Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 29 - July 1, 1976

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

USE OF LANDSAT TECHNOLOGY BY STATISTICAL REPORTING SERVICE

William H. Wigton, Statistical Reporting Service
U.S. Department of Agriculture, Washington, D. C. 20250

I. ABSTRACT

This paper describes an area sampling frame and defines the sampling error and bias of an estimate. LANDSAT data is explained in the Statistical Reporting Service framework and the essential components of computer classification are delineated. A procedure is presented that utilizes satellite data to improve an estimator with 3 percent sampling error.

II. AREA SAMPLING FRAME

The area sampling frame is used by the Statistical Reporting Service (SRS) to make production estimates at both state and national levels. In addition, the use of the area sampling frame is crucial for our application of LANDSAT technology to improve these crop acreage estimates. Therefore, it is essential to spend some time in explaining its function and use in greater detail.

The concepts of area frame sampling are very simple:

1. Divide the total area to be surveyed into N small contiguous blocks (i.e., segments) without any overlaps or omissions.
2. Select a random sample of n blocks.
3. Obtain the desired data for reporting units of the population that are in the sample blocks.
4. Estimate the population totals by multiplying the sample totals by $\frac{N}{n}$. 1/

This procedure, as outlined above, is used for crop acreage, livestock, and other farm data estimation, and is a dependable method. The use of random numbers in selecting a sample from the universe accomplishes two things:

1. It gives a basis for making inference about the total production of all farms in the U.S.
2. It provides a basis for the computation of sampling errors which will be discussed in the following section.

III. CALCULATION OF THE ACCURACY OF AN ESTIMATE

To determine the accuracy of any estimate, one requires the population target value or the actual number which is being estimated. With this information, it is not necessary to make the estimate of the target value. Therefore, it becomes mandatory to use some other method to evaluate an estimate. An example follows which illustrates the use of an area sampling frame as previously defined.

A state is divided into 30,000 pieces of land area and a random sample of 300 is selected. Data is obtained and an estimate of wheat acreage produced by multiplying the total wheat acreage by $\frac{30,000}{300} = 100$. If another 300 segments had been selected, the estimate would have been different. If the estimates do not vary considerably from one sample of size 300 to the next, then the estimate is fairly stable. However, if the estimates vary considerably, then we would conclude that our estimator has a large variance or sampling error. The variation of the estimate from one sample of 300 to other samples of 300 selected in the same manner is sampling error. Estimators with small sampling errors are most desirable. However, there is another criterion that is also important--the element of bias.

Bias

If there is a difference between the center of the distribution that defines sampling error and the true value being estimated, this difference is defined as bias.

Whether or not the true value being estimated is at the center of the sampling error distribution is controlled by:

1. The completeness of the sampling frame.

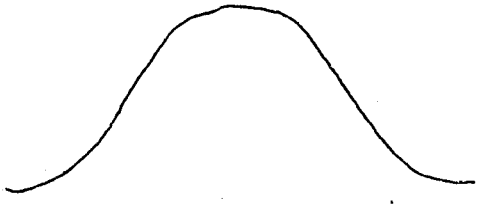
2. The importance of giving every element in the population a known positive chance of selection.

3. The use of high quality control standard of enumeration and other nonsampling errors.

If the estimator that is being generated by selecting 300 segments is centered around the true value and the variation is small, then our one estimate is an accurate one--one that is close to the true value.

Often, one cannot tell about sampling errors unless other samples of 300 segments are selected and enumerated. However, with proper sampling techniques the variation can be measured with only one sample. The segment to segment variation is used to calculate the sample to sample variation. In essence, sample to sample variation is estimated with only one sample.

Figure 1. Sampling Error Distribution



From one sample, then, the sampling distribution is estimated.

Let us assume that the distribution looks like the distribution curve illustrated in Figure 1. We do not know where in the distribution our sample lies. We only know that it was drawn from this distribution at random. We know, also, from the sampling procedure and the estimating formula that the statistic is unbiased. We have a better estimate if it comes from a distribution curve such as Figure 3, than from a curve such as Figure 2, because the values are clustered closer to the center.

