MULTIYEAR, THROUGH-THE-SEASON CROP ACREAGE ESTIMATION USING ESTIMATED ACREAGE IN SAMPLE SEGMENTS

Robert L. Sielken, Jr., Texas A&M University

Abstract. Using remote sensing technology and knowledge of crop growth behavior, NASA can estimate a crop's proportional at harvest acreage in a 5×6 nautical mile segment. Such estimates are determined at a few times during the crop's growing season for a small number of segments sampled from a large homogeneous area (stratum). Similar estimates are obtained for consecutive years on possibly different yearly samples of segments from the same stratum. A mixed effect analysis of variance model for such a multiyear data set is used as the basis for a weighted least squares estimate of the crop's proportional at harvest acreage in the stratum for the current year. Potentially useful weighting procedures and transformations of the segment estimates are also briefly discussed.

1. INTRODUCTION

NASA seeks to use its remote sensing capabilities to estimate a particular crop's proportional acreage in a large homogeneous area (stratum). Since it is not practical to process all of the satellite information on the entire stratum, a sample of relatively small segments is selected each year over a multiyear period. (In the past a segment has been a 5×6 nautical mile rectangle.) An estimate of the crop's at harvest acreage proportion is made for each sample segment at a few times during the crop's growing season. This paper focuses on the procedure for estimating the crop's current year at harvest acreage proportion in the stratum on the basis of the through-the-season segment level estimates collected over a multiyear period.

Since the same segments do not have to be in the sample every year, there is an interesting associated problem of determining an optimal multiyear sampling design. Although this problem is not specifically discussed herein, some technical reports on this topic are included in the references.

H. O. Hartley, during his years (1963-1979) as Distinguished Professor of Statistics at the Institute of Statistics, Texas A&M University, contributed greatly to NASA's research efforts pertaining to crop acreage estimation and stimulated his co-workers' efforts. Several of the more recent technical reports on crop acreage estimation which he either wrote or encouraged are listed among the references.

2. THE BASIC MODEL FOR MULTIYEAR ESTIMATION

The basic model relating the stratum's at harvest crop acreage to the crop's estimated at harvest acreage in the sample segments has the general form

$$\begin{aligned} y(\text{observation}) &= \text{year effect} + \text{segment effect} \\ &\quad + \text{season bias} + \text{noise} \end{aligned} \tag{1}$$

where $y(\cdot)$ is an appropriate transformation. The specific form of model (1) is

$$y(p_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell} \qquad \begin{aligned} t &= 1,\ldots, T, \\ s &= 1,\ldots, S, \\ \ell &= 1,\ldots, L \end{aligned} \tag{2}$$

where

$p_{ts\ell}$ = the estimated proportion of the s-th segment's acreage that will contain the crop at harvest time in the t-th year when the estimate is made at crop calendar time ℓ (for example, ℓ=1 could denote early season, ℓ=2 mid-season, and ℓ=3 at harvest time);

$y(p_{ts\ell})$ = a variate transformation of $p_{ts\ell}$;

α_t = the stratum's transformed crop acreage proportion for the t-th year;

b_s = the s-th sampled segment's departure from the stratum's transformed crop acreage proportion; the b_s's are random variables with expectations zero and variance σ_b^2;

δ_ℓ = the systematic difference between the non-harvest time estimates of the crop's transformed at harvest acreage proportion and the corresponding estimate made at harvest time ($\delta_L \equiv 0$);

$e_{ts\ell}$ = the aggregate of sampling and classifications errors in the transformed data.

The primary objective is to estimate the crop's at harvest proportion of the stratum acreage in the current year, T; that is, estimate the inverse transformation of α_T denoted by $P_T = y^{-1}(\alpha_T)$. Secondary objectives could be improved estimates of at harvest acreages in previous years or estimates of changes in the stratum's crop at harvest acreage proportion from year to year.

Estimates of the stratum's crop at harvest acreage proportion are often needed throughout the current year as well as at harvest time. For example, an early season estimate based on observations for ℓ=1,...,L for t=1,...,T-1 and only ℓ=1 for t=T is often desired.

Of course, even though the estimate $\hat{P}_T=y^{-1}(\hat{\alpha}_T)$, of the stratum's crop at harvest acreage proportion for the current year involves only $\hat{\alpha}_T$, the $\hat{\alpha}_T$ depends on the entire multiyear data set and the estimates of the segment effects and the systematic biases which are assumed to be constant from year to year.

