Construction and Assessment of Generalized Variance Functions for an Establishment Survey[1] November 2017

Kevin Roche
Julie Gershunskaya
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave, NE, Washington, DC 20212

Abstract
The Bureau of Labor Statistics Current Employment Statistics (CES) survey leads to the creation of measures which are used to create Principal Federal Economic Indicators. One of those measures is the monthly change in establishment jobs, which CES estimates for detailed industries as well as geographic domains such as states and metropolitan areas. CES calculates sampling errors using Balanced Repeated Replication, but estimates of sampling variance can be volatile and may not be available as readily as desired. Generalized Variance Functions (GVFs) are fit using regression models to existing direct estimates of sampling variance to improve estimates of those variances. A GVF should provide good fits for data used to construct the model, and it should provide good estimates of variances for other observations not used in the model. This paper develops a GVF model for ratio estimators and considers two main characteristics: accuracy in terms of confidence interval coverage, and stability. Metrics and tests for each characteristic are developed. We also consider and test new model types to further increase coverage and stability of the GVF variances.

Key Words: Generalized Variance Functions, R, Establishment Surveys, Balanced Repeated Replication, Cluster Variables

1. Introduction

The Bureau of Labor Statistics Current Employment Statistics (CES) survey provides employment estimates based on information reported from business establishments. This information leads to the creation of the Principle Federal Economic Indicator on The Employment Situation. The CES survey is based on approximately 147,000 businesses and government agencies representing approximately 634,000 worksites throughout the United States [1].

The CES sample is a stratified, simple random sample of worksites, clustered by Unemployment Insurance (UI) account number. The CES program uses a matched sample, which measures the change in employment for each establishment, but excludes establishments that reported zero employees for either month. With the matched sample, a weighted linked estimator is used for both estimates and variances. We use Balanced Repeated Replication to determine our variances, ensuring our samples use the same stratification of location, industry, and size class [2].

One issue with our estimated variances is that the true monthly variability is hard to distinguish from random irregularity of the true variances. True variances are never known

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

since a researcher can only get estimates of the true variances. If historical data are available for the whole population, the “true” variance and the respective coverage properties could be investigated based on repeated samples from the population [3]. We use GVFs to utilize information about the sampling process that can help create accurate variances that are more stable.

CES creates estimates of employment levels and monthly changes in employment for various locations and industries every month. As such, we also need to produce measures of uncertainty (e.g., variances) for these estimates. Locations could be either Metropolitan Statistical Areas (MSAs) or states, whereas industries are broken down into various categories by their North American Industry Classification System (NAICS) codes. We found that monthly variances changed dramatically from year to year, when we would otherwise expect them to stay about the same. This finding led CES to use a 3-year average of the median of months in a year as CES’s initial GVF because of its increased stability from year to year. This model has its problems such as its inability to account for large or monotonous changes in the employment over time, the necessity of multiple years of data for stability, less stability than desired even with multiple years of data, and consistent underestimation of variance in seasonal changes in employment. For these reasons, we looked at other GVFs to replace it.

The results of our models come from the use of statewide data, where industries were the estimated super sectors (e.g. all retail NAICS codes) on monthly data. Each estimated super sector was modeled separately using R, specifically the *rlm* function for robust linear modeling in the MASS package.

2. Model Creation

To create a model that is better than the 3-year average model, we wanted to create a GVF using survey information that is available to us before the estimate is even created. For each domain *d*, consider the following variables that were used to create the new GVFs:

1. Employment at benchmark month: $Y_{d,0}$
2. Average unweighted sampled employment: $Y_{d,s}$
3. Average number of respondents (UI): n_d
4. Average size of sampled establishments: $\frac{Y_{d,s}}{n_d}$
5. Part of Finite Population Correction: $\left(\frac{Y_{d,0}}{Y_{d,s}} - 1\right)$
6. Cluster effect: $X_{d,s,c}$

Employment at benchmark month refers to the total number of employees for the strata in the Quarterly Census of Employment and Wages at the benchmark month. The variance of employment is directly related to the size of employment in the strata. This lead to the creation of Model 1 (M1):

M1: $\log\{Var(\widehat{R_d})\} = \beta_0 + \beta_1 \log(Y_{d,0}) + \varepsilon_d$

