PB82-156969

Nationwide Forestry Applications Program Renewable Resources Inventory Project Multiresource Inventory Methods Pilot Test (Phase I): Volume III

Earth Satellite Corp.
Berkeley, CA

Prepared for

Forest Service
Houston, TX

30 Sep 80

U.S. Department of Commerce
National Technical Information Service
NTIS

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. NFAP-211	2.	3. Recipient's Accession No. PB82 156969
4. Title and Subtitle MULTIRESOURCE INVENTORY METHODS PILOT TEST (Phase I): Evaluation of Multiresource Analysis & Information System (MAIS) Processing Components, Kershaw County, SC Feasibility Test, Volume III			5. Report Date Published Date September 30, 1980
			6. 032043
7. Author(s) Earth Satellite Corporation			8. Performing Organization Rept. No.
9. Performing Organization Name and Address Earth Satellite Corporation Forestry Division 2150 Shattuck Avenue Berkeley, CA 94704			10. Project/Task/Work Unit No. AgRISTARS, RRI Project
			11. Contract(C) or Grant(G) No. (C) 53-3187-9-45 (G)
12. Sponsoring Organization Name and Address Nationwide Forestry Applications Program USDA, Forest Service 1050 Bay Area Blvd. Houston, TX 77058			13. Type of Report & Period Covered Interim
			14.

15. Supplementary Notes

COLOR ILLUSTRATIONS REPRODUCED IN BLACK AND WHITE

16. Abstract (Limit: 200 words)

The Forest Service Multiresource Inventory Methods Pilot Test is a planned three-phase effort, intended as a demonstration of Landsat satellite technology to supplement present methods of conducting recurrent inventories over large areas. The project creates an opportunity to apply earth resource satellite technology for testing and evaluating its potential for use by resource managers and planners. This Phase I, Volume III report describes the actual testing of components of a prototype Multiresource Analysis and Information System (MAIS) and their application to a practical test for the particular geographic area of Kershaw County, SC. The building of a loosely-linked "prototype system" is discussed in Section 1. Section 2 describes the development of estimators for use in the pilot test, and points out some of the trade-offs for different types of estimators and sampling methods. The data structure that is implied for the system is summarized. Computer programs that were generated to apply the theory of class area proportion correlations, estimates of variances, and tests of statistical significance are discussed. This is followed by a discussion of the test itself. Following a discussion of the use of Landsat data, classification procedures, and proportion extraction methods, the main results of the test analysis are presented. Section 3 summarizes the results of all the tests performed.

17. Document Analysis a. Descriptors

- inventory
- geographic information system
- Landsat

b. Identifiers/Open-Ended Terms

- multiresource inventory methods
- Multiresource Analysis & Information System
- Landsat classification

c. COSATI Field/Group Forestry (Remote Sensing)

18. Availability Statement: Release Unlimited	19. Security Class (This Report) Unclassified	21. No. of Pages
	20. Security Class (This Page) Unclassified	22. Price

(See ANSI-Z39.18) See Instructions on Reverse OPTIONAL FORM 272 (4-77) (Formerly NTIS-35) Department of Commerce

PB82-156969

AgRISTARS

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

VOL III

September 30, 1980

NATIONWIDE FORESTRY APPLICATIONS PROGRAM
Renewable Resources Inventory Project

Multiresource Inventory Methods Pilot Test (Phase I):

EVALUATION OF MULTIRESOURCE ANALYSIS AND INFORMATION SYSTEM (MAIS) PROCESSING COMPONENTS, KERSHAW COUNTY SOUTH CAROLINA FEASIBILITY TEST

EARTH SATELLITE CORPORATION
Forestry Division, Berkeley, California

USDA FOREST SERVICE
Houston, Texas

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

This document is provided in response to the requirement stated in Division 240 of Section 4.0: Statement of Work, Request for Proposal, of the Multiresource Inventory Methods and Pilot Test, Phase I, South Carolina Design and Implementation Planning Contract.

It supplements an earlier version of the component evaluation report. The preliminary version (Appendix A) contains a complete list of upper and lower level Geographic Information System (GIS) components.

The current version is of a different character; it contains a description of another step in the evolution of the MAIS system. The reason for the change is stated in the following.

1.1 Approach

The old adage: "The whole is more than the sum of the parts", is especially applicable to the evaluation of the prospective MAIS system. The assurance that each component is a state of the art component in good working condition does not imply that the same is true for the overall system. Conversely, a smoothly functioning system must necessarily consist of good working components.

An evaluation of single components is therefore not sufficient. One must also be concerned with component interactions: their method of use and the total effect on the entire system.

In the initial phases of the MAIS project, it was thought that a system could be assembled from existing components. In the course of the work, however, it became apparent that along with well-entrenched methods and programs, some new concepts in software would be required to make the system function. As a result, the main concern with the evaluation of the proposed system became focused on the simple question: Will the basic concept work?

To answer this question, the Phase IA tests described in this document were undertaken. It became clear that a positive answer was required before one could proceed with a more thorough, component-by-component evaluation to secure an optimum design. Also, it seemed wise to obtain this answer before proceeding to the related question: How well do the concepts work, as planned for Phase II?

This document, therefore, emphasizes actual testing of a "prototype" system; it does not consider the background of various components with regard to such matters as: state of the art maturity, availability, working status, and R & D requirements. The testing approach confronts one with the real problems which arise when all components must be exercised to provide a meaningful systems result, and thus concentrates the evaluation effort on those components which are currently the weakest links in the design.

The MAIS system as proposed in the MAIS concept development document basically consists of three major subsystems: the upper level GIS, the lower level GIS, and the "linear model", which relates data in the two GIS systems. In the course of the Phase IA work, it was discovered that the "linear model" was the most underrated part of these three subsystems. In this model the data from the GIS systems flow together and are combined to yield the desired estimates. It seemed that almost the entire "method of use" of the two GIS systems with their established techniques took on critical importance in this component. New methods and techniques not previously applied in this context were proposed to bring this about.

As a consequence, much of the effort for the Phase IA test was expended on this subsystem, and much of this report is devoted to the further definition of the methods and techniques formulated for its programs. The term "linear model" was replaced with "estimation subsystem" to more accurately reflect its status and complexities in the overall MAIS system.

1.2 Objectives

The main objective was to process a limited set of data through a loosely assembled "prototype system". The first concern within this overall objective was to secure a proper data flow; that is, when data are entered at one end, results will emerge from the other.

The second concern was to assess the data resulting from the data flow to ascertain that meaningful results can be obtained. Mindful of the GIGO concept, this concern seems perhaps more important than the first. However, a functioning system must necessarily exist before good results can be produced.

An important aspect of the Phase IA test was therefore to assemble the prototype system from existing components and to construct preliminary versions of those elements which had to be newly created. It is hoped that the same components can be used by the Phase II contractor to process data for the 16-county test area in South Carolina.

1.3 Scope

The scope of the Phase IA test was necessarily limited by several factors. As the main objective of this test was to assure the workability of the proposed process, the question of the quality of the estimates could not be fully answered. The time frame did not allow for a full research effort, and the area of interest may be too small to obtain an exhaustive answer. Therefore, a great deal of the Phase II effort also must be concentrated in this area.

Other limiting factors also played a role during the Phase IA test. Several resource parameters could not be completely processed through the system because of a lack of time for basic data input. The time factor was also a constraint in the analysis of those parameters for which a final data set was obtained. The overall philosophy for the Phase IA effort was to do a limited number of tasks for a limited area, but to try to complete these tasks as well as possible.

1.4 Prototype System

The meaning of the word prototype as used in this report must not be misunderstood; an integrated software package running on a single machine was not created. Rather, several methods, packages, machines and people at various locations were involved. Some components were tightly integrated packages; others were separate existing programs. Several programs were newly created for the Phase IA effort.

The upper level GIS LANDSAT analysis was performed in two separate locations. The LANDSAT clustering analysis was performed in cooperation with the remote sensing group of the Space Sciences Laboratory at Berkeley on an image processing system developed around a Data General Nova 840 computer. All other upper level GIS processing was performed on EarthSat's PRIME 450, located in the Washington, D.C., office. Lower level GIS processing was accomplished with the LANDPAK system on a PRIME 550, located at the premises of EarthSat's Berkeley office. The same system was used to integrate all data and to develop and run the estimation subsystem.

Although it would have been possible to perform all processing in-house, to ascertain a state of the art effort the assistance of experienced personnel at the Space Sciences Laboratory was procured for the LANDSAT classification.

1.5 Document Organization

The remainder of this document is divided into two main parts. The first, Section 2.1, contains the concepts and theory for the estimation subsystem as well as a description of the programs developed for it. Some of the programs can be incorporated into a permanent subsystem; most are only temporary, written for specific tasks in the Phase IA tests.

The second part, Section 2.2, contains the report of the Phase IA testing effort. The test area is described, and the input and data preparation techniques are outlined. The results and the analysis of the results for the parameters evaluated are presented.

2.0 EVALUATION

The propose. MAIS system is based on several new and innovative concepts. Before the event of current computer technology, resource information systems mostly consisted of maps, aerial photos, and data files, with severe physical limitations on the amount of numerical data that could be manipulated to provide answers to management questions. The overall emphasis was on collecting new resource data for specific problems. The resource base itself was used as a data base with limited access to resource parameters through sampling methods.

With the continuing development of computer technology, the capability to handle large statistical data files increased. Sampling and statistical techniques became more complex, and in the last decade the revolution in computer graphics has given rise to geographic information systems (GIS) which handle maps as well as descriptor data. With these systems it has become possible to create an accurate and comprehensive model of a resource base with which management problems can be evaluated, and actions can be simulated. It may, therefore, seem that the pendulum is swinging the other way and that complete enumeration of the resource model is now reasonable in many cases. In these cases, sampling methods may have become obsolete.

In the MAIS design it has been recognized that for large areas under diverse ownership, for which many different kinds of questions relating to separate disciplines must be answered, the GIS technology as well as statistical and sampling methods must play a significant role. For even if complete enumerations were possible, they might be more efficiently applied in large sample units by means of which estimates of required accuracies can be provided over large areas.

Two basic approaches are possible when combining the sampling concept with the use of GIS systems. One might model the entire resource base and sample within the GIS system to obtain answers, or one might enter only selected portions of the resource base into the GIS system and fully enumerate these to obtain specific estimates. A combination of these two methods may also be used to take advantage of the best features of both. This is the concept favored in the MAIS design.

Two GIS systems are involved. In the upper level system, the resource base is represented in its entirety in the form of a LANDSAT classification image, possibly combined with other auxiliary data of the same resolution. This model can be cheaply enumerated in different ways, this being its main advantage. The related advantage is the timeliness of the information in the LANDSAT images. Disadvantages are poor resolution and limited information with a sometimes unknown meaning.

To compensate, the lower level GIS stores high resolution detail, but only for selected sample areas. This information is also kept up-to-date, but it is mostly based on aerial photographs and ground data; hence, the update frequency is not as high as for the LANDSAT data.

2.1 Estimation Subsystem

Combining GIS technology with sampling and statistical techniques presents a unique challenge. In conventional sampling methods, randomness is mostly assured by selecting sample units in random locations. The GIS technology presumes, however, that the selection units stored in the GIS are permanent; additional units may be added, but the power of the system lies in its facility to accurately keep track of a given piece of land. The same problem has, of course, surfaced in the past in the transition from one-time inventories to the CFI approach. To reconcile these opposed concepts, some compromises need to be made, and it is worthwhile to consider what the tradeoffs are for the available options.

The first option is to shift the emphasis on the random samples in the lower level GIS to the upper level, where complete enumeration can easily be made; to tie the upper level to the lower level by means of a model; and to rely on the error distribution in this model for appropriate random effects.

A disadvantage of this approach is that a priori assumptions on the error distribution in the model must hold for it to provide unbiased estimates. The traditional approach in sampling methods has been to avoid estimators for which such a priori assumptions are required, and techniques have been developed to obtain robust, distribution-free estimates.

The second option is to use these robust estimators and forego complete location randomization by using the same sample (with possible added units as stored in the GIS subsystem on successive occasions). This approach is biased in the long run, because the population is limited to the areas represented in the GIS. This may be somewhat compensated for by the use of the upper level data, but mostly these act as an efficiency booster in the overall design.

The tradeoffs will be illustrated in more detail in the following section. For the Phase IA test, only the first option has been explored in actual test computations. The second needs further theoretical development. Hopefully, this can be accomplished during Phase II.

The remainder of the current section has been divided into three parts. The general structure of the estimation subsystem is presented first. This is followed by a more detailed discussion of the statistical formulation for the MAIS design. The section is concluded with a description of the estimation subsystem components.

2.1.1 Structure

When examining the MAIS design from a sampling point of view for the purpose of classifying the overall approach, it is clear that a multistage "sampling" design is used, with "multi" equal to three. The stages in a multistage design normally refer to the sampling steps (Murthy, 1967), with which the target population is approached. In the multistage concept as introduced to forest inventory by Langley et al (1969), each sampling step is also associated with a type of imagery of a certain scale except for the final stage, which refers to the ground level. In the MAIS design, this association is more general, as any kind of useful information stored in the GIS systems may be used at a given stage consistent with its spatial resolution. Thus, for example, at the county level LANDSAT images alone may be used, or they may be combined with NCIC data; or, one may prefer to use a general soils map instead.

2.1.1.1 Data Structure

Each stage in the multistage view of the MAIS corresponds to a set of spatial data with matching attribute data of a given resolution. The nature of the data at each of these steps is as shown in column 2 of Table 1.

