General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)

CR-171 785 c.1

(E84-10151) AREA ESTIMATION USING MULTIYEAR DESIGNS AND PARTIAL CROP IDENTIFICATION Final Report (Texas A&M Univ.) 50 p HC A03/MF A01 CSCL 02C
N84-26099
Unclas
G3/43 00151

JUN 1984
RECEIVED
NASA STI FACILITY
ACCESS DEPT.

STATISTICS

FINAL REPORT
CONTRACT NO. NAS9-16785

Area Estimation Using Multiyear Designs
And Partial Crop Identification

TEXAS A&M UNIVERSITY

Contract No. NAS9-16785

Area Estimation Using Multiyear Designs and Partial Crop Identification

Final Report

Robert L. Sielken Jr.
Department of Statistics
Texas A&M University
College Station, TX 77843

May, 1984

Contract No. NAS9-16785

Area Estimation Using Multiyear Designs and Partial Crop Identification

Final Report

This final report refers to project number 4821 entitled "Area Estimation Using Multiyear Designs and Partial Crop Identification". This project spanned the period from November 1, 1983, to March 31, 1984.

1. INTRODUCTION

Agriculture and other renewable resources can be economically inventoried over large areas using aerospace remote sensing techniques. In particular, the surface area devoted to a specific resource in a large region is especially amenable to aerospace estimation. Such resources could be as broadly defined as agriculture, forest, water, snow cover, etc. or as specifically defined as summer crops or corn. These area estimates can be combined with other measures such as estimated yield per acre to obtain production estimates. Once the appropriate estimation methodology has been successfully implemented, the successive estimates are very economical, so that frequent inventories are realistically obtainable.

During 1975-1977 NASA in conjunction with the USDA conducted the Large Area Crop Inventory Experiment (LACIE) to illustrate the potential capabilities of aerospace remote sensing techniques. This pioneering effort also served to remove many of the obstacles for future applications. A summary of the experiment is given in the proceedings of the LACIE Symposium (1978). The target resource in LACIE was the wheat acreage and production in the U. S. Great Plains.

During the transition years 1977-1979 and during 1979-1983 under the recently-terminated AgRISTARS (Agriculture and Resources Inventory Surveys through Aerospace Remote Sensing) program several advances were made in satellite imagery technology, data processing, and statistical methodologies. In addition, target resources were expanded to include other crops and other countries, as well as non-crop resources.

The research under Contract No. NAS9-16785 has focused on the statistical methodology for estimating a particular resource's acreage proportion in a large region at a specified point in time using the estimated resource acreage proportion in a sample of smaller areas. In describing this research it will be assumed that

(i) the resource is a crop,

(ii) the specified time point of interest is the harvest time for the crop,

(iii) the sample areas are all the same size (a 5x6 nautical mile rectangle called a segment), and

(iv) the sample segments are relatively "small" compared to the homogeneous region (stratum) of interest.

Also, it is assumed that in each year of a multiyear period a sample of segments is selected. The composition of the sample may vary year to year. In each year each sample segment's at-harvest crop acreage proportion is estimated at one or more times during the crop growing season. The number of estimates is not necessarily the same for all sample segments in a year and is not necessarily the same for each year. Obviously, the contract research has focused on only one part of a much larger problem. The region of concern herein is really just one stratum in a stratified sample survey

of a country or the world (see, for example, Chhikara and Feiveson (1982)). The size of the sample segment is assumed to be predetermined (see Chhikara and Feiveson (1982) and Chhikara et al. (1984)). Also, since the same segments do not have to be in the sample every year, there is an interesting associated problem of determining an optimal multi-year sampling design (see Chhikara et al. (1984), Gbur and Sielken (1980a), Gbur and Sielken (1980b), Gbur and Sielken (1981) and the discussion in Section 4). The papers by Heydorn (1984) and Hall and Houston (1984), for example, discuss the determination of the sample segment's estimated at-harvest crop acreage proportion. Finally, the estimates arising from the statistical methodology developed under this research contract and the preceding contract (No. NAS9-13894) can be input to procedures for aggregating acreage over several regions and combining acreage estimates with yield estimates to obtain production estimates. The paper by Feiveson (1984) is a good example of the research addressing these latter needs.

2. OVERVIEW OF RESEARCH ACTIVITY

The two major tasks under Contract No. NAS9-16785 were

1) the development and refinement of sampling and modeling techniques, and
2) the development and refinement of aggregation techniques.

The principle research activities associated with the development and refinement of the sampling and modeling techiques were

1) the extension of multiyear models and estimation procedures to include partial ground cover identification, and
2) the development of a procedure to determine the optimal current year sampling design as a function of previous years' results.

The major activities concerning the development and refinement of aggregation techniques were

1) the identification of statistical methodology for utilizing different weighting factors which could be assigned to the observations, and
2) the derivation of approximate variances for ground cover estimators which incorporate partially identified sampled units.

These four major activities are discussed in the next four sections respectively (Sections 3-6). Section 7 indicates some additional research results. Section 8 concludes this final report and makes a suggestion for future research.

3. EFFICIENT ACREAGE ESTIMATION USING MULTIYEAR DATA WITH BOTH PARTIALLY AND COMPLETELY IDENTIFIED SAMPLING UNITS

Each stratum at-harvest crop acreage proportion could be modeled using a regression approach with explanatory variables such as the past, present, and anticipated economic and meteorological conditions. However, the unknown form of the regression model, the large number of possible explanatory variables, and the difficulty in obtaining reasonable values for these variables makes this approach unattractive. Nevertheless, the combined effect of all of these variables is reflected in the crop acreage proportions for the stratum segments. Although it is not economical to estimate the at-harvest crop acreage proportion for every segment in the stratum, it is feasible to estimate them for a sample of segments using Landsat data (see, for example Hall and Houston (1984) and Heydorn (1984)). Hence, an alternative approach is to model the estimated at-harvest crop acreage proportion for a sample in terms of

(i) the stratum at-harvest crop acreage proportion,

(ii) stratum-wide influences which vary from year to year,

(iii) characteristics of the segment itself,

(iv) yearly influences which affect different segments differently, and

(v) the proportion of the growing season which has passed at the time the estimate is determined.

These factors may only contribute roughly additively to a transformation of the segment at-harvest crop acreage proportion and may not contribute additively to the segment proportion itself.

One specific model which is compatible with these ideas is

$$y(\hat{p}_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell} \qquad \begin{array}{l} t = 1, \ldots, T, \\ s = 1, \ldots, S, \\ \ell = 1, \ldots, L \end{array} \qquad (1)$$

where

$\hat{p}_{ts\ell}$ = 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, $\ell = 1$ could denote early season, $\ell = 2$ mid-season, and $\ell = 3$ harvest time);

$y(\hat{p}_{ts\ell})$ = a transformation of $\hat{p}_{st\ell}$;

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 independent random variables each with mean zero and variance σ_b^2;

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

$e_{ts\ell}$ = the aggregate of sampling and classification errors in the transformed data; the $e_{ts\ell}$'s are independent random variables each with mean zero.

This model is, of course, not the most general model possible. In particular, the segment effects b_s are assumed to be independent of the crop calendar time and the year. Also the departures of the transformed observations $y(\hat{p}_{ts\ell})$ on the same segment from their fixed year effects α_t and their fixed estimation time effects δ_ℓ are assumed to be positively correlated. The error terms $e_{ts\ell}$ are the composite effect of many components and need not have homogeneous variances; in particular see Heydorn (1984) for a detailed discussion of the classification error components.

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

Estimates of the stratum at-harvest crop acreage proportion are also often desired throughout the current year as well as at harvest time. For example, an early season estimate of P_T based on observations for $\ell = 1, \ldots, L$ for $t = 1, \ldots, T-1$ and only $\ell = 1$ for $t = T$ is frequently desired.

Even though the estimate $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$ of the stratum at-harvest crop acreage proportion for the current year involves only $\hat{\alpha}_T$, this estimate depends on the entire mulityear data set and not just the data from year T since the segment effects (b_s's) and systematic estimation time biases (δ_ℓ's)are assumed to be constant from year to year.

Special cases of model (1) have also been considered. For example, Chhikara et al. (1984) consider at-harvest estimates made only at harvest time, so that

their model is

$$\hat{p}_{ts} = \alpha_t + b_s + e_{ts}, \qquad t = 1, \ldots, T \text{ and } s = 1, \ldots, S.$$

For simplicity Feiveson (1984) considers only estimates of the stratum at-harvest crop acreage proportion made at harvest time during the current year; i.e.,

$$\hat{p}_{Ts} = \alpha_T + e_{Ts}, \qquad s = 1, \ldots, S.$$

When such data is not available, Feiveson (1984) utilizes historical data from agricultural reports even though previous Landsat data could also be incorporated. The methodology in both of these papers can be extended to incorporate the more general model (1).

