A GUIDE TO AREA SAMPLING FRAME CONSTRUCTION UTILIZING SATELLITE IMAGERY

William H. Wigton [1/]

and

Peter Bormann [2/]

March 1978

A GUIDE TO AREA SAMPLING FRAME CONSTRUCTION UTILIZING SATELLITE IMAGERY

A paper submitted to the United Nations at the request of the United Nations Outer Space Affairs Division based on material used and experience gained at the Second International Training Course in Remote Sensing Applications for Agriculture: Crop Statistics and Agricultural Census. (Rome: 25 April-13 May 1977)

ACKNOWLEDGEMENTS

Preparation of this report would not have been possible without the assistance of Dr. H.G.S. Murthy, Expert on Space Applications, United Nations, New York, and Dr. John A. Howard, Head, Remote Sensing Unit, Food and Agriculture Organization, Rome, Italy. Special thanks go to Ms. Berna Spiers, Technical Officer, Remote Sensing Unit, Food and Agriculture Organization, for helping teach the course for which this was prepared.

1/ William H. Wigton, Mathematical Statistician, Consultant to the United Nations, U.S. Department of Agriculture; Economics, Statistics and Cooperatives Service, Washington, D.C. 20250

2/ Peter Bormann, Geophysicist, United Nations, Outer Space Affairs Division, New York, NY 10017

CONTENTS

SUMMARY

The area sampling frame is a basic technique for collecting agricultural statistics for a quick and comprehensive agricultural information system. It is used in a number of countries to estimate all types of agricultural products as well as economical parameters such as prices and labor for the current year. This methodology provides accurate information by taking representative samples from only a small part of the total land area. Estimates can be available five to six weeks after the beginning of the data collection. These estimates are based on an objective statistical method of data collection and evaluation.

The construction of the area sampling frame is carried out in several steps. The first step is the delineation of broad areas of homogeneous land use/land forms using all types of available data and maps of the most recent date such as satellite imagery, aerial photography, topographic and/or land use maps. Areas of the same land use type form a stratum. Once these strata have been formed, one must find boundaries for them that are identifiable on the ground, such as roads, footpaths, railways and rivers. These boundaries are then marked on the map in a unique way for each stratum and the areas within each stratum are labelled.

The next step is to divide these homogeneous strata into sample units. Normally, this is done in two steps: primary sampling units (PSU's) are delineated and a small sample of PSU's are selected to be further subdivided into sample units (SU's). Again, good boundaries must be obtained on the map at each level.

One sample unit is selected from each PSU and the selected SU is called a segment. The segments vary in size depending on stratum, land use, and

population density. The general rule is that they should be small enough to be enumerated in one day. In an agricultural area a typical size is 1 km^2.

The construction of the area sampling frame ends with the selection of segments that represent the total area. Again it must be ensured that these segments have clearly recognizable boundaries that will leave the field enumerator with no problems in deciding which area is inside and outside the segment.

The desired data are then collected from these segments, usually by interviewing the farmers, measuring crop acreages and making crop cuttings. Since the segments within each stratum are statistically representative of this stratum, the results collected from these segments can be expanded to the total area of the stratum. The desired production figures for a country are obtained by summing the results for the strata of that country.

I. INTRODUCTION

Each country needs accurate, timely information on its agricultural production for proper management of its food reserves, import and export planning and many other planning activities. There are different procedures to obtain this information. All countries have some kind of agricultural production data system but there is a general need to improve them. Some of the systems depend on compulsory reporting or complete enumeration, while others rely on statistical samples.

The first kind of system works effectively and accurately in centrally planned economies while the statistical sampling system is efficient and accurate in an economy where the individual farmer decides what he plants and may change his decisions corresponding to changing market conditions. A statistical sampling frame is effective where agricultural production estimates are needed for very large areas in a short time.

