Expanding Variance Function Coverage in the Current Population Survey November 2019

Justin J. McIllece[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE #4985/02, Washington, DC 20212

Abstract
Generalized variance functions (GVFs) are used to produce official standard errors for the Current Population Survey (CPS) estimates included in *The Employment Situation*[2], a Principal Federal Economic Indicator (PFEI) published monthly by the Bureau of Labor Statistics. The parameters of the GVF models are also published and available to data users. However, variance function coverage beyond *The Employment Situation* is limited, and no GVF parameters are published for non-PFEI news releases. In this paper, recently developed models are extended in multiple dimensions, including: from time-dependent to time-robust parameters; from medians to other percentiles; from full-sample to partial-sample data; and from national to state-level estimates. Strengths and weaknesses of various research models are considered, specifically in relation to the quality of estimated standard errors and the convenience of the model form for parameter publication.

Key Words: CPS, GVF, variance functions

1. Introduction

The application of generalized variance functions in the Current Population Survey has undergone methodological transformations in recent years, including the development of a new underlying model for binomial monthly estimates (McIllece 2016) and novel GVF frameworks for estimating the variances of sample means and medians of weeks unemployed (McIllece 2018). The complexity of the multi-stage CPS sample design and motivation for employing GVFs are discussed in those papers. The impetus for prior efforts was primarily to improve the accuracy, stability, and usability of variance estimates for household tables in *The Employment Situation* monthly news release. As a PFEI, *The Employment Situation* is of central importance to the program, but other news releases and estimates are also produced, many of which have no variance function coverage.

This paper details the research and GVF modeling results of two primary objectives:

1. Improvement of generalizability across time
2. Development of GVF models for earnings percentiles

[1] Views expressed are those of the author and do not necessarily reflect the views or policies of the U.S. Bureau of Labor Statistics.
[2] https://www.bls.gov/news.release/empsit.toc.htm

The GVF model developed in McIllece (2016) is specific to binomial data, since the majority of CPS estimates are counts (levels) or rates. Recognition of the relationship between the population size N and the variance of a binomial estimate—particularly for levels—motivates the first objective.

The 2016 GVF model for a level estimate x is of the form[3]

$$v(x) = \frac{N^2}{n} p(1-p) * d \cong ax^2 + bx \qquad (1)$$

where $p = x/N$, b is a regression-smoothed parameter estimate of the national sampling interval times a complex design effect, $(N/n) * d$, and $a = -b/N$ is simply an algebraic derivation based on the formula for binomial variance. Clearly, the a, b parameterization implies that a static value of N, which will be referred to as N^*, undergirds model (1):

$$v(x) \cong ax^2 + bx = \left(-\frac{b}{N^*}\right) x^2 + bx = b\left(x - \frac{x^2}{N^*}\right) \qquad (2)$$

This was indeed the case: published a, b parameters relied on a projected annual average N^* as a static substitute of the variable monthly population total N. However, the binomial variance changes as N changes, suggesting the possibility of poor model fits from (2) when the implicit assumption $N \cong N^*$ is unsatisfied. Historical analyses are especially vulnerable to this violation.

An adjustment to model (2) is presented in Section 2 to mitigate this potential problem.

The CPS quarterly news release *Usual Weekly Earnings of Wage and Salary Workers*[4] produces earnings estimates at the 10th, 25th, 50th (median), 75th, and 90th percentile levels. The CPS annual news release *Union Membership*[5] also includes median earnings estimates.

Since earnings data in the CPS are only collected from a quarter of respondents—in the two "outgoing" of the eight rotation groups (U.S. Census Bureau 2006)—consideration of the effects of partial-sample data is necessitated for the second objective. To produce estimates of sufficient reliability and stability, weekly earnings estimates are published on either a quarterly or annual basis, significantly reducing the numbers of model observations relative to monthly series. Most of the currently published GVF parameters were fit to monthly data beginning in January 2003, giving 192 observations (2003 – 2018 for the most recent update at the time of this paper), which allow for highly accurate variance modeling results for the predominance of those series. Quarterly estimates have 48 observations (one-fourth), and annual estimates have a scant 16 observations (one-twelfth), increasing concern about the feasibility of modeling those standard errors.

Section 3 discusses the development of GVF models for weekly earnings percentiles based on these partial-sample data and a similar formulation to the McIllece (2018) model for median weeks unemployed.

[3] To simplify notation, "hats" are omitted, as every variable or parameter is a sample-based estimate subject to sampling error, except for the administrative population total N or any total derived from N, such as N^*.
[4] https://www.bls.gov/news.release/wkyeng.toc.htm
[5] https://www.bls.gov/news.release/union2.toc.htm

2. Generalizability Across Time

The GVF model (1) has been commonly used historically, especially in the CPS, and defines $a = -b/N$ as one of its two parameters (U.S. Census Bureau 2006). However, as stated in the Introduction, the fact that the variable N is implicitly built into the a parameter clearly indicates that some static substitute N^* is the actual approximation to N embedded into any such parameter intended for generalization beyond a single time point. Of course, releasing complete parameters for all time points would undermine the gain in publication efficiency, which stands as a primary motivator for the adoption of GVF models for variance estimation in the first place.

