March 31, 1989

Using the PEDITOR System for Crop Area Estimation: An Analyst's View

by Michael E. Craig

I. INTRODUCTION

The PEDITOR System was designed to make agricultural crop area estimates using satellite digital data combined with ground-gathered information. The system currently accepts data from three satellite sensors: Landsat MSS, Landsat TM, and Spot MSS. The purpose of this document is to highlight the features of PEDITOR essential to this task and to discuss them from the analyst's or implementor's 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 PEDITOR system.

PEDITOR is written mainly in PASCAL. It is currently maintained on a MicroVax 3500 computer at USDA/NASS in Washington, DC. This same version is also maintained to run on IBM-compatible personal computers using the Oregon Pascal-2 compiler. A somewhat different version is currently working on SUN/3 microcomputers; however the SUN version does not contain some of the newer modules and updates. We hope to bring the various versions back together in the future as soon as the compiler differences can be worked out.

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, an enumerator makes one or more visits during the growing season to 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 key-entered 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. Information such as county, stratum, and map number is contained (for all segments) in a segment catalog file.

III. LOCATING SEGMENTS IN SATELLITE DIGITAL DATA

The next step in analysis is to locate the segments in the satellite digital data. The segment location step includes several jobs such as: ordering satellite computer compatible tapes (CCT's), preprocessing CCT's, global calibration, local segment calibration, and finally segment mask generation. The satellite data CCT's are usually in the BIL format (Band Interleave by Line) and must be reformatted to the PEDITOR readable BIP format (Band Interleave by Pixel). This preprocessing task is usually performed at a separate supercomputer facility.

Once the reformatted tapes are available to PEDITOR, the global calibration (**scene registration**) is begun. This process involves two sets of materials: satellite image paper products and a set of maps covering each scene. Several PEDITOR subsystems are needed for registration. **Corresponding points** are selected on the image and map products, digitized, and used in a least squares regression analysis to predict the transformation between scene and map. 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 satellite digital 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 around each segment is extracted from the scene tape. The pixel data are stored in a multi-window file. Gray-scale prints of the windows are obtained using the "Subwindow and Print Window" programs. 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 one or more bands gives the actual location of the segment (versus the predicted location). Any visible difference between the actual and predicted segment locations is recorded in fractional pixels for both row and column. Shifts for all segments in a given scene are maintained in segment shifts file for later use. New developments to PEDITOR have implemented the local segment calibration process on graphics displays so that the segments can be shifted interactively, and no printing or plotting is required.

When local segment location is complete, this calibration information is used to generate a segment 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 satellite digital data by pixels. Masks are generated by the PEDITOR 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 statistical signatures for the various cover types as needed for each scene or analysis area. The USDA/NASS 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. The pixel data is stored in packed files, one for each cover type. (In PEDITOR 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 PEDITOR **"Ordinary Cluster"** program. Another clustering algorithm, called **"CLASSY Cluster"**, is sometimes used with large packed files. The CLASSY algorithm is based on a mixture of multivariate normals but is not stable for small numbers of pixels. The output statistics representing the clusters are stored in statistics files. Statistics files for the various covers are combined into one output statistics file using the "Statistic File Editor" program. This program also allows pooling of input clusters, deletion of clusters, and the addition of weighting factors (in the form of prior probabilities).

The final statistics file 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 discriminate functions. The classification process evaluates these functions to give each segment window pixel a category number corresponding to a signature in the statistics file. The output is stored in a **classified** window (either packed files or window files can be classified). 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. Tabulations can be studied by various methods to determine if the classification is acceptable. These methods might include study of individual segments, performance in a regression estimator, and percent correct classification. Several overall statistics files, representing different combinations or weightings of clusters, might be compared before a final file is chosen.

V. REGRESSION ESTIMATION

The USDA/NASS approach to utilizing satellite data 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.** Classification error rates are generally too large to accept direct pixel counts as the final estimate. 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 raw data pixels in the entire scene or analysis area are classified. This process is called **Large Scale Classification.** Assuming the analysis area is a large part of all of a scene, classification is done on a supercomputer such as a CRAY XMP. 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, called county stratum networks, which are similar to segment networks but 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 resulting estimates and variances, by strata, are stored in an estimator results file.

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 PEDITOR subsystem "Estimate Acreage". 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 PEDITOR system are essential for regression estimation using satellite 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 satellite imagery requires a registration procedure that maps ground coordinates into satellite row and column coordinates. If the satellite data is not in the correct format, some preprocessing is needed. Once ground segments are located on the satellite imagery, windows around them must be extracted from the whole frame data. Functions to prepare gray-scales and plots (either on paper or graphically) are essential to final segment location. Multivariate clustering and classification algorithms must be implemented that use traing 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 strata). 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 PEDITOR commands to process ground data and perform regression estimation distinguish PEDITOR from other image processing software. The emphasis on an output statistical estimate rather than a picture product or classification may sometimes be at odds with your remote sensing objectives. For example, someone wishing to create a classified image for input into a Geographic Information System might choose a statistics file that maximizes percent correct rather than the regression R-squared.