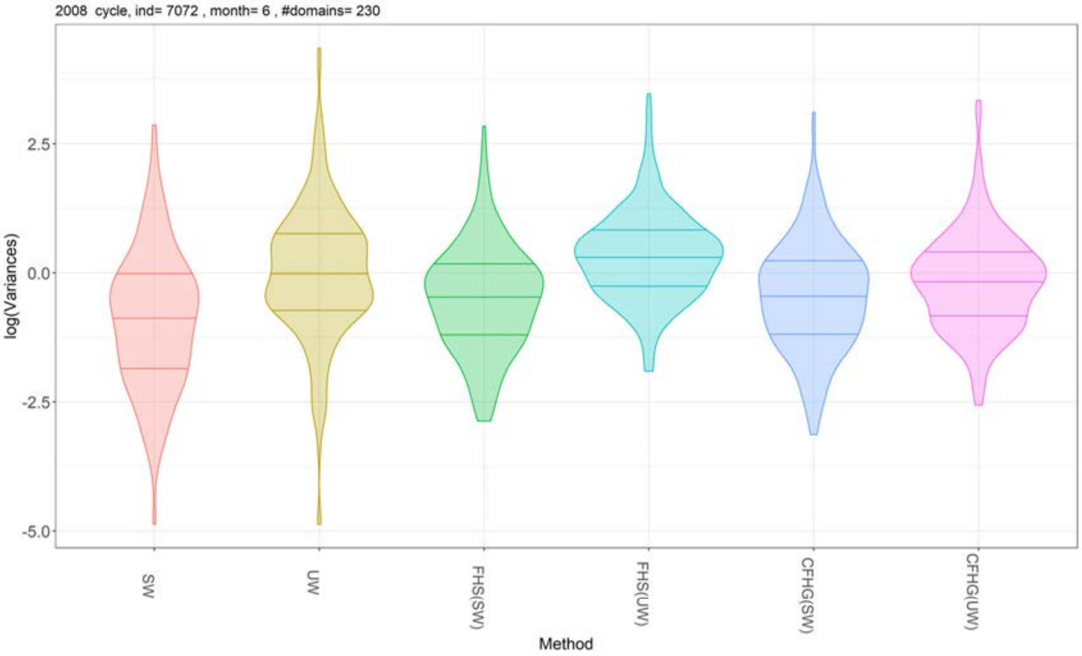

Figure 2: Distribution of log variances

Figure 3 plots FHS- and CFHG-based fitted variances of the direct point estimates versus the "observed" variance of the direct sample UW that was used as the input in both models (values are plotted on the log scale.) We also put estimated variances of SW on the same scatter plot (black dots). Note that SW variances lie on or below the 45 degree line. A large number of dots lying on the line indicates that there was a large portion of truncated variances (since we force SW variances not to exceed the UW variances). This idiosyncrasy of SW variance estimates could be a reason why we obtained better modeling results by using UW variances as proxies of SW variances for the model inputs. Green diamonds show the fitted FHS-based variances and red triangles are CFHG-based variances. Note that CFHG-based variances are generally smaller than FHS-based variances. This is consistent with the distribution plot in Figure 2. In Section 6, we provide simulation results with the aim to recreate the observed phenomenon under a simplified simulation scenario.

In Figure 4, we show respective distributions of the standardized values, where the square root of respective direct or model-fitted variances is used in the denominator for standardization. We observe that the distributions when direct variances are used has long tails, while distributions based on model-fitted variances are compact, with most of their mass lying within the (-2,2) interval. Comparison with the monthly QCEW values gives us a general idea of the form of respective distributions; however, we cannot fully rely on comparison with the monthly QCEW values because of aforementioned differences in the seasonal patterns in QCEW and CES series. In order to better understand the properties of respective confidence intervals, we consider simulations in Section 6.

