Bonnéry, Daniel; Cheng, Yang; Lahiri, Partha

Article

An evaluation of design-based properties of different composite estimators

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Bonnéry, Daniel; Cheng, Yang; Lahiri, Partha (2020) : An evaluation of design-based properties of different composite estimators, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 21, Iss. 4, pp. 166-190, https://doi.org/10.21307/stattrans-2020-037

This Version is available at:
https://hdl.handle.net/10419/236787

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 166–190, DOI 10.21307/stattrans-2020-037
Received – 31.01.2020; accepted – 30.06.2020

An evaluation of design-based properties of different composite estimators

Daniel Bonnéry[1], Yang Cheng[2], Partha Lahiri[3]

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

ABSTRACT

For the last several decades, the US Census Bureau has been applying AK composite estimation method for estimating monthly levels and month-to-month changes in unemployment using data from the Current Population Survey (CPS), which uses a rotating panel design. For each rotation group, survey-weighted totals, known as month-in-sample estimates, are derived each month to estimate population totals. Denoting the vector of month-in-sample estimates by Y and the design-based variance-covariance matrix of Y by Σ, one can obtain a class of AK estimators as linear combinations of Y, where the coefficients of a linear combination in the class are functions of the two coefficients A and K. The coefficients A and K were optimized by the Census Bureau under rather strong assumptions on Σ such as the stationarity of Σ over a decade. We devise an evaluation study in order to compare the AK estimator with a number of rival estimators. To this end, we construct three different synthetic populations that resemble the Current Population Survey (CPS) data. To draw samples from these synthetic populations, we consider a simplified sample design that mimics the CPS sample design with the same rotation pattern. Since the number of possible samples that can be drawn from each synthetic population is not large, we compute the exact Σ and the exact mean squared error of all estimators considered to facilitate comparison. To generate the first set of rival estimators, we consider certain subclasses of the broader class of linear combinations of month-in-sample estimates. For each subclass, when Σ is known, the optimum estimator is obtained as a function of Σ. An estimated optimal estimator in each subclass is then obtained from the corresponding optimal estimator when Σ is replaced by an estimator. Neither the AK estimator nor the estimated optimal estimators for these subclasses performed well in our evaluation study. In our real life data analysis, the AK estimates are constantly below the survey-weighted estimates, indicating potential bias. Our study indicates limitations of the approach that generate an estimated optimal estimator by first obtaining the optimal estimator in a class of linear combination of Y and then substituting in the optimal estimator an estimate of Σ.

Any attempt to improve on the estimated optimal estimator in any given class would require a thorough investigation of the highly non-trivial problem of estimation

[1]University of Cambridge. UK. E-mail: dbb31@cam.ac.uk.
ORCID: https://orcid.org/0000-0001-8582-7856.
[2]US Census Bureau. USA. E-mail: yang.cheng@verizon.net.
[3]JPSM, University of Maryland. USA. E-mail: plahiri@umd.edu.
ORCID: https://orcid.org/0000-0002-7103-545X.

of Σ for a complex setting like the CPS. We have not discussed this problem in this paper. Instead, we adapted the regression composite estimator used by Statistics Canada in the CPS setting. Unlike the estimated optimal estimators, the regression composite estimator does not require estimation of Σ and is less sensitive to the rotation group bias in our simulation study. Our study indicates that there is a great potential for regression composite estimation technique in improving estimation of both levels and month-to-month changes in the unemployment rates.

Key words: calibration, estimated controls, longitudinal survey, labor force statistics.

1. Introduction

In repeated surveys, including rotating panel surveys, statistical data integration plays an important role in producing efficient estimators by extracting relevant information over time. To this end, various composite estimators have been proposed; see Jones (1980), Yansaneh and Fuller (1998), Bell (2001), Singh et al. (2001), Fuller and Rao (2001) and others. Such composite estimators typically improve on the standard direct survey-weighted estimators in terms of mean squared error (MSE) and are commonly used by different government agencies for producing official labor force statistics. For example, to produce national employment and unemployment levels and rates, the U.S. Census Bureau uses the AK composite estimation technique developed using the ideas given in Gurney and Daly (1965).

Motivated by a Statistics Canada application, Singh and Merkouris (1995) introduced an ingenious idea for generating a composite estimator that can be computed using Statistics Canada's existing software for computing generalized regression estimates. The key idea in Singh and Merkouris (1995) is to create a proxy (auxiliary) variable that uses information at the individual level as well as estimates at the population level from both previous and current periods. Using this proxy variable, Singh and Merkouris (1995) obtained a composite estimator, referred to as Modified Regression 1 estimator (MR1) in the literature. However, Singh et al. (1997) noted that MR1 does not perform well in estimating changes in labor force statistics, which motivated them to propose a different composite estimator, called MR2, using a new proxy variable. Singh et al. (2001) generalized the idea of MR1 and MR2 estimators by suggesting a general set of proxy variables.

Fuller and Rao (2001) noted that the regression composite estimator proposed by Singh et al. (1997) is subject to an undesirable drift problem, i.e., it may produce estimates that drift away from the real value suggested by the underlying model as time progresses. They proposed an alternative regression composite method to rectify the drift problem. Their method differs from the method of Singh et al. (2001) in two directions. First, the idea of rectifying the drift problem by a weighted combination of the two proxy variables used for MR1 and MR2 is new. Secondly, their final regression composite estimator involves estimation of the weight assigned to MR1 or MR2 control variable in the weighted combination — this idea was not discussed in Singh et al. (2001). In short, the Fuller-Rao regression composite estimator with estimated weight cannot be viewed as a special case of Singh et al. (2001) and vice versa.

Gambino et al. (2001) conducted an empirical study to evaluate the Fuller-Rao regression composite estimator, offered missing value treatment and listed several advantages (e.g. weighting procedure, consistency, efficiency gain, etc.) of the Fuller-Rao regression composite estimator over the AK estimator. Statistics Canada now uses the Fuller-Rao method for their official labor force statistics production. Salonen (2007) conducted an empirical study to compare the currently used Finnish labor force estimator with the Fuller-Rao's regression composite and other estimators. Bell (2001) applied the generalized regression technique to improve on the Best Linear Unbiased Estimator (BLUE) based on a fixed window of time points and compared his estimator with the AK composite estimator of Gurney and Daly (1965) and with the modified regression estimator of Singh et al. (1997), using data from the Australian Labour Force Survey. Beaumont and Bocci (2005) proposed a regression composite estimator with missing covariates defined using variables of interest from the previous month.

The main goal of this paper is to compare the design-based properties of the AK estimator with different rival estimators using the CPS data. To this end, we first expand the list of potential estimators by considering two new classes of composite estimators. The first class includes the AK estimator as a member. The second class generalizes the class of estimators considered earlier by Yansaneh and Fuller (1998) to incorporate multiple categories of employment status (e.g., employed, unemployed, and not in the labor force). We obtain the best linear unbiased estimator (BLUE) for each class of estimators. We call them the best AK estimator and multivariate BLUE, respectively. As special cases of the multivariate BLUE, one can generate the univariate BLUE and the best AK estimator. If the covariance matrix between two vectors of observations corresponding to any two different variables is a null matrix, then multivariate BLUE is identical to the univariate BLUE when the design matrix is the same for the variables. However, in general they are not identical when we do not have a block-diagonal covariance structure as is the case in our problem.

The optimal estimator for a given class of estimators, derived under given model and optimality condition, cannot be used as it involves unknown model parameters (e.g., variances and covariances). The AK estimator used by the Census Bureau is obtained from the optimal estimator when variances and covariances are substituted by estimators justified under a rather strong stationary assumption. We devise an evaluation study in order to assess the exact design-based properties of different composite estimators using the CPS data and CPS sample design. We demonstrate that the optimal estimator for a given model with estimated variances and covariances can perform poorly even when the modeling assumptions are valid. We included the multivariate BLUE with estimated variances and covariances for completeness of this research. While the multivariate BLUE performs the best under the model that generates it, it performed worse than the univariate BLUE with estimated variances and covariances. Overall, we found that the Fuller-Rao estimator performed the best among all composite estimators considered in our study.

In Section 2, we discuss the population and sample design. In Section 3, we review different classes of estimators and the optimal estimator within each class. In Section 4, we describe our evaluation study to assess the design-based properties of different

estimators. In Section 5, we report the CPS data analysis. Some discussion and future research topics are given in Section 6. We defer the proofs of relevant results and description of CPS design to the Appendix. To facilitate reading of the paper, we list all the notation used in the paper in the appendix.

2. Notations

2.1. Population

Our theoretical framework uses three indices to identify three dimensions: m for month, k for individual and e for an employment status category. In this paper, we will consider three categories of employment status: employed, unemployed and not in the labor force. The theory and methods developed in this paper, however, extend to more than 3 categories of employment status. Consider a sequence of finite populations of individuals $(U_m)_{m\in\{1...M\}}$, where U_m refers to the finite population for month m. Let N denote the cardinality of $U=\bigcup_{m=1}^{M}U_m$. Let $\mathbf{y}_{m,k,e}=1$ if the kth individual belongs to U_m and has eth employment status and $\mathbf{y}_{m,k,e}=0$ otherwise, $m\in\{1,\cdots,M\}$, $k\in\{1,\cdots,N\}$, $e\in\{1,2,3\}$. Because of our three dimensional data structure, we find it convenient to introduce arrays in developing our methodology and theory. Let $\mathbf{y}=[\mathbf{y}_{m,k,e}]_{m\in\{1,...,M\},k\in\{1,...,N\},e\in\{1,2,3\}}$ denote a three dimensional $(M,N,3)$-sized array. We also define $\mathbf{x}$ as a 3-dimensional array of auxiliary variables indexed by month, individual and auxiliary variable, and an array $\mathbf{z}$, indexed the same way, which contains endogenous variables in the sense that $\mathbf{z}$ is a function of $\mathbf{x}$ and $\mathbf{y}$. Any element of an array with (m,k)-index satisfying $k\notin U_m$ is equal to 0 by convention.