The average unweighted sample employment and average number of respondents provide important information that relate to the finite population correction, but are correlated with each other and with the benchmark level of employment. Model 2 is as follows:

$$\text{M2: } \log\{Var(\widehat{R_d})\} = \beta_0 + \beta_1 \log(Y_{d,0}) + \beta_2 \log(Y_{d,s}) + \beta_3 \log(n_d) + \varepsilon_d$$

The previous model's multicollinearity is a problem for consistent parameter estimates from year to year, so a natural modification was made to the variables relating to the finite population correction. Further explanation for this modification can be found in the appendix. Model 3 is as follows:

$$\text{M3: } \log\{Var(\widehat{R_d})\} = \beta_0 + \beta_1 \log(Y_{d,0}) + \beta_2 \log\left(\frac{Y_{d,s}}{n_d}\right) + \beta_3 \log\left(\frac{Y_{d,0}}{Y_{d,s}} - 1\right) + \varepsilon_d$$

One issue that all of our models have is that they are predicting the variance based on variables that are not changing from month to month. Our model can't predict seasonality which has been shown to exist in certain industries. Looking at the average confidence interval length for transportation and education across time shows a clear seasonal effect:

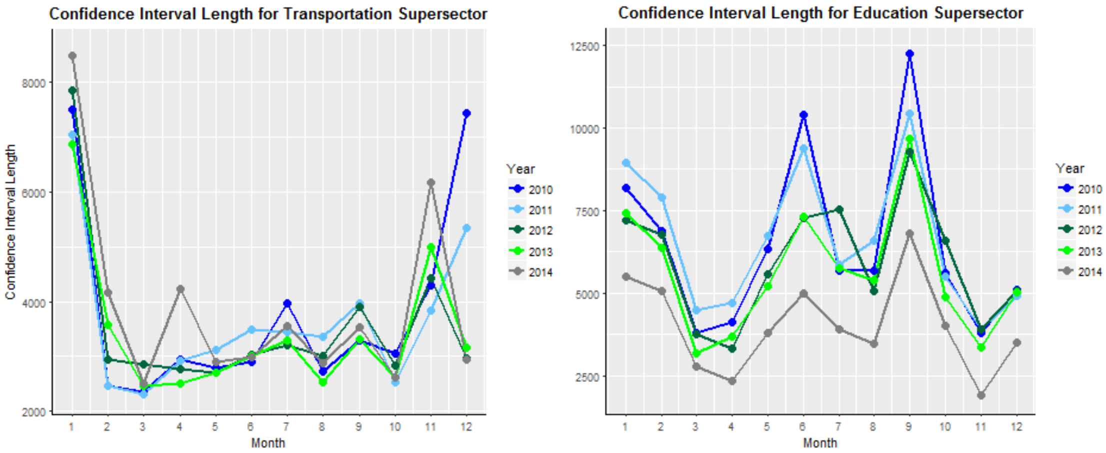
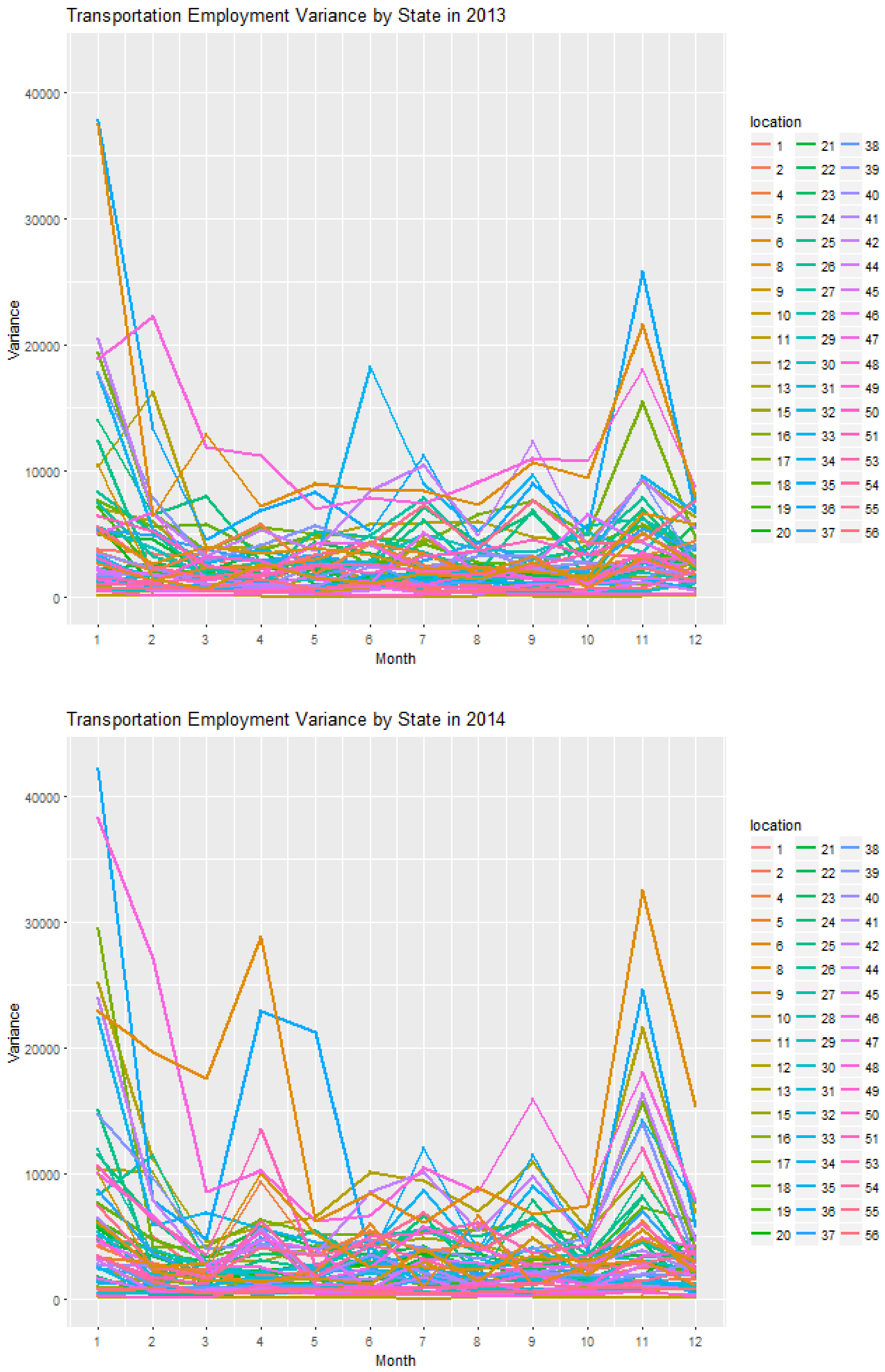


