Optimizing the Current Population Survey Composite Estimator

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

Abstract
For monthly labor force estimates, the U.S. Current Population Survey (CPS) utilizes a composite estimator that reduces the variance of over-the-month change by exploiting correlations in the rotating panel design. The gains in efficiency are substantial among highly correlated estimates, such as total employed persons, but come at the cost of a persistent bias due to unequal compositing coefficients and rotation group effects. In this paper, the complex variance of the CPS composite estimator is expressed as a series of geometric series, and the composite estimator's bias is then minimized under the theoretical constraint that the variance of the over-the-month change is not increased relative to official estimates.

Key Words: Current Population Survey; CPS; composite estimation; optimization; geometric series; variance estimation

1. Introduction

Continually since 1947, the Current Population Survey, a joint program of the U.S. Bureau of Labor Statistics (BLS) and the U.S. Census Bureau, has measured the state of the American labor force (Technical Paper 77). While the CPS has undergone numerous sample revisions over its many decades of operation, in its current form (September 2022), the multistage probability survey samples about 60,000 eligible households per month in a 16-month rotating panel design: A household is in sample the first four months; out of sample the next eight months; and in sample the final four months. The first four months (wave 1) are designated month in sample (MIS) 1 – 4, and the final four months in sample (wave 2) are designated MIS 5 – 8. This 4-8-4 rotation scheme is deployed to improve the efficiency of both over-the-month and over-the-year estimates of change (§3.1). In the literature on CPS sampling, MIS are also referred to as rotation groups or panels.

Labor force data is collected for all eligible persons in the household. The CPS universe is the civilian noninstitutional population, 16 years of age and older[2]. Among the thousands of labor force figures published by the CPS are national estimates of total employed persons, total unemployed persons, and the official U.S. unemployment rate, also known as U3. These series are included in *The Employment Situation*, a monthly publication produced by the BLS comprising data from the CPS and the CES (Current Establishment

[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] The survey also collects information on those 15 years of age; however, these individuals are not included in CPS published estimates.

Statistics) that is designated as a Principal Federal Economic Indicator by The White House Office of Management and Budget (Statistical Programs & Standards, 2022).

Two aforementioned CPS series—total employed (EM) and total unemployed (UN)—form the basis of the composite estimation research in the body of this paper. Since the unemployment rate, U3, is computed as the ratio $UN/(EM + UN)$, it is viewed here as a derivative series assumed to be optimized when its component series are optimized.

In §2 – §2.3, current CPS weighting and composite estimation are described, and a generalized compositing formula, useful for optimization, is reviewed.

In §3 – §3.4, issues from the literature relating to CPS composite estimation are summarized, and novel geometric series derivations of the CPS bias and variance estimators are presented. A nonlinear minimization of the bias is performed under the constraint that alternative composite parameterizations produce over-the-month change variance estimates that do not exceed current levels.

Finally, some conclusions are drawn, and a set of recommendations for modifying CPS composite estimation are set forth.

2. CPS Composite Estimation

The CPS is designed to meet both national and state[3] precision requirements. The national design criterion, as stated in Tech Paper 77, is "given a 6.0 percent unemployment rate…the requirement that a difference of 0.2 percentage points in unemployment rate between [two] consecutive months be statistically significant at the 0.10 (90-percent confidence) level[4]."

It is imperative that estimation methods are considerate of the design requirements[5]. Composite estimators, which are essentially weighted averages of current and previous estimates, have historically been used in the CPS to improve the efficiency of estimates of change (Bailar 1975; Breau and Ernst 1983).

Cantwell (1988) neatly summarized these results: "In many surveys, including the Current Population Survey…and the Labour Force Survey of Statistics Canada, participants are interviewed a number of times during the life of the survey, a practice referred to as a rotation design or repeated sampling. Often composite estimation—where data from the current and earlier periods of time are combined—is used to measure the level of a characteristic of interest. As other authors have noted, composite estimation can be used in a rotation design to decrease the variance of estimators of change in level."

---

[3] The state design criterion, as stated in Tech Paper 77: "The required [coefficient of variation] on the annual average unemployment level for each state, substate area, and the District of Columbia, given a 6.0 percent unemployment rate, is 8.0 percent.

[4] The sensitivity threshold is conditional on a 6.0 percent unemployment rate. This is an important distinction because the variance of the unemployment rate, ergo the variance of its over-the-month change, rises as U3 rises. The significant change threshold, at the 90-percent confidence level, can exceed 0.2 percentage points when the unemployment rate is high.

[5] This paper focuses on the national composite estimates of EM and UN and leaves the state design criterion to future discussion and research (§4).

Evidently, decreasing the "variance of estimators of change in level" tends to decrease the variance of estimators of change in the corresponding rates; i.e., a decrease in the variance of the change in total unemployed persons leads to a decrease in the variance of the change in the unemployment rate. This is empirically true in the CPS (when measuring variances by replication) and is assumed throughout this paper, without explicit demonstration, for brevity.

The current form of the CPS composite estimator, called the AK estimator (§2.2), has been in use since 1998.

2.1 CPS Weighting

Before exploring the structure of the AK estimator, CPS weighting must be understood at an elementary level. Full details of the multi-step, multi-dimensional CPS weighting procedures, including detailed benchmarking cells by demographics (such as age, sex, race, and ethnicity), are available in Tech Paper 77.

As a multistage, state-based probability survey, the CPS assigns each sample household a "base" weight equal to the inverse of its probability of selection. With some exceptions, such as design changes and special adjustment factors, most sample households within a state have the same base weight. All eligible individuals (§1) within responding households are assigned a base weight equal to that of their household.

A series of weighting adjustments[6] are then applied, first at the household level and then at the person level. At the end of the weighting procedure, each person-level respondent has a series of weights assigned to them.

The following summary of weighting steps and definitions is drawn verbatim from Tech Paper 77 (p. 67):

1. Base weighting produces simple, unbiased estimates for the basic CPS universe under ideal survey conditions, such as 100 percent response rate, zero frame error, and zero reporting error. Most sample units within a state have the same probability of selection and therefore have the same base weight.
2. Nonresponse adjustment reduces bias that would arise from ignoring [households] that do not respond.
3. First-stage weighting reduces variances due to the sampling of [non-self-representing primary sampling units].
4. State and national coverage steps and second-stage weighting reduce variances by controlling, or benchmarking, CPS estimates of the population to independent estimates of the current population.
5. Composite weighting uses estimates from previous months to reduce the variances, particularly for certain estimates of change.

[6] A subsampling adjustment, called the Weighting Control Factor (WCF), applies in certain situations "to control sample overrun," according to Tech Paper 77. Nonuniversal operational procedures or adjustments like the WCF are excluded from the list of weighting steps for clarity.

Two "estimation weights" are especially important, corresponding to steps 4 and 5 above: the second-stage weight and the composite weight. Both the second-stage weight and the composite weight can be used to produce reasonable estimates[7].

The second-stage weights in step 4 are computed by benchmarking paired MIS totals (MIS 1 & 5; MIS 2 & 6; MIS 3 & 7; MIS 4 & 8) to detailed weighting cells by:

- State/sex/age
- Ethnicity/sex/age
- Race/sex/age

Steps 4 and 5 are respectively described in Tech Paper 77:

"[T]he (second-stage) benchmark procedure adjusts the weights within each MIS pair such that the sample estimates for geographic and demographic subgroups are matched to independent population controls."

And:

"The composite weighting method involves…computation of composite estimates for the main labor force categories, classified by important demographic characteristics; and…the adjustment of the microdata weights, through a series of weighting adjustments, to agree with these composite estimates."

Second-stage weighting is calibrated to highly detailed, independent population controls produced by the Census Bureau's Population Estimates Program (Census 2021), whereas composite weighting adjusts second-stage weights to match marginal composite estimates, at a less detailed level, in a method originally proposed by Fuller (1990) and first applied to CPS data by Lent, Miller, and Cantwell (1994).

Official CPS monthly estimates are based on AK composite estimation; therefore, the AK weight is used for estimating major labor force categories published on a monthly basis, such as total employed, total unemployed, and the national unemployment rate.

Estimates published at a lesser frequency than monthly, such as quarterly or annual statistics, are based on second-stage weights, because the AK compositing procedure is designed and therefore valid only for monthly estimates (§2.2).

Furthermore, monthly estimates based on second-stage weights are included in the AK composite formula. The AK estimator itself is a recursive formula based on the current month's second-stage estimate and the previous month's AK estimate, and as such, it can be alternatively formulated as a weighted linear combination of current and past months' second-stage estimates, as expressed in §2.3. The details are slightly more complex, as the composite weights vary by MIS, which will be discussed in that section.

[7] Reasonable, in this sense, means that the weights have accounted for standard adjustments made in probability surveys, such as nonresponse and benchmarking to independent population controls. Estimates using either second-stage or AK weights are subject to usual survey limitations regarding precision. For example, a small demographic group will tend to have high coefficients of variation using either set of estimation weights.

