Using the EDITOR System for Crop Area Estimation

by Michael E. Craig

FEB. 1980

I. INTRODUCTION

The EDITOR System was designed to make agricultural crop area estimates using LANDSAT MSS data combined with ground-gathered information. The purpose of this report is to highlight the features of EDITOR essential to this task and to discuss them from the analyst or implementors point of view. For the analyst, the process of estimation begins with collection of ground data and ends with the distribution of final estimates. The various steps in the process are discussed with respect to the role of the EDITOR system.

II. GROUND DATA DESCRIPTION

This system was built around the concept of a segment. A segment is an area of land (well defined by permanent boundaries on a photograph or map) that has been randomly selected as a sample unit in some land-use stratum. For each such segment, enumerators make one or more visits during the growing season and record the crop or ground cover and size of the various fields found. Other information; such as livestock present, irrigation practice, intended use, or percent emerged; may also be collected. The segment data alone contains enough information to make area estimates with measurable precision.

Data gathered for each segment usually come in two forms: a questionnaire and a segment photo. Both of these must be transferred to digital form to be

useful for the computer analysis. Questionnaire data are keypunched and these records are used as input to the "Ground Truth Editor" subsystem. The output product is one "ground truth" file per segment containing field level information. In order to get the segment photos in a machine readable form a process called digitization is used. Using this process field boundaries are mapped into a geographical coordinate system (latitude, longitude). The "Registration and Digitization" subsystems used for this process produce one file per segment called a segment network file.

III. LOCATING SEGMENTS IN LANDSAT DATA

The next step in analysis is to locate the segments in the LANDSAT data. This step includes several jobs such as: ordering LANDSAT CCT's, preprocessing CCT's, global calibration, and finally local segment calibration. The LANDSAT CCT's as supplied by EROS Data Center are in the BIL format (Band Interleave by Line) and must be reformatted to the EDITOR readable BIP format (Band Interleave by Pixel). This preprocessing task is performed outside EDITOR at a separate computer facility.

Once the LANDSAT tapes are available to EDITOR, the global calibration (scene registration) is begun. This process involves three sets of materials: LANDSAT scene paper products, a set of maps covering each scene, and gray-scale printouts representing spectral bands for specified areas. Several EDITOR subsystems are needed for registration. In addition to the "Registration and Digitization" subsystems mentioned earlier, EDITOR programs are needed to extract windows from scene tapes and print these windows in gray-scale form. In EDITOR these programs are called "Tape Reading to Create Window Files" and

"Print Window Files". The output of the registration is a bivariate polynomial transformation between (latitude, longitude) and (row, column). The coefficients of this transformation are stored in the global calibration file.

The final task in locating segments in the LANDSAT data is called local segment calibration. The global calibration is used to predict the location of each segment and a window of this area containing a 20 pixel boundary layer is extracted from the scene tape. Gray-scale prints of the windows are obtained using the same EDITOR programs as before. Using the segment networks and the global calibration as input, the subsystem "Plot Functions" gives segment/field boundary plots at the same scale as the gray-scale print. The segment plots are then overlaid on the gray-scales at the predicted segment location. Manual interpretation of the field boundaries in the plots versus the location of field patterns visible in the various MSS bands gives the actual location of the segment (versus the predicted location). Any visible difference between the actual and predicted locations is called a segment shift and this information is contained in a local calibration file produced again by the "Registration and Digitization" subsystems.

When segment location is complete, this calibration information is used to generate a mask file for each segment which contains field containment and boundary values for each pixel found in the segment window. The mask files are the links between ground truth information and the LANDSAT MSS data by pixels. Masks are generated by the EDITOR programs "Mask File Functions".

IV. SIGNATURE ANALYSIS

After the correspondence between ground truth information and LANDSAT pixels is established the next major analysis step is to create signatures

for the various cover types as needed for each scene or analysis area. The USDA/ESCS approach has been to use a modified supervised clustering approach to determine signatures. All pixels of a known cover type are gathered together as with any supervised clustering approach. (In EDITOR this process is called packing a file and is accomplished using the "Field Selection for Analysis" subsystem). We then use an unsupervised clustering algorithm within each cover type to get one or more signatures for each cover. This algorithm was built on the LARSYS ISODATA procedure and is initiated using the EDITOR "Ordinary Cluster" program. Various combinations and poolings of these signatures are created using the "Statistics File Editor" program.

