Calculating Generalized Variance Functions with a Single-Series Model in the Current Population Survey November 2016

Justin J. McIllece[1]
[1]Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212

Abstract

In the Current Population Survey (CPS), replication methods are used to calculate variances of survey estimates. Since these are often noisy, generalized variance functions (GVFs) are used to produce published estimates of variance that are more stable over time. Recently, the calculation of GVF model parameters has been reconfigured in the CPS. Rather than cluster series and create interdependencies among variance estimates, the parameters for each series are calculated individually, based only on their own histories. Instead of an iterative refitting, a single model is constructed for each historical series, smoothing out the noisiness of replicate variances while retaining seasonality. This paper details these changes to the GVF model framework and presents the resultant improvements in CPS published variance estimates.

Key Words: CPS, variance estimation, generalized variance function, replication, GVF

1. Introduction

In complex, expansive surveys, the problem of producing accurate and efficient variance estimates associated with survey statistics cannot always be addressed by direct computation. Design-based variance estimators may not be available or easily obtained for complex survey statistics[2]. For large surveys, particularly those that make public-use data files available, anticipating all possible combinations of interest to data users is impossible. The reasons for this difficulty in the calculation and presentation of variance estimates are well-known and discussed in salient literature; see Wolter (2007). Replication methods address the issue of the estimation of variance of complex survey statistics, but these techniques are computationally intensive and tend to be noisy. Generalized variance functions (GVFs), which fit models relating estimates of variance to the expected values of their associated survey statistics, address the issue of expansive surveys; i.e., GVF models can be easily and efficiently applied to a large volume of potential survey statistics. Additionally, GVFs tend to smooth out the noisiness that may be present in the underlying data. The intersection of replication methods and GVFs, then, offers an attractive framework for handling the problem of variance estimation in complex, expansive surveys:

[1] Views expressed are those of the author and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

[2] In the context of this paper, the ambiguous term "complex survey statistics" is inclusive of both the situation of the *complex form of an estimator*, such as ratios or composite estimators, and the situation of a simple estimator, such as Horvitz-Thompson, computed from a *complex sample design*.

1. Compute replicate variance estimates for a (preferably large) primary set of survey statistics.
2. Fit GVF models and publish model parameters relating the variance estimates to the survey statistics.
3. Approximate variances for secondary statistics by selecting model parameters from the most similar primary statistics.

Indeed, the Current Population Survey (CPS) has utilized this approach since 1947. Its motivation and additional details are discussed in *Technical Paper 66* (2006).

The complex CPS sample design incorporates stratification by state, clustering of counties into self-representing and non-self-representing primary sampling units (PSUs), and systematic sampling of households within selected PSUs. Selection probabilities, and therefore basic sampling weights, differ by state, and to increase precision for certain estimates of change a monthly 4-8-4 rotation scheme is applied: sampled households are interviewed for four months, excluded for the next eight months, and interviewed again in the subsequent four months. Extensive household data is collected from survey respondents, including but not limited to information about employment, income, occupation, education, age, race, ethnicity, veteran status, disability, and nativity. Thousands of survey statistics can be calculated from the data, and public-use data files are made available.

In conjunction with the CPS sample design, the method of successive difference replication is used to construct the replicate weight files necessary for computation of variance estimates (Fay and Train, 1995). The creation of micro-level replicate weights is itself a complicated process; for the purpose of this paper, it shall be assumed that the national-level (as opposed to state-level) variance estimates computed via the replicate weight files are generally unbiased.

2. GVF Models

The generalization method utilized by the CPS is relevant to series that can be characterized as binomially distributed, such as counts of employed and unemployed from the civilian noninstitutional population (CNP), as well as associated rates. Valliant (1987) gives conditions for a class of models that supports non-binomial series, as well. Only the binomial case is considered in this paper, as it is most relevant to the Current Population Survey.