This report will deal exclusively with statistical sampling using area sampling frames. In order to be accurate, a statistical sampling frame must follow strict scientifically based procedures. A random sample will give an accurate estimate of the characteristics in a population provided the characteristics in the sample are representative of the characteristics in the population. Since the agricultural characteristics of interest are not homogeneously distributed across the whole country, the population must be divided into homogeneous areas with respect to the agricultural characteristics under study.

Imagery of the earth's surface taken from space provides a synoptic overlook of large areas and is, therefore, very useful to delineate these homogeneous areas. There are different satellite systems available. For more details on sensor characteristics see United Nations document A/AC.105/204[1].

Photographic systems provide imagery of very high spatial resolution (some 10 to 20 meters) but normally they do not provide complete and repetitive coverage of large territories. Scanning systems, as flown in the automatic LANDSAT satellites have a smaller spatial resolution (LANDSAT 1, 2 and C, some 80 meters and LANDSAT D some 30 meters), but they provide a global and repetitive data coverage every 18 days or 16 days, respectively. With more than one satellite in orbit at a time and a suitably chosen spacing between them, the time interval between repetitive coverage of the same area can be shortened significantly. Since the availability of LANDSAT data is extensive, this report will emphasize the use of LANDSAT data for the construction of the area sampling frame.

II. CONSTRUCTION OF THE AREA SAMPLING FRAME

A. Explanation of terms used.

The area sampling frame (ASF) is the total land area of a country broken down into N small parts called the sampling units (SU's). Out of these N sampling units a number n will be randomly selected for enumeration. The selected sampling units are called segments. Figure 1 shows as an example Adams County segment No. 2045. These segments must be completely enumerated by personal interview. An interviewer must go to the area on the ground and locate the owners and operators of all land inside the boundaries of the segment. The data collected by the interviewer must be recorded in a suitable way. Questionnaires used by the Economics, Statistics, and Cooperatives Service of USDA in the United States may serve as an example. Copies are given in Appendix A.

In the process of constructing the area sampling frame there are several steps which will be demonstrated in a more detailed way in the following chapters. The first step is to divide the total land into homogeneous areas with respect to the agricultural characteristics under study. These homogeneous areas are called strata. These strata will then be subdivided into intermediate parcels of land called primary sampling units (PSU's) to each of which a certain number of sampling units will be assigned. For example, the United States is broken down into approximately 3,000,000 sampling units and a sample of only 16,000 segments is selected to be interviewed. This means that only some 0.5% of the total land is sampled. Nevertheless, careful frame construction and interview procedures provide timely and sufficiently accurate estimates of the total agricultural production. The total is estimated by expanding the data collected from the n segments by the proper expansion factor N/n.

B. Survey priorities, resources and difficulties.

Before one begins the frame construction process, one should have a clear idea of which specific agricultural products are most important and must be emphasized. It is also important to know how accurate these estimates must be and how soon they must be available. Hopefully, once these questions are answered, the relative order of importance will remain fairly stable over time and the various crops and livestock items will not switch priority positions relative to each other.

Another item that needs to be determined is how much money will be available to develop the area sampling frame as well as to conduct each

necessary survey. In general it is much easier to devise a frame for a single crop than it is for a general purpose survey where a long list of priority items are being estimated, but one usually does not have this luxury, so many comprises are necessary.

The total money available to conduct the survey determines the overall sample size n and the specific list of agricultural items that are to be emphasized determines how these n segments will have to be distributed among the individual strata. The allocation of segments to strata is done according to one of several schemes outlined in most sampling books.

Normally, in agricultural surveys, most of the available segments are allocated to the intensively cultivated strata so that most of the money is spent obtaining information in those areas which have the biggest share in the total agricultural production.

Also, important for deciding on the total number of segments to be selected is the determination of the optimum segment size for each stratum. A "rule of thumb" which can be used is that the enumeration of a segment be accomplished in one day. The obvious points to consider when deciding on segment size are density and structure of the population, length of questionnaire and type of transportation used. The structure of the population is a very subtle point and is referred to in statistics books as the intra-class correlation coefficient.

In practice the careful execution of the survey plan will meet with a number of difficulties, the first being that the segments must be easy to locate in the field. This problem will be dealt with in more detail in Section II D.