Figure 1: Average of state wide confidence interval lengths of the transportation and education industries employment estimates show seasonal patterns that are typically repeated each year.

Initially, we tried models with a monthly effect via dummy variables, but this was not successful. The reason this model fit poorly is because the size of the monthly effect is not the same in all locations. Likewise, the confidence interval lengths of individual states do not always show the same yearly pattern as the average of all states. The patterns of individual states are commonly repeated across years, but not always, as Figure 2 shows with the transportation industry in 2013 and 2014.


Transportation Employment Variance by State in 2013

Transportation Employment Variance by State in 2014

Figure 2: Locations are shown by state FIPS code. The seasonal pattern of transportation variance with an increase in January and November appears in both 2013 and 2014, but the amount of variance increased differs greatly from state to state. Some states also do not follow the same seasonal pattern that others do.

This discovery led us to group observations by introducing indicator variables, which we called cluster variables. Cluster variables were created to group the relatively higher variance months of different locations together, similar to regression trees. There are many ways to create cluster variables, but these were created as dummy variables with value 1 if the variance was higher the median of its state times some value *c* and value 0 otherwise. As an example, let's say transportation in Alabama had a variance in January that was more than 3 times the median of Alabama's monthly variances in transportation, and Alaska had a variance in November that was more than 3 times the median of Alaska's monthly variances as well.

Alaska's monthly variances in transportation. The cluster variable, where c=3, would be 1 for Alabama's variance in January and Alaska's variance in November. This would allow the increase of variance for these month-state combinations without over fitting.

