# ESTIMATION OF AUTOCORRELATIONS FOR CURRENT POPULATION SURVEY LABOR FORCE CHARACTERISTICS

Tamara Sue Zimmerman and Edwin Robison
Tamara Zimmerman, Bureau of Labor Statistics, Room 4985, 2 Mass. Ave., N.E., Washington, DC 20212

The Current Population Survey (CPS) uses a rotation sample design where the sample is divided into eight parts called panels. Each month one panel is introduced into the survey. Households in a panel are interviewed in each of the four consecutive months, are dropped from the survey for eight months, and then are interviewed for four more consecutive months. This rotation pattern creates a 75 percent overlap in the sample each month and a 50 percent overlap in the sample each year for a given month, resulting in strong autocorrelation in the CPS labor force data. When sample sizes are small, these correlations can have strong effects on the analysis of the time series. This problem arises frequently in continuing surveys which are sources of important economic statistics. Since deriving correlation estimates is potentially very costly due to the involvement of complex calculations on huge micro data files, statistical agencies rarely provide data users with estimates of the autocorrelations. This paper uses a method for estimating correlations based on the readily available rotation group data. Empirical results for labor force data obtained from the CPS are presented.

KEY WORDS: Autocorrelations; Panel data; Current Population Survey

## 1.0 Introduction

The Current Population Survey (CPS), a nationwide household survey conducted monthly by the Bureau of the Census for the Bureau of Labor Statistics (BLS), provides information on the U.S. labor market. The CPS uses a rotation sample design where the sample is divided into eight parts called panels. Each month panels are rotated in and out of the sample, resulting in large month-to-month sample overlaps which induce strong autocorrelation in the CPS labor force data. When sample sizes are small, these autocorrelations can have strong effects on the analysis of the time series. For example, in signal extraction applications, ignoring the autocorrelations in the sampling error is likely to result in confounding the noise with the signal (Tiller, 1992). In particular, conventional approaches to seasonal adjustment and trend estimation break down for series with autocorrelated sampling error (Tiller, 1995). Such problems arise frequently in continuing surveys which are sources of important economic statistics. Since deriving autocorrelation estimates is potentially very costly due to the involvement of complex calculations on huge micro data files, statistical agencies rarely provide data users with estimates of the autocorrelations.

This paper uses a method for estimating CPS autocorrelations based on the readily available panel data. Using sixteen years of CPS monthly data taken from each of the eight panels, autocorrelations for both the employment and unemployment series are developed and analyzed for each of the 50 states and the District of Columbia (hereafter referred to as 51 states).

This paper is organized as follows: Section 2 discusses the CPS sample design relevant to the estimation of autocorrelations; Section 3 describes the CPS estimation scheme; Section 4 explains the methodology used for estimating the autocorrelations; Section 5 presents the results for both employment and unemployment; and Section 6 provides the conclusions.

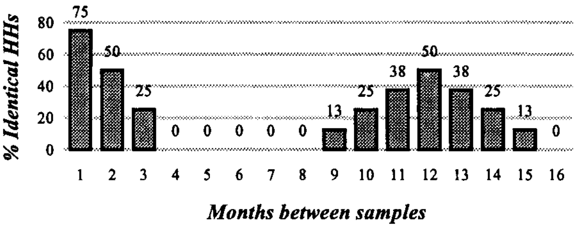
## 2.0 CPS Sample Design.

The autocorrelation structure of CPS labor force data depends upon the CPS sample design and population characteristics. The focus here is on those design features that are likely to have a major impact on the variances and autocovariances of the CPS.

One of the most important features of the CPS sample is the large overlap in sample units from month-to-month. The CPS sample is divided into eight subsamples or panels. Units are partially replaced each month according to a 4-8-4 rotation scheme. When new households are introduced into the sample, they are interviewed for four consecutive months, leave the sample during the following 8 months, return for 4 additional interviews, and then are dropped from the survey permanently. In any one month, one of the panels is interviewed for the first time, another for the second time, etc., with the eighth panel interviewed for the last time. Hence, during any given month, the panels can be uniquely identified by the number of times interviewed (the pane! being interviewed for the first time can be identified as panel 1 while the panel being interviewed for the last time can be identified as panel eight).

