MULTI-YEAR AVERAGES FROM A ROLLING SAMPLE SURVEY

Nanak Chand, Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, D.C. 20233

1. Introduction

Rolling sample surveys, such as the American Community Survey (ACS), are designed to give reliable multi-year estimates for small domains. The ACS collects the basic population and housing data using monthly rolling samples throughout the decade to update the information traditionally available from the census long form. The basic ACS estimates will be the annual averages of data obtained every month. As in the case of decennial census, this survey will sample small government units at a higher rate than other areas. This will update the long form data on characteristics of areas smaller than states. With the ACS providing the detailed annual information, the census long form will phase out after the 2000 census, the 2010 census concentrating mainly on the basic count of population.

The ACS will produce reliable annual estimates of characteristics of interest for areas with population of about 250,000 or more. The smaller areas would require cumulation of multi-year data to result in adequate sample size. The objective would then be to arrive at a fairly reasonable and simple method of cumulating three or five years' annual estimates of desired characteristics for small areas. An analogue to the census long form annual estimate would be a simple average of the ACS annual estimates.

This raises questions about what length of average is best for various applications, and about how using the multi-year averages differs from using single-year snapshots. Keeping the sampling error close to that of the census long form, would suggest cumulation of three years of data for medium-sized areas, with population between about 50,000 and 250,000, and cumulation of five years of data for smaller areas.

This paper develops methods for comparing moving averages with single-year estimates, with varying assumptions pertaining to the underlying series of data. The results provide evidence regarding the properties of regularly updated rolling averages when the goal is to compare the current characteristics of the various subdomains.

In addition to providing a broadly applicable model incorporating the use of multi-year averages, the paper contains concise general formulas showing the effect of these averages, taking into account the effects of characteristics such as jumps and spikes in the time series. Accompanying examples illustrate the general principles under different assumptions regarding the underlying variables.

Section 2. analyzes properties of estimates from series of data observed at a small number of equal time intervals. An example of such a series is the number of persons above poverty level in a geographic region or within a socio-economic subdomain in the region. Section 3. applies simulated time series to compare moving averages with the corresponding single year census estimates within the averaging period. Section 4. summarizes the conclusion that these averages generally result in better estimates of the true population values than the single point estimates.

2. A General Theory to Measure the Effect of Jumps and Spikes in the Series

2.1 Notation

The value of a population characteristic given by the census taken at a specific time point may be considered as an estimate of the unknown values at future time periods. An alternative is to estimate these values by a function of a set of observed values within a suitable time period.

This section provides a comparison of the census estimates with the moving average estimates based on their mean square errors. The analysis takes into account the non-stationary nature of the underlying series characterized by an occasional spike or a permanent jump in data observed over time.

$\{ Y_t \} = \{ Y_t,\ t = 1, ..., T \}$ is a series of a characteristic of population in a given domain. The moving average $A(Y_t)$ of 2n+1, Y variables in the interval [t-n, t+n] is given by

$$A(Y_t) = (2n+1)^{-1} \sum_{i=0}^{2n} Y_{t-n+i},$$
$$t = n+1, \ldots, T-n-1.$$

The mean square error of an estimate $\gamma(Y_t)$, used to estimate Y_{t+k}, $k \geq n+1$, is given by

$$M[\gamma(Y_t)] = (T-n-k)^{-1} \sum_{t=n+1}^{T-k} (\gamma(Y_t) - Y_{t+k})^2.$$

2.2 A Comparison of Mean Square Errors

Theorem:

Let the time series $\{ Y_t,\ t = 1, \ldots, T \}$ be of the form

$$Y_t = \mu + at + X_t \text{ for } t < S, \text{ and}$$

$$Y_t = \mu + at + X_t + \zeta \text{ for } t \geq S,$$

where a is a linear trend factor, $\{ X_t \}$ are independently distributed random variables each with mean 0 and variance σ^2, and ζ is the size of a downward or an upward jump in the series occurring at time S. Let $A(Y_t)$ be the moving average of the (2n+1), Y variables in the interval [t-n, t+n]. Then the expected value of the difference δ in mean square errors for estimating Y_{t+k} for lag k, $k \geq n+1$, by Y_t, as compared to that for the estimate $A(Y_t)$, for equally likely integers S over the interval [1, T], is positive and is given by

$$E(\delta) = \frac{2n}{2n+1}\sigma^2 + \frac{2}{3T}\frac{n(n+1)}{(2n+1)}\zeta^2,$$

where

$$\delta = M[Y_t] - M[A(Y_t)].$$

Proof:

Let

$$\delta_t = (Y_{t+k} - Y_t)^2 - (Y_{t+k} - A(Y_t))^2,$$

$$\xi_t = (X_{t+k} - A(X_t)),$$

and

$$\eta_t = X_{t+k} - X_t.$$

The interval [1, T] may be expressed as the union of the following four disjoint sets as:

$$[1, T] = B \cup C \cup D \cup E$$

with

B = [1, t-n-1], $C = \sum_{i=0}^{2n} C_i$,

D = [t+n+1, t+k], and E = [t+k+1, T],

where the set C_i contains the single time point {t-n+i}, i = 0, ..., 2n.

