AgRISTARS

Inside

E81-10182
SR-LO-00478
JSC-16378
CR-160965

JAN 2 1 1981

NASA CR-160965

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

Supporting Research

November 1980

THE MULTICATEGORY CASE OF THE SEQUENTIAL BAYESIAN PIXEL SELECTION AND ESTIMATION PROCEDURE

M. D. Pore and T. B. Dennis

(E81-10182) THE MULTICATEGORY CASE OF THE SEQUENTIAL BAYESIAN PIXEL SELECTION AND ESTIMATION PROCEDURE (Lockheed Engineering and Management) 22 p HC A02/MF A01 CSCL 12A N81-29498
Unclas
G3/43 00182

Lockheed Engineering and Management Services Company, Inc.
1830 NASA Road 1, Houston, Texas 77058

Lyndon B. Johnson Space Center
Houston, Texas 77058

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
JSC-16378; SR-LO-00478		

4. Title and Subtitle
The Multicategory Case of the Sequential Bayesian Pixel Selection and Estimation Procedure

5. Report Date
November 1980

6. Performing Organization Code

7. Author(s)
M. D. Pore and T. B. Dennis
Lockheed Engineering and Management Services Company, Inc.

8. Performing Organization Report No.
LEMSCO-14807

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
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas 77058 Technical Monitor: J. D. Erickson SH

13. Type of Report and Period Covered
Technical Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

A Bayesian technique for stratified proportion estimation and a sampling procedure based on minimizing the mean squared error of this estimator have been developed and tested on Landsat multispectral scanner data using the beta density function to model the prior distribution in the two-class case. In this paper, an extension of this procedure to the k-class case is considered. A generalization of the beta function is shown to be a density function for the general case which allows the procedure to be extended.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Bayesian techniques Stratified proportion estimation Sequential allocation	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	21	

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

SR-L0-00478
JSC-16378

THE MULTICATEGORY CASE OF THE SEQUENTIAL BAYESIAN PIXEL SELECTION AND ESTIMATION PROCEDURE

Job Order 73-306

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

PREPARED BY

M. D. Pore and T. B. Dennis

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

For

Earth Observations Division

Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

November 1980

LEMSCO-14807

CONTENTS

1. INTRODUCTION

A Bayesian technique for stratified proportion estimation and a sequential sampling procedure based on minimizing the mean squared error (MSE) of the posterior Bayesian estimator was developed by Pore (ref. 1) and tested by Lennington and Johnson (ref. 2) for the two-category case. The most favorable results were obtained when the prior distribution was modeled as a beta density function. These favorable results stemmed from a combination of the mathematical ease in developing the estimator and theoretical MSE, the ability to fairly closely model the empirical prior distribution with the beta, and the high accuracy in the data analysis. Virtually no bias and an MSE less than the proportional allocation case were reported. These results were obtained from analyses using Land Satellite (Landsat) multispectral scanner (MSS) data in which stratification was achieved by clustering picture elements (pixels) in a 9- by 11-kilometer area referred to as a segment. The two categories used were predominantly small-grains agricultural crops and nonsmall grains.

In section 2, the Bayesian development is presented for the three-category case, and in section 3, it is generalized to the k-category case. The three-category case might be used where, for example, barley is to be estimated within the small-grains category. A procedure of directly estimating barley, other small grains, and nonsmall grains might be tested if labeling practices allowed the direct labeling of barley and other small grains.

The k-category case in section 3 is presented for completeness and to document the results for future crop estimation possibilities.

The environment of these developments is as follows:

a. The segment (population) has been clustered (stratified) into several subgroups,

b. Pixels (samples) can be selected randomly within each cluster, and

c. The clustering of segments (with a given algorithm) has been performed in the past and compared to the actual labels of the pixels. Furthermore, the clustering algorithm performs somewhat uniformly across segments; that is, the rates at which different purities of clusters are generated is approximately the same from segment to segment.

