Regression Composite Estimation: An Alternative Approach for the Current Population Survey

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

[1] JPSM and US Census Bureau Research Associate, [2] US Census Bureau, [3] JPSM
[1,3]JPSM, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742, United States, [2]4600 Silver Hill Rd, Washington, MD 20233, United States

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

ABSTRACT

The Current Population Survey (CPS), a household survey sponsored by the U.S. Bureau of Labor Statistics (BLS) and conducted by the U.S. Census Bureau, is the primary source of information on the U.S. employment and unemployment levels and rates. The Census Bureau has been using the AK estimation technique for generating employment and unemployment levels and rates for the last several decades. The development of a new composite estimation method by Fuller & Rao (2001) and its subsequent adaptation by Statistics Canada for its production of official labor force statistics encourage us to conduct additional research for evaluation of the Fuller-Rao regression composite method in estimating the U.S. employment and unemployment rates. To this end, we first adapt the Fuller-Rao regression composite estimation method before applying to the CPS data. Using CPS data for the period 2005-2012, we device a Monte Carlo simulation experiment in order to compare the proposed, survey-weighted direct and the AK estimates with the simulated true employment rates. Our study also includes CPS data analysis that compares the AK estimates with different Fuller-Rao type composite estimates.

1 INTRODUCTION

In repeated surveys, different composite estimators that borrow strength over time have been proposed. Such composite estimators typically improve on the standard survey-weighted direct estimators in terms of mean squared error (MSE) and are commonly used by different government agencies. For example, to produce national employment and unemployment levels and rates, the U.S. Census Bureau uses the AK estimator, which is a composite estimator developed using the ideas given in Gurney & Daly (1965).

Motivated from a Statistics Canada application, Singh & 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 & 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 & Merkouris (1995) obtained a composite estimator, referred to as MR1 in the literature. However, Singh & Brisebois (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 & Rao (2001) noted that the regression composite estimator proposed by Singh *et al.* (2001) is subject to a drift problem in estimating changes and suggested a linear combination of the two proxy variables given in Singh *et al.* (2001) to rectify the problem. 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 (2007a,b) conducted an empirical study to compare the currently used Finnish labor force estimator with the Fuller-Rao's regression composite and other estimators.

In Section 2, we introduce a few notations used in the paper. In Section 3, we review the regression composite estimators and discuss various challenges in the implementation of the Fuller-Rao type regression composite estimators in the CPS application. In Section 4, we discuss how we adapt the Fuller-Rao method before applying to estimate the U.S. employment and unemployment rates using the CPS data, and report employment rate estimates during the period 2005-2012 by various methods. In Section 5, we report results from a Monte Carlo simulation study.