Given a survey statistic $\hat{X}$, such as total employed or unemployed, a GVF of the following form is used to estimate its variance:

$$V(\hat{X}) = a\hat{X}^2 + b\hat{X} \quad (1)$$

The subject of the methodological change affecting the estimation of CPS variances, the calculation of model parameters a and b is the focus of the paper and will be subsequently discussed in Sections 2.1 and 2.2. The final model (1) is unchanged, reflecting both the survey complexity and the underlying binomial properties of the objective series:

$$V(\hat{X}) = \frac{N^2 p(1-p)\delta}{n}$$

where

$N =$ population (or subpopulation) total

$n =$ sample size

$p = \hat{X}/N =$ estimated proportion of N with the characteristic measured in X

$\delta =$ design effect, defined as ratio of replicate variance to simple random sample variance

Expanding the numerator terms, this can be rewritten as

$$V(\hat{X}) = \frac{N\delta}{n}(\hat{X}) - \frac{\delta}{n}(\hat{X}^2)$$

with GVF model parameters

$b = N\delta/n$

$a = -\delta/n = -b/N.$

to yield the GVF form given in (1).

When N is a control total (i.e. constant), provided by the Census Bureau and utilized in benchmarking, the estimated variance of $\hat{X}$ will be zero when $\hat{X} = N$ or $\hat{X} = 0$, conforming to the properties of the binomial distribution.

2.1 Historical Parameter Estimation

Until 2015, the method used to estimate GVF parameters had three primary characteristics:

1. Relative replicate variance as the dependent variable
2. Iterative, weighted least squares regression as the modeling technique
3. Grouping, or clustering, across series to borrow strength cross-sectionally

The a and b parameters from (1) were estimated from this weighted least squares regression model relating relative variance to the survey statistic for all series within a given cluster:

$$\frac{V(\hat{X})}{\hat{X}^2} = a + b\hat{X}^{-1}$$

The inverse of the expected variance was used as the vector of series weights for this regression model. Since the expected variance resulting from the model changed after the model was fit, the process was iterated, with the weights being recalculated each time.

Wolter (2007) shows examples of how this form can be effective with respect to the relative variance, which is presumed to be stable over time. The accuracy of fit utilizing this GVF model tends to rely upon the adequacy of the clustering relative to the similarity of the series' variance properties, such as design effects, which can be formulaically defined to

encapsulate weighting adjustment arising from nonresponse and multi-dimensional benchmarking to external population controls (a.k.a. calibration). Stated simply, the GVF for a series should be effective if all the series in the cluster have similar design effects.

In practice, this method results in series grouped with others that do not meet this criterion of design effect similarity. Figures 1 and 2 display standard error estimates derived from the historical GVF model for selected Asian and Hispanic labor force characteristics (which were each clustered with many non-Asian and non-Hispanic series).