The problem of the $N^* \cong N$ assumption for historical standard error estimates of the Civilian Labor Force is illustrated in Figure 1.

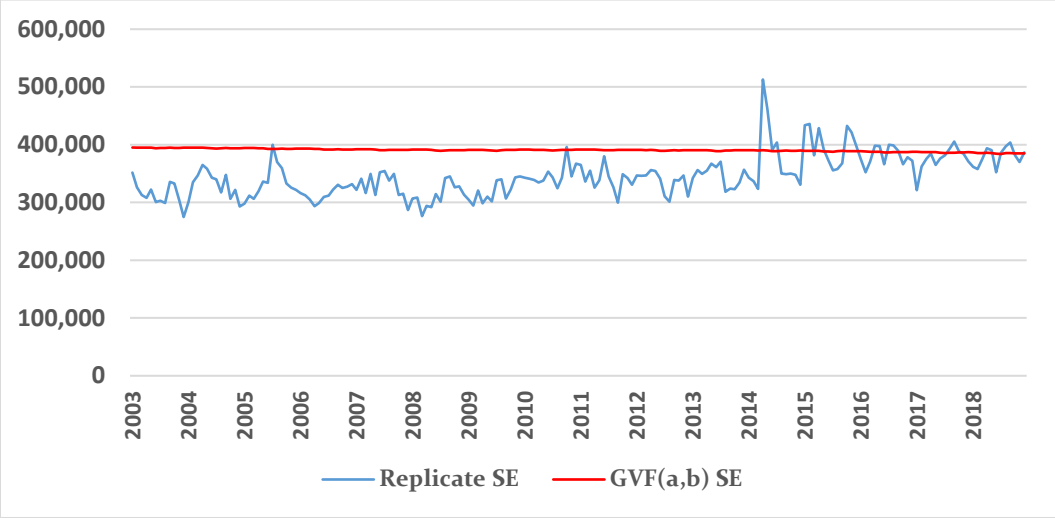

Figure 1: Standard error estimates of *Civilian labor force, 16 years and over*. The blue replicate standard errors are based on CPS successive difference replicate weights (U.S. Census Bureau 2006; Fay and Train 1995). GVF model (1) predicted standard errors are shown in the red line. The value of N^* was fixed to the value of N from December 2018 for this illustration.

If the GVF model for this series—one of the largest and most stable of all CPS estimates—were properly reflecting the replicates, the red line would better track the overall trend of the blue line. Instead, the predicted standard errors from (1) are almost uniformly just below 400,000 persons over a 16-year timeframe. In fact, the trend is slightly negative between January 2003 and December 2018.

Recall GVF model (2), which explicitly represents this practical approximation:

$$v(x) \cong b\left(x - \frac{x^2}{N^*}\right)$$

where b is estimated from an ordinary least squares (OLS) regression model against the population total of the form

$b = \alpha + \beta N$ (McIllece 2016).

Under this structure, the OLS α, β parameterization underlies the transformation into a, b parameters by defining them in relation to a population constant N^*. A seamless correction for the problem of applying a, b parameters to situations where $N^* \neq N$ is to retain the α, β parameterization and include N as an input variable, instead of substituting N^*, into the variance prediction model:

$$v(x) \cong (\alpha + \beta N)\left(x - \frac{x^2}{N}\right) \qquad (3)$$

There are two primary implications of the application of model (3) instead of the classical version given in (1):

1. The underlying model is the same, but (3) generalizes across the entire model reference period, whereas (1) was suitable only when $N^* \cong N$.
2. As a result of the change, data users must look up the additional input variable N.

The impact on retrodicting historical standard errors is demonstrated in Figure 3. The solid red line representing (3) shows an increasingly superior fit the farther back in time removed from the anchor point of N^*. (Note that in December 2018, when $N^* = N$, the two standard error predictions are equal.) At the origin of the model reference period (2003), the average bias in (1) is nearly 100,000 persons, or about 30 percent relative bias in the upward direction, when treating the average of the replicates as the objective level.