2 NOTATIONS

Consider a sequence of finite populations of individuals $(U_m)_{m\in\{1\dots M\}}$, where U_m refers to month m. Let $\mathbf{x}_m = (\mathsf{x}_{m,k})_{k\in U_m}$, and $\mathbf{y}_m = (\mathsf{y}_{m,k})_{k\in U_m}$ be vectors of auxiliary and study variables related to population U_m, respectively ($m = 1, \cdots, M$). For $k \in U_m$, dimensions of $\mathsf{x}_{m,k}$ and $\mathsf{y}_{m,k}$ can be larger than one. For example, $\mathsf{y}_{m,k} = (\mathsf{y}_{m,k,1}, \mathsf{y}_{m,k,2}, \dots)$. A vector obtained on a subset A of U_m is denoted by $\mathsf{y}_{m,A} = (\mathsf{y}_{m,k})_{k\in A}$. Total and mean over a set are denoted by the operator signs t and μ, respectively; for example, if $A \subset U_m$, $\mathrm{t}_{\mathsf{y},m,A} = \sum_{k\in A} \mathsf{y}_{m,k}$ or $\mu_{\mathsf{x},m,h} = \mathrm{mean}\,(\mathsf{x}_{m,k})_{k\in U_m}$. For month m, let s_m denote the sample of respondents. We assume that, by construction, each sample s_m is a union of households. A vector $w_m^H = \left(w_{m,h}^H\right)_{h\subset s_m}$ of weights at the household level is obtained by modifying the initial vector of sampling weights, which equal the inverse of the probability of selection of households, so that some conditions of the type $\sum_{h\subset s_m} w_{m,h}^H \mathrm{t}_{\mathsf{x},m,h} = \mathrm{t}_{\mathbf{x},m}^{\mathrm{adj}}$ are satisfied, where x is a first set of auxiliary variables and $\mathrm{t}_{\mathbf{x},m}^{\mathrm{adj}}$ are quantities specified according to some knowledge on population totals. Individual weights, denoted $w_{m,k}^I$, are computed such that equivalent conditions are satisfied for a different set of auxiliary variables. After this second step, weights of two individuals of a same household may be different. Notice that $w_{m,h}^H$ and $w_{m,k}^I$ are functions of s_m, and are thus random. We define $\hat{\mathrm{t}}_{\mathsf{y},m,U_m}^w = \sum_{k\in s_m} w_k^I \mathsf{y}_{m,k}$, an estimator of $\mathrm{t}_{\mathsf{y},m,U_m}$.

3 REGRESSION COMPOSITE ESTIMATION

In a general setting, calibration of initial set of weights, say $(d_k)_{k\in s}$, consists in choosing weights $(w_k)_{k\in s}$ that minimizes a specified distance, say $\sum_k f_k(w_k - d_k)$, subject to $\sum_{k\in s} w_k x_k = \mathrm{t}_x^{\mathrm{adj}}$, where f_k is a given function. The calibration estimator of t_y is then given by $\sum_k w_k y_k$. If f_k is well chosen, generalized regression and calibration estimators are identical.

The regression composite estimator of $\mathrm{t}_{\mathsf{y},m,U_m}$, introduced by Fuller and Rao (2001), is the Generalized Regression estimator $\hat{\mathrm{t}}_{y,m,U_m}^c$, where the variables used for the regression are the auxiliary variables $\mathbf{x}$ and control variables denoted by $\mathbf{x}^c$. The control variables x^c are computed from x'_{m,s_m} and $\mathsf{x}'_{m-1,s_{m-1}}$, where $\mathbf{x}'$ may contain variables in $\mathbf{x}$ or $\mathbf{y}$. In Fuller & Rao (2001), $\mathbf{x}' = \mathbf{y}$. The variables $\mathbf{x}^c$ and $\hat{\mathrm{t}}_{\mathsf{x}',m-1,U_m}^c$ are defined recursively:

For $m = 1$,

$$\begin{aligned}\hat{\mathrm{t}}_{\mathsf{x}',m=1,U_1}^c &= \hat{\mathrm{t}}_{\mathsf{x}',m=1,U_1}^w,\\ \hat{\mu}_{\mathsf{x}',m=1,U_1}^c &= \sum_{k\in s_m} \left(w_{m,k}\right)^{-1} \hat{\mathrm{t}}_{\mathsf{x}',m=1,U_1}^c.\end{aligned}$$

For $m \in \{2, \dots, M\}$,

$$\begin{aligned}\mathsf{x}_{m,s_m\cap s_{m-1}}^c &= \alpha \left(\tau^{-1}(\mathsf{x}'_{m-1,s_m\cap s_{m-1}} - \mathsf{x}'_{m,s_m\cap s_{m-1}}) + \mathsf{x}'_{m-1,s_m\cap s_{m-1}}\right) + (1-\alpha)\, \mathsf{x}'_{m-1,s_m\cap s_{m-1}}\\ \mathsf{x}_{m,s_m\setminus s_{m-1}}^c &= \alpha\, \mathsf{x}'_{m,s_m\setminus s_{m-1}} + (1-\alpha)\, \hat{\mu}_{\mathsf{x}',m-1,U_m}^c\\ \hat{\mathrm{t}}_{\mathsf{x}',m=1,U_1}^c &= \sum_{k\in s_m} w_{m,k}^c \mathsf{x}_{m,k},\end{aligned}$$