Once the segment has been exactly located in the field, one has to be able to extract the wanted information. In a cattle census, for example, one would need to count the cattle in the segment. This may be possible although difficult. If, on the other hand, one is interested in an insect survey, one would have to count all the insects in the segment. This task would be impossible to implement.

Some other items that are difficult to estimate are the potential agricultural production of a country, the crop production when enumerating early in the growing season of that crop, the hectares of crops when farmers do not know their own hectarage, or the number of cattle when farmers are unwilling to tell (perhaps they must pay taxes on a per head basis). If one is confronted with problems of this type, one should do serious thinking before starting to construct the frame or the results will be disappointing.

Finally, be sure that the interviewers are qualified and willing to do a good job. Training and regular instruction of the interviewers as well as quality checks of their work are indispensible.

C. Material and facilities needed.

For the construction of the area sampling frame one needs topographic and/or land use maps or aerial photography or both. Satellite imagery can be useful supplementary material. Within reason, the more material one collects and uses in the construction of the area frame, the better the finished product. All materials used should be of the most recent date.

The scales of maps and photographic products used must be related to the size of the features on the ground. Not many countries in the world have sufficiently large field sizes to allow the recognition of single fields and their boundaries in LANDSAT imagery while, on the other hand, changes in land

use pattern stand out clearly in that type of data. LANDSAT imagery can hardly be used at scales larger than 1:250,000 and fields smaller than 8 hectares are difficult to identify. LANDSAT imagery, therefore, will preferably be used in the stratification process while aerial photography at scales between 1:20,000 to 1:60,000 is particularly suitable for stratification as well as the delineation of the sampling units and their physical boundaries.

High resolution imagery from photographic satellite systems such as those flown in SOYUZ 22 and Salyut 6 can provide basic maps at scales as large as 1:50,000. Such imagery could therefore replace or supplement standard aerial photography and also be used in areas with small fields (down to 0.4 hectares).

Besides imagery and maps, light tables, magnifying glasses, various colored pencils and suitable maps storage space are needed for carrying out the construction of an area sampling frame.

D. Sequence of constructing the area sampling frame.

1. Stratification and boundary selection.

Construction of the area sampling frame is carried out in several steps. The first step is the delineation of broad areas of homogeneous land use/land forms using all types of available data as outlined in the previous section.

Areas of the same land use form a stratum. The number of strata into which the total population/land area can be or should be subdivided depends largely on the variety and distribution of various land use types in the areas under study, the significance of their visual differences in satellite imagery or aerial photography, the skill of the photointerpreter and the goal of the survey. If we wish to have 7 strata, we could, as an example,

subdivide an area into water, forest, cities (inner city), urban agriculture (suburbs), intensively cultivated land, less intensively cultivated land, and non-farmland, such as recreational areas, deserts, high mountain areas and military bases.

Cities and towns are often difficult to delineate on satellite imagery and all types of administrative and political boundaries are normally not visible but the latter can easily be derived from maps.

Many times one crop may be so important and cover large contiguous areas that a separate stratum is set up for that particular crop. In principle, however, the land-use stratification should be more general so as to accommodate different types of surveys for a number of years. Land areas smaller than 5 km^2 should not be separated out as a different stratum even though they might not fit the stratum definition.

The delineation of the strata on LANDSAT imagery should be completed without regard to physical boundaries on the ground. This allows the photo-interpreter to concentrate on pattern recognition and differentiation. Another reason is that small physical boundaries such as country roads, foot-paths, railroads, and small rivers can normally not be seen in imagery with a ground resolution coarser than some 30 m.

When transferring their strata from imagery onto maps of larger scales (scales of 1:20,000 to 1:50,000 are best suited) their initial boundaries might have to be changed slightly to coincide with physical boundaries on the ground that are easy to recognize and follow. Unique colors and roman numbers should be assigned to all strata and the strata boundaries colored correspondingly.