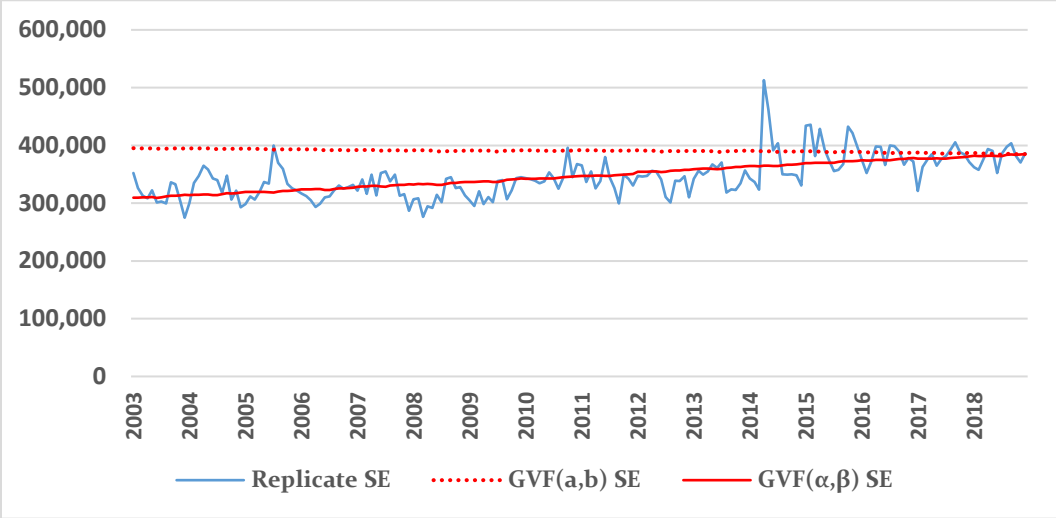

Figure 2: Standard error estimates of *Civilian labor force, 16 years and over*. This is the same as Figure 1 but displays GVF model (1) as the dashed red line and includes GVF model (3) as the solid red line.

Since the population value N is easily obtained and is the same for all level series in a given month, it was not deemed to be overly burdensome to include it as a model input, especially considering that rate series have always required the lookup of two input variables (both the numerator and denominator of the estimate). Gains are achieved in accuracy over time, especially for historical variance estimates subject to more serious violations of the $N^* \cong N$ assumption requisite to (1).

3. GVF Models for Earnings Quantiles

The CPS publishes multiple tables of weekly earnings percentiles in the quarterly news release *Usual Weekly Earnings of Wage and Salary* and the annual news release *Union Membership*. As described in the Introduction, all such quantile estimates are based on partial-sample data, since earnings data are only collected from one quarter of the sample.

The McIllece (2018) framework for creating a GVF model for a median utilized as its foundation the asymptotic variance, under Central Limit Theorem conditions, of a sample quantile $x_q \in (0,1)$ as given in the following equation:

$$v(x_q) \cong \frac{q(1-q)}{n*f(x_q)^2}$$

where n is the sample size and $f(x_q)$ is the density function. Given unequal sample weights in the multi-stage CPS sample and partial-sample collection of earnings data, the sample size term n is replaced by the sum of the outgoing rotation weights (instead of the monthly composite weights used for most estimates) for the appropriate conditional universe S of wage and salary workers[6]. Replicate outgoing rotation weights are available from the Census Bureau and facilitate the computation of replicate variances $v_r(x_q)$ of partial-sample quantile estimates. The density function, modified by a ratio adjustment d that implicitly includes a complex sample design effect, is then estimated via replication.

$$v_r(x_q) \cong \frac{q(1-q)d}{(\sum_{i \in S} w_i)*f(q)^2}$$

Since $y = \sum_{i \in S} w_i$ is the level estimate of wage and salary workers, which is treated as the population base, the equation can be rewritten and rearranged as

$$v_r(x_q) \cong \frac{q(1-q)d}{y*f(q)^2}$$

$$\frac{v_r(x_q)*y}{q(1-q)} = d * f^{-1}(q)^2 \rightarrow se_r(x_q)\sqrt{\frac{y}{q(1-q)}} \qquad (4)$$

The right-hand quantity (4) is the replication-based estimate of the d-adjusted density function, otherwise difficult to obtain formulaically. Since any replication-based measure necessarily includes the inherent period-to-period volatility of the replication procedure itself—which empirically tends to be quite high on a relative scale—modeling (4) aims to reduce said volatility to produce more stable longitudinal estimates of variance. An OLS regression model, using published estimates y and x_q as predictors, is fit to this objective quantity (4):

$$se_r(x_q)\sqrt{\frac{y}{q(1-q)}} = \alpha_0 y + \beta_0 x_q \qquad (5)$$

[6] Most tables in the *Usual Weekly Earnings of Wage and Salary Workers* and *Union Membership* news releases report weekly earnings estimates for full-time wage and salary workers. One table reports weekly earnings estimates for part-time wage and salary workers.

A concern about partial-sample data, as noted in the Introduction, is the small number of model observations, especially for annual series. Parameter estimation is not very robust against influential observations in this case. To mitigate the effects of extreme observations, outliers on the relative-variance scale $v_r(x_q)/x_q^2$ of three-plus standard deviations are removed. This relative variance quantity is more balanced and able to detect outliers in both the low and high directions, since it accounts for the level of the estimate, whereas using $v_r(x_q)$ or $se_r(x_q)$ exclusively identified outliers in the positive direction, resulting in a model that understated variance. The relative variance criterion is able to effectively screen out the most unduly influential outliers without being too intrusive—most series have zero outliers removed, while none have more than two.