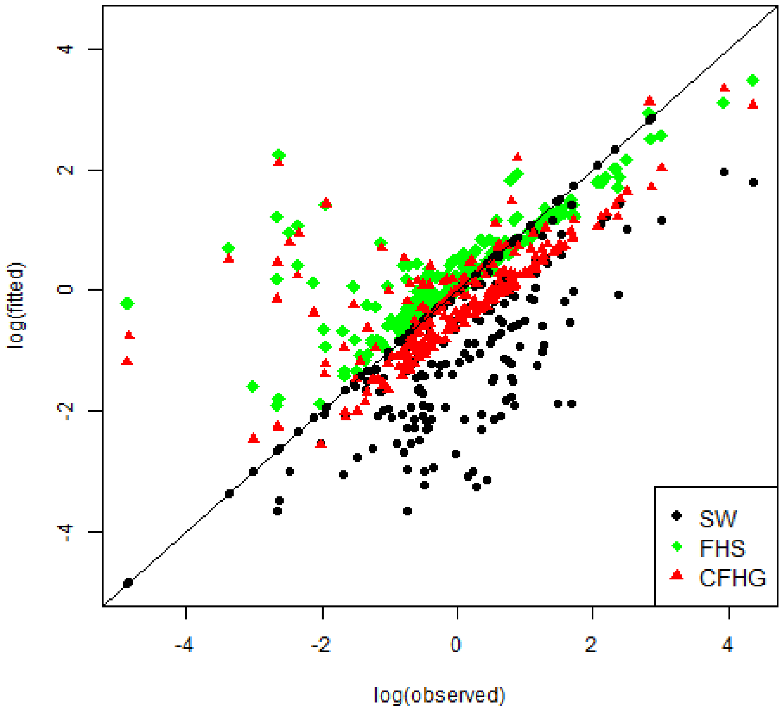

Figure 3: Log(fitted) vs Log(observed UW-based) variances of the sample-based SW estimator.

In Figure 5, we present the violin plots showing the distribution of posterior samples – which convey credibility (confidence) intervals – for the mixture probabilities π_1 , π_2, and π_3 for the case of fitting $K = 3$ mixture components for CFHG model that uses UW variances and SW point estimates as inputs. The violin plots in Figure 6 represent the distribution of posterior samples of parameters b_1 , b_2 , and b_3

for respective mixture components. The violin plots in Figure 7 represent the distribution of posterior samples of parameters μ_1, μ_2, and μ_3 for respective mixture components.

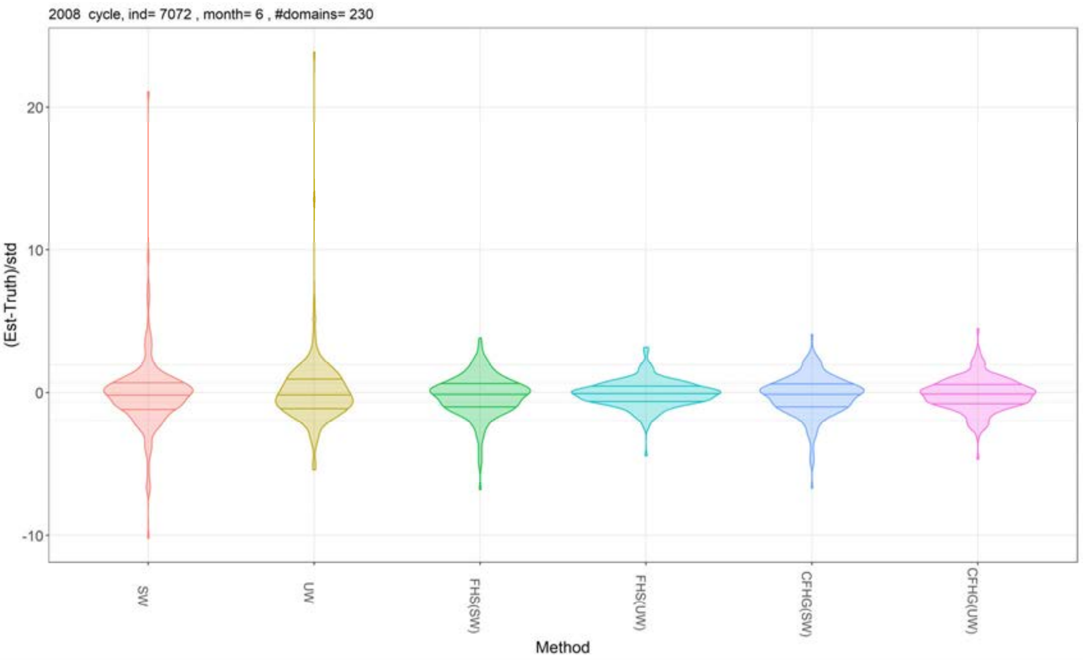

Figure 4: Distribution of Z = (SW-Truth)/sqrt(fitted_vrnc).