2.2. Notational conventions on arrays

Given subsets A, B, C of $\{1,\ldots,M\}$, $\{1,\ldots,N\}$, $\{1,2,3\}$, respectively (including the full set), we use the following notation for sub-arrays: $\mathbf{y}_{A,B,C}=[\mathbf{y}_{a,b,c}]_{a\in A,b\in B,c\in C}$, and may replace A, B, or C by "." when $A=\{1,\ldots,M\}$, $B=\{1,\ldots,N\}$ or $C=\{1,2,3\}$, respectively: for example, $\mathbf{y}=\mathbf{y}_{.,.,.}$. Let $\mathbf{t}_{\mathbf{y}}=\left[\sum_{k\in U}\mathbf{y}_{m,k,e}\right]_{m\in\{1,...,M\},e\in\{1,2,3\}}$ be the two dimensional $(M,3)$-sized array of population totals indexed by month m and employment status e. We now show we can form a vector or matrix from an array. For a p-dimensional $(a_1,\ldots,a_p)$-sized array A, define $\vec{A}$ as the vector $\left(\vec{A}_1,\ldots,\vec{A}_{\prod_{l=1}^{p}a_l}\right)$, where $\forall(i_1,\ldots,i_p)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $\vec{A}_{1+\sum_{l=1}^{p}\left[\prod_{l'<l}(a_{l'}-1)i_1\right]}=A_{i_1,\ldots,i_p}$, with the convention that a product over the empty set equals 1. By convention, when an array B is defined as an $((a_1,\ldots,a_p),(b_1,\ldots,b_q))$-sized array (with two vector of indexes), $\vec{A}$ is the matrix $\left[\vec{A}_{i,j}\right]_{i\in\{1,\ldots,\prod_{l=1}^{p}a_l\},j\in\{1,\ldots,\prod_{l=1}^{q}b_l\}}$ such that $\forall(i_1,\ldots,i_p)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $(j_1,\ldots,j_q)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $\vec{A}_{1+\sum_{l=1}^{p}\left[(i_l-1)\prod_{l'<l}(a_{l'})\right],1+\sum_{l=1}^{q}\left[(j_l-1)\prod_{l'<l}(b_{l'})\right]}=A_{(i_1,\ldots,i_p),(j_1,\ldots,j_p)}$. Given A an $((a_1,\ldots,a_n),(b_1,\ldots b_l))$ array and B a $((b_1,\ldots,b_l),(c_1,\ldots c_p))$ array, $C=A\times B$ is the $((a_1,\ldots,a_n),(c_1,\ldots c_p))$ array defined by $C_{(i_1,\ldots,i_n),(k_1,\ldots,k_n)}=\sum_{j_1,\ldots,j_l}A_{(i_1,\ldots,i_n),(j_1,\ldots,j_l)}B_{(j_1,\ldots,j_l),(k_1,\ldots,k_n)}$.

2.3. The sample design

The CPS monthly sample comprises about 72,000 housing units and is collected for 729 areas (Primary Sampling Units) consisting of more than 1,000 counties covering every state and the District of Columbia. The CPS, conducted by the Census Bureau, uses a 4-8-4 rotating panel design. For any given month, the CPS sample can be grouped into eight subsamples corresponding to the eight rotation groups. All the units belonging to a particular rotating panel enter and leave the sample at the same time. A given rotating panel (or group) stays in the sample for four consecutive months, leaves the sample for the eight succeeding months, and then returns for another four consecutive months. It is then dropped from the sample completely and is replaced by a group of nearby households. Of the two new rotation groups that are sampled each month, one is completely new (their first appearance in the panel) and the other is a returning group, which has been out of the sample for eight months. Thus, in the CPS design, six out of the eight rotation groups are common between two consecutive months (i.e., 75% overlap), and four out of eight are common between the same month of two consecutive years (i.e., 50% overlap) respectively; see Hansen et al. (1955). For month m, let S_m denote the sample of respondents. Let $S_{m,g}$ denote the set of sampled respondents in the gth sample rotation group for month m and $S_m = \bigcup_{g=1}^{8} S_{m,g}$. For a given month m, the rotation groups $S_{m,g}$, $g = 1,\ldots,8$ are indexed so that g indicates the number of times that rotation group $S_{m,g}$ has been a part of the sample in month m and before. In the US Census Bureau terminology, g is referred to as the month-in-sample (mis) index and $S_{m,g}$ as the month-in-sample g rotation group (more details on this design are given in Section 4.3). We adopt a design-based approach in this study in which variables $\mathbf{x}$ and $\mathbf{y}$ are considered fixed parameters of the underlying fixed population model for design-based inference (Cassel et al., 1977, p. 2).

3. Estimation

3.1. Direct and month-in-sample estimators

Let $\mathbf{w}_{m,k}$ denote the second stage weight of individual k in month m, obtained from the basic weight (that is, the reciprocal of the inclusion probability) after standard non-response and post-stratification adjustments. By convention, $\mathbf{w}_{m,k} = 0$ if $k \notin S_m$. Let $\mathbf{w}$ be the (M,N)-sized array indexed by m and k of $\mathbf{w}_{m,k}$. We refer to CPS Technical Paper (2006) for a detailed account of weight construction. The array of direct survey-weighted estimator of $\mathrm{t}_{\mathbf{y}}$ is given by $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}} = \left[\sum_{k\in S_m} \mathbf{w}_{m,k}\mathbf{y}_{m,k,e}\right]_{m\in\{1,\ldots,M\},e\in\{1,2,3\}}$. Define the $(M,8,3)$-sized array of month-in-sample estimates: $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}} = \left[8\times\sum_{k\in S_{m,g}} \mathbf{w}_{m,k}\mathbf{y}_{m,k,e}\right]_{m\in\{1,\ldots,M\},g\in\{1,\ldots,8\},e\in\{1,2,3\}}$. For a month-in-sample number g, $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{.,g,.}$ is called the month-in-sample g estimator of $\mathrm{t}_{\mathbf{y}}$.

3.2. An extended Bailar model for the rotation group bias

Because of differential non-response and measurement errors across different rotation groups, the direct and month-in-sample estimators are subject to a bias, commonly referred to as the rotation group bias. Bailar (1975) proposed a class of semi-parametric models on the expected values of the month-in-sample estimators. Under a model in this class, (i) the bias of each month-in-sample estimator of total of unemployed depends on the month-in-sample index g only, (ii) the bias is invariant with time, and (iii) the vector of month-in-sample biases are bounded by a known linear constraint (without this binding linear constraint, month-in-sample rotation group biases could only be estimated up to an additive constant). Note that these very strong assumptions were made in order to reveal the existence of what is known as the rotation group bias in US Census Bureau terminology. It would be highly questionable to use this model for rotation group bias correction because (i) the choice of the linear constraint would be totally arbitrary in the absence of a re-interview experiment and (ii) the stationarity assumptions are unreasonable. We propose the following model in order to extend the Bailar model to account for the rotation group biases of the multiple categories:

$$\mathrm{E}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m,g,e}\right] = (\mathrm{t}_{\mathbf{y}})_{m,e} + b_{g,e}, \tag{1}$$

where b is a two-dimensional $(8,p)$-sized array of biases such that $\forall e, C_e b_{.,e} = 0$, C_1, C_2, C_3 being known linear forms satisfying $C_e\,(1,\ldots,1)^{\mathrm{T}} \neq 0$.

3.3. Estimation of unemployment rate and variance approximation

We define the function $\mathrm{R} : (0,+\infty)^3 \to [0,1], x \mapsto x_2/(x_1+x_2)$. By convention, when applied to an array with employment status as an index, x_1, x_2 denote the subarrays for employment status 1 and 2, respectively, and / denotes the term by term division. The unemployment rate vector is defined as $\mathrm{r} = \mathrm{R}\,(\mathrm{t}_{\mathbf{y}}) = (\mathrm{t}_{\mathbf{y}})_{.,1} / \left((\mathrm{t}_{\mathbf{y}})_{.,1} + (\mathrm{t}_{\mathbf{y}})_{.,2}\right)$.
Given an estimator $\hat{\mathrm{t}}_{\mathbf{y}}^{\star}$ of $\mathrm{t}_{\mathbf{y}}$, we derive the following estimator of r from $\hat{\mathrm{t}}_{\mathbf{y}}^{\star}$: $\hat{\mathrm{r}}^{\star} = \mathrm{R}(\hat{\mathrm{t}}_{\mathbf{y}}^{\star})$. Using the linearization technique, we can approximate the variance $\mathrm{Var}[\hat{\mathrm{r}}_m^{\star}]$ of the unemployment rate estimator for month m by $J_1 \mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m,.}\right] J_1{}^{\mathrm{T}}$, where J_1 is the Jacobian matrix: $J_1 = \left(\frac{\mathrm{d}\ \mathrm{R}(t)}{\mathrm{d}\ t}\right)\left((\mathrm{t}_{\mathbf{y}})_{m,.}^{\star}\right) = \left[(\mathrm{t}_{\mathbf{y}})_{m,1}^{-1}, -(\mathrm{t}_{\mathbf{y}})_{m,1}\,(\mathrm{t}_{\mathbf{y}})_{m,2}^{-2}, 0\right]$, and the variance of the estimator of change of the employment rate between two consecutive months by $J_2 \mathrm{Var}\left[\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m,.}, \left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m-1,.}\right)\right] J_2{}^{\mathrm{T}}$, where

$$\begin{aligned} J_2 &= \left(\frac{\mathrm{d}\ \mathrm{R}(t) - \mathrm{R}(t')}{\mathrm{d}\ (t,t')}\left((\mathrm{t}_{\mathbf{y}})_{m,.}, (\mathrm{t}_{\mathbf{y}})_{m-1,.}\right)\right) \\ &= \left[(\mathrm{t}_{\mathbf{y}})_{m,1}^{-1}, -(\mathrm{t}_{\mathbf{y}})_{m,1}\,(\mathrm{t}_{\mathbf{y}})_{m,2}^{-2}, 0, -(\mathrm{t}_{\mathbf{y}})_{m-1,1}^{-1}, (\mathrm{t}_{\mathbf{y}})_{m-1,1}\left((\mathrm{t}_{\mathbf{y}})_{m-1,2}\right)^{-2}, 0\right]. \end{aligned}$$