The need for good physical boundaries applies to all further subdivision of the strata into primary sampling units and sampling units. The importance of this cannot be overemphasized. Most of those sampling frames that do not work well fail because of poor enumeration. The key to quality enumeration is to have boundaries which can easily be located by both the interviewer and a supervisor carrying our quality control in the field.

2. Primary sampling unit construction.

The next intermediate step in the construction of the area frame is to subdivide the strata into primary sampling units (PSU's). They vary in size depending on the stratum and the country. Since in the final step a specific number of sampling units (usually some 6 to 20 sampling units) will be assigned to them, they should be small enough to permit subdivision in a short time. However, they should be large enough to be useful for a variety of surveys. A statistician may be needed to help decide the optimum size (see also Chapter II B).

Again, good boundaries must be obtained on the map and marked in the color of the stratum. Primary sampling units in non-contiguous parts of the same stratum must be grouped together and all PSU's be numbered in a unique way, separately for each stratum. Each primary sampling unit can then be identified on the map by the stratum number (roman numeral), its PSU number (arabic numeral) and the size of its area (in km^2). For example, I-3-16 means the PSU number is 3 in stratum I has an area of 16 km^2.

In numbering the PSU's one could begin in the northeast corner and number in serpentine fashion from east to west so as to guarantee that no PSU is left out. The area can easily be measured using a grid or, more accurately, by using a planimeter.

After this is done, all the primary sampling units are listed on a PSU identification sheet (see Appendix C). A separate sheet is used for each stratum.

3. Sampling unit construction and primary sampling unit and segment selection.

In order to save time, not all the PSU's will actually be broken down into sampling units; rather, a certain number of sampling units will be assigned to all of them. Only a few PSU's will then be randomly selected for further subdivision into sampling units (SU's). The probability that a given PSU will be selected is proportional to the number of assigned sampling units in it. The most suitable number of PSU's to be selected (equal to the number of segments to be enumerated) depends on considerations as outlined in Chapter II B.

The number of sampling units assigned to a PSU depends on its size. The optimum SU size varies with the land use conditions in the survey area, the survey priorities and resources and the length of the questionnaire (see Chapter II B). It normally differs in different strata. As an example, the optimum SU size is given for two areas, Kings County, California and Salcedo Province in the Dominican Republic (Table 1 shown on page 12). All cities and towns, no matter how small, should have at least one sampling unit.

The actual number of SU's assigned to each PSU is determined by dividing the area of the PSU by the optimum SU size, then round the quotient down to the nearest whole number. The number of assigned sampling units in each PSU is then listed on the PSU identification sheet in the column marked

Table 1

Area	Stratum	Optimum size of the SU's (in km^2)	Range of tolerance (km^2)
Kings County, California	I (Intensively cultivated agriculture)	2.5	1.3 - 5
	II (Rangeland and desert)	15	10 - 31
	III (Non-agricultural)	25	12.5 - 51
	IV (Urban)	0.25	0.25 - 0.8
Salcedo Province, Dominican Republic	I (Intensive agriculture)	2	1 - 3
	II (Coffee)	2	1 - 3
	III (Extensive agriculture)	4	3 - 5
	IV (Non-agricultural land)	4	3 - 5
	V (Urban)	1/2	1/4 - 3/4

"S.U." and their cumulative number for each stratum in the column "Cum" S.U. and their cumulative number for each stratum in the column "Cum" S.U. (see Appendix C).

The whole PSU and segment selection procedure can be summarized as follows:

(a) Pick the random number (see 4) from 1 to N where N is the total number of sampling units in the particular stratum. Compare the random number selected with the cumulative numbers given for each PSU in the PSU identification sheet. The PSU selected for further subdivision into sampling units is the nearest one containing the random number.

(b) Find the selected PSU on the map and divide it into the assigned number of sampling units using the best available boundaries. The actual size of the sampling units to be constructed may vary within the tolerance range.

(c) Number the sampling units in the selected PSU beginning in the northeast corner and proceeding in serpentine fashion as before, select one at random and identify it with the segment number.