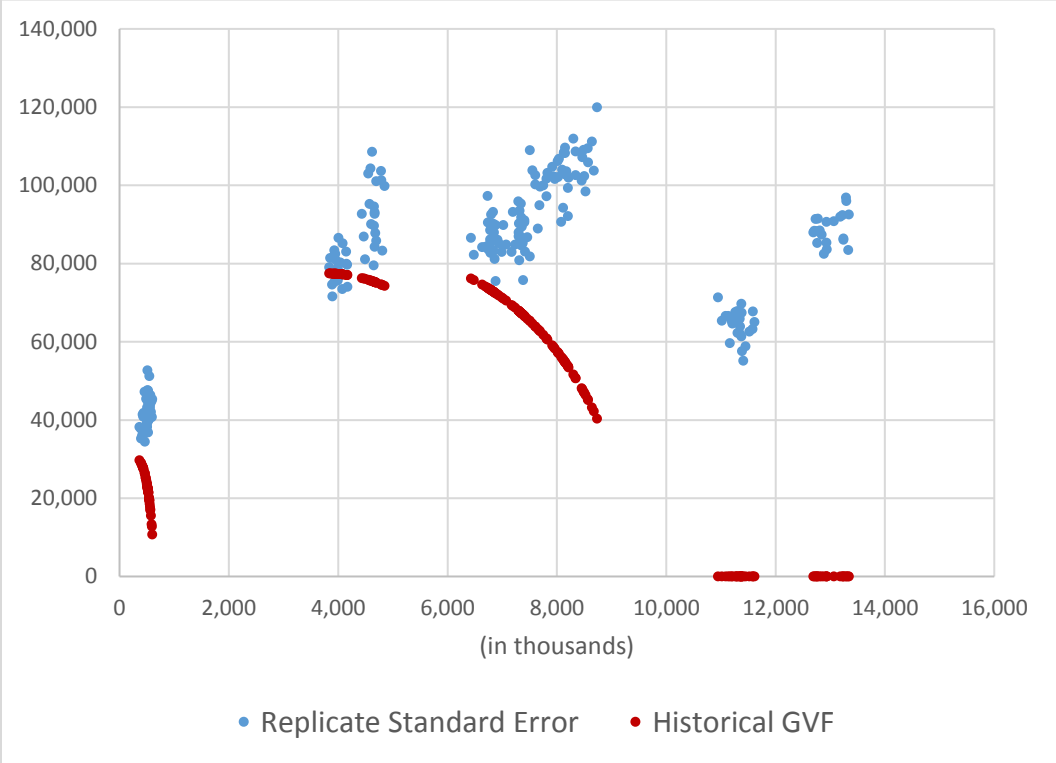

Figure 1: Replicate and GVF standard errors (vertical axis) plotted against associated survey estimates of employed, unemployed, and not in labor force for *Asian, 20 and older* and *Asian, 16 and older*.

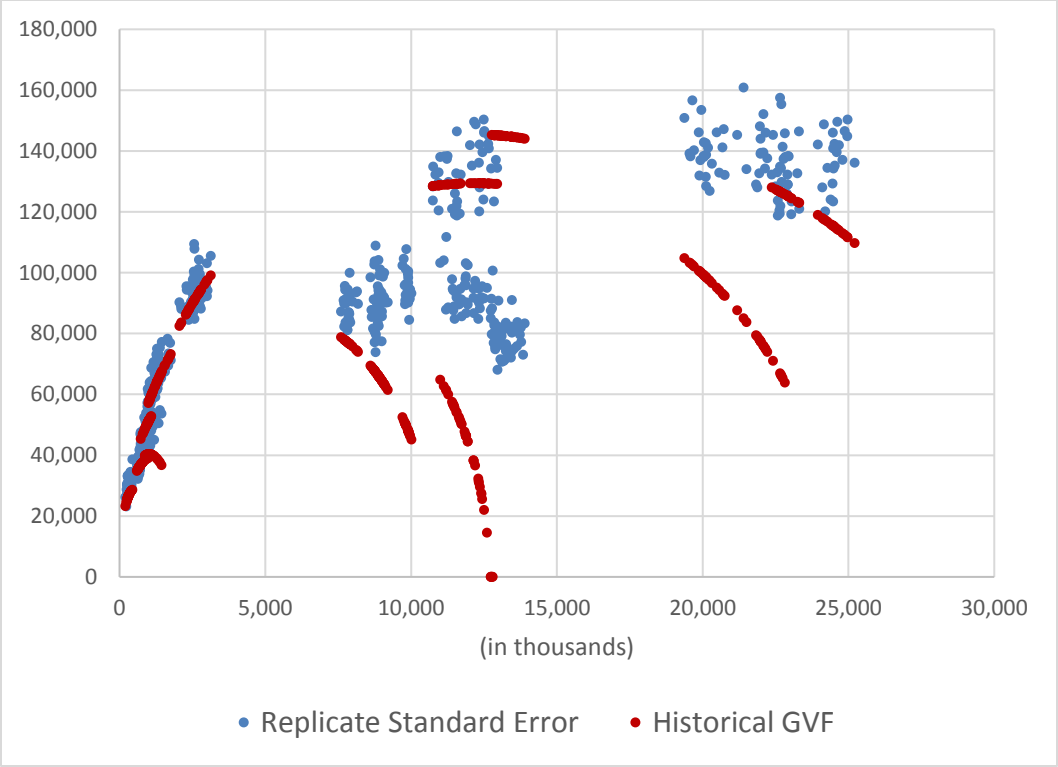

Figure 2: Replicate and GVF standard errors (vertical axis) plotted against associated survey estimates of employed, unemployed, and not in labor force for *Hispanic, 20 and older* and *Hispanic, 16 and older*.

The poor fits evident for some of the displayed series are emblematic of problems with the historical GVFs rather than exceptional. While many perform quite well, an unacceptable number veer off considerably from the cloud of replicates they are intended to represent, and in some cases are parameterized so poorly that they produce negative estimates of variance. In Figures 1 and 2, for convenience of display, these are indicated by "zero" points along the y-intercept. The modeling problems Figures 1 and 2 reveal require little explication; negative estimates of variance and irrational standard errors resulting from them are untenable for the CPS, particularly when series of considerable interest, such as Asian or Hispanic labor force, are involved.

