Using SAS/IML® for Multiyear Rotation Design Sampling and Estimation in Agricultural Surveys

Raj S. Chhikara, University of Houston-Clear Lake, Houston, TX. 77058
Lih-Yuan Deng, Memphis State University, Memphis, TN. 38152
Yilian Yuan, Memphis State University, Memphis, TN. 38152

ABSTRACT

We first describe a proposal by Chhikara and Deng (1992) which is based on an analysis of variance model that takes into account the successive sampling of units in the area frame across years as sampled by the U.S. Department of Agriculture (USDA). We then discuss some of the computational problems related to the proposed estimation procedure. An implementation using SAS/IML® (1989) for the multiyear method applied to the actual data from USDA is discussed in this paper.

1. INTRODUCTION

The area sampling frame used by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture is based on a land use stratification and provides a full coverage of the geographical area of interest. The primary sampling unit is an area segment which varies in size by land use stratum. For intensive agricultural areas, these segments are often targeted to be one square mile land areas.

The USDA estimation methodology uses the current year survey data and overlooks the fact that the area frame sampling involves multiyear rotation designs with twenty percent replacement of sample units each year. Because of a substantial overlap of sampled units from one year to another, the use of multiyear sample data would augment the sample survey information obtained in the current year and, thereby, effect an increase in the sample size. This would reduce the sampling variance of an estimate.

The estimation methodology based on successive or rotation design sampling has been considered by several investigators. The study by Patterson (1950) was the forerunner to many studies that followed. Rao and Graham (1964) and Graham (1973), among others, studied estimators derived by separating the matched and unmatched units of repeated surveys and developed a composite estimation method involving estimators for the two consecutive periods. Wolter (1979) assumed a general linear model to describe the individual panel estimators and proposed to combine these estimators into one that would have a smaller variance than the one which uses only the latest period sample data. This approach, however, would require a determination or estimation of the covariance matrix of the vector of panel estimators, which may not be feasible.

Thus far the composite estimation has been based upon certain combination(s) of periodic estimates. An alternative approach would be to pool the sample data acquired under a rotation design and construct directly a multi-period estimator. Since sample data would be cross referenced between sample units and periods, these data can be described in terms of a two-way analysis of variance model. In the context of crop surveys using satellite acquired data under a rotation design sampling, Hartley (1980) proposed the analysis of variance approach to utilize multiyear sample data for crop acreage estimation.

In Section 2, we describe the multiyear approach proposed by Chhikara and Deng (1992). Several computational problems related to their proposal are also discussed. In Section 3, we then discuss the implementation of this new procedure using SAS/IML® applied to the actual data from NASS/USDA for estimating the crop acreage and hog total in several states.

2. MULTIYEAR APPROACH

In the area frame sampling, the reporting unit is a tract which is the area of land located within a segment that is under a single operation or management. The estimator of total for a survey item is obtained by adding the tract data for each sample segment, multiplying the sum by the expansion factor for the segment, and then aggregating the expanded segment totals to the stratum and higher levels. Since its computation is based on data from the tracts confined within segments, it is sometimes
referred to as the area tract estimator. This is an un­
biasd estimator and is considered reliable for esti­
mating crop acreages. For background information 
on the land use stratification and estimation proce­
dure, readers may refer to Kuo (1989). We will next 
describe the multiyear procedure recently proposed 
by Chhikara and Deng (1992).

2.1. ANOVA model.
The area tract estimator is computed as follows:

$$\hat{Y} = \sum_{h \in H} \sum_{k=1}^{n_h} E_{hk} y_{hk}$$

where $H$ is the collection of strata, $E_{hk}$ is the expan­
sion factor for segment $k$ in stratum $h$ (which simply 
is equal to the inverse of the probability of selection 
of a segment in the stratum), $n_h$ is the number of 
segments sampled in stratum $h$, and $y_{hk}$ is the agri­
cultural item value for segment $k$ of stratum $h$.

As we mentioned earlier, the area frame sam­
pling has substantial overlap of sampled units from 
one year to another. As Hartley (1980) noted, there 
is a certain amount of consistency in area segment 
characteristics from one year to another. For ex­
ample, the suitability of the segment prevalent soil 
types will be invariant from year to year or the ca­
pabilities of certain operators in a segment to grow 
crops, etc. will be largely persistent. On the other 
hand, factors such as weather and economic condi­
tions will vary across years and will affect the farm 
outcome. Taking these aspects into consideration we 
discuss the following multiyear model approach to 
estimation of crop acreages that avoids any unwar­
ranted assumptions sometimes made, and correctly 
so, in the analysis of other ‘time series’.

