Calibration and Evaluation of Generalized Variance Functions October 2013

Julie Gershunskaya and Alan H. Dorfman[1]
U.S. Bureau of Labor Statistics,
2 Massachusetts Ave NE, Washington, DC 20212

Abstract:

The work of this paper is prompted by the particular case of the Current Employment Statistics (CES) Survey conducted monthly by the U.S. Bureau of Labor Statistics. Besides estimates at the national level, the survey yields estimates of employment for numerous domains defined by intersection of industry and geography, providing important information about the current status of the local economy. Variances of the employment estimates are estimated from the sample. However, the sample based estimated variances can be unstable, especially in smaller domains.

More stable variance estimates can be obtained using a model-based generalized variance function (GVF). The modeling is based on past years of the survey and, assuming a satisfactory model fit, the result can be applied to predict variances for the current period. However, some features of the design or population characteristics may change from one year to another, making it necessary to adjust the model parameters. We here give a method for evaluating the suitability to current data of a GVF model based on past years' data and suggest ways to calibrate the GVF to the current data.

Key Words: Balanced half samples, Current Employment Statistics Survey, replicates

1. Introduction

1.1 Reasons for indirect variance estimation

Sample based estimates of variances are usually unbiased or nearly so. However, there are reasons for avoiding estimating variances contemporaneously from the same data as is used for the point estimates:

- such estimation may take considerable time, which makes it infeasible in a tight production timeline;
- it may be desirable to have the measure of variability available and published ahead of the actual estimation;
- even when the variance estimates can be easily produced in real time, variation in these estimates can be worrisome. It is often due to random noise and does not have good substantive explanation.

Instability of the estimates is often related to the form of the distribution governing the data. Long tailed distributions are particularly prone to occasional extreme observations that can have undue effect on survey estimates. The variances could potentially be used in detecting outliers in the estimates. However, if the sample data contain extreme observations (as often happens in the establishment surveys), sample based variance estimates tend to be inflated. This creates a masking effect and renders such measures useless for detecting outliers in survey estimates.

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

Conversely, if extreme observations happen to exist only in the non-sampled part of the population, the sample based estimate of variance understates the true variance.

The result is that estimates of variance obtained from samples drawn from such outlier-prone populations may be seldom "correct", even though they are unbiased when averaged over all possible samples. The implication is that, without some sort of smoothing, these variance estimates cannot be used to characterize the quality of the point estimates. For example, raw past year estimates cannot be applied to the same domains in the current year, even when the finite population characteristics and sample design remain unchanged.

1.2 Issues in assessing quality of the indirect variance estimates

Another kind of estimate of variance is the model-based generalized variance function (GVF). There is no underlying written-in-stone theory for developing a particular GVF. Generally, it is a modeling exercise, where a set of raw estimates of variances (or some function thereof) play the role of dependent variables. The independent variables are usually related to sampling design characteristics and may also include any available auxiliary information that is deemed appropriate. Domains considered for inclusion in the GVF modeling are grouped based on some perceived similarity; for example, domains included in the model may belong to the same industry; at times, determining the grouping itself may be a non-trivial task (Valliant 1992.)

It is sometimes difficult to assess what can be considered a good model fit: as previously noted, the raw estimates involved in the modeling are very unstable and the instability can hardly be explained by the model. As a result, the usual goodness-of-fit statistics may not be informative. For example, the R-square in many real-life situations can be relatively low. Some may make the claim that one should not be concerned with explaining random noise. On the other hand, the unexplained variation may indeed have some underlying meaning that was not explained with the model at hand (see related discussion in Cho *et al*. 2002)

Our proposed solution for assessing model fit comes from the confidence interval. Upon obtaining the confidence interval coverage properties of the resulting GVF, we could evaluate the result without using the traditional goodness-of-fit measures. The approach explored in this paper concerns the possibility of evaluating coverage in real time, without knowing the true population value, and using a particular pivotal quantity.

Another twist related to the same idea is an adjustment of the GVF using the aforementioned pivot. In surveys repeated over time, the modeling involved in developing GVF usually is based on variances obtained from past years. While it is logical to expect certain continuity of the variances from one year to another, such GVF potentially would fail to account for true changes in the underlying population variability. For example, it is conceivable that the underlying population variance of the employment data vary during rapid economic growth or decline, or during periods of economic stability. Thus, we also explore evaluation of the applicability of GVF obtained from past years to current data and possible adjustments to it.

The evaluation procedure is described in Section 2. Adjustments in the case of under- or over-coverage of the confidence intervals are considered in Section 3. Section 4 describes a simulation study based on repeated samples from the real population and includes the simulation results and discussion.

2. The evaluation procedure

Suppose we are given variance estimates for a set of G independent domains. These may be, for example, GVF based variances with parameters estimated from past years of the same survey or some sort of smoothed variance estimates, or indeed variances estimated using any available direct or indirect method. The first task in the evaluation of variances is to compare the coverage properties of the corresponding confidence intervals.