2.2 CPS Composite (AK) Estimator

The universe total for the CPS, which is independently produced by the Population Estimates Program of the U.S. Census Bureau (Census 2021), is the civilian noninstitutional population, 16 years and over (CNP[8]). The CNP comprises three primary labor force categories:

- Employed (EM)
- Unemployed (UN)
- Not in labor force (NLF)

such that

$$CNP = EM + UN + NLF$$

The EM and UN series are estimated from CPS respondents each month, leaving the NLF series to be computed as a difference, since the CNP is an independent monthly control:

$$NLF = CNP - (EM + UN)$$

Thus, of the three primary labor force groups, only employed and unemployed series are directly composited[9]. NLF is indirectly composited in the sense that it is computed as the difference between a fixed population total and the sum of two directly composited series.

The focus of this paper is on the directly composited series, EM and UN, since the compositing coefficients (i.e., linear combination weights) are determined based on the properties of those two series alone.

Following notational convention from Tech Paper 77, but replacing the superscript CMP with AK for clarity, the current form of the AK estimator is expressed as:

$$\hat{Y}_t^{AK} = (1 - K)\hat{Y}_t^{SS} + K\left(\hat{Y}_{t-1}^{AK} + \Delta_t\right) + A\hat{\beta}_t \qquad [1]$$

where

$$\hat{Y}_t^{SS} = \sum_{i=1}^{8} x_{ti}$$

$$\Delta_t = \frac{4}{3}\left[\sum_{i=2}^{4}\left(x_{ti} - x_{t-1,i-1}\right) + \sum_{i=6}^{8}\left(x_{ti} - x_{t-1,i-1}\right)\right]$$

[8] This universe total is often abbreviated as CNP16+ in CPS publications to reflect that the population is restricted to persons aged 16 years and older. Here, the shorthand CNP is used to avoid formulaic confusion.

[9] The national labor force demographic series directly composited are given in Tables 2-3.5 and 2-3.6 in Tech Paper 77. State-level composites are also produced for 53 state entities, which splits New York into New York City and balance of state, California into Los Angeles County and balance of state, and includes Washington, DC.

$$\hat{\beta}_t = (x_{t1} + x_{t5}) - \frac{1}{3}\left(\sum_{i=2}^{4} x_{ti} + \sum_{i=6}^{8} x_{ti}\right)$$

i = MIS 1, 2, ..., 8

x_{ti} = sum of second-stage weights of respondents in month t and MIS i with characteristic of interest

K = 0.4 for unemployed, 0.7 for employed

A = 0.3 for unemployed, 0.4 for employed

In equation [1], the month t composite estimate of EM or UN is computed as a weighted combination of:

- $\hat{Y}_t^{SS}$, the month t second-stage estimate;
- $\hat{Y}_{t-1}^{AK} + \Delta_t$, the month $t-1$ composite estimate plus an estimate of change in the continuing rotation groups[10] (MIS 2 – 4 and MIS 6 – 8); and
- $\hat{\beta}_t$, the average difference between the second-stage estimates of the new rotation groups (MIS 1 and MIS 5) and the continuing rotation groups.

The K coefficient, which controls the recursion weight, is higher for employed than unemployed since the respective correlations between continuing rotation groups are higher for estimates of EM than UN (Tables 3 and 4). This occurs because employed persons are less likely than unemployed persons to change labor force status from one month to the next. Composite estimation only yields reduction in variance if there is a correlational structure to exploit[11], thus the amount of weight placed on past MIS estimates should be larger when the correlation is high, and vice versa.

Equation [1] is arguably unintuitive in its construction. Breau and Ernst (1983) and Cantwell (1988) established formulas for the Generalized Composite Estimator (GCE), presented in the subsequent section, of which the AK estimator is a special case.

2.3 Generalized Composite Estimator (GCE)

Recent research into the properties of the CPS composite estimator has varied between the cleaner mathematical formulation of the GCE (Erkens 2012, 2017) and the classical AK equations (Cheng, Shao, and Yu 2017). This paper follows the paradigm of the former papers, as derivations of bias and variance (§3.2 – §3.3) of the CPS composite estimator are simplified using the GCE rather than the AK estimator[12].

[10] These six rotation groups are "continuing" because all the sampled households within these MIS in month t were also sampled in month $t-1$. Due to the 4-8-4 rotation pattern, MIS 1 is the first month in sample for a household during wave 1, while MIS 5 is the first month in sample for a household during wave 2.

[11] If the MIS were completely independent, resulting in a correlation matrix of all zero entries, then the labor force estimates for month t with the lowest variances would be the second-stage estimates.

[12] In the opinion of the author. Comparative derivations using the AK and the GCE formulas are not presented.

Equation [2] below is equivalent to Cantwell (1998) with some minor modification to reconcile notation with [1]:

$$\hat{Y}_t^{GCE} = \sum_{i=1}^{8} a_i x_{i,t} - K \sum_{i=1}^{8} b_i x_{i,t-1} + K\hat{Y}_{t-1}^{GCE} \qquad [2]$$

where

$\sum_{i=1}^{8} a_i = 8$

$\sum_{i=1}^{8} b_i = 8$

i = MIS 1, 2, ..., 8

$x_{i,t}$ = sum of second-stage weights of respondents in month *t* and MIS *i* with characteristic of interest

K = 0.4 for unemployed, 0.7 for employed

The AK estimator is a special case of [2] given the *a* and *b* coefficients in Table 1, conditional on the *A* and *K* coefficients (Erkens 2012):

Table 1: GCE coefficients of the AK estimator

MIS	*a*	*b*	*MIS*	*a*	*b*
1	$1 - K + A$	4/3	5	$1 - K + A$	4/3
2	$1 + (K - A)/3$	4/3	6	$1 + (K - A)/3$	4/3
3	$1 + (K - A)/3$	4/3	7	$1 + (K - A)/3$	4/3
4	$1 + (K - A)/3$	0	8	$1 + (K - A)/3$	0

Expanding [2]:

$$\hat{Y}_t^{GCE} = \sum_{i=1}^{8} a_i x_{i,t} - K \sum_{i=1}^{8} b_i x_{i,t-1} + K\hat{Y}_{t-1}^{GCE}$$

$$\hat{Y}_t^{GCE} = \sum_{i=1}^{8} a_i x_{i,t} - K \sum_{i=1}^{8} b_i x_{i,t-1} + K \sum_{i=1}^{8} a_i x_{i,t-1} - K^2 \sum_{i=1}^{8} b_i x_{i,t-2} + K^2 \sum_{i=1}^{8} a_i x_{i,t-2} + \cdots$$

$$\hat{Y}_t^{GCE} = \sum_{i=1}^{8} a_i x_{i,t} + K \sum_{i=1}^{8} (a_i - b_i) x_{i,t-1} + K^2 \sum_{i=1}^{8} (a_i - b_i) x_{i,t-2} + \cdots$$

Letting $d_i = a_i - b_i$ and, therefore, $\sum_{i=1}^{8} d_i = 0$:

$$\hat{Y}_t^{GCE} = \sum_{i=1}^{8} a_i x_{i,t} + \sum_{l=1}^{\infty} \sum_{i=1}^{8} K^l d_i x_{i,t-l} \qquad [3]$$

under the same conditions as [2].

Equation [3], equivalent to [2] from Cantwell, further simplifies the GCE for purposes of this research, specifically by removing the recursion to express the month *t* estimate as a

weighted combination of month $t, \ldots, t-l$ second-stage estimates $x_{i,t-l}$ only; i.e., the prior month's composite estimate $\hat{Y}_{t-1}^{GCE}$ is no longer in the right-hand side of the formula.

Since official CPS estimates are generated from the AK estimator, and the AK estimator is a special case of the generalized composite estimator, then optimizing the GCE becomes a natural research pursuit. The thornier issue is defining the terms of optimization.

3. Composite Estimator Optimization

The AK and GCE each have benefits and drawbacks. Equation [1], though somewhat rigid in its construction, has only two parameters to be estimated, and it possesses historical continuity that can be important to avoid time series disruptions. In formula [2], 17 coefficients must be estimated (or preselected), providing flexibility for reducing bias and variance but also complexity and potential instability in the parameterization. Comparing the GCE to the AK, Erkens (2017) concluded that "the GCE would need to provide some substantial benefits, which depend upon the impact of the MIS effects."

As referred to in §2.1 and §2.2, direct composite estimation is applied for both employed and unemployed in the state dimension as well as at a series of predefined, demographic cells in the national dimension. The current CPS approach is to set the recursion parameters of the AK estimator equal within a labor force category: In formula [1], the parameters ($A = 0.4$, $K = 0.7$) for employed series and ($A = 0.3$, $K = 0.4$) for unemployed series were initially suggested by Lent, Miller, and Cantwell (1994) based on a grid search as an effective compromise across the target estimates of monthly levels, over-the-month changes, and annual averages. They emphasized that their "methods and results…apply to AK estimators, [and] estimators which allow more general coefficients can effect further reductions in the variances." Lent, Miller, Cantwell, and Duff (1999) followed up this research with highly detailed analyses of many important demographic estimates based on various AK parameterizations and made the same recommendation; those AK coefficients are still in use as of September 2022.

In the 1999 paper, Lent et al. further discussed the advantages of varying the A and K parameters by major labor force category (employed and unemployed), noting "(o)ptimal values of the coefficients…depend on the correlation structure of the characteristic to be estimated. Research has shown, for example, that higher values of K and A result in more reliable estimates for employment levels because the ratio estimators for employment are more strongly positively correlated across time than those for unemployment."