We now consider estimation for a stratum. Let 
$y_{tk}$ be the agricultural item value in year $t$ for seg­
ment $k$ in a stratum. Then the $y_{tk}$ can be viewed 
consisting of the stratum mean in year $t$, the segment 
effect, and an error component. One can, therefore, 
describe it in terms of an analysis of variance model:

$$y_{tk} = \alpha_t + b_k + e_{tk} \quad (1)$$

where $k = 1, 2, ..., n_t$ and $t = 1, 2, ..., T$. Here $n_t$ de­
notes the number of segments sampled for the stra­
tum in year $t$. Other terms in the model (1) are as 
follows: $\alpha_t$ is the mean value for the characteristic 
of interest over all the segments in the stratum for 
year $t$. $b_k$ is the segment effect representing the de­
viation of the $k$th segment response from that of the stratum mean. $e_{tk}$ is the model error associated with 
segment $k$ in year $t$. The error term in (1) comprises 
the sampling error and any interaction that may ex­
ist between years and segments.

We assume that the $b_k$ are independently dis­
tributed with $E(b_k) = 0$ and $\text{Var}(b_k) = \sigma^2_b$, and the 
$e_{tk}$ are independently distributed with $E(e_{tk}) = 0$ 
and $\text{Var}(e_{tk}) = \sigma^2_e$. Furthermore, $b_k$ and $e_{tk}$ are 
independent of each other.

The above linear model can be written in matrix 
form as

$$y = X\alpha + U\beta + \varepsilon,$$      \hspace{0.5cm} (2)

where $X$ is the design matrix consisting of 0's and 
1's which account for the effect due to $\alpha_t$'s and $U$ 
is the design matrix of 0's and 1's which are speci­
fied according to the rotation sampling scheme. The 
dimensions of $X$ and $U$ are $N \times T$ and $N \times S$, respec­
tively, were $N = \sum_{t=1}^{T} n_t$ and $S$ is the total number 
of distinct segments sampled in $T$ years. Note that 
$S \leq N$. Let

$$b^* = U\beta + \varepsilon,$$      \hspace{0.5cm} (3)

then $E(b^*) = 0$ and

$$\text{Var}(b^*) = \sigma^2_e (I + \gamma U'U) = \sigma^2_e W_\gamma,$$ \hspace{0.5cm} (4)

where $\gamma = \sigma^2_b / \sigma^2_e$. The weighted least-squares esti­
mator of $\alpha$ is given by

$$\hat{\alpha} = (X'W_\gamma^{-1}X)^{-1}X'W_\gamma^{-1}y.$$ \hspace{0.5cm} (5)

The parameter $\gamma$ in $W_\gamma$ can be estimated by $\hat{\sigma}^2_b / \hat{\sigma}^2_e$ 
preferably using some previously obtained survey or 
pilot study data. The variance-covariance matrix of 
$\hat{\alpha}$ is given by

$$\text{Var}(\hat{\alpha}) = (X'W_\gamma^{-1}X)^{-1}\sigma^2_e.$$ \hspace{0.5cm} (6)

This variance formula still applies asymptotically if 
$\gamma$ is replaced by a consistent estimator $\hat{\gamma}$.

The single-year estimate currently used at NASS 
and described in Section 2 can be obtained by setting 
$\gamma = 0 \text{ (i.e., no segment effect)}$ in Equation (5) for $\hat{\alpha}$, 
and are given by

$$\hat{\alpha} = (X'X)^{-1}X'y = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_T),$$ \hspace{0.5cm} (7)

where $\bar{y}_t$ is the sample mean for year $t$. In order to 
evaluate the performance of $\hat{\alpha}$ as an alternative to
\( \hat{\alpha} \), one computes the variance-covariance matrix of \( \hat{\alpha} \) under model (2), which would be

\[
\text{Var}(\hat{\alpha}) = (X'X)^{-1}(X'W\gamma X)(X'X)^{-1}\sigma_\epsilon^2.
\]

