PEDITOR--CURRENT STATUS AND IMPROVEMENTS

by
Martin Ozga
William Mason
Michael Craig
National Agricultural Statistics Service
United States Department of Agriculture
Washington, DC 20250

1992

ABSTRACT

PEDITOR is a system for processing satellite and ground gathered data with the primary goal of generating crop area estimates. PEDITOR provides functions for the entire procedure starting with scene registration and digitization of training areas through clustering and classification to the generation of regression estimates. Graphics facilities are included both to check the raw data and to display categorized results. For large-scale computations, a link to a CRAY supercomputer is available.

INTRODUCTION

PEDITOR is a system of computer programs designed to process digital satellite imagery with the goal of performing crop area estimation over large areas (Angelici 1986)(Ozga 1985). 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 outside deadlines for generation of estimates.

Although its primary goal is generation of area estimates, PEDITOR naturally includes many modules or functions generally found in 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. Certain modules, such as display modules designed for VGA graphics on IBM-compatible PCs, are coded for specific hardware. For very large computations, PEDITOR provides a link to a CRAY supercomputer. The programs on the CRAY are coded differently and are not part of the PEDITOR system, but they do read and write PEDITOR files.

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 Administrataion (NASA-Ames). The CASS system was designed

for the automation of land use mapping by displaying digital satellite data and digital map data on graphics workstations (Cotter 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.

STRUCTURE OF PEDITOR AND PORTABILITY

The PEDITOR system is divided into a large number of autonomous programs. Each program is run independently of the others and is called explicitly by the user. 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 VAX/VMS systems, IBM-compatible PCs under MS-DOS, IBM mainframes under MVS/TSO, the VACCELERATOR (an add-on faster processor for VAX sytstems), and, with some difficulty, SUN systems running variants of UNIX.

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, 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. This file has a header describing the windows in the file followed by the data. A one-window file may contain an entire scene. The data is stored in band interleaved by pixel 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, it is reformatted into this 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, and other sensors.

Certain PEDITOR modules, related to registration, digitization, and display, run on IBM-compatible PCs under MS-DOS only. The display portions of these codes require VGA graphics and certain of the display functions can make use of enhanced VGA modes if available.

INTERFACE TO THE USER

PEDITOR provides a simple alphanumeric interface to the user. No graphics interface has been attempted since settling on any one of the several available would seriously erode portability. 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 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. Certain PC-based programs make limited use of the mouse or the arrow keys when necessary to locate points or move segment outlines.

OVERVIEW OF PEDITOR PROCESSING FOR CROP AREA ESTIMATION

Crop area estimation is an important part of the overall NASS mission. The major indication used for the official estimates comes from a sample of over 16,000 areas of land, known as segments, selected throughout the United States. 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 on the ground enumerated data are generated using standard statistical techniques. Digital satellite imagery covering large areas is applied using a regression estimator approach to improve these estimates (Allen 1988).

The segments also provide an excellent source of training data for clustering since the fields have known locations and cover types. All pixels for cover types of interest are extracted for the segments and clustered, generating statistics files. After some editing, these statistics files are used to do a maximum likelihood classification of entire scenes or major portions thereof. With the strata boundaries also in digital form, an aggregation may be done by category and strata. 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 AND REGISTRATION

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. If needed, an extra copy is made to send to the CRAY facility.

After reformat, satellite scenes must be registered to a map base. The scene-to-map registration procedure consists of two steps and runs only on the PC. In the first step, a map is placed on the digitizing tablet and points which are likely to be found on the scene are digitized. These points are taken as the centers of windows. The second step displays these windows on a PC screen for the selection of an exact match between some point on the map (as digitized on the tablet) and the same point on the window. The point on the screen is selected using the mouse. 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. The primary scene is registered to a map. Then, a two-phase preliminary multitemporal registration is performed on the PC. In the first phase, sampled portions of areas from the primary and secondary scenes are displayed side by side on the PC screen. The mouse is used to point to corresponding features visible in both. These points become the centers of unsampled windows also displayed side by side on the PC screen. Matching features are marked to create control points. These control points are edited and least squares polynomials are generated to represent the preliminary registration.

