JOINT POINT AND VARIANCE ESTIMATION UNDER A HIERARCHICAL BAYESIAN MODEL FOR SURVEY COUNT DATA* September, 2023

BY TERRANCE D. SAVITSKY[1,a], JULIE GERSHUNSKAYA[2,b] AND MARK CRANKSHAW[2,c]

[1]*Office of Survey Methods Research, U.S. Bureau of Labor Statistics,* [a]*Savitsky.Terrance@bls.gov*

[2]*OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics,* [b]*Gershunskaya.Julie@bls.gov;* [c]*Crankshaw.Mark@bls.gov*

We propose a novel Bayesian framework for the joint modeling of survey point and variance estimates for count data. The approach incorporates an induced prior distribution on the modeled true variance that sets it equal to the generating variance of the point estimate, a key property more readily achieved for continuous data response type models. Our count data model formulation allows the input of domains at multiple resolutions (e.g., states, regions, nation) and simultaneously benchmarks modeled estimates at higher resolutions (e.g., states) to those at lower resolutions (e.g., regions) in a fashion that borrows more strength to sharpen our domain estimates at higher resolutions. We conduct a simulation study that generates a population of units within domains to produce ground truth statistics to compare to direct and modeled estimates performed on samples taken from the population where we show improved reductions in error across domains. The model is applied to the job openings variable and other data items published in the Job Openings and Labor Turnover Survey administered by the U.S. Bureau of Labor Statistics.

1. Introduction. Count data response variables are commonly measured by government surveys; for example, the American Community Survey administered by the U.S. Census Bureau counts the population below a poverty threshold for household domains indexed by geography (e.g., census tracts). The U.S. Census Bureau administer the Consumer Expenditures surveys of consumer units (independent households) for the U.S. Bureau of Labor Statistics (BLS) that include count variables related to local and regional locations of the consumer units. BLS administers surveys and a census instrument of business establishments related to total employment and its components (e.g., job openings, hires, separations).

As with surveys conducted for continuous data response types, surveys that include count data responses aggregate respondent-level counts, such as total employment for a business establishment respondent, to a collection of domains (such as state-by-industry classification) and produce both a point estimate and an estimated variance statistic for each domain. Small domain estimation models for the continuous response type that jointly model the point estimates and the estimated variances for the domains exist within both frequentist and Bayesian frameworks; see, for example, Maiti et al. (2014) and Sugasawa et al. (2017). These models borrow strength from the underlying correlations among the domain estimates to provide de-noised model-based estimators that are

arXiv: math.PR/0000000

*U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E, Washington, D.C. 20212 USA

Keywords and phrases: Bayesian hierarchical models, Small Area Estimation, Count data, Stan.

characterized by lower mean squared errors. The inferential goal for these models is to extract model-smoothed point and variance estimates for publication to their data users. It is important to note that the domain-indexed variances are not known, but estimated, such that small domain models provide an opportunity to enhance the quality of both point and variance estimates through their joint estimation since they are typically correlated.

Bayesian models for continuous data point and variance estimates are easily designed such that the mean of the marginal likelihood for the estimated variances represents a denoised "true" variance. The true variance, in turn, is set to be the "generating" variance for the noisy point estimate in its likelihood centered around the estimated true mean value (Sugasawa et al. 2017); for example, suppose v_d represents the estimated sampling variance for domain $d \in (1, \ldots, N)$ associated with *continuous* response, y_d. In the case of unknown, latent true domain variance, σ_d^2, one may impose a likelihood, $v_d \mid \sigma_d^2 \overset{\text{ind}}{\sim} f(\sigma_d^2)$ with mean σ_d^2. One typically chooses the conditional likelihood, $y_d \mid \theta_d, \sigma_d^2 \overset{\text{ind}}{\sim} \mathcal{N}(\theta_d, \sigma_d^2)$, where $\mathcal{N}(\cdot)$ denotes the normal distribution. We see that the variance in the conditional likelihood for continuous y_d is readily and naturally set to equal the latent true variance, σ_d^2. This connection between the point and variance estimates where the true modeled variance is set as the generating variance of the noisy point estimate ties together the likelihoods for the point and variance estimates in a single model framework.

We are not aware of a (small area) model for count data in the small estimation literature where the estimated sampling variance is modeled jointly with the direct point estimate such that the generating variance of the direct point estimate is set equal to the mean of estimated sampling variance (where we interpret the mean as the latent "true" domain variance). In their recent comprehensive review article of small estimation methods, Sugasawa & Kubokawa (2020) note the possibility to model the point estimate with a non-normal distribution and more broadly discuss the use of generalized linear models. They do not, however, explicate a count data model that incorporates estimated variances. Similarly, Rao & Molina (2015) discuss over-dispersed Poisson models for count data, but none that incorporate estimated domain variances. Tzavidis, Nikos and Ranalli, M Giovanna and Salvati, Nicola and Dreassi, Emanuela and Chambers, Ray (2015) develop a Poisson small area model for count data, but assume the domain variance is equal to the mean of the Poisson likelihood for the domain point estimate. So, they do not input domain variances, at all.

The literature does, however, provide a recent example where Bradley et al. (2016) construct a joint model for geographically-indexed point and variance estimates under a count data response. They define a Poisson likelihood such that the model conditional variance (of the point estimate) is defined as $\text{Var}\left(y_d \mid x_d\right) = \exp\left(\lambda_d\right)$ for count data response, y_d, associated to domain $d \in (1, \ldots, N)$; where x_d is a set of covariates and λ_d is the log-mean parameter. By contrast, under a normal likelihood with mean λ_d and variance φ_d for logarithm, $\log(v_d)$, of true variance v_d of y_d, the associated mean is $\mathbb{E}\left(v_d \mid x_d\right) = \sigma_d^2 = \exp\left(\lambda_d + \sigma_d^2/2\right) \neq \text{Var}\left(y_d \mid x_d\right)$. Although Bradley et al. (2016) utilize a random effect by specifying a likelihood for the log-variance, this construction does not produce a true variance that is equal to the generating variance of the count data response. All to say, the literature is more limited for small area models for count data that incorporate estimated variances and there are no implementations to our knowledge that ensure the $\sigma_d^2 = \text{Var}\left(y_d \mid x_d\right)$. Perhaps the reason for the limited literature focused on count data models for small area estimation is that for many datasets domain level counts are sufficiently large to

approximate with a continuous data distribution, though we have mentioned some data examples above with count data variables that express low counts for some domains.

This paper, by contrast, constructs a joint model for a count data point estimate y_d and its estimated variance v_d where the modeled true variance σ_d^2 (the mean of the conditional likelihood for v_d) is set equal to the variance $\text{Var}(y_d \mid x_d)$ of the point estimate likelihood. We extend a multiplicative random effect in our model specification for the point estimate as suggested by Zhou et al. (2012) for non-survey data where there is no associated variance estimate. They discuss that a Poisson-Lognormal prior set-up better fits the data with more appealing large sample theoretical properties as compared to the Negative Binomial model because the more flexible formulation of the former allows the data to learn a higher degree of over-dispersion. We extend their Poisson-Lognormal set-up in our Bayesian hierarchical model framework to indirectly induce a prior on the true variance of the likelihood for the estimated variance such that the true variance equals the generating variance for the point estimate.

Our formulation further leverages a Bayesian hierarchical probability model construction by including the variable of interest at multiple resolutions (e.g., nested geographic levels, such as states, regions, nation). The model simultaneously benchmarks modeled point estimates at higher resolutions (e.g., states) to those at lower resolutions (e.g., regions) in a fashion that sharpens the estimation quality at higher resolutions. Traditional benchmarking discussed in Rao & Molina (2015), by contrast, is often performed as a second step after modeling is completed such that it tends to add back some of the variance removed by modeling. We avoid this loss of efficiency by including the benchmarking as part of performing estimation at multiple resolutions in a single step as is done in Savitsky (2016).

