AN INTERACTIVE SYSTEM FOR AGRICULTURAL ACREAGE ESTIMATES USING LANDSAT DATA

By

Martin Ozga, Walter E. Donovan
Center for Advanced Computation, University of Illinois
Urbana, Illinois

Chapman Gleason
Statistical Reporting Service, USDA
Washington, D.C.

TO BE PRESENTED AT THE FOURTH PURDUE SUMPOSIUM ON
MACHINE PROCESSING OF REMOTELY SENSED DATA, JUNE 1977

New Techniques Section
Research and Development Branch
Statistical Reporting Service
United States Department of Agriculture
Washington, D.C.

March 1977

AN INTERACTIVE SYSTEM FOR AGRICULTURAL ACREAGE ESTIMATES USING LANDSAT DATA

By

Martin Ozga, Walter E. Donovan
Center for Advanced Computation, University of Illinois
Urbana, Illinois

Chapman Gleason
Statistical Reporting Service, USDA
Washington, D.C.

I. ABSTRACT

This paper describes interactive software systems for making agricultural crop acreage estimates using LANDSAT MSS data developed jointly by the Center for Advanced Computation of the University of Illinois and the Statistical Reporting Service of the United States Department of Agriculture. These acreage estimation procedures have been incorporated into, and use features previously developed in, EDITOR. EDITOR is an interactive file management and image processing system developed by the Center for Advanced Computation in collaboration with USGS/DI, NASA/AMES, and USDA/SRS. The crop acreage estimation software is implemented as part of the EDITOR system on TENEX, a modified DEC SYSTEM-10. The only hardware necessary to access this acreage estimation subsystem or the whole EDITOR system consists of a KSR (keyboard send-receive) terminal with acoustic coupler and a telephone link to a TENEX system on the ARPA network. An x-y coordinate digitizer and, optionally a terminal graphics plotter, are also needed for digitizing ground-truth samples and interactive registration capabilities.

II. INTRODUCTION

This paper describes additions to the EDITOR software system made to enable agricultural crop acreage estimation using LANDSAT MSS data. These additions have been developed jointly by the Center for Advanced Computation (CAC) of the University of Illinois and the Statistical Reporting Service of the United States Department of Agriculture (USDA/SRS).

Briefly, EDITOR is an interactive data analysis system for processing of LANDSAT data.[1] EDITOR is under continuing development by CAC with the aid and support of USDA/SRS, NASA/AMES, and USGS/DI. EDITOR runs on TENEX (modified DEC SYSTEM-10) systems and is currently available on three TENEX systems: two systems at Bolt, Beranek and Newman (BBN) in Cambridge, Massachusetts, and the ILLIAC IV (I4) TENEX system at the NASA Ames Research Center in Moffett Field, California. To process large data sets, EDITOR provides a link, using the ARPA Network, to the ILLIAC IV for batch processing. Also, to make the LANDSAT data more easily usable by various EDITOR subsystems, certain preprocessing functions, such as image geometric correction, are performed on various IBM 360's off line from EDITOR in batch mode.[2,3]

When performing classifier training and acreage estimation, the basic unit of data is the segment. A segment is an area of land which has been randomly selected by USDA as a sample unit in some land-use stratum. For example, there are 300 of these segments in Illinois, scattered spatially throughout the entire state. For each segment, USDA enumerators make a number of visits during the growing season and record the crop or ground cover and the size of the various fields. USDA enumerators use aerial photos of the segments on location and mark the boundaries of all fields. The locations of all segments are also recorded by USDA on USGS topographic quadrangle maps.

Using this information, the field boundaries of the segments are digitized and then mapped into a geographical coordinate system (latitude, longitude). With this mapping and the geographical registration of the LANDSAT data to latitude and longitude, a mask is generated which maps the LANDSAT pixels into the fields of the segments. These digitization and masking procedures are the first of the additions to EDITOR that we shall discuss.

Once the masking is completed, pixels may be selected automatically from groups of segments to isolate those meeting certain criteria (especially ground cover) for training. In addition, fields may be sampled using statistical techniques. Thus some fields may be used for training and the remainder for testing the results of that training. Finally, tabulation of the categorized data may be done either to evaluate the results or as an input to statistical estimation. These selection, sampling, and tabulation procedures are the second of the additions to EDITOR that we will discuss.

The actual procedures used for clustering and for statistical classification are previously developed functions of EDITOR following LARSYS algorithms and so they will not be discussed in this paper. Both clustering and classification are avaiable for small data sets on TENEX and for large data sets on the ILLIAC IV.[4,5,6,7,8,9]

Finally, using the tabulation feature, the ground truth information collected by the USDA enumerators, and the masks associated with the digitized segment data, sample estimation of crop acreages can be performed in the various land use strata using statistical methods. The segments are the sample and are selected over the region desired. Selection criteria are used to choose which ground covers will be estimated. The results of sample estimation are used both to aid in evaluating classification performance and as parameters for large scale estimation. However, before large scale estimation may be performed, an additional digitizing and masking operation is needed. In the same way as for segments, but on a larger scale and using county highway maps instead of photos, the land use strata boundaries for counties are digitized. The masks thus generated are applied to the classified data to produce aggregations by category and land use strata on a county by county basis. Using these aggregations in connection with parameters derived from the sample estimation a final or large scale estimation is performed of acres in various crops by land use strata and county. These estimation procedures are the third of the additions to EDITOR we shall discuss. One set of statistical results obtained by using this software system on a large data set is described in greater detail by Gleason,Hanuschak,Starbuck,Sigman.[10]