After outlier removal, the OLS models are fit to each series individually, utilizing longitudinal histories as the grouping mechanisms. This implicitly satisfies, presuming that the variance properties of unbroken series are stable over time, the condition that design effects are similar within modeling clusters. Empirically, dissimilarity of design effects within clusters has been observed to cause poor GVF model fits in the CPS (McIllece 2016).

Given the estimation of the initial parameters α_0 and β_0, final parameters are obtained by a simple algebraic adjustment to (5):

$$se_r(x_q) = \frac{\alpha_0 y + \beta_0 x_q}{\sqrt{\frac{y}{q(1-q)}}} = \frac{\sqrt{q(1-q)}(\alpha_0 y + \beta_0 x_q)}{\sqrt{y}}$$

Since q is a constant determined by the percentile, it can be incorporated into the parameter estimates to simplify the final GVF model for weekly earnings data:

$$\alpha = \alpha_0\sqrt{q(1-q)}$$

$$\beta = \beta_0\sqrt{q(1-q)}$$

$$se_r(x_q) = \frac{\alpha y + \beta x_q}{\sqrt{y}} \qquad (6)$$

Similarly to model (3) for binomial data, (6) relies on two published parameters and two published estimates. In the case of (3), one of the published data values is actually an administrative total, not an estimate, but the concept is the same: to obtain a standard error on an estimate, the estimate and its population base are required, along with two published GVF parameters.

Figures in Section 3.2 display results of model (6) for weekly earnings estimates, published on a quarterly or annual basis, of some wage and salary worker series.

3.1 Production Standard Error Methodology

The news release *Usual Weekly Earnings of Wage and Salary Workers* already produces standard error estimates for its earnings data. The method involves estimating a one-percent interval around the percentile estimate and multiplying the half-width *w* of that interval by an estimated standard error computed from a binomial variance formulation:

$$w\sqrt{\frac{b}{y} * 100 * q(1-q)} \qquad (7)$$

where

$$w = \frac{x_{q+.01} - x_{q-.01}}{2}$$

and b is a static parameter value that would be occasionally reestimated. Internal research has shown little bias in this formula for estimating the standard errors of percentile estimates relative to replication methods. However, model (7) is subject to fairly intense volatility, particularly for estimates of small subgroups, as will be seen in Section 3.2, rendering them less useful for longitudinal analyses.

Since b (besides occasional updates) and q are constants in the formula, and y is an estimate of the subgroup count that does not fluctuate nearly as much as the percentile estimates, the obvious driver of the volatility is w. The relative magnitude of w can change dramatically between periods and across replicates. Since w acts a multiplier, instability in w results in similarly unstable standard errors for estimates of x_q.

3.2 Comparison of Standard Errors for Weekly Earnings Quantiles

Figures 3 – 7 display selected plots of standard error estimates based on three methods:

1. Replication – Solid blue line with circles
2. Production (7) – Dotted black line[7]
3. GVF Model (6) – Solid red line

The objective of this GVF modeling research is to track the trend of the replicate standard errors well while smoothing through much of the period-to-period volatility. Within each figure, three metrics are reported: average percent bias ($\bar{e}$); variance smoothing percent (v); and total outliers removed.

The metric $\bar{e}$ is computed as the average relative difference between the GVF and replicate standard errors, multiplied by 100 to convert to a percentage basis. Ideally, this bias should be close to zero. In the presence of large positive outliers, which are more common than large negative outliers, this bias tends to be slightly negative.

The smoothing percent v is computed as a ratio of variances of over-the-quarter[8] change, with the GVF variance in the numerator and the replicate variance in the denominator, multiplied by 100 to convert to a percentage basis. As v approaches 100 percent, the more smooth the GVF standard errors relative to the replicate standard errors.

In Figure 3, standard errors are plotted for the 90th percentile weekly earnings estimates of women, 25 years and over, without a high school diploma. Both the replicate and production standard errors are quite volatile; in fact, the production method described in Section 3.1 achieves virtually no smoothing at all. In the last few years of the plot, the standard errors using these methods reaches as high as $45 before dropping precipitously

[7] Production standard error tables only dating back to 2006 were used in this research.
[8] For *Union Membership*, these would be computed as over-the-year change, since that news release has annual periodicity.

to $20 and below, sometimes in a single quarter, before shooting up to $40 or higher again. While the replicate standard errors may have desirable statistical properties (Fay and Train, 1995; Wolter 2005), their inconsistency induces some impracticality when constructing confidence intervals or creating indicators of significant change for time series data.

Comparatively, the GVF model tracks the other series well over time, mostly cutting through the replicates without drifting too high or too low for any extended time periods. GVF model (6) imparts stability into the standard error series while still producing standard error estimates that have little average bias. Additionally, the GVF standard errors are easy for data users to compute, requiring only two published estimates and two published parameters, the latter of which are viable across the entire reference period of the model.