1.1. *Job Opening and Labor Turnover Survey (JOLTS) Motivating Dataset.* This paper was motivated by the JOLTS survey conducted by the U.S. Bureau of Labor Statistics. JOLTS measures dynamic trends in the labor market by tracking job openings, hires and separations, among other variables, on a monthly basis. The survey is conducted nationally with the intent to provide a national-level estimator for a collection of industries (defined based on the North American industry classification codes (NAICS)). Users, however, desire to have state-level estimates of these labor market dynamic variables for each industry. A major challenge to produce state-level estimates of JOLTS variables is the small number of surveyed business establishments in many states; in fact, in some industries there are states that may not have any sample responses in a given month. It is cost prohibitive for BLS to increase their sample size and to use blocking by state in order to support state level estimation. Our goal in this paper is to model the collection of state-by-industry point estimates and variances constructed from the national survey to extract more efficient, higher quality estimators and to impute estimated values for state-level domains with no underlying sample. The Bayesian hierarchical model that we construct in the sequel for each industry will simultaneously impute missing point estimates and variances for those states excluded from the sample in any given month.

The remainder of this paper is organized, as follows: Section 2 provides the mathematical formulation of our joint model for state-level point and variance estimates for each month in each industry.We then extend this cross-sectional by-month model to a time-series construction that jointly models a collection of state-indexed time-series for each of the point and variance estimates in each industry. We design a simulation study in Section 3 that generates respondent level

population of count data and constructs true values for point estimates over a collection of domains. Our simulation design then takes a sample of respondents and produces domain-based point and both true and estimated variances of direct point estimates for the sampled respondents in each domain. We compare the MSE performances of the sample-based direct estimator and our modeled estimator based on the population ground truth. In Section 4, we proceed to apply our model to provide JOLTS state-based point and variance estimates and we illustrate the smoothing property of our models. We conclude with a brief discussion in Section 5.

2. Model for Count Data Point and Variance Estimates.

2.1. *Cross-sectional Model Under Assumed Known Variances.* We proceed to describe the formulation of a cross-sectional model for the joint estimation of count data point in a given month. We utilize the structure of JOLTS data to describe our model, for ease-of-understanding. Domains in JOLTS are defined by intersections of states and industries. We consider separate models by industry, each estimated over the collection of states. For each domain $d \in (1, ..., N)$ we observe sample based estimates, y_d, and respective estimates of their variances, v_d, where N denotes the total number of domains in a given industry.

We construct a model for a count data response, rather than treating point estimate, y_d, as continuous because the JOLTS variables of interest, such as job openings, are often characterized by very small counts for a given industry and state such that the conditional likelihood is expected to be very skewed (unlike a symmetric normal distribution). Our model will specify a Poisson-lognormal model that allows for an over-dispersed marginal likelihood for point estimate y_d for domain $d \in (1, \ldots, N)$ where the data may estimate the variance, $\text{Var}(y_d \mid x_d) \geq \mathbb{E}(y_d \mid x_d)$. We assume that the constructed domain variances, $(v_d)_{d=1}^N$, are known such that we treat them as fixed. So, we specify a likelihood for $y_d \mid x_d$ and set its variance equal to v_d, which we see below is not trivial.

2.1.1. *Poisson-lognormal Mixture Likelihood Formulation.* We allow for overdispersion through a Poisson-lognormal joint likelihood with:

$$y_d|\theta_d, \varepsilon_d \overset{\text{ind}}{\sim} \text{Poisson}\left(\theta_d \varepsilon_d\right), \tag{1}$$

where the use of latent random effects ε_d with a mean fixed to 1 allows for the variance to be greater than or equal to the mean by specifying the following distribution for the latent likelihood,

$$\varepsilon_d|\varphi_d \overset{\text{ind}}{\sim} \text{LN}\left(-0.5\varphi_d^2, \varphi_d^2\right). \tag{2}$$

Together, Equations 1 and 2 for observed direct sample-based estimates y_d (of job openings, hires, separations, or other JOLTS items) follow a lognormal scale mixture of Poisson distributions parameterized so that mean $E\left(y_d|\theta_d\right) = \theta_d$, where θ_d represents the parameter of interest. We accomplish setting the mean of the mixture likelihood to θ_d through our use of $-0.5\varphi_d^2$ in Equation 2 that restricts the prior mean of ε_d to be 1. If we had instead chosen a Gamma distribution for ε_d the resulting likelihood after marginalizing over ε_d would have produced a closed-form negative binomial distribution which is *not* the case for our Poisson-lognormal mixture. We, nevertheless, select the lognormal instead of the Gamma because Zhou et al. (2012) highlight that the lognormal has proven more flexible for the modeling of heavy tails that we express in the JOLTS data to the presence of domains with very small counts.

2.1.2. *Linking Model for Conditional Mean, θ_d.* The conditional mean parameter, θ_d, are used to borrow strength across domains and is constructed with the formulation

$$\theta_d = X_d \exp(\lambda_d), \tag{3}$$

where X_d is an "offset". The use of an offset allows specification of the regression model for a normalized rate, $\exp(\lambda_d)$ which allows the data to estimate correlations among domains for smoothing among domains of different sizes. The use of a magnitude offset is typical for count data models; for example, in estimating disease prevalance (Gelman et al. 2014). The offset X_d is assumed known and does not contain error. In application to JOLTS, we use the employment level, derived from the Current Employment Statistics (CES) survey conducted by the Bureau of Labor Statistics, as the offset. The total employment values are typically much larger in magnitude that the JOLTS variables (e.g., job openings, quits, hires) and the rate of JOLTS variables to total employment composes a natural ratio in a similar fashion to the number with a disease over the total population in disease mapping Gelman et al. (2014). Although the CES-based X_d is an estimate, it is based on a much larger sample than the JOLTS estimate, and so we ignore the variance associated with estimation ofX_d.

We model λ_d in Equation 3 on the log-scale such that $\lambda_d \in \mathbb{R}$ and we may specify a normal distribution prior distribution. We specify a linking model for λ_d in Equation 4 that allows "borrowing strength" across domains from a set of covariates with,

$$\lambda_d \sim \mathcal{N}\left(\beta x_d, \tau^2\right). \tag{4}$$

The prior specifies that rate λ_d follows a normal distribution, centered at βx_d with variance τ^2; x_d is a set of covariates and β is a vector of regression coefficients. Hyperparameters β and τ^2 are "global" in the sense that their values are shared for all domains.

2.1.3. *Setting Conditional Variance of y_d to Equal Known Variance, v_d.* The Poisson-lognormal mixture likelihood of Equations 1 and 2 produces the marginal variance,

$$\text{Var}\left(y_d \mid \theta_d, \varphi_d\right) = \theta_d + \theta_d^2\left(\exp(\varphi_d^2) - 1\right), \tag{5}$$

where the marginal variance is a function of parameters (θ_d, φ_d). By construction, $v_d = \text{Var}(y_d \mid X_d)$, so we need to set the marginal variance (after integrating out ε_d) of the Poisson-lognormal mixture equal to v_d (treated as known and fixed) with,

$$v_d = \text{Var}\left(y_d \mid \theta_d, \varphi_d\right) = \theta_d + \theta_d^2\left(\exp(\varphi_d^2) - 1\right). \tag{6}$$

Our inferential interest is in θ_d, the conditional mean for y_d and we specify the linking model for θ_d to borrow strength among domains to produce a smoothed estimator. So, we accomplish setting v_d to be the marginal variance for y_d by solving for φ_d^2 in Equation 7 to achieve,

$$\varphi_d^2 = \log\left(\frac{v_d - \theta_d}{\theta_d^2} + 1\right), \tag{7}$$

where we have *induced* a prior on φ_d through our linking model distribution imposed on λ_d (where we recall Equation 3). In other words, unlike the typical set-up in Bayesian models, we do not directly set a prior distribution for φ_d but induce it through θ_d and the functional relationship of

Equation 7 to guarantee that Equation 6 is achieved. As discussed in the introduction, Sugasawa & Kubokawa (2020) mention the possibility for use of non-normal likelihoods and generalized linear models, but do not include the joint modeling of point estimates and variances for count data. Similarly is the case for Rao & Molina (2015). All to say, ours is the first treatment of a count data model for domain level data that enforces the variance condition of Equation 6.