Figure 2.

Figure 3.

If we improve the current estimates from the area frame with LANDSAT, then we must alter the distribution of the possible estimates by reducing the spread.

IV. APPLICATION OF LANDSAT CLASSIFICATION

Description of LANDSAT Data

The satellite data used in this report is LANDSAT Multi-Spectral Scanner (MSS) data and is described in Section 3 of data User's Handbook. 2/

The MSS is a passive electro-optical system that can record radiant energy from the scene being sensed. All energy coming to earth from the sun is either reflected, scattered, or absorbed, and subsequently, emitted by objects on earth. 3/ The total radiance from an object is composed of reflected radiance forms, a dominant portion of the total radiance from an object at shorter wavelengths of the electromagnetic spectrum, while the emissive radiance becomes greater at the longer wavelengths. The combination of these two sources of energy would represent the total spectral response of the object. This, then, is the "spectral signature" of an object and it is the differences between such signatures which allows the classification of objects using the statistical techniques about to be discussed.

Classification Techniques

Let us suppose that we wish to classify a LANDSAT frame. The way this is done in the computer is by use of discriminant functions. Computers must differentiate between crops on the basis of reflected energy. To start, we must have two or more crops and a sample of individual pixels for each. The problem is to set up a rule using the sample pixels for each crop, which will enable us to allot some unknown crop pixel outside the sample to the correct crop type given only the amount of reflected energy of that pixel.

This can be formulated statistically, but first let me introduce some notation.

If all data in a LANDSAT frame were plotted in a scatter diagram it might appear as Figure 4.

Figure 4. Scatter Diagram of All Values in One LANDSAT Frame for Three Crops. C-Corn, S-Soybeans, W-Water

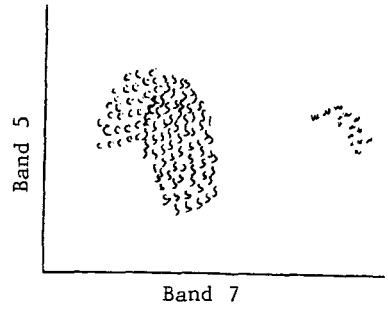
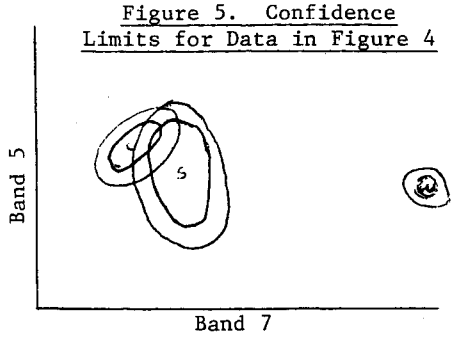


Figure 5 shows confidence limits for above data.

Figure 5. Confidence Limits for Data in Figure 4



If we study Figure 5, it does not take long to make some observations:

1. The location of the center of these concentric circles has a impact on how easy it is to set up the rules.

2. The data looks quite elliptical (often this is not the case for actual data).

3. The spread of the data varies considerably for the crops. Soybeans has wide variability for example.

4. It will be impossible to tell with certainty which crops we have, if the reflected energy comes from the overlap region of corn with soybeans, because both are possible.

5. It would be ideal if the data for each crop were as far apart as water from corn and if the spread were as small as water and elliptical in form and there were no areas of overlap.

However, it appears that these items are not under our control. The sensor (bands and bands width) determine the location of the centers of the spread of points.

The spread of the data and its contour are determined by factors such as soil conditions, varieties of crops, amount of fertilizer used, planting dates, atmospheric conditions NASA preprocessing, and many more things.

