W/comments

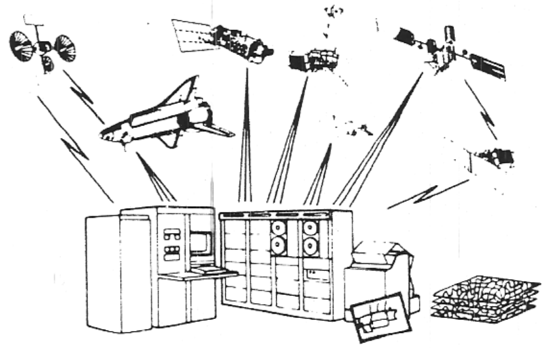
JSC-18896
LEMSCO-19985

LANDSAT DATA BASED REGRESSION ESTIMATORS IN CROP SURVEYS: IMPLEMENTATION PLAN

Prepared By
Raj S. Chhikara
Jim C. Lundgren

Lockheed Engineering and Management Services Company, Inc.
Under Contract NAS 9-15800

December 1983



EARTH SCIENCES AND APPLICATIONS DIVISION

NASA
National Aeronautics and
Space Administration

Lyndon B. Johnson Space Center
Houston, Texas 77058

1. Report No. JSC 18896	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Landsat Data Based Regression Estimators in Crop Surveys: Implementation Plan	5. Report Date December 1, 1983 6. Performing Organization Code 26-43
7. Author(s) R. S. Chhikara J. C. Lundgren	8. Performing Organization Report No. LEMSCO-19985
9. Performing Organization Name and Address Lockheed Engineering and Management Services Company, Inc. 1830 NASA Road 1 Houston, Texas 77058	10. Work Unit No. 11. Contract or Grant No. NAS 9-15800
12. Sponsoring Agency Name and Address Earth Sciences and Applications Division/SC National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, TX 77058	13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code SC

15. Supplementary Notes

16. Abstract

The report describes an implementation plan for the proposed study of Landsat based regression estimators of crop acreages. Technical issues, approaches and tasks to be performed are discussed. A preliminary task schedule is given.

17. Key Words (Suggested by Author(s)) Bayes classifier, classification errors, regression function, simulation, sample size	18. Distribution Statement USDA: Charlie Caudill, Bill Pratt, Richard Sigman, Bill Dowdy LEMSCO: B. L. Carroll, R. K. Lennington NASA: J. Erickson, B. Erb, Tech Library

19. Security Classif. (of this report) N/A	20. Security Classif. (of this page)	21. No. of Pages 16	22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161

JSC-18896

LANDSAT DATA BASED REGRESSION ESTIMATORS IN CROP SURVEYS

IMPLEMENTATION PLAN

Job Order 71-350
NAS 9-15800

Prepared by

Chhikara

R. S. Chhikara

J. C. Lundgren

LEMSCO

R. K. Lennington, Manager
Scientific Systems Department

NASA

R. Bryan Erb

R. B. Erb, Deputy Chief
Earth Sciences & Applications
Division

Prepared by

Lockheed Engineering and Management Services Company, Inc.
Houston, Texas

For

United States Department of Agriculture

National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas

December 1983

LEMSCO-19985

FOREWARD

This report has been prepared for the U. S. Department of Agriculture and is a part of the proposed study of Landsat data based regression estimators of crop acreages. It describes an implementation plan containing information on technical issues, approaches, and tasks to be performed.

LANDSAT DATA BASED REGRESSION ESTIMATORS IN CROP SURVEYS

1.0 INTRODUCTION

When a complete coverage of an area by Landsat is available, the spectral data can be used to obtain an auxiliary variable with its values known for all units comprising the area. In estimating crop acreages for a stratum, this auxiliary variable usually is the estimated crop acreages obtained using a supervised classification of Landsat data. The actual crop acreages are determined for a number of sample units/segments by a complete enumeration of their land use and cover types. The sample segment is, on the average, approximately one square mile in area. Given the actual and Landsat-estimated crop acreages for a set of sample segments, as well as the Landsat-estimated crop acreage for the entire stratum, one can consider either a ratio or a regression estimator to improve upon the standard crop acreage estimate made each year from the June Enumerative Sample Survey data.