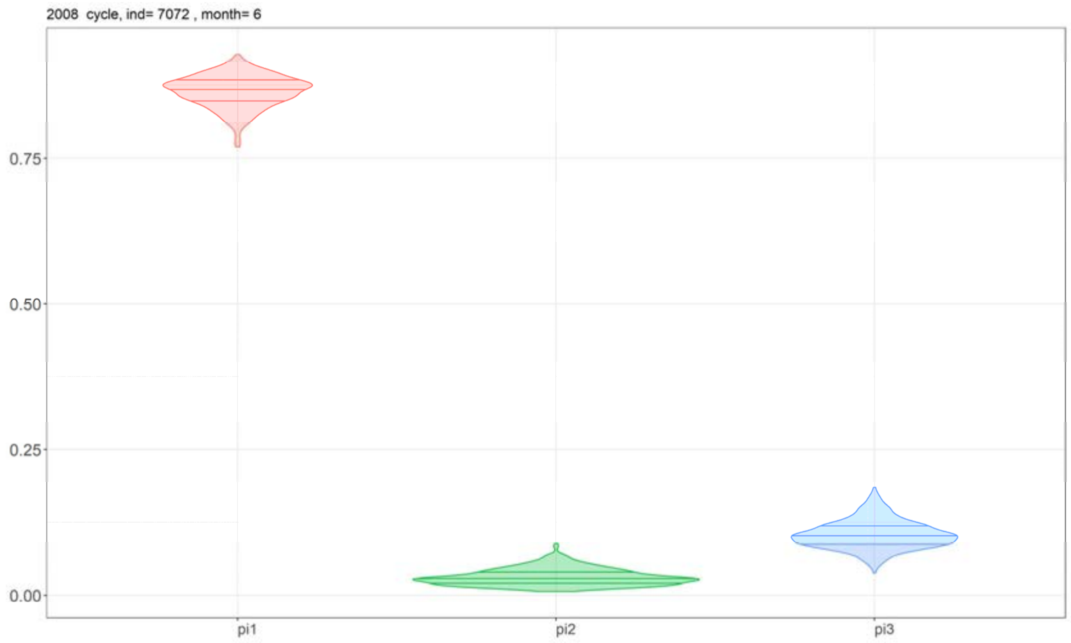

Figure 5: Estimated cluster probabilities from fitting the CFHG model with K=3.

