Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

Current Population Survey Sampling Error Autocorrelations[1]

Richard Griffiths and Khandaker Mansur
Demographic Statistical Methods Division, Bureau of the Census, Suitland, MD

Abstract

As a method for improving Current Population Survey state variance estimators, we look to modeling. One potential model, as described by Otto and Bell (1995), is based on the modeling of sampling error variance-covariance matrices for state median income and poverty estimates from March CPS data. Prior to using this model for monthly CPS state variance estimation for labor force estimates, we performed an analysis of sampling error autocorrelations of monthly state labor force estimators. In this paper we discuss the need for analyzing sampling error autocorrelations and describe the analysis which used several years of monthly CPS data. In this analysis we examined state sampling error autocorrelation patterns, time series properties of the sampling errors, and differences in sampling error autocorrelation patterns among the states. The results of this analysis are useful not only for modeling state variance estimates, but also for providing insight about CPS labor force estimates, in general.

Key words: Current Population Survey; Sampling Error; Autocorrelation; Variance Estimation.

I. Introduction

The Current Population Survey (CPS) sample design is a two-stage stratified, cluster design for each state and the District of Columbia. Within each state primary sampling units (PSUs), which are groups of counties, are stratified. A single PSU is selected into the sample in each stratum and a systematic sample of clusters of housing units is then drawn from each sampled PSU. Sampling is done independently in each state.

There are two types of strata in the CPS sample design: self-representing (SR) and non-self-representing (NSR). Each SR stratum contains a single PSU, which is selected into the sample with probability one. Each NSR stratum contains at least two PSUs, one of which is selected into the sample. The variances of CPS estimators thus have two components in NSR strata: a between-PSU component and a within-PSU component. In SR strata the estimators have only a within-PSU component of variance.

The U.S. Census Bureau currently calculates monthly estimates of variances for CPS state labor force estimators. Both successive difference replication and modified half-sample replication methods are used to calculate these state-level variance estimates. (See Fay and Train, 1995, and U.S. Census Bureau, 2000.) The method of successive difference replication is used to estimate within-PSU variances in SR strata and the half-sample replication method is used to estimate total variance in NSR strata.

Current Population Survey (CPS) state-level variance estimation suffers from two general problems:

- relatively small sample sizes at the state level and
- a bias induced by collapsing strata to estimate between-PSU variances.

As a method for improving CPS state variance (and covariance) estimates, we look to modeling. One potential model is that described by Otto and Bell (1995) for modeling sampling error variance-covariance matrices from March CPS data. Prior to examining the fit of this model to monthly CPS state variance estimates, we performed an analysis of sampling error autocorrelations of uncomposited monthly state labor force estimators. In this paper we discuss the need for analyzing sampling error autocorrelations and describe the analysis which used several years of monthly CPS data.

II. Why Study Sampling Error Autocorrelations?

We first answer the question, What are sampling errors and sampling error autocorrelations?

Assume the uncomposited CPS estimate of a characteristic Y for month t and state s may be expressed as $y_{st} = \mu_{st} + e_{st}$, where μ_{st} is the expected value of y_{st}. e_{st} is the sampling error. By sampling error autocorrelation, we mean the correlation among the e_{st}. For instance, the sampling error autocorrelation in state s, month t at lag k is $\mathrm{Corr}(e_{st}, e_{s,t-k})$.

Another natural question to ask is, Why examine sampling error autocorrelations? The answer for us is

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

threefold. The first two parts of the answer regard modeling the state-level variances. First, the original model of Otto and Bell (1995) was designed for modeling March CPS variance-covariance matrices for use in a components-of-variance model for the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. In this model, Otto and Bell (1995) assumed the sampling errors follow a stationary stochastic process. The stochastic process assumed manifests itself in the correlation of the sampling errors. That is, if we express the covariance of sampling errors at lag k as

$$Cov(\mathrm{e}_{\mathrm{st}}, \mathrm{e}_{\mathrm{s,t-k}}) = Corr(\mathrm{e}_{\mathrm{st}}, \mathrm{e}_{\mathrm{s,t-k}})\sqrt{V(\mathrm{e}_{\mathrm{st}})V(\mathrm{e}_{\mathrm{s,t-}k})},$$

$Corr(\mathrm{e}_{\mathrm{st}}, \mathrm{e}_{\mathrm{s,t-k}})$ is the autocorrelation of the assumed stochastic process at lag k. If, for instance, we assume the e_{st} follow an autoregressive AR(1) model, then $\mathrm{e}_{\mathrm{st}} = \phi\,\mathrm{e}_{\mathrm{s,t-1}} + \varepsilon_{st}$, where ε_{st} is a random error (shock) term and ϕ is the first order autoregressive parameter (see, for example, Vandaele 1983), and $Corr(\mathrm{e}_{\mathrm{st}}, \mathrm{e}_{\mathrm{s,t-k}}) = \phi^{k}$. The analysis of sampling error autocorrelations will help us determine the form of the stochastic process that best describes CPS sampling errors and will help us determine the values of the parameters in the stochastic process.

