PEDITOR - STATISTICAL IMAGE ANALYSIS FOR AGRICULTURE

Martin Ozga and Michael Craig
National Agricultural Statistics Service, USDA
3251 Old Lee Highway, Room 305
Fairfax, Virginia 22030

ABSTRACT

PEDITOR is a modular system of PC programs specifically written to estimate crop acreages with measurable precision using satellite imagery combined with ground gathered survey data. PEDITOR has been under development within NASS for several years and portions and/or offshoots of it are available for other (non-PC) platforms. Currently, most PEDITOR functions can be accomplished with a PC under MS-DOS as the primary platform. One Landsat Thematic Mapper (TM) image as used by the PEDITOR system consumes 290 megabytes of disk space; as an example, the analysis of major Mississippi River Delta crop areas in Arkansas requires twelve TM images or 3.5 gigabytes of storage and generates over one thousand files. The ground data from sampled land areas is collected during an annual NASS operational survey; for Arkansas there are over 200 samples (known as segments) in the major crop regions. Each segment, with its field boundaries and corresponding tabular data, generates at least five computer files during the estimation process. Field boundaries from sample segments are located in the satellite data and used with clustering and pattern recognition techniques to train the computer to recognize crop types. Maximum likelihood classification is applied to entire scenes to cover large areas such as entire states or major portions of states. Regression estimation is used to generate estimates of crop acreage for major crops both for large areas and by county. Area displays are provided on the PC screen and on color printers to show crop distribution. Current analysis is in the Mississippi Delta region of Arkansas, but in the past other larger areas have been processed.

INTRODUCTION

PEDITOR is a system of computer programs designed to process digital satellite imagery with the goal of estimating crop area over large areas (Angelici 1986; Ozga 1985; Ozga, et al. 1992). PEDITOR is undergoing continuing development at the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture. At NASS, PEDITOR is used mostly by a small group of analysts in a research environment allowing changes and corrections to be made. However, these same users have operational projects for certain areas which are constrained by Agency deadlines for generation of estimates (Hanuschak and Craig 1993).

Although its primary goal is crop area estimation, PEDITOR naturally includes many modules or functions (Swain and Davis 1978) generally found in commercial systems

processing satellite imagery. These functions include scene registration, clustering, classification and graphic display capabilities. PEDITOR is written in the Pascal programming language and is coded in such a way that most modules are intended to be portable to and indeed have run on various machines and operating systems, although the current emphasis is on the PC under MS-DOS. Modules which do graphical displays are coded to run only on the PC.

Besides its use for area estimation, PEDITOR has formed the basis of the Computer Assisted Stratification and Sampling (CASS) system developed by NASS and the Ames Research Center of the National Aeronautics and Space Administration (NASA-Ames). The CASS system was designed for the automation of stratification and land use mapping by displaying satellite imagery and digital map data on graphics workstations (Cotter and Mazur 1991). CASS represents a different set of capabilities and is not described here. Also PEDITOR and derivatives have found some use outside the United States, particularly in the European Economic Community Research Center in northern Italy (Annoni 1991) and, independently, in Spain.

SYSTEM STRUCTURE AND PORTABILITY

The PEDITOR system is divided into over fifty autonomous programs. Each program is run independently of the others and is called explicitly by the user. To retain flexibility in the research environment, no 'menu' of program modules is used. Communication between programs is by the various files, that is, one program will write a file and another will read that file.

The PEDITOR code is divided into a large portable portion and a small machine dependent portion. The portable portion is coded in "standard" Pascal, with two extensions found in most compilers, namely separate compilation and textual includes. The portable portion also makes the assumption that all integers and floating point values are 32 bits in length and that bytes are unsigned values between 0 and 255 occupying 8 bits.

The machine dependent portion is largely related to input-output functions and must be coded separately for each different machine and operating system. If coded in Pascal, it may use non-standard features. If coded in some other language, the routines must be callable from Pascal using the same procedure calls. Using this methodology, we have been able to successfully run the PEDITOR code on PCs under MS-DOS, VAX/VMS systems, IBM mainframes under MVS/TSO, the VACCELERATOR (an add-on faster processor for VAX systems), and, with some difficulty, UNIX-based workstation systems such as SUN and Hewlett-Packard.

PEDITOR generates large numbers of files since it is not associated with nor does it attempt to implement a data base management system. To ease the burden for the user, the more numerous files are given standard names based on their content. The user does not explicitly enter these standard names. Rather, they are generated based on the data being processed. Since the naming conventions vary between operating systems, one machine dependent module handles the standard names. This module is coded in standard Pascal, but the names generated vary.