where τ is a fixed number close to $\left(\sum_{k\in s_m\cap s_{m-1}} w^I_{m,k}\right)^{-1}\sum_{k\in s_m} w^I_{m,k}$, and w^c_{m,s_m} minimizes $\sum_k f_k(w^c_{m,k} - w_{m,k})$ under the constraints: $\sum_{k\in s_m} w^c_{m,k}\mathsf{x}^c_{m,k} = \hat{\mathrm{t}}^c_{\mathsf{x}',m-1,U_{m-1}}$ and $\sum_{k\in s_m} w^c_{m,k}\mathsf{x}_{m,k} = \mathrm{t}^{\text{adj}}_{\mathsf{x},m,U_m}$. Then, the estimator $\hat{\mathrm{t}}^c_{\mathsf{y},m,U_m}$ of $\mathrm{t}_{\mathsf{y},m,U_m}$ is defined as $\sum_{k\in s_m} w^c_{m,k}\mathsf{y}_{m,k}$. The estimators with $\alpha = 0$ and $\alpha = 1$ correspond to MR1 (MR stands for Modified Regression) given in Singh & Merkouris (1995) and MR2 given in Singh & Brisebois (1997), respectively.

Without appropriate modeling assumptions and criterion, it seems difficult to suggest an optimal choice for α. Under a simple unit level time series model with auto-regression coefficient ρ, Fuller and Rao (2001) proposed a formal expression for approximately optimal α as a function of ρ and studied a drift problem for the MR2 choice: $\alpha = 1$. They also proposed approximate expressions for variances of their estimators for levels and changes.

For various reasons, it seems difficult to obtain optimal or even approximately optimal α needed in the Fuller-Rao type regression composite estimation technique to produce 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. Even if for the sake of simplification, we define the employment status as an one-dimensional variable, complexity of the CPS design poses a challenging modeling problem. For these reasons, we do not attempt to obtain an optimal or even approximately optimal choice for α required in the Fuller-Rao type regression composite method. Instead, we evaluate the regression composite estimators for different known choices of α.

There is another challenging problem in implementing the Fuller-Rao regression composite method in the CPS setting. This is because the method requires a perfect linkage between individuals in consecutive samples. However, while the CPS micro-data files contain a unique household identifier, person number within a household may not be unique. The person number may be different for the same person from one month to another, even when the person lives in the same household. Two options are tested. The first option ignores the problem, and treats two observations with the same person number within the same household in two consecutive months as related to the same individual. The second option uses an aggregation of data at the household level.

4 THE CPS DATA ANALYSIS

4.1 The CPS and the AK estimator

The Current Population Survey has a 4-8-4 rotating panel design. A sample s_m consists of eight rotating panels with six of the panels belonging to s_{m+1} so that $\tau = 3/4$. The AK estimator of a total is defined recursively as a linear combination of the current and previous month estimates of totals obtained from each rotation group.

We apply the Fuller-Rao method using the weights $w^I_{m,h}$ obtained after the first series of adjustments at the household level and incorporating in the auxiliary variables $\mathbf{x}$ all variables used in the second series of adjustments at the individual level.

4.2 Choice of study and auxiliary variables

In the CPS, the employment status is a nominal variable with eight categories so that we can choose $\mathbf{y}_m = (\mathbf{status}_{m,U_m,1}, \ldots, \mathbf{status}_{m,U_m,8})$, where $\mathbf{status}_{m,U_m,j}$ is the vector of indicator variables corresponding to the jth category of the employment status in month m. We can also choose

$$\mathbf{y}_m = (\mathbf{status}_{m,U_m,1} + \mathbf{status}_{m,U_m,2},\ \mathbf{status}_{m,U_m,3} + \mathbf{status}_{m,U_m,4},\ \mathbf{status}_{m,U_m,5} + \ldots + \mathbf{status}_{m,U_m,8})$$

In both cases, we took $\mathbf{x}' = \mathbf{y}$. For an application of the Fuller-Rao method at the unit level, we would like to include all the variables that have been already 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 $w^I_{m,k}$ and $w^c_{m,k}$. The other extreme option is not to use any of these auxiliary variables. But then the final weights would not be adjusted. As a compromise, we choose only two variables: gender and race.