TABLE 1

DATA OVERVIEW

Stage	Type of Data	Spatial Type	GIS
1 (Primary) Sample Unit PSU	LANDSAT LANDSAT Classification NCIC DATA Digitized Map Data of Comparable Scale	Cell	Upper Level
2 (Secondary) Sample Unit SSU	Aerial Photo Classifications Digitized Map Data of Comparable Scale	Polygon Line, Point	Lower Level
3 (Ultimate) Sample Unit USU	Field Data	Point	Lower Level

Throughout the MAIS design concepts document, references are made to the upper and lower level GIS systems, connected through the "linear" model". In this document, the linear model is incorporated into a more appropriate estimation subsystem. More than one linear model may be applied to obtain one set of estimates. From this perspective, the term "upper-lower level", as used in the design concepts document, may be somewhat misleading; upper level applies to the first stage, whereas lower level applies to both second and third stages as indicated in column 4 of Table 1. The term "level", as used in the design concepts document, is related to storage and resolution requirements rather than sampling steps.

The basic resource units as stored in the GIS system have both a spatial and an attribute component. Each cell has a class, and each polygon line has an associate descriptive record. The effect of the estimation subsystem is that attribute data are divorced from their spatial equivalents to be summarized into estimates which then apply to much larger spatial units. At the first stage, cell classes are summarized into class proportions by PSU; at the second stage, polygonal areas are converted into class proportions by SSU; and at the third stage, fixed plot (USU) data are converted into averages by class. The summarized data are then related together into county estimates. The components to prepare the spatial data at each stage for use in the estimation subsystems are described in Section 2.1.3.

2.1.1.2 Processing Structure

Once the data at each stage have been properly summarized, the final estimates are derived in two steps. The first step ties the first two stages together; the second step combines the output from the first with the third stage data to produce the final results. The entire process is schematically shown in Figure 1.

DATA PREPARATION			
Stage 1	Class Proportions	STEP 1 Area Estimates	STEP 2 Final Estimates
Stage 2	Class Proportions		
Stage 3	Class Averages		

Figure 1. Processing Steps

Step 1 revolves entirely around the estimation of class or stratum areas. Step 2 integrates the area estimates with parameter estimates, by class, to obtain parameter estimates for the larger units, such as the county. The proposed "sampling" technique for Step 1 is a variation of regression sampling. The process for Step 2 can be largely characterized by stratified sampling. As in double sampling for stratification, the stratum or class areas are not fixed but are themselves random variables. This is one of the unique aspects of the MAIS design.

The reason for first arriving at a set of area estimates independently from the resource parameters estimates is the following: An alternative approach could have been taken in which various parameter estimates would have been propagated through all stages with independent linear models for continuous variables; it would have been difficult, however, to maintain known relationships between the parameters in this kind of propagation. With the current approach, relationships existing at the ultimate stage are not changed, as one only multiplies through by area.

A summary of the estimation subsystem components is presented in Table 2.

TABLE 2

ESTIMATION COMPONENTS

STAGE	DATA PREPARATION	STEP	ESTIMATION
1	LANDSAT Proportion Extraction	1	Area Estimation Component
2	Aerial Photo Proportion Extraction	2	Summary Component
3	Field Statistics Computation		

Each of these components will be discussed in Section 2.1.3. However, first it will be necessary to examine the statistical rationale for Steps 1 and 2.

2.1.2 Statistical Formulation

The Step 1 area estimation component is conceptually one of the most important components of the MAIS system. In it, a novel approach has been taken which, if proven successful, could set a new standard for incorporating LANDSAT into multistage designs.

Traditionally, the aim of most LANDSAT classifications has been to produce a class map for which there is an optimal one-to-one correspondence between its classes and resource categories defined in some other, more direct way. Results are usually expressed in contingency tables which, when both marginal classifications are identical, are referred to as confusion matrices (Colwell, 1979; Hildebrandt, 1979); or when this is not so, are referred to as co-occurrence matrices (Isaacson, et al, 1979).

It is interesting to note that in recent years a new kind of emphasis has been placed on this type of matrix. It seems to result from the realization that a perfect diagonal confusion matrix is not attainable, at least not for wildland resource classifications, and that therefore the matrix itself can be a tool to interpret the classification interpretation. Consequently, confusion matrices have been more carefully constructed using sampling techniques (Mayer, 1979; Sader, 1979; Todd, et al, 1980), and different kinds of hypothesis tests have been applied (Isaccson, et al, 1979; Sader, 1979; Todd, et al, 1980).

The MAIS approach goes along with this development but adds a new point of view which, in effect, relaxes the one-to-one correspondence concept to the extent that any unsupervised classification may be of use.

The device used is a transition "probability matrix" (Telser, 1963), or a "projection matrix" (Pielou, 1969), which translates the stage one proportions into stage two proportions as follows:

$$a = e'P \qquad (0)$$

where a and e are K and L element class proportion vectors, and P is a L x K matrix of "transition probabilities." P can be estimated from the sample proportions of a set of matching PSU-SSU's with equal areas.

This estimation is not without problems, some with published solutions only since 1976. The best known example in economics literature is an application by Telser (1963), who estimated transition probabilities for smokers switching between the brands of Lucky Strike, Chesterfield, and Camel from year to year (1923).

Being able to estimate P and its covariance matrix takes the MAIS design a step beyond the usual application of the confusion or co-occurrence matrix. Rather than subjecting it to interpretation, it is used to directly translate proportions from one stage to the other. For this reason P will be referred to as a "class transformation matrix" in the remainder of this document. In the following section we will assume that $\hat{P}$ and $\hat{\sum}\hat{P}$ the estimators for P and its covariance matrix have been computed. How this is accomplished is the subject of Section 2.1.3.1.1.

2.1.2.1 Step 1 Estimators and Variances

The formulation given in the following is all based on the assumption that sampling is with replacement.

Suppose that the PSU-SSU combinations have been located randomly and that there are n such combinations taken from an infinite population situated in the county. It is instructive to first consider the sampling properties of the SSU proportions as an independent set. The SSU can be considered a cluster of fixed size, and hence Cochran's theory for estimation of proportions in cluster sampling can be applied (Cochran, 1963).

If a_{ij} is the proportion of class i in SSU j, then the estimator for the population proportion is as follows:

$$a_i = \frac{1}{n} \sum_{j=1}^{n} a_{ij} \qquad (1)$$

An unbiased sample estimate of the variance is:

$$\hat{v}(\hat{a}_i) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (a_{ij}-\hat{a}_i)^2 \qquad (2)$$

Since there are K classes in the SSU, and the class proportions are not independent, the covariance of $\hat{a}_i$, $\hat{a}_k$ must also be considered:

$$\hat{\text{cov}}(\hat{a}_i,\hat{a}_k) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (a_{ij} - \hat{a}_i)(a_{kj} - \hat{a}_k) \qquad (3)$$

Let the vector of the above K class a_i estimates of the county be denoted by $\hat{a}$, then using (2) and (3), the estimated covariance matrix of $\hat{a}$, $\hat{\Sigma}\hat{a}$ can be computed.

Likewise, from the PSU's, the vector of L primary class proportions $\hat{e}$ can be computed, as well as its estimated covariance matrix $\hat{\Sigma}\hat{e}$.

The following estimators for the secondary county proportions are now considered (estimators B and E in the following represent the two options alluded to in Section 2.1).

A. <u>Average SSU Class Proportions</u>

The estimator is:

$$\hat{a}_{av} = \hat{a} \qquad (4)$$

The estimated covariance is defined as described above:

$$\hat{\Sigma}\hat{a}_{av} = \hat{\Sigma}\hat{a} \qquad (5)$$

If the PSU-SSU's are randomly located, the estimator is unbiased. The primary stage is not considered in this estimator.

B. Regression Prediction

The estimator is:

$$\hat{a}_{pr} = e'\hat{P} \qquad (6)$$

The estimated covariance is:

$$\hat{\Sigma}\hat{a}_{pr} = E\hat{\Sigma}\hat{P}E' \qquad (7)$$

Here e is the vector of L county proportions (enumeration). $\hat{P}$ is the estimated class transformation matrix, $\hat{\Sigma}\hat{P}$ is the estimated covariance matrix of $\hat{P}$.

P can be rearranged as a vector with M = L x K elements, and hence $\hat{P}$ is an M x M matrix. To be compatible with this matrix, e must also be rearranged into a diagonally structured matrix E, as follows:

$$E : \begin{bmatrix} e_1 & \cdots & o & e_2 & \cdots & o & \cdots & e_L & \cdots & o \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ o & \cdots & e_1 & o & \cdots & e_2 & \cdots & o & \cdots & e_L \end{bmatrix} \qquad (8)$$

K x M

The estimator is unbiased if $\hat{P}$ is an unbiased estimator of P. This is only the case if the traditional regression assumption holds; namely, the errors are uncorrelated and have constant variance over the range of the relationship.

The evaluation of $\hat{a}_{pr}$ is the subject of the Phase IA tests. Its advantage for use in MAIS is that the random distribution of the sample is not that important as long as a full range of conditions is present. The random effects which occur around the linear relationship are what count. The disadvantage is that the assumption on the error distribution in the model must hold.

It is likely that the increase in efficiency (lower variance) with this estimator is in the same order of magnitude as for the traditional regression sampling estimator, namely ρ^2 x 100% where ρ is the correlation coefficient (Zarcovic, 1964).

C. Regression Sampling

The estimator is:

$$\hat{a}_{rs} = \hat{a} + (e - \hat{e})\hat{P} \qquad (9)$$

The estimated covariance is $\hat{\Sigma}\hat{a}_{rs}$ (defined below). It is conjectured that the elements c_{ij} of the estimated covariance matrix can be computed as follows:

$$c_{ij} = \frac{1}{nv} \sum_{k=1}^{n} \left[(a_{ik} - \hat{a}_i) - \left[(e_k - \hat{e})'\hat{P}\right]_i\right] \times \left[(a_{jk} - \hat{a}_j) - \left[(e_k - \hat{e})'\hat{P}\right]_j\right] \qquad (10)$$

where $\left[(\hat{e}_k - \hat{e})\hat{P}\right]_i$ denotes the i^{th} column of the row vector resulting from the multiplication in the brackets and v is the degrees of freedom used in the regression error calculation.

A different approach to the estimated covariance is to note that (9) contains two vectors and one matrix of random variables. The partial derivatives of $\hat{a}_{rs}$ with respect to these variables can be obtained, and the "delta method" of variance propagation can be applied, given that the combined covariance matrix is available. Along the diagonal this matrix is composed of the submatrices $\hat{\sum}\hat{a}$, $\hat{\sum}\hat{e}$, $\hat{\sum}\hat{P}$. As generally the covariance of two random variables tends to zero when the sample size increases, the off diagonal submatrices could possibly be neglected for a relatively large sample.

The estimator $\hat{a}_{rs}$ is the multivariate equivalent of the traditional regression sampling estimator. All remarks about its properties are based on the description of this estimator as presented by Cochran (1963) and Murthy (1967).

The estimator is unbiased if $\hat{P}$ is a preassigned matrix. However, if $\hat{P}$ is estimated from the sample, the estimator is biased; but the bias is small and becomes smaller with increasing sample size.

The advantage of $\hat{a}_{rs}$ over $\hat{a}_{pr}$ is that it does not depend on assumptions about error distributions in the regression model. The disadvantage is that the sample must be randomly selected. A complication of this estimator when dealing with proportions is discussed in Section 2.1.3.1.

The estimator was not tested in the Phase IA tests. It may be an attractive alternative to the $\hat{a}_{pr}$ estimator because of its insensitivity to the error distribution in the regression model. It is therefore hoped that the opportunity will be present in Phase II to further develop the covariance aspect and to test and compare this estimator with the alternatives.

2.1.2.2 Step 2 Estimators and Variances

The Step 2 process integrates the stratum or secondary class area estimates with observations made in the field. Two types of observations can be made: discrete and continuous. From the USU point of view, a discrete variable is a class designation; from the class point of view, a USU is either in or out, and hence it can be considered a binary variable.

First, discrete variables will be considered.

2.1.2.2.1 Discrete Variables

One is confronted with the same problem present in Step 1, namely how to translate from one set of class designations to another. Again, this problem can be solved with a linear class transformation (Section 2.1.2), but $\hat{P}_2$ and $\hat{\sum}\hat{P}_2$ for Step 2 must be obtained differently.