3.4. The class of linear combinations of month-in-sample estimators

Here, as in Yansaneh and Fuller (1998), we consider the best estimator of counts by employment status in the class of linear combinations of month-in-sample estimators. Generalizing Yansaneh and Fuller (1998), the unbiasedness assumption of all month-in-sample estimators is:

$$\mathrm{E}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = \vec{X}\vec{\mathrm{t}}_{\mathbf{y}}, \tag{2}$$

where X is the $((M,8,3),(M,3))$-sized array with rows indexed by the triplet (m,g,e) and columns indexed by the couple (m,e) such that $X_{(m,g,e),(m',e')} = 1$ if $m' = m$ and $e' = e$, 0 otherwise. Let L be a $(p,(M,3))$-sized array with $p \in \mathbb{N}\setminus\{0\}$ and rows indexed by (m,e). By class of linear estimators of $L\mathrm{t}_{\mathbf{y}}$, we will designate the class of estimators that are linear combinations of the month-in-sample estimators, i.e., of the form $W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$, where W is a fixed (does not depend on the observations) $(p,(M\times 8\times 3))$-sized matrix.

Best linear estimator

Let $\Sigma_{\mathbf{y}} = \mathrm{Var}_{\mathbf{y}}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right]$. In the design-based approach, $\Sigma_{\mathbf{y}}$ is a function of the finite population $\mathbf{y}$. The variance of a linear transformation $W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$ of $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}$ is: $\mathrm{Var}\left[W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = W^T\Sigma_{\mathbf{y}}W$. When month-in-sample estimators are unbiased, $\Sigma_{\mathbf{y}}$ is known, and only $\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$ is observed, and $\vec{X}^{+}\vec{X} = I$, the Gauss-Markov theorem states that the BLUE of $\mathrm{t}_{\mathbf{y}}$ uniformly in t_y is the $(M,3)$-sized matrix $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{BLUE}}$ defined by

$$\vec{X}^{+}(\vec{X}\vec{X}^{+})\left(I-\Sigma_{\mathbf{y}}((I-\vec{X}\vec{X}^{+})^{+}\Sigma_{\mathbf{y}}(I-\vec{X}\vec{X}^{+}))^{+}\right)\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}, \tag{3}$$

where the $^{+}$ operator denotes the Moore-Penrose pseudo inversion, I is the identity matrix. Here the minimization is with respect to the order on the space of symmetric positive definite matrices: $M_1 \leq M_2 \Leftrightarrow M_2 - M_1$ is positive. It can be shown that $\vec{X}^{+} = \vec{X}^{\mathrm{T}}/8$ in our case and that $\vec{X}^{+}\vec{X} = I$. For more details about the Gauss-Markov result under singular linear model, one may refer to (Searle, 1994, p. 140, Eq. 3b). This is a generalization of the result of Yansaneh and Fuller (1998), as it takes into account the multi-dimensions of $\mathbf{y}$ and non-invertibility of $\Sigma_{\mathbf{y}}$. Note that $\Sigma_{\mathbf{y}}$ can be non-invertible, especially when the sample is calibrated to a given fixed population size, considered non-random, because of an affine relationship between month-in-sample estimates (e.g., $\sum_{g=1}^{8}\sum_{e=1}^{3}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m,g,e}$ is not random).

We recall the following:

(i) For any linear transformation L applicable to $\vec{\mathrm{t}}_{\mathbf{y}}$, the best linear unbiased estimator of $L\vec{\mathrm{t}}_{\mathbf{y}}$ uniformly in $\mathrm{t}_{\mathbf{y}}$ is $L\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{BLUE}}$, which ensures that the BLUE of month-to-month change can be simply obtained from the BLUE of level. Thus, there is no need for searching a compromise between estimation of level and change.

(ii) For any linear transformation L applicable to $\vec{\mathrm{t}}_{\mathbf{y}}$, any linear transformation J applicable to $L\vec{\mathrm{t}}_{\mathbf{y}}$, $L\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{BLUE}} \in \mathrm{argmin}\left\{JW\Sigma_{\mathbf{y}}(JW)^{\mathrm{T}}\middle| W, W\vec{X} = L\right\}$.Thus, the plug-in esti-

mators for unemployment rate and month-to-month unemployment rate change derived from the BLUE are also optimal in the sense that they minimize the linearized approximation of the variance of such plug-in estimators, which can be written in the form $JW\Sigma_{\mathbf{y}}(JW)^{\mathrm{T}}$.

Remark: BLUE under Bailar rotation bias model

Here, we give the expression of the BLUE under the general Bailar rotation bias model. Bailar's rotation bias model can be written in the following matrix notation:

$$\mathrm{E}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = \vec{X}\vec{\mathrm{t}}_{\mathbf{y}} + \vec{X}'\vec{b}, \tag{4}$$

where X' is a fixed known array; see also Yansaneh and Fuller (1998, equation 8). For example under Model (1), with $C_1 = C_2 = C_3 = (1,\ldots,1)$, X' is the $((M,8,3),(7,2))$-sized array such that for $m \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, $g' \in \{1,\ldots,7\}, e \in \{1,2,3\}$, $e' \in \{1,2,3\}$, $X'_{(m,g,e),(g',e')} = 1$ if $g = g' < 8$ and $e = e'$, -1 if $g = 8$ and $e = e'$, 0 otherwise. We can reparametrize Model (4) as $\mathrm{E}[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}] = X^{\star}\mu$, where $X^{\star} = [\vec{X} \mid \vec{X}']$, and the parameter $\mu = [\vec{\mathrm{t}}_{\mathbf{y}} \mid \vec{b}]^{\mathrm{T}}$. The best linear unbiased estimator of $\vec{\mathrm{t}}_{\mathbf{y}}$ under this rotation bias model is given by

$$LX^{\star+}(X^{\star}X^{\star+})\left(I - \Sigma_{\mathbf{y}}(I - X^{\star}X^{\star+})^{+}\Sigma_{\mathbf{y}}(I - X^{\star}X^{\star+}))^{+}\right)\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}},$$

with L satisfying $LX^{\star} = \vec{X}$. This is a generalization of Yansaneh and Fuller (1998) because it (i) considers non-invertible $\Sigma_{\mathbf{y}}$, (ii) does not limit to a unidimensional variable and (iii) is generalized to general Bailar's model.

3.5. AK composite estimation

Definition

We define a general class of AK composite estimators. Let $A = \mathrm{diag}(a_1, a_2, a_3)$ and $K = \mathrm{diag}(k_1, k_2, k_3)$ denote two diagonal matrices of dimension 3. The AK estimator with coefficients A and K is defined as follows: first define $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{1,.} = \left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}}\right)_{1,.}$, then recursively define for $m \in 2,\ldots,M$,

$$\begin{aligned}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.} &= (I-K)\times\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}}\right)_{m,.} \\ &+ K\times\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m-1,.} + \frac{4}{3}\sum_{k\in S_m\cap S_{m-1}}\left(\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.} - \mathbf{w}_{m-1,k,.}\mathbf{y}_{m-1,k,.}\right)\right) \\ &+ A\times\left(\sum_{k\in S_m\setminus S_{m-1}}\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.} - \frac{1}{3}\sum_{k\in S_m\cap S_{m-1}}\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.}\right),\end{aligned} \tag{5}$$

where $\setminus$ denotes the set difference operator and I is the identity matrix of dimension 3. The sum of the first two terms of the AK estimator is indeed a weighted average of the

current month direct estimator and the previous month AK estimator suitably updated for the change. The last term of the AK estimator is correlated to the previous terms and has an expectation 0 with respect to the sample design. Gurney and Daly (1965) explained the benefits of adding the third term in reducing the mean squared error. The Census Bureau uses specific values of A and K, which were empirically determined in order to arrive at a compromise solution that worked reasonably well for both employment level and rate estimation; see, e.g., Lent et al. (1999). The corresponding unemployment rate estimator is obtained as: $\hat{\mathrm{r}}_m^{\mathrm{AK}} = \mathrm{R}\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.}\right)$. Note that $\hat{\mathrm{r}}_m^{\mathrm{AK}}$ depends on a_1, a_2, k_1, k_2, but not on a_3 and k_3. Note that the class of AK estimators is a sub class of the class of linear estimators, as the AK estimator can be written as a linear combination of the month-in-sample estimators: $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.} = \sum_{m'=1}^{m}\sum_{g=1}^{8} c_{m,m',g}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m',g,.}$, where the $(3,3)$ matrices $c_{m,m,g}$ are defined recursively: $\forall g \in \{1,\ldots,8\}, c_{1,1,g} = (1/8)\times I$ and

$$\forall m \in \{2,\ldots,M\}, \begin{cases} \forall g \in \{1,5\} & c_{m,m,g} & = ((I-K)+A)/8 \\ \forall g \in \{2,3,4,6,7,8\} & c_{m,m,g} & = ((I-K)+4K/3-A/3)/8 \\ \forall g \in \{1,2,3,5,6,7\} & c_{m,m-1,g} & = c_{m-1,m-1,g}\times K-(4K/3)/8 \\ \forall g \in \{4,8\} & c_{m,m-1,g} & = c_{m-1,m-1,g}\times K \\ \forall 1 \le m' < m-1 & c_{m,m',g} & = c_{m-1,m',g}\times K \end{cases} \tag{6}$$

$\forall m' > m, g \in \{1,\ldots,8\}, c_{m,m',g} = 0.$

Let W^{AK} be the $((M,3),(M,8,3))$ array such that for $m,m' \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, $e,e' \in \{1,2,3\}$, $W^{\mathrm{AK}}_{(m,e),(m',g,e')} = c_{m,m',g}$ if $e = e'$, 0 otherwise. Then $\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{AK}} = \vec{W}^{\mathrm{AK}}\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$.