Before proceeding further, it is useful to mention briefly and in one place the various files used in this estimation process. These files are described in more detail elsewhere.[11] All files needed for a particular run are kept on disk. For each segment there are five files: the segment network file, the mask file, the local calibration file, the ground truth file, and the window file of LANDSAT MSS data. The segment network file contains the field boundary coordinates and is the direct result of digitization. The local calibration file contains the parameters to horizontally and/or vertically shift the segment from the location predicted by the global geographic registration of the image to its actual location (determined by hand registration of each segment to line printer displays of LANDSAT data). The mask file results from this process of segment to LANDSAT data registration and is a mask assigning the pixels to the various fields in the segment. The ground truth file is a computer readable version of the ground truth information collected by the USDA enumerators for a segment. The window file consists of the LANDSAT MSS data for a rectangle containing a particular segment.

The files are assigned default names based on the segment number for convenience in processing. In addition, several window files may be combined into one large file, called a multi-window file, for convenience of handling.

Also, there are two state wide files: the segment catalog file and the frame unit file. The segment catalog file has, for each segment, a list of attributes and is thus useful for obtaining groups of segments with common attributes, such as all segments in a certain county. The frame-unit file contains for each county and land-use stratum the number of frame units, i.e. the total number of possible segments in the sampling frame from which the sample segments were selected. These two state wide files are given default names based on the state name. These files and others will be discussed further in the sections of the paper dealing with their creation and use.

III. SEGMENT DIGITIZATION

A segment is an area of land. It will usually be divided into several (perhaps noncontinguous) ownership tracts, and further subdivided into fields. Each field represents an area on the ground to be considered homogeneous with respect to ground cover. The set of these areas do not overlap. There is no restriction (other than computer file storage space limitations) on the complexity of the field boundaries.

Segment digitization is the process of converting segments from fields drawn on aerial photographs or topographic maps to a file of coordinates in a geographic coordinate system. Location of points on the photos or maps are measured by hand using a data tablet digitizer, in conjunction with interactive EDITOR software subsystems, and assembled into a convenient computer-readable data structure. This data structure contains all the topological, geographic, and naming information needed to completely reconstruct the segment.

We will describe the process of digitizing agricultural segments.[12] Examples are taken from data used for the Illinois 1975 project.[13] However, it should be understood that the capabilities of this digitizing software are quite general; it has been used to digitize land-cover boundaries within both county-wide land-use strata maps and LANDSAT frame-wide classified images.

A. SOURCE DATA

Aerial photographs, mostly infrared, of each of the 300 Illinois 1975 segments were obtained in support of the Illinois crop-acreage estimation project.[13] These photographs are sufficiently orthographic that a simple linear transformation will serve to map ground points in the photo to geographic coordinates.

Some large segments require several photographs to cover completely the ground area of the segment. To avoid the problem of matching up partial segments digitized from different photographs, we simply consider segments (and partial segments) on a one-photo-at-a-time basis.

Tract and field boundaries are drawn directly on each photograph or on a plastic overlay. These boundaries and changes to them were determined on location by USDA enumerators on four dates in 1975. Boundaries of fields that changed between visits were appropriately added using a color-coded marking scheme and the ground-truth information on crop cover and field acreage updated. The field boundaries thus recorded have proved to be accurate to at least 50 feet on the ground.

To establish accurately the geographic location of all segments, 7 1/2 or 15 minute USGS topographic maps of the area were obtained, and the location of all segments marked thereupon. For Illinois, many segments are one-square-mile township sections, usually surrounded by roads, and thus were easily located on the quadrangle maps.

B. SEGMENT DATA STRUCTURE

A simple representation of the tract and field boundaries associated with a segment has proved convenient to use. Curved edges of a field are approximated by straight-line pieces, and fields themselves are a collection of one or more polygons. Fields with 'holes' (e.g., ponds within pastures) are stored as two or more polygons, one for each separate boundary. The outer boundary in such a case is denoted by a polygon whose vertices travel clockwise around the field. Inner or hole boundaries have polygons with vertices travelling counterclockwise.

For adjacent polygons, the edges in common are stored only once. Thus, the data structure for each segment is really a network of edges delineating the polygon boundaries of the fields.

C. THE SEGMENT NETWORK EDITOR SUBSYSTEM

The EDITOR command DIGITIZE A SEGMENT initiates the interactive segment digitization program. Significant commands of this program are:

ALIGN SEGMENT
READ/WRITE SEGMENT NETWORK FILES
DIGITIZE FIELDS
VIEW FIELD INFORMATION
CALIBRATE SEGMENT
CHECK FIELDS
LIST HOLES
DELETE FIELDS

Before digitizing a map or photo, it is required that the user digitize precisely the location of two fiducial points on the photograph which are to be used as reference points for the photograph coordinate system. Thus, it is possible to re-digitize errors or continue the digitization of a segment at a later date even though the original position of the map or photo on the digitizer tablet has been lost. The ALIGN command indicates that the user is ready to digitize these fiducial points.