2.2. *Model Extension for Joint Modeling of Point Estimates and Variances,* (y_d, v_d).

2.2.1. *Likelihood for Observed Variances,* v_d. We next address the case where the true, underlying domain variances are unknown such that we construct an additional likelihood statement for the observed variances, v_d, centered on the true, latent variances.

We denote the true latent variance of sample-based point estimate y_d by σ_d^2 such that $\sigma_d^2 = \text{Var}\left(y_d|\theta_d, \varphi_d\right)$. We achieve this equality by altering Equation 6 to,

$$\sigma_d^2 = \text{Var}\left(y_d \mid \theta_d, \varphi_d\right) = \theta_d + \theta_d^2\left(\exp(\varphi_d^2) - 1\right). \tag{8}$$

Note, however, that true sampling variances σ_d^2 are not observed. Instead, we observe estimated variances v_d. We choose to work with observed squared coefficient of variation, $cv_d^2 = v_d/y_d^2$, as doing so allows us to better identify both φ_d and θ_d by avoiding a multiplicative formulation for the induced prior on σ_d^2. We select a gamma distribution prior for cv_d^2 with mean

$$r_d^2 = \frac{\sigma_d^2}{\theta_d^2} = \frac{1}{\theta_d} + \left(\exp\left(\varphi_d^2\right) - 1\right). \tag{9}$$

We specify the prior precision of cv_d^2 to depend on sample size n_d and on a random scale hyperparameter a_0 that represents a latent concentration property of the population that affects the prior precision of cv_d^2:

$$cv_d^2|a_0, r_d^2 \overset{\text{ind}}{\sim} \text{Gamma}\left(0.5a_0n_d, 0.5a_0n_d\frac{1}{r_d^2}\right). \tag{10}$$

Under this Gamma distribution likelihood, the observed cv_d^2 concentrates to a greater degree on the truth, r_d^2, if the domain sample size is relatively large.

2.2.2. *Prior distribution for Overdispersion,* φ_d^2. The latent random effects φ_d of Equation 2 are interpreted as overdispersion of the Poisson with mean θ_d. We suppose that overdispersion is driven mostly because observed $y_d{}'s$ are sample-based estimates (rather than population measurements). Hence, it is natural to assume that the value of parameter φ_d depends on the (domain) sample size n_d. We let

$$\varphi_d \sim \mathcal{N}_+\left(\gamma_0\frac{1}{\sqrt{n_d}}, 0.1\right), \tag{11}$$

where $\mathcal{N}_+(\cdot)$ denotes a half normal distribution for some hyperparameter γ_0, where this prior specifies a smaller overdispersion for domains with relatively larger sample sizes because relatively larger values of n_d result in smaller values of φ_d which, in turn, produce a lower variance for ϵ_d.

Note that we induce a prior on the latent true variance, σ_d^2, from Equation 8 through our direct priors on φ_d and λ_d specified in Equations 4 and 11, respectively, to ensure that $\sigma_d^2 = \text{Var}\left(y_d|\theta_d, \varphi_d\right)$; that is, we do not directly specify a prior distribution for σ_d^2.

2.3. *Model Extension to Incorporate Regional Data.* We have now fully specified the portion of our hierarchical model for the state-level. We simultaneously model likelihoods for point and variance estimates at the regional level (where for a given industry each region nests a collection of contiguous states) since these aggregated survey estimates are more reliable. We proceed to constrain the modeled point estimates at the state level to sum to those at the region level, which accomplishes simultaneous benchmarking of the states to regions.

In addition, sample based point estimates y_r and respective estimated variances v_r are also available for regions, $r = 1, ..., R$. The United States is subdivided into four regions that nest the states, so $R = 4$.

To the degree that the regional estimates can be viewed as more reliable compared to the domain estimates, this part of the model serves as a denoised "benchmarking constraint" for domain-level estimates. Yet, because the benchmarking is performed simultaneously with estimation, it produces a lower variance estimate than separately estimating the domain modeled estimates followed by a subsequent benchmarking step. The benchmarking of modeled state estimates to modeled regional estimates also provides a practical benefit by adding stability to estimation of the model parameters. The model for region r is specified as:

Mixture likelihood for y_r

$$y_r|\theta_r \overset{\text{ind}}{\sim} \text{Poisson}\left(\theta_r \varepsilon_r\right),\ \theta_r = \sum_{d \in r} \theta_d \tag{12a}$$

$$\varepsilon_r|\varphi_r \sim \text{LN}\left(-0.5\varphi_r^2, \varphi_r^2\right) \tag{12b}$$

Benchmarking step for θ_r

$$\theta_r = \sum_{d \in r} \theta_d \tag{13}$$

Prior for overdispersion φ_r

$$\varphi_r \sim \mathcal{N}_+\left(\gamma_1 \frac{1}{\sqrt{n_r}}, 0.1\right),\ n_r = \sum_{d \in r} n_d \tag{14}$$

Likelihood for estimated (square of) coefficient of variation $cv_r^2 = v_r / y_r^2$

$$cv_r^2|a_1, r_r^2 \overset{\text{ind}}{\sim} \text{Gamma}\left(0.5a_1 n_r, 0.5a_1 n_r \frac{1}{r_r^2}\right), \tag{15}$$

where $r_r^2 = \frac{1}{\theta_r} + \left(\exp\left(\varphi_r^2\right) - 1\right)$. We accomplish a benchmarking step in Equation 13 where the mean θ_r set equal to the sum of the domain-level parameter values over the domains belonging to a given region. Thus, the model for the domain means, (θ_d) are able to borrow strength from the regional (y_r) through their links to (θ_r) in Equation 13. We emphasize that this benchmarking step is accomplished simultaneously with estimation of model parameters because the benchmarking is performed among parameters rather than the observed, estimated point estimates. So, the Bayesian implementation of benchmarking further borrows strength and aids in estimation. By contrast, the benchmarking discussed in Rao & Molina (2015) is performed *after* modeling is performed

by benchmarking the modeled estimates to external data benchmarks. This two-step process, by contrast, adds back some of the noise removed in the model estimation, so it reduces estimation quality.

We proceed to specify a further model extension that benchmarks regional estimates, θ_r, to national point and variance estimates. The form of the extension is analogous to that provided for the regions above, so we omit it for brevity and clarity of exposition.

Model hyperparameters are drawn from the following distributions: $\beta \sim \mathcal{N}\left(0, \sigma_\beta^2\right), \sigma_\beta, \tau \sim$ student-$t_{3+}(0,1)$ (half-Student with 3 degrees of freedom), $\gamma_0, \gamma_1, \sqrt{a_0}, \sqrt{a_1}, \sqrt{a_2} \sim \mathcal{N}_+(0,1)$.

2.4. *A Time-series Extension.* In our JOLTS application presented in Section 4, we compare the cross-sectional model with an extension that models a collection of (domain-indexed) time-series of point and variance estimates with the intent to borrow more strength from an autocorrelation among the monthly estimates. We use the same model set-up as the cross-sectional Poisson-lognormal mixture, but now index our parameters by *both* domain, d and time index (month), t.