Sections 2 and 3 present the development of estimators for the proportion estimation of categories within a cluster. The estimator is then applied separately to each cluster to obtain segment-level proportion estimates. The MSE is obtained in the same manner. Remarks in section 4 give additional information about obtaining segment-level estimates.

Within a cluster, the true proportion of category i is denoted θ_i, the estimated proportion, $\hat{\theta}_i$, and x_i denotes the number of pixels labeled as category i.

2. THE THREE-CATEGORY CASE

In the three-category case, $\theta_1 + \theta_2 + \theta_3 = 1$, and the conditional distribution of x_1, x_2, and x_3 is

$$f(x_1,x_2,x_3|\theta_1,\theta_2,\theta_3) = \frac{(x_1 + x_2 + x_3)!}{x_1!x_2!x_3!}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3} = M_0\theta_1^{x_1}\theta_2^{x_2}(1 - \theta_1 - \theta_2)^{x_3}$$

where $\theta_i\varepsilon(0, 1)$ and $x_i\varepsilon(0, 1, \cdots)$. This is a multinomial model: a generalization of the binomial model used in the two-category case.

We assume that, from previous experience with the clustering algorithm, the distribution of the array $(\theta_1,\theta_2,\theta_3)$ of cluster proportions can be modeled as

$$g(\theta_1,\theta_2,\theta_3) = K_0\theta_1^{a_1}\theta_2^{a_2}\theta_3^{a_3}$$

$$= K_0\theta_1^{a_1}\theta_2^{a_2}(1 - \theta_1 - \theta_2)^{a_3}$$

where $a_1,a_2,a_3 > -1 \quad ; \quad \theta_1\varepsilon\ [0,1],\ \theta_2\varepsilon\ [0,1 - \theta_1]$

and $$K_0 = \frac{\Gamma(a_1 + a_2 + a_3 + 3)}{\Gamma(a_1 + 1)\Gamma(a_2 + 1)\Gamma(a_3 + 1)}$$

The proofs that f and g are indeed probability density functions (pdf's) are given in section 3.

Now using the notation $\Theta = (\theta_1,\theta_2,\theta_3)$ and $X = (x_1,x_2,x_3)$

$$h(\Theta|X) = \frac{g(\Theta)f(X|\Theta)}{p(X)}$$

where $$p(X) = \int_0^1\int_0^{1-\theta_2} g(\Theta)f(X|\Theta)d\theta_1 d\theta_2$$

$$= M_0K_0\frac{\Gamma(x_1 + a_1 + 1)\Gamma(x_2 + a_2 + 1)\Gamma(x_3 + a_3 + 1)}{\Gamma(x_1 + x_2 + x_3 + a_1 + a_2 + a_3 + 3)}$$

Now $$\hat{\theta}_1 = E(\theta_1|X) = \int_0^1 \int_0^{1-\theta_2} \theta_1 h(\Theta|X) d\theta_1 d\theta_2$$

$$= \frac{x_1 + a_1 + 1}{x_1 + x_2 + x_3 + a_1 + a_2 + a_3 + 3}$$

$$\hat{\theta}_2 = E(\theta_2|X) = \int_0^1 \int_0^{1-\theta_2} \theta_2 h(\Theta|X) d\theta_1 d\theta_2$$

$$= \frac{x_2 + a_2 + 1}{x_1 + x_2 + x_3 + a_1 + a_2 + a_3 + 3}$$

$$\hat{\theta}_3 = E(1 - \theta_1 - \theta_2|X) = \int_0^1 \int_0^{1-\theta_2} (1 - \theta_1 - \theta_2) h(\Theta|X) d\theta_1 d\theta_2$$

$$= \frac{x_3 + a_3 + 1}{x_1 + x_2 + x_3 + a_1 + a_2 + a_3 + 3}$$

Assuming $N_0 = x_1 + x_2 + x_3$ is fixed, expressions are easily derived for the bias, variance, and mean square error (MSE):