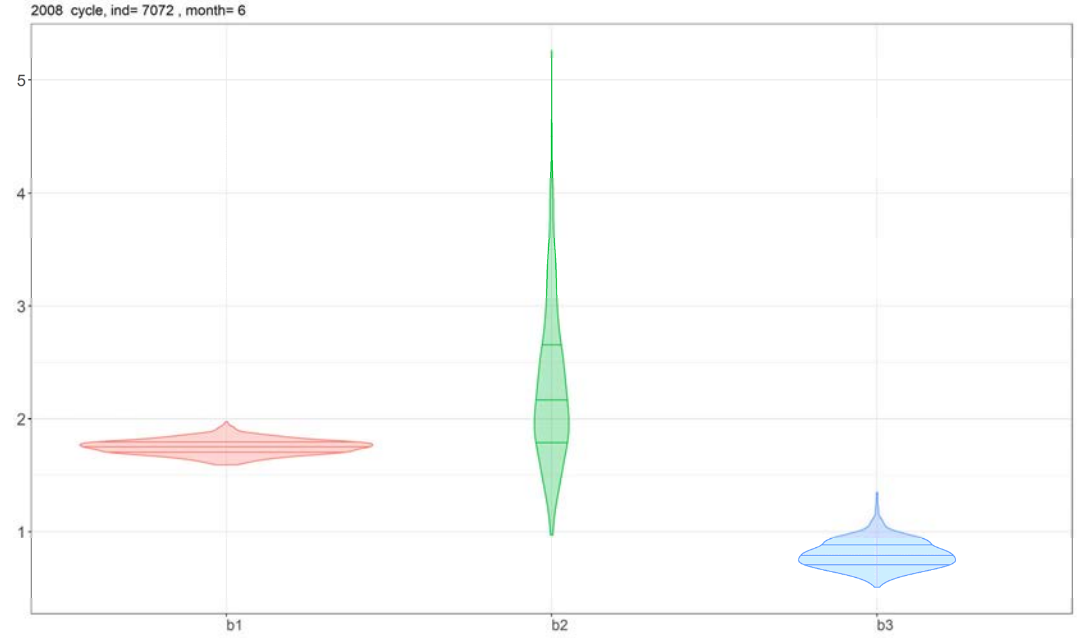

Figure 6: Bias/ridge adjustment for mixture components from fitting the CFHG model with K=3.

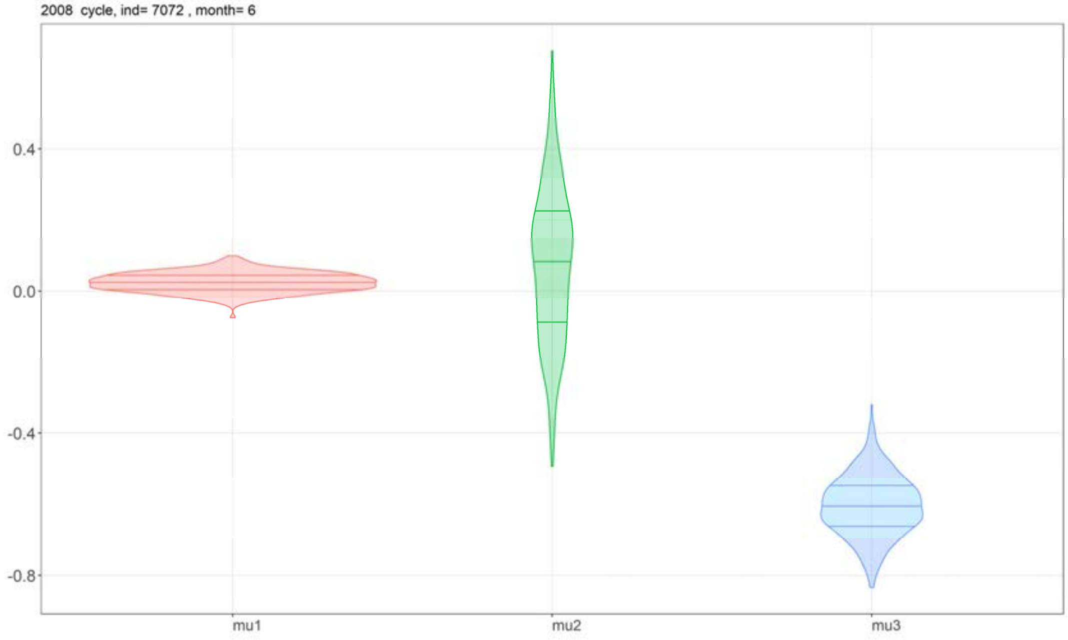

Figure 7: Intercept estimates for mixture components from fitting the CFHG model with K=3.

Note that mixture probability π_2, corresponding to uncertain wide ranging values of parameters b_2 and μ_2, is small. As for the two larger clusters, we see that (i) the first cluster b_1 adjustment accounts for a positive bias from use of UW variances; (ii) the third cluster contains domains with lower range of μ_3 and the b_3 adjustment below 1 may reflect inefficient large variances for noisy domains whose observed estimated variances were overly small.

6. Simulations

In this Section, we consider a simulation scenario to illustrate how the proposed model works.