Second, other potential state variance models could have a component which accounts for the autocorrelation present in the variance estimates. Since we are working with estimates of the sampling error variances, studying the autocorrelated nature of the sampling errors tells us something about the autocorrelated nature of the variance estimates themselves. For example, if we know the sampling errors follow an AR(1) process, then the variances of the sampling errors one month apart, $V(\mathrm{e}_{\mathrm{st}})$ and $V(\mathrm{e}_{\mathrm{s,t-1}})$, have the following relationship:

$$V(\mathrm{e}_{\mathrm{st}}) = V(\phi\,\mathrm{e}_{\mathrm{s,t-1}} + \varepsilon_{st}) = \phi^{2}V(\mathrm{e}_{\mathrm{s,t-1}}) + V(\varepsilon_{st}) \quad .$$

Mansur and Griffiths (2001) present a study of the autocorrelation of CPS state-level variance estimators.

Our third and final reason for studying sampling error autocorrelations is that it helps in developing a clear understanding of some of the general characteristics of CPS estimators themselves. This is important not only for us as statisticians working on the CPS sample design and estimation, but also for analysts who use estimates based on CPS data. In this regard, this paper provides a partial update of the important paper by Train et al (1978).

III. CPS Sampling Error Autocorrelation Patterns

Since the CPS has a rotating panel design, there is sample overlap from one month to the next. (See U.S. Census Bureau, 2000, for more detail on the CPS sample design.) As we discuss the analysis, it is helpful to keep in mind the month-to-month overlap in CPS sample, since it provides the basis for understanding the structure of sampling error autocorrelations. Table 1 displays the pattern for the CPS sample overlap.

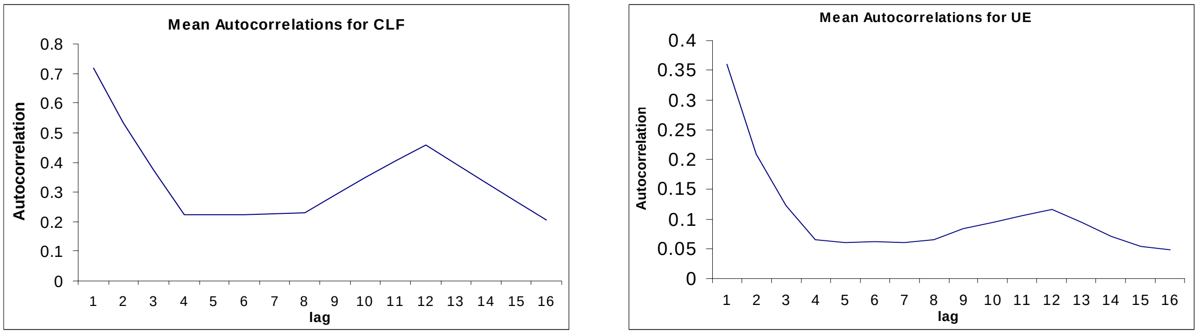
The table shows us that 75% of the households in sample in one month are in sample in the next. At a lag of two months, there is 50% overlap in the households in sample and at lag 3 months, 25% overlap. At lags 4 through 8, there is no sample overlap. Finally, the sample overlap returns at the nine month lag, increases to 50% at the 12 month lag, and then decreases to zero for lags 16 and larger.

We calculated estimated sampling error autocorrelations, using the variance estimation program VPLX (see U.S. Census Bureau, 1999), for two characteristics: estimated number of people in the civilian labor force (CLF) and estimated number of unemployed people (UE) for all states over the months January 1996 to December 1999. Figure 1 displays the average sampling error autocorrelations for lags 1 through 16 for CLF and UE.

Table 1 CPS Sample Overlap

Lag (in months)	Sample overlap (%)	Lag (in months)	Sample Overlap (%)
1	75	9	12.5
2	50	10	25
3	25	11	37.5
4	0	12	50
5	0	13	37.5
6	0	14	25
7	0	15	12.5
8	0	16+	0

Figure 1



The averages were calculated by averaging autocorrelations over all states and months.