Mixture likelihood for y_{dt}

$$y_{dt}|\theta_{dt}, \varepsilon_{dt} \overset{\text{ind}}{\sim} \text{Poisson}\left(\theta_{dt}\varepsilon_{dt}\right) \tag{16a}$$

$$\varepsilon_{dt}|\varphi_{dt} \sim \text{LN}\left(-0.5\varphi_{dt}^2, \varphi_{dt}^2\right) \tag{16b}$$

Prior distribution for smoothed point estimate θ_{dt}

$$\theta_{dt} = X_{dt}\exp\left(\lambda_{dt}\right) \tag{17a}$$

$$\lambda_{dt} \sim N\left(\beta_t x_{dt}, \tau^2\right) \tag{17b}$$

Likelihood for estimated (square of) coefficient of variation $cv_{dt}^2 = v_{dt}/y_{dt}^2$

$$cv_{dt}^2|a_2, r_{dt}^2 \overset{\text{ind}}{\sim} \text{Gamma}\left(0.5a_2 n_{dt}, 0.5a_2 n_{dt}/r_{dt}^2\right), \tag{18}$$

Prior for overdispersion φ_r

$$\varphi_{dt} \sim \mathcal{N}_+\left(\gamma_t x_{\varphi,t}, \sigma_\varphi^2\right) \tag{19}$$

Nonparametric autoregressive priors of order 1 for (β_t, γ_t)

$$\beta_t \sim \mathcal{N}\left(\beta_{t-1}, \sigma_\beta^2\right) \tag{20a}$$

$$\gamma_t \sim \mathcal{N}\left(\gamma_{t-1}, \sigma_\gamma^2\right). \tag{20b}$$

We construct $x_{\varphi,t} = \left(\mathbf{1}_N, \frac{1}{\sqrt{\mathbf{n}_t}}\right)$, $\mathbf{n}_t = (n_{1,t}, ..., n_{N,t})^T$ in Equation 19 similarly to the cross-sectional model such the degree of overdispersion is inversely proportional to $\sqrt{n_{dt}}$. The prior distributions of Equation 20 is formulated as a dynamic linear model of order 1 (West 2013) with normally distributed innovations that represents a non-parametric local smoother that allows the data to estimate varying patterns of autocorrelation.

Analogous models are formulated for regions and for the total.

3. Simulation Study. Our goal in this section is to evaluate the model performance when we know the truth (for both a population indexed by units and domains in which the units nest) under a procedure that draws samples from a known population to produce domain estimates for an observed sample that we then compare to the population ground truth.

We construct a simulation scenario that resembles important characteristics of real data and the survey sampling mechanism. We first generate a finite population stratified by "geography" and a "size class" (e.g., discretized categories for number of employees for a business establishment unit) indicator and construct domain true statistics. Secondly, we draw a stratified simple random sample with replacement from the finite population and obtain direct sample-based and model-based estimates for the population domains. We note that all domains are not guaranteed to be sampled as is the case for states in the JOLTS sample where some states have no sample in a given month. These steps are repeated a large number of times to capture the variation of population generation and the taking of a sample. The estimates are evaluated against domains true finite population totals, where biases and mean squared errors of competing estimators are calculated over the simulations. Please see Technical Supplement in Savitsky et al. (2022) for details about population generation, the taking of a sample and construction of the direct estimator and its associated variance.

We do not generally have information about the true values of the domain variables when applying candidate models to real data in BLS. Using a fit statistic to assess the performance of models on the real data would be inappropriate because the purpose of the models is to uncover a latent truth, not to reproduce the observed survey estimates. So, BLS use simulation studies to assess the relative performance of proposed models as compared to direct estimation in realistic situations that mimic the real world setting as a basis for decision making.

We proceed to use our simulated data in a Monte Carlo simulation study to compare the model formulation of Section 2.1 that treats the domain sampling variances as known and fixed with the extension of Section 2.2 that treats the underlying true domain variances as unknown and specifies a likelihood for the estimated variances. Both formulations are compared to the direct estimator and we expect to see substantial reductions in estimation error relatively small-sized domains (with a small sample size).

3.1. *Simulation Results.* The models are estimated with the Hamiltonian Monte Carlo version of Metropolis-Hastings posterior sampling of Stan (Gelman et al. 2015) where each posterior sampling iteration provides a draw from the joint posterior distribution over the model parameters. We choose Stan over the use of a Gibbs sampler primarily because of the induced priors for φ_d and σ_d^2 in Equations 7 and 8, respectively. The induced prior distributions are not of closed form (and therefore highly non-conjugate) such that a Gibbs sampler would be relatively inefficient (due to the need to use some form of slice sampling (Neal 2003)). By contrast, Stan samples the joint parameter space on each posterior sampling (MCMC) iteration in a single sweep that utilizes the unnormalized joint log posterior distribution, so it doesn't care about conjugacy. Stan tends to also be more efficient than Gibbs sampling in terms of generating a larger effective number of parameter draws per posterior sampling (MCMC) iteration because the partial suppression of random walk Metropolis-Hastings in HMC makes the sampler less sensitive to a posteriori correlation among the parameters.

Each model run employed 5,000 MCMC iterations, with 2,500 iterations reserved for the warm-up and 2,500 iterations are used to compute model estimates as respective posterior means. We

domain type	units per sample	ave no of samples	Bias			RMSE			ratio of RMSEs	
			Direct	CS-FV	CS	Direct	CS-FV	CS	CS-FV	CS
1	31.7	200.0	1,494	983	-39	83,126	67,157	62,370	0.81	0.75
2	24.6	200.0	-365	-1,096	-604	64,401	47,732	43,006	0.74	0.67
3	16.4	200.0	-617	-1,058	-973	49,855	39,159	30,334	0.79	0.61
4	8.1	199.8	-440	369	95	35,176	29,636	16,679	0.84	0.47
5	1.5	111.7	9	1,390	129	6,445	5,385	2,014	0.84	0.31
Overall		182.3	17	-6	-318	56,990	44,967	39,022	0.79	0.68

Table 1: Bias and RMSE results for domains, grouped by five domain (population size) types (based on $A = 200$ runs) where “Direct" denotes the direct estimator, “CS-FV", the cross-sectional model under known true variances and “CS", the cross-sectional model under unknown true variances.

present results after $A = 200$ Monte Carlo simulations of the finite populations and respective samples.

We compare the direct estimator that we label as “direct" with our two model estimators. We label the cross-sectional model of Section 2.2 that jointly models the point estimates and variances, (y_d, v_d), under an assumption that the true variance is unknown as “CS". We also include a simpler model that supposes the estimated variance v_d is the known true variance that is treated as fixed, such that we only model the point estimate, y_d, as specified in Section 2.1. We label this model treating the domain variances as known as “CS-FV".

The usual statistics to evaluate estimators are an empirical bias of estimator $\hat{Y}_d$, $B_d\left(\hat{Y}_d\right) = A^{-1}\sum_{\alpha=1}^{A}\left(\hat{Y}_{d(\alpha)} - Y_{d(\alpha)}\right)$, and the square root of the mean of squared errors (RMSE), $RMSE_d\left(\hat{Y}_d\right) = \sqrt{A^{-1}\sum_{\alpha=1}^{A}\left(\hat{Y}_{d(\alpha)} - Y_{d(\alpha)}\right)^2}$,where $\hat{Y}_d$'s are respective estimators, direct sample-based or model-based; $Y_{d(\alpha)}$ is true finite population value and $\hat{Y}_{d(\alpha)}$ is a realization of $\hat{Y}_d$ at simulation run α, where $\alpha = 1, ..., A$.

