Reprinted from

Eighth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Crop Inventory and Monitoring

July 7-9, 1982

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

SAMPLE DESIGN WITH IRREGULAR SAMPLING UNITS FOR A CROP PROPORTION ESTIMATION PROCEDURE BASED ON LANDSAT DATA

T.G. LYCTHUAN-LEE

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

ABSTRACT

This paper presents a sample design for the "Advanced Proportion Estimation Procedure" (APEP) developed for estimation of the population proportion of a crop or a crop class of interest. The sample design is tailored to the case in which the sampling units vary in size and shape. The precision of the estimator is discussed along with a comparison of the sampling scheme with the one proposed by Hartley-Rao-Cochran. A brief description of "APEP" is also presented.

I. BACKGROUND AND OBJECTIVE

A. TARGET POPULATION AND IRREGULAR SAMPLING UNITS

The target population is an agricultural area which is represented primarily by satellite data. The size of the area may be large or small. This area is a region on the earth's surface remotely sensed by a sensor based on the space shuttle or other space platform. The target population might be a five-by-six nautical mile areal segment, or a stratum from a stratification of a region under study, or a domain-of-study of the region, or even the region itself.

One of the characteristics of some labeling/classification procedures which are utilized in the process of estimating a crop proportion is that they could work well on any non-regular units.

Sampling units in the target population could vary in size and shape. This paper proposes a sample design dealing with such irregular sampling units. The target population itself (by its image) is composed of M elementary units which are regular in size and shape, and called "pixels." ["Pixel" stands for the word "picture element" which is originally defined as "the dual record of a single data-take during the scan of an electro-optical scanning sensor...and referenced by scan line and column. The images recorded by scanning sensors are comprised of picture elements. For example, a sample or number from the Landsat multispectral scanner representing reflected solar radiation from an area 79 square meters on the surface of the earth."[1]]

Those pixels can be clustered by a certain procedure (algorithm) into N spectral-spatial clusters. The N spectral-spatial clusters exhaustively partition the whole target population into non-overlapping sub-regions. A "cluster" (in short for "spectral-spatial cluster") is a group of contiguous pixels which could be "similar" in a certain spectral parameter. In terms of crops of interest, each cluster is hopefully not too heterogeneous and neither too large nor too small, so that it is possible to easily and precisely identify λ_{jr}, the proportion of crop j in cluster r, where $0 \leq \lambda_{jr} \leq 1$. A cluster could be composed of only one crop type [pure crop j-cluster, say; hence $\lambda_{jr} = 1$], or many crop types including the crop j of interest [non-pure crop j-cluster; hence $0 < \lambda_{jr} < 1$], or many crop types other than the crop j [pure non-crop j-cluster; hence $\lambda_{jr} = 0$]. Clusters are the sampling units of this sample design. Hence, any random sample of n clusters, chosen out of N clusters of the population, could be composed of the above three types of clusters.

B. OBJECTIVE

The objective of this paper is to present a sample design for the "Advanced Proportion Estimation Procedure" developed for estimation of the population proportion P_j of a crop j or some crops of interest.

The paper proposes a very simple and practical sampling scheme which has two steps to select n clusters out of N clusters of the target population. The proportion estimator $\hat{P}_j$ and its variances will be discussed, and the sampling scheme will be compared with the one proposed by Hartley-Rao-Cochran[2].

Two cases of λ_{jr} are considered. In the first case, λ_{jr} are the population proportions of crops j in cluster r. The APEP system identifies λ_{jr}, the relative sizes to be crops j proportions based on the whole population of M_r pixels in cluster r. In the second case, a second-stage of the sampling scheme is introduced: select a random sample of size m_r out of M_r pixels of the defined cluster r. The APEP system will work on the m_r sampled pixels to produce the estimates $\hat{\lambda}_{jr}$ of λ_{jr} with the variances $V(\hat{\lambda}_{jr})$.

II. THE "ADVANCED PROPORTION ESTIMATION PROCEDURE"

A BRIEF DESCRIPTION[3,4]

Objective. The objective of the "Advanced Proportion Estimation Procedure" (APEP) is to estimate a crop proportion (hence, acreage) without requiring manually identified training samples.

General Approach. The basic idea of the APEP is to explore scientific relationships between various agrometeorological parameters and their spectral appearances.

The major functional elements of APEP are as follows:

1. Profile models are models describing the growth and decay of spectrally derived variables (such as greenness, a satellite -- Landsat-variable) as functions of times. Features are selected from these temporal profiles.