The READ and WRITE commands are used to transfer the digitized segment information between the program and the segment network files on disk.

DIGITIZE activates the menu on the digitizer tablet. Menu commands are discussed in the next section.

Field information such as the adjacent field for each edge and the total field acreage (after calibration) is listed by the VIEW command. Digitization errors such as overlapping edges or missing vertices result in 'false edges' being generated, i.e., edges not representing boundaries between adjacent fields. These edges appear in the VIEW list with 'NONE' for the adjacent field.

CALIBRATE initiates a registration procedure by which a linear transformation between the segment coordinate system and a UTM coordinate system is computed. Once a segment is calibrated, the segment's scale and its field acreages are available. This calibration process is described in more detail below.

The CHECK command will detect certain common digitizing errors such as overlapping or impossible combinations of polygons. A list of suspicious fields, if any, is printed and these fields may be checked further with the VIEW command. Another common digitizing error leaves false holes in the middle of a group of adjacent fields. The LIST HOLES command will find and list all such polygons. There will always be at least one such 'hole' -- that one legitimately associated with the exterior boundary of the segment. Erroneous fields may be removed by the DELETE FIELDS command and then redigitized.

D. MENU COMMANDS FOR DIGITIZING FIELDS

When digitizing the field boundaries of segments, the user has a menu of available commands attached directly to the active surface of the digitizer tablet for controlling the sequence of digitizing operations. Commands are specified by marking a point in the appropriate area of the menu. Typically it is unnecessary to enter any information via the terminal keyboard while digitizing. Below is a picture of the menu:

SET TRACT & FIELD	CLOSE FIELD
NEXT FIELD	NEXT TRACT
NEXT FIELD PART	
AUTO DIGITIZE	
DELETE FIELD	QUIT

For purposes of identification, tract and field names have been assigned to each field of the segments. Tracts are a collection of fields all with the same ownership or under the same management. Tract names are one or two letters 'A', 'B', ..., 'AA', etc., and fields are numbered consecutively from 1 within tracts. A specific field is thus referenced by giving both the tract and field names, e.g., A1. It should be understood that this naming scheme is simply one convenient to USDA and does not indicate any basic restriction to the digitization capabilities of the Segment Network Editor.

Typically the user starts digitizing field A1, completes its boundary, then marks NEXT FIELD to proceed with field A2, or NEXT TRACT to begin digitizing field B1. For fields comprising more than one polygon, NEXT FIELD PART is marked between each polygon. To digitize a specific field, SET TRACT & FIELD is used.

Each polygon is digitized once, by travelling around the polygon in the appropriate direction and digitizing all of its vertices. This direction is determined by requiring that the inside of the field be to the right as you travel from vertex to vertex. The last vertex digitized is required to be the same as the first vertex digitized. This guarantees that the polygon has been closed. As a help in digitizing polygons with a large number of vertices, the CLOSE FIELD command can be used to forcibly close the polygon without actually requiring the user to mark the first vertex again.

Note that all digitization proceeds in what might be called 'point mode', as opposed to 'stream mode' in which the digitizer hardware records point coordinates at a set rate, say, every 1/50 second. Since EDITOR and the digitizer routines are meant to be used with dial-up telephone line terminals, input and output are limited to 300 baud or effectively, about 10 coordinate pairs a second. This data transmission rate is of course inappropriate for stream digitizing. Thus, we use point mode, at the expense of requiring human judgment in approximating the curved and jagged field boundaries with straight line segments.

Since edges are common to adjacent polygons, most edges are digitized twice. This serves to minimize digitization errors since the terminal's bell rings whenever the program detects an edge not digitized before. Thus, the user knows immediately that he has just digitized some vertices incorrectly if the bell rings upon digitizing an edge of a polygon adjacent to another already digitized.

To speed up the digitizing task, the menu command AUTODIGITIZE can be used. A long series of edges (representing a river boundary for example) need be digitized point by point only once. To digitize the boundary in the opposite direction, the user need merely indicate the first, next, and last vertices on the string of edges. The 'next' vertex indicates the unique path existing between the 'first' and 'last' vertices being autodigitized.

DELETE FIELD will delete the current field and allow its re-digitization. If digitization of the current field has not been started, this command will request the name of the field to be deleted.

The QUIT command exits menu command mode and returns the user to the regular command level within the Segment Network Editor.

E. PLOTTING SEGMENT NETWORK FILES

The output files of the Segment Network Editor may be plotted in various coordinate systems with the PLOT SEGMENT subsystem of EDITOR. Specifically, plots may be generated in DIGITIZER, MAP, or LANDSAT coordinates.

DIGITIZER coordinates are in units of digitizer surface inches and represent the coordinate system determined by the fiducial marks placed on the map or photo digitized. Plots in these coordinates can be superimposed directly on the map or photo allowing a direct check for digitizing errors. Additionally, tract and field names are placed on the plot for each fields. These names may be checked for accuracy against the field names on the original source material.