Notes on AK estimator

In presence of rotation bias, the bias of the AK estimator is $\vec{W}^{\mathrm{AK}}\vec{X}'\vec{b}$, which may not be equal to 0. Depending on the rotation bias model, an unbiased version of the AK estimator may not exist. Furthermore, contrary to the BLUE, the best A and K coefficients for estimation of one particular month and status may not be optimal for another month and status. Moreover, the best A and K coefficients for estimation of level may not be optimal for estimation of change. For example, it is possible to find A,K,m,e,A',K',m',e' such that $\mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,e}\right] < \mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{A}',\mathrm{K}'}\right)_{m,e}\right]$ and $\mathrm{Var}\left[\hat{\mathrm{t}}_{\mathbf{y}_{m',e'}}^{\mathrm{AK}}\right] > \mathrm{Var}\left[\hat{\mathrm{t}}_{\mathbf{y}_{m',e'}}^{\mathrm{A}',\mathrm{K}'}\right]$.

When $\Sigma_{\mathbf{y}}$ is known, let $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,level}$ and $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,change}$ denote the AK estimators obtained by minimizing (with respect to A and K) the average approximated variance of level estimators $\sum_{m=1}^{M} J_1 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{m,.}\right] J_1{}^{\mathrm{T}}$ and of change estimators $\sum_{m=1}^{M} J_2 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{\{m-1,m\},.}\right] J_2{}^{\mathrm{T}}$, respectively; let $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,compromise}$ denote the AK estimator obtained by minimizing the averaged variance

$\sum_{m=1}^{M}\left(J_1 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{m,.}\right] J_1{}^{\mathrm{T}} + J_2 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{\{m-1,m\},.}\right] J_2{}^{\mathrm{T}}\right)$. For AK estimation, note that the three objective functions are polynomial functions of A and K whose coefficients are functions of $\Sigma_{\mathbf{y}}$. By using a standard numerical method (Nelder-Mead) we can obtain the optimal coefficients.

3.6. Empirical best linear estimator and empirical best AK estimator.

Let $\hat{\Sigma}$ be an estimator of $\Sigma_{\mathbf{y}}$, and let $\hat{\mathrm{t}}_{\mathbf{y}}^{EBLUE}$ be the estimator of $\mathrm{t}_{\mathbf{y}}$ obtained from (3) when $\Sigma_{\mathbf{y}}$ is replaced by $\hat{\Sigma}$. In the same manner, we can define the empirical best AK estimators for change, level and compromise. For the CPS, optimal A and K coefficients were determined so that a compromise objective function, accounting for the variances of the month-to-month change and level estimates, would be minimum. The variances were estimated under the assumption of a stationary covariance of month-in-sample estimators; see Lent and Cantwell (1996). The method used in the Census Bureau consists in choosing the best coefficients a_1, a_2, k_1, k_2 on a grid with 9 possible values for each coefficient $(0.1,\ldots,0.9)$.

3.7. Regression Composite Estimation

In this section we elaborate on the general definition of the class of regression composite estimators parametrized by a real number $\alpha \in [0,1]$ as proposed by Fuller and Rao (2001). This class includes regression composite estimators MR1 (for $\alpha = 0$) and MR2 (for $\alpha = 1$) as defined by Singh and Merkouris (1995) and Singh et al. (2001). For $\alpha \in [0,1]$, the regression composite estimator of $\mathrm{t}_{\mathbf{y}}$ is a calibration estimator $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,.}$ defined as follows: provide calibration totals $\left(\mathrm{t}_{\mathbf{x}}^{adj}\right)_{m,.}$ for the auxiliary variables (they can be equal to the true totals when known or estimated), then define $\left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{1,.} = \left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{direct}}\right)_{1,.}$, and $\mathbf{w}_{1,k}^{\mathrm{r.c.},\alpha} = \mathbf{w}_{1,k}$ if $k \in S_1$, 0 otherwise. For $m \in \{2,\ldots,M\}$, recursively define

$$\mathbf{z}_{m,k,.}^{\mathrm{r.c.}(\alpha)} = \begin{cases} \alpha\left(\tau_m^{-1}\left(\mathbf{z}_{m-1,k,.} - \mathbf{z}_{m,k,.}\right) + \mathbf{z}_{m,k,.}\right) + (1-\alpha)\ \mathbf{z}_{m-1,k,.} & \text{if } k \in S_m \cap S_{m-1}, \\ \alpha\ \mathbf{z}_{m,k,.} + (1-\alpha)\ \left(\sum_{k\in S_{m-1}} \mathbf{w}_{m-1,k}^{\mathrm{r.c.},\alpha}\right)^{-1}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{c}}\right)_{m-1,.} & \text{if } k \in S_m \setminus S_{m-1}, \end{cases} \tag{7}$$

where $\tau_m = \left(\sum_{k\in S_m\cap S_{m-1}} \mathbf{w}_{m,k}\right)^{-1} \sum_{k\in S_m} \mathbf{w}_{m,k}$. Then the regression composite estimator of $(\mathrm{t}_{\mathbf{y}})_{m,.}$ is given by $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,.} = \sum_{k\in S_m} \mathbf{w}_{m,k}^{\mathrm{r.c.},\alpha} \mathbf{y}_{m,k}$, where

$$\left(\mathbf{w}_{m,.}^{\mathrm{r.c.},\alpha}\right) = \mathrm{argmin}\left\{ \sum_{k\in U} \frac{\left(\mathbf{w}_k^{\star} - \mathbf{w}_{m,k}\right)^2}{\mathbb{1}(k \notin S_m) + \mathbf{w}_{m,k}} \,\middle|\, \mathbf{w}^{\star} \in \mathbb{R}^U,\ \begin{array}{l} \sum_{k\in S_m} \mathbf{w}_k^{\star} \mathbf{z}_{m,k,.}^{\mathrm{r.c.}(\alpha)} = \left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{m-1,.} \\ \sum_{k\in S_m} \mathbf{w}_k^{\star} \mathbf{x}_{m,k,.} = \left(\mathrm{t}_{\mathbf{x}}^{adj}\right)_{m,.} \end{array} \right\}, \tag{8}$$

and $\left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{m,.} = \sum_{k\in S_m} \mathbf{w}_{m,k}^{\mathrm{r.c.},\alpha} \mathbf{z}_{m,k}^{\mathrm{r.c.}(\alpha)}$, where $\mathbb{1}(k \notin S_m) = 1$ if $k \notin S_m$ and 0 otherwise. Our definition of regression composite estimator is more general than the one in Fuller and Rao (2001) as it takes into account a multivariate version of $\mathbf{y}$. Modified Regression 3 (MR3) of Gambino et al. (2001) does not belong to the class of regression composite

estimators. The MR3 estimator imposes too many constraints in the calibration procedure, which leads to a high variability of the calibration weights; consequently, MR3 estimator has a larger MSE than composite regression estimators.

Choice of z and choice of α

Fuller and Rao (2001) studied the properties of the estimator $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,1}$ for the choice of $\mathbf{z} = \mathbf{y}_{.,1}$. As the employment rate is a function of $\mathbf{y}_{m,1}$ and $\mathbf{y}_{m,2}$, we investigate the properties of Regression Composite Estimator with the choice $\mathbf{z} = \mathbf{y}$. Fuller and Rao (2001) proposed a method that allows for an approximation to the optimal α coefficient for month-to-month change and level estimation, under a specific individual level superpopulation model for continuous variables. They proposed this superpopulation model to explain the drift problem of MR2 (regression composite estimator for $\alpha = 1$) and obtained the best coefficient α. Since we deal with a discrete multidimensional variable, the continuous superpopulation model assumed by Fuller and Rao (2001) is not appropriate in our situation. It will be interesting to propose an approach to estimate the best α in our situation. For our preliminary study, we examine a range of known α values in our simulations and in the CPS data analysis.

4. Simulation Experiment

4.1. Description of Simulation Study

We conducted a simulation study to enhance our understanding of the finite sample properties of different composite estimators. We generated three synthetic finite populations, each with size 100,000. In order to make the simulation experiment meaningful, we generated employment statuses for each finite population in a manner that attempts to capture the actual U.S. national employment rate dynamics during the study period 2005-2012. Moreover, in order to understand the maximum gain from the composite estimation, we induced high correlation in the employment statuses between two consecutive months subject to a constraint on the global employment rate evolution. We set the probability of month-to-month changes in employment statuses for an individual to zero in case of no change in the corresponding direct national employment rates. Samples were selected according to a rotating design with systematic selection that mimics the CPS sample design. Since the number of possible samples is only 1000, we are able to compute the exact design-based bias, variance and mean squared error of different estimators, and subsequently, the optimal linear and optimal AK estimators. We compute employment rate, total employed, and total unemployed over the 85-month period using the direct, AK and the Fuller-Rao regression composite methods. We then compare the optimal estimator in the class of regression composite estimators to those in the class of the AK and best linear estimators. Note that the simulation study can be reproduced using the R package we created for this purpose; see Bonnéry (2016c).