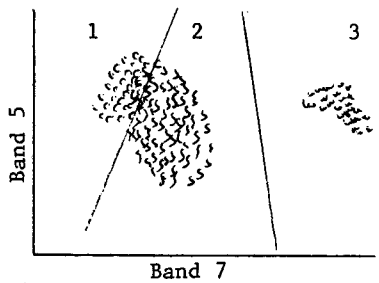
As far as the overlap areas, where mislabeling or misclassification is inevitable, nature herself is the problem. Some items that we would like to be able to tell apart reflect solar energy similarly. We look at these facts philosophically. It is not our business to change the nature of things but simply to estimate what is there.

The best we can hope for is to estimate from a sample the scatterdiagram of the population and this we know how to do if we treat it like anything else that we estimate.

We want a valid statistical estimate that requires a random sample from the population of interest. This requires that all parts of the population of interest must have a chance of selection and the size must be large enough to adequately represent the population. If the population structure is as complicated as water in Figure 4 or if estimates are needed that are quite accurate, as in corn and soybeans, then, a fairly substantial sample size is required.

The area sampling frame is perfect because a valid statistical estimate can be made for the LANDSAT frame since a random sample of all possible segments is available and reflected energy for the crop types can be determined for the sample fields inside the segments. These signatures are estimated for the scene they are in, so, it is valid to use these values for computer training of the discriminant functions. After population scatterdiagrams have been estimated, rules are set up to allot pixels with known energy readings but unknown crop labels to crop categories. Rules are simple; they amount to drawing lines that partition the space. Figure 6 shows an example of this.

Figure 6. Partitioned Space Showing Population Scatterdiagram

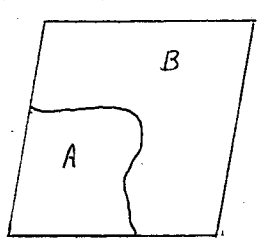


The rest is simple. All pixels that need crop labels should be plotted on the partitioned space. If they fall in partition one, give it a label of corn, even though some soybeans will creep in; obviously, we will do well with water.

Incidently, it turns out that the location, size and shape of these population scatterdiagrams shift relative to each other in different scenes and even different parts of the same scene. Hence, using a partition developed on one locale of a LANDSAT scene to label pixels from another locale is hazardous.

There are two cases, both are quite different. One is reasonable, and the other is not. Let us divide an image into two parts. Figure 7 shows a possible division of a LANDSAT scene.

Figure 7. LANDSAT Frame Divided Into Two Parts



Let us imagine that we have divided Section A into 600 small parts. We then draw a random sample of 60 parts representative of the 600. This may or may not be truly representative. If it is, then, the reflective and emitted energy (the signature) from these 60 segments adequately represents the reflected energy in all of Section A. We do not consider the use of the signature in the sample of 60 segments to classify the 600, a signature extension. This is simply a valid statistical inference. It is a commonly misunderstood notion that one does not have to sample from the population of interest to make an inference, for that population.

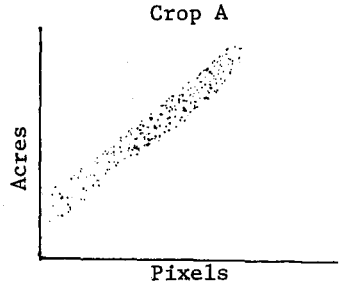
Should we wish to classify crops in Section B, it would be necessary to divide Section B into segments and draw a random sample from these segments as representative for signatures in Section B. One must sample the population of interest or the inference will be erroneous.

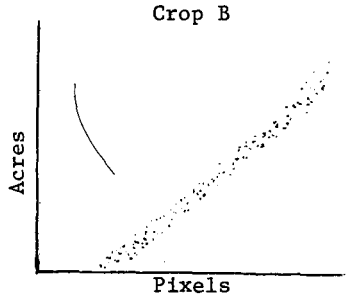
Model Utilizing LANDSAT

In order to make use of LANDSAT to reduce the sampling variation we shall first estimate the linear relationship between classified pixels for a crop and acres of the crop.