This rotation scheme generates significant overlaps. Each month, three-fourths of the sample from the previous month is interviewed, one-eighth of the sample is interviewed for the first time, and one-eighth is resuming interviews after being out of sample for 8 months. Also each month, one-half of the households being interviewed were interviewed in the same month a year ago. Figure 2.1 shows the proportion of the households in the current sample that were also in the sample k months ago. For example, 75% of the households in sample this month were in sample last month, 50% were in two months ago, etc. Note that samples from 4 to 8 months and over 15 months apart have no households in common.

*Figure 2.1.*
*Overlap of Identical Households*



The use of this rotation system requires the periodic replacement of the sample. A key feature of the replacement scheme is that successive samples are generated in a dependent way. Once an initial sample of units within a panel is selected, replacements are obtained from nearby addresses.

The overlap in the CPS sample is important because it induces strong autocorrelation in the labor force series. That is, the current value of the series (either an overestimate or an underestimate of the true value of the labor force) will depend on its own past values. For example, suppose the unemployment rate for the sampled households in the current month is higher than the rate for the entire population. Since 75% of these households will remain in the sample next month, the unemployment rate is likely to be overestimated again.

The extent of this autocorrelation depends not only on the overlap in the sample but also on the stability of the labor force characteristic being estimated. The overwhelming majority of workers spend most of their time in the labor force as employed rather than unemployed. Accordingly, the employment estimates will be more strongly autocorrelated than unemployment since employment is a more stable characteristic of the households being sampled.

## 3.0 CPS Estimation Scheme

The U.S. Census Bureau, using data from the sampled respondents, calculates CPS estimates for a given month. The estimation method involves weighting the data from each sample person by the inverse of the probability of the person being in the sample. Through a series of estimation steps, the selection probabilities are adjusted for noninterviews and survey undercoverage; data from the previous months are incorporated into the estimates through a composite estimation procedure. The composite estimate consists of a weighted average of an estimate based on the entire sample for the current month only and an estimate which is the sum of the prior month composite and change that occurred in the 6 panels common to both months. The estimate of change is based on data from sample units common to both months. A bias adjustment term is also added to the composite estimate to account for the relative bias associated with the number of times interviewed. For any given month, data from persons interviewed for the first and fifth times generally yield higher unemployment estimates than data obtained from persons during any other interview (Bailer 1975). This is often referred to as a month-in-sample (MIS) bias in the literature on rotation group bias. The CPS composite estimate can be written in terms of the panel estimates. In their research on estimating state employment and unemployment rate, Dempster and Hwang (Dempster, 1991) showed that the CPS composite estimator and be written recursively as

$$Y_t^c = \sum_{l=0}^{t-1} a^l Z_{t-l}$$

*where*

$$Z_t = \sum_{i=1}^{8} \left( b_i Y_{t-1,i} + c_i Y_{t,i} \right)$$

$$a = .4$$

$$b_i = \begin{cases} -1/15 & i = 1,2,3,5,6,7 \\ 0 & i = 4,8 \end{cases}$$

$$c_i = \begin{cases} 2/15 & i = 2,3,4,6,7,8 \\ 1/10 & i = 1,5. \end{cases}$$

$Y_{t,i}$ = *estimate of population total* $Y_t$ *at time t from panel i* (3.1).

The $Z$'s are simply linear combinations of the panel estimates for current and prior months while $Y_1^c \equiv Z_1$ is defined as the initial estimate for the month preceding the month for which the composite estimate is first made.