(d) Record the segment number, the stratum number, the PSU number, and the number of sampling units in the PSU on a segment location sheet (see Appendix C). The final column on this sheet may be used to record the name of the cities/towns for segments in the stratum "urban" or any other pertinent information.

Since only one sampling unit is selected within each selected PSU, the sample selection procedure may be thought of as two-step single stage rather than two stage cluster sampling.

4. The use of random number tables in the selection of primary sampling units and segments.

(a) Divide the random number sheet (Appendix B) into columns of the size needed.

(b) Count the number of one-digit columns, if any, and number them.

(c) Using another random number table, decide which column to begin with.

(d) Again using a random number table, decide whether to begin at the top or bottom of the column. Mark the start on the random number sheet.

(e) Again using a random number table, decide which column to go to next. Draw an arrow from the first column to the second. If you began at the top of the first column, draw an arrow from the bottom of the first to the bottom of the second.

(f) Randomly select the third column and draw an arrow from the second to the third as before.

(g) Proceed until all the one-digit columns are used up.

(h) Go to the two-digit columns and proceed as above. Continue until all the page is in order.

It may be helpful where you have a number of arrows crossing each other to use a different color for each set of different size columns.

To use the random number table, decide how many digits are in the highest possible number to be selected and use columns of the size. For numbers between one and ten, use a one-digit column (0 is ten). For numbers between eleven and one hundred, use a two-digit column (00 is 100); and so forth. Thus, if you need to select a random number between 1 and 11, go to the two-digit column and select the first number which falls between one and eleven.

As you go down the column, cross off each number considered even though these were not actually selected. In the above example, any two-digit numbers which did not fall between 01 and 11 should be marked off until a number is found. These numbers had a chance of selection and should not be used again. Don't start over at the beginning each time but begin where you left off the previous time. Random numbers are commodities to be used up and discarded.

E. Practical Example.

We have included an example to illustrate the various steps involved in the construction of an area sampling frame.

A more detailed exercise of this type can be found in a paper by Huddleston.[5]

The exercise presented here in Appendix C makes use of a LANDSAT scene for stratification and of maps for boundary selection. A more detailed treatise of this exercise can be found in a paper by Hanuschak and Morrissey.[3]

III. COST AND TIME REQUIREMENTS FOR CONSTRUCTION OF AN AREA SAMPLING FRAME

This section will need to be updated as the technique develops and as we become more experienced. As for materials, the costs vary depending on what is available. LANDSAT scenes can cost between $20.00 and $1,000.00 U.S. dollars depending on the amount of preprocessing and where the images are purchased. Costs of maps and aerial photography vary so much that it would be meaningless to try to generalize that cost. Those particular materials must be priced in each country.

People resources in terms of man/months will be presented as guidelines to follow. No country has ever constructed a sampling frame, selected the segments, developed a questionnaire and trained interviewers in less than a year and more commonly it has taken 2-3 years for a country to run regular surveys and producing agricultural estimates. It takes some time to obtain a trained staff to support an agricultural information system.

There are many variables. Obviously, countries with large land areas require proportionately more time to construct frames for than small countries. Also, intensively cultivated land requires more time than desert land. As

general rule, for the important agricultural areas, allow 1 to 2 man/months per LANDSAT image (32,000 km^2). This allows some time for training also.

Once a trained staff has its materials in hand and is working, the frame construction is quite fast. Consider Table 2 as a guide on page 17.

IV. USE OF LANDSAT DIGITIAL DATA FOR CROP ACREAGE ESTIMATION IN NEAR REAL TIME

A. Characteristics.

The satellite data used in this report is the 4-channel LANDSAT Multi-Spectral Scanner (MSS) data. It is described in Section 3 of Data User's Handbook.

The MSS is a passive electro-optical system that can record radiant energy from the scene begin sensed. All energy coming to earth from the sun is either reflected, scattered, or absorbed, and subsequently, emitted by objects on earth. The total radiance from an object is composed of two components, reflected radiance and emitted radiance. In general, it may be stated that the 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. At the wavelengths of energy LANDSAT monitors, reflected energy completely dominates the picture. This, then, is the "spectral signature" of an object and it is the differences between such signatures which allow the classification of objects using the statistical techniques about to be discussed.