4.2. Populations generation

We created three synthetic populations each with $N = 100,000$ individuals indexed by $1,\ldots,N$. For individual k of each population, we created a time series $(\mathbf{y}_{m,k,.})_{m\in 1,\ldots,M}$, where $\mathbf{y}_{m,k,.} \in \{(1,0,0),(0,1,0),(0,0,1)\}$ (for unemployed, not in labor force, employed), and $M = 85$. Each individual belongs to one household consisting of 5 individuals. The number of all households is $H = 20,000$; the set of all households is given by $\{h_i = \{(5\times(i-1)+1),\ldots,(5\times i)\} \mid i = 1,\ldots,H\}$. We created time series data under certain constraints at the population level. For each population, unemployment rates are the same as the direct estimates obtained from the CPS data. In population 1, the proportion of people who change status between two consecutive months is minimal. In populations 2 and 3, the proportions of persons who change from one status to another between two consecutive months are equal to those estimated from the CPS data. In population 2, a person with a small index has a higher probability to change status. In population 3, the probability to change status between two consecutive months is the same for all individuals with the same status.

4.3. Repeated design

We mimic the CPS design, which is described in appendix A. For month m, a sample S_m is the union of 8 rotation groups. The design and the creation of rotation groups are explained below. Rotation groups are made of $n = 20$ households with a total of 100 individuals. So for each month m, there are $\#(S_m) = 800$ individuals in the sample, and the inclusion probability of any unit is $1/125$. The selection of longitudinal sample $S_1,\ldots S_m$ is made in 3 steps:

1. Draw an integer number η between 1 and 1,000 from a uniform distribution.
2. For $\ell \in 1,\ldots,(M+15)$, create a cluster of households $\mathrm{Clu}_\ell = \bigcup_{j=1}^{n} h_{i_{\ell,j}}$, where $i_{\ell,j} = \mathrm{rem}\left((r-1+\ell-1)+\frac{H}{n}\times(j-1),H\right)+1$, and $\mathrm{rem}(a,b)$ denotes the remainder of the Euclidean division of a by b.
3. Let $\delta_1 = 0, \delta_2 = 1,\ \delta_3 = 2,\ \delta_4 = 3,\ \delta_5 = 12,\ \delta_6 = 13,\ \delta_7 = 14,\ \delta_8 = 15$. For $m \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, create the samples $S_{m,g} = \mathrm{Clu}_{m+\delta_g}$, and $S_m = \bigcup_{g=1}^{8} S_{m,g}$.

We can compute exact design-based moments by drawing all the 1000 possible samples under our sample design. For example, for $\eta = 506,\ m = 12,\ g = 3$, we have $S_{m,g} = \mathrm{Clu}_{12+\delta_3} = \mathrm{Clu}_{14}$, and $\mathrm{Clu}_{14} = \{h_{\mathrm{rem}((506-1+14-1)+\frac{20000}{20}\times(k-1),20000)+1} \mid k = 1\ldots 20\} = \{h_{19},h_{1019},h_{2019},h_{3019},\ldots,h_{19019}\}$. Table 1 displays the rotation chart for our simulation, which is identical to the CPS rotation chart (CPS Technical Paper, 2006, Figure 3-1).

4.4. Rotation bias

In each sample, we introduced a measurement error by changing employment status of 20% of employed individuals in month-in-sample group 1 from employed to unemployed, which leads to an overestimation of the unemployment rate.

Table 1: The CPS Rotation chart

	Clu$_1$	Clu$_2$	Clu$_3$	Clu$_4$	Clu$_5$	Clu$_6$	Clu$_7$	Clu$_8$	Clu$_9$	Clu$_{10}$	Clu$_{11}$	Clu$_{12}$	Clu$_{13}$	Clu$_{14}$	Clu$_{15}$	Clu$_{16}$	Clu$_{17}$	Clu$_{18}$	Clu$_{19}$	Clu$_{20}$
Jan 05	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$									$S_{1,5}$	$S_{1,6}$	$S_{1,7}$	$S_{1,8}$				
Feb 05		$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$									$S_{2,5}$	$S_{2,6}$	$S_{2,7}$	$S_{2,8}$			
Mar 05			$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$									$S_{3,5}$	$S_{3,6}$	$S_{3,7}$	$S_{3,8}$		
Apr 05				$S_{4,1}$	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$									$S_{4,5}$	$S_{4,6}$	$S_{4,7}$	$S_{4,8}$	
May 05					$S_{5,1}$	$S_{5,2}$	$S_{5,3}$	$S_{5,4}$									$S_{5,5}$	$S_{5,6}$	$S_{5,7}$	$S_{5,8}$
Jun 05						$S_{6,1}$	$S_{6,2}$	$S_{6,3}$	$S_{6,4}$									$S_{6,5}$	$S_{6,6}$	$S_{6,7}$
Jul 05							$S_{7,1}$	$S_{7,2}$	$S_{7,3}$	$S_{7,4}$									$S_{7,5}$	$S_{7,6}$
Aug 05								$S_{8,1}$	$S_{8,2}$	$S_{8,3}$	$S_{8,4}$									$S_{8,5}$
Sep 05									$S_{9,1}$	$S_{9,2}$	$S_{9,3}$	$S_{9,4}$								
Oct 05										$S_{10,1}$	$S_{10,2}$	$S_{10,3}$	$S_{10,4}$							
Nov 05											$S_{11,1}$	$S_{11,2}$	$S_{11,3}$	$S_{11,4}$						
Dec 05												$S_{12,1}$	$S_{12,2}$	$S_{12,3}$	$S_{12,4}$					
Jan 06													$S_{13,1}$	$S_{13,2}$	$S_{13,3}$	$S_{13,4}$				
Feb 06														$S_{14,1}$	$S_{14,2}$	$S_{14,3}$	$S_{14,4}$			
Mar 06															$S_{15,1}$	$S_{15,2}$	$S_{15,3}$	$S_{15,4}$		
Apr 06																$S_{16,1}$	$S_{16,2}$	$S_{16,3}$	$S_{16,4}$	
May 06																	$S_{17,1}$	$S_{17,2}$	$S_{17,3}$	$S_{17,4}$
Jun 06																		$S_{18,1}$	$S_{18,2}$	$S_{18,3}$
Jul 06																			$S_{19,1}$	$S_{19,2}$
Aug 06																				$S_{20,1}$

Source: CPS Technical Paper (2006, Figure 3-1)

4.5. Variance on month-in-sample estimators computation

As we draw all the possible samples, we are able to compute the exact variance of any estimator. Moreover, we are able to compute the true $\Sigma_{\mathbf{y}}$, which yields both the optimal best linear and AK estimators.

4.6. Estimation of $\Sigma_{\mathbf{y}}$

Define

$$\sigma^2_{m,m'} = \frac{\sum_{i=1}^{H} \left(\sum_{k\in h_i} \mathbf{y}_{m,k,.} - \frac{\sum_{i'=1}^{H} \sum_{k'\in h_{i'}} \mathbf{y}_{m,k',.}}{H} \right) \left(\sum_{k\in h_i} \mathbf{y}_{m',k,.} \right)^{\mathrm{T}}}{H-1}.$$

We estimate $\sigma^2_{m,m'}$ by

$$\hat{\sigma}^2_{m,m'} = \frac{\sum\limits_{i\in\{1,\ldots,H\} | h_i \subset S_m \cap S_{m'}} \left(\sum\limits_{k\in h} \mathbf{y}_{m,k,.} - \frac{\sum_{i=1}^{H} \sum_{k\in h_i} \mathbf{y}_{m',k,.}}{\#\{i\in\{1,\ldots,H\} | h_i \subset S_m \cap S_{m'}\}} \right) \left(\sum\limits_{k\in h} \mathbf{y}_{m',k,.} \right)^{\mathrm{T}}}{\#\{i \in \{1,\ldots,H\} \mid h_i \subset S_m \cap S_{m'}\} - 1}$$

if $S_m \cap S_{m'} \neq \emptyset$, 0 otherwise. Let $m, m' \in \{1,\ldots,M\}$, $g, g' \in \{1,\ldots,8\}$. If $m' + \delta_{g'} = m + \delta_g$ then $S_{m,g} = S_{m',g'}$ and we approximate the distribution of $S_{m',g'}$ by a cluster sampling, where the first stage is simple random sampling. We estimate $\mathrm{Cov}\left[\hat{\mathrm{t}}^{\mathrm{mis},g}_m, \hat{\mathrm{t}}^{\mathrm{mis},g}_{m'}\right]$ by $\widehat{\mathrm{Cov}}\left[\hat{\mathrm{t}}^{\mathrm{mis},g}_{\mathbf{y}_{m,e}}, \hat{\mathrm{t}}^{\mathrm{mis},g'}_{\mathbf{y}_{m',e'}}\right] = (H)^2 \left(1 - \frac{n}{H}\right) \frac{\hat{\sigma}^2_{m,m'}}{n/8}$. If $m' + \delta_{g'} \neq m + \delta_g$, then $S_{m,g} \cap S_{m',g'} = \emptyset$ and we approximate the distribution of $(S_{m,g}, S_{m',g'})$ by the distribution of two independent simple random samples of clusters conditional to non-overlap of the two samples.

Table 2: Optimal (a_1,k_1) and (a_2,k_2) values for the three synthetic populations

	Population 1	Population 2	Population 3
(a_1,k_1) (unemployed)			
Level	$(0.0471,0.85)$	$(0.0395,0.398)$	$(-0.0704,-0.619)$
Compromise	$(0.029,0.895)$	$(0.00175,0.0551)$	$(0.0038,0.0253)$
Change	$(0.0243,0.89)$	$(0.0358,0.362)$	$(-0.0239,-0.445)$
(a_2,k_2) (employed)			
Level	$(0.0714,0.752)$	$(0.0453,0.73)$	$(-0.0354,0.825)$
Compromise	$(-0.0075,-0.232)$	$(0.002,0.0598)$	$(0.0464,0.0482)$
Change	$(-0.0187,-0.256)$	$(0.0658,0.723)$	$(-0.0529,0.836)$

We estimate $\mathrm{Cov}\left[\hat{t}^{\mathrm{mis}}_{m,g,.},\hat{t}^{\mathrm{mis}}_{m',g',.}\right]$ by $\widehat{\mathrm{Cov}}\left[\hat{\mathrm{t}}^{\mathrm{mis}}_{\mathbf{y}m,g,.},\hat{\mathrm{t}}^{\mathrm{mis}}_{\mathbf{y}m',g',.}\right] = -H\hat{\sigma}^2_{m,m'}$.