For each domain i in a set of $N = 300$ domains, generate estimation targets θ_i as

$$\theta_i = \mu_i + \beta x_i + u_i, \tag{24}$$

where auxiliary data $x_i \sim Uniform(-2,2)$, $\beta = 1$, random effects $u_i \sim N\left(0,\tau^2\right)$, $\tau^2 = 1$,

and

$$\mu_i = \begin{cases} 0, & i = 1,\dots,0.95N, \\ 2, & i = 0.95N,\dots,N, \end{cases}$$

that is, 5 percent of domains have a different intercept than the bulk of the observations.

The "observed point estimates" are

$$y_i = \theta_i + e_i, \tag{25}$$

where $e_i \sim N\left(0,\sigma_i^2\right)$ and variances

$$\sigma_i^2 \sim IG\left(\lambda+1, \lambda c \exp\left(\gamma z\right)\right),$$

with $\gamma = 0.1$, $z \sim N\left(0,1\right)$, and $c = \exp\left(-0.5\gamma^2\right)$ and three scenarios for λ : $\lambda = 8,4,1$, corresponding to lower-to-higher degrees of variability of true variances around the mean value represented by the function $c\exp\left(\gamma z\right)$.

We generate "observed" variances as

$$v_i \sim b_i G\left(3, 3\frac{1}{\sigma_i^2}\right), \tag{26}$$

where b_i represent a bias in the estimate of the variance for domain i. Assume three groups of domains biases:

$$b_i = \begin{cases} 2, & i = 1,\dots,0.6N, \\ 0.3, & i = 0.6N+1,\dots,0.8N, \\ 1 & i = 0.8N+1,\dots,N. \end{cases}$$

Generate vector of covariates $g_i = \left(g_{1i}, g_{2i}\right)$ for domain i, where $g_{li} \sim N\left(m_l, 0.25\tau^2\right), l = 1,2$:

$$m_1 = \begin{cases} 0, & i = 1,...,0.95N, \\ 2, & i = 0.95N,...,N, \end{cases}$$

$$m_2 = \begin{cases} 0, & i = 1,...,0.6N, \\ 2, & i = 0.6N + 1,...,0.8N, \\ 4 & i = 0.8N + 1,...,N. \end{cases}$$

We fitted the models using $K = 6$ clusters. In Table 3 we show the mean squared errors (MSE), based on 100 simulation runs, for "direct estimator" Y and estimators based on alternative models. We included two versions of the FH model fit: FH(V), where the "observed" variances are used as "fixed and known" inputs, and FH(GVF), where a GVF used as "fixed and known" input. Here, we computed GVF using the same form and covariates as in the FHS or CFHG models.

In all λ scenarios, CFHG models has the lowest MSE.

Table 3: MSE of point estimates

λ	*Y*	*FH(V)*	*FH(GVF)*	*FHS*	*CFHG*
8	0.998	0.606	0.563	0.573	0.537
4	0.999	0.586	0.562	0.562	0.522
1	1.013	0.502	0.566	0.516	0.475

In Table 4, we present MSE of fitted variances of Y. All model-based variances have lower MSE than the "observed" variances. As expected, when $\lambda = 8$ (low variable around the "synthetic" variance), the GVF has the lowest MSE. However, with increased variability of true variances, the MSE of GVF increases. CFHG fitted variances have lower MSE than FHS fitted variances in $\lambda = 8$ and $\lambda = 4$ scenarios.

Table 4: MSE of variances of "direct estimator" Y

λ	*Observed*	*GVF*	*FHS*	*CFHG*
8	1.840	0.382	0.907	0.523
4	2.192	0.586	0.938	0.608
1	10.003	7.246	1.999	3.509

Table 5: Coverage properties of variances of "direct estimator" Y, 95% nominal

	Coverage					*Length*				
λ	*True*	*Observed*	*GVF*	*FHS*	*CFHG*	*True*	*Observed*	*GVF*	*FHS*	*CFHG*
8	0.950	0.897	0.979	0.976	0.948	3.857	4.290	4.753	4.887	4.263
4	0.950	0.896	0.976	0.978	0.957	3.798	4.228	4.768	4.850	4.298
1	0.951	0.898	0.971	0.979	0.976	3.483	3.869	4.790	4.592	4.332

In Table 5, we present the coverages of confidence intervals of "direct estimator" Y based on respective fitted variances. The "observed" variances do not give the nominal coverage. CFHG model provides nearly nominal coverage in $\lambda = 8$ and $\lambda = 4$ scenarios, with shorter lengths of the intervals, and slight overcoverage in the $\lambda = 1$ scenario.