Working with the GCE instead of the AK, Erkens (2012) used quadratic programming to minimize the mean squared error of the monthly level estimates, optimizing the coefficients a_i and b_i conditional on K. In 2017, Erkens more explicitly generalized the quadratic GCE approach to enable simultaneous optimization of multiple labor force estimates—positing bias of the monthly level estimates and variance of the over-the-month change estimates, the importance of each being "not in question" according to Lent et al. (1994), as one possibility of coincident objective functions.

This research most closely adheres to that suggestion from Erkens, fixing the *K* parameters within labor force groups to the current values for historical time series consistency[13], while attempting to minimize the bias of the monthly levels under the constraint that the over-the-month change variance not exceed the corresponding variance from the AK estimator. This construction is slightly different from the literature in that it recognizes the need to maintain over-the-month significant change thresholds, especially in an environment of declining response rates. It is unrealistic to believe any estimator would be adopted that structurally inflates the variance of over-the-month change estimates relative to their current magnitudes, particularly for the official unemployment rate. Comparatively, while the bias of the AK estimator is undesirable, it has been accepted as a tradeoff for improved over-the-month change efficiency. Thus, in this research, the variance condition is treated as a constraint, and the bias is conditionally minimized (§3.4).

The GCE coefficients are optimized for total employed and total unemployed rather than for each of the directly composited estimates. This is both an important distinction and limitation. The primary purpose of this paper is to establish the viability of its methods and minimum satisfactory criteria, at the topside level of detail, when considering adjustments to the official composite estimator parameters. However, the goal is *not* to recommend a specific change to CPS coefficients; to do so, additional research at various levels of important demographic detail is requisite.

Annual average estimates, the third primary estimate from the body of composite research by Lent et al., is excluded from the optimization here. Clearly, minimizing the bias of the monthly levels also minimizes the bias of the annual averages, so the optimization results should be beneficial rather than harmful in terms of theoretical expectations. The variance effects are unstudied but not anticipated to be problematic in a general sense; however, that analysis is left to future research.

3.1 Bias and Variance of the AK Estimator

For both total employed and total unemployed, AK compositing induces a systematic difference, or drift, from the second-stage estimates. Persistent MIS biases and unequal MIS weighting coefficients in the AK estimator combine to create this phenomenon.

Table 2: Average multiplicative bias by month in sample, January 2003 – June 2022. Relative to average second-stage estimates.

MIS	*EM Bias*	*UN Bias*
1	1.006	1.113
2	1.006	1.053
3	1.005	1.014
4	1.000	0.989
5	1.002	1.000
6	0.992	0.955
7	0.993	0.939
8	0.996	0.938

The labor force tendency for EM and UN is higher in the early MIS and declines as households remain in sample. That pattern is also apparent within each four-month wave:

[13] The Current Population Survey does not revise past estimates, except in the case of seasonal adjustment.

MIS 5 labor force tendency, after eight months out of sample, is slightly higher than MIS 4 before decreasing throughout the final months in sample. The downward trend in labor force tendency has major implications for bias, as MIS 4 and MIS 8 have the most weight assigned by the AK estimator, leading to a consistent negative bias for the employed and unemployed series.

The bias decomposition approach in McIllece (2020) additively disaggregated the AK estimator into a series of difference terms, one being a "composite drift" component, from January 2003[14] – September 2020. Figure 1 extends that decomposition through June 2022 to demonstrate how, within a few months, the AK estimates reach a difference (blue solid line) of over 200,000 employed persons from the corresponding second-stage estimates before leveling off from 2004 – 2009.

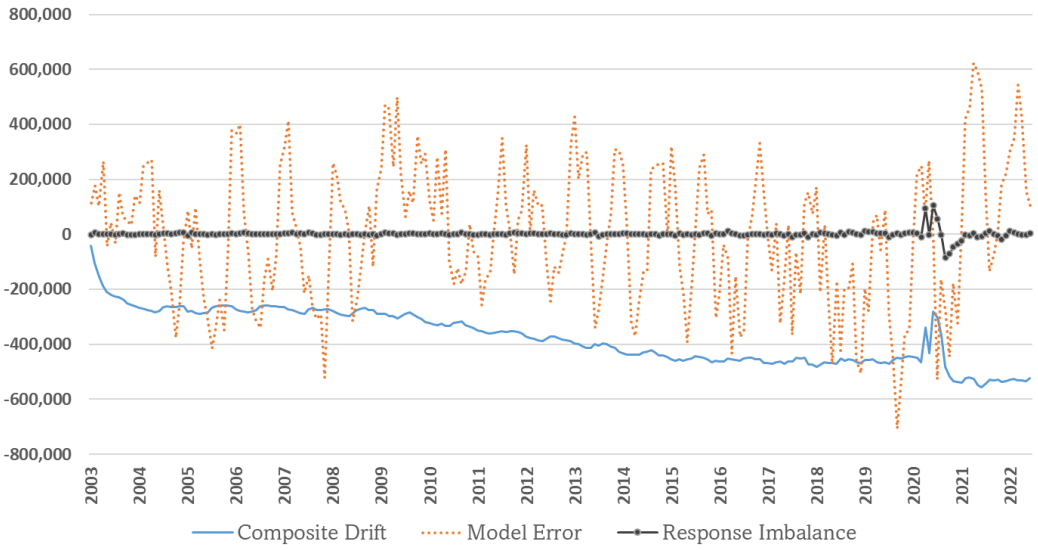

Figure 1: Bias decomposition[15] of (not seasonally adjusted) AK composite estimates of total employed, January 2003 – June 2022.

After the Great Recession of 2007 – 2009, a long-term, gradual recovery of the labor market ensued (Cunningham 2018), and during this period, the AK composite estimates drifted farther from the second-stage estimates. The drift again stabilized for a couple years until the onset of the Covid-19 pandemic, which visibly disrupted the bias components of the EM and UN time series throughout 2020.

[14] Due to changes in CPS weighting, the AK composite estimator was reset in January of 2003 (Tech Paper 77), such that the AK and second-stage estimates were equal in this month, thus marking a natural starting point for comparing the two time series.

[15] The model error (orange dotted line) and response imbalance (black dotted line) are described in McIllece (2020) but are not the subject of this paper.

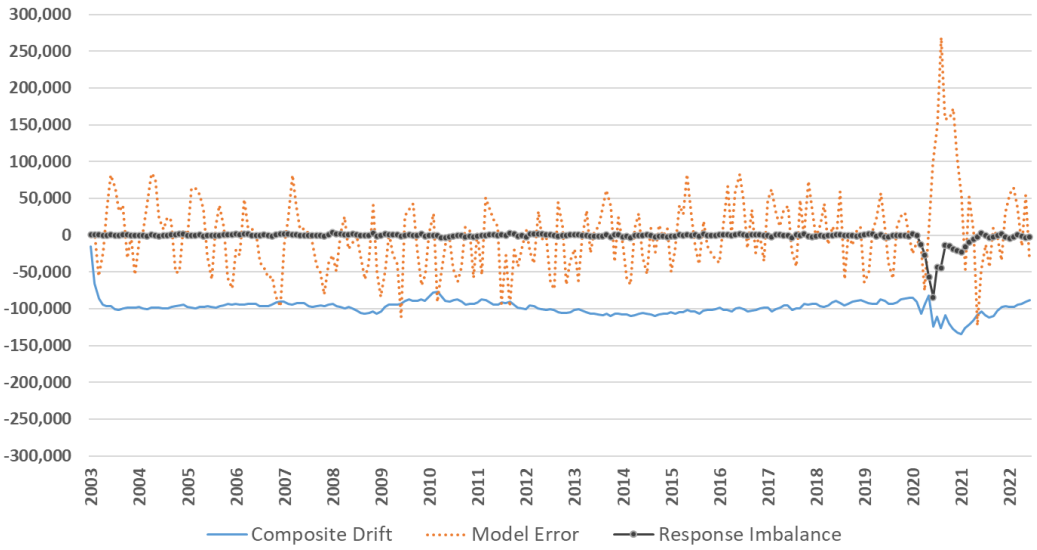

Figure 2: Bias decomposition of (not seasonally adjusted) AK composite estimates of total unemployed, January 2003 – June 2022.

The composite drift of the UN series diverges sharply in just a few months after January 2003, plateauing at a difference magnitude of approximately 100,000 unemployed persons over the course of two decades. Fluctuations are observable around the Great Recession and during the Covid-19 pandemic.

It is noteworthy that the AK estimates do not drift endlessly away from the second-stage baseline; rather, there is a structural difference between the two series that levels off rather quickly in stable economic conditions. Theoretical support for this phenomenon is given in §3.2, but in simple terms, the composite drift results from unequally weighting the eight rotation groups, each of which has its own average bias. The recursive accumulation of these bias terms produces estimates that differ predictably over time.

Following the GCE optimization discussion in §3 and given the design requirements in §2, one of which is based on variance of over-the-month change in the unemployment rate, it is important to maintain the efficiency of change estimates achieved by the AK estimator. However, it is also desirable to ensure that efficiency while reducing bias in the monthly level estimates of EM and UN, and here a common assumption is made: The second-stage estimates of total employed and total unemployed persons are approximately unbiased.