MAP coordinates are in units of meters scaled appropriately. Currently, the UTM coordinate system is used. Latitude/longitude tick marks are placed on the plot for checking the geographic registration of segments to maps.

LANDSAT coordinates are in units of pixels. An implicit transformation to inches is made, assuming 10 pixels horizontally and 8 vertically. This transformation is the same made when generating grey-scale or classified line-printer output of LANDSAT windows. Thus, LANDSAT coordinate plots will overlay directly LANDSAT line-printer output printed at 8 lines per inch. This mode allows the hand checking of the registration accuracy of segments with LANDSAT data. The next section will discuss this

registration and re-registration process in greater detail. Segments must be calibrated and a LANDSAT calibration file supplied before plotting in this mode is possible.

F. SEGMENT CALIBRATION AND REGISTRATION

The assumption is made currently that a 6 parameter linear transformation will map the segment coordinate system to the UTM coordinate system. This is a reasonable assumption given the scale of the photography and maps used for recording the ground truth information. Calibration errors for segments of large geographic area may exist. However, at the scale that these segments must be presented, the random irreducible error of digitization will be larger than the error in the linear transformation.

The parameters of the transformation are computed by a least-squares fitting procedure. Corresponding points are denoted on both map and segment and coordinate pairs obtained are used to compute the transformation coefficients. Maximum and root-mean-square errors are reported to the user on the terminal. Then the user may accept the calibration or proceed to obtain a different set of control points, based on the magnitude of the errors. RMS errors will be relatively large if the corresponding points do not truly correspond geographically or were digitized inaccurately.

Each LANDSAT image is also separately geographically registered. Together, the segment and LANDSAT registration transformations yield a mathematical mapping from digitized segment coordinates to LANDSAT pixels. The accuracy of this composite transform may be checked visually by overlaying LANDSAT coordinate plots of the segment on grey-scale printouts of the appropriate LANDSAT window. Since individual fields are usually visible within the printout, it becomes possible to specify the small shift corrections that may be necessary for the LANDSAT plot to overlay the fields exactly. The EDITOR subcommand CALIBRATE SEGMENTS LOCALLY is available for specifying these shifts and generating local calibration files storing the corrected segment to LANDSAT transformation.

G. MASK GENERATION

After the registration of each segment with respect to the LANDSAT image has been verified and any necessary local re-registration performed, it is possible to determine the corresponding LANDSAT pixels for any field of a segment. This mapping information is generated and stored in the segment mask file. The mask file generator program can also flag boundary pixels, i.e., those pixels that lie on field boundaries and are thus mixed spectrally with respect to the signatures of the two adjacent fields. Mask files are generated for each segment and LANDSAT frame pair using a scan-conversion procedure on the digitized field polygons. The resulting set of field-pixel correspondences are then compressed via run-length encoding.

IV. SELECTION, SAMPLING, AND TABULATION OF FIELDS

The selection of ground truth samples for classifier training involves several processes. First, a file may be created containing only pixels which meet certain criteria, such as having a particular ground cover (or one of a group of covers), and belonging to a certain group of segments. Such a file is called a packed file. An alternative way to create a packed file is to sample statistically from all fields which meet the criteria. This permits classification results to be evaluated with respect to the remainder of the fields. Also, once clustering or classification has been done, a feature is available for tabulation of pixels by the various classification (or clustering) categories and the ground covers actually recorded for the fields so that the results may easily be checked. This tabulation process will also produce a table file which can then be used for estimation as will be described below. These processes are described in more detail elsewhere.[14]

A. SELECTION CRITERIA

Two types of selection must be performed. First, a region or collection of segments to process must be chosen by the user. Second, criteria, which we call "options", must be selected. These two phases are called SELECT REGION and SELECT OPTIONS respectively.

A region is a collection of segments which will be processed in the course of a particular analysis. However, to avoid the tedium of entering many segment numbers, a short notation is available for describing groups of segments having common attributes, provided that a special catalog file has been created which lists the attributes of the segments of interest. The attributes include the LANDSAT frame(s) containing the segment, the land-use stratum to which the segment belongs, and the county in which the segment is found. An implicit assumption is made that a particular catalog will contain segments for one state only. Listing the attribute and appropriate parameter will then generate a list of segments satisfying that condition. For example "STRATA 11" will generate a list of segments which are in the catalog and have a land-use stratum of 11.

Once the collection of segments has been specified using SELECT REGION, the options specified under SELECT OPTIONS are used to choose the fields from within those segments to be processed. Again, to avoid entering by hand numerous field numbers, a shorthand notation is available allowing specification of field attributes, e.g. ground cover. These attributes are then used in selection of fields by checking the various fields in the ground truth file against the specified criteria for each segment

processed. The ground truth file is a computer readable representation of the ground survey data collected for a segment by USDA enumerators. In generating these ground truth files, careful checking is done to ensure consistency with the results of segment digitizations. Among the data collected for each field in the segment and stored in the ground truth file are: ground cover, field acreage, field appearance, intended use of crop, and date of field survey observation.

