STRATIFICATION OF A REMOTELY SENSED AREA SAMPLING FRAME

Ron Fecso, U.S. Department of Agriculture

Introduction

The Statistics Unit of the Economics and Statistics Service, USDA, has the responsibility of constructing the sampling frames for the AgRISTARS (Agricultural and Resources Inventory Surveys through Aerospace Remote Sensing) Foreign Commodity Production Forecasting Project. The frames are to be used in a research program to determine the suitability of using remotely sensed data collection to provide acreage and production estimates of selected agricultural commodities in various parts of the world.

Frame construction techniques are developed which considerably reduce the labor and time requirements necessary to create the requested frames and which allow for considerable flexibility in research use and sampling design.

The Population

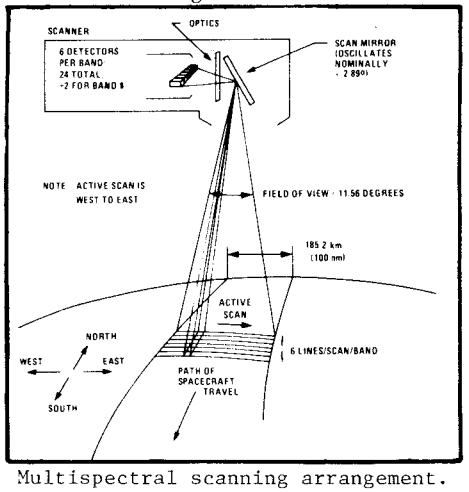
The reporting unit is defined as the remotely sensed spectral information corresponding to a satellite "pixel" reading (see figure 1). The pixel which corresponds to approximately a 57 meter square portion of the earth's surface, emits energy which is refracted into four spectral components. The response in each wavelength band is stored on magnetic tape in digital form. Thus, the sampling frame for the target area for estimation consists of sampling elements corresponding to pixels and the data is collected for the pixel by the satellite. Discriminant techniques are used to classify the land use of the element based on the multivariate spectral data recorded for the pixel. Proportion estimators for the crops of interest have been developed based on the pixel classification.

The Sample Unit

The current sampling unit is a cluster of elements which correspond to a rectangular array of 193 by 117 pixels. Research has indicated that the optimal cluster size is smaller than the one being used (Perry, 1979), but for the next few years the cluster size is fixed due to software limitations and constraints related to the discriminant function. The possibility that the sample unit size will be reduced brought out a flaw in the original frame development process.

In that process, a change in sample unit size would require the creation of a new sampling frame. Thus, the procedures for sampling frame construction were developed in a way which minimizes the work which would be required if the sample unit size changes.

Figure 1



Multispectral scanning arrangement.
Source: Landsat Users Guide

Properties of an Effective Area Sampling Frame

The development of area frames for AgRISTARS incorporates the following features which should be considered with any area frame:

1. Longevity - The coverage of a population rarely becomes out of date in an area frame, but stratum homogeneity or the correlation of auxiliary data with the items being estimated can decrease over time. An area frame which will be used over a long period of time, needs to be developed in a way which allows easy and economical updating of the auxiliary information needed for an efficient sampling design.
2. Design versatility - The frame should be developed such that there is flexibility in the sampling designs which can be used. When possible the frame development process should not preclude the possibility of future changes in strata definitions, the type of estimator or in sample unit size.

The Area Sampling Frame

Originally, defining the sample units which cover the target population was a manual process. A transparent grid overlay was made to the desired cluster size and placed over the associated base map (usually an Operational Navigational Chart) for the target area. The sample units were then defined as the area enclosed by the individual grids. Each grid was then labeled for identification and assigned some auxiliary information for use in the creation of a stratified sampling design. The auxiliary information, percent of cultivated land in the unit, was based on visual interpretation of LANDSAT 1:1,000,000 transparencies which were aligned on the base map according to physical features common to both products.

Problems associated with the original frame development process included: (1) manual burden, (2) difficult image interpretation within a sample unit and (3) inflexibility with respect to sample unit size changes.