The feature selection maps the observable variables into "feature variables" which describe the crop-related physical processes of interest. These feature variables are less contaminated by "noise" effects than the original variables.

2. As a particular case, if x is the vector of Landsat observations with the observed spectral distribution f(x), and t is of event variables having distributions $g_i(t)$ for crops i, i=1,2,..., M, which are transformed to class j -- observed spectral component distributions $f_j(x)$, j=1,...,M, by a convolution noise operator $h(\varepsilon)$, then x and t are related by the relationship

$$x = t + \varepsilon \qquad (1)$$

3. A statistical model, called a mixture model, is then used to estimate directly the proportion of the crops of interest. The general form of the mixture model is

$$f(x) = \sum_{j=1}^{M} \lambda_j f_j(x) \qquad (2)$$

where:

x is a vector of Landsat observations

f(x) is the density distribution of x for the target population (segment, for example)

$f_j(x)$ is the density distribution of x of the class j, j=1,...,M

λ_j is the size of $f_j(x)$ taken as the proportion of the area making up the j^{th} class

M is the number of classes present in the target population

4. To complete the process, each proportion estimate must be given a crop name. This is the labeling function. Labeling is achieved by matching up a predicted growth curve for a given crop, g_j^* say, derived from meteorological data, with one of the growth curves estimated from the mixture model.

III. SAMPLING PLAN

A. SAMPLING PROCEDURE

The sampling procedure proposed here is based on the assumption that there exists a certain algorithm, "clustering-algorithm," say, which has the capacity of clustering pixels being contiguous and "similar" in a certain spectral parameter into "clusters." The algorithm supposedly may be utilized to partition a target population into N clusters. The sampling scheme is comprised of two steps to select n clusters out of supposed N clusters of the target population.

Step 1. Apply the simple random sampling scheme to select n pixels out of M pixels of the target population.

Step 2. Cluster the pixels similar to each of those pixels sampled.

B. HARTLEY-RAO-COCHRAN'S UNEQUAL PROBABILITY WITHOUT REPLACEMENT SINGLE STAGE SAMPLING PROCEDURE

In [2] J. N. K. Rao, H. O. Hartley, and W. G. Cochran have proposed:

1. First split the target population at random into n groups of sizes $N_1, N_2, \ldots, N_h, \ldots, N_n$

where: $$\sum_{h=1}^{n} N_h = N$$

2. From each of these n groups, independently draw a sample of size one with probabilities proportional to p_t, the probability for drawing the t^{th} unit in the first draw from the whole population.

Note: "If the t^{th} unit falls in group h, the actual probability that it will be selected is p_t/π_h where

$$\pi_h = \sum_{\text{group } h} p_t \text{"}^2$$

C. PROBABILITY STATEMENTS

In returning to the sampling scheme proposed by this paper, the target region is supposedly comprised of N clusters and now imaginarily partitioned into n sub-regions. A sub-region h supposedly has N_h clusters where

$$\sum_{h=1}^{n} N_h = N.$$

Each, h say, of n sampled pixels is in a corresponding cluster, (t) say, created by the clustering algorithm. And such cluster (t) is imaginarily belonging to a subregion h of the population, h=1,2..., n. Each cluster (t) has M_{th} pixels where

$$\sum_{t=1}^{N_h} M_{th} = M_h^{(t)} = \text{total number of pixels in the imaginarily-created sub-region h related to cluster (t),}$$

$$\sum_{h=1}^{n} M_h^{(t)} = M = \text{the total number of pixels in the target population} \qquad (3)$$

Hence, the target region has been split at random into n sub-regions of sizes $N_1, \ldots, N_h, \ldots, N_n$ clusters mentioned in step (1) of Hartley-Rao-Cochran sampling procedure. And each sampled pixel happens to be in a created cluster which is considered as a member of a random sample of n clusters. Such a cluster will be proved here to correspond to a result of step (2) of Hartley-Rao-Cochran sampling scheme. The probability of including such a cluster (t) given a sampled pixel {h} is

$$\Pr[(t)\,|\,\{h\}] = \frac{M_{th}}{M_h^{(t)}} \qquad (4)$$

The probability of choosing such a pixel {h} which is associated with the sub-region h is

$$\Pr[\{h\}] = \frac{M_h^{(t)}}{M} \equiv \pi_h \qquad (5)$$

Therefore the probability of drawing the cluster (t) generated from the sampled pixel {h} to be in the random sample of n clusters in the first draw from the whole population will be

$$\Pr[\{h\}] \cdot \Pr[(t)\,|\,\{h\}] = \frac{M_{th}}{M} \equiv p_t \qquad (6)$$

Hence, in being related with p_t and π_h, the probability of including a cluster (t) given the corresponding sampled pixel {h} in any draw will be

$$\Pr[(t)\,|\,\{h\}] = \frac{p_t}{\pi_h} \qquad (7)$$

These probabilities statements are the same given by the Hartley-Rao-Cochran sampling scheme where they apply the Hansen-Hurwitz procedure in their second step. It will be demonstrated as follows.