The veracity of this assumption is not directly measurable, but considering that second-stage weighting, as described in §2.1, is calibrated to highly detailed, independent population controls, the assumption seems reasonable. Huang and Ernst (1981) considered three "true" values when considering the AK estimator versus a contemporaneous form of the composite estimator:

1. Ratio estimate from rotation group, or MIS, 1
2. Ratio estimate from all eight MIS
3. Ratio estimate from all eight MIS, adjusted by estimated bias from CPS reinterview responses

Regardless of which of the three they assumed to be unbiased, their "results showed that no matter which of [the] three…were assumed, … the AK composite and the current composite estimator were both negatively biased for the employed and unemployed characteristics considered." Quite a few changes have been implemented to CPS weighting since this study was conducted (1978 – 1980) and MIS biases have become more pronounced over time (Erkens 2017), yet their conclusions about the negative bias of AK compositing on estimates of employed and unemployed remain true today, as shown in Figures 1 and 2.

This result holds across changes in sample designs, questionnaires and collection instruments, expansion of racial definitions, increased dimensionality of second-stage weighting, and other CPS modifications, due to underlying MIS bias patterns that are similar now to those from previous decades; see Bailar (1975), Huang and Ernst (1981), Erkens (2012), and Table 2 (above) for an overarching view of MIS effects in the CPS from the late 1960s all the way through the global Covid-19 pandemic of 2020. The observations of these papers will not be comprehensively restated here, but as a brief sampling consider the MIS 1 unemployment bias indexes reported by these authors at different time periods:

- 1968 – 1969 (Bailar): 1.20
- 1970 – 1971 (Bailar): 1.09
- 1978 – 1980 (Huang et al.[16]): 1.09
- 2003 – 2022 (Table 2): 1.11

Additionally, Erkens showed that the magnitude of the LOESS-smoothed MIS 1 additive bias between 1980 and 2010 ranged from about +100,000 to +150,000 persons, filling the gap in the above data points with evidence that the MIS 1 bias index was consistently greater than 1.00. There have been variations in these biases over time and across rotation groups, but some fundamental findings persist across decades.

While the bias induced by compositing is not ideal, the AK estimator has been effective at reducing the variance of over-the-month change estimates, particularly for the directly composited EM series and generally for indirectly composited labor force estimates that have high month-to-month correlations. Conversely, the directly composited UN series and other labor force estimates with low month-to-month correlations have relatively small reductions in variance of over-the-month change estimates, as shown in Figures 3 and 4.

[16] This bias index was inferred from the bias percentages reported in Table 3 by Huang and Ernst (1981). Specifically: $b_1 = 6{,}729/[\frac{6{,}149}{1-0.08}]$, where b_1 is the multiplicative bias of MIS 1; 6,729 is the average unemployment estimated based only on MIS 1; 6,149 is the average unemployment of the AK estimator; and −0.08 is the relative difference of the AK estimator and the ratio estimator.

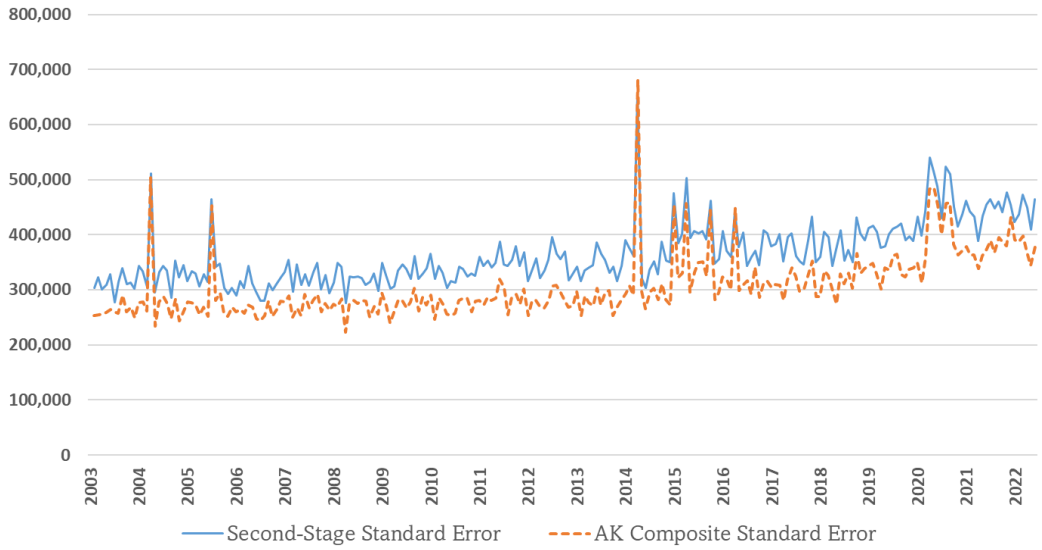

Figure 3: Replicate estimates of (not seasonally adjusted) over-the-month change standard errors for total employed, January 2003 – June 2022.

In Figure 3, the standard errors of the AK composite estimates of total employed persons, computed by replication (Tech Paper 77), are systematically less than the corresponding second-stage standard errors. The average reduction in over-the-month change standard error is 16.2 percent—an average decrease of 58,200 in magnitude.

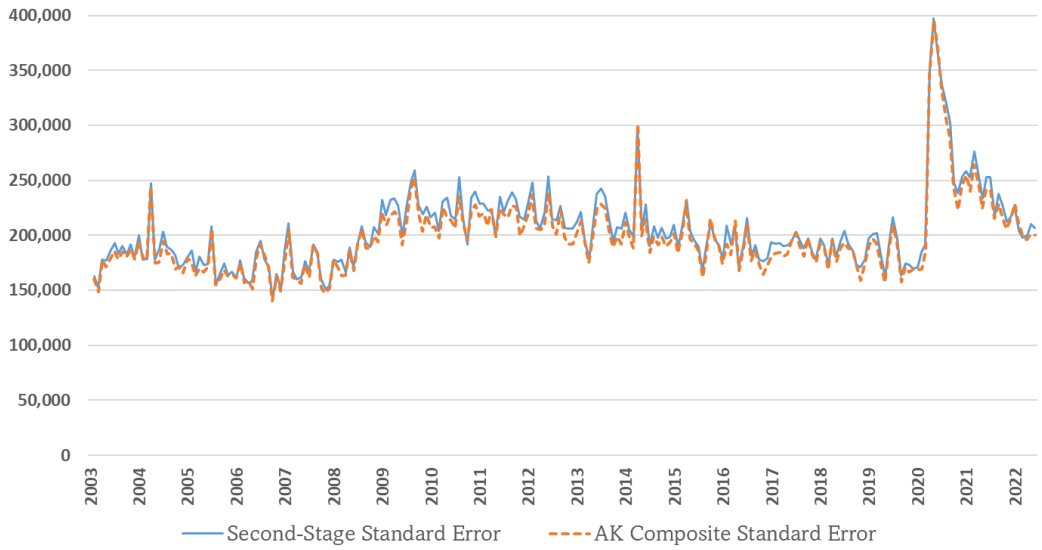

Figure 4: Replicate estimates of (not seasonally adjusted) over-the-month change variance for total unemployed, January 2003 – June 2022.

In contrast to the employed series, reductions in standard errors of changes in unemployment are scantly visible in Figure 4: The average decrease is merely 2.9 percent (6,100 in magnitude). While any gain in efficiency is useful for detecting statistically significant changes in the labor force, when combined with an underestimation bias of about 100,000 persons (Figure 2), the value of AK compositing for the UN labor force series is questionable.

As discussed in §3, a more general estimator like the GCE, with 17 parameters to utilize compared to two parameters in the AK, has the potential to improve the efficiency of CPS labor force estimation. It is also empirically true that increasing the number of parameters leads to greater inconsistency in optimization results that can become difficult to explain, as discussed by Erkens (2017). As a compromise between the understandability and stability of the AK structure and the potentially increased efficiency of the GCE, this research mimics second-stage weighting by first combining MIS into pairs (1 & 5; 2 & 6; 3 & 7; 4 & 8). Since the MIS pairs (rather than the individual MIS) are population-controlled in second-stage weighting, this pairing better stabilizes estimation of the coefficients and also simplifies the correlation matrix, reducing the number of rows from 64 to 16 while producing more correlation terms near zero. In conjunction with fixing the *K* parameter to the current AK settings, the number of parameters is decreased to eight.

§3.2 – §3.4 explore this reduced-form GCE by first deriving formulas for bias and variance and then producing optimized coefficients (under several assumptions and conditions) via nonlinear programming.

3.2 Bias of the GCE

The minimization function in this research is the bias of monthly level estimates for total employed or total unemployed persons. Bias formulas for CPS composite estimators, relative to the second-stage estimate, have been expressed in the literature for [1] and [2] but will be derived here for GCE equation [3], reduced in form to four MIS pairs, which apply to both total employed and total unemployed.