Given this recursive definition of the CPS composite estimator, and assuming covariance stationarity, the variance-autocovariance structure for the composite estimate can be expressed in terms of the variances and covariances associated with the Z's.

$$Var\left(Y_t^c\right) = \sum_{l=0}^{t-1} a^{2l} Var\left(Z_{t-l}\right) + 2\sum_{l=0}^{t-l} a^{2l} \sum_{k=1}^{t-l-1} a^k Cov\left(Z_{t-l}, Z_{t-l-k}\right)$$

$$\begin{aligned} \gamma_{\mathrm{k}} &\equiv Cov(Y_t^c, Y_{t-k}^c) \\ &= a\gamma_{\mathrm{k-1}} + \sum_{l=k}^{t-l} a^{l-k} Cov\left(Z_t, Z_{t-l}\right). \end{aligned} \qquad (3.2)$$

Defining $\eta^o$ as the variance corresponding to the panel esimates and $\eta_{ij}^l$ as the covariance between panel estimates separated by $l$ months, the variance and covariances for the Z's can be expressed in terms of the variance and covariances between the panel estimates.

$$Var\left(Z_t\right) = \sum_{i=1}^{8}\left(b_i^2 + c_i^2\right)\eta^o + 2\sum_{i=1}^{8}\sum_{j=1}^{8} c_i b_j \eta_{ij}^1$$

$$Cov\left(Z_t, Z_{t-l}\right) = \sum_{i=1}^{8}\sum_{j=1}^{8} b_i c_j \eta_{ij}^{l-1} + \sum_{i=1}^{8}\sum_{j=1}^{8}\left(c_i c_j + b_i b_j\right)\eta_{ij}^l$$

$$+ \sum_{i=1}^{8}\sum_{j=1}^{8} c_i b_j \eta_{ij}^{l+1}$$

$$\eta_{ij}^o = \begin{cases} \eta^o & \text{if } \mathrm{i} = \mathrm{j} \\ 0 & ow \end{cases}$$

$$\left.\begin{aligned} \eta^o &= Var(Y_{t,i}) \\ \eta_{ij}^l &= Cov(Y_{t,i}, Y_{t-l,j}) \end{aligned}\right\} \text{for i = 1,2,...,8} \qquad (3.3)$$

## 4.0 CPS Autocorrelation Estimation Methodology

Since the autocovariance structure for the composite estimate depends upon the covariance structure for the Z's (which are linear combinations of the panel estimates), we can estimate autocovariances for the composite estimates by producing variance and covariances for the Z's. In other words, given estimates for the variance and covariances for the panel estimates, $\tilde{\eta}^o$ and $\tilde{\eta}_{ij}^l$, the variance and covariances for the Z's can be estimated by substituting these into formulae (3.3). These then can be substituted into formulae (3.2) resulting in estimates for the variance and autocovariances for the composite estimates. Our estimation methodology focuses on estimating the variances and covariances for the panel estimates.

For each state, we considered the analysis of variance (ANOVA) model below:

$$\begin{aligned} Y_{t,i} &= \mu + \theta_i + \beta_t + \varepsilon_{t,i} \\ \mu &= overall\ mean \\ \theta_i &= MIS\ effect \\ \beta_t &= time\ effect \\ \varepsilon_{t,i} &= sampling\ error. \end{aligned}$$

Here we are modeling the expectation of the *i-th* panel estimate at time *t* with fixed main effect for time and a fixed main effect for the relative bias associated with the number of times the panel has been interviewed (MIS effect). Furthermore, we are assuming that the sampling error is covariance stationary, that is:

$$\begin{aligned} E(\varepsilon_{t,i}) &= 0 \\ Var(\varepsilon_{t,i}) &\equiv \eta^0 \\ Cov(\varepsilon_{t,i}, \varepsilon_{t-l,j}) &\equiv \eta_{i,j}^l. \end{aligned}$$