Multiyear estimation models provide the ability to make estimates of the current year's acreage on the basis of not only the current year's sampled data, but also the previous years' sampled data. In the past such multiyear models have been developed and used when the sampled data is the proportional acreage of a single crop of interest. In such cases the use of multiyear models can easily reduce the variation in the current year's estimate to one half of what it would have been if the previous years' sampled data were ignored.

In Sielken (1981) techniques were developed and tested for estimating the acreage for a particular crop when there is only sampled data from a single year and some of the segments have been only partially identified. A segment is said to only be partially identified as opposed to completely identified if only the proportion of the segment containing some unknown percentage mixture of two or more ground covers (including the specified crop of interest) is estimated. In developing these estimations techniques consideration has been given to the following approaches:

a) maximum likelihood estimation,

b) least squares methods,

c) weighted least squares methods, and

d) a combination of a least squares ratio estimator of the specified crop's acreage percentage, say R, within the combined acreage of all crops in the mixture and a maximum likelihood estimator of the mixture acreage.

The empirical behavior of approach (d) based on the combination of the least squares ratio estimator, denoted by $\hat{R}$ say, of the specified crop's acreage percentage within the combined acreage of all crops in the mixture and a maximum likelihood estimator of the mixture acreage has been usually as good as, if not better than, approaches (a) - (c).

The following procedure is recommended for estimating a crop's current year acreage within a stratum based on both the current year's data and previous year's data when these data involve both partially and completely identified sampling units:

i) Determine the least squares ratio estimator, $\hat{R}$, for the crop of interest using the current year's partially and completely identified sampling units.

ii) Transform each multivariate segment observation into a univariate observation by combining all of the acreages for the crops involved in the mixture of crops creating the partial identification. Call this combination of crop acreages the mixture acreage.

iii) Apply the multiyear modeling and estimation procedures to the multiyear data set consisting of the observed segment mixture acreages. Let $\hat{P}_M$ denote the corresponding estimated proportion of the stratum's

current year acreage containing crops in the mixture.

iv) Estimate the stratum's current year acreage proportion for the crop of interest by the product $\hat{R} * \hat{P}_M$.

Time series or regression models can be used to augment step (i) in the above procedure if trends over time in the specified crop's ratio R are anticipated or if covariates for R can be identified.

4. Optimal Current Year Sampling Designs Based on Previous Years' Data

Here a sampling design is a plan which defines the way in which the sample of segments is to be chosen from a stratum's population of segments. An optimal design yields estimates which have optimal properties. In the past, sampling designs in support of ground cover proportion estimation have specified at the outset of the study how the sampling is to be done in each year of the study. As these designs are being implemented considerable information is gathered. For example, cloud cover may have eliminated particular observations and improved estimates of relevant variances may have become available. Such information is not incorporated in the original non-sequential design. However, a sequentially determined sampling design which allows information from previous years to influence the current year's design should produce sampling designs leading to better estimates.

The use of a multiyear mixed model weighted analysis of variance to estimate a stratum's at-harvest crop acreage proportion based on estimated proportions from sampled segments has been described in Dahm and Sielken (1980). The selection of multiyear sampling designs as described in Gbur and Sielken (1980a, 1980b, and 1981) was based on two simplifying assumptions. First, the design selection procedure did not take into account any previous sampling information

on the stratum nor did it allow sampling information obtained during the early periods of the design to affect sampling in subsequent periods. Second, in an attempt to reduce the number of competing designs to a manageable level, it was assumed that the number of segments to be sampled in each future year was the same.

Yearly changes in economic conditions, measurement techniques, equipment characteristics, and reliability requirements suggest that a more realistic approach would be to sequentially select each year's sampling pattern. Such a sequential approach would utilize the information collected from all previous years' sampling of the stratum. It would allow for the selection of a sampling pattern for each year which reflects the effects of missing observations in previous years' samples as well as changes in factors such as those mentioned above.

In a new technical report (Gbur and Sielken (1983)) a computer program called OPTDESIGN is documented which enables the user to obtain a list of the best sampling patterns for a stratum for the current year based on the segment proportion information from all previous years. Two criteria for design selection have been implemented. These are the minimization of the variance of the current year's estimated stratum transformed proportion and the minimization of the variance of the estimated change from the previous year's stratum transformed proportion.

Since the variances of the $y(\hat{p}_{ts\ell})$'s in model (1) are not necessarily equal, a weighted form of the model (1) has been used. In matrix notation this model can be expressed as

$$WY = WX\begin{bmatrix}\alpha\\ \delta\end{bmatrix} + WUb + I\varepsilon \ , \qquad (2)$$

where

$Y = [y_{111}, y_{112}, \ldots, y_{TSL}]'$,

$\alpha = [\alpha_1, \ldots, \alpha_T]'$,

$\delta = [\delta_1, \ldots, \delta_{L-1}]'$, $(\delta_L \equiv 0)$

$b = [b_1, \ldots, b_S]'$,

W = weight matrix = $[w_{ts\ell}]$,

X = design matrix for the fixed effects (α_t's and δ_ℓ's),

U = design matrix for the random effect (b_s's),

$\varepsilon = [\varepsilon_{111}, \varepsilon_{112}, \ldots, \varepsilon_{TSL}]' = We$

= vector of transformed errors.

In the weighted model (2), the estimates of α_t are obtained from the appropriate entries of $(X'W'V^{-1}WX)^{-1}\ X'W'V^{-1}Y$ and the covariance matrix of the vector $\hat{\alpha}$ is the upper left block of the matrix

$$\hat{\Sigma} = (X'W'V^{-1}WX)^{-1}\ \sigma_\varepsilon^2 ,$$

where

$$V = I + W'U'UW\gamma ,$$

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

The stratum at-harvest crop acreage proportion is estimated by $\hat{P}_t = y^{-1}(\hat{\alpha}_t)$.

In determining the optimal sampling design, it is assumed that information from years t = 1, 2, ..., T is available for the stratum under consideration. Since within season segment proportion estimates are often not available and are not particularly important in design selection, our procedure only utilizes the at-harvest segment proportion estimates made at harvest time. Therefore, the within season biases δ_ℓ in model (2) are eliminated.

The required information for OPTDESIGN for each stratum consists of

i) segment identification numbers for each segment sampled in each previous year,

ii) the final estimated weight, $\hat{w}_{tsL}$, associated with each estimated transformed acreage proportion, $y(\hat{p}_{ts\ell})$, and

iii) an estimate of the variance component ratio γ.

The estimated segment proportions $\hat{p}_{ts\ell}$ are not required for design selection, except insofar as they may be needed to calculate the estimated wieghts $\hat{w}_{tsL}$. Since the covariance matrix for each design contains the same (unknown) multiplier σ_ε^2, the particular value of σ_ε^2 does not need to be considered in the design selection process.

The optimality measures implemented in OPTDESIGN are

i) minimize $\mathrm{var}(\hat{\alpha}_{T+1})$, the variance of the estimated stratum transformed at-harvest crop acreage proportion for the current year,

ii) minimize $\mathrm{var}(\hat{\alpha}_{T+1} - \hat{\alpha}_T)$, the variance of the estimated change in the stratum transformed at-harvest crop acreage proportion from the previous year.

The minimizations in (i) and (ii) are determined over the set of all possible T+1 year designs containing the specified number of segments to be sampled in the current T+1$\underline{\text{th}}$ year and for which the parent T year design is given by the sampling history of the stratum.

OPTDESIGN is a self-contained computer program for determining the best designs according to the optimality criteria described above. It is written in Fortran and contains numerous comment cards which provide extensive internal documentation. A listing of OPTDESIGN and a flowchart of the program logic, as well as sample inputs and corresponding outputs for OPTDESIGN are given in Gbur and Sielken (1983).

For each stratum the following information is printed in the output from OPTDESIGN:

(1) Initial Information including

(a) stratum number,

(b) number of years of prior information,

(c) number of segments sampled in each previous year,

(d) number of segments to be sampled in the current year,

(e) estimate of the variance component ratio γ,

(f) weight to be assigned to all current year segments for the purpose of computing the optimality measures.

(2) For each previously sampled segment,

(a) segment label,

(b) year the segment was sampled,

(c) weight attached to that observation.

(3) A list of the NOPT best designs for the stratum for each optimization above, along with the value of the criterion for each design.

The program OPTDESIGN has been written to allow for as much flexibility as possible in the sampling history of the stratum. The only unchangeable restriction is that at least one year of prior information is required. The program will accept any positive numbers as weights and arbitrary samples sizes for each previous year in which the stratum has been sampled.