The user inputs SELECT REGION and SELECT OPTIONS statements using a simple Boolean expression language having the operators AND, OR, and NOT and allowing parentheses to facilitate grouping. A selection statement is such a Boolean expression followed by the terminator symbol "#". We now describe the SELECT REGION and SELECT OPTIONS separately in more detail.

1. SELECT REGION

SELECT REGION has the following operands: SEGMENT, FRAME, COUNTY, and STRATA. Each of the operands is followed by an appropriate parameter:

SEGMENT followed by a segment number;
COUNTY followed by a county name;
FRAME followed by a frame number or name;
STRATA followed by a land-use stratum number.

To get the segments desired, these basic descriptor variables are combined using AND, OR, or NOT. Thus, the expression "FRAME 2194-16042 AND STRATA 11" indicates all segments which are in the frame 2194-16042 and also in land-use stratum 11. On the other hand, "COUNTY CHAMPAIGN OR COUNTY DOUGLAS OR COUNTY MOULTRIE" generates segments which are in (any) one of Champaign, Douglas, or Moultrie counties. In fact, we may abbreviate the expression as "COUNTY ONEOF (CHAMPAIGN, DOUGLAS, MOULTRIE)" or more simply "COUNTY (CHAMPAIGN, DOUGLAS, MOULTRIE)".

Combining the above to obtain a more complex expression, we obtain "FRAME 2194-16042 AND STRATA 11 AND COUNTY (CHAMPAIGN, MOULTRIE, DOUGLAS)" which generates all segments in frame 2194-16042 and strata 11 and one of the counties Champaign, Moultrie, or Douglas.

2. SELECT OPTIONS

SELECT OPTIONS has a similar Boolean expression language with the operands COVER, ALL, BACKGROUND, FIELD, TRACT, SIZE, APPEARANCE, and INTENDEDUSE. The operands ALL (the entire area covered by the mask for a segment) and BACKGROUND (areas outside the boundaries of all digitized fields) have no parameters. The remaining operands must be followed by a parameter as follows:

COVER	followed by a ground cover name;
FIELD	followed by a full field name including the segment number, the tract (one or two letters), and the field number within the segment and tract;
TRACT	followed by a segment number and tract (one or two letters);
SIZE	followed by a relational operator and size in acres;
APPEARANCE	followed by an appearance code name;
INTENDEDUSE	followed by an intended use code name.

Except for FIELD, these return all fields selected from the various segments generated by SELECT REGION. FIELD gives a specific field in one segment and is used chiefly for packing using the fields generated by field sampling.

In SELECT OPTIONS additional operators are available to handle boundary pixels. To improve classifier training, it is often useful to eliminate boundary pixels. To eliminate such pixels, the unary operator "-" is used. The default is to include boundaries. For completeness, the operator "+" is provided to specify explicitly the inclusion of boundary pixels.

As an example of SELECT OPTIONS usage, "COVER CORN" or equivalently "+COVER CORN" returns all fields with a ground cover of corn and will also return the boundary pixels when generating a packed file. To exclude boundary pixels, we would use "- COVER CORN".

As in SELECT REGION, the ONEOF construct may be used so that "-(COVER CORN OR COVER WHEAT OR COVER SOYBEANS)" may be entered as "-COVER ONEOF(CORN,WHEAT,SOYBEANS)" or simply as "-COVER(CORN,WHEAT,SOYBEANS)".

Of course, we may use the AND operator as in "-(COVER CORN AND APPEARANCE MATURE)" to indicate all fields with a cover of corn and an appearance code of mature, all without boundary pixels.

B. PACKING

A packed file is one which contains pixels of either raw or categorized data taken from a collection of segments obtained using SELECT REGION and from fields meeting the criteria specified within SELECT OPTIONS. These pixels may then be used for any kind of further analysis as representing only pixels of those categories of interest. Spatial relationships between pixels in a packed file are lost; however, sufficient information is retained in the packed file to restore the spatial relationship if needed; this is done by unpacking a packed file.

The packing process requires a mask and a ground truth file for each segment processed as well as a data (window) file from which the

pixels to be packed will be extracted. The files are given agreed-upon default names used by this and other programs in the EDITOR system to avoid user entering of several file names for each segment. In addition, the data or pixels to be used for several segments may be placed in one file, known as a multi-window file, for convenience in handling and to save disk space.

The packing process then consists of using SELECT REGION to get a list of segments and then using SELECT OPTIONS code on each segment to generate a list of fields which meet specified criteria. The mask file, obtained by digitizing the segment, is used to determine which pixels fall within the specified fields. Pixels thus masked are moved into the packed file.

C. TABULATION

Once categorization (clustering or classification) has been performed on packed files of LANDSAT MSS data, it is often useful to see a tabulation of the categories and the various ground covers. For this purpose, the TABULATE option is used on a packed categorized file. Tables may be generated individually for each segment or for all segments in the region combined. The rows of the table displayed represent verified ground covers and the columns spectral categories. This allows checking the accuracy of the categorization. The tabulation process uses the mask and ground truth files along with information representing the user-entered SELECT REGION and SELECT OPTIONS to assign the pixels to the various fields. Thus, the correspondence between the category (of the pixel in the packed categorized file) and the ground cover (of the field as recorded in the ground truth file) is obtained.