$\hat{P}_2$ can be obtained by constructing a table with the secondary class definitions in the left margin and the ultimate definitions at the top. Each element in the table is the proportion of ground plots in the ultimate class j identified as secondary class i, or $p_{ij} = n_{ij}/n_i$. (11)

If there are K classes at the second stage and I classes at the ultimate stage, then P is a K x I matrix; and its corresponding covariance matrix is of dimension (K x I) x (K x I). An estimate of this matrix can be obtained from the ground sample by computing the diagonal elements as

$$c^{on} = \frac{1}{n_i} \frac{n_{ij}(n_i - n_{ij})}{n_i^2} \qquad (12)$$

using the variance formula for binomial distribution, and the off diagonal elements as

$$c^{off} = \frac{1}{(n_i+n_k)} \frac{n_{ij}n_{k\ell}(n_i n_k - n_{ij}n_{k\ell})}{n_i^2 n_k^2} \tag{13}$$

In the first step, $\hat{P}$ and $\hat{\Sigma}\hat{P}$ were obtained from the primary and secondary class proportions. The ground sample provides $\hat{P}_2$ and $\hat{\Sigma}\hat{P}_2$. An estimate for the proportions of the ultimate (ground) classes is now obtained as follows:

$$\hat{g} = \hat{a}'P_2 \tag{14}$$

Where $\hat{g}$ is a vector of ultimate class percentage estimates: $\hat{a}$ can be obtained either as $\hat{a}_{rs}$ or $\hat{a}_{pr}$. In the latter case (14) can be rewritten as:

$$\hat{g} = \hat{e}'\hat{P}_1\hat{P}_2 \tag{15}$$

The covariance matrix of $\hat{g}$ again can be estimated with the delta method. Assuming that $\hat{a}$ and $\hat{P}_2$ are independent, one can derive:

$$\hat{\Sigma}\hat{g} = \hat{A}\,\hat{\Sigma}\hat{P}_2 A' + \hat{P}_2'\,\hat{\Sigma}\hat{a}\,\hat{P}_2 \tag{16}$$

Where $\hat{A}$ is a diagonally banded version of $\hat{a}$. A subject of further investigation must be whether the independent assumption holds. It is also possible that the estimated covariance matrix of $\hat{P}_2$ can be obtained in a better way.

2.1.2.2.2 Continuous Variables

For continuous variables, two important types may be distinguished: by class (for the county) and by county. In the by class category, we may obtain estimates by primary, secondary or ultimate class definitions. A further distinction can be made as to whether the estimate is a "per acre" value or a total value. Figure 3 presents a hierarchical organization of these types of estimates.

In the preceding section we have used e, a and g to denote primary, secondary and ultimate stage proportions and proportion vectors. In the following we will denote per acre estimates in each of these categories by $\bar{e}$, $\bar{a}$, $\bar{g}$, and total estimates by $\widetilde{e}$, $\widetilde{a}$, $\widetilde{g}$. Not all of these kinds of estimates are useful. The most important ones are discussed in the following.

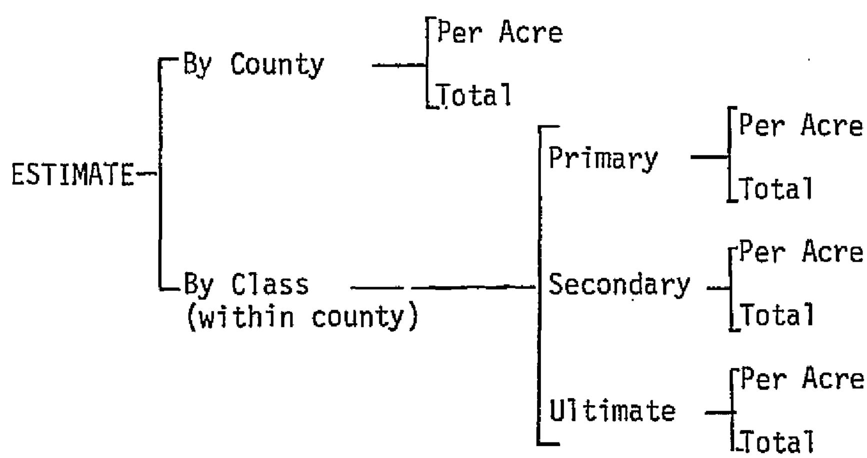


Figure 3. Types of Estimates

A. Ultimate Class, Per Acre

The estimate is simply the mean of the desired variable as observed on the ground, computed over all plots in the ultimate class designation. Let this mean be denoted by y_i (class i). And let $s^2_{\bar{y}_i}$ be its estimated variance. Then,

$$\hat{g}_i = \bar{y}_i \quad \text{and} \quad s^2_{\hat{g}} = s^2_{\bar{y}_i} \tag{17}$$

B. Ultimate Class, Total

The area for Class i is estimated as $\hat{g}_i$ (Section 2.1.2.2.1), and hence the estimate for the total is:

$$\hat{\hat{g}}_i = \hat{g}_i \, \bar{y}_i \, C \tag{18}$$

Where C is the county area. An estimate for its variance is:

$$s^2_{\hat{\hat{g}}_i} = C^2(\bar{y}_i^2 \; s^2_{\hat{g}_i} + \hat{g}_i^2 \; s^2_{\bar{y}_i}) \tag{19}$$

where $s^2_{\hat{g}_i}$ is the appropriate diagonal element of $\hat{\Sigma}_{\hat{g}}$ (Section 2.1.2.2.1).

C. Secondary Class, Per Acre

The estimator is similar to the one described under A., but the ground plots are sorted according to secondary class designations, and the means are computed for these classes.

D. Secondary Class, Total

Similar to B.: replace g with a.

E. Primary Class, Per Acre

A different approach has to be used for this kind of estimate. The ground plots cannot be directly related to the cells at the primary stage because of misregistration of primary maps and images with respect to the ground coordinate system.

Instead, the relation $e = a'\hat{P}$ can be used in reverse, namely:

$$e' = \hat{P}^{-1}a \qquad (20)$$

where P^{-1} is the generalized inverse of P. Using this relation on the vector of secondary class per acre estimates (this vector can be considered a constant times a set of proportions, with the constant automatically carried through the matrix multiplication), one obtains a vector of per acre estimate as follows:

$$\bar{e}' = \hat{P}^{-1}\bar{a} \qquad (21)$$

F. Primary Class, Total

The same transformation method (24) can also be applied to the vector of total estimates by class.

The application for the primary class estimates is that they be used to construct a precise legend for a primary stage classification map. This may be a very important use for this type of estimate.

G. Entire County, Per Acre

County per acre and total estimates can be based on any of the three stratifications: primary, secondary or ultimate.

The per acre estimate of the ultimate classification by class was $\hat{\bar{g}}_i$. The per acre estimate for the entire county is simply a weighted average over all classes of these estimates:

$$\hat{\bar{g}} = \sum_{i=1}^{I} \hat{g}_i \ \hat{\bar{g}}_i \tag{22}$$

or by (17):

$$\hat{\bar{g}} = \sum_{i=1}^{I} \hat{g}_i \ \bar{y}_i \tag{23}$$

where I is the number of ultimate classes, and the $\bar{y}_i$'s are the average parameter estimates for the ultimate classification.

The variance can be estimated as:

$$s^2_{\hat{\bar{g}}} = g \hat{\Sigma}_{\bar{y}} g + \bar{y} \hat{\Sigma}_{\hat{g}} \bar{y}' \tag{24}$$

where $\Sigma_{\bar{y}}$ is a diagonal matrix with $\hat{s}^2_{\bar{y}}$ on the diagonal and zero's elsewhere.

H. Entire County, Total

The estimator of the previous paragraph is simply multiplied by C, the total county area. Its estimated variance must be multiplied by C^2.

2.1.3 Components

Proposed components for the estimation subsystem are shown in Table 2. In the following subsections, each of these components will be discussed. The area estimation component ("linear model" of the design concepts document) will be treated somewhat in depth, since much of the Phase IA effort was spent on developing a prototype. The function of the other proposed components was "simulated" by calculating needed intermediate results with "throwaway" programs. Hence, each of these components will be discussed only briefly, mostly with respect to its role in a future MAIS.

2.1.3.1 Area Estimation Component

Three approaches are possible to obtain an estimate of the co-occurrence or class transformation matrix. First, one may spatially overlay the two classification maps and compute the area of each of the combined categories. The proportions of these area in terms of total area for the classes of one of the classifications are the elements of the transformation matrix.

If a complete overlay is not possible, one can resort to obtaining a sample of points and inspect each point for classification. This is the method taken in Step 2 for deriving ultimate class estimates. It is also commonly described in the literature (Isaacson, et al, 1979; Sader, 1979; Todd, 1980).

The third approach is to use a regression model to estimate the transformation matrix. This is basically possible if there is a sufficient number of sample units with proportion vectors a and e. However, working with proportions presents some basic difficulties, the most notable one resulting from the requirement that the predicted proportions must add to one. Another difficulty is that a large number of coefficients must be estimated.

The complete overlay of two complete classifications, if not prohibitive, would be extremely time-consuming and costly, especially in the polygonal mode. An overlay using a cell approach would be more reasonable, but then a polygon to cell conversion would be necessary. The sample approach also requires an overlay if automated and, if used with individual pixels, must suffer from registration problems.

The regression method seems to offer several advantages: the entire data set is used; a linear model is developed to which a large body of statistical "know-how" applies; and the developed matrix can be used directly to translate from one classification to another. For these reasons the regression approach was selected. How the associated problems were solved in the Phase IA development is the subject of the following section.

2.1.3.1.1 <u>Theory</u>

A. <u>Conditions</u>

The linear model $a = e'P$ can be set down in terms of its individual elements as:

$$\underset{1 \times L}{\begin{bmatrix} e_1 & e_2 & \cdots & e_L \end{bmatrix}} \underset{L \times K}{\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & & & \vdots \\ p_{L1} & p_{L2} & \cdots & p_{LK} \end{bmatrix}} = \underset{K \times 1}{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}}' \qquad (25)$$

Since e is a vector of proportions, the following conditions hold:

$$\sum_{i=1}^{L} e_i = 0 \qquad 0 \leq e_i \leq 1 \qquad i = 1, \ldots\ldots, L \quad (26)$$

When the model is used in the prediction mode, the following conditions must hold for the output vector:

$$\sum_{i=1}^{K} a_i = 1 \qquad 0 \leq a_i \leq 1 \qquad i = 1, \ldots\ldots, K \quad (27)$$

The problem to be solved is how to constrain the elements of P such that (27) will hold true, not only for members of the sample but for <u>any</u> prediction.

First it can be seen that the inequality in (27) can be reduced to: $a_i \geq 0$ for $i = 1, \ldots\ldots, K$.

Now each element of a is computed as:

$$a_j = \sum_{i=1}^{L} e_i \; P_{ij} \qquad (28)$$

The following condition must therefore be satisfied:

$$\sum_{i=1}^{L} e_i \; P_{ij} \geq 0 \qquad (29)$$

But, $e_i \geq 0$ for $i = 1, \ldots\ldots, L$.

It is, therefore, necessary and sufficient that $P_{ij} \geq 0$ for all i and j, for (29) to hold. The first constraint on the elements of P is thus a nonnegativity constraint.

The second requirement: $\sum_{i=1}^{K} a_i = 1$ can be translated into a constraint on the elements of P, as follows:

$$\sum_{i=1}^{K} a_i = \sum_{j=1}^{K} \sum_{i=1}^{L} e_i p_{ij} = \sum_{i=1}^{L} \sum_{j=1}^{K} e_i p_{ij} = \sum_{i=1}^{L} e_i \sum_{j=1}^{K} p_{ij} \qquad (30)$$

Because $\sum_{i=1}^{L} e_1 = 1$, a necessary and sufficient condition for the a_i's to add to 1 is that $\sum_{i=1}^{K} p_{ij} = 1$.

Summarizing, for the estimator $\hat{a}_{pr}$ (Section 2.1.2.1), the following constraints must be enforced in the regression:

$$\sum_{i=1}^{K} p_{ij} = 1 \text{ and } p_{ij} \geq 0 \qquad (31)$$

The same conditions are also reported by Judge and Takayama (1966) in this discussion of the cigarette example. Goodman (1953) demonstrated that the condition $\sum p_{ij} = 1$ is automatically satisfied in ordinary least squares.

Without special precautions, negative estimates of p_{ij} will occur however.

A different estimator may need different constraints. To examine the requirements for $\hat{a}_{rs}$ (Section 2.1.2.1) one may first observe that the following condition must hold:

$$\bar{a}_j + \sum_{i=1}^{L} (e_i - \bar{e}_i) p_{ij} \geq 0 \qquad (32)$$

and

$$\sum_{j=1}^{K}(\bar{a}_j + \sum_{i=1}^{L}(e_i - \bar{e}_i)p_{ij}) = 1 \qquad (33)$$

where $\bar{e}_i$ and $\bar{a}_j$ represent the primary and secondary sample average proportions and e_i is the primary class percentage for the entire county. (The bars over the letters here mean sample averages; they do not indicate per acre estimates as elsewhere in this document.)

Now it can be seen that the earlier derived constraints do not guarantee (31) because $e_i - \bar{e}_i$ may be negative. Therefore, a general solution for $\hat{a}_{rs}$ is not available, but for a specific set of e_i's (county proportions of primary classes), (32) is a linear condition in the p_{ij}'s which can be enforced with the regression approach described below. Some algebraic manipulation shows that (33) holds when (31) is enforced. Thus, $\hat{a}_{rs}$ presents an additional complication when estimating proportions with solutions that are less general than those obtained for $\hat{a}_{pr}$.

B. <u>Regression Model</u>

Suppose that n PSU-SSU pairs are available from which the P matrix can be estimated. The following equation system can then be set up:

$$
N\!\times\!K \underbrace{\begin{bmatrix} [e_{11}\ I][e_{21}\ I] & \cdots & [e_{L1}\ I] \\ [e_{12}\ I][e_{22}\ I] & & [e_{L2}\ I] \\ \vdots & & \\ [e_{1N}\ I][e_{2N}\ I] & & [e_{LN}\ I] \end{bmatrix}}_{L\times K}
\underbrace{\begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1K} \\ p_{21} \\ p_{22} \\ \vdots \\ p_{2K} \\ p_{L1} \\ p_{L2} \\ \vdots \\ p_{LK} \end{bmatrix}}_{(L\times K)\times 1}
=
\underbrace{\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{K1} \\ a_{12} \\ a_{22} \\ \vdots \\ a_{K2} \\ \vdots \\ a_{1N} \\ a_{2N} \\ \vdots \\ a_{KN} \end{bmatrix}}_{(N\times K)\times 1}
+ \qquad (34)
$$

where, for example, $[e_{11}I]$ has the following structure:

$$
\underbrace{\begin{bmatrix} e_{11} & 0 & \cdots & 0 \\ 0 & e_{11} & & \vdots \\ \vdots & & e_{11} & \vdots \\ 0 & \cdots & \cdots & e_{11} \end{bmatrix}}_{K\times K} \qquad (35)
$$

The P matrix has been rearranged as a vector, and the secondary proportions have been joined into a single column. The vector ξ is made up of random disturbances.