When using cluster variables, we first use M3 to explain the yearly average of a particular state and industry's variance. As such, we are only predicting the residuals. We tested many different ways to utilize more cluster variables than data points, such as elastic net and a naïve average of parameters ensemble. The average ensemble was conducted such that only one or two of the cluster variables were being modeled at any given time. For example, when predicting one variable at a time, with 10 different cluster variables, the end result model would have all 10 clusters with 1/10 of the parameter estimate of the one model it was used in.

In our use of the naïve average ensemble model, each model of one variable is predicted separately and then combined with each of the linear models into another model made from the average of the coefficients for each cluster variable.

$$\frac{1}{n}\sum_{i=1}^{n} Residuals \sim C_i = \text{Residuals} \sim \frac{1}{n}\sum_{i=1}^{n} C_i$$

This would allow us to use a minimum of 1 data point per model (as intercept is not used) while allowing us to create *n* models. That is obviously not ideal, but it allowed us to create 22 cluster variables in our model (11 per year) even if our data includes only 10 different states in any particular industry class. For this data, we have as many as 50 points per month, but sometimes a state doesn't have employment in a particular industry classification.

Depending on the goals (stability or accuracy) different sets of cluster variables could be used, such as using more than one year's cluster variables or creating models with more than one cluster variable at a time. We will show the results of using 11 cluster variables per year for two years (22 variables), where "22 choose 1" variable models were used, and where "22 choose 2" variable models were used. Penalized regression models were attempted that gave good, but less desirable results, compared to the naïve average ensemble.

3. Measuring Accurate Variances

When making an estimate of the variance with a GVF, we need a way to determine if the variance estimate is better. We used 3 different measures to determine how good a variance estimate is: coverage properties of confidence intervals, confidence interval length, and stability.

In this paper, we do not compute "true" coverage over repeated samples for each domain. Instead, we use an approximate method considered by Gershunskaya and Dorfman [3]. The method uses one or a few replicates from the Balanced Repeated Replication procedure to form certain pivotal quantities. If the variances used to form the pivot are close to true sampling variances, the pivot should have the standard normal distribution. Thus, under certain assumptions, the pivot allows us to evaluate the characteristics of the variance estimates using a set of domains instead of performing full scale simulations by drawing repeated samples. Using the replicate estimate, estimate based on the whole sample, and the modeled variance estimate, we form the pivot and create respective 95% confidence interval assuming the normality of the pivot. Then we compute how many times over the set of domains the interval contains zero. Here we would expect 95% of all replicate estimates to be within the 95% confidence interval of the whole sample estimate.

With a limited number of subsamples, we will not be able to test each individual GVF estimate for 95% coverage. Therefore, the average or total confidence interval length needs to be addressed, as we will be estimating the variances for many industries and locations at once. If variance estimates were excessively high, we could always have 95% or higher coverage. Therefore, when comparing models, it is important to realize that coverage isn't the only factor necessary to compare the accuracy of a model's predictions.

To simplify the process of comparing models, we compare the coverage percentage of models with the same average confidence interval length. To fix the confidence interval length to be the same, we multiply all variances by the average confidence interval we want divided by the average confidence interval of the model. The average confidence interval we want is based on the previous year's calculated variances. We do this process for each industry super sector. The reason we modify each industry separately is because each linear model we made was created at the industry level, for all locations. With coverage at a set confidence interval length, we now have a way to measure accuracy in a way that is comparable to other models.

Fixing the confidence interval length is also a possible solution to consistent underestimation of the variance, which we saw in our 3-year average model that uses the median of months. If we use the most recent year's variance estimates to guide the average confidence interval length, we can adjust the 3-year average to increase the coverage with reasonable assumptions. The adjustment has no change with a multiplicative value of 1. Depending on the model method adjustments either increased or decreased them. For example, the 3-year average had an average 1.14 multiplicative increase but ranged from .77 to 1.51 over all industries for the 4-year period. The other 4 models had average adjustment values between .86 and .90, typically lowering their variances. Given those adjustments, it may seem like our models overestimated the variance, but the models were predicting the next year's variance that is typically increasing with employment, not the previous year.