In addition to displaying the table, it is also possible to create a table file containing the tabulated information. Such a table file may then be used to produce another tabular printout showing the percentage of correct categorization by cover type. Such table files are also saved for later use as one of the inputs to the sample estimation procedure to be described below.

D. SAMPLING OF FIELDS

Sampling allows selection of a sub-sample of fields from the larger sample of fields in the segments chosen by SELECT REGION and satisfying the criteria specified within SELECT OPTIONS. There is a slight difference in the use of SELECT OPTIONS here in that boundary specifications are ignored. Thus, a field is in the sampling universe if SELECT OPTIONS returns its interior, boundaries, or both.

The user may select one of two types of sampling: random sampling or systematic sampling. Once the type of sampling has been chosen, the user may select one of five types of probabilities for selection. All of these result in the selection of a sample proportional to field size (see Hansen, et al.[15]) These probabilities in effect give different weights to fields in different land-use strata. The five different selection probabilities are:

probabilities all equal
probabilities proportional to unexpanded digitized acres
probabilities proportional to expanded digitized acres
probabilities proportional to unexpanded reported acres
probabilities proportional to expanded reported acres

Here "reported acres" refers to the acreage for the field recorded in the ground truth file. "Digitized acres" refers to the acreage computed for the digitized field and stored in the mask file. The modifier "expanded" refers to an expansion factor for each segment which is an attribute for that segment. This expansion factor is derived by USDA based on their statistical sampling methodology and is stored in the catalog file. Then, "unexpanded" indicates that this expansion factor is to be ignored, or equivalently, set to 1.0 for all segments.

1. SELECTION PROBABILITIES

The type of selection probabilities chosen is used to build a vector P which is then used in the sampling. In the following, let

N = total number of fields in the generated sampling universe

e_i = the expansion factor for field i

d_i = the digitized acres for field i

r_i = the reported acres for field i

k = any number between 1 and N.

The order in which the fields occur in the sample universe is arbitrary (but not random).

For equal probabilities we have

$$P_k = \sum_{i=1}^{k} 1.$$

For probabilities proportional to unexpanded digitized acres we have

$$P_k = \sum_{i=1}^{k} d_i.$$

For probabilities proportional to expanded digitized acres we have

$$P_k = \sum_{i=1}^{k} e_i d_i.$$

For probabilities proportional to unexpanded reported acres we have

$$P_k = \sum_{i=1}^{k} r_i.$$

For probabilities proportional to expanded reported acres we have

$$P_k = \sum_{i=1}^{k} e_i r_i.$$

3. TYPES OF SAMPLING

Using the above notation, for random sampling we generate a random integer I between 1 and P_N and find L such that $P_{L-1} < I \leq P_L$. Field L is then selected to be in the sample. We make the assumption that an extra element $P_0=0$ exists. The process is repeated until the number of fields requested by the user is generated for the sample, taking care to enter any one field in the sample only once.

For systematic sampling, we generate a random integer m between 1 and P_N/N. Then, if V is a vector with N elements, let

$$V_i = m + (i-1)P_N/N, \text{ for } i=1,2,\ldots,n$$

Let S be the requested number of fields in the sample. Then for j=1,2,...,S find L such that

$$P_{L-1} < V_j \leq P_L$$

and select field L for the sample, letting $P_0=0$ as before. Care is taken that no field is entered into the sample more than once. If need be to fulfill the size of the sample requested by the user, the process is repeated as often as necessary over only those fields not yet included in the sample.

V. ACREAGE ESTIMATION SUBSYSTEMS

This section describes the software currently available in EDITOR to do acreage estimation based on a stratified sampling design. The mathematical basis of the system will be described as well as its implementation details.

The acreage estimation command has two main functional subcommands which perform the following functions:

1. Sample estimation
2. Large scale estimation

We will describe these functions together since they depend on the same statistical theory.

Sample estimation computes regression, ratio, and reciprocal of probability of selection (commonly called direct expansion) estimates for any region (set of segments) selected by the user with SELECT REGION. The region selected is usually all segments in a set of contiguous counties. Acreages may be estimated for any cover type selected by the user with SELECT OPTIONS. SELECT OPTIONS works as described above except that any input relating to boundaries are ignored so that a particular field enters into the acreage computation if its interior, boundaries, or both are returned. A table file (see Section IV.C) is necessary to do the regression and ratio estimation. Any or all land use strata may be pooled and some, such as non-agricultural or urban, may be deleted.

When a regression or ratio estimator is specified, the sufficent statistics (means, variances by strata or pooled strata) may be output to a file called an estimator parameter file for use in large scale estimation. Large scale estimation is used to estimate acreages for large geographic regions using all LANDSAT imagery for that area and not just a sample. Printed output from sample estimation is as follows:

1. coefficient of determination (r-square),
2. an estimate of the variance of the estimated acreage for the cover type in the segments selected using SELECT REGION, and
3. the relative statistical efficiency of the regression estimator vs. the direct expansion estimator for the dependent variable.