The plan is to form a pivotal quantity in each domain and evaluate its properties based on a set of domains. To form the pivot, we obtain replicate estimates from the sample similar to the way it is done in a replication based variance estimation procedure. The difference from the whole-scale replication exercise is that, since we use the assumption that the pivots in the G domains are independent, as little as a single replicate along with the original sample estimate will suffice here.

To clarify the idea, let us consider the following simple setup.

Let $y_1,\ldots,y_n$ be a sample of independent measurements with $\hat{y}=\frac{N}{n}\sum_{j=1}^{n} y_j$, the estimate of the population total.

Suppose n is an even number. The set can be randomly divided into halves. Denote by $\alpha 1$ and $\alpha 2$ the sets of units in half 1 and 2 for a given random subdivision α.

Let $y_{\alpha 1}=\frac{2N}{n}\sum_{j\in\alpha 1} y_j$ and $y_{\alpha 2}=\frac{2N}{n}\sum_{j\in\alpha 2} y_j$, so that $\hat{y}=\frac{1}{2}\left(y_{\alpha 1}+y_{\alpha 2}\right)$.

Next, for a chosen constant K (say, $K=0.5$; this is similar to Fay's factor in the balanced repeated replication procedure), let us adjust the weights of units in half $\alpha 1$ by $2-K$; adjust the weights of units in $\alpha 2$ by K.

Let $\hat{y}_\alpha=\frac{1}{2}\left(2-K\right)y_{\alpha 1}+Ky_{\alpha 2}$. The proposed approach is based on the readily seen fact that quantity $\left(\hat{y}_\alpha-\hat{y}\right)/\left(1-K\right)$ has mean 0 and variance $Var\left(\hat{y}\right)$.

Suppose now we have $g=1,\ldots,G$ independent domains and measurements $y_j^{(g)}$, $\hat{y}^{(g)}=\frac{N^{(g)}}{n^{(g)}}\sum_{j=1}^{n^{(g)}} y_j^{(g)}$ for domain g and $v^{(g)}=Var\left(\hat{y}^{(g)}\right)$ the true variance of $\hat{y}^{(g)}$.

Let $\hat{y}_\alpha^{(g)}$ be a replicate estimate for domain g. For example, this can be a replicate from the setup similar to balanced half sample replication (BHS). Again, let K denote the Fay's factor that is often used with the BHS method (Judkins 1990; Rao and Shao 1999.) We focus on this setup because this is the way variances are estimated in the Current Employment Statistics (CES) survey that motivated the research. Alternatively, the setup may be similar to the one used in the bootstrap scheme as described in Rao *et al.* (1992).

Consider a set of G independent observations

$$z^{(g)}=\frac{\hat{y}_\alpha^{(g)}-\hat{y}^{(g)}}{\left(1-K\right)}. \qquad (1)$$

For a large enough sample in a domain, we usually assume normality of $\hat{y}^{(g)}$ when constructing confidence intervals. Thus, the same normality assumption holds for $z^{(g)}$:

$$z^{(g)} \overset{ind}{\sim} N\left(0, v^{(g)}\right). \tag{2}$$

Next, suppose a set of proposed estimates $\chi\pi^{(g)}$ of variances $v^{(g)}$ is available from an earlier study. To evaluate $\chi\pi^{(g)}$ based on a group of domains $g = 1,..., G$, we compute the percentage of times interval $\left[-t_\alpha, t_\alpha\right]$ contains

$$z^{(g)} \Big/ \sqrt{\chi\pi^{(g)}}, \tag{3}$$

where t_α is a quantile of the normal distribution. The nominal coverage is $<\mathrm{I}\left(t_\alpha\right) - <\mathrm{I}\left(-t_\alpha\right) = 1 - 2\alpha$, where <I is the standard normal distribution function.

Remark: In the case of sampling without replacement, when forming $z^{(g)}$ we need to account for the fact that the variance under evaluation accounts for a non-negligible sampling fraction.

3. The adjustment procedure

If the coverage of the confidence intervals described in the previous section deviates from the nominal level, we may think of some sort of adjustment to the set of proposed variances $\chi\pi^{(g)}$. In this Section, we consider several alternatives.

3.1 A simple adjustment

Let us assume that $Var\left(z^{(g)}\right) = \psi\chi\pi^{(g)}$, where factor ψ is not domain-specific (Model 1).

An unbiased estimate of ψ can be found by solving equation

$$\frac{1}{G}\sum_{g=1}^{G}\frac{z^{(g)2}}{\psi\chi\pi^{(g)}} = 1, \tag{4}$$

which gives us

$$\hat{\psi} = \frac{1}{G}\sum_{g=1}^{G}\frac{z^{(g)2}}{\chi\pi^{(g)}}. \tag{5}$$

Note: The normality of $z^{(g)}$ is not required here.