We present biases and RMSE results in Tables 1-3. Table 1 shows results for domains, grouped by five domain (population size) types, as defined above, where type 5 contains the smallest number of population units and type 1, the largest. The employment size classes are equally distributed across each domain. While units (business establishments) containing relatively more employees are over-sampled in each domain, the probability that a domain appears in the sample (through the sampling of at least one nested unit) is proportional to its population size types (number of units). Domains containing more units are more likely to appear in each (Monte Carlo) sample than domains containing relatively few units. The number of units sampled within each domain is random such that it varies across realized samples.

Table 1 presents an average number of sampled domains over the repeated 200 Monte Carlo samples. Note that domains of type 5 are small and are not necessarily represented in each sample. This is reflected in the third column of Table 1, showing an average number of times a domain

units per sample	Bias			RMSE			ratio of RMSEs	
	Direct	CS-FV	CS	Direct	CS-FV	CS	CS-FV	CS
163	11,838	9,011	9,381	173,344	138,994	121,371	0.80	0.70
163	4,301	7,326	-304	175,199	147,740	128,409	0.84	0.73
163	6	-4,864	-3,085	204,108	157,705	143,692	0.77	0.70
163	-14,651	-11,566	-20,781	256,677	213,853	196,755	0.83	0.77
Overall	373	-23	-3,697	205,113	167,145	150,483	0.81	0.73

Table 2: Bias and RMSE results, regional (based on $A = 200$ runs) where “Direct" denotes the direct estimator, "CS-FV", the cross-sectional model under known true variances and "CS", the cross-sectional model under unknown true variances.

units per sample	Bias			RMSE			ratio of RMSEs	
	Direct	CS-FV	CS	Direct	CS-FV	CS	CS-FV	CS
817	1,494	-93	-14,789	403,256	360,231	358,282	0.89	0.89

Table 3: Bias and RMSE results, national (based on $A = 200$ runs) where “Direct" denotes the direct estimator, "CS-FV", the cross-sectional model under known true variances and "CS", the cross-sectional model under unknown true variances.

of type 5 gets into a sample: this value is 111.7, whereas the larger domains are represented in all 200 sample realizations. In the case a domain does not appear in a sample realization in some Monte Carlo iteration, its point estimate is imputed under both the CS and CS-FV models and the sampling variance is also imputed under CS. Table 1 reveals that the RMSE ratios between both models and survey direct estimate are always less than 1 for every domain type, but the improvement in error is particularly large for domains of types 4 and 5 at the expense of a relatively small amount of increased bias as compared to the survey-based direct estimator. The CS model produces a lower RMSE ratio to the direct estimator than does CS-FV for every size class owing to the additional estimation strength borrowed by co-modeling the domain variances.

Table 2 presents results for domains rolled up to “regions”, and Table 3 shows results for the overall population total.

Figure 1 presents results in a graphical form for the comparison of fit performance between the survey-based direct estimate and the modeled estimate by plotting the distribution over the Monte Carlo iterations of the log of relative deviations constructed as, $\log\left(\frac{\hat{Y}_{d(\alpha)}}{Y_{d(\alpha)}}\right)$, for domains that are present in each sample, over simulations α, $\alpha = 1, ..., A$, where the numerator is the estimated and the denominator is the truth. A distribution is rendered for each domain type where type 5 represents domains containing the smallest number of population units and type 1, the largest. Each distribution is presented in the form of a “violin" plot that mirrors the distribution. Its interpretation is similar to a box plot with additional information for the distribution shape. The distribution plots on the right represent the survey-based direct estimates and those in the middle represent

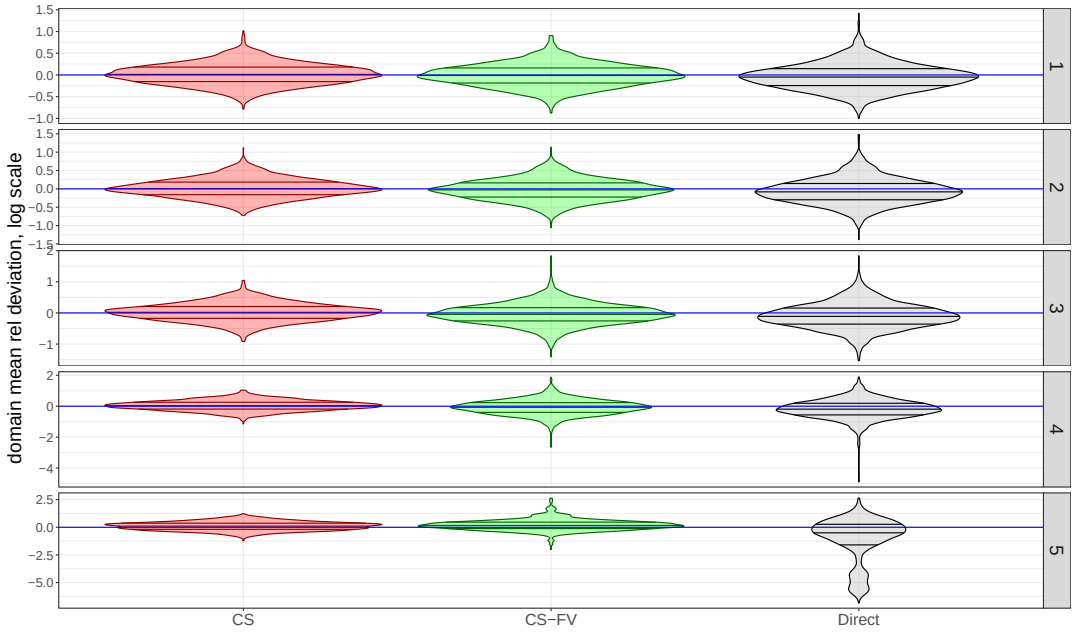


Fig 1: Relative errors of point estimates, over simulations, by domain type (where 5 contains the smallest number of population units and type 1 contains the largest), presented on a log scale. ”CS" denotes the cross-sectional model under unknown true variances, ”CS-FV", the cross-sectional model under known true variances and “Direct" denotes the direct estimator, from left-to-right.

the CS-FV model-based posterior means, while those on the left represent the CS model-based posterior means. The solid horizontal line in each plot panel represents the population true value of the domain point estimates. We see that while the Monte Carlo distribution over all estimators is centered on the true value, the distributions for the model-based estimators are more concentrated on the true values, with the CS model expressing the smallest variance around the truth.

Figure 2 is an analogous presentation of $\log\left(\frac{\hat{V}_{d(\alpha)}}{\bar{V}_{d(\alpha)}}\right)$ for the sample-based estimates of *variances* of direct estimator and posterior means of the CS model fitted/smoothed estimates of sampling variances (σ_d^2). The solid horizontal line contained in each plot panel of Figure 2 represents the true sampling variance of the point estimator that is approximated over the Monte Carlo iterations by each of the direct sampling variance and the CS model. In the case that there are no sampled units for small domains (that contain relatively few sampled units) the sampling variances are imputed by the CS model for those missing domains. The CS-imputed variances represent the quality of the point estimator based on the pattern of borrowing strength from other domains and their underlying estimated variances. The figure reveals that the modeled variances under the CS model (in the left-hand set of plots) are more concentrated than the direct estimates (in the right-hand set of plots), similar to the point estimates.