$$A_0 = a_1 + a_2 + a_3$$

$$\hat{\theta}_i = \frac{x_i + a_i + 1}{N_0 + A_0 + 3}$$

$$E(\hat{\theta}_i) = \frac{N_0\theta_i + a_i + 1}{N_0 + A_0 + 3}$$

$$\text{bias }(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i) = \frac{a_i + 1 - \theta_i(A_0 + 3)}{N_0 + A_0 + 3}$$

$$\text{Var }(\hat{\theta}_i) = E(\hat{\theta}_i - E\hat{\theta}_i)^2 = E \frac{x_i - N_0\hat{\theta}_i}{N_0 + A_0 + 3}^2$$

$$= \frac{N_0\hat{\theta}_1(1 - \hat{\theta}_2)}{(N_0 + A_0 + 3)^2}$$

$$\mathrm{MSE}\ (\hat{\theta}_i) = \mathrm{Var}\ \hat{\theta}_i + \left[\mathrm{bias}\ (\hat{\theta}_i)\right]^2$$

$$= \frac{N_0\hat{\theta}_i(1 - \hat{\theta}_i) + \left[a_i + 1 - \theta_i(A_0 + 3)\right]^2}{(N_0 + A_0 + 3)^2}$$

3. THE K-CATEGORY CASE

The K-category case is merely an extension of the three-category case. Proofs have been omitted from section 2 since they are special cases of those presented in this section.

We begin by assuming that the prior distribution, the distribution of the array $\Theta = (\theta_1, \theta_2, \cdots, \theta_k)$, can be modeled as a generalized beta pdf.

<u>Theorem 1</u>: The function

$$g(\Theta) = g(\theta_1, \cdots, \theta_k) = K \cdot \prod_1^k \theta_i^{a_i}$$

$$= K \cdot \left(1 - \sum_1^{k-1} \theta_j\right)^{a_k} \prod_1^{k-1} \theta_i^{a_i}$$

where $\sum_1^k \theta_i = 1$, $\theta_i > 0$ for $i \in \{1, \cdots, k\}$

$$\text{and } K = \frac{\Gamma\left(\sum_i^k (a_1 + 1)\right)}{\prod_1^k \Gamma(a_j + 1)}$$

is a probability density function for each set of $\{a_i\}$ such that $a_i > -1$, $i \in \{1, \cdots, k\}$.

<u>Proof</u>: The function is obviously nonnegative and continuous for $0 < \theta_i < 1$ for each i. Hence, it remains only to show that it integrates to 1. Notice that if $k = 2$, g reduces to the well-known beta pdf; i.e., the theorem is true for $k = 2$ since

$$\int_0^1 t^{a_i}(1 - t)^{a_2} dt = \frac{\Gamma(a_i + 1)\Gamma(a_2 + 1)}{\Gamma(a_1 + a_2 + 2)}$$

for any choices of a_1, and $a_2 > -1$.

From here, we proceed by induction on k. We assume that the theorem is true for k = n; i.e.,

$$\int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-\sum_{r=1}^{n-2}\theta_r} \prod_{i=1}^{n-1} \theta_i^{a_i} \left(1 - \sum_{j=1}^{n-1}\right)^{a_n} \prod_{p=1}^{n-1} d\theta_p$$

$$= \frac{\left(\prod_{i=1}^{n-1} \Gamma(a_i + 1)\right)\Gamma(a_n + 1)}{\Gamma\left[\sum_{i=1}^{n-1} (a_j + 1) + a_n + 1\right]}$$

for all choices of $a_1, a_2, \cdots, a_n > -1$.

We then use the substitution $t = \dfrac{\theta_n}{1 - \sum_{j=1}^{n-1} \theta_j}$ to evaluate the integral in question for the values $a_1, a_2, \cdots, a_{n+1}$ in the case k = n + 1. This integral is given in equation (3-1) on page 8.