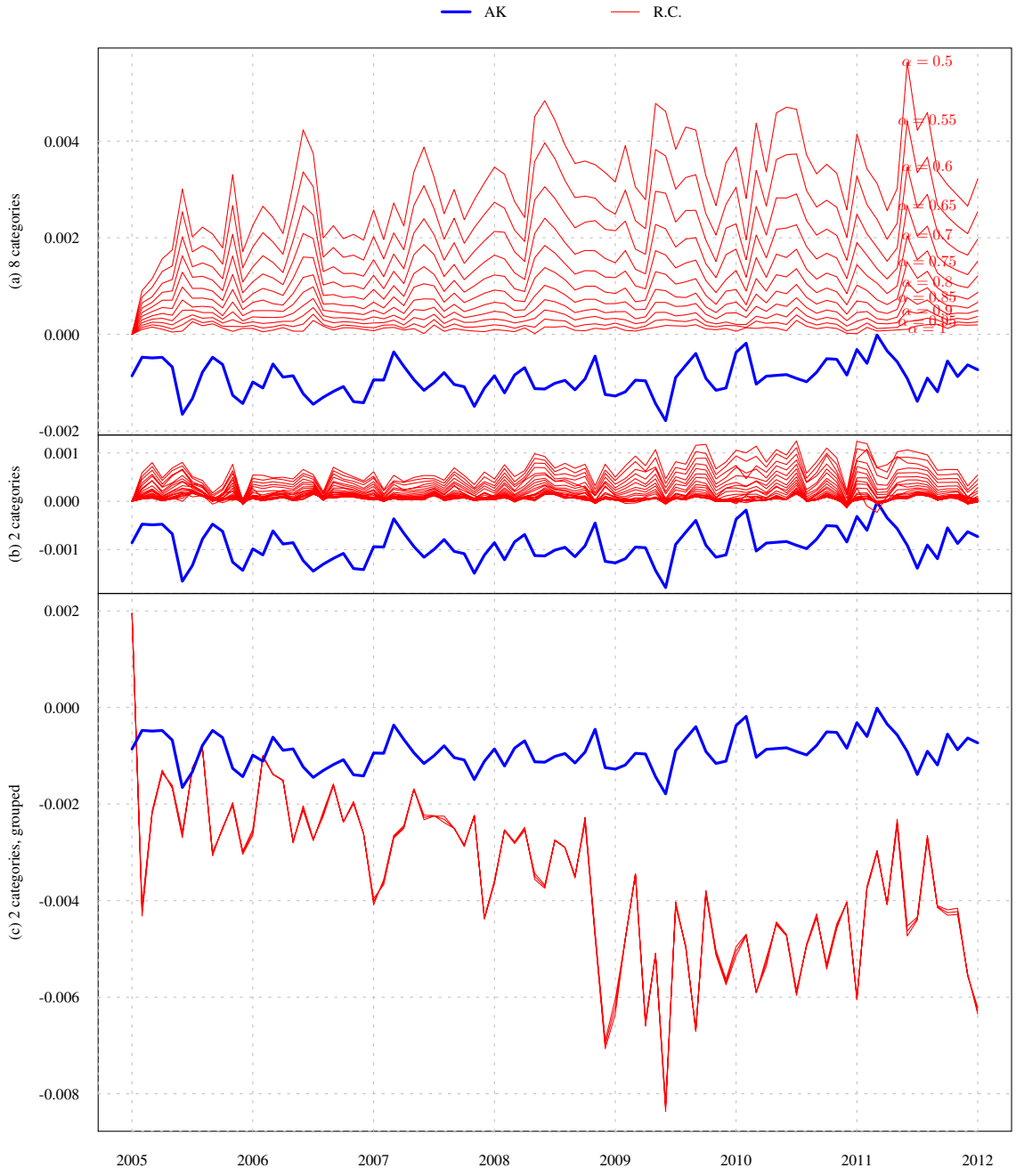
4.3 Results

We report results obtained after applying the Fuller-Rao method at the individual level without grouping the covariates. The results obtained after grouping the covariates and/or at the household level are similar with respect to the estimates. However, the weight variation after adjustment will be larger at the household level.

Figure 1(a) displays $\hat{\mathrm{r}}_m^{\mathrm{AK}} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ and $\hat{\mathrm{r}}_m^{\mathrm{MR},8,\alpha} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ versus month m, where $\hat{\mathrm{r}}_m^{\mathrm{AK}}$, $\hat{\mathrm{r}}_m^{\mathrm{Direct}}$, $\hat{\mathrm{r}}_m^{\mathrm{MR},8,\alpha}$ are the AK, direct and regression composite unemployment rate estimates using eight categories for the employment status, respectively, and for $\alpha \in \{0.5, 0.55, \ldots, 1\}$. Figure 1(b) displays $\hat{\mathrm{r}}_m^{\mathrm{AK}} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ and $\hat{\mathrm{r}}_m^{\mathrm{MR},2,\alpha} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ versus month m, where $\hat{\mathrm{r}}_m^{\mathrm{MR},2,\alpha}$ are the regression composite unemployment rate estimates when two categories for the employment status are used, and for $\alpha \in \{0.05, 0.1, \ldots, 1\}$. Figure 1(c) displays $\hat{\mathrm{r}}_m^{\mathrm{AK}} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ and $\hat{\mathrm{r}}_m^{\mathrm{MR},2,\mathrm{h},\alpha} - \hat{\mathrm{r}}_m^{\mathrm{Direct}}$ versus month m, where $\hat{\mathrm{r}}_m^{\mathrm{MR},2,\mathrm{h},\alpha}$ are the regression composite unemployment rate estimates, computed at the household level, using two categories for the employment status, and for $\alpha \in \{0.75, 0.85, 0.95\}$.

It is interesting to note that the AK estimates are always lower than the direct estimates in both Figures 1(a) and 1(b). To our knowledge, such a behavior of AK estimates has not been noticed earlier. The composite regression estimates do not exhibit such behavior for two categories in Figure 1(b) and also for eight categories in Figure 1(a) when α is close to 1. However, the regression composite estimates with eight categories deviate from the direct estimates upward as α deviates from 1. 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 it also leads to an increase in the variances of the estimates.

Figure 1: Estimated series of differences of different composite estimates with the corresponding direct estimates



5 MONTE CARLO SIMULATION EXPERIMENT TO COMPARE DIFFERENT ESTIMATORS

5.1 Description of Simulation Study

Encouraged by our real CPS data analysis, we conducted a Monte Carlo simulation experiment to enhance our understanding of different composite estimators and to study their finite sample properties. The simulation procedure involved generating finite populations, each with size 100,000, for 96 months during the study period 2005-2012 with the same variables and categories given in the original CPS micro-data maintained by the U.S. Census Bureau. In order to make the simulation experiment meaningful, employment statuses for 96 finite populations were generated in a manner that attempt 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, the employment statuses between two consecutive months were made highly correlated subject to a constraint on the global employment rate evolution. The probability of month-to-month changes in employment statuses for an individual was assumed to be zero in case of no change in the corresponding actual national employment rates.