3.2 The model when the direct estimates of variances are available

In this subsection, we assume that the direct estimates of variances, denoted $v^{(g)}$, are available. In addition, we assume the design variance $V^{(g)}$ of these estimates is known. As discussed in the introduction, such a favorable setup is not expected in real time. Nevertheless, we considered this ideal situation and the corresponding estimator in our simulation study of Section 4.

In situations where the current year's direct estimates are not available, we would view the procedure as follows. The development in this subsection can be considered as an updating step for an "old" set of functions $\chi\pi^{(g)}$ based on the most recent available set of direct variance estimates (usually, the year immediately preceding the current one). The use of historical $\chi\pi^{(g)}$'s (rather than modeling "from the scratch", i.e., from the updated set of the auxiliary variables) aims at ensuring continuity of the GVF. This step can be followed by the contemporaneous evaluation and adjustment based on the $z^{(g)}$'s, as described in earlier subsections.

Model 2:

$$\nu^{(g)} \mid \psi^{(g)} \overset{ind}{\sim} \left(\psi^{(g)}\chi\pi^{(g)}, V^{(g)}\right), \tag{6}$$

$$\psi^{(g)} \overset{ind}{\sim} \left(\psi, I^2\right). \tag{7}$$

In reality, the variance $V^{(g)}$ is not known; for this research, we approximated it by using simulations based on repeated sampling from past years.

The marginal expectation of $\nu^{(g)} / \chi\pi^{(g)}$ is

$$E\left(\frac{\nu^{(g)}}{\chi\pi^{(g)}}\right) = E\left[E\left(\frac{\nu^{(g)}}{\chi\pi^{(g)}} \mid \psi^{(g)}\right)\right] = E\left(1 \cdot \psi^{(g)}\right) = \psi. \tag{8}$$

Hence, $E\left(\frac{1}{G}\sum_{g=1}^{G}\frac{\nu^{(g)}}{\chi\pi^{(g)}}\right) = \psi$ and an unbiased estimate of ψ can be found as

$$\hat{\psi} = \frac{1}{G}\sum_{g=1}^{G}\frac{\nu^{(g)}}{\chi\pi^{(g)}}. \tag{9}$$

Our goal is to find a set of optimal weights $w^{(g)C}$ that minimize the mean squared error of the following composite estimator (superscript C stands for "Composite"):

$$\hat{v}^{(g)C} = \left(1 - w^{(g)C}\right)\hat{v}^{(g)} + w^{(g)C}\nu^{(g)}, \tag{10}$$

where component $\hat{v}^{(g)}$ is the estimate of variance based on the adjustment factor given by (9):

$$\hat{v}^{(g)} = \hat{\psi}\chi\pi^{(g)}. \tag{11}$$

The optimal weights are expressed in terms of the mean squared errors of the estimators involved in the composite form (10) (see Rao 2003, pp. 57-58). In our case, the weights are

$$w^{(g)C} = \frac{E\left(\hat{v}^{(g)} - v^{(g)}\right)^2}{E\left(\hat{v}^{(g)} - v^{(g)}\right)^2 + V^{(g)}}. \tag{12}$$

To find an estimate of the mean squared error of $\hat{v}^{(g)}$, we have

$$E\left(\hat{v}^{(g)} - v^{(g)}\right)^2 = \pi^{(g)2} E\left(\hat{\psi} - \psi^{(g)}\right)^2. \tag{13}$$

Note that

$$E\left(\hat{\psi} - \psi^{(g)}\right)^2 = E\left(\hat{\psi} - \psi\right)^2 + E\left(\psi^{(g)} - \psi\right)^2 - 2E\left[\left(\hat{\psi} - \psi\right)\left(\psi^{(g)} - \psi\right)\right]. \tag{14}$$

Let us consider each term of the above expression:

(a) $$E\left(\hat{\psi} - \psi\right)^2 = Var\left(\hat{\psi}\right) = \frac{1}{G^2}\sum_{g=1}^{G} \frac{Var\left(v^{(g)}\right)}{\pi^{(g)2}} = \frac{1}{G^2}\sum_{g=1}^{G} \frac{V^{(g)}}{\pi^{(g)2}}. \tag{15}$$

(b) $$E\left(\psi^{(g)} - \psi\right)^2 = Var\left(\psi^{(g)}\right) = T^2, \tag{16}$$

and thus T^2 can be estimated as

$$\hat{T}^2 = \frac{1}{G}\sum_{g=1}^{G}\left(\hat{\psi}^{(g)} - \hat{\psi}\right)^2, \tag{17}$$

where $\hat{\psi}^{(g)} = v^{(g)} / \pi^{(g)}$ and $\hat{\psi}$ is defined by (9).

(c) The covariance term is zero:

$$E\left[\left(\hat{\psi} - \psi\right)\left(\psi^{(g)} - \psi\right)\right] = E\left[\left(\hat{\psi} - \psi\right) E\left[\left(\psi^{(g)} - \psi\right) \mid \psi\right]\right] = 0 \tag{18}$$

Weights $\hat{w}^{(g)C}$ are obtained by using (15), (17), and (18) to estimate (14).