If we were to use last year's variances to estimate the next year's estimates, instead of our GVFs, the estimates would change a lot from year to year, much more than we would realistically expect them to change. The initial GVF, the 3-year average, reduces the amount any particular estimate will change, but can still lead to some dramatic increases. We measure the change by using the absolute relative difference with the minimum function as the denominator. For example, a variance estimate using the 3-year average can differ by over 200% from one year to the next. In other words, the variance estimate could increase nine fold, or decrease to 1/9th in size from one year to the next.

We measure stability by looking at the change in a model's prediction from year to year. When calculating the absolute relative difference for many variance estimates, you end up with a distribution of the stability. To compare stability, we use measures of the distribution, such as the mean, median, or max. We would like to reduce the maximum change in predictions, as dramatic changes are not to be expected without an equally drastic change in sample size or employment in that industry.

For the same level of accuracy, we would want to have as little change in predictions as possible. That being said, the economy and employment is constantly changing, so we would not expect the variance of our estimate to be a constant over time, either.

4. Comparison of Models

Now that we have measures for a comparable accuracy and stability of model predictions, we can compare the effectiveness of each model. The coverage of each model and method is compared on a fixed CI based on the previous year's average CI length, for each industry super sector. Even the current year's variance estimate is modified to use the same average CI length so that we can compare the model to the coverage of the calculated variances for that year.

The first thing we can see is that the sample variance is able to get approximately 95% of all subsample variance estimates, as we expected. The sample variance percent did not change much when using the previous year's CI length, which shows that the average super sector variance does not change much from year to year, even if individual variances can change at a median rate of 38%.

Using the coverage graph in figure 3, the large ensemble had the highest coverage, followed by the smaller ensemble, M3, the 3-year average, and then M1. If stability was not important, the large ensemble would be the best choice, as it is the most accurate at predicting the next year's variance. When we consider stability as well, we see that the order is almost exactly the opposite from coverage. The exception to this would be that the 3-year average's max change is larger than any other model.

One fact that shows the natural variation in the calculated variances is that the previous year's variance was the worst predictor of the next year's variance subsample estimates. Since one year's variance estimates are made from the same year's subsample variances, it makes sense why it is the best predictor. When looking at stability, using last year's variance value to predict the next year's variance is not stable at all, with a higher median change than any other method at 38%.