Since the weights assigned to each previously sampled segment are, from the program's viewpoint, arbitrary positive numbers, they can be used to reflect many different factors. The weights need not be computed solely as functions of the estimated segment proportions. They could be used to account for such factors as changes in measurement techniques, classification algorithms, and equipment characteristics as well as factors such as differences in the level of difficulty of classification for the AI, number and quality of the

set of "photographs" used to obtain the estimate, and differences in AI personnel.

The current version of the program assumes that the weight matrix is diagonal. However, the data input format can be easily modified to allow for arbitrary nonnegative weight matrices.

The sample sizes for previous years sampling are arbitrary positive integers. This allows for differences in sample sizes caused by factors such as missing observation in one or more years, budgetary changes, and the targeting of selected strata for more intensive sampling in certain years.

It is conceivable that the sampling history of a stratum contains T* years in which no sampling occurred. Since OPTDESIGN requires the previous years to be labeled as 1, 2, ..., T, the years in which the stratum was sampled could be numbered consecutively as 1, 2, ..., T-T* and all years' information utilized.

"Ground truth" data could be combined with the stratum's sampling history. The weights for such "ground truth" estimates should reflect any differences in their quality and variability as compared to the remotely sensed segment estimates.

The multiyear model (2) on which the program OPTDESIGN is based on relatively simple. Additional fixed effects and covariates could be incorporated to improve the estimates. Modification of OPTDESIGN to reflect the expanded model can be achieved in a straightforward manner by substitution of a new subroutine for computing the fixed effects design matrix X. The inclusion of additional random effects in the model would require more extensive modification of the program, but could also be accomplished.

5. INCORPORATING WEIGHTING FACTORS INTO THE STATISTICAL METHODLOGY FOR MULTIYEAR DATA

Current multiyear estimation methodology uses observations as if their variability was only dependent upon the true underlying proportion being estimated. In practice, however, the variability of an observation is dependent upon many other factors; for example, the season in which the observation is made, the amount of previous satellite imagery available, the quality of that imagery, the satellite being used, the "closeness" of the spatial-spectoral-temporal patterns observed in the sampled units to their classical prototypes (say for corn, soybeans, pasture, forest, etc.). Better use of the observations can be made in aggregation if greater weight can be given to the more precise observations and lesser weight given to less precise observations. Hence, better aggregation estimates should be obtainable if the precision of the observations is more accurately assessed and then incorporated into the multiyear area estimation techniques. This is particularly important in the multiyear environment where satellite technology, analyst and computer methodologies, etc. are hopefully improving from year to year.

The suggested approach is to characterize precision or confidence in the observations in terms of their variances and weight the observations proportionately to the inverse of their variances. The Var $(\hat{p}_{ts\ell})$ can be approximated on the basis of information such as

(i) the type of satellite being used,

(ii) the sharpness of the satellite imagery,

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

(iv) the number of satellite images successfully obtained by the time the segment proportion is estimated.

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

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

(vii) the physical characteristics of the segment.

In addition, recognizable segment characteristics which make it either easier or harder to estimate the segment's crop proportion can be incorporated. Obvious differences in the amount of information going into the $\hat{p}_{ts\ell}$'s can also be reflected. These latter differences can be due to the estimation times themselves as well as due to loss of satellite imagery from cloud cover, machine failure, etc.

The statistical procedures for area estimation documented in Dahm and Sielken (1981) can easily incorporate as input both the observation and its weight (confidence measure). The weighted form of the multiyear model (1) is the model (2) discussed in Section 4. Detailed procedures for implementing the statistical analyses associated with model (2) are given in Dahm and Sielken (1981).

The advantages and disadvantages of doing weighted analyses of linear models as opposed to unweighted analyses when the observations have unequal reliability or variances is well documented in the statistical literature (see, for example, Draper and Smith (1981), Kleijnen (1981), and Scheffe (1959)).

6. VARIANCES FOR GROUND COVER ESTIMATORS INCORPORATING PARTIALLY IDENTIFIED SAMPLED UNITS

In the past, large scale ground cover area estimation techniques have been developed for a single year's data which may include partially identified sampled units. In order for these techniques to support aggregation activities some statement of the uncertainty of the estimate must be conveyed. This is

best handled by providing an estimate of the variance of the ground cover area estimator. Within a large homogeneous area (called a stratum) a sample of segments (currently 5 by 6 nautical mile rectangles) is observed. These observations are collected as satellite imagery and are available for a period of a few years and at several times during the crop growing seasons. Using these segment acreages segment proportions for several crops are estimated. Difficulties in distinquishing between crops leads to partially identified segments as opposed to completely identified segments. Herein, a segment will be considered to be planted in two major crops, crop 1 and crop 2, the remainder of the segment will be pooled under crop 3, "other". When crop 1 and crop 2 are distinguishable the segment is completely identified. If it is not possible to distinguish them, the segment is partially identified. Both types of segments can be combined to estimate a crop's proportional acreage in the stratum.

Methods of estimating individual crop acreage using a mixture of completely and partially identified segments have been discussed in Sielken (1981) and (1982).

The assumption used in Sielken (1982) is that the number of acreage units harvested in a segment follows a multinomial distribution. An acreage unit will be hereafter referred to as a block. The number of blocks within a segment planted in crop i is denoted by Y_i, and the total number of blocks in a segment is denoted by N. Under the multinomial assumption the Y_i's have the distribution

$$P(Y_1 = y_1, Y_2 = y_2) = (N!/(y_1!y_2!y_3!))\, p_1^{y_1}\, p_2^{y_2}\, p_3^{y_3} ,$$

when

$$N = y_1 + y_2 + y_3$$

and p_i is the at-harvest proportion of the stratum planted in crop i. This assumption is correct if every decision maker acts independently and allocates each block independently to crop 1, crop 2, or "other" with probabilities p_1, p_2, and $p_3 = (1 - p_1 - p_2)$ respectively.

A random sample of J segments is to be observed. Let Y_{ij} = number of blocks in segment j containing crop i i = 1, 2, 3 and j = 1, ..., J. Assume the segments j = 1, ..., J_C are completely identified and segments j = J_C + 1, ..., $J_C + J_P$ are partially identified. Therefore, $J = J_C + J_P$. Let

$$Z_{Ci} = \sum_{j=1}^{J_C} Y_{ij}, \; i = 1, 2, 3,$$

$$Z_{P12} = \sum_{j=J_C+1}^{J_P} (Y_{1j} + Y_{2j}),$$

$$Z_{P3} = \sum_{j=J_C+1}^{J_P} Y_{3j}.$$

Thus, Z_{Ci} is the total number of blocks containing crop i in the completely identified segments. The total number of blocks containing either crop 1 or crop 2 in the partially identified segments is Z_{P12}. The total number of blocks containing crop 3 in the partially identified segments is Z_{P3}. The total number of blocks completely (partially) identified is $N_C(N_P)$. Thus, if N is the number of blocks in one segment,

$$N_C = J_C N,$$

$$N_P = J_P N.$$

As noted in Vidart and Sielken (1984) the results from Hocking and Oxspring (1971) can be used to show that the maximum likelihood estimators are

$$\hat{p}_1 = [Z_{C1}/(Z_{C1} + Z_{C2})][(Z_{C1} + Z_{C2} + Z_{P12})/(N_C + N_P)].$$

$$\hat{p}_2 = [Z_{C2}/(Z_{C1} + Z_{C2})][(Z_{C1} + Z_{C2} + Z_{P12})/(N_C + N_P)].$$

and

$$\hat{p}_3 = \widehat{1-p_1-p_2} = 1-\hat{p}_1-\hat{p}_2 = (Z_{C3} + Z_{P3})/(N_C + N_P).$$

The form of these estimates is fairly intuitive since

$\hat{p}_1$ = [Estimated proportion of crop 1 and 2 that is crop 1 in the completely identified segments] x

[Estimated proportion of crop 1 and 2 in all the segments].

The asymptotic variances (AV) of these estimates are

$$AV(\hat{p}_1) = p_1(1-p_1)/N_C-[p_1^2p_3N_P]/[N_C(N_C + N_P)(p_1 + p_2)], \text{ and}$$

$$AV(\hat{p}_2) = p_2(1-p_2)/N_C-[p_2^2p_3N_P]/[N_C(N_C + N_P)(p_1 + p_2)].$$

The second term of these expressions shows the improvement obtained by using the partially identified segment. After some simplification, the asymptotic variances can b rewritten as

$$AV(\hat{p}_1) = p_1(1-p_1)/(N_C+N_P) + [N_Pp_1p_2]/[N_C(N_C+N_P)(p_1+p_2)],$$

$$AV(\hat{p}_2) = p_2(1-p_2)/(N_C+N_P) + [N_Pp_1p_2]/[N_C(N_C+N_P)(p_1+p_2)],$$

and for $p_3 = 1-p_1-p_2$

$$AV(\hat{p}_3) = p_3(1-p_3)/(N_C+N_P).$$

A computer program has been implemented to test the accuracy of these asymptotic variances when N_C and N_P are not both arbitrarily large. Samples with the prescribed number of partially and completely identified segments were simulated following a multinomial distribution. For each sample the maximum likelihood (ML) estimates of the p's were computed. Finally the sample variance of these ML estimates were compared to the asymptotic variances. The details of the evaluation of the applicability of the asymptotic variance formulas to small sample sizes are given in Vidart and Sielken (1984). The conclusion was that the asymptotic variance can be used as a good approximation of the actual variance under the multinomial decision process even for relatively small sample sizes.