4. Simulation using repeated samples from real population

In this Section, we describe the simulation experiment that we carried out in order to assess the usefulness of the proposed approach. We focus on the monthly estimation of employment from the Current Employment Statistics (CES) survey of the U.S. Bureau of Labor Statistics (BLS).

For the simulation, we use the population of businesses as reflected in the Quarterly Census of Employment and Wages (QCEW) dataset. The QCEW dataset closely matches the target population of the CES survey, and also provides the sampling frame and benchmark values for the CES. The QCEW contains administrative data for all businesses covered by the Unemployment Insurance program. It is released quarterly, several months after the publication of the corresponding CES estimates (which are designed to give more timely information). Although there are some differences between the available historical QCEW employment data and the population that was actually targeted by CES, these differences are considered minor and disregarded for the purposes of this simulation study.

We start with a brief overview of the CES sampling design and estimation. Next, we describe the simulation setup and present results followed by discussion.

4.1 CES sampling design and variance estimation

A stratified sample of unemployment insurance (UI) accounts is selected from the QCEW based frame. Strata are defined by the 50 States and DC, industrial supersectors (high level industrial aggregations based on North American Industry Classification System, NAICS), and employment size classes of UI accounts. The size class is determined based on the over-the-year maximum of the monthly total employment for each UI account. Sample allocation is determined to minimize, for a given cost, the variance of the over-the-month employment change at the State level. Within strata, Metropolitan Statistical Areas (MSA) define an additional, "implicit" level of stratification; units are selected systematically to ensure that the MSA sample sizes are proportional to the number of population units in MSAs. The sample is selected annually using the first quarter QCEW-based frame and is updated with the sample of new businesses ("births") when the third quarter of QCEW becomes available.

Establishments under a UI account may belong to different industries. The sampling procedure is based on the dominant industrial supersector ascribed to the UI account, while estimation is done using the establishment-based industry definition.

In this paper, we consider variances at month m in domain g for the estimate of the relative over-the-month change in employment level. These variances can be sufficiently approximated by the variance of the ratio of two survey weighted sums:

$$\hat{R}^{(g)}_{m-1,m} = \frac{\hat{Y}^{(g)}_{m}}{\hat{Y}^{(g)}_{m-1}}, \qquad (19)$$

where $\hat{Y}^{(g)}_{m} = \mathbf{I}'_{jES^{(g)}} w^{j}\, y^{j,m}$ and $\hat{Y}^{(g)}_{m-1} = \mathbf{I}'_{jES^{(g)}} w^{j}\, y^{j,m-1}$; $y^{j,m}$ and $y^{j,m-1}$ are employment levels reported by a unit j at months m and $m-1$ and w_j is its sampling weight; $S^{(g)}_{m}$ is a subset of units in the domain that report positive employment in both months.

CES uses a replication-based Repeatedly Grouped Balanced Half Samples (RGBHS) method for variance estimation. The method is an extension of the Balanced Half Samples (BHS) methodology for the case where there are more than two sampled clusters per stratum (Rao and Shao 1996). In addition, instead of using a half of the sample for each replicate estimate, CES employs Fay's method (Judkins 1990), thus using the whole sample with perturbed weights (the perturbation factor being 0.5).

4.2 The simulation setup

From the sampling frame constructed based on the third quarter of 2009 QCEW data, we selected 1,000 samples using the same sampling design used in CES. From each sample, we obtained estimates $\hat{I}^{(g)}_{m-1,m}$ at the State supersector level for 12 months from October 2010 through September 2011.

For each of these estimates, we computed estimates of their variances, $v^{(g)}_{m}$. During the actual production of estimates, the variances are computed using replication. For this simulation exercise, however, we employed a Taylor linearization formula. This provides

results close to the replication outcome at a far cheaper computational cost, which is helpful for large-scale simulations.

Let $u_{j,m} = y_{j,m} - \hat{R}^{(g)}_{m-1,m}\, y_{j,m-1}$, (20)

The Taylor linearization based variance estimate of $\hat{R}^{(g)}_{m-1,m}$ is

$$\tilde{v}^{(g)}_{m} = \frac{1}{\left(\bar{Y}^{(g)}_{m-1}\right)^2} \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{l=1}^{n_h} \left(u_{l,m} - \bar{u}_{h,m}\right)^2, \quad (21)$$

where $h = 1,..., H$ are strata; n_h and N_h are respectively the numbers of sample and population UI accounts in stratum h; $u_{l,m} = \sum_{j \in l} u_{j,m} I_{j \in g}$ is the cluster l total for domain g; $I_{j \in g}$ is the indicator that establishment j belongs to domain g; $u_{h,m}$ is the stratum average of $u_{l,m}$: $\bar{u}_{h,m} = \frac{1}{n_h} \sum_{l=1}^{n_h} u_{l,m}$.