The current USDA approach is to use the linear regression estimate obtained by regressing the actual crop acreage y of a segment onto its estimate x obtained from its Landsat data. The basis of this estimate is the linear model,

$$y_i = \alpha + \beta x_i + e_i \qquad (1)$$

with $E[e_i] = 0$, var $[e_i] = \sigma^2$, a constant, for all i and the e_i are uncorrelated. When the model (1) holds true (in the sense there is a high correlation between the two variables, x and y), the gain in efficiency is substantial from the use of the linear regression estimate compared to the sample mean for the stratum crop acreage.

In the present context, the variable x is subject to variability in training (spectral) data from the sample segments which are used in estimating the class parameters and hence, in training a classifier. So the large sample properties known for the linear regression estimator hold true conditionally, given that the training data are fixed. The small sample properties of the estimator are analytically difficult to establish even for the standard situation.

The main objective of the proposed study is to investigate both the large and small sample properties of the linear regression estimator developed by USDA using auxiliary variables derived from Landsat data. Specifically, the model (1) will be studied with its regression coefficients and error term characterized in terms of both the spectral and crop acreage variabilities for the stratum. The gain in efficiency due to regression estimator will be investigated and a comparison will be made between its efficiency and those of its competitors, in particular the ratio ~~and post-stratified~~ estimator~~s~~. Both the analytic and empirical approaches will be used. The technical approaches and the plan to carry out the research are outlined next.

2.0 ERROR ANALYSIS AND LARGE SAMPLE PROPERTIES OF REGRESSION ESTIMATOR

2.1 Background and Technical Issues

When the classifier is completely specified, the segment crop proportion obtained from the classification of its Landsat data can be written as a linear function of its actual crop proportion and vice versa. But this functional relationship between the two proportions varies across segments unless the classification errors of omission and commission are the same for all segments with respect to each crop of interest. Since the spectral class separability is usually variable from one segment to another, these classification errors are not expected to remain constant across segments. Empirical studies involving Landsat data have shown that there can be a substantial between-segment variation in a stratum for each error type. Thus, when a linear model is considered for describing the relationship between the actual and the estimated crop proportions across segments, it is assumed that the model coefficients, intercept and slope, are functions of the average errors of omission and commission for the stratum and the error term in the model has a conditional mean of zero. However, the conditional variance of the variable y with value of the variable x specified, may be a function of the specified value; and hence, the error term in the model may not have a constant variance. In addition, this error variance is a function of the variances of the classification errors of omission and commission and the covariance between them.

When the classifier is trained using spectral data from sample segments, the error term is subject to sampling error in training statistic and thus an additional variability is introduced in model (1). This would certainly weaken the linear model and, if the relationship between y and x is less precise, there may not be any gain in efficiency due to the regression estimator. When the spectral class distributions (corresponding to distinct crop types) are well separated, the effect on the model due to sampling error in training statistics may not be significant, particularly in the case of large samples.

Thus, an evaluation of the regression estimator consists of (a) a study of the classification errors associated with a classifier and their effect on a segment crop acreage estimate, and (b) an investigation of the regression function and the mean square error of the regression estimator as a function of the sample size, the overlap between spectral class distributions and the relative size of the class of mixed pixels. The problem in full generality is analytically intractable, and so the present study will be carried out by treating it as a two-class discrimination problem with the Bayes linear classifier used. The size of training data will be assumed large, an assumption likely to hold true when spectral data from all sample segments are pooled for the training set.

2.2 Technical Approach

Let C_1 denote the class of interest and C_0 be the other class. Suppose the spectral observation $\underset{\sim}{Z}$ is a p-dimensional vector and has the multivariate normal distribution with mean $\underset{\sim}{\mu}_i$ and covariance matrix $\underset{\sim}{\Sigma}$ for C_i, i=0,1. Without any loss of generality, assume that

$$\underset{\sim}{\mu}_0 = \begin{bmatrix} -\Delta/2 \\ \underset{\sim}{0} \end{bmatrix} , \quad \underset{\sim}{\mu}_1 = \begin{bmatrix} \Delta/2 \\ \underset{\sim}{0} \end{bmatrix}$$

and $\Sigma = I$, where $\Delta = [(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_0)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_0)]^{1/2}$.