Thirty independent samples, each consisting of employment statuses for 96 months during the study period, were drawn from the corresponding finite populations using the CPS rotating panel design. For each of these thirty samples, employment rate series over the 96-month period were computed using the direct, AK and the Fuller-Rao composite regression methods. Guided by our CPS real data analysis given in Section 4.2, we chose the best Fuller-Rao regression composite method with $\alpha = 0.95$ and linkage at the individual level. For the comparison purpose, we produce average series over the study period, average being taken over the thirty independent samples.

5.2 Results

Figure 2(a) displays the average of the different estimates over different sample replicates and the simulated finite population unemployment rates. From the graph, it is clear that both the AK composite and the Fuller-Rao regression composite methods are outperforming the direct estimates. It is difficult to distinguish between the AK composite and the Fuller-Rao regression composite methods from this graph. Thus, for each of the three estimated series, we draw Figure 2(b), which displays the average difference between the estimated series and the corresponding simulated population unemployment series. The direct estimator exhibits the largest bias. Overall the Fuller-Rao is doing better than the AK composite in terms of the simulated bias criterion. Figure 2(c) displays simulated mean squared error of different estimated series. It is clear that overall the Fuller-Rao regression composite estimated series has the smallest simulated MSE. The direct method is performing the worst. Figure 3 displays the simulated biases of different estimated series of month-to-month changes. The Fuller-Rao regression composite estimator is again the clear winner. Figure 4 displays the simulated mean squared error of different estimated series of month-to-month changes in unemployment rates. The Fuller-Rao regression composite estimator is showing the smallest mean squared error.

Figure 2: Simulated means, biases, and mean squared errors of true and different estimated series