We see from Figure 1 that for CLF and UE, the autocorrelations are largest at lags 1 through 3 and 9 through 15, as we'd expect from the CPS sample design. We also note that for the other lags – lags which correspond to no sample overlap in the CPS design – the autocorrelations do not die off to zero. Both of these phenomena need to be reflected in the form of the stochastic process model assumed for the sampling errors when we fit the Otto/Bell model.

IV. Time Series Models for Sampling Error Autocorrelations

The type of stochastic process model we'll use to describe the sampling errors will be an autoregressive moving average (ARMA) model. ARMA models describe the sampling errors as functions of their past values and random error terms:

$$e_{st} = \phi_1 e_{s,t-1} + \ldots + \phi_p e_{s,t-p} + \varepsilon_{st} + \theta_1 \varepsilon_{s,t-1} + \ldots + \theta_q \varepsilon_{s,t-q} \tag{1}$$

(1) is an ARMA(p,q) model.

The following two ARMA models evidently describe the CLF and UE sampling errors well:

- ARMA(1,0)(0,1)$_{12}$:

$$e_{st} = .75 e_{s,t-1} + \varepsilon_{st} + .7\varepsilon_{s,t-12}$$ and

- ARMA(1,1)(0,1)$_{12}$:

$$e_{st} = .61 e_{s,t-1} + \varepsilon_{st} + .3\varepsilon_{s,t-1} - .11\varepsilon_{s,t-12} + (-.3)(.11)\varepsilon_{s,t-13}$$

The autocorrelation patterns of sampling errors that follow these models are very similar to the observed CPS sampling error autocorrelation patterns of Figure 1. We can see this in Figure 2, where we graph the observed CPS sampling error autocorrelations and the sampling error autocorrelations for the two ARMA processes (called theoretical process 1 and theoretical process 2, respectively, in the graphs). In these graphs the solid lines represent the observed CPS sampling error autocorrelation patterns and the dashed lines represent the sampling error autocorrelations for sampling errors that follow these ARMA processes.

In terms of fitting the Otto/Bell model to state variance and covariance estimates, this indicates that we can assume the CLF sampling errors follow an ARMA(1,0)(0,1)$_{12}$ process and the UE sampling errors follow an ARMA(1,1)(0,1)$_{12}$ process. This provides the structure for the autocorrelations needed to model the covariances. The fit of the Otto/Bell model to the observed CPS state variances and covariances will then

Figure 2

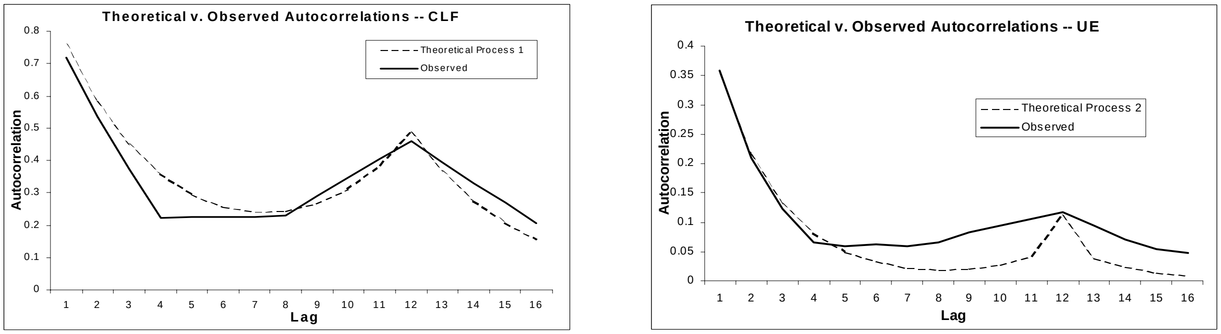


Table 2 State Effects: SS(Effect)/SS(Total) x 100

	state | lag	lag | state	R^2
CLF	11.4%	39.3%	50.7%
UE	4.2%	14.2%	18.4%

produce the estimates of the parameters in these ARMA processes.

V. Grouping States

One of the possibilities we need to allow for in fitting the Otto/Bell model is that of separate fits for different groups of states due to state effects. That is, some states may have different sampling error autocorrelation patterns than others. To check for the existence of state effects, we conducted an analysis of the variance in the sampling error autocorrelations across states and lags. Table 2 exhibits the results of this analysis.