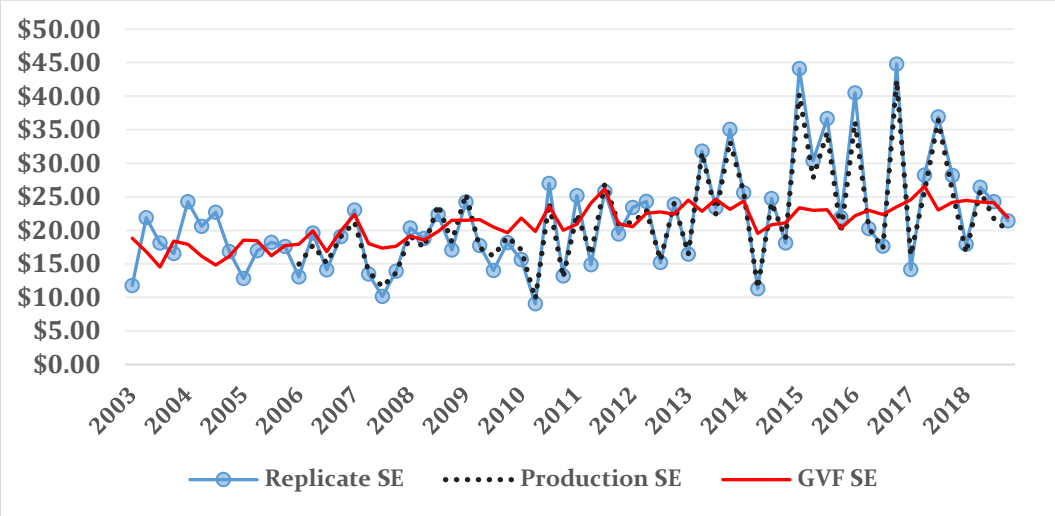

Figure 3: Replicate, production, and GVF standard errors of weekly earnings estimates for *Women, less than a high school diploma, 25 years and over, 90th percentile*. $\bar{e} = -2.7; v = 96.2$. Two outliers were identified.

In Figure 4, standard errors are plotted for the 50th percentile weekly earnings estimates of men, 25 years and over. Since there are no educational breakouts, this is a considerably larger subgroup than in Figure 3, and as such the standard errors are much lower. The volatility here is less pronounced, but relative changes as high as 25 percent or 50 percent from one quarter to the next are not uncommon for the replication series. The GVF model again tracks the replicates well, both in the 2003 – 2009 upward sloping area of the chart and in the ensuing flatter period, which demonstrates how the model can appropriately react based on shifts in the underlying estimates and population size.

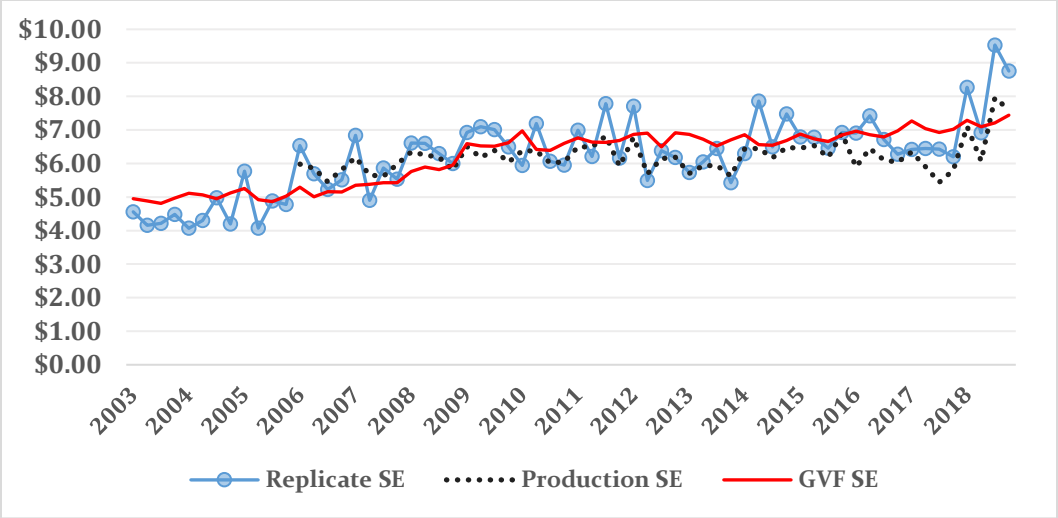

Figure 4: Replicate, production, and GVF standard errors of weekly earnings estimates for *Men, 25 years and over, 50th percentile*. $\bar{e} = 0.2; v = 95.8$. Zero outliers were identified.

In Figure 5, standard errors for first quartile weekly earnings for all in-universe persons of Hispanic or Latino ethnicity show a large, rounded peak in the replicates in 2016 and 2017. However, the underlying first quartile estimates are quite consistent throughout the entire period, displaying steady and stable growth over time, which is reflected in the GVF standard errors.