By the substitution of the values a_n, a_{n+1}, in the known case k = 2, and the values $a_1, a_2, \cdots, a_{n+1}$, and $(a_n + a_{n+1} + 1)$ in the assumed induction hypothesis, this integral reduces to

$$\frac{\left[\prod_{i=1}^{n-1} \Gamma(a_i + 1)\right]\Gamma(a_n + a_{n+1} + 2)}{\Gamma\left[\sum_{j=1}^{n-1} (a_j + 1) + a_n + a_{n+1} + 2\right]} \cdot \frac{\Gamma(a_{n+1})\Gamma(a_{n+1}+1)}{\Gamma(a_n + a_{n+1} + 2)} = \frac{\prod_{i=1}^{n} \Gamma(a_i + 1)}{\Gamma\left[\sum_{j=1}^{n} (a_j + 1)\right]}$$

and hence, $g(\Theta)$ is a pdf. QED.

The parameters $\{a_i\}$ are to be determined by an empirical fitting procedure using previous experience with the clustering algorithm.

$$\int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-\sum_{r=1}^{n-1} \theta_r} \left(\prod_{i=1}^{n} \theta_i^{a_i} \right) \left(1 - \sum_{i=1}^{n} \theta_i \right)^{a_{n+1}} \prod_{p=1}^{n} d\theta_p$$

$$= \int_0^1 \int_0^{1-\theta_i} \cdots \int_0^{1-\sum_{r=1}^{n-1} \theta_r} \left(\prod_{i=1}^{n-1} \theta_i^{a_i} \right) \int_0^{1-\sum_{q=1}^{n-1} \theta_q} \theta_n^{a_n} \left[\left(1 - \sum_{i=1}^{n-1} \theta_i \right) - \theta_n \right]^{a_{n+1}} d\theta_n \prod_{p=1}^{n-1} d\theta_p$$

$$= \int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-\sum_{r=1}^{n-2} \theta_r} \left(\prod_{i=1}^{n-1} \theta_i^{a_i} \right) \left(1 - \sum_{i=1}^{n-1} \theta_i \right)^{a_n + a_{n+1}} \int_0^{1-\sum_{q=1}^{n-1} \theta_q} \left(\frac{\theta_n}{1 - \sum_{i=1}^{n-1} \theta_i} \right)^{a_n} \left(1 - \frac{\theta_n}{1 - \sum_{i=1}^{n-1} \theta_i} \right)^{a_{n+1}} d\theta_n \prod_{p=1}^{n-1} d\theta_p$$

$$= \int_0^1 \int_0^{1-\theta_i} \cdots \int_0^{1-\sum_{r=1}^{n-2} \theta_r} \left(\prod_{i=1}^{n-1} \theta_i^{a_i} \right) \left(1 - \sum_{i=1}^{n-1} \theta_i \right)^{a_n + a_{n+1} + 1} \int_0^1 t^{a_n} (1-t)^{a_{n+1}} dt \prod_{p=1}^{n-1} d\theta_p \qquad (3\text{-}1)$$

The conditional distribution of the observed frequencies $X = (x_1, x_2, \cdots, x_k)$, given the true proportions $\Theta = (\theta_1, \cdots, \theta_k)$, is the well-known multinomial distribution:

$$f(X|\Theta) = M \cdot \prod_i^k \theta_i^{x_i} = M \cdot \left(1 - \sum_1^{k-1} \theta_j\right)^{x_k} \prod_1^{k-1} \theta_i^{x_i}$$

where $0 < \theta_i < 1$, $\sum_i^k \theta_i = 1$, $x_i \in \{0, 1, 2, \cdots\}$ and

$$M = \frac{\Gamma\left(1 + \sum_i^k x_i\right)}{\prod_i^k \Gamma(x_j + 1)} = \frac{\left(\sum_i^k x_i\right)!}{\prod_1^k x_j!}$$

Now the posterior distribution of Θ is

$$h(\Theta|X) = \frac{g(\Theta)f(X|\Theta)}{p(X)}$$

where
$$p(X) = K \cdot M \cdot \frac{\prod_i^k \Gamma(x_i + a_i + 1)}{\Gamma\left[\sum_i^k (x_j + a_j + 1)\right]}$$

Theorem 2: The marginal distribution of X is p, given above, when the prior distribution of Θ is the generalized beta pdf given by g, and the conditional distribution of X is the multinomial f.