### 2.2. Computation Efficiency.

Note that the dimension of \( W_\gamma \) is \( N \times N \), which, in general, is a large matrix, especially when \( T \) is large. The direct computation of the inverse of \( W_\gamma \) may present a numerical problem. Using the special structure of \( W_\gamma \) as defined in (4), we can find the inverse as following: Using the matrix inversion formula (see Searle, 1982, p. 151)

\[
(I + AB)^{-1} = I - A(I + BA)^{-1}B,
\]

we have

\[
W_\gamma^{-1} = I - UD_1 U', \tag{8}
\]

where \( m_k \) is the total number of yearly survey occasions for which segment \( k \) falls in the sample and the diagonal matrix,

\[
D_1 = \text{diag}\left(\frac{\gamma}{1 + \gamma m_1}, \frac{\gamma}{1 + \gamma m_2}, \ldots, \frac{\gamma}{1 + \gamma m_S}\right).
\]

Making use of (8), one can simplify the matrix expressions in Equations (5) and (6) as follows.

\[
X'W_\gamma^{-1}X = D_2 - X'UD_1 U'X,
\]

where

\[
D_2 = \text{diag}(n_1, \ldots, n_T).
\]

Clearly, this would avoid a direct tedious computation of \( W_\gamma^{-1} \). Similarly,

\[
X'W_\gamma^{-1}y = X'y - X'UD_1 U'y
\]

where \( y_t \) is the total of \( y \)'s for year \( t \), \( y_k \) is the total of \( y \)'s for segment \( k \), and \( d_i = \frac{T}{1 + \gamma m_i}, i = 1, 2, \ldots, S. \)

### 2.3. Estimation of \( \gamma \).

Note that the multiyear estimator \( \hat{\alpha}_T \) is the best linear unbiased estimator (BLUE) of \( \alpha_T \), provided \( \gamma \) is specified correctly in matrix \( W_\gamma \). In general, \( \gamma \) is unknown and needs to be estimated from some survey data. We will use \( \hat{\alpha}(\gamma_0) \), and \( \hat{\alpha}_T(\gamma_0) \) for the latest period, to denote the multiyear estimators when \( \gamma \) is specified as \( \gamma_0 \).

If it is necessary to estimate \( \gamma \) using the regular survey data, one can first get an initial estimate \( \hat{\gamma} \) from the multiyear rotation sampling design data using the two-way analysis of variance method. This is done first by computing the mean square due to segment and the mean square due to error as follows:

\[
MS_b = \frac{\sum_{k=1}^S m_k (\bar{y}_k - \bar{y}_{..})^2}{S - 1}
\]

and

\[
MS_e = \frac{\sum_{t=1}^T \sum_{k \in S_t} (y_{tk} - \bar{y}_{..})^2}{N - S + 1},
\]

where \( S \) is the number of distinct segments sampled in \( T \) years, \( S_t \) is the set of sampled segments in the \( t \)th year, \( m_k \) is the number of times the \( k \)th segment is observed in \( T \) years, \( \bar{y}_{tk} \) is the sample average for segment \( k \), and \( \bar{y}_{..} \) is the grand average of all segments in \( T \) years. Although this computation of \( MS_e \) is most suitable for a full randomized block design, we use it here to compute simply an initial estimate of \( \gamma \).

When there is a full randomized block design, it is well-known from the random-effect ANOVA model that

\[
E(\text{MS}_b) = \sigma_\epsilon^2 + T\sigma_\epsilon^2.
\]

Accordingly, a reasonable estimate of \( \sigma_\epsilon^2 \) may be obtained by

\[
\hat{\sigma}_\epsilon^2 = \frac{1}{T} \left[ \text{MS}_b - \text{MS}_e \right] +
\]

where \( [x]_+ = \max(x, 0) \) and \( T' = N/S \). Since \( N = \sum_{t=1}^T n_t \), \( T' \) represents on average the number of years each segment is observed. Hence, an initial estimate of \( \gamma \) is obtained by

\[
\hat{\gamma} = \frac{1}{T'} \left[ \frac{\text{MS}_b}{\text{MS}_e} - 1 \right] +
\]