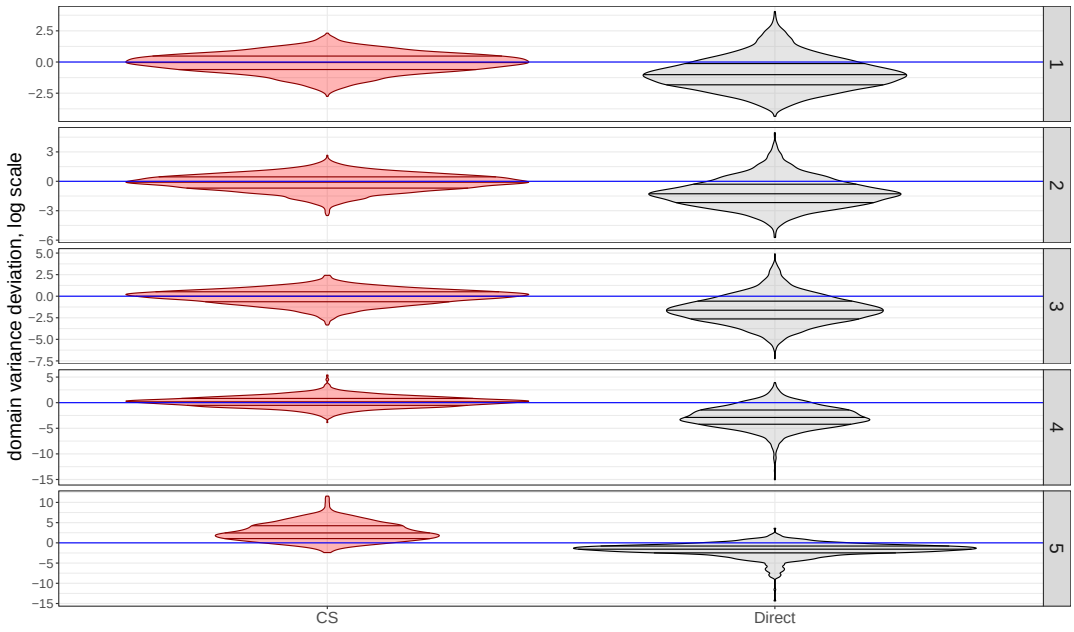
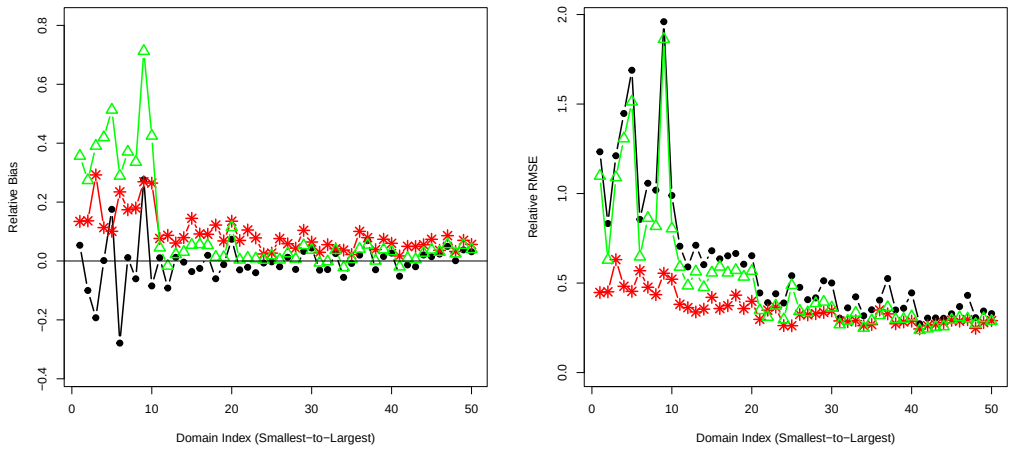


Fig 2: Relative errors of variances of direct estimates, over simulations, by domain type (presented on a log scale). ”CS" denotes the cross-sectional model under unknown true variances and “Direct" denotes the direct estimator, from left-to-right.

There is the appearance of a slight positive bias in the fitted sample-based variance estimates (on the left) for domains of types 4 and 5 with fewer population units shown in Figure 2, but this apparent bias is likely not real. Our Monte Carlo approximation procedure for the true variance suffers from the exclusion of the null sample realizations for type 4 and 5 domains. The procedure excludes the null cases for any domain because we are not able to form a sample-based point estimate in any case where the realized sample in any Monte Carlo iteration excludes a domain. As a result, our Monte Carlo approximation method is expected to under-estimate the true sampling variance for such smaller domains by excluding the null sampling event. The higher is the probability of the event that a sample realization excludes a domain, the larger would be the under-estimation for our Monte Carlo approximation for the true variance. Such is likely why we see the appearance of a larger bias for type 5 domains than type 4 (since the probability of a null sample for type 5 domains is higher).

Figure 3a draws a line plot of relative biases $relB_d\left(\hat{Y}_d\right) = A^{-1}\sum_{\alpha=1}^{A}\left(\frac{\hat{Y}_{d(\alpha)}-Y_{d(\alpha)}}{Y_{d(\alpha)}}\right)$ and Figure 3b draws a line plot of relative root MSE $relRMSE_d\left(\hat{Y}_d\right) = \sqrt{A^{-1}\sum_{\alpha=1}^{A}\left(\frac{\hat{Y}_{d(\alpha)}-Y_{d(\alpha)}}{Y_{d(\alpha)}}\right)^2}$, where the x-axis indexes domain labels sorted by the average number of sampled units in each domain from left-to-right. The line connecting dots represents the direct estimator, the line con-



(a) Relative bias (b) Square root of relative MSE

Fig 3: Relative biases and square root of relative MSE of point estimates for 50 domains sorted by the number of respondents. The line connecting dots represent our baseline direct estimator, stars represent the CS model, whereas triangles represent the CS-FV model.

necting triangles represents the CS-FV model estimator and the line connecting stars represents the CS model estimator.

Figure 1 demonstrates that the modeled estimates produce lower true sampling variances than do the survey-based direct estimators. Figure 3a highlights that the absolute bias is slightly higher, however, particularly for domains with smaller number of sampled units, though the relative MSE improvement for these domains with lower sample sizes are much larger as seen in Figure 3b.

Our model would be expected to outperform direct estimation under any positive correlation between x_d and y_d because in the worst case setting of no correlation there would be less shrinkage away the direct estimator in the mean parameter, θ_d, that we use our model-based estimator. The correlation between y_d and v_d, however, would provide additional estimation for θ_d.

As the sample size per domain increases, the model estimate will generally contract on the direct estimate, particularly in a flexible model such as ours. In our experience, one may use the direct estimate in lieu of the model estimate when the number of units per domain is greater than 100, though some may focus on a target coefficient of variation for favoring the direct estimate. The direct estimate has the advantage that it is both design unbiased and immune to model misspecification.

In summary, our Poisson-lognormal joint model for (y_d, v_d) would be expected to perform well in the case where there are domains with small counts as we see in the JOLTS due to its ability to handle distribution skewness induced by overdispersion. The performance on domains with larger counts would also be robust, but at the expense of less efficient computation than with a model that treats these domain point estimates as continuous. We would recommend our model under the set-up where some domains express small counts that induce overdispersion.

4. Application to JOLTS. We now consider state/industry by-month estimation for the JOLTS job openings variable using a predictor formed from a census instrument that we describe below. Our model estimations for job openings covers the period from January 2014 to December 2018 for up to 52 states (including Puerto Rico and Washington, DC) depending on the industry being modeled. We compare the smoothing and imputation performances of our 3 model formulations: 1. The cross-sectional (CS) model is separately estimated for each industry and month. In other words, each model is fitted on a set of States in a given industry at a given month; 2. The cross-sectional model under fixed and known variances (CS-FV); 3. The multivariate time series (MV) model jointly models the collection of months in the measurement period for the set of states in each industry, unlike the CS model which estimates one month, at-a-time.