Proof: The joint distribution of X and Θ can be expressed in terms of g and f

as
$$t(\Theta,X) = g(\Theta) \cdot f(X|\Theta)$$

and
$$p(X) = \int t(\Theta,X)d\Theta = \int g(\Theta)f(X|\Theta)d\Theta$$

$$= \int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-\sum_1^{k-1}\theta_i} g(\Theta)f(X|\Theta)d\theta_{k-1} \cdots d\theta_2 d\theta_1$$

where $$g(\Theta)f(X|\Theta) = K \cdot M \cdot \left(1 - \sum_1^{k-1} \theta_j\right)^{x_k + a_k} \prod_1^{k-1} \theta_i^{x_i + a_i}$$

From the induction hypothesis proven in Theorem 1, it is seen that $g(\Theta)f(X|\Theta)$ integrates to

$$P(X) = K \cdot M \cdot \frac{\prod_{i=1}^{k} \Gamma(x_i + a_i + 1)}{\Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1)\right]}$$

Now, using the same integration techniques, we derive the estimators.

<u>Theorem</u> 3: For f and g, as defined above, and using $N = \sum_{i=1}^{k} x_i$ and $A = \sum_{i=1}^{k} a_i$,

$$\hat{\theta}_p = E(\theta_p|X) = \int \theta_p h(\Theta|X)d\Theta$$

$$= \frac{x_p + a_p + 1}{\sum_1^k (x_i + a_i + 1)} = \frac{x_p + a_p + 1}{N + A + k}$$

for each $p \in \{1, \cdots, k\}$.

<u>Proof:</u> It can be seen that

$$\int \theta_p \, h(\Theta|X)d\Theta = \frac{\theta_p \, \Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1)\right]\left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{x_k + a_k} \prod_{i=1}^{k} \theta_i^{x_i + a_i}}{\prod_{i=1}^{k} \Gamma(x_i + a_i + 1)}$$

Thus $$\hat{\theta}_p = \frac{\Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1)\right]}{\prod_{k=1}^{k} \Gamma(x_i + a_i + 1)} \cdot \frac{\left[\prod_{\substack{i=1\\ i\neq p}}^{k} \Gamma(x_i + a_i + 1)\right]\Gamma(x_p + a_p + 2)}{\Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1) + x_p + a_p + 2\right]}$$

$$= \frac{\Gamma(x_p + a_p + 2)}{\Gamma(x_p + a_p + 1)} \cdot \frac{\Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1)\right]}{\Gamma\left[\sum_{i=1}^{k} (x_i + a_i + 1) + 1\right]} = \frac{x_p + a_p + 1}{\sum_{i=1}^{k} (x_i + a_i + 1)}$$

QED

The MSE of the Bayes posterior estimator is easily derived:

$$E(\hat{\theta}_i) = \frac{N\theta_i + a_i + 1}{N + A + k}$$

$$\text{bias } (\hat{\theta}_i) = \frac{a_i + 1 - \theta_i(A + k)}{N + A + k}$$

$$\text{Var } (\hat{\theta}_i) = E[\hat{\theta}_i - E(\hat{\theta}_i)]^2 = E\left[\frac{x_i - N\hat{\theta}_i}{N + A + k}\right]^2$$

$$= \frac{N\hat{\theta}_i(1 - \hat{\theta}_i)}{(N + A + k)^2}$$

$$\text{MSE } (\hat{\theta}_i) = \frac{N\hat{\theta}_i(1 - \hat{\theta}_i) + [a_i + 1 - \theta_i(A + k)]^2}{(N + A + k)^2}$$