One Monte Carlo study of the empirical behavior of the crop acreage estimation procedure utilizing partially identified data was already available in Sielken (1982). The sample variances of the maximum likelihood estimates of the p_i's in Sielken (1982) can be compared to the asymptotic variances under the multinomial assumption. In order to computer the asymptotic variances of the maximum likelihood estimators of the p's, the number of blocks N contained in a segment must be determined. This information is not available since the CAMS estimates are given in percentages rather than blocks. Therefore, N was estimated separately for i = 1, 2, and 3 and for different combinations of J_C and J_P. For a particular crop the estimated value of N is nearly the same for the different combinations of J_C and J_P. However the value of N seems to vary with the crop. In other words, the theoretical variances under the multinomial decision process differed markedly from the observed sampled variances. Consequently, the multinomial decision process is not applicable.

Since the multinomial assumption does not hold, another decision process must be considered. The estimated N values suggest that some crops are planted in a larger "standard area" than others. The standard area of a "large" crop has more blocks than a "small" crop does. Crop 3 "other" appeared to be a "large" crop and crop 1 a "small" crop. This suggested the following approach. A block will now denote a particular fixed number of acres corresponding to the smallest decision possible. Let K_i denote the theoretical number of blocks in a standard area of crop i. This conceptualization envisions crop i being planted only in integer multiples of K_i blocks.

If p_i is equal to the overall proportion of the stratum planted with crop i, then this alternative decision process independently allocates each K_3 blocks of acreage according to the following sequential procedure:

1. Allocate K_3 blocks to crop 3 with probability p_3.
2. If the K_3 blocks are not allocated to crop 3, then allocate those K_3 blocks to crops 1 and 2 as follows:

   2-1. Allocate K_2 blocks to crop 2 with probability

   $$p_2' = p_2/(1-p_3).$$

   2-2. If these K_2 blocks are not allocated to crop 2 during step 2-1, then allocate these K_2 blocks to crop 1.

Obviously, it is assumed that $K_1 = K_2$, N is an integer multiple of K_3, and K_3 is an integer multiple of K_1. For simplicity K_1 and K_2 are defined to be 1 block and K_3 is renamed K. The resulting decision process can be summarized as

1. Allocate K blocks to crop 3 with probability p_3.

2. If these K blocks are not allocated to crop 3 during step 1, allocate those K blocks to crop 1 and 2 using a binomial decision process with probabilities p_1' and $p_2' = 1-p_1'$ where

$$p_1' = p_1/(1-p_3).$$

3. Repeat steps 1 and 2 until all N blocks are allocated.

This alternative decision process will be referred to as the KDP. A group of K blocks will be called a superblock. The particular case where K = 1 is the multinomia decision process, 1DP.

The parameters in KDP include N, the number of blocks in a segment, and K, the number of blocks in a superblock, as well as p_1, p_2, and p_3. In Vidart and Sielken (1984) it is shown that the maximum likelihood estimates for the p_i's under the KDP are the same as under 1DP and do not depend upon N or K. However, the asymptotic variances of the $\hat{p}_i$'s do depend upon N and K which are both unknown. In Vidart and Sielken (1984) estimators for N, K, and approximations for the variances of the $\hat{p}_i$'s are derived. Also a simulation check on the approximate expressions for the variances of the $\hat{p}_i$'s is reported there. The sample variances of the $\hat{p}$'s were very close to their approximating expressions.

In Vidart and Sielken (1984) the KDP is also extended to the situation where the sampling units have variable sizes instead of the constant size typified by 5x6 nautical mile segments. Such a situation could easily occur if the sampling units were political subdivisions such as counties.

One objective of the contract research was to determine the improvement brought about through the use of the KDP instead of the 1DP when CAMS estimates are studied. Some improvement in the prediction of the variance of the crop acreage estimators is achieved by considering the new decision process. For

fairly large samples, typically 50 segments, the use of the KDP as opposed to the 1DP leads to an improvement in the prediction of the variances of the ML estimators of the crop proportions based on CAMS data. For smaller, more realistic size samples, typically 5 segments, the variance estimation techniques were not very accurate. However, the empirical results indicate a better performance under the KDP, than under the 1DP. The variance estimates under KDP have a distribution with more spread but centered much closer to the sample variance than the corresponding distribution under 1DP. The greatest overall improvement is associated with the estimated variance for the smallest crop (i.e., the crop planted in the smallest size blocks) while the other estimated variances improve just slightly overall.

7. ADDITIONAL RESEARCH RESULTS

A special issue of Communications in Statistics concerning statistical applications at NASA is being prepared under the coordination of Dr. Raj Chhikara, Lockheed Engineering and Management Services Company, Inc. R. L. Sielken, Jr. and E. E. Gbur have prepared a contribution entitled "Multiyear, Through the Season Crop Acreage Estimation Using Estimated Acreage in Sample Segments" for that special issue. That contribution has been refereed and accepted. A copy of that paper is attached to this final report.

Some additional research has been done on the empirical behavior of the transformations $y(\hat{p})$ used in conjunction with model (1) and (2). The simplest transformation $y(\hat{p})$ of the estimated segment crop acreage proportion p to use in (1) or (2) is the identity transformation

$$y(\hat{p}) = \hat{p}.$$

However, it is very doubtful that the additive model (1) would hold for $y(\hat{p}) = \hat{p}$ particularly if the $\hat{p}$'s exhibit a large variation within the stratum. On the other hand a multiplicative model for $\hat{p}$ may be more reasonable. For instance, if

(i) 30% of the stratum contains wheat at the time wheat is harvested in year t;

(ii) the s-th segment's wheat acreage proportion averages only 80% of the stratum's wheat acreage proportion 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

$$\hat{p}_{ts\ell} = (.30)\ (.80)\ (.70)\ (1.10).$$

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

$$y(\hat{p}_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell}$$

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

The logit transformation,

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

is another useful transformation which approximately converts a multiplicative model for $\hat{p}$ into an additive model for $y(\hat{p})$. A small advantage of the logit transformation is that it guarantees that

$$0 \leqslant \hat{P}_T = y^{-1}(\hat{\alpha}_T) \leqslant 1\ ,$$

whereas the logarithmic transformation only guarantees

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

and the identity transformation makes no guarantees.

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

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

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

(iii) confidence intervals on P_T .

These derivations are all similar and are based upon Taylor series approximations (statistical differentials). For instance, if $y(\hat{p}) = \ell n(\hat{p})$, then

$$\hat{P}_T = y^{-1}(\hat{\alpha}_T) = y^{-1}(\hat{\alpha}_T) + (\hat{\alpha}_T - \alpha_T) \left[\frac{dy^{-1}(\hat{\alpha}_T)}{d\hat{\alpha}_T}\right]_{\hat{\alpha}_T = \alpha_T}$$

$$= P_T + (\hat{\alpha}_T - \alpha_T)P_T \ ,$$

so that

$$\text{MSE } (\hat{P}_T) \equiv E\,[(\hat{P}_T - P_T)^2] \cong P_T^2 \ \text{Var}(\hat{\alpha}_T) \ .$$

A small simulation study was conducted in order to observe the empirical behavior of the estimators of the components of models (1) and (2) (namely, $\hat{\sigma}_b^2$, $\hat{\sigma}_\varepsilon^2$, $\hat{\gamma} = \hat{\sigma}_b^2/\hat{\sigma}_\varepsilon^2$) and the estimators of the stratum's crop acreage proportions over the years $t = 1, \ldots, T$ (namely, $y^{-1}(\hat{\alpha}_1), \ldots, y^{-1}(\hat{\alpha}_T)$). In this simulation study each of the three transformations (identity, log, and logit) were used to generate a random data set corresponding to each of four underlying situations. Each of the twelve data sets was analyzed three times: once using the identity transformation, once using the log transformation, and once using the logit transformation. Thus each data set was analyzed once using the "correct" transformation and twice using an "incorrect" transformation. Since the "correct"

transformation is unknown in practice, the simulation study provided a limited evaluation of the sensitivity of the estimators to the "correctness" of the transformation being used in the statistical analysis. All underlying simulated situations involved

i) 3 years with the stratum crop acreage proportions being 0.6, 0.6, and 0.4 for years 1, 2, and 3 respectively;

ii) 3 seasons with the seasonal biases being $\delta_1 = -0.3$, -0.1, and $\delta_3 = 0$ respectively;

iii) 10 segments observed in each season in each year; and

iv) no partial identification.