$\hat{Y}_t^{GCE} = \sum_{i=1}^{4} a_i x_{i,t} + \sum_{l=1}^{\infty} \sum_{i=1}^{4} K^l d_i x_{i,t-l}$ = CPS GCE estimate in month *t*

$\hat{Y}_t^{SS} = \sum_{i=1}^{4} x_{i,t}$ = CPS second-stage estimate in month *t*

$B\left(\hat{Y}_t^{GCE}\right) = B\left(\sum_{i=1}^{4} a_i x_{i,t} + \sum_{l=1}^{\infty} \sum_{i=1}^{4} K^l d_i x_{i,t-l}\right) = \frac{E\left(\hat{Y}_t^{GCE}\right)}{\hat{Y}_t^{SS}} =$ multiplicative bias of $\hat{Y}_t^{GCE}$

where

i = MIS pair (i,j) where $i \in [1,2,3,4]$; $j = i + 4$

$x_{i,t}$ = second-stage estimate of MIS pair *i* in month *t*

$\bar{x}_t = \frac{1}{4} \sum_{i=1}^{4} x_{i,t}$ = average of all four MIS pair second-stage estimates in month *t*

$b_{i,t} = \frac{x_{i,t}}{\bar{x}_t}$ = multiplicative bias of MIS pair *i* in month *t*

$\bar{b}_i = \frac{1}{T} \sum_{t=1}^{T} b_{i,t}$ = average multiplicative bias of MIS pair *i* across all months *T*

Then making the following assumptions:

1. $E\left(b_{i,t}\right) = \bar{b}_i$
2. $E\left(x_{i,t-l}\right) = \bar{x}_{i,t}$

The first of these two assumptions is reasonable when the MIS bias patterns are fairly stable over time, as was shown in §3.1. The second assumption, *prima facie*, may appear unreasonable, as obviously there is labor force growth as populations grow (EM) or sudden change during economic instability (UN). Yet it works well empirically due to the exponential decay in the *K* parameters of the GCE. Considering $K = 0.4$ for unemployed, then lagged merely four months, the second-stage estimate from month t-4 has an exponential factor of $K^4 = 0.4^4 = 0.0026$, or nearly zero. Viewed from this perspective, the assumption about the expectation of the second-stage estimate of MIS i seems less worrisome[17].

Given this pair of assumptions, the multiplicative bias of the monthly level estimate can be derived by first computing the expectation of $\hat{Y}_t^{GCE}$:

$$E\left(\hat{Y}_t^{GCE}\right) = E\left(\sum_{i=1}^{4} a_i x_{i,t} + \sum_{l=1}^{\infty}\sum_{i=1}^{4} K^l d_i x_{i,t-l}\right)$$

$$= \sum_{i=1}^{4} a_i E\left(x_{i,t}\right) + \sum_{l=1}^{\infty}\sum_{i=1}^{4} K^l d_i E\left(x_{i,t-l}\right)$$

$$= \sum_{i=1}^{4} a_i \bar{b}_i \bar{x}_t + \sum_{l=1}^{\infty}\sum_{i=1}^{4} K^l d_i \bar{b}_i \bar{x}_t$$

Note the definition of a geometric series: $\sum_{t=0}^{\infty} ar^t = \frac{a}{1-r} \;\forall\; |r| < 1$.

Since K is a parameter between 0 and 1 in the GCE, and the a_i and d_i coefficients are invariant, the infinite sum within the expectation is geometric:

$$E\left(\hat{Y}_t^{GCE}\right) = \sum_{i=1}^{4} a_i \bar{b}_i \bar{x}_t + \sum_{t=0}^{\infty}\sum_{i=1}^{4} K^t K d_i \bar{b}_i \bar{x}_t$$

$$= \sum_{i=1}^{4} a_i \bar{b}_i \bar{x}_t + \sum_{i=1}^{4} \frac{K}{1-K} d_i \bar{b}_i \bar{x}_t$$

$$= \sum_{i=1}^{4} \left(a_i + \frac{K}{1-K} d_i\right) \bar{b}_i \bar{x}_t$$

Recalling the definition of the second-stage estimate $\hat{Y}_t^{SS} = \sum_{i=1}^{4} x_{i,t} = 4\bar{x}_t$:

$$E\left(\hat{Y}_t^{GCE}\right) = \sum_{i=1}^{4} \left(a_i + \frac{K}{1-K} d_i\right) \bar{b}_i \bar{x}_t = \left(\frac{1}{4}\hat{Y}_t^{SS}\right) \sum_{i=1}^{4} \left(a_i + \frac{K}{1-K} d_i\right) \bar{b}_i$$

$$\therefore\; B\left(\hat{Y}_t^{GCE}\right) = \frac{E\left(\hat{Y}_t^{GCE}\right)}{\hat{Y}_t^{SS}} = \left[\left(\frac{1}{4}\hat{Y}_t^{SS}\right) \sum_{i=1}^{4} \left(a_i + \frac{K}{1-K} d_i\right) \bar{b}_i\right] \Big/ \hat{Y}_t^{SS}$$

$$= \frac{1}{4} \sum_{i=1}^{4} \left(a_i + \frac{K}{1-K} d_i\right) \bar{b}_i \qquad [4]$$

[17] The $K = 0.7$ parameter for employed decays less rapidly, but short-term changes in employment tend to be relatively small. For example, a sizable over-the-month change of one million employed persons would still represent less than one percent of the employed labor force. An average growth ratio can be added into the expectation to allow for increasing employment, but this inclusion has negligible effects on optimization results.

3.3 Variance of the GCE

The optimization constraint in this research is based on the variance of over-the-month change estimates for total employed or total unemployed persons. Variance formulas have been derived for equation [2] in the literature, including by Cantwell (1990).

Equation [3], which equivalently expresses the GCE as only a weighted linear combination of second-stage estimates, lends itself to a useful theoretical derivation presented below.

First, the over-the-month change estimate, $\Delta\hat{Y}_t^{GCE}$, is reduced:

$$\Delta\hat{Y}_t^{GCE} = \hat{Y}_t^{GCE} - \hat{Y}_{t-1}^{GCE} = \left[\sum_{i=1}^{4} a_i x_{i,t} + \sum_{l=1}^{\infty}\sum_{i=1}^{4} K^l d_i x_{i,t-l}\right] - \left[\sum_{i=1}^{4} a_i x_{i,t-1} + \sum_{l=1}^{\infty}\sum_{i=1}^{4} K^l d_i x_{i,(t-1)-l}\right]$$

$$= \sum_{i=1}^{4} a_i x_{i,t} + \sum_{i=1}^{4}\left(K d_i x_{i,t-1} - a_i x_{i,t-1}\right) + \sum_{l=2}^{\infty}\sum_{i=1}^{4}\left(K^l d_i x_{i,t-l} - K^{l-1} d_i x_{i,t-l}\right)$$

$$= \sum_{i=1}^{4} a_i x_{i,t} + \sum_{i=1}^{4}(K d_i - a_i) x_{i,t-1} + \sum_{l=2}^{\infty}\sum_{i=1}^{4}\left(K^l - K^{l-1}\right) d_i x_{i,t-l}$$

$$= \sum_{i=1}^{4} a_i x_{i,t} + \sum_{i=1}^{4}(K d_i - a_i) x_{i,t-1} + \sum_{l=2}^{\infty}\sum_{i=1}^{4} K^l \left(\frac{K-1}{K}\right) d_i x_{i,t-l} \qquad [5]$$

Two simplifying assumptions are made to facilitate computation of the variance of [5]:

1. $V\left(x_{i,t}\right) = \sigma^2 \;\forall\; i,t$
2. $Cov\left(x_{i,t}, x_{j,t-l}\right) = r_{i\,j,l}\sigma_t^2 \;\forall\; i,j,l,t$

where *t* is the current month, $l \in [0,\cdots,\infty)$ is the monthly lag, and $i \in [1,\ldots,4]$ and $j \in [1,\ldots,4]$ are MIS pair indexes.

The underlying ideas to this pair of assumptions are that the second-stage estimates have equal variances and that the correlations between MIS pairs (i,j) at time $(t-l)$ do not depend on *t*. Neither is strictly true, but they function well empirically, as will be shown in Table 6. The first assumption is aided by the exponential decay of recursion in the composite estimator, such that the difference between assuming $V\left(x_{i,t}\right) = \sigma^2 \;\forall\; i,t$ and $V\left(x_{i,t}\right) = \sigma_t^2 \;\forall\; i$ is practically inconsequential. The second assumption is effective because the correlational patterns have changed little since 2003—using the average correlations by MIS pair at incremental monthly lags fits the structure well whether at the beginning or the end of the time span in question.

$$V\left(\Delta\hat{Y}_t^{GCE}\right) = V\left(\sum_{i=1}^{4} a_i x_{i,t} + \sum_{i=1}^{4}(K d_i - a_i) x_{i,t-1} + \sum_{l=2}^{\infty}\sum_{i=1}^{4} K^l \left(\frac{K-1}{K}\right) d_i x_{i,t-l}\right)$$

Given assumptions 1 and 2, the variance of [5] can be computed as a sum of infinite series at each monthly lag $l = 0$, $l = 1$, …, $l = \infty$.