Figure 8 illustrates this relationship.

Figure 8. Population Relationship Between Classification Results and Reported Acres of the Same Crop for One LANDSAT Scene





Again, these relationships are population relationships that we do not know, so we wish to estimate them from a sample.

Our area frame sample segments can be used to estimate this relationship. The sample observations for Crop A are shown in Figure 9 and Figure 10.

Figure 9. Sample Data Points for Crop A Showing Relationship Between Pixels and Acres

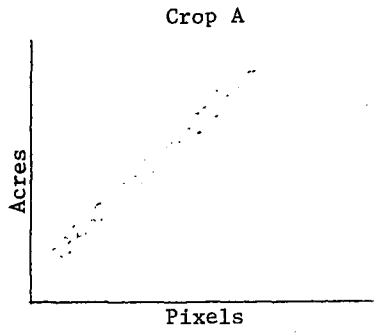


Figure 10. Estimated Population Linear Relationship Based on Sample Data in Figure 9

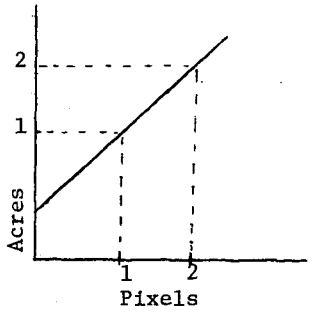


Figure 10 illustrates the relationship that is needed in order to use LANDSAT results.

This is on a per segment relationship. Therefore, we can locate a segment in LANDSAT, classify the segment and count the pixels of Crop A. If the pixels for Crop A turn out to be at point 1 then we read the corresponding acres on the y-axis. If on the other hand, the classified pixels for the segment turn out to be at point 2 then we read that value on the y-axis.

This procedure could be completed for each segment in the population and we could sum up all the segments to get an estimate using satellite information across the whole area. However, all this is unnecessary.

Since we know N, the total number of segments in the LANDSAT frame, we can classify every pixel in the frame and divide the total number of pixels in Crop A by the number of segments in the frame. This then would equal the average number of pixels in Crop A for the average segment.

Also, we know total number of pixels of Crop A in sample segments (n). With this information we can adjust the direct expansion estimate for the difference between the pixels in Crop A for the sample (n) versus the total of the population (N).

Figure 10 illustrates how the adjustments would be made. Say a difference between the average pixels for Crop A for the sample is at point 1 and the average for the universe is at point 2. The adjustment in acres is made on the y-axis. The formula is:

$$\hat{Y}_{reg} = \bar{Y} + b\ (\bar{X}_{total} - \bar{x}_{sample})$$

$\hat{Y}_{reg}$ is the adjusted number of acres in the average segment. $\hat{Y}_{reg}$ is then multiplied by N to get an estimate for the total.

The variance for $\hat{Y}_{reg}$ is $\frac{n - 1}{n - 2}\ (1 - r^2)$ times the variance of the direct expansion. This regression model reduces the spread of the sampling error distribution by a factor of $(1 - r^2)$.

In summary, we have ground data for a properly selected statistical sample, as well as the computer classification for the same. Thus, the necessary information is available to adjust a full frame classification for all systematic errors. If there is a good linear relationship between ground data and what the computer classifies as being on the ground, the sampling error will be materially reduced as compared to not having remotely sensed data.

REFERENCES

1/ Houseman, Earl E., Area Frame Sampling in Agriculture, Washington, D.C., United States Department of Agriculture: 1975

2/ LANDSAT Multi-Spectral Scanner Data User's Handbook, Goddard Space Flight Center, Greenbelt, Maryland: 1971

3/ Baker, J. R. and Mikhaul, E. M., Geometric Analysis and Restitution of Digital Multispectral Scanner Data Arrays, LARS Information Note 052875

4/ Von Steen, Donald and Wigton, William, Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery, Statistical Reporting Service, United States Department of Agriculture, Washington, D.C. 20250