All PEDITOR files, except a few related to scene registration, are binary. Each binary file has a header with some description of the contents followed by the data. The description in the header is sufficient to determine the length of the file. Thus, PEDITOR programs never read data until end of file is reached. Also, no PEDITOR program ever modifies a file in place. If information must be changed, a copy of the file is made with the updated information.

PEDITOR binary files are portable in the sense that when a file is transferred from one machine to another by a binary file transfer method such as KERMIT or FTP, the file may be immediately read without having to do a conversion. All files have four standard types of data: byte, integer, floating point, and character. Byte data is values of 0 to 255 stored in 8 bits and is used for satellite imagery and other things. Integer data is 32-bit integers in twos-complement form. Floating point data is in the IEEE 32-bit format. Character data is ASCII. The machine dependent portion of the code converts this data, as required, between the standard format and the internal format of the host machine.

There is only one format for satellite imagery, the 'window' file. A window is a collection of pixels extracted from the satellite imagery; it is generally a rectangular set or subset of the image defined by the row and column coordinates of its top left and bottom right pixels. A one-window file may contain an entire scene, while a multi-window file contains several subsets from the scene. This file has a header describing the windows in the file followed by the data. The data is stored in band interleaved by pixel (BIP) format so that all the bands for a particular pixel are contiguous. The number of bands is specified in the header. Each band of each pixel is assumed to occupy one byte.

As satellite imagery is received at NASS, it is reformatted into the BIP format before being used by any PEDITOR programs. The reformatting process allows the same program code to handle data from the Landsat Thematic Mapper (TM), Landsat Multispectral Scanner (MSS), SPOT Multispectral Scanner, Indian Remote Sensing satellites instruments and other raster-based sensors.

Certain PEDITOR modules, related to registration, digitization, and graphics display, run on PCs under MS-DOS only. The display portions of these codes require VGA graphics, preferably the enhanced or "super" modes. High resolution options require graphics cards with one megabyte or greater video memory (VRAM).

USER INTERFACE

PEDITOR generally provides a simple alphanumeric interface to the user. A few programs related to registration provide a simple graphical interface. All commands may be entered in upper or lower case and may be abbreviated to as few characters as will make the command unique. Generally, entering a question mark (?) will give a short help message about the type of input expected. Sequences of command inputs which are repeated often may be placed in an ASCII text file and that file called by preceding it with an exclamation mark (!) when an input is requested.

Future plans call for considerable simplification of the user inputs. In the past, the user has been assumed to have had considerable experience with remote sensing and so PEDITOR has required the user to enter many parameters to fine tune the processing. In the future, many of these parameters will be automatically supplied to make the programs easier to use by less experienced users. This will involve encoding the standard operating procedures used in analyses in the past and, in some cases, expert system techniques. Further in the future, we plan to put PEDITOR into the WINDOWS environment.

PEDITOR PROCESSING PROCEDURES

Overview of Crop Estimation

Crop area estimation is an important part of the overall NASS mission. The major indication used for the official crop area estimates comes from a sample of over 16,000 areas of land, known as segments, selected throughout the United States. All segments are visited every year in early June and completely enumerated. In order to statistically select these areas of land for surveys, all land in each state is stratified based on land use and/or percent cultivation. Substrata may be created for special needs and for ease in sampling. The collection of strata and substrata boundaries is known as the area sampling frame (Cotter 1987).

Each substratum is further divided into similar sized sampling units. A number of these units, known as segments, are randomly chosen for surveys. Segments average approximately one square mile each, although this varies by stratum. Enumerators visit the segments and collect information about the crops in various fields by interviewing the farmers and by personal observation. The field boundaries of these segments are drawn on aerial photographs as a quality control measure. Crop area estimates based only on the ground enumerated data are generated using standard statistical techniques in the normal NASS operational procedure. Digital satellite imagery covering large areas is applied using a regression estimator approach to improve these estimates as described below.

Area frame strata and sample field boundaries are put in digital form and located on a map base using processes called **digitization** and **calibration**. Satellite scenes are **registered** to the same map base using polynomial regression. In some areas, two scenes over the same area but with different acquisition dates are combined to create a multitemporal scene. Since the scene registration is only approximate, segment boundaries must be locally calibrated using a technique known as **segment shifting**. The segments provide an excellent source of training data, or **ground truth**, since the fields have known locations and cover types.