A. IMPLEMENTATION DETAILS

The direct expansion option of sample estimation is capable of using three different variables to estimate the acreages for a region selected by the user:

1. Digitized acreage -- This is the acreage computed by digitizing each field from an aerial photograph and having the crop type specified from the ground-truth file. This acreage is computed and stored in the mask file.
2. Reported acreage -- This is the acreage reported by USDA enumerators and stored in the ground-truth file.
3. Pixel acreage -- If the user so desires, all pixels classified into a single category or several categories from a table file may be used to estimate the acreages for the region selected. A user supplied constant is needed to convert pixels to acres.

The computations performed for the direct expansion as well as the ratio and regression options will be given below. For the segments generated by the SELECT REGION specified, let

L = total number of different strata to which the segments generated by SELECT REGION belong

N_h = total number of frame units in the SELECT REGION, for the h-th strata.

The set of all possible segments (called frame units) for a state is stored in a file called the frame_unit file by county and strata.

$N = \sum_{h=1}^{L} N_h$ = total number of frame units (over all strata)

n_h = total number of segments selected using SELECT REGION for the h-th strata

$n = \sum_{h=1}^{L} n_h$ = total number of sample segments (over all strata) generated by SELECT REGION

y_{hi} = { Total digitized acreage for all cover types selected for the i-th segment in the h-th strata. or Total reported acreage for all cover types selected for the i-th segment in the h-th strata.

x_{hi} = total number of pixels corresponding to the cover type selected for the i-th segment in the h-th strata.

The set of pixels corresponding to the cover type(s) selected using SELECT OPTIONS is stored in a table file for each segment (described above). The user must specify the table file name and the category number(s) associated with the cover type(s).

The direct expansion estimate of the total acreage in the region for a particular cover type is:[16]

$$\hat{Y} = \sum_{h=1}^{L} N_h \left(\sum_{j=1}^{n_h} y_{hj} \right) / n_h$$

and the estimated variance of (1) is:[16]

$$v(\hat{Y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h (n_h - 1)} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

A ratio estimate of the total acreage in the region for a particular cover type is:[17]

$$\hat{Y}_{RATIO} = \sum_{h=1}^{L} \left(\bar{y}_h / \bar{x}_h \right) X_h$$

$$= \sum_{h=1}^{L} r_h X_h \quad , \text{where } r_h = \bar{y}_h / \bar{x}_h$$

This ratio estimate is computed in the large scale estimation process.

Note that X_h is the total pixels by stratum in an entire large region, typically a county or group of counties, classified into the categories corresponding to the cover type(s). The computation of totals by stratum with a strata mask is called aggregation. The aggregation process is similar to the tabulation process described earlier; however larger mask files must be handled for large data sets and thus the distinction between these two processes. Aggregation may be performed on the ILLIAC IV in batch mode for very large files or on TENEX in a timesharing environment for other files. The results of aggregation are stored in an aggregation file.

The variance of the ratio estimate is:[17]

$$V(\hat{Y}_{RATIO}) = \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h} \left(S_{h,y}^2 + r_h^2 S_{h,x}^2 - 2 r_h \rho_h S_{h,y} S_{h,x} \right)$$

where,

ρ_h = sample correlation coefficient between x and y for the h-th strata

$S_{h,y}^2$ = sample variance for the h-th strata for the y variate

$S_{h,x}^2$ is similarly defined.

The regression estimate for a SELECT REGION for a particular cover is:[18]

$$\hat{Y}_R = \sum_{h=1}^{L} N_h \cdot \bar{y}_{h(reg)}$$

$$\bar{y}_{h(reg)} = \bar{y}_h + \hat{b}_h (\bar{X}_h - \bar{x}_h)$$

$$\hat{b}_h = \frac{\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)(y_{hj} - \bar{y}_h)}{\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2}$$

and the estimated variance is:[18]

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{N_h^2}{n_h} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 \cdot \frac{1 - \rho_h^2}{n_h - 2}$$

Note that $Var(\hat{Y}_R)$ is computed from only the sample and from the frame units in the region selected (which are known apriori); thus classification strategies that minimize this variance (rather than maximize the percent correct) can be determined.

Of course the computation of $\bar{X}_h$, the mean number of pixels classified into the categories for the cover of interest, is for the population and not just the sample. Thus large

subsets corresponding to political boundaries such as counties must be classified and aggregated by strata to compute these estimates. The estimation of acreages using aggregation files is done using large scale estimation.

In large scale estimation an option exists for regression estimation with subsets of the area originally selected. With this option one can estimate the acreages for an individual county and estimate the variance of that estimate. Since the regression estimator is linear, a further property of this estimator is that the individual counties add to the total for the SELECT REGION.

The regression estimate for the acreage of an individual county or a proper subset of counties is:

Let $S^2_{h,y}$, $S^2_{h,x}$, $\hat{b}_h$, $\sum_{i=1}^{n_h} x^2_{hi}$, $\sum_{i=1}^{n_h} x_{hi}$, n_h

be computed from the SELECT REGION sample described earlier, and let:

$N_{h,c}$ = total number of frame units in the h-th strata for the subset of c counties in the SELECT REGION, and let

$\bar{X}_{h,c}$ = total number of pixels classified as a cover for the subset of c counties divided by $N_{h,c}$.