Suppose n sample segments (clusters of pixels) are randomly selected and their (completely enumerated) spectral data are utilized in estimating the class parameters. Let $\overline{\underset{\sim}{Z}}_0$ and $\overline{\underset{\sim}{Z}}_1$ denote the sample means and $\underset{\sim}{S}$ the pooled sample covariance matrix for the two classes. Then a spectral observation $\underset{\sim}{Z}$ can be classified on the basis of linear discriminant functions given by

$$\lambda\ (\underset{\sim}{Z}) = b_0 + \underset{\sim}{b}'\underset{\sim}{Z} \qquad (2)$$

where

$$b_0 = -(1/2)\ (\overline{\underset{\sim}{Z}}_1 - \overline{\underset{\sim}{Z}}_0)'\ \underset{\sim}{S}^{-1}\ (\overline{\underset{\sim}{Z}}_1 + \overline{\underset{\sim}{Z}}_0)$$

$$\underset{\sim}{b} = S^{-1}\ (\overline{\underset{\sim}{Z}}_1 - \overline{\underset{\sim}{Z}}_0) \qquad (3)$$

The classification procedure is to regard the observed value, $\underset{\sim}{Z}$ coming from C_0 or C_1, as the discriminant value, $\lambda\ (\underset{\sim}{Z}) \leq 0$ or >0, respectively.

Define the random variable

$$\psi(\underset{\sim}{Z}) = \begin{cases} 1 & \text{if } \lambda(\underset{\sim}{Z})>0 \\ 0 & \text{if } \lambda(\underset{\sim}{Z})\leq 0 \end{cases} \qquad (4)$$

Then an estimate of the actual crop acreage y_i for the ith segment is given by

x_i is a r.v.

$$x_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \psi\ (\underset{\sim}{Z}_{ij}) \qquad (5)$$

where M_i is the total number of pixels in segment i. Here $\underset{\sim}{Z}_{ij}$ denotes the observation vector for the jth pixel in segment i.

$$\hat{\Theta}_{0i} = \frac{\Sigma \psi(Z_{ij}) I_{C_0}(Z_{ij})}{\Sigma I_{C_0}(Z_{ij})} \leftarrow m_i - y_i$$

r.v.'s (statistics) vs parameters

It follows from (4) that the number of pixels classified in C_1, given training statistics $\underset{\sim}{\overline{Z}}_1$, $\underset{\sim}{\overline{Z}}_0$, $\underset{\sim}{S}$, is

$$x_i = (1-\hat{\theta}_{1i} - \hat{\theta}_{0i})y_i + \hat{\theta}_{0i} \qquad (6)$$

$E[\hat{\Theta}_{11} | Z_1, Z_0, \Sigma] = \Theta_0$

where

$$\hat{\theta}_{0i} = P_r[\psi(\underset{\sim}{Z}) = 1 | \underset{\sim}{\overline{Z}}_1, \underset{\sim}{\overline{Z}}_0, \underset{\sim}{S}, \underset{\sim}{Z} \in C_0]$$
$$\hat{\theta}_{1i} = P_r[\psi(\underset{\sim}{Z}) = 0 | \underset{\sim}{\overline{Z}}_1, \underset{\sim}{\overline{Z}}_0, \underset{\sim}{S}, \underset{\sim}{Z} \in C_1] \qquad (7)$$

no i on right side

For the classifier in (2), the first two asymptotic moments of $\hat{\theta}_{0i}$ and $\hat{\theta}_{1i}$ can be obtained using the approach given by Efron (JASA, 1975) provided the training data are assumed to be random.