The system (34) can be expressed in the usual regression notation as:

$$X\beta = Y + \xi \qquad (36)$$

As was shown, this system must be subject to a set of a priori constraints to yield admissable predictions.

Noting that an inequality constraint can be written as two inequality constraints, for instance a = c: a $\geq$ c and -a $\geq$ -c, the condition (31) can be expressed as the following set of inequalities:

$$\begin{array}{c} L\times K \\ \\ \\ L \\ \\ L \end{array}
\underbrace{\left[\begin{array}{l}
1\ 0\ 0\ 0\ \cdots\cdots\cdots\cdots\ 0 \\
0\ 1\ 0\ 0\ \cdots\cdots\cdots\cdots\ 0 \\
0\ 0\ 1\ 0\ \cdots\cdots\cdots\cdots\ 0 \\
0\ 0\ 0\ 1\ \cdots\cdots\cdots\cdots\ 0 \\
\vdots \\
\hline
1\ 1\ 1\ \cdots\ 1\quad 0\ \cdots\cdots\cdots\ 0 \\
0\ \cdots\cdots\ 1\ 1\ 1\ \cdots\ 1\ 0\ \cdots\cdots\ 0 \\
\vdots \\
0\ \cdots\cdots\cdots\cdots\ 1\ 1\ 1\ ..1...0 \\
\hline
-1\ -1\ -1\ \cdots\ 1\ 0\ \cdots\cdots\cdots\ 0 \\
0\ \cdots\cdots -1\ -1\ -1\ \cdots\ 1\ 0\ \cdots\ 0 \\
\vdots \\
0\ \cdots\cdots\cdots -1\ -1\ -1\ -1..0
\end{array}\right]}_{((L\times K)\ +\ 2L)\ \times\ (L\times K)}
\underset{(L\times K)\times 1}{\beta} \geq
\underbrace{\left[\begin{array}{c}
0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \hline 1 \\ 1 \\ \vdots \\ 1 \\ \hline -1 \\ -1 \\ \vdots \\ -1
\end{array}\right]}_{((L\times K)+2L)\ \times 1}
\qquad (37)$$

Or in matrix form:

$$A\beta \geq c \tag{38}$$

If b is the ICLS estimate for β, then with the least squares criterion the following problem must be solved: minimize

$$Z = 1/2(y-Xb)'(y-Xb) \tag{39}$$

subject to

$$Ab \geq c \tag{40}$$

or

$$Ab - v \geq c \tag{41}$$

where v is a slack vector.

Several approaches to solving the general inequality constraint least squares problem (ICLS) can be found:

(1) Mixed estimation (Zelner, 1971; Theil, 1963)

(2) Quadratic programming (Lemke, 1962; Liew, 1976; Dantzig, 1967)

(3) Branch and bound Methods (Gentle, et al, 1980)

(4) Statistical testing of negative proportions (O'Reagan, 1980)

(5) The condition equation approach (van Roessel, 1974)

The quadratic programming approach was selected for use in the MAIS. It provides a correct solution under all circumstances, and since 1976, an approach to the calculation of the covariance matrix has been available (Liew, 1976).

C. Inequality Constrained Least Squares

In this subsection, the ICLS approach will be outlined briefly with specific emphasis on the derivation of the covariance matrix as proposed by Liew (1977).

The Kuhn Tucker conditions specify that at the optimum point the following conditions hold:

$$(X'X)b - Xy - A'\lambda = 0 \tag{42}$$

$$\lambda'(Ab - c) = 0 \tag{43}$$

$$\lambda \geq 0 \tag{44}$$

where λ is a vector of Lagrangian multipliers. If λ is known, b can be computed as:

$$b = (X'X)^{-1} (Xy - A'\lambda) \tag{45}$$

substituting into (41), one obtains

$$A(X'X)^{-1}Xy - c + A(X'X)^{-1}A'\lambda = v \tag{46}$$

or defining

$$q = A(X'X)^{-1}Xy - c = A\hat{\beta} - c \tag{47}$$

where $\hat{\beta}$ is the OLS estimate and:

$$W = A(X'X)^{-1}A' \tag{48}$$

one obtains the so-called fundamental problem, namely: find v and λ, such that:

$$v = W\lambda + q \ , \quad v'\lambda = 0 \ , \qquad \lambda \geq 0 \tag{49}$$

Two algorithms for solving this problem are available, one by Dantzig and Cottle (1967) and the other by Lemke (1962). The Lemke algorithm was programmed for the Phase IA test. The equation $v = W\lambda + q$ can be written in matrix form as:

$$\underset{m \times 2m}{[I \quad -W]} \underset{2m \times 1}{\begin{bmatrix} v \\ \lambda \end{bmatrix}} = q \tag{50}$$

If there are m constraints, then v and λ are m x 1 vectors, and hence $[v\ \lambda]'$ is 2m x 1. But since the inner product of v and λ is zero, the near zero elements can be eliminated from v and λ. and (50) can be reduced to:

$$[I_1 - W_1] \begin{bmatrix} v^\circ \\ \lambda^\circ \end{bmatrix} = \underset{m\times 1}{q} \tag{51}$$

Now one can "solve" for $[v^\circ\ \lambda^\circ]'$ as follows:

$$\begin{bmatrix} v^\circ \\ \lambda^\circ \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} q \tag{52}$$

or, in particular, by (47)

$$\lambda^\circ = M_2(A\hat{\beta} - c) \tag{53}$$

The following breakdown can also be made:

$$A'\lambda = (\tilde{A}_1', \tilde{A}_2') \begin{bmatrix} 0 \\ \lambda^\circ \end{bmatrix} = \tilde{A}_2' \lambda^\circ \tag{54}$$

Substituting (53) and (54) into (45), one obtains:

$$b = \hat{\beta} + (X'X)^{-1} \tilde{A}_2' M_2 (A\hat{\beta} - c) \tag{55}$$

or

$$b = \beta\,(I + (X'X)^{-1}\tilde{A}_2^{-1}\, M_2 A) - (X'X)^{-1}\tilde{A}_2^{-1}\, M_2 c$$

Designating the multiplier of $\hat{\beta}$ as M, the covariance matrix of b is:

$$\Sigma b = MV(\hat{\beta})M' = \sigma^2 M(X'X)^{-1}M' \tag{56}$$

and its estimator is as follows:

$$\hat{\Sigma} b = s^2 M(X'X)^{-1}M \tag{57}$$

where $s^2 = \sum_{\ell=1}^{nK} e_\ell^2/v$ and $e_\ell = \hat{a}_\ell - a_\ell$.

Normally v is computed as n - k, where n is the number of observations and k is the number of estimated parameters. In the proportion estimation case, v is computed as follows:

$$v = n(K - 1) - K(L - 1) \tag{58}$$

since each vector of proportions has a implied mean. $\hat{a}_\ell$ represents a secondary class sample proportion prediction, and a_ℓ is its observed value.

A coefficient of multiple determination can be computed as:

$$R^2 = \frac{\sum_{\ell=1}^{nK} (\hat{a}_\ell - 1/K)^2}{\sum_{\ell=1}^{nK} (a_\ell - 1/K)^2} \tag{59}$$

where 1/K is the average of all secondary class proportions in the sample.

Similarly, an F statistic for the significance of the linear relation can be computed as:

$$\tilde{F} = \frac{1}{s^2} \frac{\sum_{\ell=1}^{nK} (\hat{a}_\ell - 1/K)^2}{K(L-1)} \tag{60}$$

2.1.3.1.2 Programs

The theory of the previous section was incorporated in the program ESTPRP written in Fortran IV. This program currently runs on the PRIME 550 at the contractor's Berkeley facility. As was mentioned in Section 2.1.1.2, one of the problems associated with proportion estimation is the large number of coefficients to be estimated. For instance, if L = 10 and K = 14, 140 coefficients must be determined, and the associated covariance matrix is of size 140 x 140. Several other matrices of this dimension are also required in the course of computation. The memory requirements of the current program are approximately 800,000 bytes, allowing for the estimation of a maximum of 300 coefficients. This presents no problem on the PRIME 550, because it has a virtual memory; however, should it be desired to run a program on a machine with a more modest, fixed-size memory, an additional programming effort will be necessary.

The ESTPRP program is based on a general ICLS solution, as implemented in subroutine ICLS. This was done because the exact form of the constraint matrix was not known at the beginning of the project. Therefore, considerable space savings can be achieved by implementing a more specific proportion estimation solution.

Already, a considerable reduction of needed memory was obtained by an algorithmic computation of the initial X'X matrix, rather than using straightforward matrix multiplication. The X'X matrix is a sparse matrix, and very likely, a special inversion algorithm designed for this type of matrix, can perform in a fraction of the time currently used by the generalized inverse routine RUST. The routine ICLS calls the routine LEMKE, which contains the Lemke algorithm.

At the beginning of the project it was not known whether the row sum constraint of the P matrix would hold implicitly, or whether the constraint was to be expressed in the A matrix. The latter course was chosen, and rightly so, as it appeared that these constraints were active in most cases tested.

Because a computer is not a perfect mathematical machine, the entire ICLS approach was implemented with appropriate tolerances. For example, the row sum constraint is enforced as follows:

$$1 - \xi \leq \sum_{j=1}^{K} p_{ij} \leq 1 + \xi$$

with $\xi = 0.0001$. The error in the sum of predicted proportions was generally less than 0.0005.

The PRIME 550 is a 32-bit machine; all computations are therefore performed in double precision.

The program ESTPRP writes $\hat{P}_1$, and $\hat{\sum}\hat{P}_1$ to disk for further processing by program ESTCNT. ESTCNT takes the county proportions (see Section 2.2.2.2.6) and the output matrices of ESTPRP and computes secondary class proportions for the county according to equations (6) and (7), Section 2.1.2.1. It, in turn, writes $\hat{a}_{pr}$ and $\hat{\sum}\hat{a}_{pr}$ to the disk for further processing.

2.1.3.1.3 Program Testing

Although the "linear model" concept of the Step 1 area computations is simple and elegant, its details as implemented through the ICLS approach are quite complex. The greatest precautions were taken, therefore, to assure that the programming was correct and that the program will continue to function properly. This was accomplished by (a) computing known test cases, and (b) incorporating test computations into the program.

Lemke (1963) provides a numerical example, beginning with the matrix W. It was used to test the LEMKE subroutine. Liew (1976) gives an example of a covariance computation for an economics case. An attempt was made to reconstruct this test computation, but it failed as the reference material in the paper was inadequate. An example was therefore requested directly from the author; a data set was received together with a test set made with the author's package at the University of Oklahoma (Liew and Shim, 1978). The covariance matrix of this set was duplicated on the PRIME 550. The differences between the computed elements of the covariance matrices was generally less than 0.0003 for elements with a magnitude of 0.02.

The test computations built into the program check various relations that should exist in the course of the computation, such as (49), (52) and (54), of Section 2.1.3.1.1. In addition, matrix inverse computations are checked by multiplying the result by the original matrix, and matrices to be inverted are inspected for rank deficiency. During development of the program, test computations provided insight into the tolerances to be used; in future use, they will quickly abort the program in the eventuality of storage problems or other environmental defects.

2.1.3.2 Field Statistics Component

This component, as currently implemented in programs CSTAT and DSTAT, computes the sample averages, by class, and the estimated standard deviations for the parameters of interest (CAI). (See Section 2.1.2.2.2, subparagraph A.) It also computes $\hat{P}_2$ and $\sum\hat{P}_2$ (See Section 2.1.2.2.1).

The procedure currently in use is as follows: each field plot is interpreted as to its secondary (aerial photo) attributes. A file is then constructed which assigns two sets of attributes to each plot. The first set is the PI code; the second set is derived from the basic plot and field observations and consists of the plot estimates of the parameters of interest. Currently this second set contains: CAI, Ground Land Use Code and Forest Type Code; the latter as defined by McClure, Cost and Knight (1979).

Plots can then be sorted on various combinations of PI attributes in the identical manner in which the SSU resource units are sorted for a desired stratification. (See Section 2.1.3.5) For continuous variables, averages and standard deviations are computed by these classes (Program CSTAT) For discrete data, a P matrix and a covariance matrix are computed (Program DSTAT).

If, for security reasons, plot data are not to be stored in the lower level GIS, the field statistics component can be the primary receptacle for all field data and can be the main instrument for all field data-related processing. Thus, all field data manipulations currently performed by the Plot Summary Program at the Southeastern Experiment Station can be a part of the field statistics component.

2.1.3.3 Summary Component

The function of this component simulated in the Phase IA test are currently performed by program ESTSMR. The possible functions are the computations of the estimators and variances described in Section 2.1.2.2. Program ESTSMR takes the outputs of programs ESTCNT, CSTAT and DSTAT, and computes the final estimates and their estimated variances.