This table shows the percent of variation in the sampling error autocorrelations due to the state effect (state | lag column), due to the lag effect (lag | state column), and due to both effects combined (R^2 column). We see that for CLF, the state effect explains 11.4% of the variation in the sampling error autocorrelations and for UE, the state effect explains 4.2%. While these percentages may look small, we note that we are looking for any indication of differences in sampling error autocorrelation patterns due to state effects and this table provides that evidence, more clearly for CLF than for UE.[2]

Taking this to be an indication that we need to fit the Otto/Bell model separately for different groups of states, we undertook an analysis to determine potential groupings of states. One criteria that we used in trying to determine such groupings was that the groups must have some rational justification. That is, we wanted to be able to point to a characteristic shared by the states in each group other than that they had similar observed sampling error autocorrelation patterns. This lead us to postulate two possible groupings:

- Former direct-use states and non-direct-use states.[3]
- States with relatively large rural populations and states with relatively large urban populations.

We calculated mean autocorrelations for each group separately. Figure 3 displays the mean autocorrelations at lags 1 through 16 for the former direct-use and non-direct-use states for CLF and UE. From this figure we see that the differences in autocorrelation patterns between these two groups of states is minimal for both CLF and UE. To put these visual observations on a more formal basis, we conducted an analysis of the variance in the sampling error autocorrelations. The results of this analysis are presented in Table 3.

The results given in Table 3 confirm our visual observations: very little of the variation in the sampling error autocorrelations is explained by the differences between former direct-use and non-direct-use states: 0.04% for CLF and 0.2% for UE.

We conducted a similar analysis for the rural-urban grouping of states. Figure 4 displays the mean autocorrelations at lags 1 through 16 for the rural and urban states for CLF and UE. From this figure we see that for this grouping, also, the differences in autocorrelation patterns is minimal for both CLF and UE. Again, to put these visual observations on a more formal basis, we conducted an analysis of the variance in the sampling error autocorrelations. The results of this analysis are presented in Table 4.

The results given in Table 4 confirm our visual observations: very little of the variation in the sampling error autocorrelations is explained by the differences

[2] We didn't perform the typical ANOVA F-tests for significance of effects due to the fact that the usual normality and homoscedasticity assumptions are probably violated by sampling error autocorrelations.

[3] Former direct-use states are those states that, at the inception of the most recent CPS redesign, had a requirement on the coefficient of variation (CV) for the monthly estimator of UE. The requirement was that the CV for the monthly UE, assuming a 6 percent unemployment rate, could be no more than 8 percent. These states were then termed "direct-use states," because their monthly UE estimates were precise enough to be used directly. Due to subsequent sample reductions in the CPS, this requirement no longer holds; thus, we use the term "former direct-use states." The former direct-use states are California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, and Texas.

Figure 3

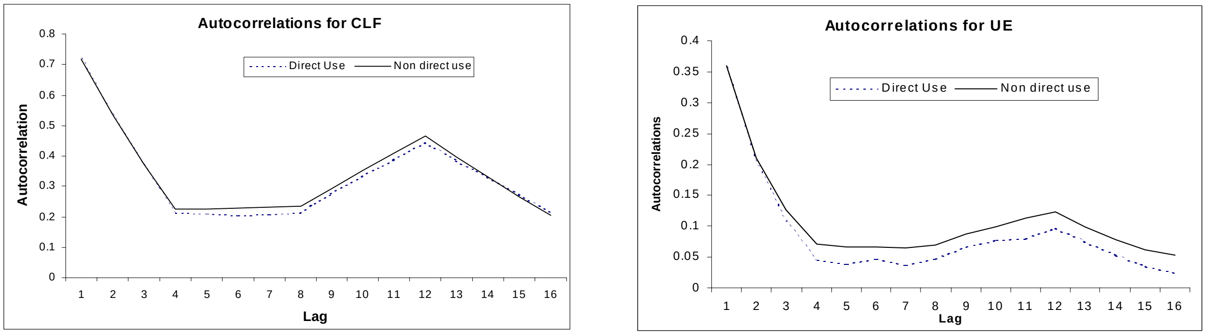


Table 3 Direct-Use State Effect: SS(Effect)/SS(Total) x 100

	direct-use status | lag	lag | direct-use status	R^2
CLF UE	0.04% 0.2%	39.3% 14.2%	39.3% 14.4%

between rural and urban states: 0.006% for CLF and 0.5% for UE.

The conclusion we then make from these analyses for the different grouping of states is that while there do appear to be state effects in the sampling error autocorrelations, they are not attributable to differences between either the former direct-use and non-direct use states or the rural and urban states.

VI. Stationarity

We conclude our preliminary analyses with an examination of one of the assumptions of the Otto/Bell model: namely, the stationarity of the sampling errors. If the sampling errors are weak-sense stationary, then $Corr(e_{st}, e_{s,t-k})$ depends only on the lag k and not on the month t. In other words, there are no month effects in the sampling error autocorrelations.