The variance among segments σ_b^2, variance within segments σ_ε^2, and $\gamma = \sigma_b^2/\sigma_\varepsilon^2$ took on different values in each data set; the four combinations were ($\sigma_b^2 = 0.0004$, $\sigma_\varepsilon^2 = 0.001$, $\gamma = 0.4$), ($\sigma_b^2 = 0.0004$, $\sigma_\varepsilon^2 = 0.0001$, $\gamma = 4$), ($\sigma_b^2 = 0.004$, $\sigma_\varepsilon^2 = 0.001$, $\gamma = 4$), and ($\sigma_b^2 = 0.004$, $\sigma_\varepsilon^2 = 0.0001$, $\gamma = 40$). The estimators of σ_b^2, σ_ε^2, and γ are shown in Tables 1-4 for each of the four data sets. Also in these tables are the estimators and approximate 90% confidence intervals for the stratum's crop acreage proportions P_1, P_2, and P_3 for the three years.

In the simulation study the estimators of σ_b^2, σ_ε^2, and γ were not precise. However, these estimators are usually of only secondary importance. The primary conclusion from the simulation studies was that the estimators and 90% confidence intervals for the stratum's crop acreage proportion which are of primary importance behaved quite well and were relatively robust with respect to the transformation used.

ORIGINAL PAGE IS
OF POOR QUALITY

Table 1. Situation No. 1 in the Simulation Study of the Multiyear Model Estimators and their Robustness to the Transformations Involved

	Transformation used in Data Generation								
	Identity			Log			Logit		
Transformation used in Model Fitting:	Id.	Log	Logit	Id.	Log	Logit	Id.	Log	Logit
Quality Being Estimated									
Stratum Proportions									
$P_1 = .6$	.594	.605	.606	.587	.595	.589	.588	.593	.591
$P_2 = .6$	.598	.612	.610	.592	.600	.594	.597	.603	.599
$P_3 = .4$	.396	.348	.373	.414	.397	.409	.400	.382	.396
90% Confidence Intervals on									
$P_1 = .6$	[.58,.61]	[.57,.64]	[.58,.63]	[.57,.60]	[.58,.60]	[.58,.60]	[.57,.60]	[.57,.61]	[.57,.61]
$P_2 = .6$	[.58,.61]	[.57,.65]	[.59,.63]	[.58,.60]	[.59,.61]	[.58,.61]	[.58,.61]	[.58,.62]	[.58,.62]
$P_3 = .4$	[.38,.41]	[.32,.38]	[.35,.39]	[.40,.43]	[.39,.40]	[.40,.42]	[.38,.42]	[.36,.39]	[.38,.41]
Variances									
$10{,}000 * \sigma_b^2 = 4$	3.4	69.0	36.0	2.0	6.7	7.7	3.9	19.4	17.5
$1{,}000 * \sigma_\varepsilon^2 = 1$	0.8	9.7	1.0	1.3	0.7	0.3	3.0	3.4	0.7
$\gamma = \sigma^2/\sigma_\varepsilon^2 = 0.04$	0.5	0.7	36.0	0.2	0.9	2.9	0.1	0.6	2.3

ORIGINAL PAGE IS
OF POOR QUALITY

Table 2. Situation No. 2 in the Simulation Study of the Multiyear Model Estimators and their Robustness to the Transformations Involved

	Transformation used in Data Generation								
	Identity			Log			Logit		
Transformation used in Model Fitting:	Id.	Log	Logit	Id.	Log	Logit	Id.	Log	Logit
Quality Being Estimated									
Stratum Proportions									
$P_1 = .6$	.597	.608	.608	.589	.597	.591	.593	.599	.598
$P_2 = .6$	.598	.610	.609	.591	.599	.593	.596	.602	.596
$P_3 = .4$	.397	.349	.374	.415	.398	.411	.403	.385	.402
90% Confidence Intervals on									
$P_1 = .6$	[.59,.61]	[.58,.64]	[.59,.63]	[.58,.60]	[.59,.60]	[.59,.60]	[.59,.60]	[.59,.61]	[.59,.61]
$P_2 = .6$	[.59,.61]	[.58,.64]	[.59,.63]	[.58,.60]	[.59,.61]	[.59,.59]	[.59,.60]	[.59,.61]	[.59,.60]
$P_3 = .4$	[.39,.41]	[.33,.37]	[.33,.39]	[.41,.42]	[.39,.40]	[.40,.42]	[.40,.41]	[.38,.39]	[.40,.41]
Variances									
$10{,}000 * \sigma_b^2 = 4$	2.8	63.0	31.0	1.2	3.2	4.3	1.0	5.9	6.2
$10{,}000 * \sigma_\varepsilon^2 = 1$	0.8	87.0	7.9	6.2	0.7	0.9	3.5	6.7	0.8
$\gamma = \sigma^2/\sigma_\varepsilon^2 = 4$	3.7	0.7	3.9	0.2	4.4	4.1	0.3	0.9	7.8

Table 3. Situation No. 3 in the Simulation Study of the Multiyear Model Estimators and their Robustness to the Transformations Involved

	Transformation used in Data Generation								
	Identity			Log			Logit		
Transformation used in Model Fitting:	Id.	Log	Logit	Id.	Log	Logit	Id.	Log	Logit
Quality Being Estimated									
Stratum Proportions									
$P_1 = .6$	.59	.59	.60	.59	.59	.59	.59	.59	.59
$P_2 = .6$	.59	.60	.61	.59	.60	.59	.59	.60	.60
$P_3 = .4$	.39	.35	.37	.41	.40	.41	.40	.38	.39
90% Confidence Intervals on									
$P_1 = .6$	[.56,.62]	[.54,.64]	[.56,.64]	[.57,.61]	[.57,.61]	[.57,.61]	[.56,.61]	[.56,.62]	[.56,.61]
$P_2 = .6$	[.56,.62]	[.55,.65]	[.57,.64]	[.57,.61]	[.58,.62]	[.57,.61]	[.57,.62]	[.57,.63]	[.57,.62]
$P_3 = .4$	[.36,.42]	[.32,.38]	[.33,.40]	[.39,.43]	[.38,.41]	[.39,.43]	[.38,.42]	[.36,.40]	[.37,.42]
Variances									
$10{,}000 * \sigma_b^2 = 4$	2.5	20.7	17.0	0.8	3.3	3.3	1.0	4.6	4.3
$10{,}000 * \sigma_\varepsilon^2 = 1$	0.9	9.5	1.2	1.6	0.7	0.3	3.0	3.5	0.7
$\gamma = \sigma^2/\sigma_\varepsilon^2 = 4$	2.8	2.2	15.0	0.5	4.5	11.0	0.3	1.3	5.8

ORIGINAL PAGE IS OF POOR QUALITY

Table 4. Situation No. 4 in the Simulation Study of the Multiyear Model Estimators and their Robustness to the Transformations Involved

	Transformation used in Data Generation								
	Identity			Log			Logit		
Transformation used in Model Fitting:	Id.	Log	Logit	Id.	Log	Logit	Id.	Log	Logit
Quality Being Estimated									
Stratum Proportions									
$P_1 = .6$	.59	.59	.61	.59	.60	.59	.59	.60	.59
$P_2 = .6$	.59	.59	.61	.59	.60	.59	.59	.60	.60
$P_3 = .4$	.39	.35	.37	.41	.40	.41	.40	.38	.40
90% Confidence Intervals on									
$P_1 = .6$	[.56,.62]	[.54,.64]	[.57,.64]	[.57,.60]	[.58,.61]	[.57,.60]	[.57,.60]	[.57,.62]	[.58,.61]
$P_2 = .6$	[.56,.62]	[.55,.65]	[.57,.64]	[.57,.60]	[.58,.61]	[.57,.61]	[.57,.61]	[.58,.62]	[.58,.61]
$P_3 = .4$	[.36,.42]	[.32,.38]	[.33,.41]	[.40,.43]	[.38,.41]	[.39,.43]	[.39,.42]	[.37,.40]	[.38,.41]
Variances									
$10{,}000 * \sigma_b^2 = 4$	2.4	20.1	16.7	0.7	2.9	2.9	0.7	3.1	2.9
$10{,}000 * \sigma_\varepsilon^2 = 1$	2.0	89.6	10.0	8.8	0.7	1.4	3.9	8.3	0.7
$\gamma = \sigma^2/\sigma_\varepsilon^2 = 40$	12.0	2.3	16.7	0.8	39.0	21.0	1.7	3.7	39.0

8. Concluding Remarks

The primary purpose of this research effort has been to identify and develop statistical procedures for large area assessments using both satellite and conventional data. Crop acreages, other ground cover indices, and measures of change have been the principal characteristics of interest. The characteristics are capable of being estimated from samples collected possibly from several sources (different satellites, aerial surveys, ground measurements, etc.) at varying times (different years, seasons, crop calendar days, etc.) with different levels of identification (for example, vegetation, crops, summer crops, corn). The overall objective has been to be able to obtain the most precise large area estimates from multiyear samples including possibly partially identified sample units. Included in this research have been

a) extensions of multiyear analysis techniques to include partially identified samples, and

b) the determination of the best current year sampling design corresponding to a given sampling history,

c) determination and utilization of observation weights reflecting the precision or confidence in each observation, and

d) quantification of the variation in estimates incorporating partially identified samples.