4. REMARKS

The cluster-specific results presented in sections 2 and 3 can be assimilated into segment-level statistics by the following equations:

s = the number of clusters or strata

M_q = the number of pixels (samples) in cluster (strata) q

ToT = the total number of pixels in the segment

$$= \sum_{i=1}^{s} M_i$$

$\theta_{i,q}$ = the true proportion of category i in strata q

P_i = the proportion of pixels in category i in the segment

$$= \sum_{q=1}^{s} \frac{M_q}{ToT} \cdot \theta_{i,q}$$

$\hat{\theta}_{i,q}$ = the estimated proportion of category i in strata q

$\hat{P}_i$ = the estimated proportion of category i in the segment

$$= \sum_{q=1}^{s} \frac{M_q}{ToT} \cdot \hat{\theta}_{i,q}$$

$$\text{bias}\ (\hat{P}_i) = \sum_{q=1}^{s} \frac{M_q}{ToT}\ \text{bias}\ (\hat{\theta}_{i,q})$$

$$\text{Var}\ (\hat{P}_i) = \sum_{q=1}^{s} \left(\frac{M_q}{ToT}\right)^2 \text{Var}\ (\hat{\theta}_{i,q})$$

$$\text{MSE}(\hat{P}_i) = \text{Var}(\hat{P}_i) + [\text{bias}(\hat{P}_i)]^2$$

$$= \sum_{q=1}^{s} \left(\frac{M_q}{ToT}\right)^2 \text{Var}(\hat{\theta}_{i,q}) + \left[\sum_{j=1}^{s} \frac{M_j}{ToT} \cdot \text{bias}(\hat{\theta}_j)\right]^2$$

$$= \sum_{q=1}^{s} \left(\frac{M_q}{ToT}\right)^2 \text{Var}(\hat{\theta}_{i,q}) + \sum_{i=1}^{s} \left(\frac{M_j}{ToT}\right)^2 [\text{bias}\ \hat{\theta}_{i,q}]^2$$

$$+ \sum_{q=1}^{s} \sum_{j=q+1}^{s} 2\frac{M_q M_j}{ToT^2} \text{bias}(\hat{\theta}_{i,q})\ \text{bias}(\hat{\theta}_{i,j})$$

$$= \sum_{q=1}^{s} \left(\frac{M_q}{ToT}\right)^2 \text{MSE}(\hat{\theta}_{i,q})$$

$$+ \sum_{q=1}^{s} \sum_{j=q+1}^{s} 2\frac{M_q M_j}{ToT^2} \text{bias}(\hat{\theta}_{i,q})\ \text{bias}(\hat{\theta}_{i,j})$$

One application of the theory developed in this report is to randomly select a predesignated number of pixels from a segment, note the pixel labels and breakdown by clusters, and implement the Bayesian approach (above) to calculate $\hat{\theta}_{i,q}(i = 1, \cdots, k)$, $\hat{P}_i$, and MSE $(\hat{P}_i)$. One problem with this approach is that each cluster may not contain two samples; thus, MSE $(\hat{\theta}_{i,q})$ cannot be estimated, and the MSE evaluation of the estimator, $\hat{P}_i$, will not exist in this case. Another problem is that the samples may be inefficiently allocated to obtain a small MSE $(\hat{P}_i)$. In an attempt to resolve these problems, the alternate sampling strategy of sampling in proportion to cluster size can be used. Again, however, since the MSE $(\hat{\theta}_{i,q})$ is a function of cluster size, number of samples, and the proportion $\theta_{i,q}$, the optimal sampling strategy will depend on cluster purity (as well as size). Sampling in proportion to cluster size cannot be optimal. The following approach is a first attempt at addressing the problem of stratified sampling within a segment.