For this simulation, the set of domains is defined as a set of States inside a given supersector. Since we have 13 supersectors, there are 13 different sets of domains in a given month. For each set, consider the following linear model:

$$\log\left(\bar{v}^{(g)}\right) = \beta_0 + \beta_1 \log\left(T_0^{(g)}\right) + \beta_2 \log\left(\frac{r^{(g)}}{T_0^{(g)}}\right) + \varepsilon, \quad (22)$$

where

$$\bar{v}^{(g)} = \frac{1}{12} \sum_{m=1}^{12} \tilde{v}^{(g)}_{m}, \quad (23)$$

$r^{(g)}$ is the over-the-year average number of reporting UI accounts in the domain; $T_0^{(g)}$ is the domain true population level at the benchmark month of a given year. (Alternatively, we could have taken the month specific variances in place of the over-the-year averages. The current version works sufficiently well and is also convenient for demonstration of how the proposed adjustment works when a particular month deviates from the average.)

We fit the above model using a robust linear regression function available in R software and obtain estimates of parameters for each repeated sample. The set of GVF functions is defined as

$$\chi\pi^{(g)} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \log\left(T_0^{(g)}\right) + \hat{\beta}_2 \log\left(\frac{r^{(g)}}{T_0^{(g)}}\right) + \frac{\hat{\sigma}^2}{2}\right), \quad (24)$$

where $\hat{\sigma}^2$ is the model MSE.

We select 1000 samples from the third quarter of the 2010 frame, which is the frame used in the following year. The GVF set for this year is defined as

$$\chi\pi^{(g)}_{*} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \log\left(T^{(g)}_{0^*}\right) + \hat{\beta}_2 \log\left(\frac{r^{(g)}}{T^{(g)}_{0^*}}\right) + \frac{\hat{\sigma}^2}{2}\right), \quad (25)$$

where subscript "*" in $r_*^{(g)}$ and $T_{0*}^{(g)}$ signifies that the information is specific to the new year; the model parameters, however, are estimated using the "old" year sample.

Next, we obtain direct variance estimates using this "new" year sample. Using repeated samples, we obtain the empirical variances of the estimates of variances and use them as $V^{(g)}$ of (6) in Model 2. Thus we obtain the set of composite estimates

$$\hat{v}_m^{(g)C} = \left(1 - w_m^{(g)}\right) \hat{\psi}_m \chi\pi_*^{(g)} + w_m^{(g)} v_m^{(g)}, \tag{26}$$

where $w_m^{(g)} = \frac{\square_m I_*^2 \chi\pi^{(g)2}}{V_m^{(g)} + I_m^2 \chi\pi^{(g)2}}$.

For each month m, we form $z_m^{(g)}$ variables based on the "new" year and compute adjustment for GVF $\chi\pi_*^{(g)}$ using $\hat{\psi}_m$ of (5), as described in Section 3.1. Using $z_m^{(g)}$, we also compute coverages, as described in Section 2, for each of the alternative variance estimators.

We evaluate the results for nine months of the "new" year. The results are summarized in the next subsection.

4.3 Simulation results

Below is the notation summary of the variance estimators considered in the simulation:

$\hat{I}_m^{(g)}$	**Estimator**
$v_m^{(g)}$	Direct
$v_{m-12}^{(g)}$	Empirical, m-12
$v_{m-12}^{(g)}$	Direct, m-12
$\bar{v}^{(g)}$	Direct, last year average
$\chi\pi_*^{(g)}$	Unadjusted GVF
$\hat{\psi}_m \chi\pi_*^{(g)}$	Adjusted GVF
$\hat{v}_m^{(g)C}$	Composite

Tables 1-3 show properties of several estimators of variances, based on the simulation results for a single month in the "current" year (January of 2012.)

Let $y_{m,s}^{(g)}$ denote a point estimator, based on simulation run s, of the true population value $y_m^{(g)}$ for domain g at month m. The true coverage of Table 1 is calculated as $1000^{-1} \prime \mathbf{I}_{s=1}^{1000} G^{-1} \prime \mathbf{I}_{g=1}^{G} I\left(\left|\hat{y}^{(g)} - y^{(g)}\right| < 1.96\sqrt{\hat{E}_{m,s}^{(g)}}\right)$, where $\hat{E}_{m,s}^{(g)}$ is the variance

estimate based on simulation run s . Similarly, the z-estimated coverage is calculated as $1000^{-1}\prime\sum_{s=1}^{1000} G^{-1}\prime\sum_{g=1}^{G} I\left\{\left|z_{m,s}^{(g)}\right|<1.96\sqrt{\hat{E}_{m,s}^{(g)}}\right\}$.

In Table 1, we observe that although the past year empirical variances provide close to nominal average coverage, the corresponding direct variance estimates give low coverage. Undercoverage is also observed for the confidence intervals which are based on the averaged (across 12 months) direct variance estimates, as well as for the GVF before the adjustment. The confidence intervals based on the adjusted GVF and the composite estimator provide satisfactory average coverage.