For each state series, we used sixteen years of CPS monthly data taken from each of the eight panels to produce least squares estimates for the overall mean, MIS, and time effects. Given the residuals,

$$\tilde{\varepsilon}_{t,i} = Y_{t,i} - (\tilde{\mu} + \tilde{\theta}_i + \tilde{\beta}_t)$$

the variance and covariance between panels i and j, k months apart can be estimated by averaging the cross-products of the residuals corresponding to the panel estimates:

$$\tilde{\eta}^o = \frac{\sum_{i=1}^{8}\sum_{t=1}^{192}\tilde{\varepsilon}_{t,i}^2}{8*191} \qquad (4.1)$$

$$\tilde{\eta}_{i,j}^k = I_{A_k}\frac{\sum_{i=1}^{8}\sum_{t=1}^{192}\tilde{\varepsilon}_{t,i} * \tilde{\varepsilon}_{t-k,j}}{192-k}$$

where

$$I_{A_l}=\begin{cases}1 \ if(i,j)\in A_l\\ 0 \ otherwise.\end{cases}$$

for $l=8m+s$, $m=0,1,2,3...$, and $s=1,2,...$

$$A_{m8+1}=\{(i,j):(1,8),(2,1),(3,2),(4,3),(5,4),(6,5),(7,6),(8,7)\}$$

$$A_{m8+2}=\{(i,j):(1,7),(2,8),(3,1),(4,2),(5,3),(6,4),(7,5),(8,6)\}$$

$$A_{m8+3}=\{(i,j):(1,6),(2,7),(3,8),(4,1),(5,2),(6,3),(7,4),(8,5)\}$$

$$A_{m8+4}=\{(i,j):(1,5),(2,6),(3,7),(4,8),(5,1),(6,2),(7,3),(8,4)\}$$

$$A_{m8+5}=\{(i,j):(1,4),(2,5),(3,6),(4,7),(5,8),(6,1),(7,2),(8,3)\}.$$

$$A_{m8+6}=\{(i,j):(1,3),(2,4),(3,5),(4,6),(5,7),(6,8),(7,1),(8,2)\}$$

$$A_{m8+7}=\{(i,j):(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,1)\}$$

$$A_{m8+8}=\{(i,j):(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),(7,7),(8,8)\}$$

The indicator function above simply imposes the restriction that some of the covariances between the panel estimates were assumed to be zero due to the design of the panel samples. Our research indicated that covariances indicated in the sets, $A_k$, above were in fact zero.

To summarize, variance and covariance estimates for the panel estimates result in the following variance and autocovariances for the composite estimates

$$\tilde{\gamma}_k=a\tilde{\gamma}_{k-1}+\sum_{l=k}^{t-1}a^{l-k}\left\{\sum_{ij}^{8}b_ic_j\tilde{\eta}_{ij}^{l-1}+\sum_{ij}^{8}(c_ic_j+b_ib_j)\tilde{\eta}_{ij}^{l}+\sum_{ij}^{8}c_ib_j\tilde{\eta}_{ij}^{l+1}\right\}$$

$$k=1,2,3,...$$

$$\tilde{\gamma}_o=\sum_{l=0}^{t-1}a^{2l}\left\{\sum_{i=1}^{8}(b_i^2+c_i^2)\tilde{\eta}^{o}+2\sum_{ij}^{8}c_ib_j\tilde{\eta}_{ij}^{1}\right\}$$

$$+2\sum_{l=0}^{t-1}\sum_{k=1}^{t-l-1}a^{k}\left\{\sum_{ij}^{8}b_ic_j\tilde{\eta}_{ij}^{k-1}+\sum_{ij}^{8}(c_ic_j+b_ib_j)\tilde{\eta}_{ij}^{k}+\sum_{ij}^{8}c_ib_j\tilde{\eta}_{ij}^{k+1}\right\}. \quad (4.2)$$

