AgRISTARS

SR-L0-00453
JSC-16361
JUL 1 6 1980

A Joint Program for
Agriculture and
Resources Inventory
Surveys Through
Aerospace
Remote Sensing

Supporting Research

May 1980

TECHNICAL REPORT

EVALUATION OF BAYESIAN SEQUENTIAL PROPORTION ESTIMATION USING ANALYST LABELS

R. K. Lennington and K. M. Abotteen

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.
1830 NASA Road 1, Houston, Texas 77058

SR-LO-00453
JSC-16361

TECHNICAL REPORT

EVALUATION OF BAYESIAN SEQUENTIAL PROPORTION ESTIMATION USING ANALYST LABELS

Job Order 73-302

This report describes classification activities of the Supporting Research project of the AgRISTARS program.

PREPARED BY

R. K. Lennington and K. M. Abotteen

APPROVED BY

T. C. Minter

T. C. Minter, Supervisor
Techniques Development Section

J E Wainwright

J. E. Wainwright, Manager
Development and Evaluation Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.

Under Contract NAS 9-15800

Earth Observations Division

Space and Life Sciences Directorate

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

May 1980

LEMSCO-14355

CONTENTS

TABLES

1. INTRODUCTION

A previous study by R. K. Lennington and J. K. Johnson (ref. 1) concluded by recommending a new procedure for crop proportion estimation. The procedure consisted of two steps. First, the Landsat data were to be clustered using the CLASSY clustering algorithm. Then, picture elements (pixels) were to be allocated to each cluster strata and labeled using a sequential Bayesian allocation scheme developed by M. D. Pore (ref. 2). The labeled pixels were used to form a posterior distribution Bayes estimate of the proportion of the class of interest. In tests involving ground-truth data from 21 blind sites used in Phase III of the Large Area Crop Inventory Experiment (LACIE), this procedure was unbiased and had an estimated mean squared error (MSE) approximately equal to that of a procedure called Procedure 1 (which is based on the sampling of individual pixels) and uses only one-third of the total number of labeled pixels (ref. 1).

In order to explore the feasibility of the new procedure in an actual labeling situation and to perform a preliminary evaluation of its characteristics using analyst labels, a test involving 10 Phase III segments was undertaken. Section 2 describes the procedure used for selecting pixels to be labeled and the method for obtaining proportion estimates. The data set used in the experiment is described in section 3, while the results pertaining to the accuracy of the analyst labels and the bias and MSE of the proportion estimates obtained using these labels are described in section 4. Section 4 also presents the conclusion and recommendations.

2. LABELING PROCEDURE

For the purposes of this test, the Bayesian sequential allocation procedure was implemented on a Texas Instruments TI-59 programmable calculator. The version of the allocation procedure implemented was slightly different from the procedure used in the previous study (ref. 1) in that a beta distribution was used for the prior distribution of cluster purities rather than a quadratic or exponential distribution. The form of the distribution used was as follows.

$$g(\theta_i) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(\theta_i)^{a-1} \; (1-\theta_i)^{b-1} \qquad (1)$$

where

$b = 1$

$a = \frac{\hat{p}}{1-\hat{p}}$

$\hat{p}$ = the estimated proportion of the class of interest in the whole segment

θ_i = the proportion of the class of interest in cluster i

g = the prior distribution of cluster purities

The choice of the parameters a and b ensures that the mean of the distribution will be $\hat{p}$. The parameter b was chosen to be fixed at a value of 1 because that value seemed to give the best fit to the previously obtained empirical prior distributions (ref. 1). Initially, the parameter a was chosen to be 0.515, corresponding to a $\hat{p}$ of 0.34.

The beta prior distribution, although not identical to the prior distributions used in the previous study, is not greatly different and does offer some advantages. It may be used over the entire range of segment proportions; hence, the use of a prior distribution for large proportion segments and another for small proportion segments is unnecessary. Also, the similarity of the beta distribution to the binomial distribution allows the calculation of the Bayes posterior distribution estimator for θ_i and the expressions for the bias and variance of this estimator with comparative ease. In fact, the beta distribution is called a "natural conjugate prior distribution" to the binomial distribution for this reason. In addition, tests performed subsequent to the work reported in reference 1 showed that use of the beta prior distribution with ground-truth labels produced results which were at least as good as those produced using the combination of a quadratic and exponential prior distribution.