3. TRANSFORMATIONS OF THE ESTIMATED SEGMENT PROPORTIONS

The simplest transformation, $y(p)$, of the estimated segment crop acreage proportion, p, to use in (2) is the identity transformation

$$y(p) = p \quad .$$

However, it is very doubtful that the additive model (2) would hold for $y(p)=p$ particularly if the p's exhibit a large variation within the stratum. On-the-other-hand a multiplicative model for p may be more reasonable. For instance if

(i) 30% of the stratum's acreage is planted in wheat at the time wheat is harvested in year t;

(ii) the s-th segment's wheat acreage averages only 80% of the stratum's wheat acreage at harvest time;

(iii) the at harvest acreage estimate made at mid-season is only 70% of the at harvest estimate made at harvest time; and

(iv) the sampling and classification errors cause the estimated at harvest acreage to be 110% of what it would be without these errors,

then

$$p_{ts\ell} = (.30)(.80)(.70)(1.10) \quad .$$

Here a logarithmic transformation, $y(p) = \ell n(p)$, would be appropriate and

$$y(p_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell}$$

$$= \ell n(.30) + \ell n(.80) + \ell n(.70) + \ell n(1.10) \quad .$$

The logit transformation,

$$y(p) = (1/2)\ \ell n[p/(1-p)] \quad ,$$

is another useful transformation which approximately converts a multiplicative model for p into an additive model for y(p). In addition, the logit transformation has the property that

$$0 \leq y^{-1}(\hat{\alpha}_T) \leq 1$$

whereas the logarithmic transformation guarantees only

$$y^{-1}(\hat{\alpha}_T) \geq 0$$

and the identity transformation makes no guarantees.

All three of the above transformations are considered in Sielken and Dahm (1981). There approximate expressions are derived for

(i) the variance of $y(p)$,

(ii) the bias of $y^{-1}(\hat{\alpha}_T)$,

(iii) the mean squared error of $y^{-1}(\hat{\alpha}_T)$, and

(iv) confidence intervals for P_T

under the assumption that p arises from a binomial random variable.

4. THE WEIGHTED LEAST SQUARES ANALYSIS OF THE SEGMENT ESTIMATES

When estimating the parameters (α_t, b_s, δ_ℓ) in model (2), it is not particularly reasonable to assume that the variance of $y(p_{ts\ell})$ is the same for all t,s,ℓ. Hence a <u>weighted</u> least squares analysis is called for as opposed to the usual unweighted least squares analysis.

The first step in the weighted least squares analysis of model (2) proceeds as follows:

(i) The data is the segment estimates $p_{ts\ell}$.

(ii) For all $p_{ts\ell}$ satisfying $|y(p_{ts\ell})| < \infty$, define $y_{ts\ell} = y(p_{ts\ell})$. (This is all $p_{ts\ell}$ if $y(p) = p$ but only $p_{ts\ell} > 0$ if $y(p) = \ell n(p)$ and only $0 < p_{ts\ell} < 1$ if $y(p) = 1/2\ \ell n[p/(1-p)]$).

(iii) Using only $y_{ts\ell}$ with $|y(p_{ts\ell})| < \infty$, determine $\hat{\alpha}_t$, $\hat{b}_s$, $\hat{\delta}_\ell$ for the fixed effects model

$$y_{ts\ell} = \alpha_t + b_s + \delta_\ell + e_{ts\ell}$$

from a weighted least squares analysis of

$$w_{ts\ell}\, y_{ts\ell} = w_{ts\ell}\alpha_t + w_{ts\ell}b_s + w_{ts\ell}\,\delta_\ell + \varepsilon_{ts\ell}$$

with

$$\sum_s b_s \equiv 0 \ , \ \delta_L \equiv 0 \quad ,$$

and

$$w_{ts\ell} = \{\mathrm{Var}[y(p_{ts\ell})]\}^{-\frac{1}{2}} \quad .$$

(iv) For all t, s, ℓ calculate

$$\hat{y}_{ts\ell} = \hat{\alpha}_t + \hat{b}_s + \hat{\delta}_\ell \text{ and } \hat{p}_{ts\ell} = y^{-1}(\hat{y}_{ts\ell}) \quad .$$

(v) (a) If all $p_{ts\ell}$ satisfy $|y(p_{ts\ell})| < \infty$, then redefine $p_{ts\ell} = \hat{p}_{ts\ell}$ and either return to (iii) or stop this first step after a "sufficient" number of iterations.