4.7. Choice of optimal estimator in each class

In our simulations, the best linear unbiased estimator turned out to be exact in the sense that for the three different choices of $\mathbf{y}$ (population 1, population 2, population 3), the $(1000,2040)$-matrix Y whose rows are the 1000 probable values of $\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}}$ is of rank 1000, so for all (m,e), we can find a 2040-sized vector $x_{m,e}$ such that $Yx_{m,e} = (\mathrm{t}_{\mathbf{y}})_{m,e}.\mathbb{1}$, where $\mathbb{1}$ is 1000-sized vector of ones. Then, we define W_o as the $((M\times 8\times 3),(M\times 3))$-sized array whose rows are the vectors $x_{m,g}$ such that $W_0\, Y^{\mathrm{T}} = \vec{\mathrm{t}}_{\mathbf{y}}$. This surely implies $W_o\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}} = \vec{\mathrm{t}}_{\mathbf{y}}$, and hence the BLUE is necessarily equal to $W_o\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}}$, a result that we were able to reproduce in our simulations. This situation is particular to our simulation setup, which allows a small number of possible samples, but with a design for which the number of probable samples is larger than the number of month-in-sample estimates, the best linear unbiased estimator would likely have a strictly positive variance. We computed the objective functions for $\alpha \in \{0,0.05,\ldots,1\}$ only. Table 2 shows the optimal values for a_1, k_1, a_2, and k_2 for the three different populations and the best empirical estimator for level, change and compromise. The Census Bureau uses the coefficients $a_1 = 0.3$, $k_1 = 0.4$, $a_2 = 0.4$ and $k_2 = 0.7$ for the CPS. We notice that for each population, the best set of coefficients for change, level and compromise is very close, which means that the optimal choice for level is also almost optimal for change for those three populations. Table 3 shows the best coefficient α for the regression composite estimators.

4.8. Analysis without measurement error

Figure 1 displays the relative mean squared errors of different estimators of unemployment level and change over time: $\left(\frac{\mathrm{MSE}[\hat{\mathrm{r}}^{\star}_m]}{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\mathrm{direct}}_m\right]}\right)_{m\in\{1,\ldots,M\}}$, and $\left(\frac{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\star}_m-\hat{\mathrm{r}}^{\star}_{m-1}\right]}{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\mathrm{direct}}_m-\hat{\mathrm{r}}^{\mathrm{direct}}_{m-1}\right]}\right)_{m\in\{2,\ldots,M\}}$, for $\star\in\{\mathrm{direct},\mathrm{AK},\mathrm{r.c.}\}$. In this figure, the best representative in each class is chosen in the sense that the coefficients of Tables 2 and 3 are used.

Table 3: Optimal regression composite estimator's α parameter value for three synthetic populations

	Population 1	Population 2	Population 3
Level	0.55 (0.6)	0.45 (0.6)	0
Change	1	0.75	0.8
Compromise	0.55 (0.6)	0.45 (0.6)	0

Numbers in the parentheses indicate parameter values in presence of rotation with bias when different

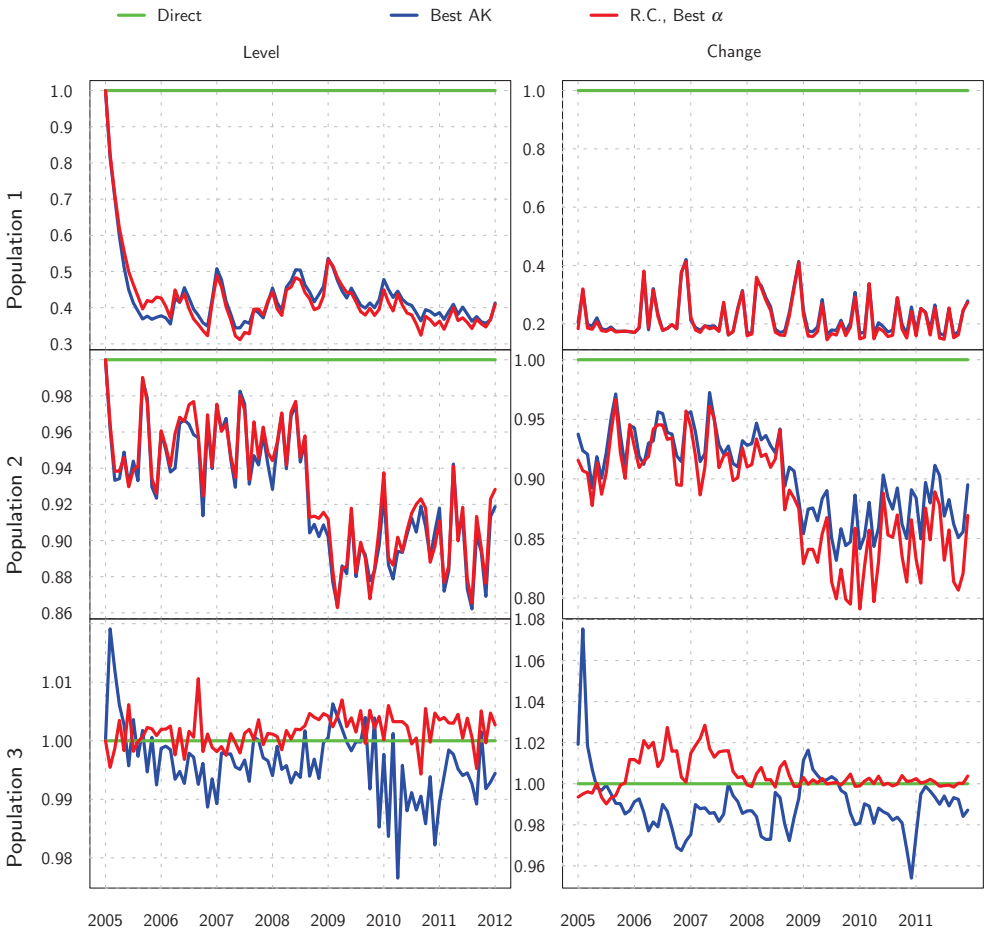
Table 4: For three synthetic populations, quantiles and means (over months) of the relative mean squared errors of unemployment level estimators

	Population 1					Population 2					Population 3				
	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.
0%	0.318	1	1	0.322	0.477	0.87	1	1	0.863	0.885	0.983	1	1	0.994	0.994
25%	0.377	1.52	2.59	0.38	0.546	0.906	1.35	2.56	0.913	0.94	0.996	1.08	1.03	1	1.01
50%	0.409	1.6	2.64	0.42	0.591	0.929	1.41	2.7	0.945	0.974	0.997	1.14	1.04	1	1.02
75%	0.454	1.95	2.74	0.472	0.663	0.951	1.49	2.79	0.969	0.989	1	1.26	1.07	1	1.02
100%	1	2.09	2.86	1	1	1	1.68	3.08	1	1.02	1.01	1.65	1.14	1.01	1.15
Mean	0.431	1.72	2.64	0.443	0.613	0.926	1.42	2.66	0.94	0.966	0.997	1.19	1.05	1	1.02

Note that in the absence of measurement error, the performances of all best "estimators" are comparable.

When trying to estimate the best A and K, the results differ. For different synthetic populations, Table 4 and 5 report the quantiles of relative mean squared errors of the best AK estimator, the empirical best AK estimator, the AK estimator with coefficient taken arbitrarily equal to the CPS AK coefficients (Arb. AK column), the best regression composite estimator (r.c.column) and the Regression Composite estimator with α taken arbitrarily equal to 0.75 (Arb. AK column) for the level and change estimation, respectively. For all three synthetic populations, both the estimated best AK estimator and arbitrary AK estimator perform worse than the direct estimator. Moreover, the arbitrary regression composite estimator seems to behave much better than the estimated best AK estimator and arbitrary AK estimators. We observe (not reported here) that the estimated best linear estimator performs worse than the estimated best AK estimator. This underlines the weakness of the AK and Yansaneh-Fuller type estimators: without a good estimator of the variance-covariance matrix, they perform very poorly. We note that the regression composite estimator with arbitrary α performs better without requiring any estimation of the variance.

Figure 1: Relative mean squared errors of different estimated series of unemployment level and of month-to-month changes



4.9. Analysis with measurement error

Under (2), a solution to the rotation group bias for adapting the AK estimator consists in estimating the rotation bias parameter vector b and then applying AK coefficients to corrected month-in-sample estimates, to obtain $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}*}\right)_{m,.} = \sum_{m'=1}^{m}\sum_{m'=1}^{m}\left(c_{m,m',g}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis},g}\right)_{m,g,.} - \hat{b}_g\right)$. The question of how to adapt the regression composite estimator to take into account measurement error is more complicated. Besides, the model used for rotation bias is itself questionable. The linear constraint on b ($\sum b_{g,.} = 0$ or $b_{1,.} = 0$) is imposed to address an identifiability problem, but one cannot assess its validity. As a result we think it is not a good way to deal with the rotation bias. We have not investigated how to adapt the regression composite estimator to address the problem of rotation bias. Instead we studied its behaviour in presence of rotation bias. To this end, we systematically (for all months, all samples) changed the status of up to 2 unemployed persons of month-in-sample group 1 from unemployed to employed. For different populations, Tables 6 and 7 display quantiles and means of the relative mean squared errors of the best AK estimator and the best regression composite estimator for both level and change. We applied the best AK and best regression composite estimators to the cases without measurement error and with measurement error. We notice that AK estimator is very sensitive to rotation bias, whereas regression composite estimator is not. A reason may be that introducing a variable not correlated to the study variables in the calibration procedure does not much change the estimation of the study variable. Rotation bias weakens the correlation between $\mathbf{z}$ and $\mathbf{y}$, and yet the performance of the regression composite estimator is comparable to the performance of the direct.