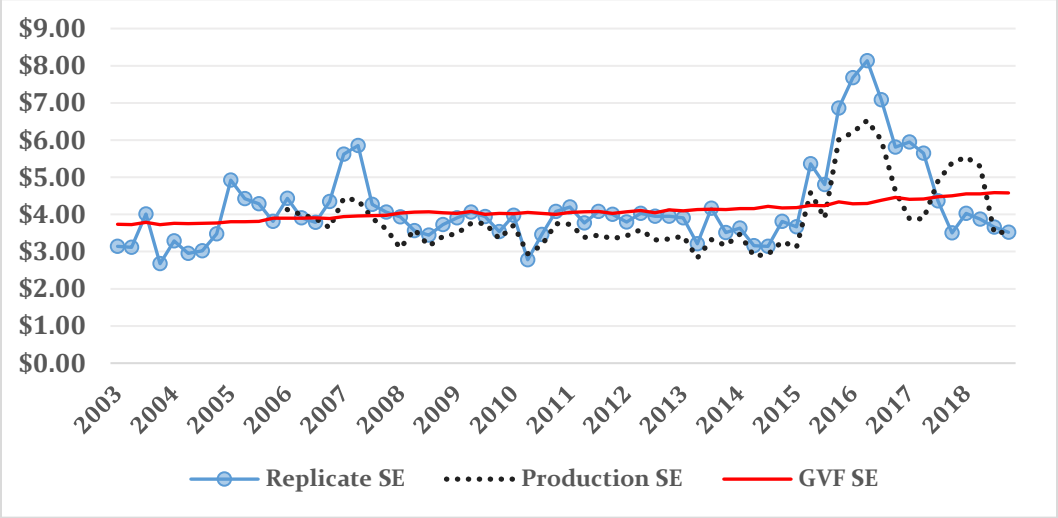

Figure 5: Replicate, production, and GVF standard errors of weekly earnings estimates for *Hispanic or Latino ethnicity, 16 years and over, 25th percentile*. $\bar{e} = -2.4; v = 99.7$. Two outliers were identified.

The final selected series from the quarterly earnings tables is for 10th percentile weekly earnings for men with an advanced degree, 25 years and older, as shown in Figure 6. The GVF standard errors track the trend of the replicates well, while the production standard errors overfit the replicates—under the rubric of smoothing through the sampling error of the variance—until the final year of the time series, when a clear outlier is effectively ignored by both the GVF model and the production model.

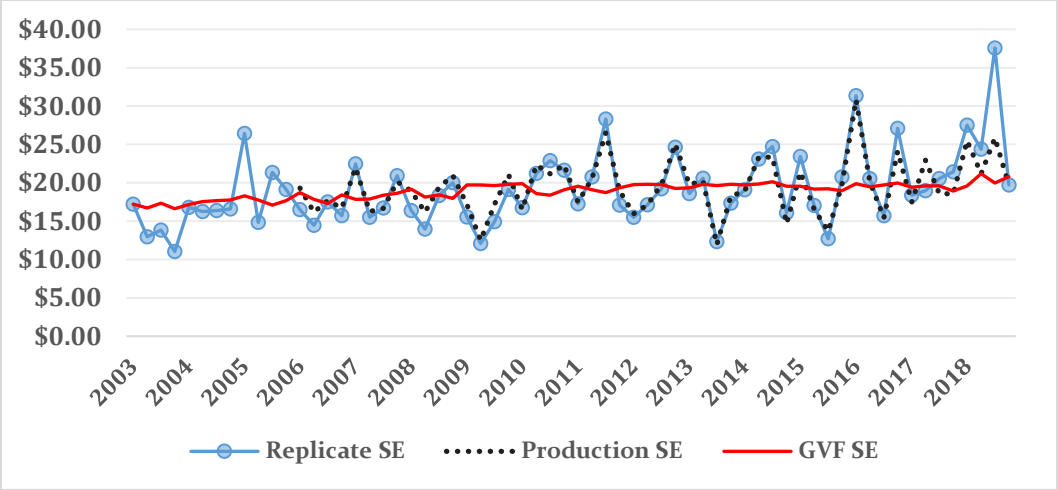

Figure 6: Replicate, production, and GVF standard errors of weekly earnings estimates for *Men with an advanced degree, full-time wage and salary workers 25 years and over, 10th percentile*. $\bar{e} = -1.1; v = 99.0$. One outlier was identified.