Our model formulations, in addition to state levels, includes likelihood contributions for regional and national levels (for each industry). As a result, model-fitted estimates for regions and national estimates for job openings are also available as a by-product of the model. It is expected that at higher levels / lower resolutions model-fitted estimates should be close to respective sample-based estimates. A benefit of incorporating these lower resolution regional and national sample-based direct estimates into the model is that higher resolution state-level model-fitted estimates are forced to add up to respective model-fitted regional estimates, which in turn add up to the model-fitted national estimates. This construction is similar to traditional benchmarking, except that we do not require additivity exactly to sample-based lower resolution estimates; instead, we require additivity to model-based lower resolution estimates, which, as we have mentioned are close, yet not exactly equal to, sample-based estimates.

Another feature of the models is that they simultaneously imputes a value for those months where sample-based estimates are not available (i.e., when there is zero number of respondents.)

4.1. *Synthetic Predictor, x_d, for JOLTS Estimation.* JOLTS has designed a synthetic estimator for each state-level domain that we use as a predictor in our modeling. We now briefly describe this estimator. The JOLTS synthetic estimator is constructed by leveraging a census instrument of business establishments, the Quarterly Census of Employment and Wages (QCEW), maintained by BLS for the purpose of measuring employment. The QCEW is administered quarterly and processing time results in a lag of many months before results are published. The QCEW does not publish the components of employment measured by JOLTS, but to obtain the synthetic estimator, the necessary JOLTS variables are "imputed" in a Hot Deck-like procedure to each record in the QCEW (in a 1 year lag to the current month). This is done, for each month, by stratifying the QCEW records by the intersection of employment change trend (e.g., increasing, decreasing, flat), industry classification and employment size. A data record is randomly drawn from the JOLTS survey data within each stratum and the ratio job openings to total employment for that JOLTS data record is used to impute job openings values of every QCEW record (that contains the employment level for some business establishment) in that stratum. The procedure is repeated until the level of job openings is imputed for every record in the QCEW. The sum of such imputed job openings, across the QCEW records, to each state – by - industry (labeled state/industry) domain provides a "synthetic" estimate. These imputed state/industry synthetic estimates constructed for job openings are further benchmarked to the actual JOLTS sample estimates, by region, to ensure the imputed state levels for job openings equates to the actual JOLTS survey openings levels at the regional

and national levels. Similar procedure is performed for other JOLTS data items. As noted, such a synthetic estimator is available based on the lagged QCEW data and does not use the current period JOLTS information.

The synthetic predictor encodes historical expectations (on a one year lag) for each state/industry domain, while the "direct" sample-based estimates supply current information. Synthetic estimates are less variable but they may be biased. By contrast, direct estimates are considered unbiased, but they have large variance when the sample is small. Model-fitted estimates represent a compromise between "synthetic" and "direct" estimates. The model automatically "decides" on the balance between the synthetic and direct components. This "decision" is made based on the relative variability of the direct sample-based domain estimates and quality of the model fit, which, roughly speaking, depends on how well overall synthetic estimates explain sample-based results.

4.2. *Results.* Figures 4 and 5 illustrate and compare the relative smoothing and imputation performances of our CS, CS-FV and MV model formulations with the sample-based direct estimate. Each figure shows alternative estimates for the January 2014-December 2018 period for a single state/industry domain. The dotted line represents CS-FV, the dotdash line represents the CS estimates and the dashed line, the MV estimates. The solid line represents sample-based direct estimates. Figure 4 presents results for a relatively smaller domain with fewer population business establishment units such that we note the monthly gaps in the solid line where there are no sample-based direct estimates available for that state/industry domain from the national sample in those months. We observe that all of the models, nevertheless, provide imputed values for these missing months. The models borrow strength among states with similar patterns of job openings, as well as additionally borrowing strength from the associated regional estimates.

The MV borrows further information from the autocorrelation of the monthly estimates co-modeled under this formulation for each state/industry domain. This additional time-indexed dependence induces a higher degree of over-the-month smoothing. While the MV model estimation is too computationally-intensive to include within our Monte Carlo simulation study (due to repeated estimations), we ran a few iterations under a simulated dataset generated with autocorrelations and observed that the greater smoothing provided by the MV relative to CS resulted in a slightly, but not notable, improvement in performance as compared to the CS, which is what we also observe in the JOLTS application. For the models estimated on the mining and logging industry, the CS model estimates in about 500 seconds, while the 12 month MV model estimates in about 5500 seconds. For this domain containing relatively few business establishment units we see that both models substantially smooth the noisy sample-based direct estimate.

Figure 5 presents the estimation results for a relatively large domain that contains many business establishment population units such that all months are observed. As expected, the models perform relatively less smoothing of the survey-based direct estimates as these direct estimates are more reliable such that they express lower variances. For this domain, the CS and MV are relatively coincident with one another with only a small degree of additional smoothing induced by MV. The CS-FV shows somewhat less smoothing of the direct estimator with more pronounced peaks due to excluding the variances that are useful to strengthen estimation. The MV model is expected to outperform the CS model in the presence of any autocorrelation so long as the time series pattern doesn't dramatically change due to a "regime shift" event, such as an economic downturn. During

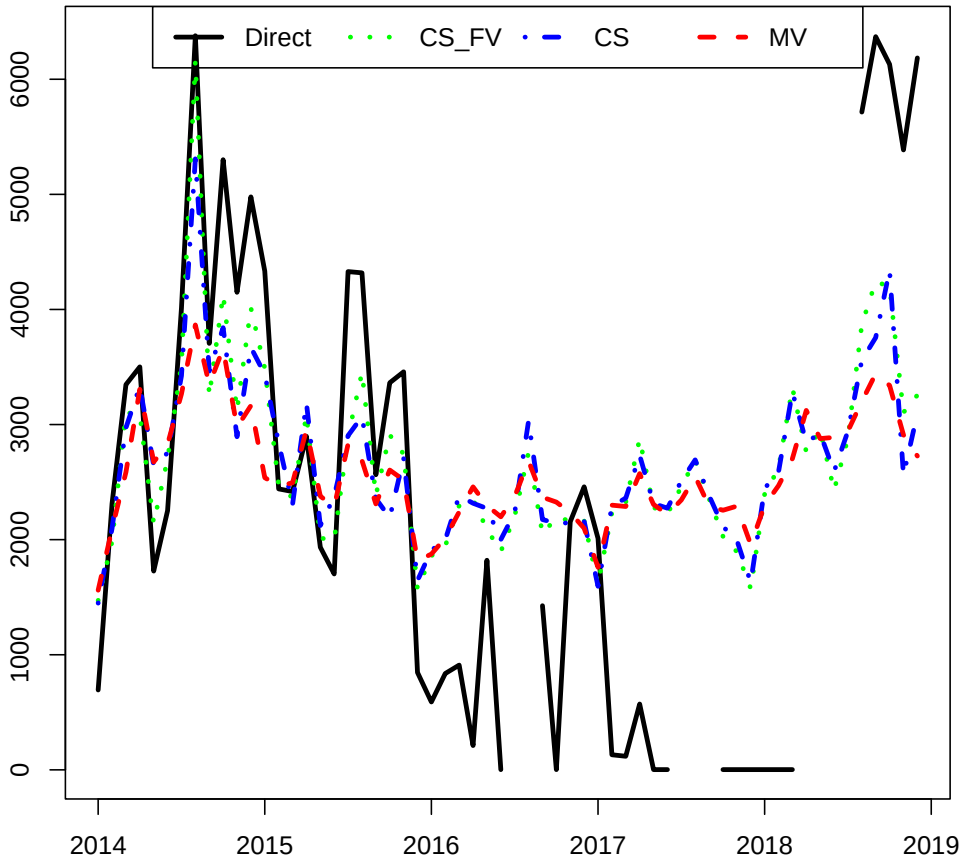


Fig 4: Example 1 of domain direct vs cross-sectional (CS) and cross-sectional under known variances (CS-FV) vs multivatiate (MV) time series model fitted estimates, over the period from January 2014 – December 2018, average number of respondents over the period is 4.5

and after a regime shift event, the correlation structure over time between JOLTS variables would be expected to change such that past month observations before the regime shift would no longer be as useful. In such a case the CS model would outperform.