Multiple factors can contribute to poor GVF fits. As mentioned in Section 1, dissimilarity of design effects between series in the same cluster can lead to biased estimates of variance. Additionally, these models treated the data statically, pooling together at least one year of monthly observations. Standard errors of totals, however, can change substantially over time, as the size of the CNP and relevant subpopulations grow.

Interventions can be made when problems arise, but a more robust solution to the problem of poor GVF fitting was desired.

2.2 Replicate Variance Components

The current GVF methodology employed by the CPS was implemented in 2015. The framework of model (1) is still utilized, but the estimation of parameters a and b now has three primary characteristics differentiating it from the historical method:

1. The b parameter, $N\delta/n$, as the dependent variable, which can be viewed as the design effect times the sampling interval (a.k.a. base weight)
2. Ordinary least squares regression as the modeling technique
3. "Grouping" within series over time to borrow strength longitudinally

The motivation for this modeling strategy is primarily drawn from the long-term stability but short-term volatility of the *b* parameter. (Stability, in this context, refers to a series that is well-behaved, or predictable, but not necessarily static. Trends are not precluded.) Empirically, it has been observed that national CPS design effects do not drastically or obviously change over extended periods of time, nor during periods of economic upheaval, such as the Great Recession of the late 2000s. Additionally, the sampling intervals have been quite stable over time. Any substantial changes to the national sampling interval, such as those resulting from significant sample maintenance reductions, are known in advance and can be accounted for in modeling, but that possibility will not be extrapolated upon in the remainder of this paper.

On a monthly basis, the design effects are quite noisy, since they are calculated as ratios of replicate variance (which tend to be noisy, as indicated in Section 1) to theoretical simple random sample variance. The simple random sample, in this case binomial, variance component is much more stable; modeling the design effect component, which is included in the *b* parameter, appealingly mitigates the short-term volatility while leveraging predictable long-term behavior. Figure 3 overlays these two components to visually express their relative stability for variance estimates of total persons in the civilian labor force.

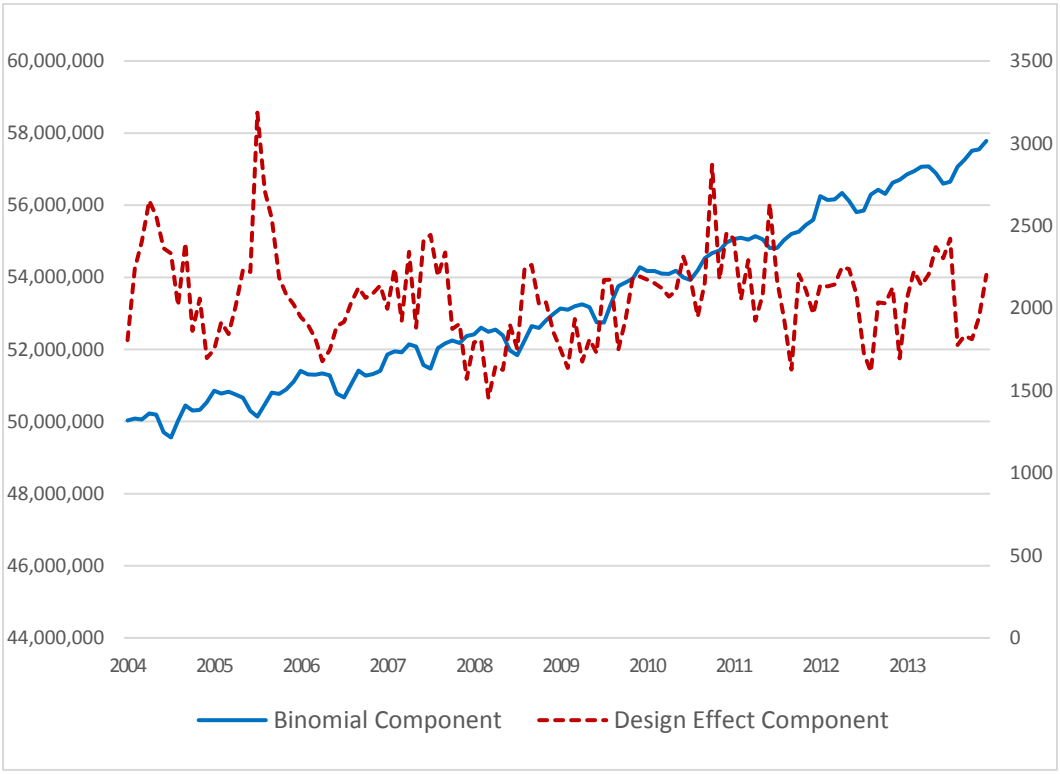

Figure 3: Decomposition of civilian labor force replicate variance into "binomial" [Np(1-p)] and "design effect" [Nδ/n] components. The design effect component is on the secondary (right-hand) scale.