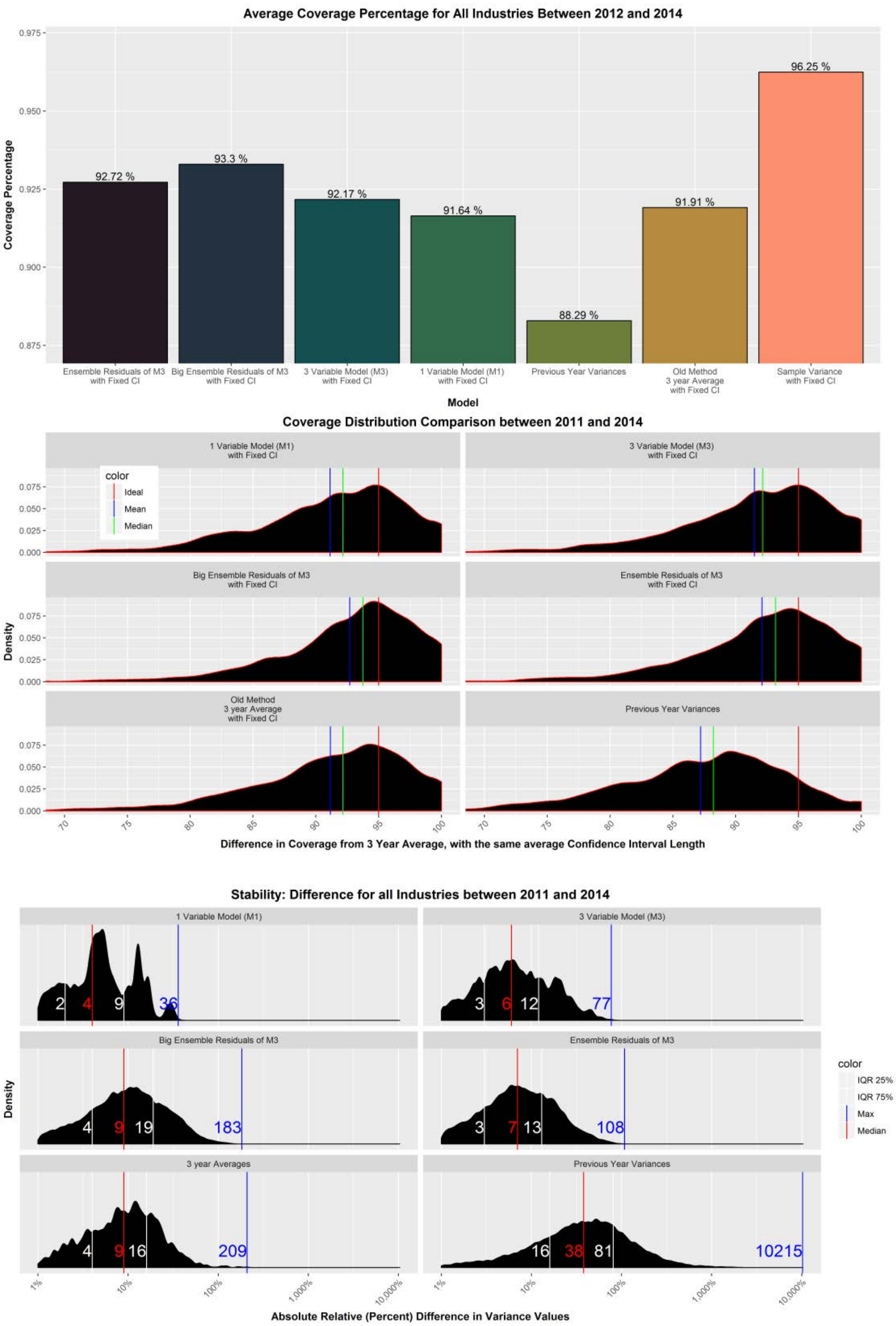


Figure 3: Metrics of model comparison from top to bottom: coverage, coverage distribution, and stability distribution.

The coverage distribution chart is important because we want to ensure that each industry and state is represented accurately, not just all states being represented more accurately on average. The values were cut off around 70% in the graph, but the large and small ensembles had a minimum coverage of 59.8% and 58.8% while the other models had between 50-55% minimum coverage, and the previous year's variance had as low as 34.2% coverage for any state and industry combination's yearly average coverage.

Conclusion

The use of GVFs allows us to predict the next year's variance with a high degree of accuracy while allowing us to also have more stable variance estimates than the use of the calculated sample variances themselves. Depending on importance of stability or accuracy, different models are able to provide more accurate or more stable results, but in our case, rarely both. The use of cluster variables allows us to have a higher degree of accuracy when a seasonal effect exists, but there is too little data for each state and industry to have a time series.

References

1. "CES Frequently Asked Questions." bls.gov. Bureau of Labor Statistics, 7 Apr. 2017. Web.
2. "Current Employment Statistics Overview." bls.gov. Bureau of Labor Statistics, 3 Feb. 2017. Web.
3. Gershunskaya, J. and Dorfman, A.H. (2013) Calibration and Evaluation of Generalized Variance Functions. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 2655-2669

Appendix

Generally, consider *a model*

$$g\left(V_d\right) = f\left(X_{d1},...,X_{dp}\right) + \varepsilon_d \,,$$

for a set of domains $d = 1,...,D$, where

$g\left(V_d\right)$ are direct variance estimates or a function thereof, e.g., log,

$f\left(X_{d1},...,X_{dp}\right)$ is a function of a vector of auxiliary variables (e.g., sampling design characteristics, response rate, other),

ε_d are the model residual.

Consider stratified design with SRSWOR in each stratum.