Let $V_l\left(\Delta\hat{Y}_t^{GCE}\right)$ be the partial covariance terms corresponding to lag l:

$$V_0\left(\Delta\hat{Y}_t^{GCE}\right) = \sum_{i=1}^{4}\sum_{j=1}^{4} a_i a_j r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)\left(Kd_j - a_j\right)r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^4\left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^6\left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,0}\,\sigma^2 + \cdots$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{4} a_i a_j r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)\left(Kd_j - a_j\right)r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^2(K-1)^2 d_i d_j r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^4(K-1)^2 d_i d_j r_{ij,0}\,\sigma^2 + \cdots$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{4} a_i a_j r_{ij,0}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)\left(Kd_j - a_j\right)r_{ij,0}\,\sigma^2 + \sum_{t=0}^{\infty}\sum_{i=1}^{4}\sum_{j=1}^{4}(K^2)^t K^2(K-1)^2 d_i d_j r_{ij,0}\,\sigma^2$$

Note the definition of a geometric series: $\sum_{t=0}^{\infty} ar^t = \frac{a}{1-r}\ \forall\ |r| < 1$.

Since K is a parameter between 0 and 1 in the GCE, and the various coefficients are invariant, the triple sum in the partial covariance $V_0\left(\Delta\hat{Y}_t^{GCE}\right)$ is geometric. Thus, it can be easily reduced, and the overall $V_0\left(\Delta\hat{Y}_t^{GCE}\right)$ formula can be written in closed form:

$$V_0\left(\Delta\hat{Y}_t^{GCE}\right) = \left(\sum_{i=1}^{4}\sum_{j=1}^{4} a_i a_j r_{ij,0} + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)\left(Kd_j - a_j\right)r_{ij,0} + \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{K^2(K-1)^2}{1-K^2} d_i d_j r_{ij,0}\right)\sigma^2$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{4}\left(a_i a_j r_{ij,0} + (Kd_i - a_i)\left(Kd_j - a_j\right)r_{ij,0} + \frac{K^2(K-1)^2}{1-K^2} d_i d_j r_{ij,0}\right)\sigma^2$$

For the lag $l = 1$ covariance terms:

$$V_1\left(\Delta\hat{Y}_t^{GCE}\right) = 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} a_i\left(Kd_j - a_j\right)r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)K^2\left(\frac{K-1}{K}\right)d_j r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^5\left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^7\left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,1}\,\sigma^2 + \cdots\right]$$

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} a_i\left(Kd_j - a_j\right)r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K(K-1)(Kd_i - a_i)d_j r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^3(K-1)^2 d_i d_j r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^5(K-1)^2 d_i d_j r_{ij,1}\,\sigma^2 + \cdots\right]$$

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} a_i\left(Kd_j - a_j\right)r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K(K-1)(Kd_i - a_i)d_j r_{ij,1}\,\sigma^2 + \sum_{t=0}^{\infty}\sum_{i=1}^{4}\sum_{j=1}^{4}(K^2)^t K^3(K-1)^2 d_i d_j r_{ij,1}\,\sigma^2\right]$$

Resolving the geometric series:

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} a_i\left(Kd_j - a_j\right)r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K(K-1)(Kd_i - a_i)d_j r_{ij,1}\,\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{K^3(K-1)^2}{1-K^2} d_i d_j r_{ij,1}\,\sigma^2\right]$$

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4}\left(a_i(Kd_j - a_j)r_{ij,1} + K(K-1)(Kd_i - a_i)d_j r_{ij,1} + \frac{K^3(K-1)^2}{1-K^2} d_i d_j r_{ij,1}\right)\right]\sigma^2$$

.

For the $l \in [2,\cdots,L]$ covariance terms, where L is defined to be large enough such that $V_{L+1}\left(\Delta\hat{Y}_t^{GCE}\right) + \cdots + V_{\infty}\left(\Delta\hat{Y}_t^{GCE}\right) \approx 0$:

$$V_l\left(\Delta\hat{Y}_t^{GCE}\right) = 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} a_i K^l \left(\frac{K-1}{K}\right) d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4}(Kd_i - a_i)K^l K\left(\frac{K-1}{K}\right) d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^l K^4 (K^2)^0 \left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^l K^4 (K^2)^1 \left(\frac{K-1}{K}\right)^2 d_i d_j r_{ij,l}\sigma^2 + \cdots\right]$$

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} K^l \left(\frac{K-1}{K}\right) a_i d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^l (K-1)(Kd_i - a_i) d_j r_{ij,l}\sigma^2 + \sum_{t=0}^{\infty}\sum_{i=1}^{4}\sum_{j=1}^{4} K^l K^2 (K^2)^t (K-1)^2 d_i d_j r_{ij,l}\sigma^2\right]$$

Resolving the geometric series:

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4} K^l \left(\frac{K-1}{K}\right) a_i d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} K^l (K-1)(Kd_i - a_i) d_j r_{ij,l}\sigma^2 + \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{K^l K^2 (K-1)^2}{1-K^2} d_i d_j r_{ij,l}\sigma^2\right]$$

$$= 2\left[\sum_{i=1}^{4}\sum_{j=1}^{4}\left(K^l \left(\frac{K-1}{K}\right) a_i d_j r_{ij,l} + K^l (K-1)(Kd_i - a_i) d_j r_{ij,l} + \frac{K^l K^2 (K-1)^2}{1-K^2} d_i d_j r_{ij,l}\right)\right]\sigma^2$$

Then combining the partial covariance terms:

$$V\left(\Delta\hat{Y}_t^{GCE}\right) = \sum_{l=0}^{L} V_l\left(\Delta\hat{Y}_t^{GCE}\right) \qquad [6]$$

3.4 GCE Optimization

As previously stated, the optimization goal of this paper is to minimize the bias of the monthly level estimates under the constraint that the over-the-month change variance of the reparameterized GCE be no larger than the corresponding AK variance. A fundamental component of the $V_l\left(\Delta\hat{Y}_t^{GCE}\right)$ covariance terms is the correlation matrix {R}, which can be estimated directly through CPS second-stage replicate weights and averaged over time to meet the condition of the static correlation terms specified in equations [4] and [6].

Tables 3 and 4 display the {R} matrices for employed and unemployed, respectively. The monthly lag $t - l$ is given horizontally, beginning with lag $l = 0$; i.e., the correlation between MIS pair i and MIS pair j in the same month, which explains why the correlations for (i, j) pairs (1,1), (2,2), (3,3), and (4,4) in the lag $l = 0$ column are equal to 1.0, as they represent the same second-stage estimates of total employed.

Table 3: Average second-stage replicate correlations between MIS pairs (i,j) at monthly lag $l \in [0,\cdots,16]$ for estimates of total employed, January 2003 – June 2022.

(i,j)	t-0	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12	t-13	t-14	t-15	t-16
(1,1)	1.00	0.01	0.02	0.01	0.05	0.01	0.01	0.00	0.04	0.01	0.01	0.01	0.23	0.00	0.02	0.00	0.05
(2,1)	0.01	0.78	0.01	0.02	0.01	0.05	0.01	0.01	0.00	0.04	0.00	0.00	0.01	0.22	0.00	0.01	0.00
(3,1)	0.01	0.01	0.71	0.01	0.02	0.00	0.05	0.00	0.01	0.01	0.05	0.00	0.01	0.00	0.22	0.00	0.02
(4,1)	0.01	0.02	0.01	0.66	0.01	0.01	0.00	0.04	0.01	0.01	0.01	0.04	0.00	0.00	0.01	0.21	0.00
(1,2)	0.01	0.02	0.01	0.05	0.01	0.01	0.00	0.05	0.01	0.01	0.01	0.25	0.01	0.02	0.00	0.05	0.01
(2,2)	1.00	0.01	0.02	0.01	0.05	0.01	0.01	0.01	0.05	0.00	0.01	0.01	0.24	0.00	0.01	0.00	0.04
(3,2)	0.01	0.80	0.01	0.02	0.00	0.05	0.01	0.01	0.01	0.05	0.00	0.01	0.00	0.24	0.01	0.01	-0.01
(4,2)	0.02	0.01	0.73	0.02	0.02	0.00	0.05	0.01	0.01	0.00	0.05	0.00	0.01	0.01	0.23	0.01	0.01
(1,3)	0.01	0.01	0.06	0.01	0.01	0.00	0.04	0.01	0.01	0.02	0.26	0.01	0.02	0.01	0.05	0.01	0.00
(2,3)	0.01	0.02	0.01	0.06	0.01	0.01	0.00	0.04	0.00	0.01	0.01	0.24	0.00	0.01	0.00	0.04	0.00
(3,3)	1.00	0.01	0.02	0.00	0.05	0.01	0.01	0.00	0.05	0.00	0.01	0.01	0.24	0.01	0.02	0.00	0.04
(4,3)	0.01	0.81	0.01	0.02	0.00	0.05	0.01	0.01	0.01	0.04	0.00	0.00	0.01	0.24	0.01	0.02	0.00
(1,4)	0.01	0.06	0.01	0.01	0.00	0.04	0.01	0.01	0.02	0.26	0.01	0.02	0.01	0.05	0.01	0.01	0.00
(2,4)	0.02	0.01	0.06	0.00	0.01	0.00	0.04	0.00	0.01	0.02	0.25	0.00	0.01	0.00	0.04	0.01	0.01
(3,4)	0.01	0.02	0.00	0.05	0.01	0.02	0.00	0.05	0.00	0.01	0.01	0.25	0.01	0.01	0.00	0.04	0.00
(4,4)	1.00	0.01	0.02	0.00	0.05	0.00	0.01	0.00	0.04	0.00	0.01	0.01	0.25	0.01	0.01	0.00	0.04

In the lag $l = 1$ column, the correlation is highest for MIS pairs one month apart. In the total employed correlation matrix, for example, MIS pair 2 in month t is highly correlated (r = 0.78) with MIS pair 1 in month $t - 1$, because these two pairs represent the same set of sample households moving through the 4-8-4 rotation pattern. Particularly for employed respondents, labor force status does not change often, especially at short time lags. The correlation gradually declines for the overlapping sample cells as the monthly lag increases, as evident in the highlighted cells in {R} dropping from an average 0.80 correlation one month apart to approximately 0.25 twelve months apart.