After placing the first N_1 clusters into group 1, the second N_2 clusters into group $2, \ldots, N_h$ into group h, ..., and the last N_n clusters into group n, within each group h a cluster (t) will have its relative sizes defined as follows:

	Relative sizes
	$p_1 \equiv p_{1h} = \frac{M_{1h}}{M}$
	...
	$p_t \equiv p_{th} = \frac{M_{th}}{M}$
	...
	$p_{N_h} \equiv p_{N_h h} = \frac{M_{N_h h}}{M}$
Total	$\pi_h = \sum_{t=1}^{N_h} \frac{M_{th}}{M} = \frac{\sum_{t=1}^{N_h} M_{th}}{M}$ (8)

Then the Hansen-Hurvitz procedure continues on selecting at random a random number R from 1 to π_h, $1<R<\pi_h$, and the cluster (t) is selected for

which

$$\sum_{i=1}^{t-1} p_i < R < \sum_{i=1}^{t} p_i \qquad (9)$$

This procedure gives to the t^{th} cluster in the population a probability of selection proportional to size M_{th}, where the probability of selecting the t^{th} cluster at any draw is

$$\Pr[\text{Selecting } t^{th} \text{ cluster at any draw}] = \frac{p_t}{\pi_h}, \qquad (10)$$

since R having a uniform distribution on $(0,\pi_h)$ has

$$\Pr[\sum_{i=1}^{t-1} p_i < R < \sum_{i=1}^{t} p_i] = \frac{1}{\pi_h}\int_{\sum_i^{t-1} p_i}^{\sum_i^{t} p_i} du = \frac{p_t}{\pi_h} \qquad (11)$$

which is the same as $\Pr[(t)|\{h\}]$ in (7). Therefore the following estimation formulae are easily proved.

IV. PROPORTION ESTIMATION

A. THE CASE OF THE POPULATION PROPORTION λ_{jr} IDENTIFIED DIRECTLY BY APEP

(1)(i). The unbiased estimator of the proportion of the crop j in the target population is

$$\hat{P}_j = \sum_{r=1}^{n} \frac{M_h^{(r)}}{M}\lambda_{jr} \qquad (12)$$

where $M_h^{(r)}$ is defined in (5), λ_{jr} is the proportion of crop j in cluster r given by APEP.

Proof.

The unbiased estimator of the population total number X of pixels which are of crop j is given in [2] as

$$\hat{X}_j = \sum_{r}^{n} \frac{X_{jr}}{P_r/\pi_h} \qquad (13)$$

where

$X_{jr} = M_{rh}\lambda_{jr}$ are the total number of crop j pixels in cluster (r)

$$\frac{P_r}{\pi_h} = \frac{M_{rh}}{M_h^{(r)}}$$

Hence, the unbiased estimator of the proportion of the crop j in the population is

$$\hat{P}_j = \sum_{r}^{n} \frac{M_h^{(r)}}{M}\lambda_{jr} \qquad (14)$$

(1)(ii). Since $M_h^{(r)}$ does not exist in the case of this paper, however $M, M_h^{(r)}$, and n can be so large that the sampling scheme of choosing n pixels first then locating them in n sub-regions can be treated as a similar sampling scheme mentioned in [5].

That is a post-stratified large sample of n pixels (pixels are treated as sampling units) out of M pixels [whereat the sample stratum size which is n_h is one, the population stratum size which is N_h is $M_h^{(r)}$], hence the basic theorem in [5] states that

$$\frac{M_h^{(r)}}{M} \simeq \frac{1}{n} \qquad (15)$$

when M, $M_h^{(r)}$, and n are large.