Table 5: Quantiles and means (over months) of the relative mean squared errors for different populations and unemployment month-to-month change estimators

	Population 1					Population 2					Population 3				
	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.
0%	0.0959	2.77	5.43	0.0279	0.0936	0.845	0.872	2.72	0.774	0.791	0.973	0.998	1.01	0.984	0.994
25%	0.123	3.31	6.35	0.0455	0.112	0.887	0.953	3.07	0.835	0.847	0.99	1.02	1.03	0.992	1
50%	0.142	3.68	6.64	0.0552	0.127	0.914	0.998	3.33	0.885	0.89	0.993	1.02	1.04	0.997	1
75%	0.215	5.21	6.93	0.146	0.201	0.932	1.03	3.62	0.916	0.919	0.996	1.03	1.06	1	1
100%	0.395	6.12	7.59	0.355	0.383	0.971	1.13	3.92	0.965	0.967	1.04	1.06	1.14	1.11	1.01
Mean	0.174	4.21	6.68	0.102	0.163	0.909	0.993	3.33	0.876	0.883	0.993	1.03	1.04	1	1

Table 6: Quantiles and means (over months) of the relative mean squared errors of unemployment level estimators for different populations.

	Population 1		Population 2		Population 3		Pop. 1 (bias)		Pop. 2 (bias)		Pop. 3 (bias)	
	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.
0%	0.318	0.322	0.87	0.863	0.983	0.994	1	0.0521	1	0.117	0.919	0.158
25%	0.377	0.38	0.906	0.913	0.996	1	46.1	0.0735	2.78	0.155	1.5	0.739
50%	0.409	0.42	0.929	0.945	0.997	1	47.9	0.0949	2.81	0.179	1.59	0.768
75%	0.454	0.472	0.951	0.969	1	1	52.5	0.115	2.86	0.254	1.86	0.786
100%	1	1	1	1	1.01	1.01	57.6	0.162	2.92	0.3	2.18	0.843
Mean	0.431	0.443	0.926	0.94	0.997	1	45.6	0.0957	2.77	0.203	1.64	0.754

5. The CPS Data Analysis

5.1. Implementation of regression composite estimator for the CPS

5.1.1 Choice of α

Under a simple unit level times series model with auto-regression coefficient ρ, Fuller and Rao (2001) proposed a formal expression for an approximately optimal α as a function of ρ and studied the so-called drift problem for the MR2 choice: $\alpha = 1$. They also proposed approximate expressions for variances of their estimators for the level and change. For various reasons, it seems difficult to obtain the optimal or even an approximately optimal α needed for the Fuller-Rao type regression composite estimation technique to produce the U.S. employment and unemployment rates using the CPS data. First of all, the simple time series model used by Fuller and Rao (2001) is not suitable to model a nominal variable (employment status) with several categories. Secondly, the complexity of the CPS design poses a challenging modeling problem. Before attempting to obtain the optimal or even an approximately optimal choice of α required for the Fuller-Rao type regression composite method, it will be instructive to evaluate regression composite estimators for different known choices of α. This is the focus of this section.

5.1.2 Choice of $\mathbf{x}$ and $\mathbf{z}$

In our study, we considered two options for $\mathbf{z}$: (i) $\mathbf{z} = \mathbf{y}$, (ii) a more detailed employment status variable with 8 categories. As the use of this more detailed variable reduces

Table 7: Quantile and means (over months) of the relative mean squared errors of unemployment month-to-month change estimators for different populations.

	Population 1		Population 2		Population 3		Pop. 1 (bias)		Pop. 2 (bias)		Pop. 3 (bias)	
	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.
0%	0.0959	0.0936	0.845	0.791	0.973	0.994	0.422	0.0298	0.898	0.457	1.19	0.935
25%	0.123	0.112	0.887	0.847	0.99	1	0.477	0.0385	0.938	0.552	1.48	0.994
50%	0.142	0.127	0.914	0.89	0.993	1	0.515	0.05	0.954	0.583	1.5	1.01
75%	0.215	0.201	0.932	0.919	0.996	1	0.563	0.093	0.971	0.613	1.52	1.02
100%	0.395	0.383	0.971	0.967	1.04	1.01	3.96	0.209	1.25	0.673	1.58	1.05
Mean	0.174	0.163	0.909	0.883	0.993	1	0.665	0.0671	0.958	0.581	1.48	1

the degrees of freedom in the calibration procedure and leads to estimates with a higher mean squared error, we report results for option (i) only. For an application of the Fuller-Rao method, one might think of including all the variables that have already been used for the weight adjustments in the $\mathbf{x}$ variables. However, this would introduce many constraints on the coefficients and thus is likely to cause a high variability in the ratio of $\mathbf{w}_{m,k}$ and $\mathbf{w}^{\text{r.c.}}_{m,k}$. The other extreme option is not to use any of the auxiliary variables, but then the final weights would not be adjusted for the known totals of auxiliary variables $\mathbf{x}$. As a compromise, we selected only two variables: gender and race.

5.2. Results

Figure 2(a) displays the difference $\widehat{\text{r}}^{\text{AK}}_m - \widehat{\text{r}}^{\text{direct}}_m$ between different composite estimates and the corresponding direct estimates against months m. For the regression composite estimator, we considered three choices: (i) $\alpha = 0.75$ (suggested by Fuller and Rao), (ii) $\alpha = 0$ (corresponding to MR1), and (iii) $\alpha = 1$ (corresponding to MR2). We display similar graphs for month-to-month change estimates in Figure 2(b). Notice that $\alpha = 0$ and $\alpha = 1$ correspond to MR1 and MR2, respectively. We display similar graphs for month-to-month change estimates in Figure 2.

It is interesting to note that the AK composite estimates of unemployment rates are always lower than the corresponding direct estimates in Figure 2(a). To our knowledge, this behavior of AK composite estimates has not been noticed earlier. In contrast, the regression composite estimates MR1 are always higher than the corresponding direct estimates. However, such deviations decrease as α gets closer to 1 as shown in Figure 2(a). Application of the Fuller-Rao method at the household level causes an increase in the distance between the original and calibrated weights and one may expect an increase in the variances of the estimates. Figure 2(b) does not indicate systematic deviations of the composite estimates of the month-to-month changes from the corresponding direct estimates. Deviations of the regression composite estimates from the corresponding direct estimates seem to decrease as α approaches 1.

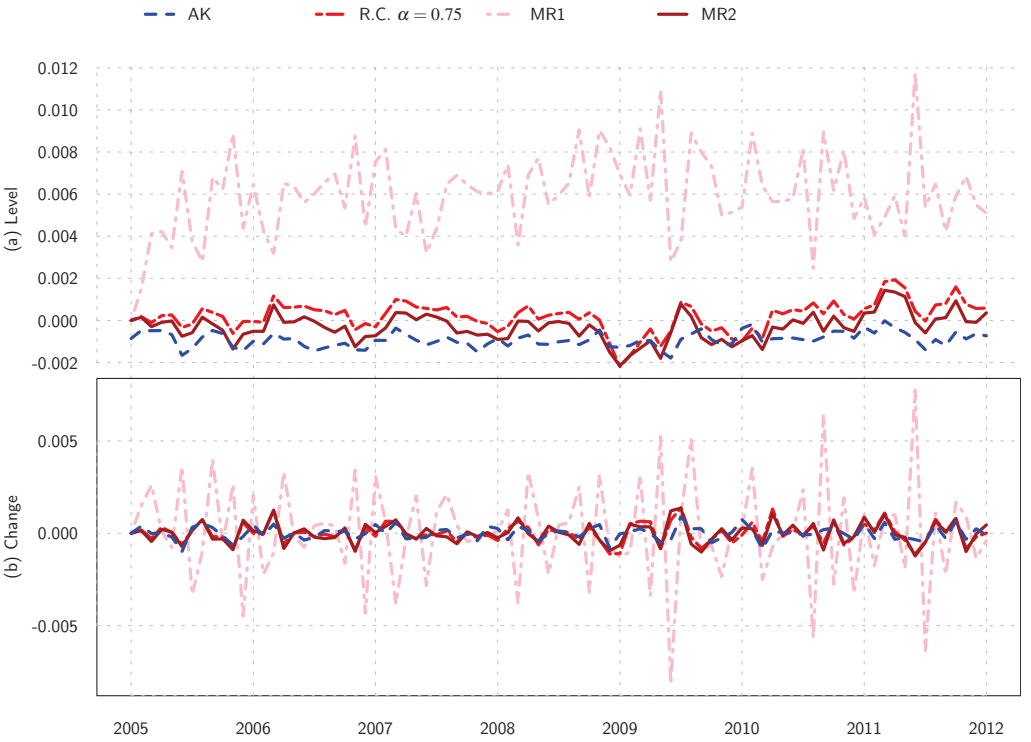
6. Discussion

Our study reveals that there is ample scope for improving the AK estimator used by the Census Bureau. We would like to emphasize the following undesirable features of the AK estimation method:

(i) The method used to compute optimal coefficient is crude — the best coefficients are selected from 9 different values. Our R package, based on the built in R Nelder-Mead algorithm, can provide the optimal coefficients within 8 digits of precision in a reasonable time.

(ii) The stationarity assumption on the variances and covariances of the month-in-sample estimators over a period of 10 years does not seem realistic, and to our knowledge, has not been tested before. Moreover, even if the stationary model was reasonable, the complexity of the CPS sample design makes it difficult to evaluate the quality of the estimators used for that model. The difficulty in proposing a stochastic model for the

Figure 2: Estimated series of differences between different composite estimates and the corresponding direct estimates



best linear estimators in the CPS was pointed out earlier by (Jones, 1980, Sec. 4). Our evaluation study shows that the AK estimators is very sensitive to the choices of A and K and that the errors in the estimation of the variances and covariances may lead to poor performance of the AK estimators. Moreover, estimators of variances and covariances of month-in-sample estimators affect the performances of empirical best linear unbiased estimators.