2.1.3.4 Primary Proportion Extraction

One alternative is to consider the proportion extraction tasks as a part of GIS-related processing. However, it is proposed that these tasks are to be a part of the estimation subsystem. A compelling reason for this recommendation is that the calculated proportions are for classifications directly related to the types of estimates to be made, to be specified at request time. These classifications can then be defined as the estimation subsystem is invoked, and the proportions can be generated accordingly. The manner in which primary (LANDSAT) PSU proportions were extracted for the current test is described in Section 2.2.2.2.4. A program EPREP was written to further combine classes generated by the program COUNT described in this section.

2.1.3.5 <u>Secondary Proportion Extraction</u>

The LANDPAK report generator is a versatile program capable of sorting and summarizing resource unit data in a variety of ways. For the purpose of the test, an output option was built into this generator which allows one to summarize data onto a file for further processing. Using this option, a file is generated which for each RU contains the RU number, the LANDPAK-computed area, the SSU number and the PI code for the RU. A program JPREP was then created that sorts the RU's in this file according to a given classification, computes the proportions of the total area of each class of the SSU, and outputs a list of these proportions by SSU to the disk for further processing. Program JPREP must be modified for each additional classification, and as such is actually part of the file. A corresponding modification must be made in CSTAT. A user friendly way of specifying a desired classification must be part of a permanent MAIS.

2.2 Prototype Test

The Phase IA test is described in this section. The report of this effort is divided into two main parts: input and data preparation of the upper and lower level GIS, and analysis of the Step 1 and Step 2 computations of the prototype estimation subsystem. Once again, it should be emphasized that the current version of this subsystem, with the exception of the program ESTPRP, consists mainly of "throwaway" programs specially created for the Phase IA test.

The description of the testing effort is preceded by a short description of the test area: Kershaw County in South Carolina.

2.2.1 Test Area Description

Kershaw County, South Carolina, was selected as the site for the Phase IA pilot test. Situated in the north-central region of the state, it offers a variety of forest types and land use regimes, making it suitable for evaluating new, multiresource inventory methods.

Kershaw County has a total land area of 499,840 acres, of which approximately 395,000 are forested (G. C. Craver, 1978). It contains three physiographic regions: the Southern Coastal Plain, Carolina Sand Hills and the Southern Piedmont.

The predominant forest types are loblolly-shortleaf pine, oak-gum-cypress, oak-hickory, and oak-pine. The distribution of these types is controlled mainly by soil type and moisture. Broadly, hardwoods occur in the drains with increasing occurrence of pine on drier ground. The influence of man on the distribution of forest species is significant.

Reforestation of abandoned cotton fields throughout the county has been encouraged, beginning in the 1950's. Species planted consist chiefly of loblolly and shortleaf pines. Much of the land on upland sites is cropland and pasture.

The range in elevation is from 100 feet to about 500 feet. The climate is mild, due to the influence of the Gulf of Mexico and the Atlantic Ocean. Winters are humid and mild, with average January temperatures of 45°F, although occasional periods of frost and freezing temperatures occur. Summers are warm and humid, with average July temperatures of 80°F. Mean annual total rainfall is approximately 50 inches.

2.2.2 Input and Data Preparation

2.2.2.1 Lower Level GIS

2.2.2.1.1 Sample Selection

The sampling configuration decided upon includes the 210 Forest Service field plots. These are termed Ultimate Sample Units (USU's). In addition, a random sample of aerial photo sample units were selected. These are termed Secondary Sample Units (SSU's). The SSU's are square, measuring one mile on an edge on the ground.

The SSU's were selected by first selecting (randomly, without replacement) a subset of the 210 ground plots. All of the ground plots classified by the Forest Service as other than forest or cropland were deliberately included in the sample. This was an attempt to improve the distribution of sample plots among the classes, as these classes were inadequately represented in the ground plot sample.

Around each USU selected (we chose a total of 60), an SSU was randomly located so that the USU was contained within the SSU boundary and oriented orthogonal to the directions of the compass.

2.2.2.1.2 Photo Interpretation

The aerial photographs provided are panchromatic black and white, taken in April, 1975. The 1:20,000 scale prints were enlarged from 1:40,000 nine-inch negatives. The image quality ranges from fair to poor, with graininess and low contrast being the main deficiencies. Incomplete stereo coverage for more than half of the SSU's was also a problem. High quality color IR optical bar photography, flown in 1979, was used to supplement photo coverage of these problem areas.

A list of land use classes was devised based on the requirements described in the RFP. Image quality in conjunction with the interpreters' ability to discriminate between these classes was the basis for formulating the set of aerial photo classes used. (See Appendix B.)

Training was accomplished by first selecting training sites, which covered the range of conditions identifiable on the photos. These sites were then interpreted using a 1-3 power mirror stereoscope.

A field trip was made to the test area in July, 1980. EarthSat personnel spent 6 days in the area. The training sites were checked for correct interpretation, and new training sites were established. The training sites were described and documented using stereo pairs of color photos taken from ground stations. Some training sites were documented by low-altitude, oblique aerial photos taken during a reconnaissance and photo flight chartered by EarthSat. These documents, keyed to the aerial photos, were the main reference and training aid used in the photo interpretation.

In addition to the interpretation of the SSU's, the remaining USU's were classified as to photo class. Thus, information was compiled for 60 SSU's and 210 USU's. An interpreted photo for SSU No. 20 is illustrated in Figure 2.

2.2.2.1.3 Data Entry

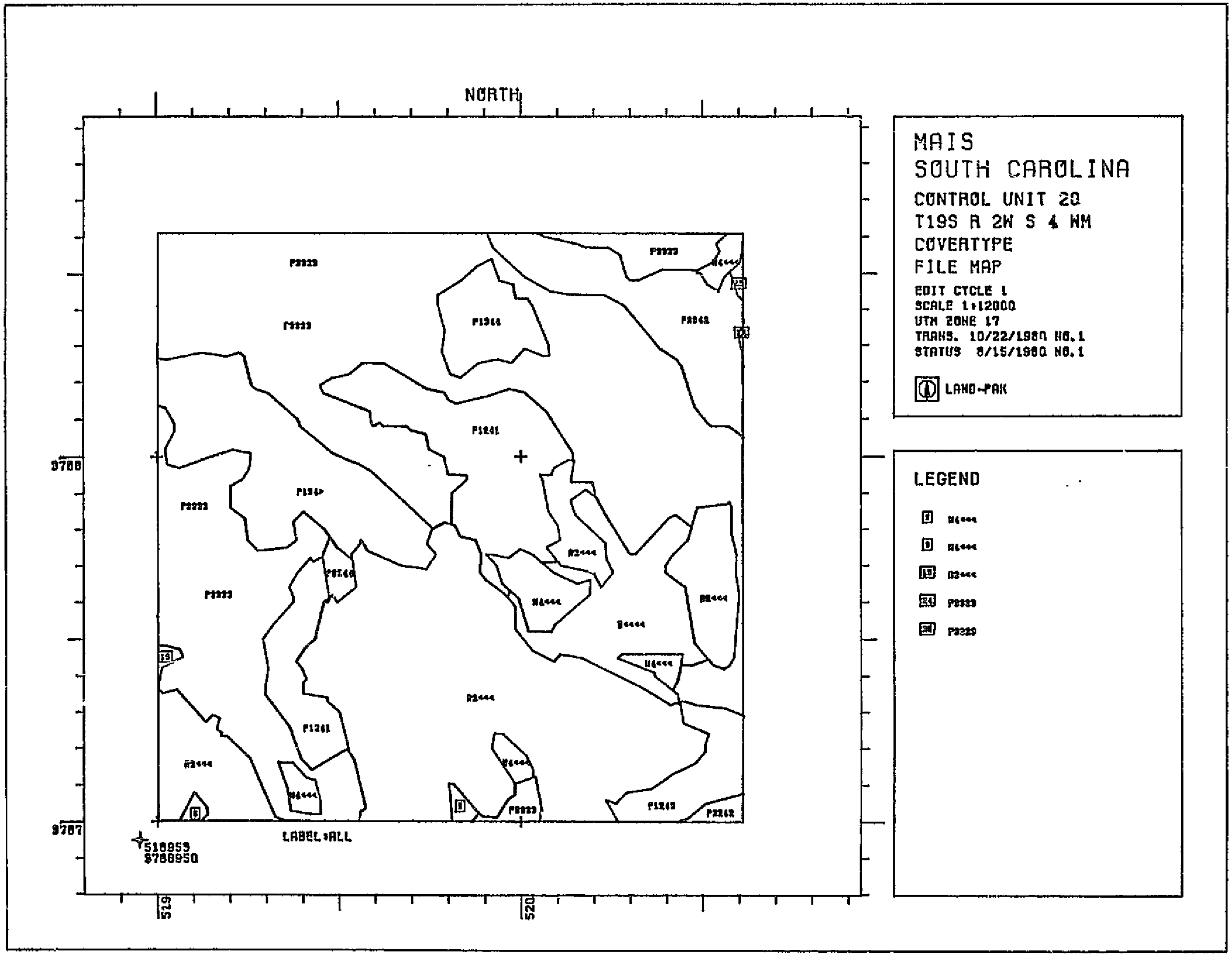
The lower level GIS consists chiefly of the LANDPAK system. The first step in entering data from source maps is to establish geometric control for that map. This consists of establishing enough points on the source map, of known ground coordinates, to ensure a good transformation from digitizer table coordinates to ground coordinates. This is between 4 and 8 points, usually 4 if the source map is a controlled map and 6 to 8 if the source map is a delineated photo or other uncontrolled map.

Control points were derived from 7½-minute U.S.G.S. topographic maps if available; otherwise, 15-minute maps were used. An error of 15 meters RMS was the tolerance allowed for any transformation.

During the control procedure, the ground locations of the SSU centers were established. These centers were used to generate control units, the control layer for LANDPAK data. At this point, digitizing and entry into the LANDPAK system of the SSU type-maps could proceed. These type-maps, when entered, constitute the covertype layer. Similarly, elevation contour lines were digitized and entered to form the topography layer.

The data items entered for the covertype layer consist of elements of the PI codes assigned to each delineation or resource unit (RU). Thus for covertype, 5 data items were entered for each RU, including values denoting unknown or irrelevant status. Stored data items for the topography layer consisted of contour line elevations. Figures 3 and 4 illustrate covertype and topography

Figure 2.



NORTH
MAIS
SOUTH CAROLINA
CONTROL UNIT 20
T19S R 2W S 4 WM
COVERTYPE
FILE MAP
EDIT CYCLE 1
SCALE 1:12000
UTM ZONE 17
TRANS. 10/22/1980 NO.1
STATUS 8/15/1980 NO.1
LAND-PAK
LEGEND
3788
3787
LABEL:ALL
518959
3786950
519
520

Figure 3.

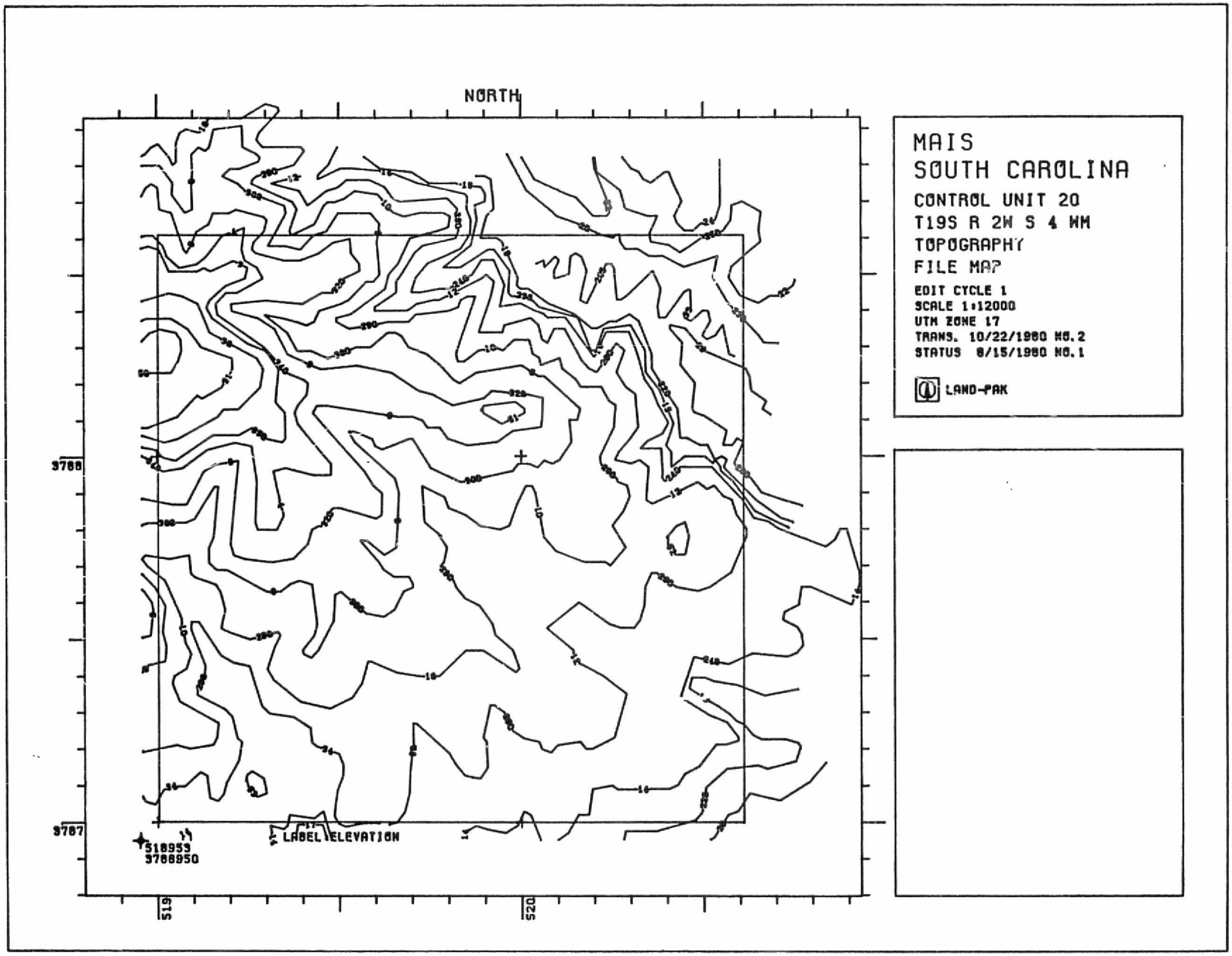

NORTH
MAIS
SOUTH CAROLINA
CONTROL UNIT 20
T19S R 2W S 4 WM
TOPOGRAPHY
FILE MAP
EDIT CYCLE 1
SCALE 1:12000
UTM ZONE 17
TRANS. 10/22/1980 NO.2
STATUS 8/15/1980 NO.1
LAND-PAK
3788
3787
LABEL ELEVATION
518953
3788950
519
520

Figure 4.

layer images, respectively, which were retrieved from the database and plotted for SSU No. 20.