$E(X|y_i) = f(y_i)$

These moments are functions of Δ, p and n, among others, and their evaluation would provide the mean and variance of x_i for given y_i. Inversely, the y_i can be written in terms of the expected value of x_i and some functions of the expected values of $\hat{\theta}_{0i}$ and $\hat{\theta}_{1i}$. When this relationship is modeled across segments, it may be feasible to characterize the regression functions in terms of parameters Δ, p, and n and study the mean square error of the linear regression estimator as functions of these basic parameters.

what about non-basic parameters e.g. γ estimate from data

One, however, needs to take into account the presence of mixed pixels, which sometimes constitute a large part of a Landsat scene. Recently Chhikara (J. of Remote Sensing of Environment, 1984 - to appear) has extended the approach of Efron to the case that includes the class of mixed pixels as well. Suppose C_m is the class of mixed pixels and $\hat{\theta}_{mi} = P_r[\psi(\underset{\sim}{Z})=1|\underset{\sim}{\overline{Z}}_1,\underset{\sim}{\overline{Z}}_0,\underset{\sim}{S},\underset{\sim}{Z} \in C_m]$.

If π_{mi} is the size of the mixed pixels class, then the number of pixels classified in C_1 is given by

$$x_i = (1-\pi_{mi})[\hat{\theta}_{oi} + (1-\hat{\theta}_i - \hat{\theta}_{oi})\, y_i] + \pi_{mi}\, \hat{\theta}_{mi}. \qquad (8)$$

?

Following the approach of Chhikara (1984) and subsequently the line of argument outlined in the previous paragraph, a study of the linear regression function and the resulting estimator can be made.
In specific, the asymptotic distributions of $\hat{\theta}_{oi}$, $\hat{\theta}_{1i}$ and $\hat{\theta}_{mi}$ will be obtained and their parameters will be related to the true regression coefficients and the error variance so that the model as given in (1) can be fully described. The analytical study will be followed by a Monte Carlo study where Landsat data will be simulated and crop acreages estimated using the classification procedure given in (2). A comparison between the two results will permit an evaluation of the assumptions the classifier and the linear regression model are based on. The results of this evaluation will then be used to study the large sample properties (i.e., bias and variance) of the stratum crop acreage estimator.

2.3 Task Description

Expressions for the classification error rates will be derived for the linear Bayes and maximum likelihood classifiers which are estimated/trained using sample data. Their asymptotic distributions and first two moments will be given. The coefficients and the error variance of the model in (1) will be described in terms of the spectral class separability, the sample size and the actual crop acreages. The linearity of the model will be examined and the regression estimator will be compared to the ratio and ~~post-stratified~~ estimators for gain in efficiency. The theoretical results will be numerically illustrated for some special cases and a sensitivity study conducted.

2.4 Task Schedule

Error Characterization and Model Investigation	March 31, 1984
Regression Estimation Study	August 15, 1984
Reporting of Results	September 1, 1984
Documentation	September 30, 1984

3.0 PROPERTIES OF REGRESSION ESTIMATOR

3.1 Background and Technical Issues

When the number of sample segments available for regression in a stratum is small, the sampling properties, (e.g., bias and variance) of the regression estimator are unknown. In particular, no general analytical results are available on the accuracy of the approximate variance formula mentioned above and its estimate obtained from the residual mean square error. Rao (J. Ind. Statist. Assoc., 1968, pp. 160-168), among others, has studied by Monte Carlo techniques, the underestimation of the variance when either the approximate variance formula or its usual estimate are utilized for variance estimation. No such investigation has been made for the particular situation currently addressed, where the classifier introduces an additional source of error in the independent variable x and the sample size is usually small.

3.2 Technical Approach

No analytical study of the performance of a classifier and the regression estimator is possible in the case of a small to moderate sample size. To investigate the small sample properties of the estimator, the simulation study will include cases of small sample sizes. It has been argued (Cochran, Sampling Techniques, Wiley, 1977, and Rao, 1968) that the variance of the regression estimator is usually underestimated when the approximate variance formula is utilized. The results of earlier studies of Rao and others reported in the literature do not apply to the case of linear models with measurement errors. These cannot be used to evaluate the Landsat-data based regression estimator; an evaluation study which would take the classification aspects into full consideration will be made using simulations.

Lockheed has developed a computer algorithm which simulates Landsat satellite Multispectral Scanner imagery over a randomly generated segment and field structure. The simulation program can be divided into two major parts. The first part creates a simulated segment structure in which fields are fit together to form segments and a pixel grid is overlaid onto this structure. The second part simulates the spectral values for each pixel.