The clustering process produces one or more collections of signatures (called statistics files) to represent the cover types found in the analysis area. These statistics files along with the segment raw data windows are the inputs to the small scale classification(s) which is the next step in analysis. Each statistics file defines a set of maximum likelihood descriminant functions which are used to give each segment window pixel a category number corresponding to a signature in the statistics file. Tabulations by cover type and category are made using the "Field Selection for Estimation" program for each segment and for the collection of all segments. These tabulations show the number of pixels of a known cover type that were classified to the various categories (i.e. cover types) in the statistics file. This information is used by the "Percent Correct Classification" program to calculate omission and comission errors for each classifier (as defined by a specific statistics file). The error rates may be one source of deciding which of the statistics files used is the "best".

IV. REGRESSION ESTIMATION

The USDA/ESCS approach to utilizing LANDSAT for crop estimation is to use it as an auxiliary or independent variable and the ground truth as the dependent variable in a regression estimator. A simple linear regression is calculated using the ground truth files and getting the independent pixel data from the segment level tabulation of the small scale classification. The "Estimate Acreages" subsystem is used to calculate these regressions and creates an estimator parameter file for use in actual estimation. This parameter file contains the coefficients of the regression plus sums of squares of the segment data for later variance calculation.

Using the segment level data to calculate the regression coefficients is referred to as small scale estimation. A small scale regression estimator is made for each statistics file used to classify the segment data. The R-squared values by stratum measure the relationship between the ground truth data and the classified pixels at the segment level. Our procedure is to choose as the "best" classifier the statistics file with the highest R-squared values for the crop or crops of interest.

Once the final classifier is chosen, all LANDSAT raw data pixels in the analysis area are classified. This process is called Large Scale Classification. Assuming the analysis area is a large part or all of a scene, classification is done at the ILLIAC IV computer in California. When this machine is not available, classifications are done using the window approach utilized in small scale classification for segment windows.

The output of the large scale classification must be tabulated by the various land-use stratum present in the estimator parameter file. This process is called aggregation and is initiated using the "Aggregation Functions"

subsystems. In order to do this, aggregation mask files are created from a digitization of land-use stratum boundaries (similar to segment digitization using stratum boundaries to define fields). These masks then map each pixel in an analysis area to a land-use stratum. The aggregation output file is then combined with the estimator parameter file and the final large scale regression estimate is calculated.

The large scale regression estimate is then compared to the estimate made using ground data only. One final measure of the effectiveness of the regression is calculated by the EDITOR subsystem "Estimate Acreages". This measure, called the Relative Efficiency or RE, is the ratio of the variance of the ground data estimate to the variance of the regression estimate. Both ground data estimates and regression estimates can then be distributed to the various interested parties.

VI. SUMMARY

The following areas that are addressed by the EDITOR system are essential for regression estimation using LANDSAT remotely sensed data. A method is needed to get ground gathered data (both photos and questionnaires) into digital form for computer use. Location of segment data in the LANDSAT imagery requires a registration procedure that maps ground coordinates into LANDSAT row/column coordinates. If the LANDSAT data is not in the correct format, some preprocessing is needed. Once ground segments are located on the MSS imagery, windows around them must be extracted from the whole frame data. Functions to prepare gray-scales and plots are essential to final segment location. Multivariate clustering and classification algorithms must be

implemented that use training data to create signatures and eventually categorize individual pixels. Allowances must be made for classification of large amounts of data. Once pixels are classified they must be tabulated to the correct level (segment or land-use). Software is needed to calculate ground data only estimates, regression parameters, R-squared values, and eventually large scale regression estimates.

VII. NOTE OF CAUTION

The EDITOR commands to process ground data and perform regression estimation distinguishes EDITOR from other image processing software. Unfortunately, the computer programs for these EDITOR commands cannot be easily transferred to other computers (non-DEC-10) on other applications. Conversion to non-DEC-10 computers is difficult if not impossible because of extensive program use of operating system features for input and output, the use of the SAIL programming language which presently can be compiled only on DEC-10 computers, and becuase of the use of DEC-10 assembly language. Also, by design EDITOR's ground-data processing is specific to USDA's June Enumerative Survey. Other types of ground data, such as different cover types, sampling plans, or sample-unit identification schemes cannot be handled by EDITOR without program modifications. In view of these restrictions, direct conversion of EDITOR programs to other machines on non-USDA applications is not recommended. Instead EDITOR output and user documentation should be used as guides by new users in such situations to develop their own computer programs.