In the second phase this preliminary registration is used to obtain a large number of block pairs, which 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 CRAY, but work has begun in implementing it in PEDITOR.

DIGITIZING SEGMENTS AND COUNTIES

Two approaches are available to obtain digital representation of ground boundaries. Segment enumeration and county area frame boundaries may be digitized into 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 of the segments or counties to a map base. County area frame files in the same format may also be obtained directly from the CASS system.

Segment boundaries may also be digitized 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 a digitizing tablet.

The ground data information about the size and crop for each field as collected by the enumerators is retrieved and stored in a ground truth file for each segment. The 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.

SEGMENT SHIFTING

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 PC-based 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 FILES TO GET TRAINING DATA

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 as well as possibly other conditions.

CLUSTERING

The 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. The ISODATA cluster method is used (Bellow 1991), but with an additional modification allowing cluster splitting as well as merging. The statistics files are combined, with some editing, into one statistics file to be used for the maximum likelihood classification. The editing is an attempt to remove categories that do not represent the crop in question. Such categories may still occur if a field is mixed or has small, non-contiguous areas which are not the labelled cover.

CLASSIFICATION AND AGGREGATION

Ordinary maximum likelihood classification is used. Prior probabilities, reflecting the likelihood of certain classes being in an area, may be used to modify the classification. Since some classifications are quite time consuming, particularly of multitemporal data with large numbers of categories, the CRAY is often 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 aggregations by strata

and category to be created. These aggregation files are an important input to estimation. 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.

ESTIMATION

Estimation is done in four phases: small scale, large scale, accumulation and county estimation. 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 one or more satellite images acquired on the same date or in the same pass. Regression parameters are calculated for each stratum 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. The final sample estimation 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.

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) 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.

DISPLAYS

Currently all displays of satellite imagery in PEDITOR are PC-based using the VGA graphics system. The registration programs and the video digitization labelling program use only standard VGA. The program for doing segment shifting, which also has many other display capabilities, can use enhanced or "super" VGA. Among these other display capabilites are the ability to show sampled areas representing large areas of scenes or even entire scenes. County boundaries 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. County level categorized displays in hard copy form will be provided to accompany numerical estimates in 1992.

INTEGRATION OF THE CRAY

Programs written for the CRAY supercomputer are not in the PEDITOR standard form even when, as with maximum likelihood classification, they perform the same function as PEDITOR programs. The CRAY programs are written mostly in FORTRAN, but with some CRAY assembly language, and have been coded to take advantage of the vectorization features of the CRAY. However, the CRAY programs do read and write PEDITOR files.

The CRAY currently used is at the Idaho National Enginerring Laboratory (INEL) of the Department of Energy in Idaho Falls, Idaho. Since NASS does not currently have a network connection to INEL, the MicroVAX acts as an RJE station to the CRAY. Using the HASP protocol, the jobs must pass through an IBM mainframe as well as a Control Data Cyber machine to reach the CRAY. Since the job control language to do this, as well as actually execute the job on the CRAY, is complex and long, a special PEDITOR program has been created to assemble CRAY jobs. The user only has to enter which jobs are to be performed and the files needed. Currently, only multitemporal file creation and maximumn likelihood classification with aggregation are done on the CRAY.

OPERATIONAL USE OF PEDITOR

PEDITOR has been used in various operational NASS estimation projects, mostly in the midwest states and California. The current operational project is in the Mississippi River Delta region. In 1991, state and county estimates for Arkansas and Mississippi were produced in a timely manner using unitemporal and multitemporal Landsat TM data (Bellow 1992), despite problems with late delivery of much of the data. The 1992 Delta Project will cover Arkansas, Mississippi and Louisiana. Currently PEDITOR is being used for several research projects as well as the 1992 Delta Project.

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 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, in Proceedings of the IGARSS '86 Symposium, Zurich, Switzerland, pp. 265-269.

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 ans Satellite Data in Journal of the American Statisticsl Association, vol. 83, no. 401, pp 28-36.

Bellow, M.E. and Graham, M.L., 1992, Improved Crop Area Estimation in the Mississippi Delta Region Using Landsat TM Data in 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
in 1991 Proceedings of the Section on Survey Research Methods, American Statistical Association, Atlanta, GA, pp 466-471.

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.

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