Table 1. Coverage properties of confidence intervals, for 95% nominal.
("T" denotes true coverage over the repeated samples; "Z" denotes z-estimated coverage averaged over repeated samples.)

Industry		10	20	31	32	41	42	43	50	55	60	65	70	80
Direct	**T**	90	95	94	93	94	94	93	91	94	94	95	95	95
	Z	87	94	93	91	95	94	91	90	95	92	93	92	95
Empirical, m-12	**T**	94	96	96	94	95	95	95	95	94	95	95	95	95
	Z	89	95	94	90	95	94	92	94	95	94	93	93	94
Direct, m-12	**T**	83	93	92	89	89	93	89	83	89	91	92	94	92
	Z	79	92	91	86	89	92	86	83	89	90	89	91	91
Direct, last year average	**T**	84	90	90	87	85	86	89	84	85	85	89	94	91
	Z	80	89	87	84	85	83	86	83	85	82	87	91	90
Unadjusted GVF	**T**	88	91	91	88	86	87	91	89	87	85	89	94	91
	Z	85	90	88	87	86	84	89	87	88	83	87	92	91
Adjusted GVF	**T**	94	95	95	93	93	95	95	93	93	94	95	96	94
Composite	**T**	93	95	95	94	95	95	95	93	94	94	95	95	95

Table 2 shows the mean of respective variances relative to the mean of the direct variance estimates, computed as $G^{-1}\prime\sum_{g=1}^{G}\sqrt{E_m^{(g)}/v_m^{(g)}}$, where $E_m^{(g)}=1000^{-1}\prime\sum_{s=1}^{1000}\hat{E}_{m,s}^{(g)}$ is the empirical mean for variance estimator $\hat{I}_m^{(g)}$ and $v_m^{(g)}$ is the empirical mean of the direct variance estimator (which coincides with the empirical variance, since the direct variance estimator is unbiased.)

In most industries, the last year averaged direct variance estimates were somewhat lower than the current year variances. As expected, this property of the mean variance estimates translates into the lower mean of the unadjusted GVF. The lower mean may explain the cases of undercoverage shown in Table 1. The means of the adjusted GVF and the composite estimator are close to the current year's variance.

On the other hand, in almost all industries, true (empirical) variances for the same month of the past year are slightly higher than the current year variances. The same is true for the past year direct variance estimates. Thus, undercoverage of the past year direct

variances cannot be explained by the lower mean. Discussion of the reasons for undercoverage in this case is given in Section 4.4 below.

Table 2. Mean of respective estimates of variances relative to the mean of the direct estimates, averaged across States

Industry	10	20	31	32	41	42	43	50	55	60	65	70	80
Direct	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Empirical, m-12	1.1	1.1	1.1	1.0	1.2	1.1	1.1	1.2	1.1	1.1	1.1	1.1	1.0
Direct, m-12	1.0	1.1	1.1	0.9	1.2	1.1	1.1	1.0	1.1	1.1	1.1	1.1	1.0
Direct, last year average	0.8	0.9	0.9	0.7	0.7	0.8	0.9	0.9	0.7	0.8	0.8	1.0	1.0
Unadjusted GVF	0.8	0.8	0.8	0.7	0.6	0.8	0.8	0.6	0.7	0.7	0.8	1.0	0.9
Adjusted GVF	1.0	1.0	1.0	0.9	0.9	1.1	1.0	0.7	0.9	1.0	1.0	1.1	1.0
Composite	0.9	1.0	1.0	0.9	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0

Table 3 shows relative variability of the variance estimators, as $G^{-1}\prime\mathbf{I}_{g=1}^{G}\ sd\left(\hat{E}_m^{(g)}\right)\Big/ sd\left(v_m^{(g)}\right)\Big\}$, where $sd\left(\hat{E}_m^{(g)}\right)=\sqrt{\prime\mathbf{I}_{s=1}^{1000}\left(\hat{E}_{m,s}^{(g)}-E_m^{(g)}\right)^2\Big/999}$.

The GVF-based and composite estimators are substantially less variable than the direct estimator.

Table 3. Variability of respective estimates of variances relative to the variability of the direct estimates, averaged across States

Industry	10	20	31	32	41	42	43	50	55	60	65	70	80
Direct	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Empirical, m-12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Direct, m-12	1.3	2.5	2.4	1.3	1.8	1.4	4.4	3.7	2.3	3.5	2.0	4.1	1.7
Direct, last year average	0.6	1.9	0.8	0.6	0.4	0.7	1.0	1.3	0.8	1.1	0.6	2.4	1.9
Unadjusted GVF	0.4	0.1	0.1	0.2	0.0	0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.1
Adjusted GVF	0.8	0.6	0.6	0.6	0.2	0.6	0.6	0.4	0.4	0.4	0.7	1.0	0.5
Composite	0.9	0.6	0.7	0.8	0.7	0.8	0.7	0.8	0.7	0.7	0.8	0.8	0.6

As noted above, Tables 1-3 display the *average* coverage across all domains in each industry for a particular month. Table 4 displays the distribution of coverage across domains and months. It shows the number of cases where the observed coverage is below 90% (for 95% nominal). For the true (empirical) variances, there are only a few low-coverage cases. However, for the variances that are based exclusively on the past data, we observe a large percentage of such cases. This effect is observed even with the true (empirical) variance of the same month of the past year. Estimators using current year information, such as the direct variance estimator and the composite estimator have significantly fewer low coverage cases.