2.2.2.1.4 Data Compilation for Soil Loss Estimates

A model was selected to demonstrate an application of these methods to soil erosion potential estimation. The model, the Universal Soil Loss Equation (USLE) is expressed as:

$$A = RKLSCP$$

where:

A is the average annual soil loss in tons/acre/year from the site

R is the value of the rainfall erosivity index for the site

K is the value of the soil erodibility index for the site

L is the length of slope factor

S is the steepness of slope factor

C is the vegetation cover influence factor

P is the erosion control practice factor.

(For more information on applying the USLE, refer to USDA, SCS Technical Notes, January 1, 1978.)

Average values for R, K and L for each of two major regions in Kershaw County were taken from a previous study (Dissmeyer and Stump, 1978). Each SSU fell into one of the regions for purposes of assigning these factors. Efforts to include soil-type maps in the database were abandoned after delays in receiving requested data. This would have facilitated site-specific determinations for the K factor. Attempts to approximate K values by SSU were also abandoned for lack of data.

Pertinent information existed in the LANDPAK database in the form of the covertype and topography layers. This provided a means of assigning C and S values. Slope-class maps were generated using the topography layer and LANDPAK subsystems. The following class limits, in per cent slope, were devised:

1. 0 - 2
2. 2 - 4
3. 4 - 8
4. 8 - 16
5. 16 - 30
6. 30+

The slope-class maps were inserted into the database and constitute the slope-class layer. A slope-class map retrieved from the database and plotted can be seen in Figure 5. Values for S, for assignment to each slope-class (Table 1, Technical Notes, 1978) were selected.

Representative values for C were assigned based on each RU's covertype PI code. These values were approximated from published suggested values (Technical Notes, 1978). We feel it should be noted here that the PI classes were not configured for determining these values. We chose a C value which was our best approximation for an average value for that PI class. Similarly, we were not in a position to determine P values. Since P applies chiefly to agriculture and is really an adjustment to account for efforts to minimize erosion, we feel its omission is relatively unimportant but should ultimately be included to account for certain forest practices.

2.2.2.1.5 <u>Derivation of Soil Loss Sample Set</u>

This section describes the processing of the soil loss data using LANDPAK

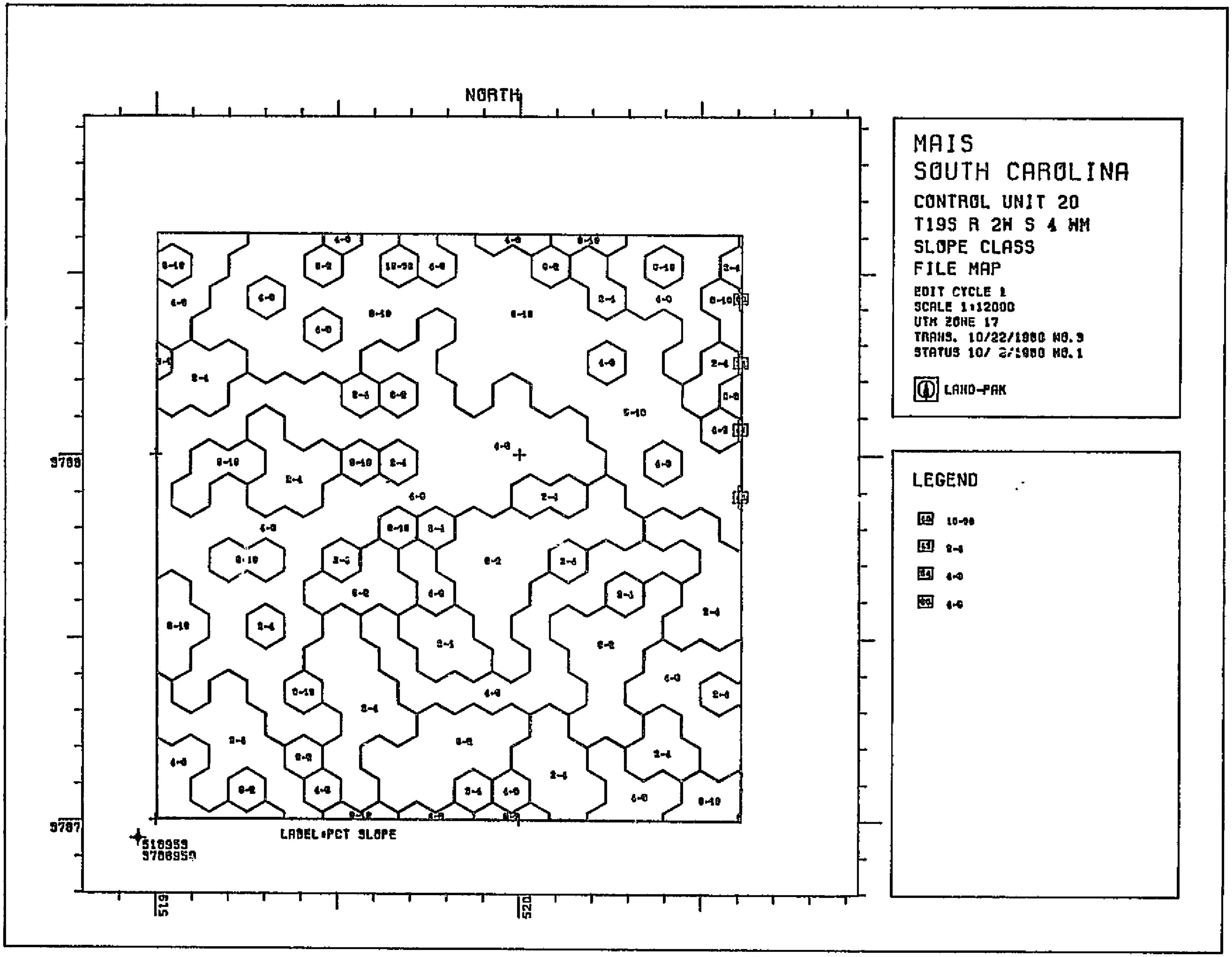

NORTH
MAIS
SOUTH CAROLINA
CONTROL UNIT 20
T19S R 2W S 4 WM
SLOPE CLASS
FILE MAP
EDIT CYCLE 1
SCALE 1:12000
UTM ZONE 17
LAND-PAK
LEGEND
LABEL:PCT SLOPE
519
520

Figure 5.

and special programs up through the interface with the estimation subsystem.

The slope-class and covertype layer images were retrieved from the database, and an overlay operation was performed to intersect these two map layers. The resulting map, a new layer referred to as the selection map, consists of new RU's of homogeneous slope-class and PI class. Figure 6 is a plot of this selection map for SSU No. 20. A LANDPAK report was generated for this selection map and stored in an auxiliary file. This report contains information on each RU as to land area, PI code and slope-class.

This report was input into EROS, a modified version of computer program JPREP2. EROS assigns the appropriate values to factors of the USLE. EROS solves the USLE and assigns the RU to a soil-loss class. The soil-loss class limits were assigned (in tons/acre/year) as follows:

1. 0 - 2
2. 2 - 4
3. 4 - 8
4. 8 - 16
5. 16 - 32
6. 32 - 64
7. 64+

The program accumulates areas by each of these classes. The proportion of the total area contained in each of these classes is then computed. This table of class proportions is the output from EROS, which is stored for use by the estimation subsystem. Essentially, this process is carried out by SSU. The whole process was carried out for only 12 SSU's due to constraints of time and resources.

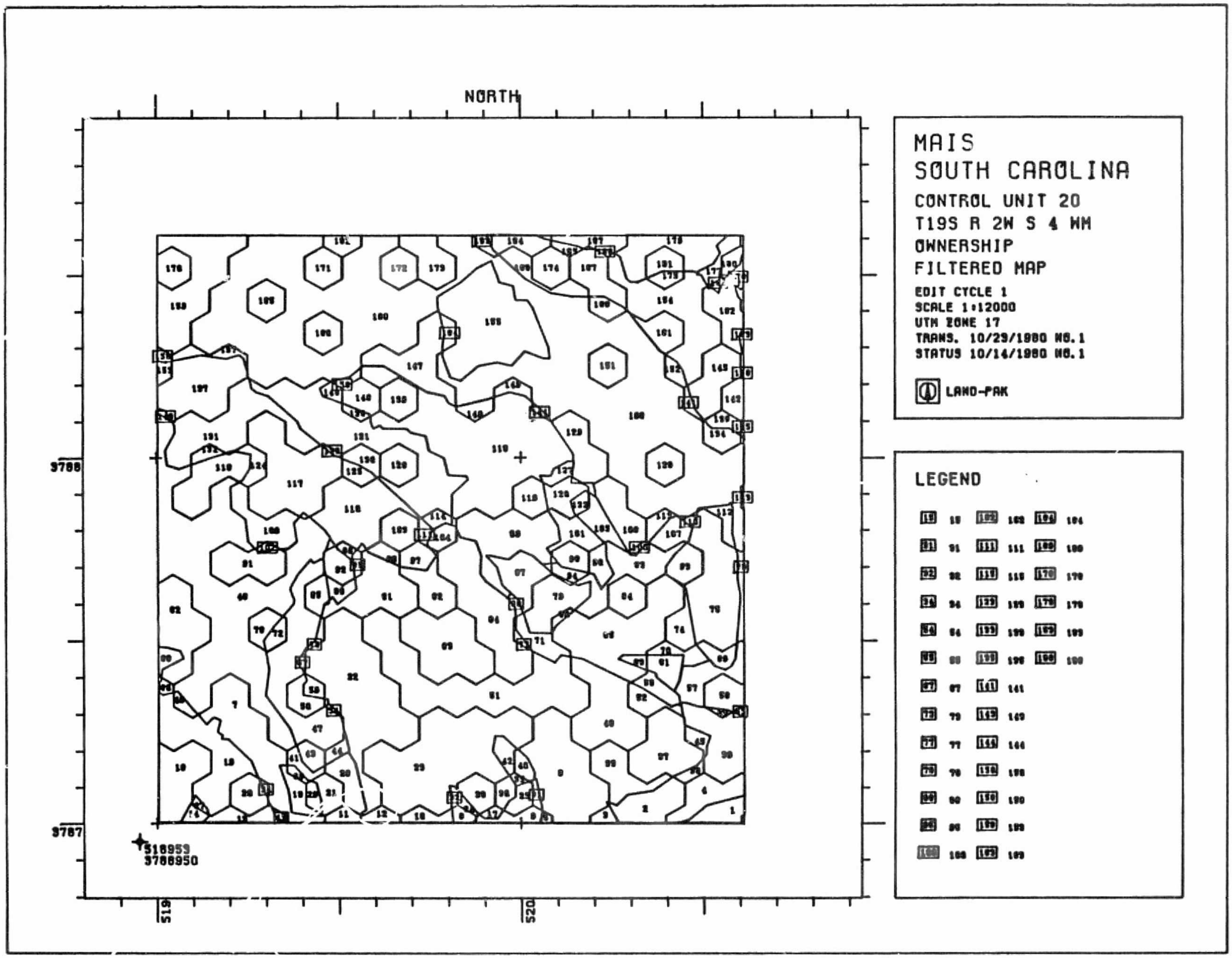

NORTH
MAIS
SOUTH CAROLINA
CONTROL UNIT 20
T19S R 2W S 4 WM
OWNERSHIP
FILTERED MAP
EDIT CYCLE 1
SCALE 1:12000
UTM ZONE 17
TRANS. 10/29/1980 NO.1
STATUS 10/14/1980 NO.1
LAND-PAK
LEGEND
3788
3787
518953
3786950
519
520

Figure 6.

2.2.2.2 Upper Level GIS

This section describes the procedures used for the input and manipulation of LANDSAT data in the upper level GIS. The upper level functions used in this project included LANDSAT preprocessing, unsupervised classification, image registration, statistical tabulation and map generation. With the exception of unsupervised classification and digitization of the county boundary, all processing was performed at EarthSat's Washington, D.C., office on a PRIME 450 minicomputer. Classification was performed at the Remote Sensing Laboratory at the University of California at Berkeley. The digitizing of Kershaw County boundary was done by EarthSat, Berkeley, personnel using LANDPAK.