Figures 3 – 6 plotted results from the quarterly tables, which are based on 64 model observations. GVF models for weekly earnings series in the annual *Union Membership* news release are based on only 16 observations, which raises a concern for model quality. However, the model form (6) with relative variance outliers removed seems to perform well in this limited-data case. Figures 7 – 9 respectively report results from three median weekly earnings of full-time wage and salary workers series published annually: professional and related occupations, total; construction and extraction occupations, members of unions; and transportation and warehousing, non-union.[9]

Likely to due to the small number of model observations, outliers are rarely identified for the annual series. No outliers were removed when modeling the series in Figures 7 – 9. The GVF standard errors seem to effectively represent the target quantities while smoothing through some of the sampling error, although the smoothing percentages for some series—given by the quantity v in the figures—can dip lower as compared to the quarterly publication. Specifically, $v = 88.8$ percent in Figure 8, while the quantities in Figures 3 – 6 ranged from 95.8 to 99.7 percent. This is still a fairly high level of smoothing, and the bias results tend to be close to zero.

Empirical review of the GVF models for all percentile estimates included in these news releases—10th, 25th, 50th, 75th, and 90th—demonstrates promising consistency of standard error prediction quality, all based on model (6). The application of a consistent model form should be beneficial for data users.

The predicted standard errors for both the quarterly and annual series appear to be of reliable quality and, based on GVF model (6), are no more complex than existing GVF models for monthly binomial estimates.

[9] The *Union Membership* news release includes both occupation—e.g., Figures 7 and 8—and industry estimates—e.g., Figure 9—of median weekly earnings. No production standard errors were included in the research and review of fitting model (6) to the union tables.

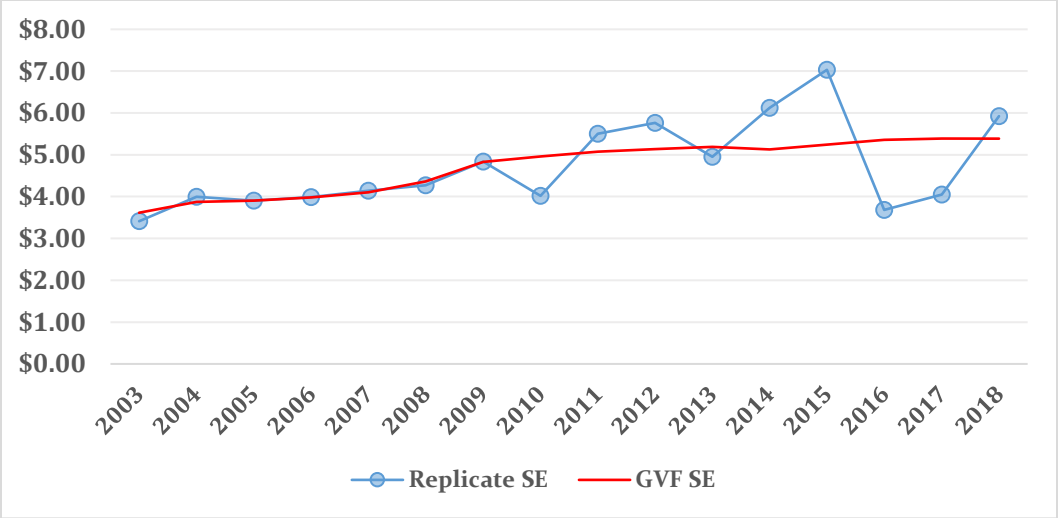

Figure 7: Replicate and GVF standard errors of median weekly earnings estimates for *Professional and related occupations, full-time wage and salary workers, 16 years and over, total*. $\bar{e} = -0.1; v = 98.9$. Zero outliers were identified.

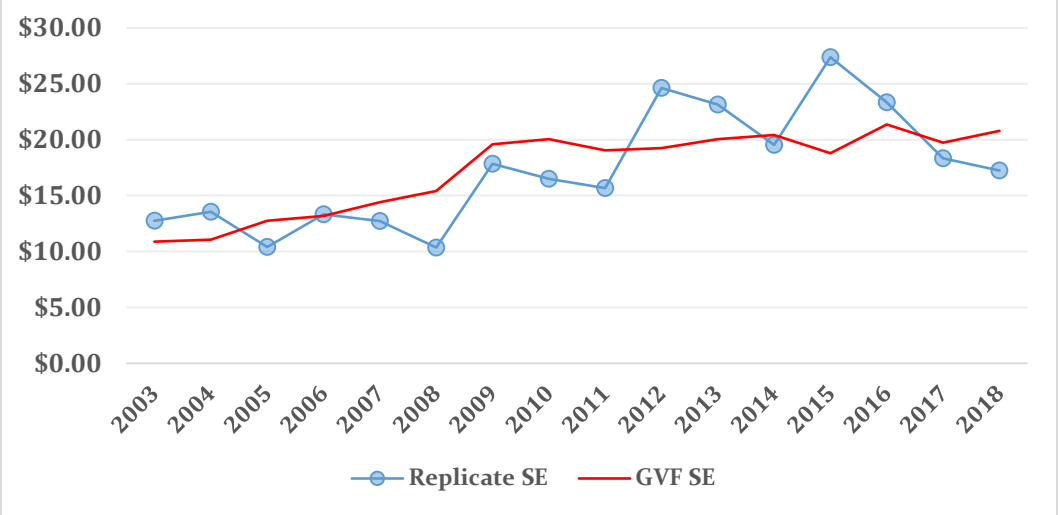

Figure 8: Replicate and GVF standard errors of median weekly earnings estimates for *Construction and extraction occupations, full-time wage and salary workers, 16 years and over, members of unions*. $\bar{e} = 0.0; v = 88.8$. Zero outliers were identified.