All pixels for cover types of interest are extracted (or **packed**) from the segment windows of image data and **clustered** separately by type, generating files containing cluster statistics. Statistics files by type are concatenated together to make one overall file; after some editing, this statistics file is used to do a **maximum likelihood classification** of entire scenes or major portions thereof. With the strata boundaries also in digital form, **aggregation** may be done by category and strata to reflect the sample design. This aggregation, along with the segment data, is input to a **regression estimator** to obtain the final estimates. The various steps needed to arrive at the estimates will now be described in more detail.

Scene Reformat, Registration, and Multitemporal Imagery

As mentioned above, all scenes must be reformatted before they can be used by any PEDITOR programs. There is a reformat program for each type of sensor. The reformat programs also extract the registration supplied with the scene. Unfortunately, this registration is generally not satisfactory but serves as a useful starting point to get the final registration. The reformatted scenes are written to tape to be saved.

After reformat, satellite scenes must be registered to a map base. Registration is done on the PC using Digital Line Graph (DLG) data from USGS. First, the DLG data must be converted to a PEDITOR format. This conversion need only be done once and the files saved. The entire scene is displayed in a sampled form. Using the mouse, the user selects a likely point. The area around that point is displayed, unsampled, on the left of the screen and the corresponding DLG, as estimated by the registration supplied with the scene, is displayed on the right of the screen. The user selects a point in the scene and then the same point in the DLG. If such a correspondence cannot be found, the point is ignored and another tried. A number of points are so selected, being careful that they are, as much as possible, well distributed over the scene. This gives a collection of corresponding points, allowing, after some editing, least squares polynomials to be generated representing the final registration.

If a multitemporal (two-date) scene is to be created, one scene is designated as the primary scene and the other as the secondary scene. The coordinate system of the primary scene is used for the multitemporal scene. Both the primary scene and secondary scene are registered to a map base, although the registration of the secondary scene may be of lower quality than that of the primary scene, requiring fewer corresponding points.

A large number of blocks are taken on a grid from the primary scene. The registrations of the primary and secondary scenes are used to find the block location in the secondary scene for each block in the primary scene. These block pairs are then correlated to generate the final overlay. This final overlay, which is also expressed by least squares polynomials, assigns to each pixel in the primary scene a pixel in the secondary scene using the nearest neighbor rule. The correlation and overlay is presently done only on the VAX but will be moved to the PC in the near future.

Digitization of Segment and County Strata Boundaries

Two approaches are available to obtain digital representation of ground boundaries. Segment enumeration and county area frame boundaries may be digitized into vector-based polygons representing the fields or strata boundaries with a digitizing tablet connected to a PC. This is known as manual digitization. A PEDITOR program is used for this digitization and also for registration (or calibration) of the segments or counties to a map base using DLG or paper maps on a digitizing tablet. County area frame files in the same digital vector format may also be obtained directly from the CASS system.

Segment boundaries may also transferred to digital form using an approach known as video digitization. Video digitization runs on the PC and requires that the boundaries of the segment be traced on acetate. Using a video camera and a commercial frame grabber package, the image is then captured as a raster image for further processing. A standard

portable PEDITOR program is used to thin the lines and perform connectivity analysis on the image. The fields are labelled by displaying the outline of the segment on the PC screen and having the user enter the name of each field. The segment must also be registered to a map base using DLG or with paper maps on a digitizing tablet. The boundaries of the fields of a segment may be changed by deleting old boundaries and/or drawing new ones, either over satellite data or on a blank screen. The same techniques may be used to create additional segments based on features clearly visible on the satellite data, such as bodies of water.

Ground Truth and Segment Shifting

The ground data information about the size and crop for each field is collected by the enumerators in June. Some fields require follow up visits if the crop is not planted in June. Sample field information is retrieved from the operational survey and stored in a ground truth file for each segment. The survey data in the ground truth file are checked against that from the digitized segment files. Any discrepancies are either resolved or else the fields in question are marked as 'bad' and not used for training.

Despite registering both the scene and the segment, minor misregistration is often seen when the segment boundary is displayed on the scene. This is remedied by another program allowing the user to move the segment around on the scene to obtain the best fit. The distance of movement, referred to as the shift value, is used in generating the overlay of the segment onto the scene. It is important that the segment be correctly registered to the scene so that the pixels extracted for training truly represent the crops desired.