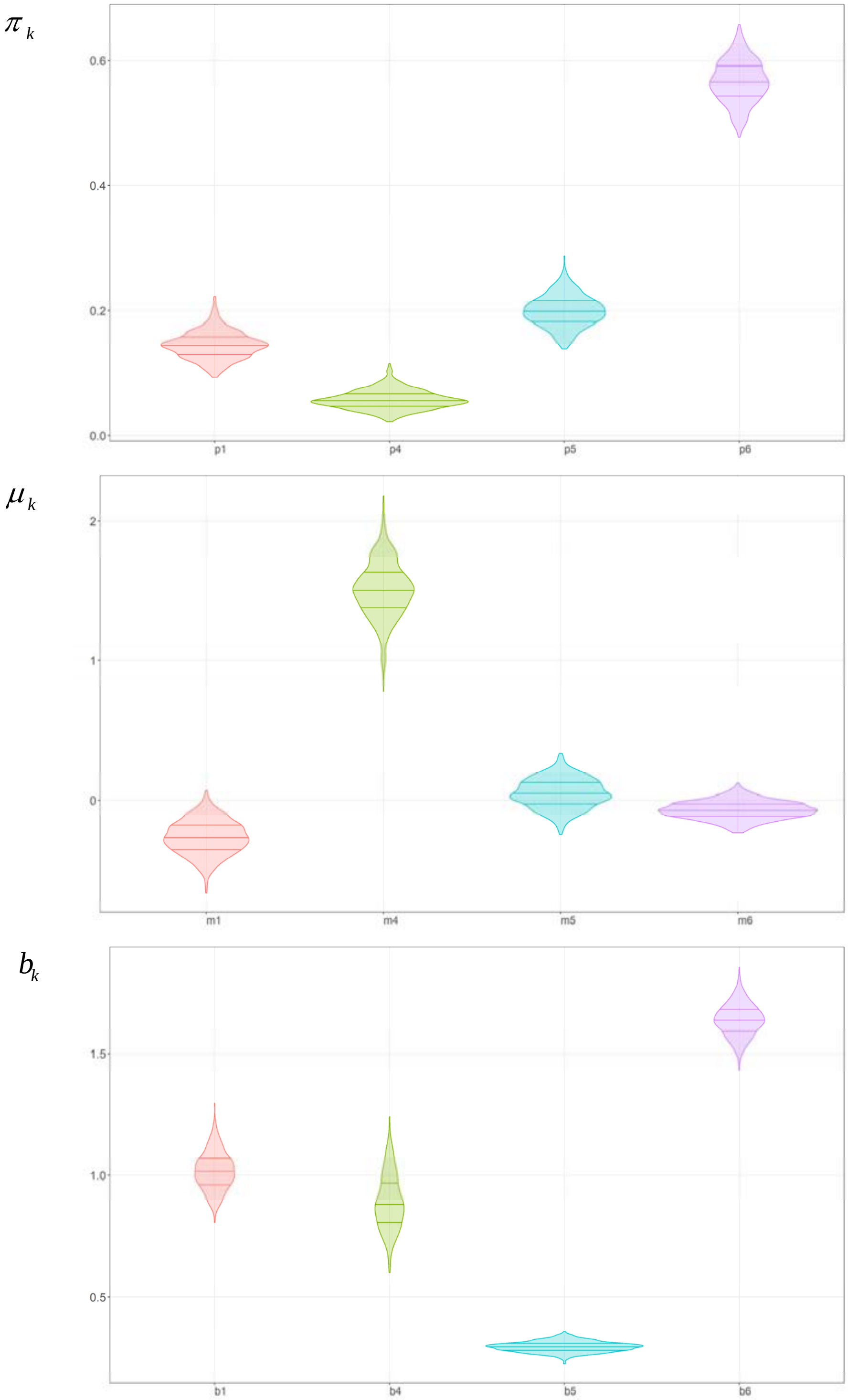

Figure 8: Cluster probabilities, intercepts, and bias/ridge adjustments, from fitting CFHG with K=6.