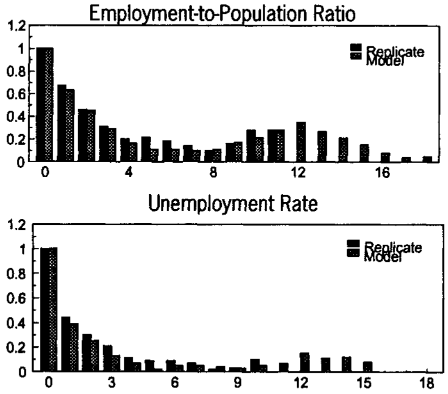
Combining these two results provides autocorrelation estimates for the composite estimates

$$\tilde{\rho}_l=\frac{\tilde{\gamma}_l}{\tilde{\gamma}_o} \quad l=0,1,2,... \quad (4.3)$$

## 5.0 Results

In order to check the consistency of our autocorrelation estimates to those produced in earlier studies, we graphed the autocorrelation estimates for the national labor force series obtained from our ANOVA (Model) approach with those produced through a generalized replication method (Fisher, 1993). Figure 5.1 shows the autocorrelation estimates for both the employment-to-population ratio, the CPS total employed divided by the adult civilian non-institutional population, and the unemployment rate.

*Figure 5.1*
*National Autocorrelations*



As can be seen, the ANOVA estimates are similar to those produced by the replication method. Since the overwhelming majority of the labor force spend most of their time as employed rather than unemployed, the errors in the employment estimates are more strongly autocorrelated than those in the unemployment estimates. However, the overall pattern in the autocorrelation estimates is the same for both labor force series. As expected, the autocorrelation is highest at low lags and falls at the higher lags as the proportion of identical households common to both samples decline. Note the strong peak at the 12-month lag corresponding to the 50 percent overlap in identical households from year to year. Also, after a lag of 15 months when there is no longer any overlap of identical households, the autocorrelations fail to completely dampen out for the employment data. Again this is not unexpected since the replacement of households is rotated into the sample from the same neighborhood. This correlation could persist for a full decade until a new sample is selected.

Figures 5.2 and 5.3 provide some preliminary autocorrelation estimates for the 11 most populous states' employment and unemployment estimates. These state autocorrelation estimates show some similarities to the national autocorrelation estimates. The autocorrelations are strongest at the first 3 lags and decline from lags 4 through 8, where there is no overlap of households in the two samples. Even with no overlap there is still some dependency between non-identical households in the same panel since they are selected from the same neighborhood. The autocorrelations begin to rise at higher lags where the two samples overlap again. The state autocorrelation estimates, like the national estimates, show the peak at the 12 month lag which corresponds to the local peak in the sample overlap.