Table 2

Estimated Time Required to Construct a Land Use Area Sampling Frame for Various Countries

Area	Total Area	Agriculture Area	Time to construct area frame
			Man-years
Africa			
Egypt	1,000,000	30,000	1
Ethiopia	1,220,000	800,000	2
Morocco	446,600	150,000	2
South Africa	1,221,000	1,100,000	4
North America			
Canada	9,970,000	700,000	4
Mexico	2,000,000	1,000,000	4
South America			
Argentina	2,780,000	1,700,000	4
Bolivia	1,098,600	300,000	2
Brazil	8,500,000	1,300,000	4
Asia			
China	9,600,000	3,270,000	4
India	3,300,000	2,000,000	4
Indonesia	1,910,000	300,000	2
Europe			
Bulgaria	110,000	50,000	1
Czechoslovakia	120,000	70,000	1
France	550,000	320,000	2
Hungary	93,000	70,000	1
Italy	300,000	170,000	1
Poland	310,000	200,000	1
Spain	505,000	340,000	2
Yugoslavia	256,000	150,000	1

B. Computer Classification Techniques.

Suppose that LANDSAT digitial data is available to classify in a computer. This can be done in the computer by use of discriminant functions. Computers must differentiate between crops on the basis of reflected energy. Before starting, a sample of data from two or more crops must be available that represents how those particular crops reflect energy. 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 let me introduce some notation.

If all data in a LANDSAT frame were plotted in a scatter diagram it might appear as Figure 2.

Figure 2. Scatter Diagram of All Values in One LANDSAT Frame for Three Crops. C-Corn, S-Soybeans, W-Water.

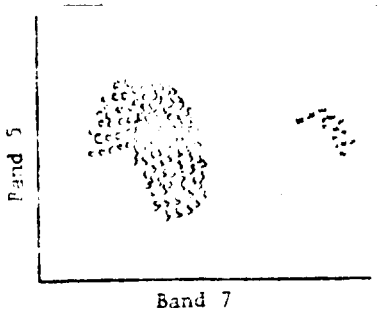
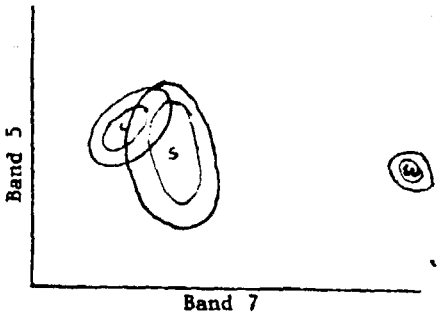


Figure 3 shows confidence limits for above data.

Figure 3. Confidence Limits for Data in Figure 2.



If one studies Figure 2, the following observations can be made:

1. The location of the center of these concentric circles has an 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 have 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 resion of corn with soybeans, because both are possible.
5. It would be ideal if the data for each crop as far apart as water from corn and if the spread were as small as water and elliptical in form and there were not areas of overlay.

However, it appears that these items are not under our control. The position of sensor bands and their band width determines the locations 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, and data preprocessing.

As far as the overlapping areas are concerned where mislabeling or misclassification is inevitable, the complexity of nature herself is the reason. It is impossible to identify unambiguously all types of targets on the ground on the basis of their reflectance characteristics in the visible and near infrared part of the spectrum along without considering additional criteria such as shape, size, texture, pattern and association. This the more so, since spectral reflectance of natural targets normally varies with time and environment. Often the differences between various targets are significant only in very narrow wavelength bands, therefore, when comparing the reflectance values of different targets in the broad spectral bands of the LANDSAT multi-spectral scanner, these differences may no longer be recognizable.

One has also to take into account the relation between pixel and feature size. Most of the natural targets on the ground are smaller than a LANDSAT picture element. The spectral radiation value of a given pixel, therefore, represents normally a mixed signature. Only when a sufficiently large area of the surface is densely and homogeneously covered by the feature under study can one expect to have a predominant signature representative of this feature.