In Figure 3, it is observable that the binomial component is stable, while the design effect component is volatile short-term and fairly flat long-term (indicating little change of the "true" underlying design effect). Considering these are multiplicative factors, the volatility of the design effect component impels the volatility of the replicate variance estimate. Smoothing out the design effect component's short-term volatility is therefore the modeling objective for the GVF.

2.3 Single-Series Parameter Estimation

It was observed in Section 2.1 that combining series with dissimilar design effects can lead to poor estimates of variance. It was further observed in the decomposition in Figure 3 how the design effect within a series does not change much long term. A single-series model that avoids clustering and interdependencies altogether, leveraging the long-term stability of a series' variance components, offers some advantages over the historical method. However, Figure 3 also emphasizes what is theoretically known about the binomial variance component: it changes as the population changes over time. Thus, a single-series model, necessarily dependent upon a substantial longitudinal history, must account for population dynamics. Omission of the changing population results in GVF parameters only accurate for the center of the time history used in the model fit.

All these concepts informed the construction of a new GVF model for the estimation of *a* and *b* parameters. This model assumes that historical growth rates for relevant subpopulation groups in the U.S. accurately reflect future growth rates. The resulting GVF parameters tend to produce variance estimates robust against mild to moderate projection errors. Large projection errors relative to the size of the subpopulation lead to more substantial bias in the GVF variance estimates. Most series in the primary CPS tables do not have large subpopulation projection errors when using this method.

Given up to a 10-year series history of monthly (*t*) variance estimates, an ordinary least squares regression model that accounts for population growth is constructed.

Let:

$$\phi_t = N_t \delta_t / n_t$$

$$\underline{\phi}_t = \frac{\left(\phi_t - \overline{\phi}\right)}{\overline{\phi}}$$

$$\underline{N}_t = \frac{\left(N_t - \overline{N}\right)}{\overline{N}}$$

where *T* is the number of months in the series history (up to 120), and

$$\overline{\phi} = \frac{1}{T}\sum_{t=1}^{T} \phi_t$$

$$\overline{N} = \frac{1}{T}\sum_{t=1}^{T} N_t$$

Then construct the following zero-intercept regression model based on the sample data:

$$\underline{\hat{\phi}}_t = \hat{\beta}\underline{N}_t = \hat{r}_{\underline{\phi},\underline{N}}\left(\frac{\hat{\sigma}_{\underline{\phi}}}{\hat{\sigma}_{\underline{N}}}\right)\underline{N}_t$$

where

$\hat{r}_{\underline{\phi},\underline{N}}$ = series correlation between $\underline{\phi}_t$ and $\underline{N}_t$

$\left(\frac{\hat{\sigma}_{\underline{\phi}}}{\hat{\sigma}_{\underline{N}}}\right)$ = ratio of standard deviations of predictor and dependent variables

This model can then be expanded to yield an estimated value for the *b* parameter from model (1):

$$b = \hat{\phi}_t = \overline{\phi} + \overline{\phi}\left[\hat{r}_{\underline{\phi},\underline{N}}\left(\frac{\hat{\sigma}_{\underline{\phi}}}{\hat{\sigma}_{\underline{N}}}\right)\right]\frac{\left(N_t - \overline{N}\right)}{\overline{N}}$$

If the correlation between the design effect component and subpopulation size is zero, then the *b* parameter is equal to $\overline{\phi}$ for all months. Otherwise, the *b* parameter changes as the population changes. For projecting future parameters, the subpopulation projection is used in place of the actual population value, as discussed above.

Since this model still conforms to (1), the *a* parameter is defined in relation to *b*:

$$a = -b/N$$

The model in this section relates to variances of levels, such as total employed or unemployed. The parameters for rates are developed analogously, assuming that the additional denominator (base) term of the binomial variance is a constant. In practice, this is only true when the base is a fixed population control. Many CPS rate statistics have random variables in the denominator, but empirical review suggests that treating them as constants has negligible impact for most variance estimates.

3. Results

CPS tables A1 through A16, available at www.bls.gov/cps, contain over 600 series of primary importance in the development of an alternative GVF methodology. Figures 4 - 8 show selected series with varying properties.