*Figure 5.2*
*State Employment-to-Population Ratio Autocorrelations*

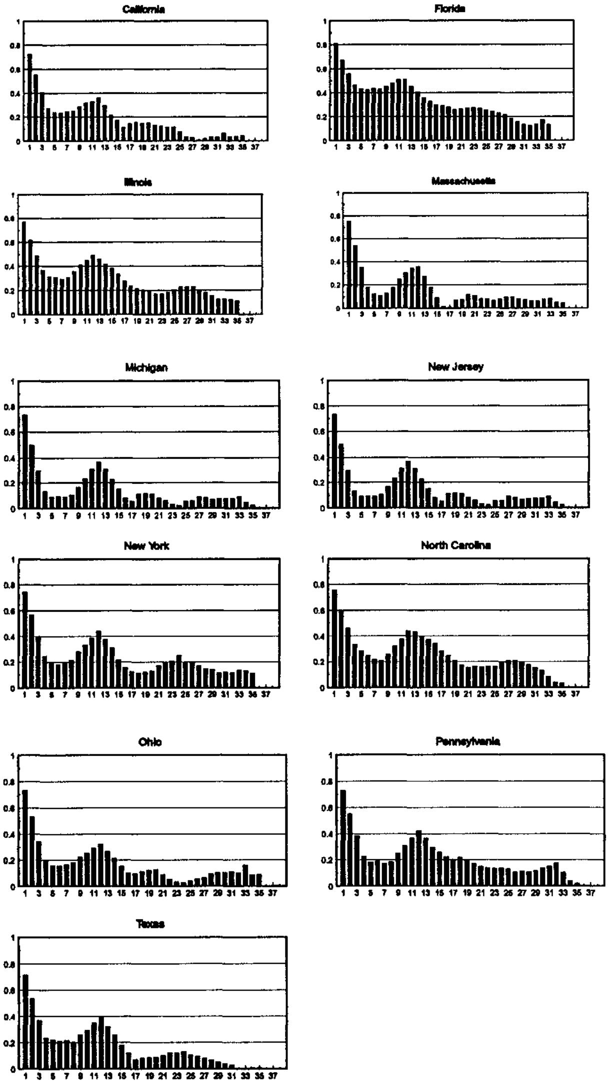
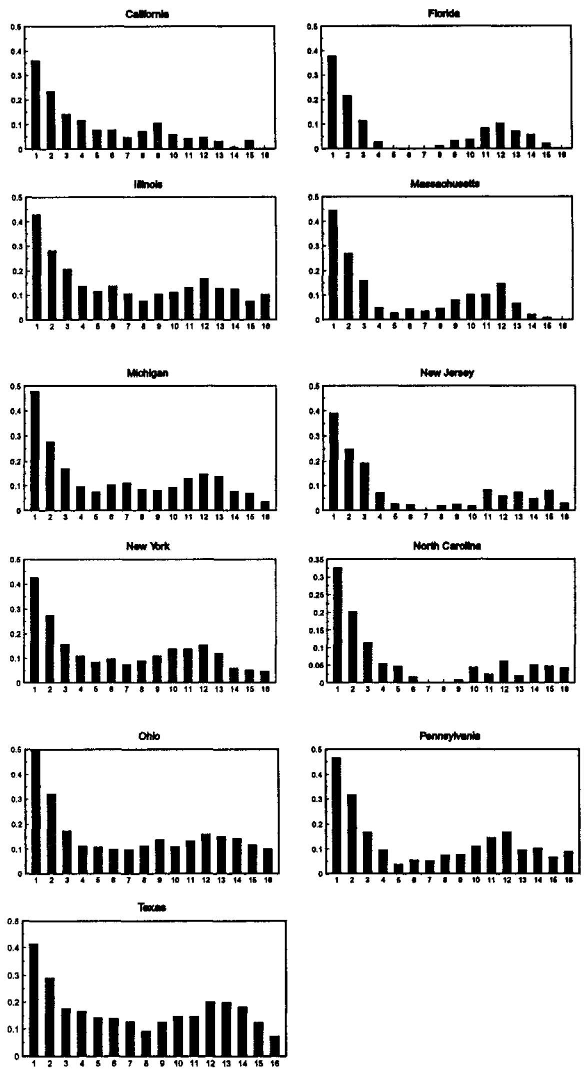


*Figure 5.3*
*State Unemployment Rate Autocorrelations*



## 6.0. Conclusions

In this paper, we proposed a fairly simple procedure for estimating state autocorrelations for both employment and unemployment. Although not as efficient as a procedure based on the individual sampling units such as replication, the method has a cost advantage and provides information on the autocorrelation structure of the labor force series at higher lags. We also compared our autocorrelation estimates for the national labor force series to those obtained through a replication procedure and found them to be quite similar, indicating that this method does offer some promise for estimating state level autocorrelation for CPS labor force series.

## References

Bailer, B.A. (1975), "The Effects of Rotation Group Bias on Estimates from Panel Surveys," *J. Amer. Statist. Assoc.* 70, 23-30.

Dempster, A.P., and Hwang, J.S. (1991), "A Sampling Error Model of Statewide Labor Force Estimates from the CPS," paper prepared for the U.S. Bureau of Labor Statistics.

Fisher, R., and McGuinness, R., (1993), Bureau of the Census memorandum.

Tiller, R.B. (1992), "Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey," *Journal of Official Statistics*, 8, 149-166.

Tiller, R.B. (1995), "Trend Estimation for Continuous Survey Data with Correlated Errors," unpublished paper presented at 1995 ISI meetings

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.