With this result, the simple form of the unbiased estimator of the proportion of crop j in the population according to the sampling scheme proposed by this paper will be

$$\hat{P}_j^{*} \simeq \frac{1}{n}\sum_{r=1}^{n}\lambda_{jr}$$

(2)(i). The population variance of the proportion estimator $\hat{P}_j$ of P_j will be

$$V(\hat{P}_j) = \frac{\sum^{n} N_h^2 - N}{N(N-1)}\left(\sum_{i}^{N} \frac{M_{ih}}{M}\lambda_{ji}^2 - P_j^2\right) \qquad (17)$$

Proof.

The population variance of the estimator $\hat{X}$ is given in [2] as

$$V(\hat{X}_j) = \frac{\sum^{n} N_h^2 - N}{N(N-1)}\left(\sum_{i}^{N} \frac{X_{jr}^2}{P_i} - X_j^2\right) \qquad (18)$$

Hence,

$$V(\hat{P}_j) = \frac{\sum^{n} N_h^2 - N}{N(N-1)}\left(\sum_{i}^{N} \frac{M_{ih}}{M}\lambda_{ji}^2 - P_j^2\right) \qquad (19)$$

(2)(ii). The unbiased estimator of $V(\hat{P}_j)$ is

$$v(\hat{P}_j) = \frac{\sum^{n} N_h - N}{N^2 - \sum^{n} N_h^2}\left(\sum_{r}^{n} \frac{M_{rh}}{M}\lambda_{jr}^2 - \hat{P}_j^2\right) \qquad (20)$$

Proof.

The unbiased estimator of $V(\hat{X}_j)$ is given in [2] as

$$v(\hat{X}_j) = \frac{\sum^{n} N_h^2 - N}{N^2 - \sum^{n} N_h^2}\left(\sum_{r}^{n} \frac{X_{jr}^2}{P_r} - \hat{X}_j^2\right) \qquad (21)$$

Hence,

$$v(\hat{P}_j) = \frac{\sum_{h}^{n} N_h^2 - N}{N^2 - \sum_{n}^{n} N_h^2} \left(\sum^{n} \frac{M_{rh}}{M} \lambda_{jr}^2 - \hat{P}_j^2 \right) \quad (22)$$

(2)(iii). In applying the same principle in [5] on the random sample of n clusters in the post-stratified sampling scheme where $n_h=1, n=n$, $\frac{N_h}{N} \simeq \frac{n_h}{n} = \frac{1}{n}$,

$$\frac{\sum_{h}^{n} N_h^2 - N}{N^2 - \sum_{n}^{n} N_h^2} \simeq \frac{\frac{n}{\Sigma}(\frac{1}{n})^2}{1 - \frac{n}{\Sigma}(\frac{1}{n})^2} = \frac{1}{n-1} \; ; \quad (23)$$

henceforth, the estimated variance of $\hat{P}_j^*$ will be

$$v(\hat{P}_j^*) \simeq \frac{1}{n-1} \left(\sum^{n} \frac{M_{rh}}{M} \lambda_{jr}^2 - \hat{P}_j^{*2} \right) \quad (24)$$

B. THE CASE OF THE POPULATION PROPORTION λ_{jr} ESTIMATED BY APEP BASED ON A SECOND-STAGE SAMPLE

If for any reason (such as clusters that are too large or too heterogeneous) APEP will not work efficiently on the whole clusters but on a random sample of pixels chosen out of each sampled cluster, then a second-stage sampling plan needs to be introduced. The APEP working on this second-stage sample of pixels would give the estimator $\tilde{\lambda}_{jr}$ with its expectation $E_2[\tilde{\lambda}_{jr}]$ and its variances $V_2(\tilde{\lambda}_{jr})$ as well as $v_2(\tilde{\lambda}_{jr})$ which is the estimator of $V_2(\tilde{\lambda}_{jr})$.

This part of the paper will not specify any particular second-stage sampling plans but only discuss a general one. However, in any case, the proportion estimator of P_j will be

$$\hat{P}_j^* \simeq \frac{1}{n} \sum_{r=1}^{n} \tilde{\lambda}_{jr}, \quad (25)$$

where $\tilde{\lambda}_{jr}$ are the estimators of λ_{jr} given by the APEP working on the second-stage sample.