(b) If all $p_{ts\ell}$ do not satisfy $|y(p_{ts\ell})| < \infty$, then use a Taylor series expansion of y(p) about $p = \hat{p}$,

$$y(p) = y(\hat{p}) + (p-\hat{p}) \left[\frac{dy(p)}{dp}\right]_{p=\hat{p}} \quad ,$$

to create "working y's." For the logarithmic transformation, define

$$y_{ts\ell} = \hat{y}_{ts\ell} + (p_{ts\ell} - \hat{p}_{ts\ell})/\hat{p}_{ts\ell} \ ;$$

for the logit transformation

$$y_{ts\ell} = \hat{y}_{ts\ell} + (p_{ts\ell} - \hat{p}_{ts\ell})/[2\hat{p}_{ts\ell}(1-\hat{p}_{ts\ell})].$$

Using these $y_{ts\ell}$'s redefine

$$p_{ys\ell} = y^{-1}(y_{ts\ell})$$

for all $ts\ell$ and return to (iii).

The final estimate $\hat{\alpha}_T$, which is of primary interest, is obtained in the next step from a mixed effects model weighted least squares analysis of the final $y_{ts\ell}$'s determined in the first step, namely the iterative weighted least squares analysis using a fixed effects model. The mixed model weighted analysis assumes the b_s is a random factor with mean 0 and variance σ_b^2 and is outlined below.

When the basic mixed effects model

$$y_{ts\ell} = \alpha_t + b_s + \delta_\ell + e_{ts\ell}$$

is analyzed, it is considered in the form

$$w_{ts\ell}y_{ts\ell} = w_{ts\ell}\,\alpha_t + w_{ts\ell}b_s + w_{ts\ell}\delta_\ell + \varepsilon_{ts\ell}$$

with $w_{ts\ell}$ again proportional to $[\mathrm{Var}(y_{ts\ell})]^{-\frac{1}{2}}$. In matrix notation

$$Wy = WX\,\binom{\alpha}{\delta} + WU\,b + I\varepsilon$$

where

$y = (y_{111}, y_{112}, \ldots)^T$,

$\alpha = (\alpha_1, \ldots, \alpha_T)^T$,

$\delta = (\delta_1, \delta_2, \ldots, \delta_{L-1})^T$, $(\delta_L \equiv 0)$

$b = (b_1, b_2, \ldots)^T$,

W = matrix containing the $w_{ts\ell}$'s ,

X = matrix of 0's and 1's corresponding to the α's and δ's,

U = sampling design matrix of 0's and 1's corresponding to which b_s's appear in which years, and

I = identity matrix.

The random portion of Wy is $WUb + I\varepsilon$ which has covariance

$$V\sigma_\varepsilon^2 = I\sigma_\varepsilon^2 + WUU^TW^T\sigma_b^2$$
$$= (I + WUU^TW^T\gamma)\sigma_\varepsilon^2$$

where

$$\sigma_\varepsilon^2 = \mathrm{Var}\ (\varepsilon_{ts\ell})\ ,$$

$$\sigma_b^2 = \mathrm{Var}\ (b_s)\ ,$$

$$\gamma = \sigma_b^2/\sigma_\varepsilon^2\ .$$

The weighted least squares estimator of $(\alpha,\delta)^T$ is

$$\widehat{\binom{\alpha}{\delta}} = (X^TW^TV^{-1}WX)^{-1}X^TW^TV^{-1}\,Wy$$

and

$$\mathrm{Var}[\widehat{\binom{\alpha}{\delta}}] = (X^TW^TV^{-1}WX)^{-1}\,\sigma_\varepsilon^2\ .$$

In particular

$$\mathrm{Var}\ (\hat{\alpha}_T) = (X^TW^TV^{-1}WX)^{-1}_{T,T}\ \sigma_\varepsilon^2$$

where $(\ )^{-1}_{T,T}$ denotes the T-th element on the diagonal of the matrix inverse.