Thorough review of the model fits for these series has shown accuracy across the historical data used for model construction and in projecting future parameters. Further, this GVF model reflects the natural seasonality of unadjusted series, for which they are developed. Evans, McIllece, and Miller (2015) showed that the impact of seasonal adjustment on variance is not negligible for some primary labor force series, but currently all GVF

parameters are constructed from the unadjusted series. The GVF model developed in this paper naturally extends to an analogous application using seasonally-adjusted data, insofar as replicate variance estimates for the seasonally-adjusted series are available.

In Figure 4, the replicate standard errors of civilian labor force (16 years and over) estimates show a great deal of monthly and yearly volatility. The GVF parameters, calculated over the 2004 to 2013 sample data and projected forward for 2014, smooth out this undesirable noise. Note that the spike in April 2014 is artificial, not economic, resulting from the beginning of the phase-in of the 2010 sample from the 2000 sample.

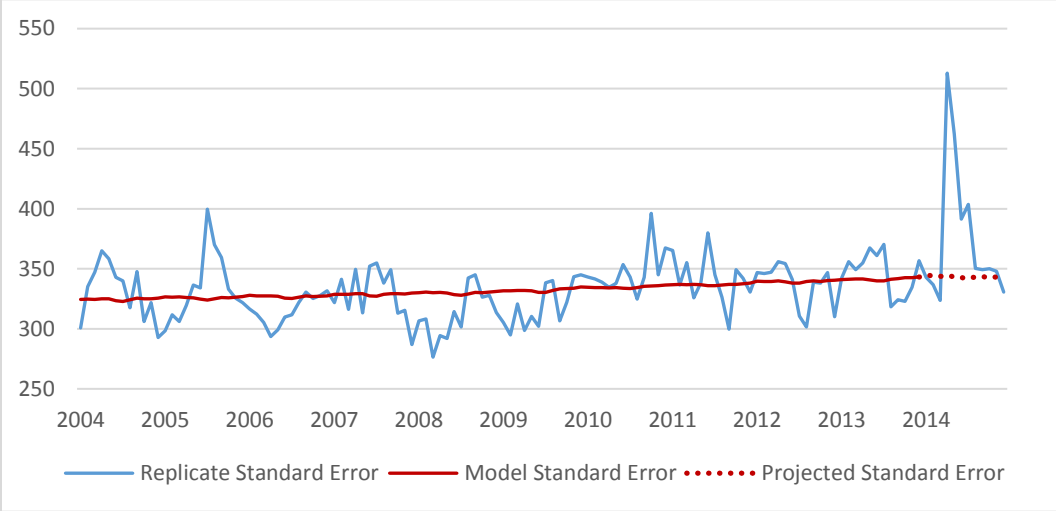

Figure 4: Replicate, model, and projected standard errors (in thousands) for *civilian labor force*. Not seasonally adjusted.

In Figure 5, the standard errors for Hispanic or Latino employment (16 years and over), a series that was among those poorly parameterized in the historical method because of ineffective clustering (see Figure 2), are accurately reflected by single-series parameter estimation.