In Figure 8, we present the violin plots showing the distribution of posterior samples for π_k, μ_k, and b_k for those clusters where estimated cluster probability is greater than 0.05 (based on a single simulation.) .

The CFHG model, in this example recognizes at least 4 distinct clusters having cluster probabilities from 0.05 to about 0.60. The first plot shows posterior distribution of cluster probabilities. The second plot shows posterior distributions of μ_k's. The third plot shows the distribution of respective cluster biases b_k. We observe that μ_k is distributed around 1.5-2 in the cluster where π_k is around 0.05 and b_k is around 1. The cluster that has probability π_k about 0.6 also has estimated bias adjustment of about 1.5-2, which corresponds to the overestimated group; the other two clusters (each distributed roughly around 0.15-0.2) correspond to bias adjustments b_k centered around 1 and 0.3. Thus, based on the values of estimated biases, we should expect to recover at least part of the bias in the estimated variances.

7. Summary

In this paper, we proposed clustering model that utilizes additional covariates *to inform clustering.* The model is robust to *deviations from the linearity* and can correct for a *bias in the variance estimates.*

Results from application on ten years of data from the Current Employment Statistics survey suggest that direct SW estimates are more efficient than UW estimates. Models based on SW estimates perform better than models based on UW estimates. Joint point estimates and variances models, FHS or CFHG, perform better than the classical FH model. Overall, CFHG performs slightly better than FHS, although results vary by industry. FHS model with SW point estimates and UW-based variances as inputs performs better than the FHS model that uses as inputs SW point estimates and estimated SW variances. MAD from the CFHG model based on these pairs of inputs are close.

References:

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. Biometrika, 95, 3, pp. 539–553

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.

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

Gershunskaya, J. and Savitsky, T.D. (2018) Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. Accepted for publication in the *Journal of Survey Statistics and Methodology*

Gershunskaya, J. and Sverchkov, M. (2014) On Weight Smoothing in the Current Employment Statistics Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association

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.

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

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.

Appendix: Proof of Proposition 1

Let $\delta = \left(b, \tau_u^2, \theta_1, ..., \theta_N, \sigma_1^2, ..., \sigma_N^2\right)$ denote the vector of parameters and let $D = (d_1, ..., d_N)$, where $d_i = (y_i, v_i)$ is the data vector for domain i. .

The posterior distribution of δ is

$$f(\delta \mid D) \propto p(\delta) \prod_{i=1}^{N} f(d_i \mid \delta_i),$$

where

$$f(d_i \mid \delta_i) = \frac{1}{(2\pi)^{1/2} \tau_u} \exp\left(-\frac{1}{2}\frac{\theta_i^2}{\tau_u^2}\right) \frac{1}{(2\pi)^{1/2} \sigma_i} \exp\left(-\frac{1}{2}\frac{(y_i - \theta_i)^2}{\sigma_i^2}\right)$$

$$\times v_i^{\frac{an_i^*}{2}-1} \left(\frac{an_i^*}{2}\frac{1}{b\,\sigma_i^2}\right)^{\frac{an_i^*}{2}} \exp\left(-\frac{an_i^*}{2}\frac{v_i}{\sigma_i^2}\frac{1}{b}\right) \times \left(\frac{1}{\sigma_i^2}\right)^{\frac{\beta_\sigma}{2}} \exp\left(-\frac{\beta_\sigma}{2}\frac{1}{\sigma_i^2}\gamma_i\right) \times \left(\frac{1}{b}\right)^{\frac{\beta_b}{2}} \exp\left(-\frac{\beta_b}{2}\frac{1}{b}\right)$$