The estimate of the ratio is

$$\hat{R}_{t-1,t} = \frac{\sum_{h=1}^{H}\sum_{j\in h} w_h y_{j,t}}{\sum_{h=1}^{H}\sum_{j\in h} w_h y_{j,t-1}}, \text{ where } w_h = \frac{N_h}{n_h}.$$

The variance is

$$Var\left(\hat{R}_{t-1,t}\right) \approx \frac{1}{Y_{t-1}^2}\sum_{h=1}^{H}\left(1-f_h\right)\frac{N_h^2}{n_h}S_{hu}^2 \,, \qquad (1)$$

where $f_h = \frac{n_h}{N_h}$ and $S_{hu}^2 = \frac{1}{N_h - 1}\sum_{j\in h}\left(u_{j,t} - \bar{u}_{h,t}\right)^2$,

$$u_{j,t} = y_{j,t} - R_{t-1,t}y_{j,t-1}, \; \bar{u}_{h,t} = \frac{1}{N_h}\sum_{j=1}^{N_h}u_{j,t}, \; R_{t-1,t} = \frac{\sum_{j=1}^{N}y_{j,t}}{\sum_{j=1}^{N}y_{j,t-1}}$$

Assume the superpopulation model:

$$y_{j,t} = \beta_t y_{j,t-1} + \varepsilon_{j,t}, \qquad (2)$$

where, for $j \in h$: $\varepsilon_{j,t} \overset{iid}{\sim} N\left(0, \sigma^2 y_{j,t-1}^{\alpha}\right)$

for some fixed but unknown σ^2 and $\alpha \geq 1$.

Consider formula (1) under model (2):

$$E\left(S_{hu}^2\right) = \sigma^2\frac{1}{N_h}\sum_{j\in h}y_{j,t-1}^{\alpha}$$

Thus, under the model, we have

$$Var\left(\hat{R}_{t-1,t}\right) \approx \frac{1}{Y_{t-1}^2}\sigma^2\sum_{h=1}^{H}\left(1-f_h\right)\frac{N_h}{n_h}\sum_{j\in h}y_{j,t-1}^{\alpha}$$

If we used simple random sampling (SRS), the variance would be

$$Var_{SRS}\left(\hat{R}_{t-1,t}\right) \approx \frac{1}{Y_{t-1}^{2}}\sigma^{2}\left(\frac{N}{n}-1\right)\sum_{j\in p} y_{j,t-1}^{\alpha}$$

Consider the following analogue (sort of) to design effect:

$$l = \frac{\log\left(\sum_{h=1}^{H}\left(1-f_h\right)\frac{N_h}{n_h}\sum_{j\in h} y_{j,t-1}^{\alpha}\right)}{\log\left(\frac{N}{n}-1\right)\sum_{j\in p} y_{j,t-1}^{\alpha}} \qquad (3)$$

It is expected that $l<1$. We assume it is approximately common to all cells included in the modeling.

For some $k<1$,

$$\sum_{j\in p} y_{j,t-1}^{\alpha} = Y_{t-1}^{\alpha k}$$

We hope that αk is approximately constant across States for a given industry.

From (3):

$$\sum_{h=1}^{H}\left(1-f_h\right)\frac{N_h}{n_h}\sum_{j\in h} y_{j,t-1}^{\alpha} = \left(\frac{N}{n}-1\right)^{l}\left(\sum_{j\in p} y_{j,t-1}^{\alpha}\right)^{l}$$

Then

$$Var\left(\hat{R}_{t-1,t}\right) \approx \sigma^{2}\left(\frac{N}{n}-1\right)^{l} Y_{0}^{l\alpha k-2}$$

Suppose

$$\left(\frac{N}{n}-1\right) \approx \left(\frac{Y_0}{Y_s}-1\right)\left(\frac{Y_s}{n}\right)^{k}$$

Y_s - averaged (over-the-year) unweighted sampled employment

$\frac{Y_s}{n}$ - average size of sampled establishments (UIs)

Thus, the model we consider is

$$\log\left\{Var\left(\hat{R}_d\right)\right\} = \beta_0 + \beta_1 \log Y_{d,0} + \beta_2 \log\frac{Y_{d,s}}{n_d} + \beta_3 \log\left(\frac{Y_{d,0}}{Y_{d,s}}-1\right) + \varepsilon_d$$

Results show that the regression coefficients l and $l\alpha k-2$ vary by industry. But their values are relatively stable over the years. In almost every industry, as expected $l<1$ and $l\alpha k-2<0$.