Of course, since $\gamma = \sigma_b^2/\sigma_\varepsilon^2$ is unknown, an estimate of γ must be substituted in the weighted least squares analysis of the mixed effects model. The estimation of the variance components σ_b^2 and σ_b^2 and their ratio γ can be based upon equating certain sums of squares from the fixed effects model with their expectations under the mixed model and then solving the equations for σ_b^2 and σ_ε^2. Basically this is Henderson's Method 3. In particular, if the segment effects $b_1, \ldots, b_S$ are treated as fixed effects in

$$Wy = WX\ \binom{\alpha}{\delta} + WUb + I\varepsilon,$$

then the residual sum of squares is

$$(Wy)'[I-(WX,WU)\{(WX,WU)'(WX,WU)\}^{-1}(WX,WU)'](Wy) \tag{3}$$

which under the mixed effects model has expectation

$$[n-(T + L + S - 2)]\ \sigma_\varepsilon^2 \tag{4}$$

where n is the number of observations. Equating (3) to (4) provides a nonnegative estimate, $\hat{\sigma}_\varepsilon^2$, of σ_ε^2 . Similarly, the regression sums of squares due to the b's given the α's and δ's, namely

$$(Wy)'(WX,WU)\{(WX,WU)'(WX,WU)\}^{-1}(WX,WU)'(Wy)$$
$$-(Wy)'(WX)\{(WX)'(WX)\}^{-1}(WX)'(Wy)\ , \tag{5}$$

under the mixed effects model has expectation

$$\sigma_b^2\ \mathrm{trace}\{(WU)'[I-(WX)\{(WX)'(WX)\}^{-1}(WX)'](WU)\}$$
$$+\ \sigma_\varepsilon^2\ (S - 1)\ . \tag{6}$$

Substituting $\hat{\sigma}_\varepsilon^2$ for σ_ε^2 in (6) and equating (5) to (6) provides an estimate, $\hat{\sigma}_b^2$, of σ_b^2. The estimate of γ is $\hat{\gamma} = \hat{\sigma}_b^2/\hat{\sigma}_\varepsilon^2$ as long as $\hat{\sigma}_b^2 > 0$. Since $\sigma_b^2 \geq 0$, an estimate $\hat{\sigma}_b^2 \leq 0$ strongly suggests that γ is in reality a small positive number.

5. COMPUTER IMPLEMENTATION: ACRE

A self-contained computer implementation of the above weighted least squares estimation procedure has been given to Lockeed, NASA, and ERIM.

The computer program has been nicknamed ACRE. ACRE will accept up to 5 years of data ($T \leq 5$) on up to 20 segments ($S \leq 20$) with up to 5 observations per segment per year ($L \leq 5$). These dimension restrictions could of course be easily increased. With the current dimensions ACRE requires 768K bytes of core memory on an AMDAHL 470/V6. The program is written in straightforward FORTRAN and is extensively internally documented.

ACRE will allow the user to choose any one of the three transformations

(i) $y(p) = p$,

(ii) $y(p) = \ell n(p)$, or

(iii) $y(p) = .5\ \ell n[p/(1-p)]$.

ACRE regularly provides the following output:

1. A listing of the observations that have been input.
2. An indication of the transformation y(p) that

the user has requested.

3. The estimated effects $(\hat{\alpha}, \hat{b}, \hat{\delta})$ in the final iteration of the weighted least squares analysis of the fixed effects model.
4. An analysis of the residuals in the fixed effects model.
5. The variance component estimates $\hat{\sigma}_b^2$, $\hat{\sigma}_\varepsilon^2$, and $\hat{\gamma}$.
6. The estimates of the fixed effects $(\hat{\alpha}, \hat{\delta})$ in the weighted least squares anlysis of the mixed model.
7. An analysis of the residuals in the mixed effects model.
8. The stratum's estimated at harvest crop acreage proportion, $P_t = y^{-1}(\hat{\alpha}_t)$, for each year $t = 1, \ldots, T$.
9. Indications of the precision of the estimated at harvest crop acreage proportions for all years. Namely, approximate values for the bias and mean squared error of $\hat{P}_t$, and an approximate 90% confidence interval on P_t for each year $t = 1, \ldots, T$.