The expectation and variance of $\hat{P}_j^*$ will be computed according to the following general formulae:

$$E[\hat{P}_j^*] = \underset{12}{E}[\hat{P}_j^*] = \underset{1}{E}[\frac{1}{n} \sum_{r=1}^{n} \underset{2}{E}[\tilde{\lambda}_{jr}]], \quad (26)$$

$$V(\hat{P}_j^*) = \underset{1}{E}[\underset{2}{V}(\hat{P}_j^*)] + \underset{1}{V}[\underset{2}{E}(\hat{P}_j^*)], \quad (27)$$

where E and V are the expectation and variance of $(\frac{1}{n} \sum_{r=1}^{n} \lambda_{jr})$ in the first-stage sampling plan mentioned in section (4.a).

Depending on the second-stage sampling scheme to be specified, the above formulae of the corresponding (two-stage) sample design will be chosen accordingly.

C. FOR ANOTHER PROPORTION ESTIMATION PROCEDURE

Now considered is the case where this sample design would support another estimation procedure still exploring satellite data in order to estimate the proportion of a crop j in a region, such a proportion estimation procedure could be based directly on counting the number of pixels which are of crop j in each defined cluster. In this case, the crop j proportion in cluster r will be

$$\lambda_{jr} = \frac{1}{M_r} \sum_{t=1}^{M_r} y_{tr} \quad (28)$$

where

$$y_{tr} = \begin{cases} 1 & \text{if the } t^{th} \text{ pixel of cluster selected at draw r is a pure crop j pixel} \\ \gamma_t & \text{if the } t^{th} \text{ pixel of cluster selected at draw r is a mixed pixel which proportion of crop j is } \gamma_t,\ 0<\gamma_t<1 \\ 0 & \text{if the } t^{th} \text{ pixel of cluster selected at draw r is a pure non-crop j pixel} \end{cases} \quad (29)$$

Note: It need not say the assumption of having solved the problem of proportion estimation in mixed pixels. Moreover, if a random sample of m_r pixels drawn from each defined cluster is used to estimate λ_{jr}, then the estimator $\tilde{\lambda}_{jr}$ will be

$$\tilde{\lambda}_{jr} = \frac{1}{m_r} \sum_{t=1}^{m_r} y_{tr} \; ; \quad (30)$$

and the sample design will be built on a two-stage sampling scheme.

There are many different combinations of two single sampling schemes to make up various two-stage sampling plans. Many such (two-stage) sample designs will be discussed in other papers.

V. CONCLUSIONS

The formulae of the estimator and its variance appear to be familiar and computationally simple.

This sample design offers a very simple and practical method of selecting clusters without formally listing all the clusters in the population before

randomly drawing them with probabilities proportional to the sizes of the clusters. However, it will be applicable only in the case of the large sample design where the population and sample sizes are large and the probability of the sub-region sizes being null is zero.

This sample design can support various estimation methods or estimators. The proportion estimator is used here for the purpose of demonstration as well as for achieving a goal of a certain project. The Advanced Proportion Estimation Procedure is also used here for the same purposes even though it is still under development.

This sample design should help to improve the results given by the APEP itself. Instead of working on the whole region, which may be too heterogeneous and which gives APEP many classes in its mixture model, APEP will need to work only on the clusters population, which is less heterogeneous and which gives APEP a few (maybe two) classes in its mixture model. Hence, this sample design is recommended to support the APEP to estimate the population proportion P_j of a crop j or some crops j, $j=1,1...,M$, of interest.

REFERENCES

[1] "First LACIE Dictionary of Remote Sensing Terminology" (working copy) (1976) National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas.

[2] Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). "On a Simple Procedure of Unequal Probability Sampling without Replacement," The Journal of the Royal Statistical Society, Series B (Methodological), V.24, No. 2, pp 482-491.

[3] Heydorn, R. P., et al (1981). "Project Implementation Plan for the Small Grains Consortium. The Advanced Estimation Procedure." NAS 9-15800, Lockheed-EMSCO-16853.

[4] FY 1981 AgRISTARS Annual Report. AP-J2-04225, January 1982, National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas

[5] Lycthuan-Lee, T. G., and Hallum, C. R. (to be published) "Post-Stratified Large Sample Design for the Estimation of a Crop Proportion."

Thomas Gayle Lycthuan-Lee is a scientist of Lockheed Engineering and Management Services Company. He was an Outstanding Awarded Mathematical Statistician (U.S.) Summer Intern in 1978 and a National Statistician of the Republic of Vietman before May 1975. In addition to university degrees from Vietnam, he has an M.A. in Mathematics (Statistics) from University of Texas at Austin, an M.S. in Statistics from Texas A & M University, and wrote his Ph.D. dissertation under Professor H. O. Hartley. Areas of special interest include survey design, linear model, and multivariate analysis.