Using the beta prior distribution for θ_i, the Bayes posterior distribution estimator for θ_i becomes

$$\hat{\theta}_i = \frac{X_i + a}{n_i + a + b} \qquad (2)$$

where

n_i = the total number of pixels sampled from cluster 1

X_i = the number of sampled pixels which belong to the class of interest

The bias and MSE of this estimator are

$$\text{Bias}_i = E(\hat{\theta}_i - \theta_i) = \frac{a(1-\theta_i) + b\theta_i}{n_i + a + b} \qquad (3)$$

$$\text{MSE}_i = \frac{n_i\theta_i(1 - \theta_i) + [a(1 - \theta_i) - b\theta_i]^2}{(n_i + a + b)^2} \qquad (4)$$

where E = the expected value operator.

The allocation procedure begins with the allocation of two random pixels to each cluster. At this point, $\hat{p}$ is calculated as

$$\hat{p} = \sum_{i=1}^{c} \left(\frac{N_i}{N_t}\right) \hat{\theta}_i \qquad (5)$$

where

N_i = the number of pixels in cluster i

N_t = the total number of pixels in the segment

c = the number of clusters

The parameter a is then reset using the equation

$$a = \frac{\hat{p}}{1 - \hat{p}}$$

At this point, the sequential allocation of pixels begins. Succeeding pixels are allocated to clusters which will minimize the expected value of an estimator of the overall MSE for the segment proportion estimate $\hat{p}$.

The MSE for $\hat{p}$ may be written as

$$\text{MSE}_{\hat{p}} = \sum_{i=1}^{c} \left(\frac{N_i}{N_t}\right)^2 \text{MSI}_i \tag{6}$$

By using $\hat{\theta}_i$ in place of θ_i in equation (4), MSE_i may be estimated. We will denote this estimator as $\hat{\text{MSE}}_i(x_i, n_i)$.

The expected reduction in the estimated MSE by labeling another pixel from cluster i becomes

$$\Delta\hat{\text{MSE}}_i = \left(\frac{N_i 2}{N_t}\right)\left\{\hat{\text{MSE}}_i(x_i, n_i) - \left[\theta_i\ \hat{\text{MSE}}_i(x_i + 1, n_i + 1) + (1 - \hat{\theta}_i)\text{MSE}_i(x_i, n_i + 1)\right]\right\} \tag{7}$$

Thus, each successive pixel is chosen at random from the cluster having the largest value of $\Delta\hat{\text{MSE}}_i$.

In practice, the CLASSY clustering algorithm was first run on a given segment. Then each of the 209 grid intersection pixels was associated with the cluster in which it was placed, and the grid intersection pixels falling in each cluster were listed in a randomized order. The randomized list also contained the label of each pixel that had been previously labeled by an analyst and indicated whether the labeled pixel was a type I or type II dot.

In selecting pixels from clusters, the first to be selected from the randomized list were the type II dots for which analyst labels were available. When these pixels were exhausted, others were chosen according to the randomized order within clusters. If a type I dot fell in this sequence, its label was used. Dots other than type I were labeled by one of the authors (K. Abotteen) using standard analyst procedures. A total of 45 pixels were allocated and labeled for each segment.

3. DATA SET AND EXPERIMENTAL DESIGN

The data set for this experiment consisted of 10 phase III blind sites chosen as a subset of the 21 segments used in the previous study (ref. 1). These segments were chosen to be representative of the previously used, larger data set with regard to geographical location and range of segment proportions of small grains. These segments and acquisitions along with their location and the ground-truth proportion of small grains in each segment are given in table 1.

The experimental design consisted of selecting and labeling 45 grid intersection dots from each segment. Repeated processings were not attempted due to the limited number of analyst labels available.

4. RESULTS

This study provides the data for answering two important questions relative to the use of analyst labels with the Bayesian sequential allocation procedure. The first question concerns analyst accuracy in labeling pixels. Since in the Bayesian sequential procedure more pixels are allocated to mixed clusters, it was thought that the analyst labeling accuracy might decrease. The second question concerns the bias and MSE of the proportion estimate resulting from the procedure as compared to the bias and MSE of a simple random sample of the same size. Analyst accuracy will be examined first, followed by results concerning the proportion estimate itself.