Table 4. Number of domains with CI coverage lower than 90% (95% nominal)

Industry	Direct	Empirical, m-12	Adjusted GVF	Composite	Empirical	Out of
10	148	76	112	86	3	396
20	7	39	35	15	1	396
31	29	50	74	30	0	432
32	54	84	70	46	0	432
41	20	36	63	22	0	459
42	8	42	40	13	0	459
43	32	65	71	32	1	459
50	118	86	86	84	3	459
55	11	61	89	21	0	459
60	17	43	47	21	0	459
65	21	60	52	20	0	459
70	21	38	20	14	0	459
80	8	34	39	10	0	450

An alternative visualization the distribution of coverage is shown in Figure 1. The plot shows the distribution of coverage across domains in industry 42 (Retail Trade.) A point on the plot corresponds to a States at a given month (out of the 9 months considered in the simulation.) The direct variance estimator coverage (black dots) is distributed around the 95% line and is mostly above the 90% reference line. Similarly, the composite estimator coverage (blue squares) is distributed around the 95 % line, with only 13 cases found below the 90% line. The adjusted GVF (red triangles), however, has 40 cases below the 90% coverage. Similarly, a high percentage of undercoverage, 42 cases below 90%, is observed when the true (empirical) variance from the same month of the past year (green diamonds) is applied to the current year.

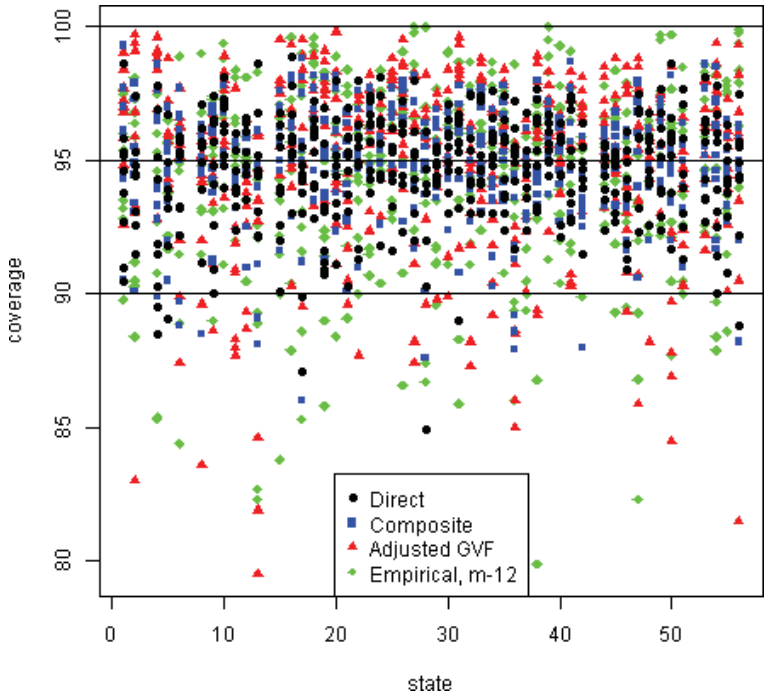

Figure 1. Confidence interval coverage in Industry 42 (Retail Trade)

4.4 Discussion

As noted in Section 4.3, confidence intervals that are based on the past year direct variance estimator provide low average coverage, even when they are slightly longer on average (Tables 1 and 2.) We conjecture that this effect is due to the properties of the employment data distribution. The distribution of the monthly employment change is a long-tailed distribution, prone to the appearance of extreme observations. The direct variance estimates depend on the realized sample. Since extreme observations occur randomly and generally may appear in one domain in the past but in a different domain in the current year (i.e., a different realized set of extreme observations across domains and years), the past year direct variance estimates are not suitable for the current year.

To demonstrate the phenomenon, we set up the following simple simulation. Observations are generated from a contaminated normal distribution

$$y_j^{(g)} \overset{iid}{\sim} 0.03N(0,1) + 0.97N(0,10^2),$$

for a set of $G = 50$ domains; $g = 1,...,G$; $j = 1,..., n$.