We then iteratively search for the best \( \gamma_0 \) in the neighborhood of \( \hat{\gamma} \) which will minimize

\[
\text{Var}(\hat{\alpha}_T(\gamma_0)) = \nu_T^2 (X'W_{\gamma_0}^{-1}X)^{-1}v_T\sigma_\epsilon^2.
\]

Then \( \sigma_\epsilon^2 \) can be estimated by

\[
\hat{\sigma}_\epsilon^2(\gamma_0) = \frac{y'W_{\gamma_0}^{-1}y - \hat{\alpha}'(\gamma_0)X'W_{\gamma_0}^{-1}y}{N - T}.
\]

This search for the best \( \gamma_0 \) is justified because if \( \gamma_0 \) is the true \( \gamma \) in model (2), then \( \hat{\alpha}(\gamma_0) \) is BLUE and \( \text{Var}(\nu_T^2 \hat{\alpha}(\gamma_0)) \) should be smaller than the variance of any other estimator \( \nu_T^2 \hat{\alpha}(\gamma) \) of \( \alpha_T \).
4. Implementation Using SAS/IML®

The SAS/IML® code implementing Chhikara and Deng (1992) multiyear estimation procedure, as described in Section 2, is listed below:

```sas
proc iml;

* subroutine gami(y) computes the initial estimate of gamma. The input is a matrix y which includes the information of one stratum. The output is the estimate of initial gamma. */

start gami(y); /* gets number of columns of y */
t=ncol(y); /* s represents number of different segments */
s=nrow(y); /* nt = number segments in each year */

ymiss=(y<0); nt=j(1,t,s)-ymiss[+,];

n=nt[ , +];

mk=j(s,l,t)-ymiss[ , +]; /* y = replace missing with 0 */
y=0<>y;

segavr=y[ , +]/mk;

yavr=y[+,] In;

llt=y[+,];
yt=j(1,t,O);

doi=ltot;

if nt[1,i]>0 then yt[1,i]=llt[1,i]/nt[1,i];
end;

ms1=shape(yt,s,t);
ms2=(shape(segavr',t,s)');
ms3=j(s,t,yavr);

ms=ssq((1-ymiss)*(y-ms1-ms2+ms3));

if ((n-s-t+1)>0) then mse=ms/(n-s-t+1);
else mse=0.0;

sl=segavr-j(s,1,yavr);

s2=mk#(sl##2);

if ((s-1)>0) then mssb=s2b/(s-1);
else mssb=0.0001;

if (mssb>mse) then gamma=(mssb/mse-1)/(n/s);
else gamma=0.1;

return(gamma);
finish;
```

```sas
/* subroutine varian(y, gamma) calculates the estimate of alpha and variance of it. The input matrix y and an estimate of gamma. The output is a matrix z as follows: z[1,1]=variance of alpha corresponding to the last year z[1,2]=alpha corresponding to the last year z[1,3]=the last diagonal value of inverse of x'wx z[1,4]=estimate of variance of error (se2) */

start varian(y,gamma); /* get number of columns of y */
t=ncol(y); /* s represents number of different segments */
s=nrow(y); ymiss=(y<0);

nt=j(1,t,s)-ymiss[+,]; n=nt[ , +];

mk=j(s,1,t)-ymiss[ , +]; ux=j(s,t,1)-ymiss;

llt=ux#y;

ux=ux';

di=diag(gamma/(1+gamma#mk));

d2=diag(nt);

xwx=d2-xu*di*ux;

doi=ltot;

if nt[1,i]=0 then xwx[i,i]=0.01;
end;

a=det(xwx);

if a=0 then ixwx=inv(xwx);
else ixwx=inv(t);

xy=(wt[ , j])';

uy=wt[ , +];

xwxy=ux*di*uy;

w2=ssq(wt);

yw=wt2-uy*di*uy;

alpha=xwxy*xwxy;

z=j(1,4,O);

z[1,2]=alpha[t,1];

if (n>t) then se2=(yw-alpha*xwxy)/(n-t);
else se2=0.0001;

z[1,4]=se2;

vv=ixw[t,t]*se2;

z[1,1]=vv;

z[1,3]=ixw[t,t];

return (z);
finish;
```

```sas
/* subroutine

1129
```
total(y,fpcl,alphal,ixvx,se2) figures out the estimate of the total of the stratum for the last year and the variance for the last year using both methods proposed by usda and uhcl. the input are a matrix y which includes the the information of the stratum, fpcl which equals to the expansion factor of the last year, alphal which denotes the estimate of alpha for the last year, ixvx for the last diagonal value of inverse of x'wx, and se2 which is estimate of variance of error. the output is re as defined below:

re[1,1]=total of the stratum of the last year using method of usda
re[1,2]=variance of the estimate using method of usda
re[1,3]=total of the stratum of the last year using method of uhcl
re[1,4]=variance of the estimate using method of uhcl