Keeping in mind the problems associated with the original frame development methods and the properties desired in an area frame, an automated system for sample unit creation and assignment of its auxiliary information was developed. Treating the earth as an ellipsoid of revolution (Colvocoresses, 1974), the following parameters are required:

Equatorial radius $a = 6378206$ meters
Polar radius $b = 6356584$ meters
First eccentricity $e = (a^2 - b^2)\cdot^5/a$
Second eccentricity $e_b = (a^2 - b^2)\cdot^5/b$

Then, the distance on the surface of the ellipsoid from the equator to a specified latitude, t, is (in series expansion)

$$S(t) = a_o t + a_1 \sin(2t) + t_2 \sin(4t) + \ldots$$

where

$$a_o = a(1-e^2)(1 + (3/4)e^2 + (45/64)e^4 + (175/256)e^6 + (11025/16384)e^8 + \ldots)$$

$$a_1 = -(1/2)a(1-e^2)((3/4)e^2 + (15/16)e^4 + (525/512)e^6 + (2205/2048)e^8 + \ldots)$$

$$a_2 = (1/4)a(1-e^2)((15/64)e^4 + (105/256)e^6 + (2205/4096)e^8 + \ldots)$$

For the creation of AgRISTARS sampling units, the inverse of S(t) is necessary. The following inverse computes the latitude (in radians) of a specified distance, H, from point t_i along a line of longitude.

$$T(H,t_i) = t_i + (S-S_o)\, q_1 + (S-S_o)^2\, q_2 + \ldots$$

with $H = (S-S_o) = (S-S(t_i))$,

$$q_1 = 1/(a_o + 2a_1\cos(2t_i) + 4a_2\cos(4t_i)),$$

and

$$q_2 = (1/2)q_1^3\,(4a_1\sin(2t_i) + 16\, a_2\, \sin(4t_i)).$$

The distance along a line of latitude, t=T, between two lines of longitude is determined by

$$D((T,g_2), (T,g_1)) = (r \cos T)(g_2-g_1)$$

where $g_2 > g_1$ are the lines of longitude and

$$r = (((\cos^2\theta)/a) + ((\sin^2\theta)/b)^2))^{-2}.$$

Thus, the angle of longitude (in radians) subtended by an arc of length, L, along the line of latitude, t_i, is

$$G(L,t_i) = \frac{L}{r \cos t_i}.$$

Each of the distance functions has an inverse mapping in which, given a point, P_c, and a desired distance along a given latitude or longitude, the latitude and longitude of the point meeting the specifications can be computed. To grid a spheriod using these distance measures and inverses, it is necessary to define starting points. Using the equator and the 0^o line of longitude will be convenient for computer applications.

Creating the Sampling Frame

The "grid" of sample units can be established by first defining planes of latitude which slice the surface of the sphere into strips which have latitude defined by t_i, such that,

$$S(t_{i+1}) - S(t_i) = H \quad i = 0, 1, 2, \ldots I$$

where

H = height of desired sample unit, $t_o = 0$ (the equator), $t_{i+1} > t_i$, and

$t_I < (2/9)$ radians (approx. 80^o).

Note that the slicing algorithm for general application is done from the equator toward the "North" Pole and the appropriate reflections will be used to define the slices in the Southern Hemisphere.

With the strips which will contain the sample units defined, all that is required is to slice the strip into sample units of the appropriate length (L). The algorithm starts the slicing by placing the westernmost edge of the first sample unit approximately along the 0^o line of longitude. The center line of latitude is computed for the strip and increments of length equal to the sample unit length are made along this center line. Thus, the center points of this latitude/longitude slicing process define the sample units.

This outlines the approach used. The exact algorithm and notes on the limits of the grid system are presented in Fecso (1981).

To understand the gridding algorithm, note that there are (I+1) unique values for latitude, t_i, which slice the quadrant. For each of the $j = 1, 2, \ldots, I$ "center latitudes" of the slices there are N_j values of longitude in the quadrant which "chop" the slice into sampling units. Notice that N_j is nonincreasing as the center latitudes, Y_j, increase.