2.2.2.2.1 LANDSAT Preprocessing

The Kershaw County study area is contained within a single LANDSAT frame of path 17 or 18, row 36. Two scenes were available: 11035-15054 (LANDSAT 1, May 1975) and 30515-15201 (LANDSAT 3, August 1979). The LANDSAT 1 scene was chosen because it was available in the old CCT format and could be destriped with EarthSat's scan line suppression software. The LANDSAT 3 scene was available only in corrected MDP format which cannot be destriped because the geometric correction process destroys the detector identities.

EarthSat's LANDSAT preprocessing program, CCTRFM, performs 3 functions: reformatting, scan line suppression and geometric correction, in a 2-pass procedure. In the first pass the four spectral bands are separated into different files, and the four vertical strips are pieced together. At the same time, the radiometric calibration introduced by NASA is removed and a histogram is acquired for each of the 24 detectors. Four calibration lookup tables are computed in order to match the six detectors for each spectral band. In the second pass, the image is recalibrated and resampled (nearest neighbor) to correct for mirror velocity profile, earth curvature and panoramic distortions. Synthetic pixels, extra pixels inserted by NASA to attain consistent line length, are deleted during this pass.

The corrected LANDSAT tape generated by CCTRFM was then used as input to the classification and registration steps.

2.2.2.2.2 Classification

Classification of the LANDSAT digital data was performed on the data general NOVA 840 of the Space Sciences Laboratory of the University of California at Berkeley, with program CLUSTER. It employs the ISODATA algorithm developed by Ball and Hall, and is about five generations removed from the original program (Ritter and Kaugars, 1978).

An envelope for Kershaw County was defined using the interactive image processing system, and an initial clustering was performed on a 4% sample of the points in the envelope.

The maximum standard deviation for the splitting of a cluster was set at 0.6, and the minimum distance for combining clusters was defined at 3.2. The run terminated cleanly without the iterative splitting and combining that occurs at times, with an average cluster standard deviation of 0.3779, and an average intracluster distance of 15.04. A total of 14 clusters was generated.

TABLE 3

Percentage of Points in Kershaw County Envelope Assigned to Each Cluster

Cluster	Per Cent
1	0.3
2	25.5
3	22.2
4	19.3
5	6.7
6	7.9
7	4.0
8	0.6
9	2.2
10	8.6
11	0.7
12	0.8
13	0.5
14	0.3

Table 3 lists the clusters and the percentages of points assigned to each. It is interesting to note that clusters 2, 3 and 4 contain almost 70% of the points in the sample.

After the initial cluster run, the program was restarted at a later date, and all points in the area were assigned a cluster number. An output tape with the resultant image was forwarded for processing to the EarthSat, Washington, D.C., office.

In the following, we will refer to the cluster and its associated points as a spectral class. By means of an inspection of the classification image, initial land use assignments were made for each spectral class.

TABLE 4

Initial Land Use
Cluster Assignments

Cluster	Land Use Category
1	Bare Soils
2	Hardwood Pine Mix
3	Bottom-land Hardwoods
4	Pine Hardwood Mix
5	Cropland
6	Pine
7	Bare Soil
8	Wetland
9	Bare Soil
10	Cropland
11	Water
12	Bare Soil
13	Water
14	Bare Soil

The number of spectral classes was judged too high for further analysis, and so, after the SSU proportions were extracted, spectral classes were combined on the basis of their intracluster distance, the number of points in each class, and the initial land use class assigned to the cluster.

Two sets of classifications were ultimately used, as shown in Table 5, one with 7 classes and one with 10 classes. We will refer to these classifications as the 7- and 10-class spectral classifications.

TABLE 5

Composition of Spectral Classifications Used in ICLS Estimations

	7 Spectral Classes	10 Spectral Classes
Spectral Class	Cluster(s)	Cluster(s)
1	1, 12, 14	1, 12, 14
2	2	2
3	3	3
4	4	4
5	5, 7, 9, 10	5
6	6	6
7	8, 11, 13	7, 9
8		8
9		10
10		11, 13

2.2.2.2.3 Image Registration

In this step, ground control points (GCP's) were located to develop a relationship between LANDSAT coordinates (line, column) and the UTM coordinate system (northing, easting). 7½-minute maps were obtained for Kershaw County and the surrounding area. Unfortunately, a large section of the county is covered only by a 15-minute sheet (Camden) which was last updated in 1939. This was unacceptable for our purposes, so no ground control was available for that area of the county.

Five GCP's were located on the 7½-minute maps and the LANDSAT image. Since the study area covered about 1/9 of a LANDSAT scene, this was equivalent to 45 GCP's for a full scene. EarthSat's program SHADE was used to produce line printer maps of those portions of the LANDSAT image (band 5) which contained the control points. The line and column coordinates for each point were measured from these line printer maps. UTM coordinates were measured from the U.S.G.S. maps.

Ground Control Points

LANDSAT		UTM	
Line	Column	N	E
1471	371	3,810,874	520,071
1126	116	3,840,521	512,163
1094	167	3,842,405	515,627
1711	659	3,789,234	531,929
1367	193	3,821,109	511,790

These points were entered into EarthSat's program AFFINE to determine the affine coefficients which would map them with the minimum RMS error. The affine transformation has the form:

$$X_2 = V_1*X_1 + V_2*Y_1 + V_3$$

$$Y_2 = V_4*X_1 + V_5*Y_1 + V_6$$

where X_1, Y_1 is the location in coordinate system 1, X_2, Y_2 is the location in coordinate system 2, and V_1 through V_6 are the coefficients. The actual transformation used was:

UTM N = -77.095*LINE - 11.468*COLUMN + 3928650

UTM E = -19.565*LINE = 57.590*COLUMN + 527457

This transformation yielded a 67-meter RMS mapping error at the control points.

2.2.2.2.4 SSU Proportion Extraction

With a suitable mapping transformation defined, it was then possible to locate the 60 SSU's on the classified LANDSAT image and calculate the class proportions.

EarthSat's program RESAMP was used to resample and extract each SSU. The data were resampled to a 20-meter UTM grid using a nearest neighbor algorithm. The 20-meter grid was chosen over a coarser grid to allow for a more accurate splitting of border pixels (a 20 x 20 meter cell is approximately 1/12 of a LANDSAT pixel). Each resampled SSU consisted of an 80 x 80 array of class numbers. EarthSat's program COUNT was used to count the occurrences of each class and to convert them into proportions. A color coded map of a resampled SSU classification image is shown in Figure 7. This is the identical SSU shown in

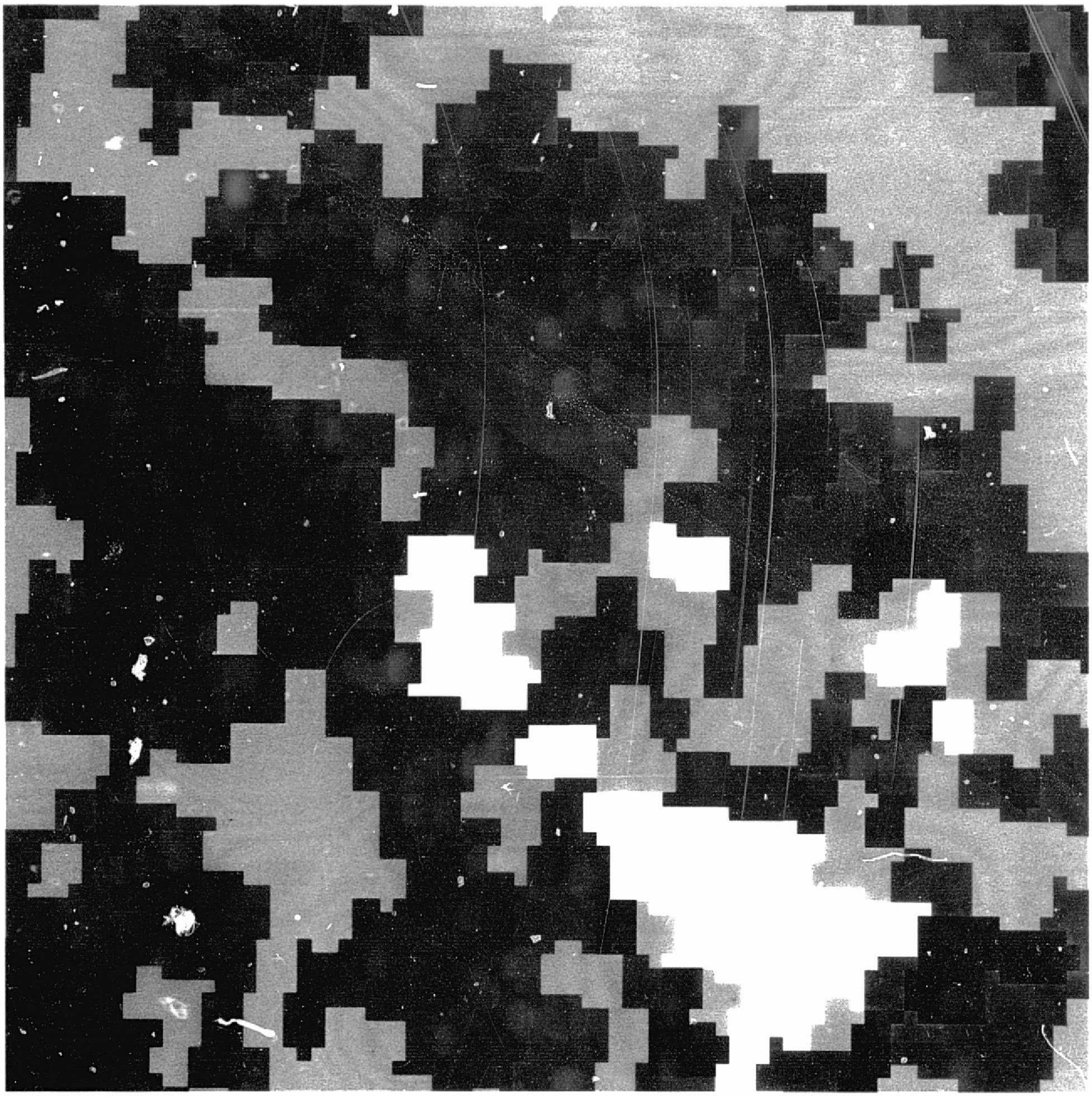


FIGURE 7.

Figures 2, 3, 4, 5 and 6.

2.2.2.2.5 Digitizing of County Boundary

The county boundary was digitized from U.S.G.S. 15-minute quadrangle maps, using EarthSat's program TALOS. Files produced for each quadrangle were edited and merged for input to program CPOINT, a LANDPAK auxiliary program. CPOINT was used to convert the machine coordinates to UTM coordinates using a projective transformation computed from digitized map control points.

The file of UTM coordinates was then input to program KCC to produce a simulated binary image to be used as a mask for the classification image. A tape with this county boundary image was forwarded to EarthSat's Washington, D.C., office.

2.2.2.2.6 County Proportion Extraction

In this step the digitized county boundary was logically combined with the classified LANDSAT image to generate class statistics for the entire Kershaw County area. RESAMP was again used, this time to resample the entire classified image to the 50-meter UTM grid of the county boundary file. EarthSat's program CMBHST was used to compute statistics for those LANDSAT pixels which fell within the Kershaw County boundary.

2.2.2.2.7 Map Generation

The map generation capability of the upper level GIS was used to create a LANDSAT classification map of Kershaw County. EarthSat's program COMBIN was used to combine the resampled class map with the county boundary map, masking out all pixels which lay outside the county. The masked class map was then processed by EarthSat's program OPTRONIC, to assign a color to each class and create three annotated film recorder compatible tapes (one tape for each primary color). An optronics film recorder was used to create three black and white transparenci-s from these tapes. EarthSat's photo lab composited the transparencies to produce a color negative and color photomaps at various scales.

2.2.3 Analysis

The most significant results of the estimation computation of the Phase IA test are presented in this section. Three primary categories of variables are of interest: namely, land use, current annual increment, and soil erosion. This section is divided into three corresponding subsections, and for each of these we will discuss the Step 1 and 2 computations.

2.2.3.1 Land Use, Step 1

Six land use classifications were defined at the secondary level. They are summarized in Table 6.

The first is a land use classification with three broad forest type-classes. The second has three additional forest type-classes obtained by subdividing the original classes according to tree size. The third classification is similar, but forest classes are split according to tree densities. In the fourth classification, the three basic forest types are split two ways: by tree density and by tree size. For this classification, all other non-forest classes are collapsed into three: namely, wetland, disturbed and other. The fifth and sixth classifications were made to test the notion that better estimates can be obtained for any given class when the class is treated by itself rather than in conjunction with a number of other classes.

All the classifications are referred to as land use classifications; specifically, we will refer to each individually as indicated at the top of Table 6: namely, land use, forest types 1, forest types 2, forest types 3, grass and water.