To examine the validity of this assumption, we conducted an analysis of the variance in the sampling error autocorrelations, looking at the amount explained by a month effect. The results of this analysis appear in Table 5. From this table, we see that very little variation in the sampling error autocorrelations is due to a month effect: 0.2% for CLF and 0.3% for UE. We thus have very little evidence of nonstationarity in the sampling errors and will feel safe in making the assumption that sampling errors are stationary.

Table 4 Rural-Urban State Effect: SS(Effect)/SS(Total) x 100

	rural-urban status | lag	lag | rural-urban status	R^2
CLF UE	0.006% 0.5%	39.3% 14.2%	39.3% 14.7%

Figure 4

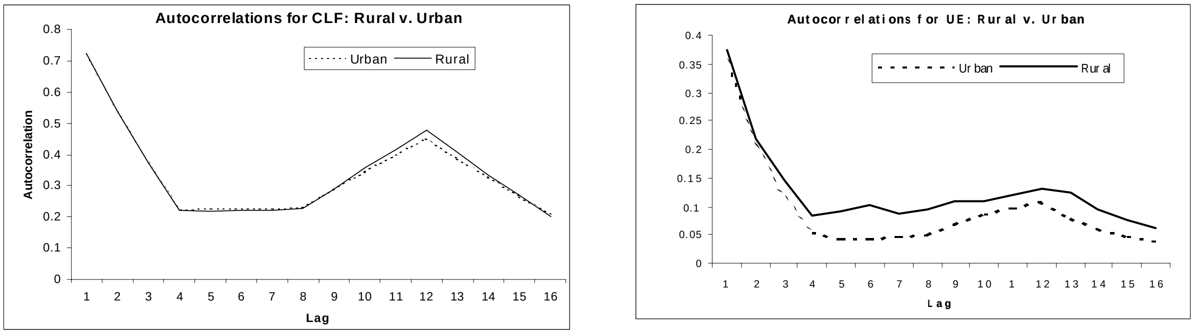


Table 5: Month Effects: SS(Effect)/SS(Total) x 100

	month | lag	lag | month	R^2
CLF UE	0.2% 0.3%	37.6% 13.0%	39.5% 14.5%

VII. Conclusions

In this paper we've examined CPS sampling error autocorrelations as part of a preliminary analysis for fitting the Otto/Bell model to state variances and covariances. We have learned the following from this analysis:

- Sampling errors for CLF appear to follow a pattern which can be adequately modeled by an ARMA(1,0)$(0,1)_{12}$ process; those for UE appear to follow a pattern which can be adequately modeled by an ARMA(1,1)$(0,1)_{12}$ process.
- State effects in the sampling error autocorrelations appear to exist, though they are not traceable to the differences between either former direct-use and non-direct-use states or rural and urban states.
- There is little evidence of nonstationarity of the sampling errors.

All three of these points are important for fitting the Otto/Bell model. Assuming sampling errors follow the given ARMA processes will give us the structure for the autocorrelations needed to fit the model. The existence of state effects indicates that we may need to fit the model separately for different groups of states, though we didn't uncover an appropriate grouping of states in this paper. This is an area of further research. And finally, we have found that the Otto/Bell model's assumption of stationarity for the sampling errors is probably fulfilled.

Acknowledgments

The authors wish to thank Harland Shoemaker, Elizabeth Huang, Bac Tran, and Bob Abramson of the Census Bureau for helpful comments on drafts of this paper. The authors would also like to thank those who offered helpful comments on preliminary presentations of this paper.

References

Fay, R.E. and G.F. Train (1995), "Aspects of Survey and Model-based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section, American Statistical Association, 154-159.

Mansur, K. and R. Griffiths (2001), "Analysis of the Current Population Survey State Variance Estimates," paper presented at the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association, August 2001.

Otto, M.C. and W.R. Bell (1995), "Sampling Error Modelling of Poverty and Income Statistics for States," Proceedings of the Government Statistics Section, American Statistical Association, 160-165.

Train, G., L. Cahoon, and P. Makens (1978), "The Current Population Survey Variances, Inter-Relationships, and Design Effects," Proceedings of the Section on Survey Research Methods, American Statistical Association, 443-448.

U.S. Census Bureau (1999), "VPLX: Variance Estimation for Complex Samples," retrieved October 10, 2001 from http://www.census.gov/sdms/www/vwelcome.html (last revised 8/17/1999).

U.S. Census Bureau, Bureau of Labor Statistics (2000), Current Population Survey: Design and Methodology, Technical Paper 63, Washington, DC.

Vandaele, W. (1983), Applied Time Series and Box-Jenkins Models, Academic Press.