For each domain, we compute the mean, $\bar{y}^{(g)} = n^{-1} \prod_{j=1}^{n} y_j^{(g)}$, and the direct sample variance of the mean, $\hat{v}\left(\bar{y}^{(g)}\right) = n^{-1}(n-1)^{-1} \prod_{j=1}^{n}\left(y_j^{(g)} - \bar{y}^{(g)}\right)^2$. We then randomly re-assign the variances to the domains. We repeat this procedure 1,000 times and compute the percent of the confidence interval coverage for each version of the variances. The result is given in Table 5. The confidence intervals based on the direct and true variances

have approximately nominal coverage. However, the confidence intervals based on the reshuffled variances (standing for the "past year variance" situation) give low coverage.

Table 5. Average CI coverage in the case of the contaminated normal distribution

	Ave coverage (95% nominal)		
n	**Direct**	**True**	**Reshuffled**
30	96	94	83
50	96	94	85
100	96	95	89

Next, we attempt to explain the case of the observed undercoverage in a large number of domains as exhibited in Table 4. Our simulations of Section 4 use a fixed population for each year. This means that the number of extreme observations falling in each domain is also fixed for a particular year. However, in subsequent years the extreme observations may be randomly redistributed across the domains. It is generally not possible to predict the pattern of "reassignment" of extreme observations to domains in a new year. If the estimates of variances are based solely on the past data, domains where the percentage of extreme observations is higher in the current year are at risk of having a low coverage. Note that even the true (empirical) variance from the past year has this same property.

To illustrate this, we use the same simulation set up as describe above, in conjunction with Table 5. The GVF in this case is simply the average of the direct variance estimates, $\chi\pi = G^{-1}\prime\prod_{g=1}^{G} v\hat{}^{(g)}$. Although we repeat the simulations 1,000 times, we keep the number of extreme observations fixed in each of the G domains. The result is presented in Table 6. For the direct variance estimator, there are no domains where the coverage is below 90%. However, the true variance and the GVF, have a substantial number of domains having the low coverage. On average, on the other hand, all the estimators provide the nominal coverage.

Table 6. Average CI coverage and percentage of domains with low coverage, in the case of the contaminated normal distribution

	Ave coverage (95% nominal)			Percent of groups with coverage < 90%		
N	**Direct**	**True**	**GVF**	**Direct**	**True**	**GVF**
30	96	94	94	0	21	24
50	96	94	94	0	18	18
100	96	95	95	0	17	17

Summary

We proposed a method of evaluation and adjustment of a previously designed GVF using the pivotal quantity obtained from the current sample data. The GVF based variances tend to be more stable than the direct variance estimates, and they also have the advantage of being available before the actual estimation. However, the GVF estimates may be biased. The bias can be evaluated using the proposed pivotal quantity method. Under certain assumptions, it is also possible to adjust the variance estimates to reduce the bias.

If the data distribution is prone to extreme observations, direct variance estimates are correlated with the point estimates. In such a case, even if the true variance is

the same in the past and current years, the confidence intervals based on the past year direct variance estimates are not applicable for the current year, as they would result in low coverage.

If extreme observations appear randomly across domains, true variances from the past year work on average, over all domains; however, simulations based on repeated samples from the *fixed* finite population for a given year would produce a number of domains with the low confidence interval coverage. The same phenomenon is also observed with the GVF estimates. This effect creates difficulties evaluating the estimates: the simulation results neither produce a definitive reassurance of the quality of the variance estimator nor indicate a problem.

The estimates based on the composite estimator represent a compromise between the unbiased direct and stable GVF-based estimates. The downside of the composite estimator is that it is not available before the actual estimation process. It also assumes knowledge of the variance of the direct variance estimates.

Consider the case where direct variance estimates are available for the year of interest. As noted before, in the long-tailed distribution these variance estimates are correlated with the point estimates. A manifestation of this is wider confidence intervals when extreme observations are present in the sample. On one hand, the wider intervals have a greater chance of covering the true value. On the other hand, they mask outliers in the point estimates; in this sense, the direct variance estimates hardly provide a satisfactory measure of quality of the point estimates, even though they are unbiased.

References

Bureau of Labor Statistics (2011). Employment, hours, and earnings from the establishment survey. Chapter 2 of BLS Handbook of Methods, U.S. Department of Labor, http://www.bls.gov/opub/hom/pdf/homch2.pdf

Cho, M. J., Eltinge, J. L., Gershunskaya, J. and Huff, L. (2002). Evaluation of the Predictive Precision of Generalized Variance Functions in the Analysis of Complex Survey Data. In JSM Proceedings, the Section on Survey ResearchMethods. American Statistical Association, 534-539.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics* **6**(3), 223–239.

Rao, J.N.K. (2003). *Small Area Estimation*. JohnWiley & Sons, Hoboken, New Jersey.

Rao, J.N.K., Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of American Statistical Association* **91**, 343–348.

Rao, J.N.K., Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika* **86**(2), 403–415.

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18**, 209–217.

Valliant, R. (1992). Smoothing Variance Estimates for Price Indexes Over Time. *Journal of Official Statistics* **8**(4), 433–444