Packing and Clustering to Extract Training Information

Using the scene registration, the segment registration, and the shift, an overlay of the segment onto the coordinate system of the scene is created. This overlay is referred to as a mask file. The mask file specifies which pixels from the scene are contained in the various fields of the segment. In conjunction with the ground truth files, the mask files allow all pixels for any cover or group of covers to be extracted and placed into files called packed files.

The user specifies packing criteria by a list of segments and a boolean expression specifying which crops are to be included plus specific conditions for exclusion for this crop. Such exclusions might be pixels from fields identified as 'bad' from the ground editing or 'mixed' pixels which fall on or near field boundaries. Additional pixels are edited out after packing by the use of outlier detection and principle components techniques.

The final packed files are clustered to obtain statistics files, containing the means and covariances, for each crop. Generally, there will be several categories for each crop. This process is called **modified supervised clustering** and has been in use by NASS since the 1970's. The ISODATA cluster method is used, but with an additional modification allowing cluster splitting as well as merging (Bellow and Ozga 1991). The statistics files are combined, generally with little or no editing, into one statistics file to be used in the classification process. The editing, if required, is an attempt to remove categories that do not represent the crop in question. Such categories may still occur if a field is mislabeled, mixed or has small, non-contiguous areas which are not the labelled cover.

Classification and Aggregation

Maximum likelihood classification is used to determine the output category of each pixel. Prior probabilities, reflecting the likelihood of certain classes being in an area, may be used to modify the classification. Although classifications can be quite CPU intensive, particularly if a large number of categories are used, the newer PCs, particularly high-end Pentiums, may be used.

In much the same manner as with segment files, except with no graphics-based shifting, mask files are generated for the counties. The mask files allow summation of classified pixels by strata and category to be created to match the stratified sample design. This summation process is known as aggregation. Aggregation is generally performed on a county-by-county basis first, then county files are summed or aggregated to the desired region or analysis area level. These aggregation files are important inputs to estimation as population values. Once aggregation has been completed, the categorized file is no longer needed for estimation. However, it is generally saved on tape for use in displays.

Regression Estimation

Crop area estimation is done in four phases: small scale, large scale, accumulation and county estimation (Allen 1988, Bellow 1992). The small scale phase involves the calculation of single variable regression parameters based on the sample segments only. Segment ground truth information provides the independent variable and classification of the pixels found within segments provides the auxiliary variable in this regression. Analysis districts, which are the land areas to be included in a specific regression estimation, are defined by counties and parts of counties contained in one or more satellite images acquired on the same date.

Regression parameters are calculated for each stratum with a sufficient number of segments within an analysis district. Various classification approaches may be compared in this phase to select the "best" analysis district statistics file for large area classification. For example, classifications based on apriori versus equal prior probabilities might be compared. The final sample estimation selected as 'best' is then used to set parameters for the large scale estimation.

In the second phase, known as large scale estimation, entire counties and/or scenes are first classified using the "best" statistics file. These classifications are then aggregated to strata level and input as the auxiliary variable population values in a regression estimator. This provides the estimation for the covered area in each analysis district as well as providing a results file for accumulated estimation of larger areas.

A final state-level estimation is done with the accumulate estimation program. This program first calculates prorated estimates, based only on ground information, for areas not covered by satellite imagery. These areas include: strata within analysis districts with insufficient segments for regression, cloud covered regions within scenes and areas outside scene boundaries. The accumulate program then pulls together regression estimates from the various analysis districts and summarizes all types of estimates at the state level.

Finally, county or small area estimation may be done using the Battese-Fuller method (Battese 1988, Bellow 1994) based on the accumulated estimates for the analysis districts.

The Battese-Fuller approach uses the analysis district regression slopes and calculates individual county regression intercepts. The sample size (number of segments) in individual counties is too small to allow completely separate regressions for each county. County estimation has not been completely integrated into PEDITOR since some of the code is written in FORTRAN, which is called by a Pascal main program. The county estimation program reads PEDITOR files, however.

GRAPHICS DISPLAY

Currently all graphics displays of satellite imagery, segment boundaries, country strata, and DLG in PEDITOR are PC-based using the VGA graphics system, preferably in the enhanced higher resolution or "super" modes. The previously described registration and segment shifting programs are examples of the use of graphical displays. The segment shifting program has many additional capabilities. Sampled areas representing large areas of scenes or even entire scenes may be shown. This sampling is performed on-the-fly from the window file.

County boundaries, segment boundaries or DLG vectors may also be overlaid on these displays, with the county boundaries taken either from the digitized or mask files. Also, categorized data may be displayed, with the user selecting the colors to be assigned to the various categories from a menu of colors. Categorized display is useful in showing the distribution of various crops in an area.