(iii) Using the Bailar model for the bias in our study, we showed that AK estimator is very sensible to rotation group bias. There is currently no satisfactory way to correct the AK estimator for the rotation bias. The Bailar model relies on an arbitrary constraint on the month-in-sample biases and a strong stationarity assumption of the month-in-sample bias and should not be used unless some re-interview study can justify the Bailar's model. One possible option would be to study the rotation bias at the individual level using resampling method. In this paper, we have not investigated how to adapt the regression composite estimator to address the problem of rotation bias. This could be a good problem for future research.

(iv) The computation of composite weights in CPS to calibrate the weights on the AK

estimators will affect all other weighted estimators. Although Lent and Cantwell (1996) showed that there was not a big effect on the estimates, considering the concerns about AK estimators listed before, we do not think that the use of those composite weights is a good option.

(v) The CPS data analysis shows that the AK estimates are consistently smaller than the corresponding direct survey-weighted estimates for the period 2005-2012. This is also a source of concern.

The composite regression estimator does not rely on an estimation of the variances and covariances matrix. In our simulation study, it appears to be less sensitive to rotation group bias, and bounces around the survey-weighted estimates when applied to the real CPS data. Our study encourages the use of the regression composite method in the US labor force estimation.

To facilitate and encourage further research on this important topic, we make the following three R packages, developed under this project, freely available: (i) the package dataCPS can be used to download CPS public data files and transform them into R data set (Bonnéry (2016b)); (ii) the package CompositeRegressionEstimation can be used to compute the AK, best AK, composite regression, linear and best linear estimators (Bonnéry (2016a)); (iii) the package pubBonneryChengLahiri2016 can be used to reproduce all computations and simulations of this paper (Bonnéry, 2016c).

Acknowledgements

We thank Editor-in-Chief Professor Wlodzimierz Okrasa and an anonymous referee for reading an earlier version of the article carefully and offering a number of constructive suggestions, which led to a significant improvement of our article. We thank Ms. Victoria Cheng for helping us proofreading the entire manuscript. The research of the first and third authors has been supported by the U.S. Census Bureau Prime Contract No: YA1323-09-CQ-0054 (Subcontract No: 41-1016588). The programs used for the simulations have been made available on the github repository Bonnéry (2016c). The first author completed the research as a postdoctoral research associate of P. Lahiri at the University of Maryland, College Park, USA.

References

BAILAR, B. A., (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70(349), pp. 23–30.

BEAUMONT, J.-F. and BOCCI, C., (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey for Change Estimates. *SSC Annual Meeting, Proceedings of the Survey Methods Section*, (June), pp. 1–6.

BELL, P., (2001). Comparison of alternative labour force survey estimators. *Survey Methodology*, 27(1), pp. 53–63.

BONNÉRY, D. B., (2016a). R package CompositeRegressionEstimation. https://github.com/DanielBonnery/CompositeRegressionEstimation.

BONNÉRY, D. B., (2016b). R package dataCPS. https://github.com/DanielBonnery/dataCPS.

BONNÉRY, D. B., (2016c). R package pubBonneryChengLahiri2016. https://github.com/DanielBonnery/pubBonneryChengLahiri2016.

CASSEL, C., SÄRNDAL, C., and WRETMAN, J., (1977). *Foundations of inference in survey sampling*.

CPS Technical Paper, (2006). Design and Methodology of the Current Population Survey. Technical Report 66, U.S. Census Bureau.

FULLER, W. A. and RAO, J. N. K., (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27(1), pp. 45–51.

GAMBINO, J., KENNEDY, B., and SINGH, M. M. P. M. M. P., (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27(1), pp. 65–74.

GURNEY, M. and DALY, J. F., (1965). A multivariate approach to estimation in periodic sample surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, volume 242, p. 257.

HANSEN, M. H., HURWITZ, W. N., NISSELSON, H., and STEINBERG, J., (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50(271), pp. 701–719.

JONES, R. G., (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2), pp. 221–226.

LENT, J. and CANTWELL, S., (1996). Effect of Composite Weights on Some Estimates from the Current Population Survey. *Proceedings the American Statistical Association, Section on Survey Research Methods*, pp. 130–139.

LENT, J., MILLER, S. M., CANTWELL, P. J., and DUFF, M., (1999). Effects of composite weights on some estimates from the current population survey. *Journal of Official Statistics-Stockholm*, 15(1), pp. 431–448.

SALONEN, R., (2007). Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Second Baltic-Nordic Conference on Survey Sampling, 2.–7. June 2007, Kuusamo, Finland*, 8(3): 503–517.

SEARLE, S., (1994). Extending some results and proofs for the singular linear model. *Linear Algebra and its Applications*, 210, pp. 139–151.

SINGH, A. C., KENNEDY, B., and WU, S., (2001). Regression composite estimation for the Canadian Labour Force Survey: evaluation and implementation. *Survey Methodology*, 27(1), pp. 33–44.

SINGH, A. C., KENNEDY, B., WU, S., and BRISEBOIS, F., (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 300–305.

SINGH, A. C. and MERKOURIS, P., (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 420–425.

YANSANEH, I. S. and FULLER, W. A., (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*, 24, pp. 31–40.

APPENDIX

A. Description of CPS design

This section uses CPS notations for rotation groups. Let U be the intersection of a given basic primary sampling unit component (BPC) and one of the frames used in CPS (see CPS Technical Paper (2006)). The BPC is a set of clusters of about four housing units, the clusters are the ultimate sampling units (USU). Let N be the number of clusters in U. The clusters in U are sorted according to geographical and demographic characteristics and then indexed by $k = 1 \ldots N$. In the sequence, we will designate a cluster by its index. Let SI_w be the adjusted within-PSU sampling interval, as defined in CPS Technical Paper (2006, p. 3-11). Let $n = \lfloor (21 \times 8 * SI_w)^{-1} N \rfloor$, where $\lfloor . \rfloor$ is the floor function. The number n is the sample size for a sample rotation group. The drawing of the USU within the PSU consists in the generation of a random number X according to the uniform law on $[0,1]$. For $i = 1 \ldots n$, $j = 1 \ldots 8$, $\ell = 85 \ldots (85+15)$, let $k_{i,j,\ell}$ denote the cluster $k_{i,j,\ell} = \lfloor (X + 8 \times (i-1) + j) \times SI_w + (\ell - 85) \rfloor$. Then, with the notations of CPS Technical Paper (2006) for $\ell = 85 \ldots 100$, $j = 1 \ldots 8$, the rotation group j of sample A_ℓ is given by

$$A_{\ell,j} = \left\{ k_{i,j,\ell} \mid i = 1 \ldots n \right\}.$$

For a given month the sample consists of 8 rotation groups. There are 120 months in a period of 10 years. For $m = 1 \ldots 120$, $j' \in \{1, \ldots, 8\}$, $\ell_{m,j'}$ and $j_{m,j'}$ are given by: $j_{m,j'} = t + j' - 1 - 8 \times \lfloor (t + j' - 2)/8 \rfloor$. If $j' \in \{1, \ldots, 4\}$, $\ell_{m,j'} = 85 + \lfloor (t + j' - 2)/8 \rfloor$. If $j' \in \{5, \ldots, 8\}$, $\ell_{m,j'} = 86 + \lfloor (t + j' - 2)/8 \rfloor$.

The sample of the mth month, counting from November 2009, is given by

$$s_m = \bigcup_{j'=1}^{8} A_{\ell_{m,j'}, j_{m,j'}}.$$

For example, June 2013 corresponds to $m = 44$, counting from November 2009. Then

$$\begin{aligned}
\ell_{m,1} &= 85 + \lfloor 43/8 \rfloor = 90 & j_{m,1} &= 44 - 8 \times \lfloor 43/8 \rfloor = 4 \\
\ell_{m,2} &= 85 + \lfloor 44/8 \rfloor = 90 & j_{m,2} &= 45 - 8 \times \lfloor 44/8 \rfloor = 5 \\
\ell_{m,3} &= 85 + \lfloor 45/8 \rfloor = 90 & j_{m,3} &= 46 - 8 \times \lfloor 45/8 \rfloor = 6 \\
\ell_{m,4} &= 85 + \lfloor 46/8 \rfloor = 90 & j_{m,4} &= 47 - 8 \times \lfloor 46/8 \rfloor = 7 \\
\ell_{m,5} &= 86 + \lfloor 47/8 \rfloor = 91 & j_{m,5} &= 48 - 8 \times \lfloor 47/8 \rfloor = 8 \\
\ell_{m,6} &= 86 + \lfloor 48/8 \rfloor = 92 & j_{m,6} &= 49 - 8 \times \lfloor 48/8 \rfloor = 1 \\
\ell_{m,7} &= 86 + \lfloor 49/8 \rfloor = 92 & j_{m,7} &= 50 - 8 \times \lfloor 49/8 \rfloor = 2 \\
\ell_{m,8} &= 86 + \lfloor 50/8 \rfloor = 92 & j_{m,8} &= 51 - 8 \times \lfloor 50/8 \rfloor = 3
\end{aligned}$$

We can check from the CPS rotation chart (CPS Technical Paper, 2006, Fig. 3-1) that the sample of June 2013 consists of the 4th, 5th, 6th, 7th rotation groups of A90, of the 8th rotation group of A91, and of the 1st, 2d and 3rd rotation groups of A92:

$$S_{\text{June 2013}} = A_{90,4} \cup A_{90,5} \cup A_{90,6} \cup A_{90,7} \cup A_{91,8} \cup A_{92,1} \cup A_{92,2} \cup A_{92,3}.$$

Index