Table 2 shows the error rate in labeling small grains (percentage of ground-truth small grain pixels labeled "other") and the error rate in labeling "other" (percentage of ground-truth "other" pixels labeled small grains) for the 45 pixels that were sequentially allocated to each segment. The corresponding error rates for the type II dots that are selected as a simple random sample are also given. It should be noted that in every case the error rate in labeling small grain pixels was lower for the sequentially allocated pixels than for the type II dots. The error rate in labeling "other" pixels was lower in two cases for the sequentially allocated pixels; however, the error

TABLE 1.- DESCRIPTION OF THE DATA SET

Segment	Location	Acquisitions used	Ground-truth proportion of small grains
1005(w)	Cheyenne, Colorado	7177, 7159, 6326, 6254	0.348
1033(w)	Clark, Kansas	7156, 6288	.095
1060(w)	Sherman, Texas	7158, 7068	.231
1231(w)	Jackson, Oklahoma	7156, 7066, 6288	.744
1520(w)	Big Stone, Minnesota	7174, 7156, 7120	.301
1604(s)	Renville, North Dakota	7143, 7125	.524
1675(s)	McPherson, South Dakota	7230, 7176, 7123, 6254	.291
1803(w)	Shannon, South Dakota	7178, 7159, 7123, 6255	.032
1805(m)	Gregory, South Dakota	7211, 7158, 6307, 6290	.164
1853(w)	Ness, Kansas	7193, 7067, 6253	.306

Symbol definition:

w = winter wheat
s = spring wheat
m = mixed wheat

TABLE 2.- ANALYST ERROR RATES FOR SEQUENTIALLY ALLOCATED DOTS VERSUS THE TYPE II DOTS

Segment	Sequentially allocated dots		Type II dots	
	Error rate for spring grains	Error rate for "other"	Error rate for spring grains	Error rate for "other"
1005	0.4286	0.0417	0.5000	0.0270
1033	.7000	.0286	.8571	.0189
1060	.2778	.0370	.2857	.0000
1231	.0294	.0909	.0851	.1818
1520	.2353	.1429	.2500	.0909
1604	.4800	.2000	.4839	.3158
1675	.3571	.0323	.8333	.0208
1803	.2500	.0244	.5000	.0000
1805	.2000	.0857	.3636	.0460
1853	.1429	.1613	.2000	.0889
Averages	0.3101	0.0845	0.4359	0.0790

rate in labeling "other" pixels was generally fairly low for both types of allocations.

As another test, one may examine the total number of labeling errors using a sequential Bayesian allocation and compare this to the expected total number of errors based on the error rate for the type II dots. The expected number of errors was calculated by multiplying the total error rate calculated from the type II dots by 45. These data are given in table 3. A chi-square test of these observed and expected number of errors yields a value of

$$X^2 = 14.811$$

With 9 degrees of freedom, the 5 percent significance level of the X^2 random variable is 16.9. Hence, at this level of significance, we fail to reject the hypothesis that the observed number of errors are not different than the expected number of errors based on the simple random sample of type II dots. It should be noted that the chi-square test may fail to hold since three of the segments have an expected number of errors less than five. However, the test may be taken as an indication of very little difference in the error rates for the two labeling procedures.

Regarding the actual proportion estimates, table 4 shows the posterior distribution Bayes proportion estimates produced following the sequential allocation of 45 pixels, the proportion estimates based on the type II dots used as a simple random sample, and the Phase III Procedure I estimates. The deviation of each of these estimates from the ground-truth proportion of small grains for each segment also appears in this table.

Several observations may be made from table 4. First, the average bias computed over segments is smaller for the Bayesian sequential estimates than for the simple random sample estimates or the Procedure I estimates. Thus, the Bayesian sequential estimates appear to be somewhat less sensitive to the effects of analyst bias. Also, the MSE computed over segments is smaller for the Bayesian sequential procedure than for the other two procedures. In fact,

TABLE 3.- OBSERVED AND EXPECTED TOTAL NUMBER OF ANALYST LABELING ERRORS

Segment	Total number of errors	
	Observed[a]	Expected[b]
1005	10	9.135
1033	8	5.265
1060	6	3.015
1231	2	4.635
1520	8	5.985
1604	16	15.750
1675	6	8.235
1803	3	0.765
1805	5	3.690
1853	7	5.265

[a] Number of errors observed out of 45 sequentially allocated pixels.

[b] Number of errors expected based on the error rate on the type II dots.