The development and utilization of observation weights reflecting the observation's precision may be a very fruitful area for additional research.

Referenecs

Chhikara, R.S. and A.H. Feiveson (1982). A sample survey of global wheat acreage using satellite (LANDSAT) Data. Sankhya: The Indian Journal of Statistics, B, 44, 304-329.

Chhikara, R.S., C.R. Hallum and T.G. Lycthuan-Lee (1984). Stratification and sampling design for Landsat crop surveys. To appear in Communications in Statistics.

Dahm, P.F. and R.L. Sielken, Jr. (1981). Multiyear estimation of the at-harvest crop acreage proportion: Methodology and implementation. Technical Report S-20, Institute of Statistics, Texas A&M University.

Draper, N.R. and H. Smith (1981). Applied Regression Analysis. John Wiley and Sons, Inc., New York.

Feiveson, Alan H. (1984). Weighted ratio estimation in agricultural surveys. To appear in Communications in Statistics.

Gbur, E.E. and R. L. Sielken, Jr. (1980). Optimal rotation designs for multi-year estimation, I: Unweighted estimation. Technical Report S-18, Institute of Statistics, Texas A&M University.

Gbur, E.E. and R.L. Sielken, Jr. (1980). Optimal rotation designs for multi-year estimation, II: Weighted estimation. Technical Report S-19, Institute of Statistics, Texas A&M University.

Gbur, E.E. and R.L. Sielken, Jr. (1981). Missing observations in multiyear rotation sampling designs. Technical Report S-22, Institute of Statistics, Texas A&M University.

Gbur, E.E. and R.L. Sielken, Jr. (1983). Implementation of the multiyear model methodology: Sequential selection of optimal sampling designs. Technical Report S-23, Department of Statistics, Texas A&M University.

Hall, F.G. and A.G. Houston (1984). Use of satellite data in agricultural surveys. To appear in Communications in Statistics.

Heydorn, Richard P. (1984). Using satellite remote sensing data to determine crop proportions in a sampling unit. To appear in Communications in Statistics.

Hocking, R.R. and H.H. Oxspring (1971). Maximum likelihood estimation with incomplete multinomial data. Journal of the American Statistical Association, 66, 65-70.

Kleijnen, J., R. Brent, and R. Brouwers (1981). Samll-sample behavior of weighted least squares in experimental design applications. Communications in Statistics, B, 10:303-313.

Proceedings of Technical Sessions--The LACIE Symposium (1978). JSC-16105, Vols. 1 and 2, NASA/JSC, Houston, Texas.

Scheffe, H. (1959). The Analysis of Variance. John Wiley and Sons, Inc., New York.

Sielken, R.L., Jr. (1981). Incorporating partially identified sample segments into NASA acreage estimation procedures: Estimates using only observations from the current year. Technical Report S-21, Institute of Statistics, Texas A&M University.

Sielken, R.L., Jr. (1982). Incorporating partially classified sample segments into NASA acreage estimation procedures. American Statistical Association (1982) Proceedings of the Section on Survey Research Methods, 63-69.

ATTACHMENT

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

Robert L. Sielken, Jr.
and
Edward E. Gbur

Institute of Statistics, Texas A&M University
College Station, Texas

Key Words and Phrases: Mixed models; remote sensing; sampling; weighted least squares.

ABSTRACT

Large scale crop surveys can be made frequently and inexpensively during a crop growing season using Landsat data. A crop's estimated at-harvest acreage in a stratum can be estimated from the crop's estimated at-harvest acreage in a small sample of the stratum's segments. The stratum estimate can utilize Landsat imagery obtained during the current crop growing season and in previous years. A mixed effects analysis of variance model is used to generate a weighted least squares estimate of the stratum at-harvest acreage proportion for the current year. Similar Landsat based stratum crop proportion estimates can be combined with historical information on non-sampled (or unsuccessfully sampled) strata to provide crop acreage estimates for large regions. These regional estimates of the at-harvest acreage can be determined early in the crop growing season, at different intermediate points, and at harvest time.

1. INTRODUCTION

Agriculture and other renewable resources can be economically inventoried over large areas using aerospace remote sensing techniques. In particular, the surface area devoted to a specific resource in a large region is especially amenable to aerospace estimation. Such resources could be as broadly defined as agriculture, forest, water, snow cover, etc. or as specifically defined as summer crops or corn. These area estimates can be combined with other measures such as estimated yield per acre to obtain production estimates. Once the appropriate estimation methodology has been successfully implemented, the successive estimates are very economical, so that frequent inventories are realistically obtainable.

During 1975-1977 NASA in conjunction with the USDA conducted the Large Area Crop Inventory Experiment (LACIE) to illustrate the potential capabilities of aerospace remote sensing techniques. This pioneering effort also served to remove many of the obstacles for future applications. A summary of the experiment is given in the proceedings of the LACIE Symposium (1979). The target resource in LACIE was the wheat acreage and production in the U. S. Great Plains.

During the transition years 1977-1979 and during 1979-1983 under the recently-terminated AgRISTARS (Agriculture and Resources Inventory Surveys through Aerospace Remote Sensing) program several advances were made in satellite imagery technology, data processing, and statistical methodologies. In addition, target resources were expanded to include other crops and other countries, as well as non-crop resources.

This paper focuses on the statistical methodology for estimating a particular resource's acreage proportion in a large region at a specified point in time using the estimated resource acreage proportion in a sample of smaller areas. It will be assumed that

(i) the resource is a crop,

(ii) the specified time point of interest is the harvest

time for the crop,

(iii) the sample areas are all the same size (a 5x6 nautical mile rectangle called a segment), and

(iv) the sample segments are relatively "small" compared to the homogeneous region (stratum) of interest.

Also, it is assumed that in each year of a multiyear period a sample of segments is selected. The composition of the sample may vary year to year. In each year each sample segment's at-harvest crop acreage proportion is estimated at one or more times during the crop growing season. The number of estimates is not necessarily the same for all sample segments in a year and is not necessarily the same for each year. Obviously, this paper is focusing on only one part of a much larger problem. The region of concern herein is really just one stratum in a stratified sample survey of a country or the world (see, for example, Chhikara and Feiveson (1982)). The size of the sample segment is assumed to be predetermined (see Chhikara and Feiveson (1982) and in this issue Chhikara et al. (1984)). Also, since the same segments do not have to be in the sample every year, there is an interesting associated problem of determining an optimal multi-year sampling design (see Chhikara et al. (1984) and the technical reports listed in the bibliography). The papers by Heydorn (1984) and Hall and Houston (1984) in this issue discuss the determination of the sample segment's estimated at-harvest crop acreage proportion. Finally, the estimates arising from the statistical methodology in this paper can be input to procedures for aggregating acreage over several regions and combining acreage estimates with yield estimates to obtain production estimates. The paper by Feiveson (1984) in this issue addresses these latter needs.

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 his ideas have frequently stimulated his co-workers' efforts. The seeds for many of the sampling and modeling techniques utilized in several of the papers in this issue were sown by him.

2. BASIC MODEL FOR MULTIYEAR ESTIMATION

Each stratum at-harvest crop acreage proportion could be modeled using a regression approach with explanatory variables such as the past, present, and anticipated economic and meteorological conditions. However, the unknown form of the regression model, the large number of possible explanatory variables, and the difficulty in obtaining reasonable values for these variables makes this approach unattractive. Nevertheless, the combined effect of all of these variables is reflected in the crop acreage proportions for the stratum segments. Although it is not economical to estimate the at-harvest crop acreage proportion for every segment in the stratum, it is feasible to estimate them for a sample of segments using Landsat data (see, for example Hall and Houston (1984) and Heydorn (1984), both in this issue). Hence, an alternative approach is to model the estimated at-harvest crop acreage proportion for a sample segment in terms of

(i) the stratum at-harvest crop acreage proportion,

(ii) stratum-wide influences which vary from year to year,

(iii) characteristics of the segment itself,

(iv) yearly influences which affect different segments differently, and

(v) the proportion of the growing season which has passed at the time the estimate is determined.