5. Discussion. We introduce a novel general Bayesian hierarchical model formulation for the analysis of survey variables of the count data type that extends the automatic property from con-

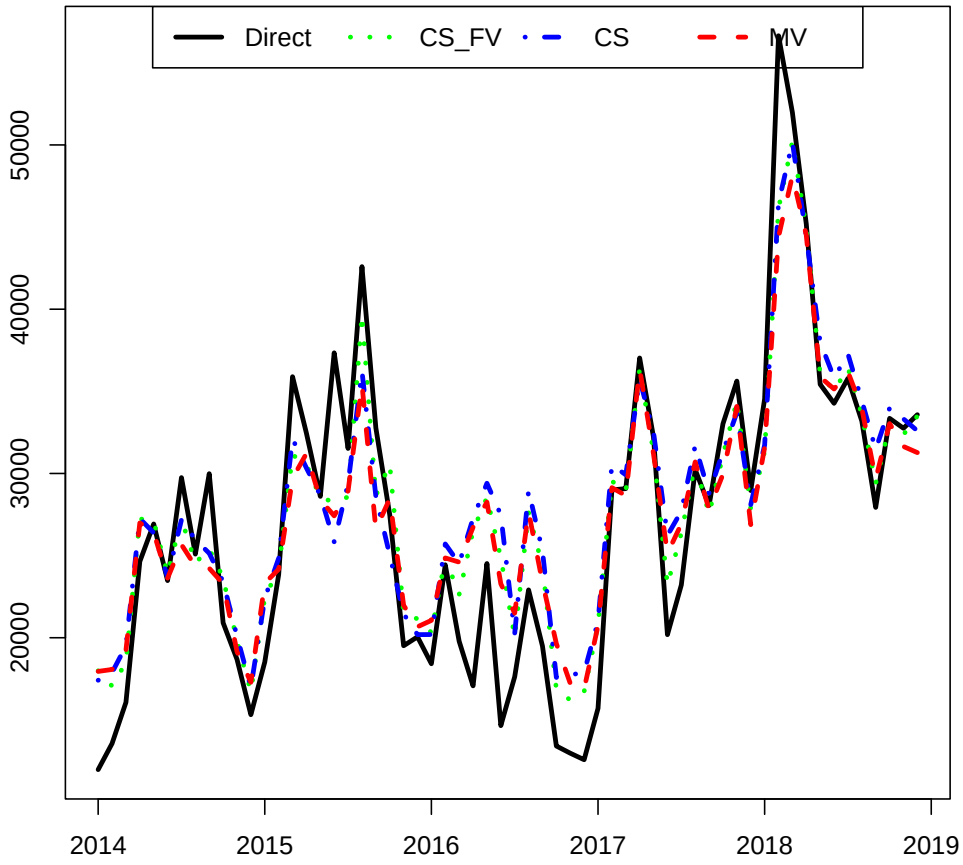


Fig 5: Example 2 of domain direct vs cross-sectional (CS) and cross-sectional under known variances (CS-FV) vs multivatiate (MV) time series model fitted estimates, over the period from January 2014 – December 2018, average number of respondents over the period is 53.2

tinuous data model formulations that the estimated true variance is set equal to the generating variance for the point estimate. We set the variance of our Poisson-lognormal construction for the point estimate to the true variance mean of the gamma distribution prior on the observed survey variances. The prior on the true variances is induced by from the prior on the mean and overdispersion parameter of the Poisson-lognormal set-up.

Our modeling approaches further incorporates benchmarking of higher resolution (e.g., state level) estimated totals (which are functions of model parameters) to lower resolution (e.g., regions) totals simultaneously with their estimation. This multiresolution construction both sharpens estimates of the higher resolution totals through their nesting in more reliable lower resolution aggregations of domains linked to lower resolution survey-based direct estimates and collapses the usual two-step process of estimating the higher resolution quantities in a model and benchmarking those to lower resolution totals. Our procedure avoids the reintroduction of noise to lower resolution estimates that typically occurs under the two-step process.

We demonstrated in a simulation study that constructed ground truth estimates for population domains containing varying numbers of underlying units that our count data modeling framework produces lower error estimates than the sample-based direct survey estimates and those improvements are dramatic in domains contain relatively few population units. The model provides these reduction in errors while simultaneously producing imputed estimates for non-sampled domains. Lastly, our application to the JOLTS job openings variable demonstrated that our multivariate model extension provides more smoothing than our base, cross-sectional model by borrowing information from over-the-month correlations.

SUPPLEMENTARY MATERIAL

Technical Supplement: Technical Appendices
The online supplement contains three technical appendices with detailed material on the following topics:

1. Simulation study detailed design;
2. Stan (Gelman et al. 2015) scripts for models;

REFERENCES

Bradley, J. R., Wikle, C. K. & H, S. (2016), 'Bayesian spatial change of support for count-valued survey data with application to the american community survey', *Holan* **111**(514), 472–487.

Gelman, A., Carlin, B., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2014), *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*, Chapman and Hall/CRC.

Gelman, A., Lee, D. & Guo, J. (2015), 'Stan: A probabilistic programming language for bayesian inference and optimization', *In press, Journal of Educational and Behavior Science* .
URL: *http://www.stat.columbia.edu/*

Maiti, T., Ren, H. & A. (2014), 'Prediction error of small area predictors shrinking both means and variances,', *Sinha* **41**, 775–790.

Neal, R. M. (2003), 'Slice sampling', *The Annals of Statistics* **31(3)**, 705–767.

Rao, J. & Molina, I. (2015), *Small Area Estimation*, Wiley Series in Survey Methodology, Wiley.
URL: *https://books.google.com/books?id=i1B_BwAAQBAJ*

Savitsky, T. D. (2016), 'Bayesian nonparametric multiresolution estimation for the american community survey', *The Annals of Applied Statistics* **10**(4), 2157–2181.
URL: *http://www.jstor.org/stable/44252230*

Savitsky, T. D., Gershunskaya, J. & Crankshaw, M. (2022), 'Supplement to "Bayesian Model for Survey Count Data Point and Variance Estimation"', *Annals of Applied Statistics* .

Sugasawa, S. & Kubokawa, T. (2020), 'Small area estimation with mixed models: a review', *Japanese Journal of Statistics and Data Science* **3**(2), 693–720.
URL: *https://doi.org/10.1007/s42081-020-00076-x*

Sugasawa, S., Tamae, H. & Kubokawa, T. (2017), 'Bayesian estimators for small area models shrinking both means and variances', *Scand J Statist* **44**, 150–167.

Tzavidis, Nikos and Ranalli, M Giovanna and Salvati, Nicola and Dreassi, Emanuela and Chambers, Ray (2015), 'Robust small area prediction for counts', **24**(3), 373–395.

West, M. (2013), Bayesian dynamic modelling, *in* P. Damien, P. Dellaportas, D. Polson & N. G. Stephens, eds, 'Bayesian theory and applications, chapter 8'.

Zhou, M., Li, L., Dunson, D. & Carin, L. (2012), Lognormal and gamma mixed, *in* 'Negative Binomial Regression. Proc Int Conf Mach Learn', Vol. 2012, PMID: 25279391; PMCID: PMC4180062, p. 1343–1350.