Finally, one should mention that spectral signature clusters overlapping in a two-dimensional color space might be completely separated in color spaces of more than 2 dimensions. In general, it can be said that the more spectral bands that are compared and the narrower they are, the better the separability of the targets. Digital LANDSAT MSS data principally allows to analyze the signature patterns in a four-dimensional color space.

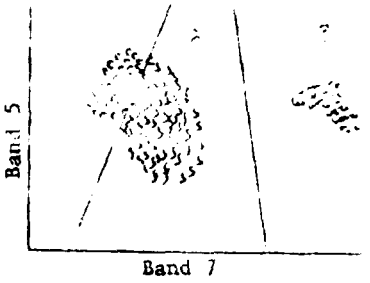
The best that can be hoped for in practice is to estimate from a sample the scatter diagram of the population and this we know how to do if it is dealt with using scientific sampling procedures.

A valid statistical estimate is needed and this requires a random sample from the population of interest. 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 2, or if estimates are needed that are quite accurate, as for 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 crops can be determined for the fields inside the segments. These signatures are estimates for the scene they are in, so, it is valid to use these values for computer training of the discriminant functions. After population scatter diagrams 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 4 shows an example of this.

Figure 4. Parititioned Space Showing Population Scatter Diagram.

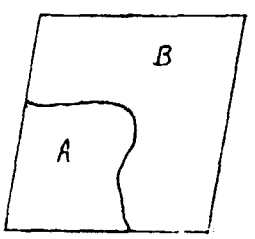


All pixels that need crop labels should then be plotted on the partitioned space. If they fall in partition one, give it a label of corn, even though some soybeans will creep in obviously, water will be no problem.

Incidently, it turns out that the location, size and shape of these population scatter diagrams 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. For illustration we divide LANDSAT image into two parts as shown in Figure 5.

Figure 5. LANDSAT Frame Divided Into Two Parts.



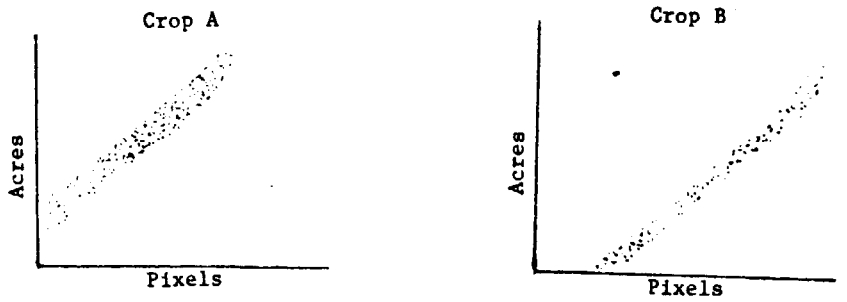
We further assume that Section A has been divided into 600 small parts and draw a random sample of 60 parts representative of the 600. This may or may not be truly representative. If it is, then, the reflective energy (the signature) from these 60 segments adequately represents the reflected energy for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

Should one wish to classify crops in Section B, it would be necessary to divide the 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.

C. Acreage Estimates Using Classifed LANDSAT Data.

In order to make use of LANDSAT to reduce the sampling variation one must first estimate the linear relationship between classified pixels for a crop and acres of that crop. Figure 6 illustrates this relationship.

Figure 6. Population Relationship Between Classification Results and Reported Acres of the Same Crop for One LANDSAT Scene.



Again, these relationships are population relationships that are unknown, so they must be estimates from a sample.

Our area frame sample segments can be used to estimate this relationship. For example, sample observations for Crop A are shown in Figure 7 and Figure 8.

Figure 7. Sample Data Points for Crop A Showing Relationship Between Pixels and Acres.