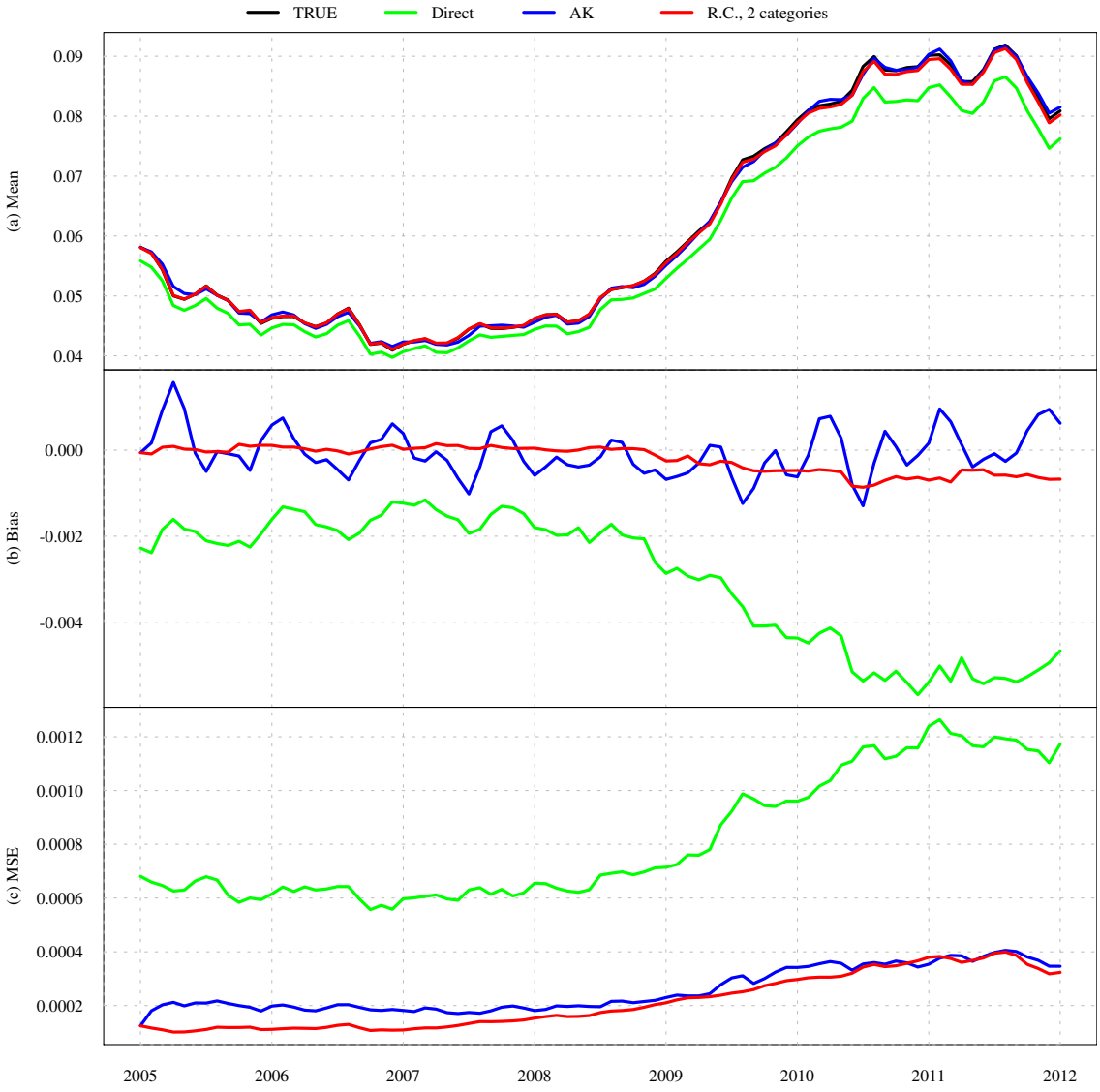


Figure 3: Simulated biases of different estimated series of