6. CURRENT RESEARCH

The sensitivity of the estimate, $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$, of the stratum's at harvest crop acreage proportion to such things as the transformation used, the accuracy of the weights, and the reliability of the estimate of $\gamma = \sigma_b^2/\sigma_\varepsilon^2$ is under study. The empirical behavior of the approximate expressions for the bias of $\hat{P}_T$ and the mean squared error of $\hat{P}_T$ as well as the approximate confidence intervals on P_T is also being evaluated. The extension of the basic model (1) to include year-segment interactions and segment-season interactions is being considered. Another possibility is to replace the seasonal bias term in (1) by a covariate in terms of something like the number of "crop calendar days" passed by the date of the last satellite imagery used in determining the estimated segment at harvest crop acreage proportion.

Another important line of research concerns the nature of the weights themselves. If the true segment at harvest crop acreage proportion were p* and the estimated p's were binomial in nature, then the variance of a segment estimate p would be proportional to p*(1-p*). Furthermore, the variance of y(p) could be derived for a given y, and the appropriate weight in the weighted least squares procedure could be straight-forwardly approximated using the estimated p. However, the variance of the estimate of the segment's at harvest crop acreage proportion may not be binomial in nature but rather depend on such things as

(i) the satellite being used,

(ii) the sharpness of the satellite imagery,

(iii) the amount of satellite imagery available at the time of the segment estimate,

(iv) the nearness of the segment's observed behavior to classical crop profiles,

(v) the season during which the estimate is being made,

(vi) the weather conditions during the crop's growing season,

(viii) the composition of the segment, etc.

The derivation of appropriate weights under this latter scenario is being investigated.

REFERENCES

Technical reports prepared for NASA:

(1) Hartley, H. O. June, 1974. The Sample Design for the ERTS Satellite.

(2) Hartley, H. O. July, 1974. The "ERSATZ" Sampling Plan for the ERTS Satellite II (representing a revision of an earlier report under the same title)

(3) Feiveson, A.H. and H. O. Hartley. August, 1974. Proposed LACIE Sampling Plan.

(4) Hartley, H. O. and D. A. Lamb, April, 1975. Formulas for Variance Estimation in Proposed LACIE Sampling Plan.

(5) Hartley, H. O. April 1975. The Treatment of the Defective Area Segments in the LACIE Survey for the U.S.A.

(6) Hartley, H. O. and D. A. Lamb. September, 1975. Formulas for Variance Estimation in Proposed LACIE Sampling Plan II.

(7) Book, D., Hartley, H. O.,Lamb, J., Larsen, G. and Michael Trenchard. March, 1976. Small Area Estimates of Historical Wheat Acreages for Russia and China.

(8) Hartley, H. O. March, 1976. An Outline of Multiyear Estimates in the LACIE Sampling Plan.

(9) Hartley, H. O. and David Lamb. August, 1976. Estimation of Wheat Acreage Via Small Grain Crop Acreage.

(10) Hartley, H. O. March, 1977. The Use of the 1974 Ag Census for LACIE Estimates.

(11) Hartley, H. O. October, 1976. Bias Through Engineering Restrictions.

(12) Hartley, H. O. and Ly Cong Thuan. Multiyear Estimates for the LACIE Sampling Plans.

(13) Hartley, H. O. Improvements in Multiyear Estimates of Wheat Acreage.

(14) Hartley, H. O. November 1978. Control of Classification Bias Through Ground Truth Data Bank.

(15) Hartley, H. O. February, 1979. A Multicrop-Multicountry Sampling Strategy.

(16) Freund, R.J.,Hartley, H. O. and T. Lee. May 1979. Gains of Precision Achievable by Multi-Year Estimation.

(17) Hartley, H.O., Hughes, T.H. and R. L. Sielken, Jr. May, 1979. A Computer Implementation of the

Multicrop Sampling Strategy.

(18) Gbur, E.E. and R. L. Sielken, Jr. December 1980. Optimal Rotation Designs for Multiyear Estimation, I. Unweighted Estimation.

(19) Gbur, E.E. and R. L. Sielken, Jr. December 1980. Optimal Rotation Designs for Multiyear Estimation, II. Weighted Estimation.

(20) Dahm, P.F. and Robert L. Sielken, Jr. March 1981. Multiyear Estimation of the At Harvest Crop Acreage Proportion: Methodology and Implementation.

(21) Sielken, Robert L. Jr., December 1981. Incorporating Partially Identified Sample Segments Into Acreage Estimation Procedures: Estimates Using Only Observations From the Current Year.

(22) Gbur, E. E. and R. L. Sielken, Jr. December 1981. Missing Observations in Multiyear Rotation Sampling Designs.