The posterior density of parameters is:

$$f(\delta \mid D) \propto p(\delta)\left(\tau_u^2\right)^{-N/2} \exp\left(-\frac{\beta_b}{2}\frac{N}{b}\right) \prod_{i=1}^{N} v_i^{\frac{an_i^*}{2}-1} \left(\frac{1}{\sigma_i^2}\right)^{\frac{1}{2}+\frac{an_i^*}{2}+\frac{\beta_\sigma}{2}} \left(\frac{1}{b}\right)^{\frac{an_i^*}{2}+\frac{\beta_b}{2}} \exp\left(\begin{array}{c} -\frac{1}{2}\frac{\theta_i^2}{\tau_u^2} - \frac{1}{2}\frac{(y_i-\theta_i)^2}{\sigma_i^2} \\ -\frac{an_i^*}{2}\frac{v_i}{\sigma_i^2}\frac{1}{b} - \frac{\beta_\sigma}{2}\frac{\gamma_i}{\sigma_i^2} \end{array}\right)$$

The full conditional distributions for σ_i^2 and b:

$$\sigma_i^{-2} \mid \tau_u^2, \theta, \sigma_{(-i)}^2, b, D \sim G\left(\frac{1}{2}\left(1 + an_i^* + \beta_\sigma\right) + 1, \frac{1}{2}\left\{(y_i - \theta_i)^2 + an_i^* \frac{v_i}{b_i} + \beta_\sigma \gamma_i\right\}\right), i = 1, ..., N,$$

$$b^{-1} \mid \tau_u^2, \theta, \sigma^2, D \sim G\left(\frac{N}{2}\left(a\bar{n}^* + \beta_b\right) + 1, \frac{N}{2}\left\{a\bar{n}^* N^{-1} \sum_{i=1}^{N} \frac{n_i^*}{\bar{n}^*}\frac{v_i}{\sigma_i^2} + \beta_b\right\}\right).$$

Let

$$a_i = \frac{3}{2} + \frac{an_i^*}{2} + \frac{\beta_\sigma}{2}$$

$$c_i = \frac{1}{2}\left((y_i - \theta_i)^2 + an_i^* \frac{v_i}{b} + \beta_\sigma \gamma_i\right)$$

Then, integrating out σ_i^2, gives us:

$$f\left(b,|\,\tau_u^2,\theta,D\right)\propto\left(\frac{1}{b}\right)^{\frac{N}{2}\left(a\bar{n}^*+\beta_b\right)}\exp\left(-\frac{\beta_b}{2}\frac{N}{b}\right)\prod_{i=1}^{N}c_i^{-a_i}=\left(\frac{1}{b}\right)^{\frac{N}{2}\left(a\bar{n}^*+\beta_b\right)}\exp\left(-\frac{\beta_b}{2}\frac{N}{b}\right)\exp\left(-\sum_{i=1}^{N}a_i\ln c_i\right).$$

Consider

$$\ln\frac{1}{2}\left(\left(y_i-\theta_i\right)^2+an_i^*\frac{v_i}{b}+\beta_\sigma\gamma_i\right)=\ln\frac{\left(\left(y_i-\theta_i\right)^2+\beta_\sigma\gamma_i\right)}{1+\beta_\sigma}\left(\frac{3}{2}+\frac{an_i^*}{2}+\frac{\beta_\sigma}{2}\right)$$

$$+\ln\frac{1+\beta_\sigma}{\left(\left(y_i-\theta_i\right)^2+\beta_\sigma\gamma_i\right)}\frac{\frac{1}{2}\left(\left(y_i-\theta_i\right)^2+an_i^*\frac{v_i}{b}+\beta_\sigma\gamma_i\right)}{\left(\frac{3}{2}+\frac{an_i^*}{2}+\frac{\beta_\sigma}{2}\right)}$$

The first summation term doesn’t depend on b. Using condition (19) and approximation $\ln x\approx x-1$, we can write (after leaving out the multiplicative terms that don’t depend on b):

$$f\left(b,|\,\tau_u^2,\theta,D\right)\propto\left(\frac{1}{b}\right)^{\frac{N}{2}\left(a\bar{n}^*+\beta_b\right)}\exp\left(-\frac{1}{2}\frac{N}{b}\left(\beta_b-\left(1+\beta_\sigma\right)\frac{a\bar{n}^*}{N}\sum_{i=1}^{N}\frac{\frac{n_i^*}{\bar{n}^*}v_i}{\left(\left(y_i-\theta_i\right)^2+\beta_\sigma\gamma_i\right)}\right)\right).$$