In the two-category case, two samples were selected from each cluster (to insure an estimate of the variance). Then, additional samples were selected sequentially so that, at each sampling, the sample was selected from the cluster that was expected to maximally minimize the weighted cluster MSE for the one proportion estimate. The weighting is the square of the cluster size as a proportion of the segment. Therefore, the expected change for each cluster q is as follows.

$$\hat{\theta}_{i,q} = \hat{\theta}_{i,q}(n,x) = \frac{x_{i,q} + a_i + 1}{N_q + A + k} \quad ; \quad k = 2$$

$$\mathrm{MSE}^* [\hat{\theta}_{i,q}(n,x)] = \left(\frac{M_q}{\mathrm{ToT}}\right)^2 \left\{ \frac{N\hat{\theta}_{i,q}(n,x)[1 - \hat{\theta}_{i,q}(n,x)]}{(N_q + A + k)^2} + [a_i + 1 - \hat{\theta}_{1,q}(n,x) \cdot (A + k)]^2 \right\}$$

$$\Delta\mathrm{MSE}^* = \mathrm{MSE}^*[\hat{\theta}(n,x)] - [1 - \hat{\theta}(n,x)] + \mathrm{MSE}^*[\hat{\theta}(n + 1,x)] - \hat{\theta}(n,x) \cdot \mathrm{MSE}^*[\hat{\theta}(n + 1, x + 1)]$$

Notice that $\Delta\mathrm{MSE}^*$ is a function of the crop being estimated, though this is hidden since there are only two categories. In the k-category case, this dependence can be averaged out for each cluster q by using

$$\Delta = \sum_{i=1}^{k} \Delta\mathrm{MSE}^*(\hat{\theta}_{i,q})$$

For $k = 2$, $\Delta = \Delta\mathrm{MSE}^*(\theta_i, q)$ for either i, and this problem does not exist.

Also, although $\Delta\mathrm{MSE}^*$ is the weighted cluster MSE, it does not exactly represent the cluster contribution to the segment MSE: $\mathrm{MSE}(\hat{P}_i)$. It would be preferable to calculate a $\Delta\mathrm{MSE}(\hat{P}_i)$ for each cluster and sampling, but earlier experiments used $\Delta\mathrm{MSE}^*(\theta_{i,q})$ as a computational expedience and an approximation to $\Delta\mathrm{MSE}(\hat{P}_i)$.

The exact relationship of the two is given in the last $\mathrm{MSE}(\hat{P}_i)$ equation given above.

The ΔMSE criterion, either $\Delta MSE^*(\hat{\theta}_{i,q})$ or $\Delta MSE(\hat{P}_i)$, would appear to be the optimum approach in extending to multicategory ($k > 2$) proportion estimation also. The unresolved issue is the determination of which categories to include and by what weighting.

That is

$$\Delta MSE(\hat{P})q = \sum_{i=1}^{k} \alpha_i \ \Delta MSE(\hat{P}_{i,q})$$

$$\alpha_i > 0, \ \sum_{i=1}^{k} \alpha_i = 1$$

or

$$\Delta MSE^*(\hat{\theta})q = \sum_{i=1}^{k} \alpha_i \ \Delta MSE^*(\theta_{i,q})$$

The weightings $\{\alpha_i\}$ will determine the relative importance of the respective crops, or vice versa. Another possibility would be to select from the cluster q with the largest $\Delta MSE^*(\hat{\theta}_{i,q})$, $i = 1, 2, \cdots, k$. The particular criterion selected should be tailored to each specific application and determined through empirical studies.

5. SUMMARY

A Bayesian technique for stratified proportion estimation is presented for the multicategory case, and detailed equations are derived for the case of a generalized beta prior distribution. Additionally, a technique of sequentially sampling from the clusters to achieve minimum mean squared error segment proportion estimates for the categories of interest was presented, and some computational issues were identified.

6. REFERENCES

1. Lennington, R. K.; and Johnson, J. K.: Clustering Algorithm Evaluation and the Development of a Replacement for Procedure 1. LEC-13945, NASA/JSC (Houston), November 1979.

2. Pore, M. D.: Bayesian Techniques in Stratified Proportion Estimation. LEC-13940, American Statistical Association, 1979 Proceedings of the Business and Economic Statistics Section, August 1979.