month-to-month changes

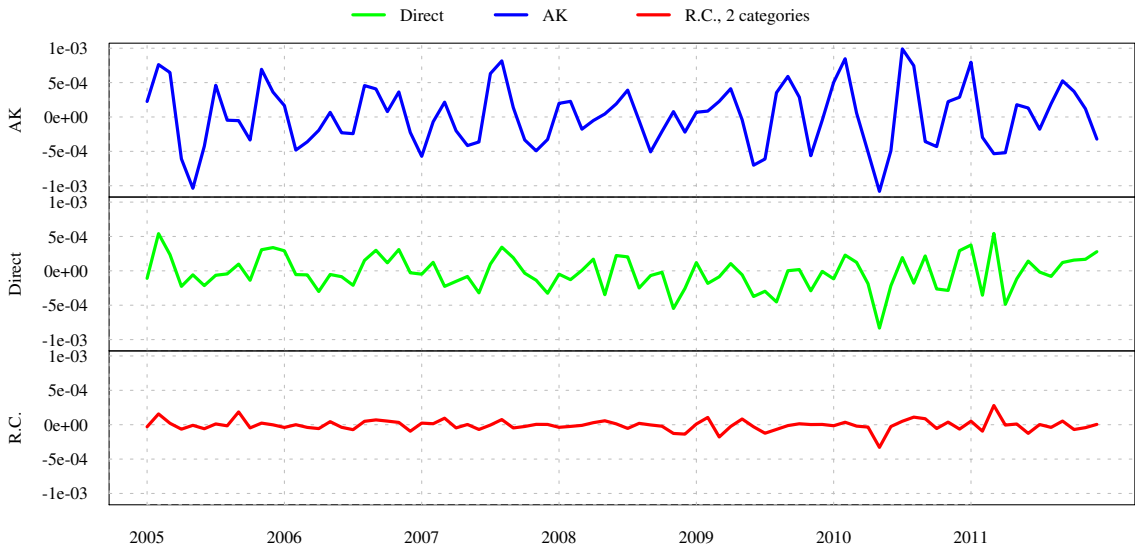
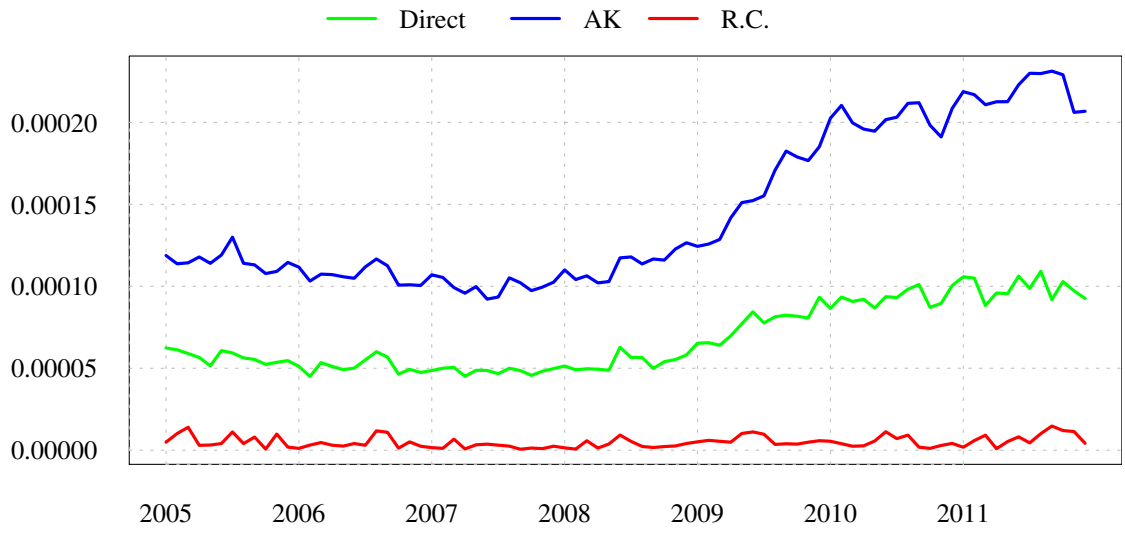