The Y_j are found incrementally by first defining $t_o = 0$ and then incrementing along $g = 0$ to find $t_1, t_2, \ldots t_I$. The latitude of a center point of a sample unit, Y_j, is found by

$$Y_j = (t_j + t_{j-1})/2, \; j=1, 2, \ldots, I$$

where

$$t_j = T(H, t_{j-1}).$$

The N_j sample units along latitude Y_j are computed iteratively. First let $g_o = 0$. Then the line of latitude is chopped into lengths the size of the sample units by creating

$$g_k = g_{k-1} + G(L, Y_j), \; k=1, 2, \ldots, N_j.$$

Thus, the center point longitude of the sample unit, X_k, is computed as follows:

$$X_k = (g_k + g_{k-1})/2, \; k=1, 1, 2, \ldots N_j.$$

As a result, the iteratively created collection of ordered pairs, (X_k, Y_j), defines the center point lat-long for each sample unit in the quadrant. It should be noted that the actual program creates a file of sample units which consists of the intersection of these center points and the lat-longs of the area for which the frame is being created. That is, only the sample units in the target area are created.

Assigning Auxiliary Information

The sampling frame does not yet have auxiliary information to use in developing the sampling design. To allow freedom in choosing the sample unit size, the auxiliary information must be developed independently of the sample unit location. Inspection of LANDSAT imagery and other materials is used to develop a land use data base which contains multivariate

auxiliary information for use in survey design.

For the target region, the useful auxiliary data must be determined. "Useful auxiliary data" for our statistical purposes is information which can be used to reduce the error of the estimator through sampling designs such as stratification or PPS selection or through ratio and regression estimation. For AgRISTARS, information is needed which is useful for stratification with respect to the agricultural commodities to be estimated in the target region.

The estimator of production for an agricultural commodity is a product of an acreage and yield per acre estimate. Thus, the information necessary for an effective stratification of the area frame must include the density of the crop and the yield potential for the crop.

The land use characteristics which were identified as being useful for stratification and available through inspection of LANDSAT images and other materials are: (1) percentage of land cultivated, (2) percentage of the cultivated land which is utilized for the crop being estimated, (3) field size, and (4) soil type. After drafting a latitude-longitude overlay for the LANDSAT images, the land areas are partitioned into blocks which are homogeneous with respect to the four auxiliary variables. These blocks of land are then digitized to create a data base which contains the boundary points of the block and the associated auxiliary data.

At this point the sampling frame and the land use data files are merged in an automated process which determines the land use information contained in each sample unit by comparing the sample unit location to the boundaries of the land use blocks. With sampling frame and associated auxiliary information defined, the researcher is left with the task of deciding on the best design. Currently a stratified design which clusters the auxiliary information is being considered in AgRISTARS sampling.

Benefits of the Automated Frame Development

1. The land use stratification needs to be done once for any size sample unit. Thus, pilot surveys can be used to determine the best sample unit size, and changes can be made for the survey by altering the software rather than reworking the land use stratification.
2. Additional auxiliary information can easily be input into the land use blocks allowing better stratification or estimators. For example, if it is desired to estimate an additional commodity from a frame, an auxiliary variable could be estimated for each block and used in improved estimators.
3. Frame updates due to land use changes over time will be simplified. The auxiliary information for the block needs only be changed, then improved estimators can be used or a new stratification developed.

References

[1] Colvocoresses, Alden P., (1974) "Space Oblique Mercator," Photogrammetric Engineering, pp. 921-6.

[2] Donovan, Walt, (1974), "Oblique Trandformations of ERS Images to Approximate North-South Orientation," Center for Advanced Computations, Technical Memorandum No. 38, Univ. of Illinois, Urbana, Illinois.

[3] Fecso, Ron, (1981) "Frame Development and Sampling Design for Remotely Sensed Observations," USDA, ESS.

[4] Perry, Charles R., (1979) "Sampling Unit Size Considerations in Large Area Crop Inventorying Using Satellite-Based Data," Proceedings of the Survey Research Section of the Annual ASA meeting, Washington, D.C.