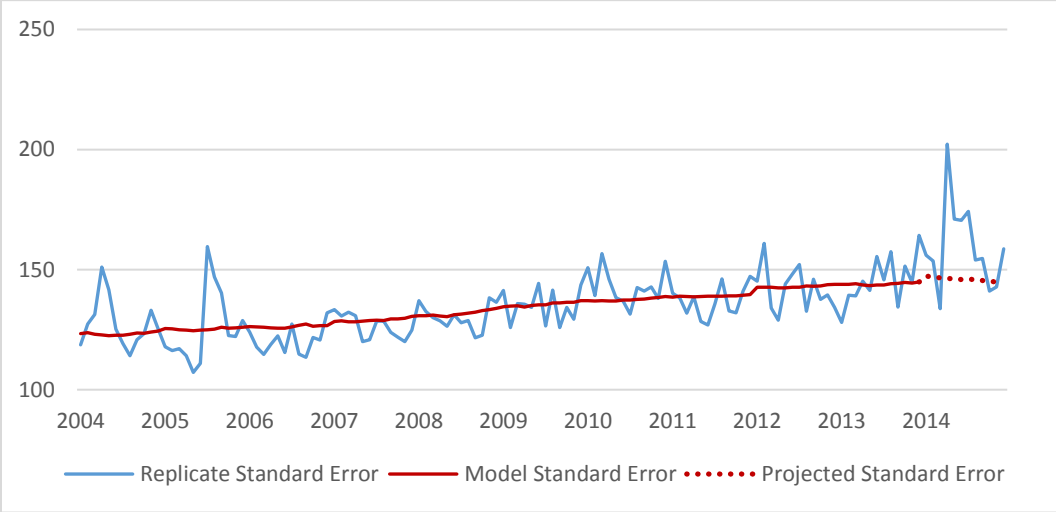

Figure 5: Replicate, model, and projected standard errors (in thousands) for *Hispanic or Latino employment*. Not seasonally adjusted.

Figure 6 displays the standard errors of the official unemployment rate. Seasonality is more easily observed; the model, fit on the non-seasonally adjusted series and adhering to the binomial properties underlying the data, reflects the seasonality. The sharp increase in standard errors resulting from the Great Recession (and gradual decrease since) are evident and modeled accurately. Periods of economic change could be problematic in the historical method, which relied on data from a relatively short period of time.

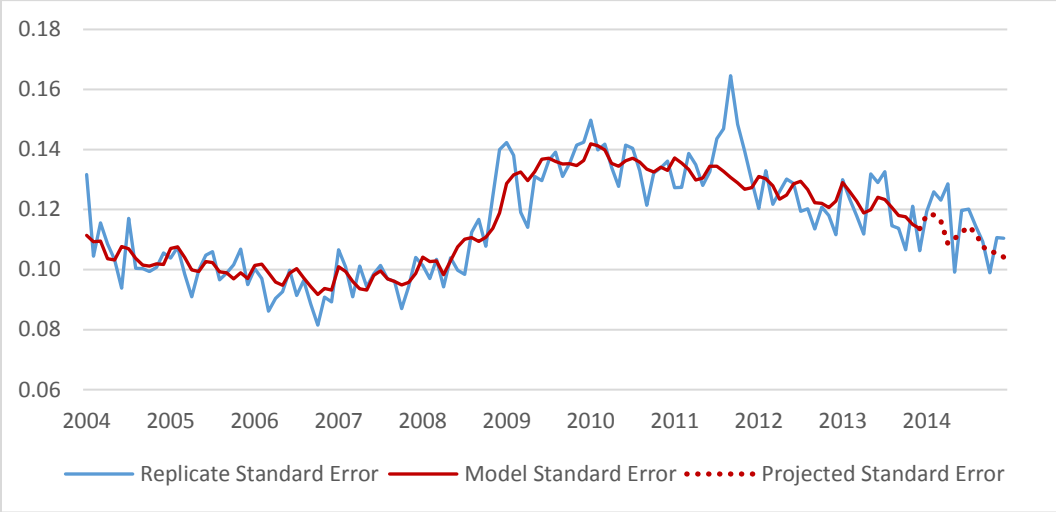

Figure 6: Replicate, model, and projected standard errors (in thousands) for *unemployment rate*. Not seasonally adjusted.

Figures 4 - 6 are for large series, but the GVF models must be effective for smaller series, as well. Figure 7 displays standard errors for the percent of persons in the civilian labor force unemployed at least 15 weeks. The model again accurately reflects the series, including tracking the surge during the Great Recession as more people were long-term unemployed. As in the prior figures, the projection forward fits the trend observed from the replicates.