The objective of this program is to create simuated segments which are similar to 33 actual Missouri segments which were sampled by the Statistical Reporting Service of the USDA in 1979. Specifically, for the first major part, the simulated segments are designed to have the same expected crop acreage proportions, distribution of field sizes, and segment size. Other parameters which are considered in the design of the simulation are the distribution of segment crop proportions, distribution of segment sizes, and percentage of edge pixels.

The first two objectives are met by randomly selecting fields according to the empirical distribution of fields in the 33 actual segments. These fields are joined together by assuming a rectangular shape for all fields with a constant width, arranging these fields into long arrays with parallel sides, and placing these arrays side by side in a formation of fields. By randomly locating a segment pattern on this formation and assigning fields which have their centers within that pattern to be part of the segment, we have a simulated segment which has the same expected size as the pattern and which has the same expected crop proportions and distributions of field sizes as the actual data.

A pixel grid is overlaid on this simulated segment such that the angle between the pixel and field edges is similar to the corresponding angle in the actual data.

The specific objectives of the second part of the simulation procedure are to create spectral values with the same within-field variance, between-field (within segment) variance, and between segment variance as the actual spectral values for each crop. In addition, it was desired to simulate spectral data in each of the four MSS channels, for each of two acquisitions, and for both pure and mixed pixels. The first step in the spectral simulation is to find the principal components of the eight spectral values (four channels in each of two acquisitions) for pure pixels in each crop. Since the principal components are uncorrelated by definition, it was decided to simulate each principal component independently.

In studies previously conducted with real data, the first four principal components accounted for the vast majority of the variability in the data. Therefore, it was decided to simulate only the first four principal components before transforming the data back into eight spectral values. A univariate unbalanced nested analysis of variance model was considered for each of the first four principal components within each of the major crops:

$$Y_{ijk} = \mu + S_i + F_{ij} + \xi_{ijk}$$

where Y_{ijk} = a principal component value for pixel k, in field j, and segment i,

μ = the mean principal component value,

S_i = the effect of segment i,

F_{ij} = the effect of field j within segment i,

ξ_{ijk} = the effect pixel k, within field j, within segment i.

The analysis of variance was conducted to test for significant segment and field effects, and to partition the sums of squares in computing the variance components. Since both the segment and field effects were found to be significant, the same model was used in the simulation procedure.

In this simulation model, each effect (segment, field, and pixel) is assumed to come from a normal distribution with a mean of zero and variance equal to the computed variance component. By generating a random effect for each segment, for each field, and for each pixel from the appropriate distribution, a random principal component value can be constructed for every pure pixel. These simulated principal component values are then transformed into their spectral values through use of the inverse eigenvector matrix. It has been shown that distributions of these simulated values do not significantly differ from the distributions of actual spectral values.

A mixed pixel is simulated by generating pure pixel values for each of the applicable fields, then linearly combining these values with weights equal to the proportion of the pixel in each field.

ranging from 25 to 1000 segments, together with sample sizes of 4, 6, 10, 15 and 100 segments (depending on population size). For each combination of population and sample size, several estimates of the crop proportion will be made for each of several replications. These estimates include the sample mean as well as both the regression and ratio estimators derived from using both Bayes and maximum likelihood classifiers. Through use of replication, we will estimate the bias and variance of each of these procedures and relate these estimates to the sample/population sizes.

A second similar regression/classifier study is planned in which the separability of the spectral distributions of the crop classes is either increased or decreased through adjustments in the means and variances of the simulation model.

This simulation approach should be useful not only in the evaluation of bias and variance of the estimator as a function of sample size, but it will also provide the information to verify the assumptions underlying a classifier and the regression model. This, in turn, can be used to study the robustness of the estimator to any deviations from the model assumptions.

3.4 Task Schedule

Software Development	variance estimator ratio estimator	January 31, 1984
Initial Sample Size Studies of Regression Estimators		March 31, 1984
		report april 10
Regression Studies with Varying Separability		July 31, 1984
Reporting of Results		September 1, 1984
Documentation		September 30, 1984