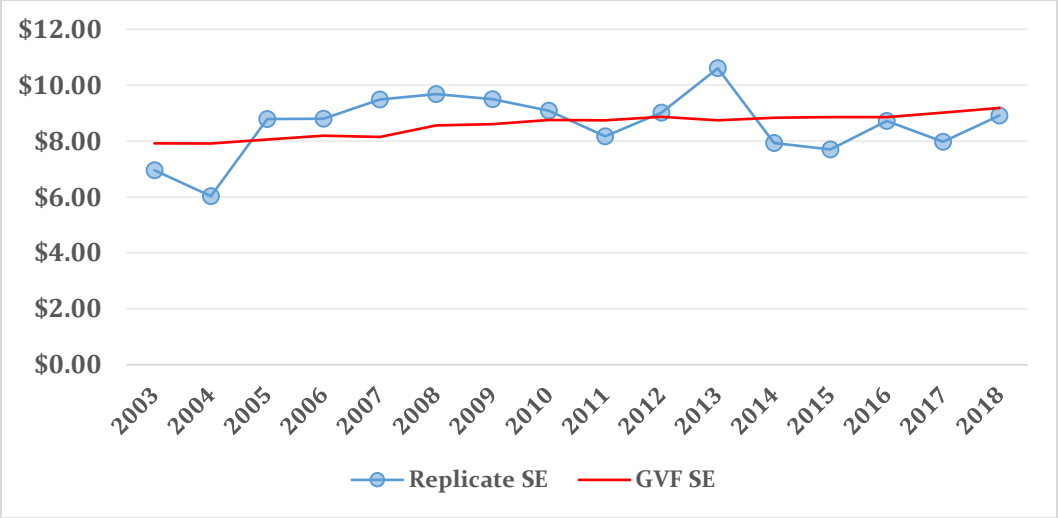

Figure 9: Replicate and GVF standard errors of median weekly earnings estimates for *Transportation and warehousing, full-time wage and salary workers, 16 years and over, non-union*. $\bar{e} = -0.1; v = 99.1$. Zero outliers were identified.

4. Summary

The first primary objective of this paper was to present improvements to the generalizability of GVF models across time. In Section 2, which focused on binomial series, a slight modification of the estimation process—defining the population value by the input variable N to compute α, β parameters, instead of a static N^* associated with a, b parameters—allowed for more accurate standard error estimation across the entire modeling period. Figures 1 and 2 show the improvement, specifically the reduction in bias, of the change in methodology from model (1) to model (3). GVF model (3) has been officially adopted by the CPS; published α, β parameter tables for *The Employment Situation* household tables are available, along with documentation on the CPS reliability webpage[10].

The second primary objective of this paper was to present research GVF models for estimating the standard errors of weekly earnings percentiles. Most GVF models in the literature tend to focus on binomial data (Wolter 2005; McIllece 2006). However, utilizing the asymptotic variance of a sample quantile and estimating the density function by replication, practical GVF models for earnings percentiles were constructed and shown to work well empirically. A comparatively small number of observations was an initial concern for model development, but the results demonstrated that GVF standard errors under model (6) generally captured the trend of the objective replicate series while smoothing through much of the associated sampling error.

A tertiary objective mentioned in the abstract—extending variance models from national- to state-level data series—was abandoned in the early phase of this project. The CPS program publishes very few state-level estimates itself. Instead, the Local Area Unemployment Statistics (LAUS) program is primarily responsible for publishing state-level and other small area estimates[11]. CPS data and sampling error information are used

[10] https://www.bls.gov/cps/documentation.htm#reliability
[11] https://www.bls.gov/lau/laumthd.htm

as inputs to complex small area models for estimates and variances. It was determined to be of little overall value to the program to pursue this direction as a research objective.

The overarching goal of extending variance function coverage to more CPS tables is a continuing process, aimed at making accurate standard errors more widely available, both internally and externally. Models (3) and (6) contribute to this objective and fit within the current publication structure (two input variables; two published parameters) that should be familiar to past users of CPS GVF models. Lastly, the novel characteristics of these models may be of interest to researchers or in the production of other official statistics.

References

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.

McIllece, J.J. (2016). "Calculating Generalized Variance Functions with a Single-Series Model in the Current Population Survey," in *Proceedings of the 2016 Joint Statistical Meetings*, Survey Research Methods Section.

McIllece, J.J. (2018). "On Generalized Variance Functions for Sample Means and Medians," in *Proceedings of the 2018 Joint Statistical Meetings*, Survey Research Methods Section.

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

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