Table 4: Average second-stage replicate correlations between MIS pairs (i,j) at monthly lag $l \in [0,\cdots,16]$ for estimates of total unemployed, January 2003 – June 2022.

(i,j)	t-0	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12	t-13	t-14	t-15	t-16
(1,1)	1.00	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.02	-0.01	0.00	0.01	0.08	0.00	0.00	0.01	0.02
(2,1)	0.01	0.47	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.06	0.00	0.00	0.01
(3,1)	0.00	0.00	0.36	0.00	0.01	0.00	0.02	0.01	0.00	0.00	0.01	0.00	-0.01	0.01	0.06	0.01	0.00
(4,1)	0.01	0.00	0.00	0.30	-0.01	0.01	0.01	0.02	0.00	-0.01	0.00	0.02	0.01	0.00	0.01	0.06	0.01
(1,2)	0.01	0.00	0.01	0.03	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.08	0.01	0.00	0.00	0.02	0.01
(2,2)	1.00	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.00	0.01	0.00	0.08	0.00	0.00	-0.01	0.02
(3,2)	0.00	0.49	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.02	-0.01	0.00	0.01	0.08	0.01	0.00	0.00
(4,2)	0.01	0.01	0.39	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.01	0.01	0.00	0.07	0.01	0.00
(1,3)	0.00	0.01	0.02	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.07	0.00	0.00	0.01	0.02	0.00	-0.01
(2,3)	0.00	0.01	0.01	0.02	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.01	0.01
(3,3)	1.00	0.01	0.01	0.00	0.02	0.01	0.00	0.00	0.02	0.00	0.00	0.01	0.07	0.00	0.00	0.01	0.02
(4,3)	0.01	0.51	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.07	0.00	0.00	0.00
(1,4)	0.01	0.02	0.00	0.00	0.00	0.02	0.00	0.01	-0.01	0.08	0.01	0.00	0.00	0.03	0.00	0.00	0.00
(2,4)	0.01	0.01	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.09	0.01	0.00	-0.01	0.01	0.01	0.00
(3,4)	0.01	0.00	0.01	0.01	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.09	0.01	0.00	0.00	0.01	0.01
(4,4)	1.00	0.01	0.01	0.01	0.02	0.00	-0.01	0.00	0.02	0.01	0.01	0.00	0.08	0.00	0.00	0.00	0.01

Comparatively, for the total unemployed matrix in Table 4, the analogous correlations are about 0.49 and 0.08. A reported labor force status of unemployed is more likely to change during a respondent's time in sample, weakening lag correlations for panel estimates of unemployment.

Since the 4-8-4 design comprises 16 months including month t, at lag 16, there is no sample overlap among any of the MIS pairs. At lag $l = 16$ and beyond, the entries of $\{R\}$ are near zero for all MIS pairs. Considering the exponential decay of the recursion parameter K in the GCE and the minimal correlations, the effects of including lag $l = 16$ and beyond are negligible in estimation; as such, the maximum lag in equation [6] is set to $L = 15$.

It is necessary to compute the theoretical over-the-month change variance of the AK estimator for both total employed and total unemployed to set the optimization constraints. The theoretical bias and variance of the AK estimator can be calculated using formulas [4] and [6], respectively, yielding the results in Table 5:

Table 5: Multiplicative bias of monthly level estimates and variance reductions of over-the-month change variances, AK estimator. All biases and variance reductions relative to second-stage estimates.

Labor Force Estimate	*Multiplicative Bias*	*Theoretical Variance Ratio*
Employed	0.997	0.713
Unemployed	0.989	0.962

The table results are consistent with Figures 1 – 4: The biases are below 1.00, indicating the AK composite underestimates both employed and unemployed persons, while the variance ratios show a substantial reduction for employed and much less of an improvement for unemployed monthly changes. As a comparison to the theoretical results, the average biases and average replicate variances are computed in Table 6:

Table 6: Multiplicative bias of monthly level estimates and variance reductions of over-the-month change variances, AK estimator, theoretical versus observed. All biases and variance reductions relative to second-stage estimates.

Labor Force Estimate	*AK Theoretical Multiplicative Bias*	*AK Theoretical Variance Ratio*	*AK Average Multiplicative Bias*	*AK Average Replication Variance Ratio*
Employed	0.997	0.713	0.997	0.705
Unemployed	0.989	0.962	0.989	0.942

The comparison demonstrates that the theoretical bias accurately reflects the observed bias, and the variance ratios are quite similar, especially for employed. For unemployed, a difference of 0.02 is still quite small and is a likely consequence of the simplifying assumptions and the variance of the replicate variance, which itself can be quite noisy (McIllece 2016). Overall, the theoretical results appear well supported and sufficient for use in optimization.

Applying the nonlinear programming solver in the SAS *OPTMODEL* procedure (SAS Institute 2022) to minimize the bias from equation [4], under the constraint that the variance in [6] is no larger than the theoretical AK variances implied by Table 5, yields alternative parameterizations for the GCE, subject to the MIS-pair constraints $\sum_{i=1}^{4} a_i = 4$ and $\sum_{i=1}^{4} d_i = 0$. The optimized GCE coefficients are reported in Tables 7 and 8. Table 9 reprints the theoretical bias and variance of the AK composite estimator and compares them to the optimized GCE results.

Table 7: Optimized GCE coefficients for total employed by MIS pair, under prespecified conditions. $K = 0.7$.

MIS pair	a_i	d_i
(1,5)	0.6001	−0.0323
(2,6)	0.7736	−0.7771
(3,7)	1.2871	−0.3666
(4,8)	1.3392	1.1760

Table 8: Optimized GCE coefficients for total unemployed by MIS pair, under prespecified conditions. $K = 0.4$.

MIS pair	a_i	d_i
(1,5)	0.9563	0.1935
(2,6)	1.0284	−0.1228
(3,7)	0.9826	−0.6642
(4,8)	1.0327	0.5935

Table 9: Multiplicative bias of monthly level estimates and variance reductions of over-the-month change variances, AK estimator versus optimized GCE. All biases and variance reductions relative to second-stage estimates.

Labor Force Estimate	*AK Theoretical Multiplicative Bias*	*AK Theoretical Variance Ratio*	*GCE Multiplicative Bias*	*GCE Theoretical Variance Ratio*
Employed	0.997	0.713	0.999	0.711
Unemployed	0.989	0.962	1.000	0.959

Based on formulas [4] and [6], relative to the AK estimator, the optimized GCE achieved reductions in monthly level bias while preserving the over-the-month change variance. For total employed persons, the bias is reduced by approximately two-thirds, and for total unemployed persons, the bias is effectively eliminated. Figures 4 and 5 display the composite drift longitudinally, using the same bias decomposition[18] applied in Figures 2 and 3.

[18] The bias decomposition from McIllece (2020) uses an "adjusted" second-stage estimate, which accounts for MIS pair response imbalance and bias-smoothing models, as the unbiased quantity. As a result, the composite drift in Figures 2 – 5 is not guaranteed to be zero at the origin. This relatively small decomposition effect does not affect the results of Tables 5 – 9.

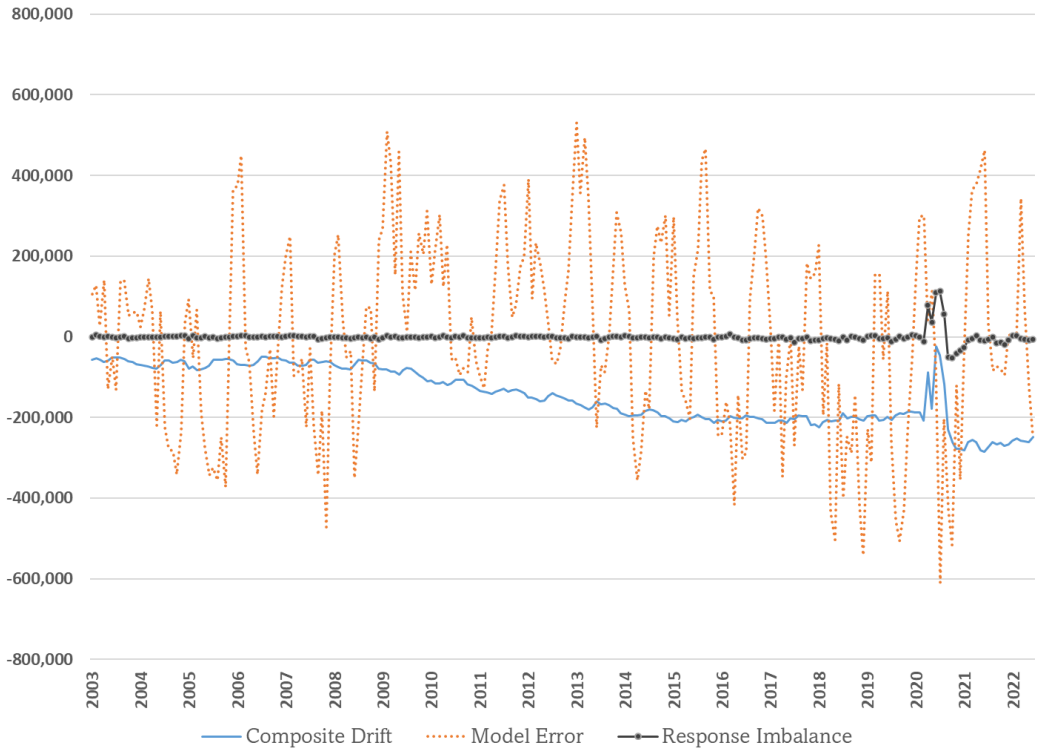

Figure 4: Bias decomposition of (not seasonally adjusted) optimized GCE estimates of total employed, January 2003 – June 2022.