Let P(A) be the probability of S being in set A and let δ_{tA} be the corresponding difference in the mean square errors conditional on this event. We then have,

$$E(\delta_t) = E(\delta_{tB})P(B) + \sum_{i=0}^{2n} E(\delta_{tC_i})P(C_i) + E(\delta_{tD})P(D) + E(\delta_{tE})P(E)$$

$$= E[(ak+\eta_t)^2 - (ak+\xi_t)^2]\frac{t-n-1}{T} + \sum_{i=0}^{n} E[(ak+\eta_t)^2 - (ak+\xi_t+\frac{i\zeta}{2n+1})^2]\frac{1}{T} + \sum_{i=n+1}^{2n} E[(ak+\eta_t+\zeta)^2 - (ak+\xi_t+\frac{i\zeta}{2n+1})^2]\frac{1}{T} + E[(ak+\eta_t+\zeta)^2 - (ak+\xi_t+\zeta)^2]\frac{k-n}{T} + E[(ak+\eta_t)^2 - (ak+\xi_t)^2]\frac{T-k-t}{T}.$$

Since

$$E(\xi_t) = 0,\ E(\eta_t) = 0,$$
$$E(\xi_t^2) = Var(\xi_t) = \frac{2n+2}{2n+1}\sigma^2, \text{ and}$$
$$E(\eta_t^2) = Var(\eta_t) = 2\sigma^2,$$

the expression for $E(\delta_t)$ simplifies to

$$E(\delta_t) = \frac{2n}{2n+1}\sigma^2$$
$$+ \frac{1}{T}\sum_{i=1}^{n}\left[(ak)^2 - \left(ak+\frac{i\zeta}{2n+1}\right)^2\right]$$
$$+ \frac{1}{T}\sum_{i=n+1}^{2n}\left[(ak+\zeta)^2 - \left(ak+\frac{i\zeta}{2n+1}\right)^2\right]$$
$$= \frac{2n}{2n+1}\sigma^2 - \frac{\zeta^2}{T(2n+1)^2}\sum_{i=1}^{2n} i^2$$
$$- \frac{2ak\zeta}{T(2n+1)}\sum_{i=1}^{2n} i + \frac{1}{T}(n\zeta^2+2akn\zeta)$$
$$= \frac{2n}{2n+1}\sigma^2 + \frac{2}{3T}\frac{n(n+1)}{2n+1}\zeta^2$$

The result follows since

$$E(\delta) = E[E(\delta/t)]$$
$$= (T-k-n)^{-1}E\left[\sum_{t=n+1}^{T-k} E(\delta_t)\right]$$
$$= E(\delta_t).$$

Corollary 1:

Let the series $\{ Y_t,\ t = 1, ..., T \}$ be of the form

$$Y_t = at + \sum_{j=1}^{m}\alpha_j X_{jt} \text{ for } t < S, \text{ and}$$

$$Y_t = at + \sum_{j=1}^{m}\alpha_j X_{jt} + \zeta \text{ for } t \geq S,$$

where $\{ X_{jt},\ j = 1, ..., m \}$ are independently distributed random variables with means $\mu_1, \ldots, \mu_m$ and variances $\sigma_1^2, \ldots, \sigma_m^2$. Then the expected value of difference δ between the mean square errors $M[Y_t]$ and $M[A(Y_t)]$ is given by

$$E(\delta) = \frac{2n}{2n+1}\sigma^2 + \frac{2}{3T}\frac{n(n+1)}{(2n+1)}\zeta^2,$$

where

$$\sigma^2 = \sum_{j=1}^{m}\alpha_j^2\sigma_j^2.$$

Proof:

The theorem applies with

$$X_t = \sum_{j=1}^{m}\alpha_j X_{jt} - \mu,$$

$$\mu = \sum_{j=1}^{m}\alpha_j\mu_j,$$

and

$$Var(X_t) = \sigma^2.$$

Corollary 2:

Let $\{ Y_{t-n+i},\ i = 0, ..., 2n \}$ be the sample statistics corresponding to the census variables $\{ Y_{t-n+i} \}$ and let $a(y_t)$ be the resulting moving average. Then the expected value of the difference d in the mean square errors for estimating Y_{t+k} for lag k, $k \geq n+1$, by Y_t as compared to that by $a(y_t)$ is given by

$$E(d) = \left(1 - \frac{c^2}{2n+1}\right)\sigma^2 + \frac{2}{3T}\frac{n(n+1)}{(2n+1)}\zeta^2,$$

where

$$d = M[Y_t] - M[a(Y_t)].$$

and

$$c^2 = Var(y_{t-n+i}) / \sigma^2, \quad i = 0, ...,2n.$$

Proof:

Let

$$d_t = (Y_{t+k} - Y_t)^2 - (Y_{t+k} - a(y_t))^2,$$

$$\psi_t = (X_{t+k} - a(X_t)),$$

we have,

$$E(\psi_t) = 0, \text{ and}$$

$$Var(\psi_t) = \sigma^2 [1 - \frac{c^2}{2n+1}],$$

Since

$$Var(\eta_t) - Var(\psi_t) = \sigma^2 [1 - \frac{c^2}{2n+1}],$$

the corollary follows by replacing ξ_t *by* ψ_t in the proof of the theorem.

Example:

Let $\{Y_1, ..., Y_{10}\}$ be a series of annual numbers of persons with income above a certain level L, in a domain of population, for a ten year period, with

a. A general average μ of numbers of persons with income greater than L, subject to a positive linear trend of five percent per year.

b. A spike in the curve represented by a random variable X_{st} given by

X_{st} = +3 with probability .025
= -3 with probability .025
= 0 with probability .950

c. A jump of size plus or minus 3.5 with probability of occurrence equal to .1 in each of the ten years, and

d. White noise consisting of unit normal variates $\{e_t\}$.

A spike represents a sudden temporary change in the characteristics of the area, while a jump represents a sudden permanent change such as closing of a large factory or a military base.

Given $\{Y_1, ..., Y_5\}$, the expected value of the difference δ in the mean square errors of estimating each of $Y_6, ..., Y_{10}$ by Y_3 as compared to that by the moving average

$$A(Y_t) = \frac{1}{5} \sum_{i=1}^{5} Y_i$$

is calculated by applying Corollary 1. with the following parameters:

a =.05, n =2, ζ = + *or* -3.5

$$X_t = X_{st} + e_t,$$

and

$$\sigma^2 = Var(X_t) = 1.45$$

This gives

$$E(\delta) = \frac{4}{5}(1.45)^2 + \frac{2}{25}(3.5)^2$$

$$= 2.662$$

3. An Example of Evaluation of Multi-Year Averaging of Data

3.1 Analysis of Simulated Time Series for Small Areas

The first full test of ACS occurred in 1996, and the actual multi-year data for all areas will not be available for some time. Testing the appropriateness of the proposed estimates on other surveys may not be suitable because of the differences in measurement errors among the various surveys.

An alternative testing procedure to avoid the above limitations involves simulating time series of annual estimates for a hypothetical small area using known time series models to generate the true population values. The three year or five year averages will then provide analogues to one year census estimates for comparison with the true population values.

The multiyear averages based on fresh data are clearly different from the traditional time series projections which would require many years of observed ACS data. While the latter projections have their own optimality properties under assumed models, our present objective is to assess measurement errors of multiyear average estimates derived from fairly fresh data as compared to the single year estimates of characteristics of interest.

An appropriate model for simulation is that of general autoregressive integrated moving average (ARIMA) time series (Anderson (1971), Box and Jenkins (1976), Dickey and Fuller (1979), Fuller (1976), Harvey (1981), Kendall (1976), and Priestley (1981)). Two examples of particular interest from this general class are the second order autoregressive (AR(2)) process given at time t by

$$Y_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} = e_t$$

where e_t is normally distributed with mean 0 and variance σ^2, and is independent of Y_{t-1} and Y_{t-2}; and an integrated moving average (IMA(1,1)) process, given by

$$Y_t = Y_{t-1} + e_t - \alpha e_{t-1}$$

where $\{e_t\}$ are independent and identically distributed random variables each with mean 0 and variance σ^2.

We perform simulation on the AR (2) model along with an alternative process containing contamination of occasional random spikes separate from the autocorrelated components, and consider both the three and five year averages.