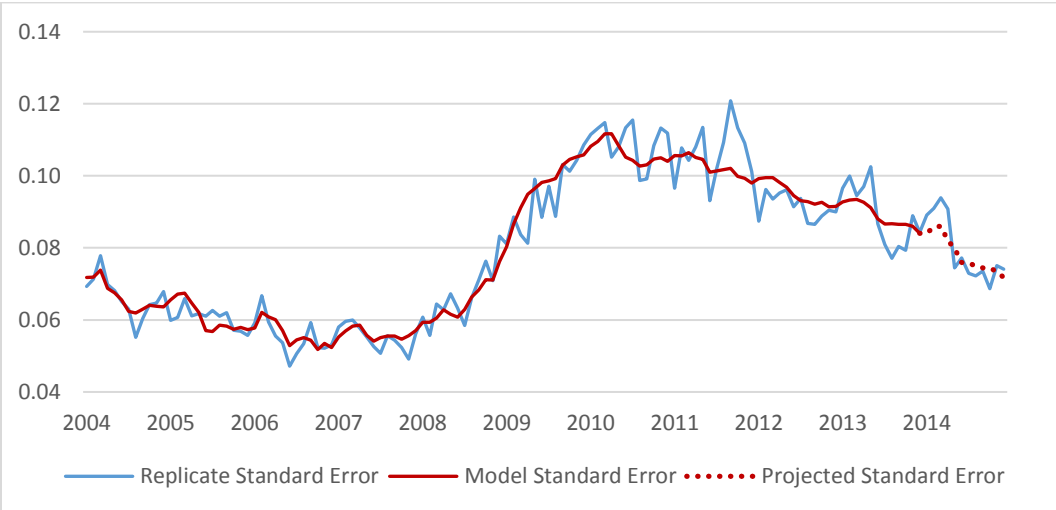

Figure 7: Replicate, model, and projected standard errors (in percent) for *unemployed 15 weeks and over, as a percent of civilian labor force*. Not seasonally adjusted.

In Figure 8, a relatively small series with highly seasonal standard errors is shown to be well represented by the single-series GVF model. A slight underestimation in the projection is the result of a subpopulation projection that was not as accurate as for the larger series. However, the model is fairly robust against this subpopulation projection error, as it is still a close representation of the replicate standard errors.

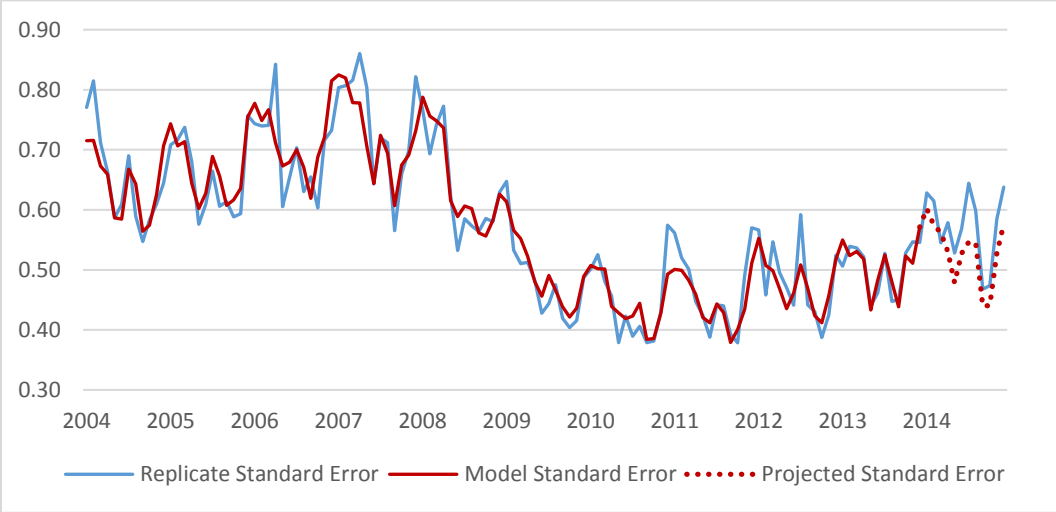

Figure 8: Replicate, model, and projected standard errors (in percent) for *job losers on layoff, as a percent of total unemployed*.

The model fits of these series are indicative of the general quality of fits for the 600+ series in tables A1-A16. Due to the overall accuracy of the model standard errors, the single-series parameter estimation was found to be an acceptable replacement for the historical method of clustering and iterative refitting. Continuing work includes parameter estimation for additional series and developmental research of GVFs to effectively model the standard errors for non-binomial series, such as means and medians.

References

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2nd ed.), New York, NY, Springer.

U.S. Census Bureau (2006). *Design and Methodology, Current Population Survey, Technical Paper 66* (2006). Washington, DC, by authors.

Fay, R.E. and Train, G.F. (1995). "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties", in *Proceedings of the Joint Statistical Meetings*, Government Statistics Section.

Valliant, R.L. (1987). "Generalized Variance Functions in Stratified Two-Stage Sampling", *Journal of the American Statistical Association* Vol. 82, No. 398, pp. 499-508.

Evans, T., McIllece, J., and Miller, S. (2015). "Variance Estimation by Replication for National CPS Seasonally Adjusted Series", in *Proceedings of the Joint Statistical Meetings*, Statistical Methods Section.