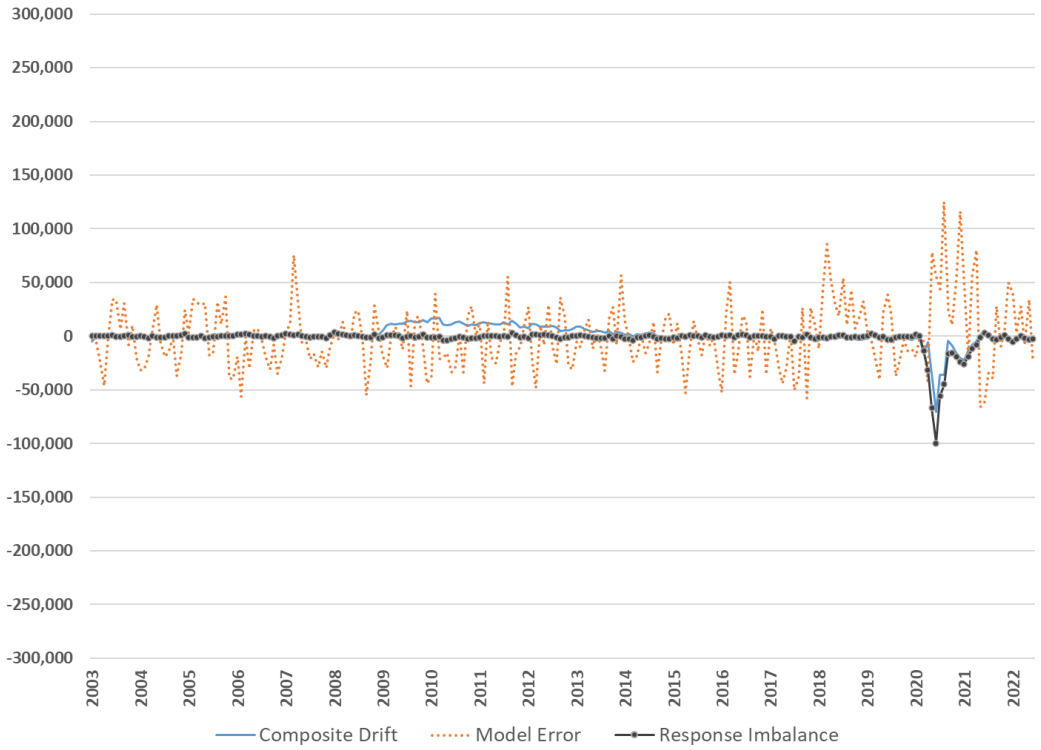

Figure 5: Bias decomposition of (not seasonally adjusted) optimized GCE estimates of total unemployed, January 2003 – June 2022.

In June 2022, the GCE optimization reduced the magnitude of the total employed bias by about 300,000 and essentially eliminated the bias for the estimate of total unemployed persons. For both labor force series, the over-the-month precision was maintained or even slightly improved, indicating that allowing the additional parameters of the GCE can improve CPS estimates at the topside level.

4. Conclusion

Regarding CPS composite estimation, there are always tradeoffs.

The AK estimator, its current coefficients implemented in the 1990s, has consistently reduced variances of over-the-month change at the expense of inducing a negative bias for the major labor force categories of total employed and total unemployed persons. Conversely, since the not in labor force estimate is computed as the difference between an independent population control—the civilian noninstitutional population, 16 years and older—and the sum of employed and unemployed, the not in labor force estimate has been positively biased over that same timeframe.

Structurally, the GCE, of which the AK is a special case, is less rigid, allowing for optimizations that reduce the bias of the major labor force groups relative to AK estimation while preserving over-the-month change variances (Table 9). However, with its many free coefficients, GCE parameterization tends to be more variable such that multiple "optimal" solutions, depending on the researchers' assumptions and frameworks, may produce similar bias and variance results but with considerable volatility in the parameterization vectors.

There is also hereditary value in preserving the properties of a time series, particularly considering that the CPS does not revise past estimates outside of seasonal adjustment, and avoiding breaks in series is imperative for critical labor force series such as those studied here. However, it is worth adding that the biases themselves could lead to a disruption of the series should the composite estimator eventually need to be reinitialized. In that situation—which could result from a sample design change, weighting modifications, a lapse in budgetary appropriations, etc.—it would be desirable if the composite series is unburdened by excessive bias relative to the second-stage estimator.

These summary thoughts conclude with several recommendations to be observed before changes are made to official CPS composite estimation procedures:

Any changes to compositing coefficients must not exacerbate the negative bias for the total employed and total unemployed series and should, ideally, reduce them. For example, it is well known in the literature that increasing K tends to reduce variance at the expense of increasing bias, and this should be avoided.

Modifications to CPS composite estimation should, at minimum, preserve the AK reductions in the over-the-month change variances of total employed and total unemployed and state-level annual averages of unemployment. This is especially important given the reality of declining CPS response rates and the national and state design requirements in §2.

Complete analysis of the impacts of modifying the CPS composite estimator should include the effects on national demographic estimates, minimally those series directly composited in the current estimation procedure.

Lastly, the impacts on seasonal adjustment, such as differences in trends or seasonal effects, should be analyzed. It is important that a reparameterized composite estimator not harm the ability to seasonally adjust important demographic series.

References

Bailar, B. (1975). "The Effect of Rotation Group Bias on Estimate from Panel Surveys," in *Journal of the American Statistical Association*, 70:349, pp. 23 – 30.

Breau, P. and Ernst, L. (1983). "Alternative Estimators to the Current Composite Estimator," in *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp. 397 – 402.

Cheng, Y., Shao, J., and Yu, Z. (2017). "Optimal AK Composite Estimators in Current Population Survey," in *Statistical Theory and Related Fields*, 1:2, pp. 257 – 264.

Cunningham, E. (2018). "Great Recession, Great Recovery? Trends from the Current Population Survey," in *Monthly Labor Review*, April 2018. U.S. Bureau of Labor Statistics.

Current Population Survey Technical Paper 77, Design and Methodology (2019). *https://www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf*. U.S. Census Bureau.

Erkens, G. (2012). "Changes in Panel Bias in the U.S. Current Population Survey and its Effects on Labor Force Estimates," in *Proceedings of the 2012 Joint Statistical Meetings, Section on Survey Research Methods*, pp. 4220 – 4232.

Erkens, G. (2017). "Practical and Theoretical Considerations of Panel Effects for the Current Population Survey's Composite Estimator," in *Proceedings of the 2017 Joint Statistical Meetings, Section on Government Statistics*, pp. 1449 – 1467.

Fuller, W. (1990). "Analysis of Repeated Surveys," in *Survey Methodology*, 16:2, pp. 167 – 180.

Huang, E. and Ernst, L. (1981). "Comparison of an Alternative Estimator to the Current Composite Estimator in CPS," in *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp. 303 – 308.

Lent, J., Miller, S., and Cantwell, P. (1994). "Composite Weights for the Current Population Survey," in *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp. 867 – 872.

Lent, J., Miller, S., Cantwell, P., and Duff, M. (1999). "Effects of Composite Weights on Some Estimate from the Current Population Survey," in *Journal of Official Statistics*, 15:3, pp. 431 – 448.

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

McIllece, J. (2020). "Covid-19 and the Current Population Survey: Response Rates and Estimation Bias," in *Proceedings of the 2020 Joint Statistical Meetings, Section on Survey Methods Research*.

Methodology for the United States Population Estimates: Vintage 2021 (2021). *https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020-2021/methods-statement-v2021.pdf*. U.S. Census Bureau Population Estimates Program.

Statistical Programs & Standards (2022). *https://www.whitehouse.gov/omb/information-regulatory-affairs/statistical-programs-standards/*. The White House Office of Management and Budget.

The Employment Situation. *https://www.bls.gov/bls/news-release/empsit.htm*. Bureau of Labor Statistics. Issued monthly.

The OPTMODEL Procedure, Procedures in Online Documentation (2022). *https://support.sas.com/rnd/app/or/procedures/optmodel.html*. SAS Institute Inc.