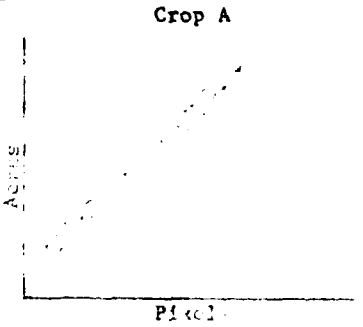


Figure 8. Estimated Population Linear Relationship Based on Sample Data in Figure 7.

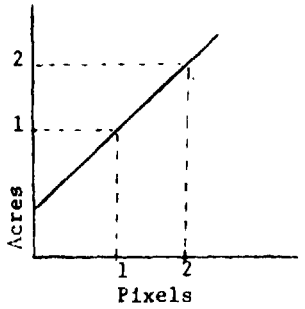


Figure 8 illustrates the relationship that is needed in order to use classified LANDSAT CCT results. (CTT stands for "Computer Compatible Tape").

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 number of pixels for Crop A turns out to be at point 1 then we read

the corresponding acres on the y-axis. If on the other hand, the number of 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 sgments to get an estimate using satellite information across the whole areas. 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, the total number of pixels of Crop A in sample segments (n) is known. 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 8 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, one needs ground data for a properly selected statistical sample, as well as the computer classification for the same areas. Thus, the necessary information is available to adjust a full frame classification for all linear relationships between ground data and what the computer classifies as being on the ground, the sampling error will be substantially reduced as compared to not having remotely sensed data.

V. CONCLUSIONS

The area sampling frame is a basic means used in a number of countries for collecting statistical data in agriculture. It allows one to derive estimates of economical parameters and of all types of agricultural products from samples that cover only a small part of the land under survey. A statistical sampling frame is therefore fast and cost-effective, particularly in very large areas or in cases where a complete enumeration is not practicable or economically feasible due to other reasons such as socio-economic conditions or lack of infrastructure. The area sampling frame technique may therefore help to establish a quick and comprehensive agricultural information system in developed and developing countries alike.

The accuracy of the estimates depends on whether the characteristics in the sample are representative of the characteristics in the whole population. In order to ensure this, the construction of the area sampling frame must follow scientifically based procedures. The subdivision of the total land into homogeneous areas (strata) with respect to their agricultural characteristics, as well as the further subdivision of these strata into sampling units with clearly recognizable physical boundaries are two decisive steps in the construction of the area sampling frame. Topographic and/or land use maps or aerial photography or both are needed for this purpose.

Satellite imagery is a useful supplementary material in the construction of the area sampling frame particularly for the stratification of large areas. High resolution imagery from photographic satellite systems is also suitable for further subdivision of the strata into sampling units and the identification and delineation of their physical boundaries. This might be of particular importance for developing countries which are still lacking aerial photography and accurate small scale maps. If satellite data are available also in digital form on computer compatible tapes (CCT's) they can additionally be used for crop classification and the improvement of acreage estimates.

Philosophy and procedure of the construction of the area sampling frame are outlined in this paper and demonstrated by example.

References

1. United Nations Document A/AC.105/204. Characteristics and Capabilities of Sensors for Earth Resources Survey (1977).

2. Caudill, Charles E., Current Methods and Policies of the Statistical Reporting Service, LARS Symposium on Machine Processing of Remote Sensing, 1975.

3. Hanuschak, George and Morrissey, Kathleen, Pilot Study of the Potential Contributions of LANDSAT Data in the Construction of Area Sampling Frames, Statistical Reporting Service, USDA, Washington, D.C. 20250.

4. Houseman, Earl E., Area Frame Sampling in Agriculture, Washington, D.C., USDA: 1975.

5. Huddleston, Harold F., A Training Course in Sampling Concepts for Agricultural Surveys, Statistical Reporting Service, April 1976.

6. Wigton, William H., Use of LANDSAT Technology by Statistical Reporting Service, LARS Symposium on Machine Processing of Remote Sensing, 1975.

7. Data User's Handbook, Earth Resources Technology Satellite, NASA, Document No. 71SD4249, Goddard Space Flight Center, Greenbelt, Maryland, 1972.