These factors may only contribute roughly additively to a transformation of the segment at-harvest crop acreage proportion and may not contribute additively to the segment proportion itself.

One specific model which is compatible with these ideas is

$$y(\hat{p}_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell} \qquad \begin{array}{l} t = 1, \ldots, T, \\ s = 1, \ldots, S, \\ \ell = 1, \ldots, L \end{array} \qquad (1)$$

where

$\hat{p}_{ts\ell}$ = 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, $\ell = 1$ could denote early season, $\ell = 2$ mid-season, and $\ell = 3$ harvest time);

$y(\hat{p}_{ts\ell})$ = a transformation of $\hat{p}_{ts\ell}$;

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 independent random variables each with mean zero and variance σ_b^2;

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

$e_{ts\ell}$ = the aggregate of sampling and classification errors in the transformed data; the $e_{ts\ell}$'s are independent random variables each with mean zero.

This model is, of course, not the most general model possible. In particular, the segment effects b_s are assumed to be independent of the crop calendar time and the year. Also the departures of the transformed observations $y(\hat{p}_{ts\ell})$ on the same segment from their fixed year effects α_t and their fixed estimation time effects δ_ℓ are assumed to be positively correlated. The error terms $e_{ts\ell}$ are the composite effect of many components and need not have homogeneous variances; in particular see Heydorn (1984) for a detailed discussion of the

classification error components.

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

Estimates of the stratum at-harvest crop acreage proportion are also often desired throughout the current year as well as at harvest time. For example, an early season estimate of P_T based on observations for $\ell = 1, \ldots, L$ for $t = 1, \ldots, T-1$ and only $\ell = 1$ for $t = T$ is frequently desired.

Even though the estimate $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$ of the stratum at-harvest crop acreage proportion for the current year involves only $\hat{\alpha}_T$, this estimate depends on the entire multiyear data set and not just the data from year T since the segment effects (b_s's) and systematic estimation time biases (δ_ℓ's) are assumed to be constant from year to year.

Special cases of model (1) have also been considered. For example, Chhikara et al. (1984) consider at-harvest estimates made only at harvest time, so that their model is

$$\hat{p}_{ts} = \alpha_t + b_s + e_{ts}, \qquad t = 1, \ldots, T \text{ and } s = 1, \ldots, S.$$

For simplicity Feiveson (1984) considers only estimates of the stratum at-harvest crop acreage proportion made at harvest time during the current year; i.e.,

$$\hat{p}_{Ts} = \alpha_T + e_{Ts}, \qquad s = 1, \ldots, S.$$

When such data is not available, Feiveson (1984) utilizes historical data from agricultural reports even though previous Landsat data could also be incorporated. The methodology in both of these papers can be extended to incorporate the more general model (1).

3. TRANSFORMATIONS OF THE ESTIMATED SEGMENT PROPORTIONS

The simplest transformation $y(\hat{p})$ of the estimated segment crop acreage proportion $\hat{p}$ to use in (1) is the identity transformation

$$y(\hat{p}) = \hat{p}.$$

However, it is very doubtful that the additive model (1) would hold for $y(\hat{p}) = \hat{p}$ particularly if the $\hat{p}$'s exhibit a large variation within the stratum. On the other hand a multiplicative model for $\hat{p}$ may be more reasonable. For instance, if

(i) 30% of the stratum contains wheat at the time wheat is harvested in year t;

(ii) the s-th segment's wheat acreage proportion averages only 80% of the stratum's wheat acreage proportion 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

$$\hat{p}_{ts\ell} = (.30)\ (.80)\ (.70)\ (1.10).$$

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

$$y(\hat{p}_{ts\ell}) = \alpha_t + b_s + \delta_\ell + e_{ts\ell}$$
$$= \ell n(.30) + \ell n(.80) + \ell n(.70) + \ell n(1.10).$$

The logit transformation,

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

is another useful transformation which approximately converts a multiplicative model for $\hat{p}$ into an additive model for $y(\hat{p})$. A small advantage of the logit transformation is that it guarantees that

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

whereas the logarithmic transformation only guarantees

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

and the identity transformation makes no guarantees.

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

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

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

(iii) confidence intervals on P_T .

These derivations are all similar and are based upon Taylor series approximations (statistical differentials). For instance if $y(\hat{p}) = \ell n(\hat{p})$, then

$$\hat{P}_T = y^{-1}(\hat{\alpha}_T) \cong y^{-1}(\alpha_T) + (\hat{\alpha}_T - \alpha_T) \left[\frac{dy^{-1}(\hat{\alpha}_T)}{d\hat{\alpha}_T} \right]_{\hat{\alpha}_T = \alpha_T}$$

$$= P_T + (\hat{\alpha}_T - \alpha_T) P_T ,$$

so that

$$\text{MSE } (\hat{P}_T) \equiv E\,[(\hat{P}_T - P_T)^2] \cong P_T^2 \text{ Var}(\hat{\alpha}_T) .$$

4. THE WEIGHTED LEAST SQUARES ANALYSIS OF THE SEGMENT ESTIMATES

The probable heteroscedasticity of the $y(\hat{p}_{ts\ell})$'s suggests that the mixed effects model (1) should be analyzed in the form

$$w_{ts\ell}\, y(\hat{p}_{ts\ell}) = w_{ts\ell}\alpha_t + w_{ts\ell}\, b_s + w_{ts\ell}\, \delta_\ell + \varepsilon_{ts\ell} \qquad (2)$$

where $w_{ts\ell}$ is proportional to $\{\text{Var}\,[y(\hat{p}_{ts\ell})]\}^{-\frac{1}{2}}$.

In matrix notation (2) can be written as

$$Wy = WX\binom{\alpha}{\delta} + WU\, b + I\varepsilon \qquad (3)$$

where

$y = (y_{111}, y_{112}, \ldots, y_{TSL})'$,

$\alpha = (\alpha_1, \ldots, \alpha_T)'$,

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

$b = (b_1, b_2, \ldots, b_S)'$,

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

X = design matrix of 0's and 1's corresponding to the fixed effects α_t and δ_ℓ ,

U = sampling design matrix of 0's and 1's corresponding to the sampling pattern for the distinct segments, and

I = identity matrix.

In (3) the random portion of Wy is $WUb + I\varepsilon$ which has covariance

$$V\sigma_\varepsilon^2 \equiv I\sigma_\varepsilon^2 + WUU'W'\sigma_b^2$$
$$= (I + WUU'W'\gamma)\sigma_\varepsilon^2 \quad ,$$

where $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_{ts\ell})$ and $\gamma = \sigma_b^2/\sigma_\varepsilon^2$. Hence, the usual weighted least squares estimator of $(\alpha, \delta)'$ is

$$\hat{\binom{\alpha}{\delta}} = (X'W'V^{-1}WX)^{-1}X'W'V^{-1}Wy \tag{4}$$

and

$$\mathrm{Var}[\hat{\binom{\alpha}{\delta}}] = (X'W'V^{-1}WX)^{-1}\sigma_\varepsilon^2 \ .$$

In particular

$$\mathrm{Var}\ (\hat{\alpha}_T) = (X'W'V^{-1}WX)^{-1}_{T,T}\ \sigma_\varepsilon^2, \tag{5}$$

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

Although the formulas in (4) and (5) are fairly standard, there are several obstacles to be overcome before they can be applied. The detailed procedures for overcoming them are given in Dahm and Sielken (1981). Only the nature of obstacles and the basic approach to overcoming them are discussed here.

An initial obstacle is that the y vector is not computable if any $y(\hat{p}_{ts\ell})$ corresponds to either the logarithmic transformation with $\hat{p}_{ts\ell} = 0$ or the logit transformation with $\hat{p}_{ts\ell} = 0$ or 1. Although $\hat{p}_{ts\ell} = 1$ would be highly unexpected, $\hat{p}_{ts\ell} = 0$ is quite common. This obstacle can be overcome through the use of "working y's" as in Finney (1964); that is, by

(i) estimating the parameters in (3) using only the data for which y is calculable;

(ii) substituting the estimated parameters from (i) into (1) along with $e_{ts\ell} \equiv 0$ to obtain approximate $y_{ts\ell}$'s, say $y^*_{ts\ell}$, and approximate $\hat{p}_{ts\ell}$'s , say $\hat{p}^*_{ts\ell} = y^{-1}(y^*_{ts\ell})$; and finally

(iii) creating working values for $y(\hat{p}_{ts\ell})$ using a first order Taylor series expansion of $y(\hat{p})$ about $\hat{p} = \hat{p}^*_{ts\ell}$.