Figure 4: Simulated mean squared errors of different estimated series of month-to-moth changes



CONCLUSION

Our CPS data analysis shows higher values of unemployment rates for the Fuller-Rao regression composite estimates when eight categories of employment status are used and linkage is performed at the individual level. Situation is just the opposite when two categories of the employment status are used and linkage is performed at the household level. In this preliminary work, we have tried a couple of somewhat crude linking methods. The Fuller-Rao regression composite estimator outperforms the AK composite estimator in our simulation study, which suggests further research on regression composite method. As a part of this research, we have developed a R package to facilitate computation of AK, MR1, MR2, and the Fuller-Rao regression composite estimates. In the future, we plan to develop a regression composite estimator that incorporates advanced record linkage method to link individuals over time and uses an appropriate time series hierarchical model for nominal variables and to capture various salient features of the CPS complex design.

References

Fuller, Wayne A, & Rao, JNK. 2001. A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, **27**(1), 45–52.

Gambino, Jack, Kennedy, Brian, & Singh, Mangala P. 2001. Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, **27**(1), 65–74.

Gurney, M, & Daly, JF. 1965. A multivariate approach to estimation in periodic sample surveys. *Page 257 of: Proceedings of the Social Statistics Section, American Statistical Association*, vol. 242.

Salonen, Riku. 2007a. Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Statistics in Transition*, **8**, 503–517.

Salonen, Riku. 2007b. Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Page 63 of: Proceedings of the Second Baltic-Nordic Conference on Survey Sampling June 2–7, Kuusamo, Finland*.

Singh, A.C., & Merkouris, P. 1995. Composite estimation by modified regression for repeated surveys. *Pages 420–5 of: ASA Proc. Surv. Res. Meth. Sec.*

Singh, A.C., Kennedy, B., & Wu, S. 2001. Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, **27**(1), 33–44.

Singh, A.C., Kennedy B. Wu S., & Brisebois, F. 1997. Composite estimation for the Canadian Labour Force Survey. *Pages 300–305 of: ASA Proc. Surv. Res. Meth. Sec.*

ACKNOWLEDGEMENTS

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).