$X_{h,c}$ is computed by aggregating the individual aggregation files for each county. Then, the estimate of the total for the set of c counties is:

$$\hat{Y}_{REG,c} = \sum_{k=1}^{L} N_{k,c}(\bar{y}_k + \hat{b}_k (\bar{X}_{k,c} - \bar{x}_k))$$

And the variance of this estimate is:

$$v(\hat{Y}_{REG,c}) = \sum_{h=1}^{L} N^2_{h,c} \frac{N_h - n_h}{n_h} S^2_{h,y} \frac{n_h - 1}{n_h - 2} \cdot$$

$$(1 - \rho_h^2)\,(I(c) + \frac{1}{n_h} + \frac{(\bar{X}_{h,c} - \bar{x}_h)^2}{\sum_{i=1} (x_{hi} - \bar{x}_h)^2})$$

where

I(c) = 1 if O(c) < total number of counties wholly contained in the SELECT REGION
= 0 otherwise

O(c) is the cardinality of the set c.

REFERENCES

1. Robert M. Ray, Martin Ozga, Walter E. Donovan, John D. Thomas, Marvin L. Graham, 'EDITOR An Interactive Interface to ILLIAC IV - ARPA Network Multispectral Image Processing Systems', CAC Technical Memo No. 114, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, June 1975

2. Walter E. Donovan, 'Oblique Transformation of ERTS Images to Approximate North-South Orentation', CAC Technical Memo No. 38, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, November 1974.

3. Walter E. Donovan, Martin Ozga, Robert M. Ray, 'Compilation and Geometric Registration of ERTS Multitemporal Imagery', CAC Technical Memo No. 52, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, May 1975.

4. John Thomas, 'An ILLIAC IV Algorithm for Cluster Analysis of ERTS-1 Imagery', CAC Technical Memo No. 17, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, May 1974.

5. John Thomas, 'An ILLIAC IV Algorithm for Statistical Classification of ERTS-1 Imagery', CAC Technical Memo No. 18, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, May 1974.

6. Robert M. Ray, John D. Thomas, Walter E. Donovan, and Phillip H. Swain, 'Implementation of ILLIAC IV Algorithms for Multispectral Image Interpretation, Final Report', CAC Document No.112, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, June 1974.

7. Phillip H. Swain, 'Implementation and Evaluation of ILLIAC IV Algorithms for Multispectral Image Processing', LARS Technical Memorandum T-16 111273, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, July 1974.

8. Martin Ozga, John Thomas, 'An ILLIAC IV Algorithm for Statistical Classification of 8-channel Multispectral Data', CAC Technical Memo 54 Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, November 1974.

9. Martin Ozga, 'An ILLIAC IV Algorithm for Cluster analysis of 8-channel Multispectral Data', Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, June 1975

10. C.P. Gleason, G.A. Hanuschak, R.R. Starbuck, R.S. Sigman, 'Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment', Forth Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue university, West Lafayette, Indiana, June 1977.

11. Martin Ozga, Walter E. Donovan, Robert M. Ray III, John D. Thomas, Marvin L. Graham, 'Data File Formats for Processing of Multispectral Image Data' CAC Technical Memorandum No. 19, Center for Advanced Computation, University of Illinois at Urbana-Champaign Urbana, Illinois, October 1976 (revised).

12. Walter E. Donovan, Martin Ozga, 'Retrieval of LANDSAT Image Samples by Digitized Polyonal Windows and Associated Ground Data Information', CAC Technical Memo No. 57, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, August 1975.

13. Robert M. Ray, Harold F. Huddleston, 'Illinois Crop-Acreage Experiment', Third Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, July 1976.

14. Martin Ozga, 'Selection, Sampling, and Tabultaion of Masked Files in EDITOR', CAC Technical Memo No. 79, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, December 1976

15. Morris H.Hansen, William N. Hurwitz, William G. Madow, Sample Survey Methods and Theory, John Wiley and Sons, Vol. I, pp. 341-348.

16. William G. Cochran, Sampling Techniques, (2nd Ed.), John Wiley & Sons, pp. 91-92.

17. William G. Cochran, Sampling Techniques, (2nd Ed.), John Wiley & Sons, pp. 167-175.

18. William G. Cochran, Sampling Techniques, (2nd Ed.), John Wiley & Sons, pp. 200-208.

ACKNOWLEDGEMENTS

The authors wish to acknowledge several individuals for their contributions: Robert M. Ray for his systems management and leadership in the development of EDITOR, Harold F. Huddleston and William H. Wigton for their statistical leadership and guidance, Robert R. Starbuck and Richard S. Sigman for their work on the sampling and acreage estimation command in EDITOR, and Paul W. Cook for his initial ideas on segment digitization, registration and plotting. Special support was also received by the CAC authors from NASA and USGS/DI for the use of the ARPANET and the Illiac IV.