/*
start total(y,fpcl,alphal,ixvx,se2);
fpcl=1-1/fpcl;
if (alphal < 0) then alpha=0;
else alpha=alphal;
t=ncol(y);
s=nrow(y);
ymiss=(y<0);
nt=j(1,t,s)-ymiss[+];
ntt=nt[1,t];
m1=j(s,t,1)-ymiss;
wt=m1*y;
yt=j(1,t,0);
do i=1 to t;
   if nt[1,i]>0 then yt[1,i]=wt[1,i]/nt[1,i];
end;
ytt=yt[1,t];
y1=j(s,1,ytt);
y2=m1[i,t]*wt[1,t]-yt[1,t];
y3=ssq(y2);
if (ntt-1) then s2=y3/(ntt-1);
else s2=0;
tot2=yt+ntt;
var2=s2*ntt*fpcl;
tot1=alpha+ntt;
var1=se2*(ntt*2)*fpcl*ixvx;
re=j(1,4,0);
re[1,1]=tot1; /*uhcl total */
re[1,2]=var1; /*uhcl variance*/
return(re);
finish;

/*main program begins*/

/*input1 contains stratay, seg, and sampling data of each year*/
use input1;
read all into w;
do i=3 to ncol(w);
do j=1 to nrow(w);
   if (w[j,1]=.) then w[j,1]=-1;
end;
end;

/*input2 contains the stratum and corresponding value of mexpfctr*/
use input2;
read all into f;

/*input3 contains the stratum and corresponding number of segments*/
use input3;
read all into s;
ns=nrow(s);
s1=j(ns+1,1,0);
s1[1:ns,1]=s[ ,2];
sggv=j(ns,4,0);
sggv[ ,1]=s[ ,1];
saisf=j(ns,4,0);
saisf[ ,1]=s[ ,1];
totv=j(ns,5,0);
totv[ ,1]=s[ ,1];
k1=0;
k2=s1[1,1];
c=ncol(w);
do i=1 to ns;
   /*w contains stratum and segment id as the first 2 columns*/
y=w[(k1+i):k2,3:c];
print y;
   /*call gami to get the estimate of initial gamma value*/
g=gami(y);
sggv[i,2]=g;
/* create a neighborhood of the initial gamma */
gam = do (0.005*5g, 3*g, 0.003*g);
vi = gam;
do j = 1 to ncol(gam);
q = varian(y, gam[l, j]);
vi[j] = q[l, l];
end;

/* find the minimum variance */
vmin = vi[<>];
sggv[i, 4] = vmin;
/* find position of gamma corresponding to the minimum value */
index = loc (vi = vmin);
/* find the best gamma */
gammabe = gam[l, index[l]];
sggv[i, 3] = gammabe;

/* get the information related to the best gamma */
al = varian(y, gammabe);
saisf[i, 2] = al[l, 2];
saisf[i, 3] = al[l, 3];
saisf[i, 4] = al[l, 4];
totv[i, 2:5] =
total (y, f[i, 2], al[l, 2], al[l, 3], al[l, 4]);
k1 = k2;
k2 = k2 + s1[i + 1, 1];
end;

namesg = {straty gammaigammabest variance};
namesa = {straty alphaxxxxt se2fpsc};
nametv = {straty usdat usdav uhclt uhclv};

print sggv[colname = namesg];
print saisf[colname = namesa];
print totv[colname = nametv];

uhcl = totv[+, +];
uhcl[1, 1:4] = uhcl[1, 2:5];
/* ratio = var (usda) / var (uhcl) */

Acknowledgement
The authors' work was partially supported with the funds provided by the National Agricultural Statistics Service of the U.S. Department of Agriculture under a cooperative research program at the University of Houston—Clear Lake. The authors would like to thank Mr. Bill Iwig of USDA for his insightful explanation of the USDA data collection and estimation procedures.

References