3.2 Assumptions

The Y series of true population values is assumed to follow the AR (2) process and is given at time t by

$$Y_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} = e_t.$$

The Z series is defined as

$$Z_t + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} = e_t + \varepsilon_t + \varepsilon_t$$

where e_t has a normal distribution with mean 0 and variance σ^2, and is independent of Z_{t-1} and Z_{t-2}. To generate a fairly general pattern of spikes, the variable ε_t is taken as the product of three independent random variables given by

$$\varepsilon_t = U_t V_t W_t,$$

and

$$\varepsilon_t = \sum_{\tau=1}^{\tau=t} \overline{U}_\tau \overline{V}_\tau \overline{W}_\tau,$$

U_t assumes values 1 and -1 each with probability .5, V_t is a Bernoulli random variable with probability of success equal to .05, and $\overline{W}_t$ has a Poisson distribution with mean 3. Similarly, $\overline{U}_\tau$ assumes values 1 and -1 each with probability .5, $\overline{V}_\tau$ is a Bernoulli random variable with probability of success equal to .05, and $\overline{W}_\tau$ has a Poisson distribution with mean 3. and each of the six random variables are independent of each other.

3.3 Mean Square Error Comparisons

Let $A_3\ (Y_t)$ and $A_5\ (Y_t)$ respectively denote the simple average of the Y values for the three and five year periods (t-1, t+1) and (t-2, t+2). The root mean square errors for estimating Y_{t+k} for lag k, by Y_t as compared to $A_3\ (Y_t)$ and $A_5\ (Y_t)$ are given by

$$R_1(\alpha_1, \alpha_2, \sigma, Y, k, 1) = \frac{1}{\sqrt{(n-m+1)}} \sqrt{\sum_{t=m}^{t=n} (Y_t - Y_{t+k})^2}$$

$$R_3(\alpha_1, \alpha_2, \sigma, Y, k, 3) = \frac{1}{\sqrt{(n-m+1)}} \sqrt{\sum_{t=m}^{t=n} (A_3(Y_t) - Y_{t+k})^2}$$

$$R_5(\alpha_1, \alpha_2, \sigma, Y, k, 5) = \frac{1}{\sqrt{(n-m+1)}} \sqrt{\sum_{t=m}^{t=n} (A_5(Y_t) - Y_{t+k})^2}$$

where m and n are respectively the starting and ending points of the time series selected to measure the mean

square errors. $R_i(\alpha_1, \alpha_2, \sigma, Y, k, i)$, i = 1, 3, 5 ;

for the Z-series are similarly defined.

3.4 Numerical Values

The following table shows the percent reduction in the average root mean square errors obtained by taking three or five year averages as compared with the single point estimates. These reductions are given for i =3, 5,

$$\Delta_{U,k,i} = \frac{M_1(U, k, 1) - M_i(U, k, i)}{.01 M_1(U, k, 1)},$$

where, for j = 1, 3, 5,

$$M_j(U, k, j) = \frac{1}{N}\sum R_j(\alpha_1, \alpha_2, \sigma, U, k, j) ,$$

where U represents either the Y or the Z series, and the summation ranges over all possible permutations of $(\alpha_1, \alpha_2, \sigma)$, N being the number of such

permutations.

Thus the table entries are the $\Delta_{U,k,i}$ values for the Y

and Z series with elements of the vector $(\alpha_1, \alpha_2, \sigma)$

ranging from (.1, .1, .1) to (.9, .9, .9), for the simulated time series of three hundred terms, for lag k = 3, 4, and 5.

Depending on the lag period k, five year averages generally result in a larger reduction in the mean square errors than the three year averages. Larger averages smooth noise and spikes, smaller lag periods are better for trends and jumps.

Table
Percent Reduction in Root Mean Square Errors

LAG /	3	4	5
	Five Year Averages		
Y - Series	40.30	49.10	25.56
Z - Series	26.94	30.55	21.76
	Three Year Averages		
Y - Series	33.86	29.42	35.59
Z - Series	23.04	21.79	24.58

REFERENCES

[1] Anderson, T.W. (1971), The Statistical Analysis of Time Series, New York: Wiley.

[2] Box, G.E.P., and Jenkins, G.M. (1976), Time Series Analysis: Forecasting and Control, Oakland, CA: Holden Day.

[3] Dickey, D.A., and Fuller, W.A. (1979), Distribution of the Estimators for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 427-431.

[4] Fuller, W.A. (1976), Introduction to Statistical Time Series, New York: Wiley.

[5] Harvey, A.C. (1981), Time Series Models, Oxford: Philip Allan Publishers, Ltd.

[6] Kendall, M.G. (1976), Time Series, New York: Hafner Press.

[7] Priestley, M.B. (1981), Spectra Analysis and Time Series, Volume 1: Univariate Series, New York: Academic Press.

(This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.)