These working y's can then be used in (4).

A second obstacle to using both (4) and (5) is that V^{-1} contains the unknown variance component ratio $\gamma = \sigma_b^2/\sigma^2_\varepsilon$. If γ is replaced by an independent consistent estimator $\hat{\gamma}$, then (5) is asymptotically correct. When such a $\hat{\gamma}$ is unavailable, a reasonable alternative is to treat (3) as if it were a fixed effects model and obtain estimates of σ_b^2 and σ_ε^2 (and hence their ratio γ) by equating certain sums of squares from the fixed effects model analysis with their expectations under the mixed model. This is basically Henderson's Method 3 (see, for example, Searle (1971)).

Finally, the weight matrix W is unknown since Var $[y(\hat{p}_{ts\ell})]$ is unknown. A first order Taylor series approximation can be used to relate Var $[y(\hat{p}_{ts\ell})]$ to Var $(\hat{p}_{ts\ell})$. For example, if $y(\hat{p}) = \ell n(\hat{p})$ and $\hat{p}$ is distributed with mean p and variance σ_p^2 , then

$$y(\hat{p}) \cong \ell n(p) + (\hat{p}-p)\left[\frac{dy(p)}{d\hat{p}}\right]_{\hat{p}=p}$$

$$= \ell n(p) + (\hat{p}-p)/p \ ,$$

so that

$$E\,[y(\hat{p})] \cong \ell n(p)$$

and

$$\mathrm{Var}\,[y(\hat{p})] \cong E\,[(\hat{p} - p)^2/p^2]$$

$$\sigma^2_{\hat{p}}/p^2 \ .$$

In this manner the form of W can be identified. Replacing p by $\hat{p}$ would yield an estimate of W if $\sigma^2_{\hat{p}}$ could be estimated. One approach to estimating $\sigma^2_{\hat{p}}$ is to assume that $N\hat{p}$ is binomially distributed for some unknown value of N which is constant for all segments. Then $\sigma^2_{\hat{p}}$ is proportional to $p(1 - p)$ and in the above example Var $[y(\hat{p})]$ is proportional to $(1 - p)/p$ which can be estimated by $(1 - \hat{p})/\hat{p}$. A slight improvement can sometimes be obtained by iterating on the estimates of W and the p's. An alternative method of obtaining an

estimate of W is currently under investigation. Here Var ($\hat{p}_{ts\ell}$) is approximated primarily on the basis of information such as

(i) the type of satellite being used,

(ii) the sharpness of the satellite imagery,

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

(iv) the number of satellite images successfully obtained by the time the segment proportion is estimated,

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

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

(vii) the physical characteristics of the segment.

This alternative approach may be particularly appropriate for multiyear data sets where the remote sensing technology and segment proportion estimation methodology is changing from year to year. In addition, recognizable segment characteristics which make it either easier or harder to estimate the segment crop porportion can be incorporated. Obvious differences in the amount of information going into the $\hat{p}_{ts\ell}$'s can also be reflected. These latter differences can be due to the estimation times themselves as well as due to loss of satellite imagery from cloud cover, machine failure, etc.

5. AN EXAMPLE

The technical reports cited in the bibliography as well as the paper by Chhikara et al. (1984) in this issue indicate the theoretical advantages of basing estimators on the full multiyear data set as opposed to only the data from a single year. Even when there are only 2 or 3 years' data available, the accuracy of the current year's at-harvest crop proportion estimate can often be improved by as much as 50% by utilizing the multiyear estimation procedures. Of course, the improvement depends on the multiyear sampling design and the underlying value of $\gamma = \sigma_b^2/\sigma_\varepsilon^2$.

Some of the potential benefits of the multiyear estimation procedure in a real-world setting are seen in the following

example. The Landsat based estimates of the at-harvest wheat acreages computed at harvest times during each of 1976, 1977, and 1978 for 108 sample segments in the Great Plains states were available to the authors. Although these sample segment estimates were determined for other purposes, they can also be used to evaluate proposed statistical procedures. In an experiment the following procedure was repeated 200 times:

(a) Randomly select (without replacement) 40 segments from the 108 available.

(b) Treat this sample of 40 segments with their 3 years of estimated at-harvest wheat acreages as the simulated "stratum" whose at-harvest wheat acreage proportion is to be estimated. Determine the true at-harvest wheat acreage proportion for 1978 for this "stratum". This proportion is the estimation target for this repetition.

(c) Assume the following multiyear rotation sampling design. In 1976 a random sample of 5 segments from the stratum of 40 segments is observed. In 1977 three of these five are observed again along with two new randomly selected segments. Finally in 1978 one of the three segments observed in both 1976 and 1977 is observed a third time, the two new segments in 1977 are observed a second time in 1978, and finally two totally new randomly selected segments are observed. Schematically the sampling design of 5 segments per year is as follows:

Segment Number	1976	1977	1978
1	x		
2	x		
3	x	x	
4	x	x	
5	x	x	x
6		x	x
7		x	x
8			x
9			x

(d) The multiyear estimation procedure described in section 4 is carried out using $y(p) = \ln(p)$. The multiyear estimate, $y^{-1}(\hat{\alpha}_3)$, of the stratum's at-harvest wheat acreage proportion in 1978 is computed. The corresponding single-year estimate is also computed using only the 1978 sample data.

(e) The corresponding estimation errors are the differences between the simulated stratum's 1978 at-harvest wheat acreage proportion and the multiyear and single-year estimates.

The average absolute value of the errors was 0.046 for the multiyear estimator and 0.072 for the single-year estimator. Thus, the average absolute error for the single-year estimator was approximately 1.6 (0.072/0.046 =1.57) times as great as the average absolute error for the multiyear estimator. All of the other measures of empirical behavior considered also favored the multiyear estimator. The average squared errors for the multiyear and single year estimators were 0.0033 and 0.0073, respectively. The average biases relative to the average 1978 at-harvest wheat acreage proportion for the entire 108 segments were 0.002 and -0.047. The sample standard deviations of the multiyear and single-year procedures were 0.061 and 0.076, respectively. Thus, the multiyear estimation procedure provided a substantial percentage improvement over the single-year estimator.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support provided by NASA under contracts NAS9-13894 and NAS9-16785. The cooperative research spirit generated by the co-contractors and NASA personnel was also greatly appreciated.

BIBLIOGRAPHY

Chhikara, R.S. and A.H. Feiveson (1982). A sample survey of global wheat acreage using satellite (LANDSAT) Data. *Sankhya: The Indian Journal of Statistics*, B, 44, 304-329.

Chhikara, R.S., C.R. Hallum and T.G. Lycthuan-Lee (1984). Stratification and sampling design for Landsat crop surveys. *Communications in Statistics*, A,

Dahm, P.F. and R.L. Sielken, Jr. (1981). Multiyear estimation of the at-harvest crop acreage proportion: Methodology and implementation. Technical Report S-20, Institute of Statistics, Texas A&M University.

Feiveson, Alan H. (1984). Weighted ratio estimation in agricultural surveys. *Communications in Statistics*, A,

Finney, D.J. (1964). *Statistical Methods in Biological Assay*. Hafner Publishing Company, New York.

Freund, R.J., H.O. Hartley and T. Lee (1979). Gains of precision achievable by multi-year estimation. Technical Report S-16, Institute of Statistics, Texas A&M Unviersity.

Gbur, E.E. and R.L. Sielken, Jr. (1980). Optimal rotation designs for multiyear estimation, I: Unweighted estimation. Technical Report S-18, Institute of Statistics, Texas A&M University.

Gbur, E.E. and R.L. Sielken, Jr. (1980). Optimal rotation designs for multiyear estimation, II: Weighted estimation. Technical Report S-19, Institute of Statistics, Texas A&M University.

Gbur, E.E. and R.L. Sielken, Jr. (1981). Missing observations in multiyear rotation sampling designs. Technical Report S-22, Institute of Statistics, Texas A&M University.

Hall, F.G. and A.G. Houston (1984). Use of satellite data in agricultural surveys. *Communication in Statistics*, A,

Heydorn, Richard P. (1984). Using satellite remote sensing data to determine crop proportions in a sampling unit. *Communication in Statistics*, A,

Proceedings of Technical Sessions--The LACIE Symposium (1978). JSC-16105, Vols. 1 and 2, NASA/JSC, Houston, Texas.

Searle, S.R. (1971). *Linear Models*. John Wiley and Sons, Inc., New York.