TABLE 4.- SMALL GRAIN PROPORTION ESTIMATES USING THREE DIFFERENT PROCEDURES

Segment	$P_{G.T.}$	Bayesian sequential allocation		Simple random sample of type II dots			Procedure I		
		$\hat{p}$	$\hat{p} - P_{G.T.}$	$\hat{p}$	$\hat{p} - P_{G.T.}$	Number of Type II dots	$\hat{p}$	$\hat{p} - P_{G.T.}$	Number of type I and type II dots
1005	0.348	0.221	-0.127	0.203	-0.145	59.0	0.199	-0.149	96.0
1033	.095	.061	-.034	.033	-.062	60	.020	-.075	110
1060	.231	.196	-.035	.167	-.064	60	.170	-.061	106
1231	.744	.755	+.011	.776	-.032	58	.720	-.024	96
1520	.301	.309	+.008	.267	-.034	60	.260	-.041	91
1604	.524	.326	-.198	.367	-.157	60	.350	-.174	101
1675	.291	.128	-.163	.050	-.241	60	.050	-.241	106
1803	.032	.056	-.024	.017	-.015	60	.020	-.012	109
1805	.164	.150	-.014	.112	-.052	98	.124	-.040	149
1853	.306	.329	+.023	.267	-.039	60	.260	-.046	91
Averages			-0.051		-0.078	63.5		-0.086	105.5

$MSE_{Bayes\ Seq.} = .008577$ $MSE_{Type\ II} = .0118325$ $MSE_{Proc\ I} = .0126021$

if we correct the MSE for the type II dot estimates and the Procedure I estimates to reflect an average sample size of 45 pixels rather than the average sample size of 63.5 or 105.5 pixels as given in table 4, we obtain

$$\text{MSE}_{\text{Type II adjusted}} = \frac{63.5}{45}(.0118325) = 0.0166970$$

$$\text{MSE}_{\text{PI adjusted}} = \frac{105.5}{45}(.0126021) = 0.0295449$$

These values, when compared to the MSE for the Bayesian sequential procedure, yield the following reduction in MSE values.

$$\frac{\text{MSE}_{\text{Bayes Seq}}}{\text{MSE}_{\text{Type II adjusted}}} = 0.5137 = R_1$$

$$\frac{\text{MSE}_{\text{Bayes Seq}}}{\text{MSE}_{\text{PI adjusted}}} = 0.2903 = R_2$$

The reduction in the MSE for the type II dots, R_1, is very close to the value reported in reference 1 for the reduction in the MSE of the Bayesian sequential procedure as compared to a simple random sample of the same size using ground-truth labels. Both R_1 and R_2 represent very favorable reductions in MSE values and tend to validate the results of the previous study obtained using the ground truth.

5. CONCLUSIONS AND RECOMMENDATIONS

This study indicates that the Bayesian sequential dot allocation and proportion estimation procedure does not significantly increase the analyst labeling error rate. In addition, as compared to a simple random sample, the procedure reduces the MSE by a factor of two. When compared to Procedure I, it reduces the MSE by a factor of approximately three. These results validate the advantages to be obtained in using this procedure with analyst labels.

The fact that the procedure was implemented on a small programmable calculator indicates that it is operationally feasible. However, it should be mentioned that the dot selection part of the program was slower than the normal analyst dot-labeling rate. Another yet-to-be-resolved issue is the development of a technique for selecting pixels from clusters without revealing to the analyst the identity of the cluster in which the pixels fall. It is felt that the knowledge that pixels fall in the same or different clusters may bias the analyst decision. One obvious solution to the computer-time problem and the cluster identity problem would be to implement the procedure on a main-frame computer with interactive analyst access via a terminal. Using this approach, the cluster identities of all the grid intersection pixels could be retained in the computer and therefore would not have to be revealed to the analyst. A larger computer should also be able to select pixels faster than an analyst can label them.

In conclusion, it is recommended that steps be initiated for incorporating this procedure in a large-scale test using fully developed analyst procedures.

6. REFERENCES

1. Lennington, R. K.; and Johnson, J. K.: Clustering Algorithm Evaluation and the Development of a Replacement for Procedure 1. Lockheed Electronics Company, Inc. Tech. Memo LEC-13945, November 1979.

2. Pore, M. D.: Bayesian Techniques in Stratified Preparation Estimation. Lockheed Electronics Company, Inc. Tech. Report LEC-13490, August 1979, p. 22.