Any PEDITOR graphics program screen can be captured to a standard TARGA image file format for printing or later display. PEDITOR can currently print directly to certain thermal wax color printers; a conversion to PostScript(c) format is also available. County level categorized displays in hard copy form are a useful adjunct to the numerical estimates; DLG boundaries overlaid on categorized images can be used to create land cover maps.

OPERATIONAL USE OF PEDITOR

PEDITOR has been used in various operational NASS estimation projects, previously in the Midwest states, Michigan and California and more recently in the Mississippi River Delta region of Arkansas (Graham 1993, Allen 1988). Recent research has included the use of Landsat, SPOT, and Indian Remote Sensing satellites' data in this region. Current work includes an intensive comparison study of one county in Arkansas using all of these sensors. Additional recent work has included a study of Native American reservations in Montana.

CONCLUSIONS

PEDITOR is a complete system for producing crop area estimates using satellite imagery. This system starts with scene registration and reformat, covers digitization of segment and county boundaries, and produces state and county estimates based on regression and/or

direct expansion of ground information. Although the focus is on estimation, many general features needed for any system processing satellite imagery are provided. Much of the PEDITOR system, including its files, is portable as has been demonstrated by porting it to quite different machines. Although PEDITOR was originally written to produce numeric estimates, PC-based display facilities have been added both to aid in certain procedures and to provide additional output for users.

REFERENCES

Allen, J.D. and Hanuschak, G.A. 1988. The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987, NASS Staff Report No. SRB-08-88, U.S. Department of Agriculture, Washington, DC.

Angelici, G.; Slye, R.; Ozga, M.; and Ritter, P. 1986. "PEDITOR--A Portable Image Processing System", Proceedings of the IGARSS '86 Symposium, pp. 265-269, Zurich, Switzerland.

Annoni, A.; Dicorato, F.; Stakenborg, J. 1991. Manual for the Use of Software for Agricultural Statistics Using Remotely Sensed Data, Commission of the European Communities, Institute for Remote Sensing Applications, Agriculture Project, Ispra (VA), Italy.

Battese, G.E.; Harter, R.M.; and Fuller, W.A. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data", Journal of the American Statistical Association, vol. 83, no. 401, pp 28-36.

Bellow, M.E. 1994. Application of Satellite Data to Crop Area Estimation at the County Level, NASS Research Report No. STB-94-02, U.S. Department of Agriculture, Washington, DC.

Bellow, M.E. and Graham, M.L. 1992. "Improved Crop Area Estimation in the Mississippi Delta Region Using Landsat TM Data", Technical Papers, 1992 ACSM-ASPRS Annual Convention, Washington, DC.

Bellow, M.E. and Ozga, M. 1991. "Evaluation of Clustering Techniques for Crop Area Estimation Using Remotely Sensed Data", 1991 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 466-471, Atlanta, GA.

Cotter, J.J. and Mazur, C. 1991. Automating the Development of Area Sampling Frames Using Digital Data Displayed on a Graphics Workstation, NASS Staff Report, U.S. Department of Agriculture, Washington, DC.

Cotter, J. and Nealon, J. 1987. Area Frame Design for Agricultural Surveys, NASS Staff Report, U.S. Department of Agriculture, Washington, DC.

Graham, M.L. 1993. "State Level Crop Area Estimation Using Satellite Data in a Regression Estimator," Survey Methods for Businesses, Farms, and Institutions, ICES Part

I, NASS Research Report Number SRB-93-10, U.S. Department of Agriculture, Washington, DC.

Hanuschak, G.A. and Craig, M.E. 1993. "Remote Sensing Program of the National Agricultural Statistics Service: From A Management Perspective," Survey Methods for Businesses, Farms, and Institutions, ICES Part II, NASS Research Report Number SRB-93-11 U.S. Department of Agriculture, Washington, DC.

Ozga, M. 1985. "USDA/SRS Software for Landsat MSS-Based Crop Acreage Estimation", Proceedings of the IGARSS '85 Symposium, pp. 762-772, Amherst, MA.

Ozga, M.; Mason, W.W.; Craig, M.E. 1992. "PEDITOR -- Current Status and Improvements," Proceedings of ASPRS/ACSM/RT'92 Convention, pp. 175-183, Washington, DC.

Swain, P.H. and Davis, S.M. 1978. Remote Sensing - The Quantitative Approach, McGraw-Hill, Inc., New York, NY.