TABLE 6

Secondary Land Use Classifications

Class	Land Use	Forest Types 1	Forest Types 2	Forest Types 3	Grass	Water
1	FC	FCL	FCS	FCLS	GR	W
2	FH	FCH	FCB	FCLB	NG	NW
3	FM	FHL	FHS	FCHS		
4	AG	FHH	FHB	FCHB		
5	WL	FML	FMS	FHLB		
6	WT	FMH	FMB	FHHS		
7	GR	AGI	AGI	FHHB		
8	UR	AGN	AGN	FMLS		
9	DS	WL	WL	FMLB		
10	BS	WT	WT	FMHS		
11		GR	GR	FMHB		
12		UR	UR	WL		
13		DS	DS	DS		
14		BS	BS	OT		

Meaning of first two characters:

FC = Forested Conifer
FH = Forested, Hardwood
FM = Forested, Mixed
AG = Agriculture
GR = Grass
UR = Urban
DS = Disturbed
BS = Bare Soil

NG = Non-Grass
NW = Non-Water
GT = Other

Meaning of third and fourth characters:

L = Low Tree Density
H = High Tree Density
S = Small Trees
B = Big Trees
I = Irrigated
N = Non-Irrigated

2.2.3.1.1 Land Use

Two runs were made with program ESTPRP, one for each of the spectral classifications mentioned in Section 2.2.2.2.2. The county proportions of the land use classes for each of these runs are listed in the left-hand side of Table 7. Proportions for the same classes were also computed from the sample set with the estimator $\hat{a}_{av}$ (Section 2.1.2.1). These proportions are reported in the right-hand side of Table 7. It can be seen in this table that the estimator $\hat{a}_{av}$, for which the estimated proportions are the SSU averages, is a fairly good estimator compared with $\hat{a}_{pr}$, which makes use of LANDSAT. The standard errors for $\hat{a}_{pr}$ are fairly uniform for all classes, whereas for $\hat{a}_{av}$ the standard errors vary proportionately with the class proportions. For the larger classes, $\hat{a}_{pr}$ seems to be superior to $\hat{a}_{av}$; however, the standard errors are small to start with, so that an efficiency increase obtained with LANDSAT amounts to only a few per cent in increased accuracy.

One interesting effect that can be observed in Table 7 is that $\hat{a}_{av}$ estimates water at 3.24%. The LANDSAT estimator $\hat{a}_{pr}$, however, reduces this estimate to 1.72%, indicating that the sample is not representative of the entire county. The probable reason is that one SSU is situated on the Wateree Reservoir.

The correlation coefficients are surprisingly high for Landsat-based analyses. The highest correlations obtained in an earlier study relating LANDSAT data to ground data (van Roessel, 1976) was 0.67. The $\tilde{F}$ values of 11.68 and 8.29 must be compared with $\tilde{F}_{0.95}$ for the indicated degrees of freedom: namely, 1.32 and 1.25, respectively. The linear relationships are, therefore, highly significant, even satisfying a criterion mentioned by Draper and Smith (1966), namely that for a useful prediction, the value of $\tilde{F}$ must be at least four times the critical value.

TABLE 7

Land Use Proportion Estimates
For Kershaw County

	Estimator: $\hat{a}_{pr}$				Estimator: $\hat{a}_{av}$	
	7 Spectral Classes		10 Spectral Classes			
Land Use Class	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error
FC	33.03	1.34	33.44	1.43	32.24	2.85
FH	6.71	1.38	6.89	1.38	6.23	1.02
FM	28.74	1.40	29.11	1.43	28.36	2.14
AG	14.36	1.20	12.04	1.36	13.83	2.24
WL	5.84	1.31	5.88	1.26	5.78	1.25
WT	1.72	1.07	1.82	1.39	3.24	1.56
GR	6.33	1.40	6.77	1.39	6.60	0.99
UR	1.92	1.17	2.50	1.39	2.10	0.80
DS	1.25	1.32	1.37	1.39	1.39	0.33
BS	0.05	1.20	0.13	1.39	0.22	0.12
R	0.7677		0.7740			
$\tilde{F}$	11.68		8.29			
D.F.	(63,477)		(90,450)			

The standard errors in Table 7 can be used to construct individual confidence intervals for each class using a $t_{0.025}$ value with approximately 450 degrees of freedom: 1.96. However, one must not make a simultaneous interpretation of these intervals.

It is interesting to note that the smaller number of spectral classes provides as good a correlation as the larger number, with a higher $\tilde{F}$ value.

2.2.3.1.2 Forest Types 1

Again, two runs were made with ESTPRP, one for each spectral classification. The results are shown in Table 8. The additional breakdown of the forest classes by tree size does not seem to be meaningful, as the correlation is down by .1 in both cases, and the $\tilde{F}$ values are considerably less than those of Table 7. This makes sense when considering that the LANDSAT signatures are probably related more to tree density than to tree size.

2.2.3.1.3 Forest Types 2

Here the additional breakdown is by tree density, rather than by tree size. Results are shown in Table 9. Only one spectral classification was used. The standard errors are generally lower, and the correlation coefficient and $\tilde{F}$ value are considerably higher than those for the breakdown by tree size.

2.2.3.1.4 Forest Types 3

The original three forest types FC, FH and FM are all divided by tree size and tree density, yielding twelve different classes. However, one class has been eliminated due to the lack of any occurrence for this class in the aerial photo sample. The total absence of a class in the SSU proportions

TABLE 8

Forest Types 1 Proportion
Estimates for Kershaw County

	Estimator: $\hat{a}_{pr}$				Estimator: $\hat{a}_{av}$	
	7 Spectral Classes		10 Spectral Classes			
Land Use Class	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error
FCL	16.20	1.25	16.81	1.25	14.38	2.56
FCH	17.53	1.24	17.85	1.24	17.86	2.10
FHL	0.05	1.17	0.09	1.17	0.12	0.10
FHH	6.51	1.27	6.72	1.27	6.11	1.03
FML	2.75	1.25	2.86	1.25	2.68	0.69
FMH	25.92	1.29	26.21	1.29	25.67	2.00
AGI	14.08	1.09	11.88	1.22	13.80	2.24
AGN	0.00	0.00	0.00	0.00	0.03	0.02
WL	5.79	1.21	5.84	1.20	5.78	1.25
WT	1.68	1.17	1.75	1.17	3.24	1.56
GR	6.23	1.29	6.44	1.30	6.60	0.99
UR	1.87	1.23	2.08	1.22	2.10	0.80
DS	1.22	1.17	1.27	1.17	1.39	0.33
BS	0.11	1.17	0.16	1.17	0.22	0.12
R	0.6813		0.6877			
$\tilde{F}$	7.14		5.11			
D.F.	(91,639)		(130,650)			

TABLE 9

Forest Types 2 Proportion
Estimates For Kershaw County

Estimator: $\hat{a}_{pr}$, 7 Spectral Classes		
Land Use Class	Proportion (Percent)	Standard Error
FCS	2.94	1.17
FCB	29.91	1.13
FHS	0.00	0.00
FHB	6.80	1.16
FMS	3.38	1.12
FMB	25.41	1.17
AGI	14.32	0.99
AGN	0.00	0.00
WL	5.82	1.09
WT	1.67	0.88
GR	6.44	1.17
UR	1.94	1.12
DS	1.27	1.15
BS	0.05	1.12
R	0.7643	
$\tilde{F}$	11.49	
D.F.	(91,639)	

causes singularity in the X'X matrix. The non-forest classes were compressed into three categories: wetland, disturbed and other. The estimated proportions for the classification are shown in Table 10. Note that the bulk of the forest category is divided between two classes, FCHB and FMBH.

2.2.3.1.5 Grass

Because the evaluation of winter range is of specific interest for the Phase III test, it was thought to be of special value to determine the efficiency with which the grass category can be estimated by itself, with all the other land use classes lumped in a non-grass class. The experiment was also of general interest because it represents an extreme of the possible number of secondary classes. Results are shown in Table 11.

Correlation coefficients and F statistics are extremely high; however, the standard error is somewhat higher than in the previous cases where grass was estimated in conjunction with other classes. The reason for this effect is not clear at present. To verify that the high values are not the result of an induced correlation present when working with proportions (Chayes and Kruskal, 1966), a random set of proportions of identical size to those used for the grass category was generated and processed through program ESTPRP. The resultant correlation coefficient was 0.2676 and $\tilde{F}$ was 0.59, correctly reflecting the random nature of the input data.

It is interesting to inspect the P matrix for the ten spectral classes case. This matrix is shown in Table 12.

Most of the grass proportions is due to spectral class 8 (54.66%), which was initially identified on the classification image as wetland.

TABLE 10

Forest Types 3 Proportion
Estimates For Kershaw County

Land Use Class	7 Spectral Classes			
	Estimator: $\hat{a}_{pr}$		Estimator: $\hat{a}_{av}$	
	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error
FCLS	0.40	1.07	0.46	0.31
FCLB	2.67	1.15	2.45	0.69
FCHS	6.39	1.15	5.30	1.77
FCHB	24.12	1.19	24.02	2.18
FHLB	0.00	0.00	0.03	0.02
FHHS	0.00	0.00	0.01	0.01
FHHB	6.82	1.17	6.20	1.03
FMLS	0.08	1.13	0.15	0.11
FMLB	3.36	1.14	2.97	0.70
FMHS	0.32	1.10	0.34	0.19
FMHB	25.07	1.19	24.90	2.05
WL	5.77	1.09	5.78	1.25
DS	1.27	1.13	1.39	0.33
OT	23.68	1.10	25.99	2.78
R	0.7635			
$\tilde{F}$	11.71			
D.F.	(91,689)			

TABLE 11

Grass Proportion
Estimate For Kershaw County

Land Use Class	Estimator: $\hat{a}_{pr}$				Estimator: $\hat{a}_{av}$	
	7 Spectral Classes		10 Spectral Classes			
	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error
G	6.63	1.49	6.52	1.46	6.60	0.99
NG	93.32	1.49	93.42	1.46	93.40	0.99
R	0.9842		0.9828			
$\tilde{F}$	285.21		195.15			
D.F.	(7,53)		(10,50)			

TABLE 12

Grass P Matrix
(Percent)

Cluster Combination for Spectral Class	Secondary Class	
	G	NG
1, 12, 14	0.00	100.00
2	0.00	100.00
3	7.66	92.33
4	19.46	80.55
6	9.49	90.50
7, 9	2.80	97.21
8	54.66	45.33
10	25.33	74.68
11, 13	0.00	100.00

2.2.3.1.6 Water

The capability to distinguish water from non-water is also a major requirement for MAIS. A similar analysis as that for grass was therefore undertaken for water. Results are shown in Table 13.

Here the correlations are almost equal to unity, indicating an almost perfect correspondence with water as shown by the classification image and the aerial photo interpretations. This could be expected, as water is the land use class most discernible on LANDSAT images. However, the result demonstrates that a very good PSU-SSU registration was obtained.

In this case, unlike the analysis for grass, the low standard error does seem to reflect the high correlation. The estimator $\hat{a}_{pr}$ is clearly superior to $\hat{a}_{av}$. Again, Table 13 shows how LANDSAT introduces a global correction for the water proportion of Kershaw County, as contrasted with the proportion in the sample which is high because of an SSU located over the Wateree Reservoir.

2.2.3.2 Land Use, Step 2

Once the secondary proportions have been obtained, another class projection can be applied, and proportion estimates and corresponding covariance matrices for ground classes can be obtained using the formulation of Section 2.2.2.2.1. These are the estimates presented in the following sections.

Two ground classification proportion estimates were attempted, one for general land use classes, the other for more specific forest type-classes. Class designations are explained in Table 14.

2.2.3.2.1 Ground Land Use Classes

The results for the more general land use classes are shown in Table 15.

TABLE 13

Water Proportion
Estimate For Kershaw County

Land Use Class	Estimator: $\hat{a}_{pr}$				Estimator: $\hat{a}_{av}$	
	7 Spectral Classes		10 Spectral Classes			
	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error	Proportion (Percent)	Standard Error
W	1.86	0.11	1.87	0.15	3.24	1.56
NW	98.08	0.11	98.08	0.15	96.76	1.56
R	0.9994		0.9993			
$\tilde{F}$	21989.08		14707.36			
D.F.	(7,53)		(10,50)			

TABLE 14

Ultimate Land Use Classification
Definitions

Class	Land Use
CF	Commercial Forest
CR	Cropland
IP	Improved Pasture
IF	Idle Farmland
OF	Other Farmland
UR	Urban and Other
WT	Water

Class	Forest Types
LLP	Longleaf Pine
SHP	Slash Pine
LBP	Loblolly Pine
SLP	Shortleaf Pine
PDP	Pond Pine
OYP	Oak-young Pine
OHI	Oak-hickory
SCO	Southern Scrub Oak
OGC	Oak-gum Cypress
EAC	Elm-ash-cottonwood
NC	Not-commercial Forest

TABLE 15

Ground Land Use Class
Proportion Estimates
For Kershaw County

Estimator: $\hat{g}$, usir $\hat{a}_{pr}$		
Land Use Class	Proportion (Percent)	Standard Error
CF	72.28	3.63
CR	13.28	2.19
IP	4.74	1.74
IF	2.16	1.01
OF	2.69	1.40
UR	2.70	1.30
WT	2.03	1.13

TABLE 16

Ground Forest Types
Proportion Estimates
Kershaw County

Estimator: $\hat{g}$, using $\hat{a}_{pr}$		
Land Use Class	Proportion (Per Cent)	Standard Error
LLP	2.44	1.14
SHP	11.05	2.60
LBP	20.85	3.33
SLP	2.18	0.96
PDP	0.84	0.58
OYP	10.49	2.40
OHI	11.07	2.51
SCO	3.85	1.26
OGC	6.00	1.88
EAC	3.76	1.37
NC	27.04	2.92

The standard errors are generally larger than those obtained for the secondary classifications, because the estimates are computed from two sets of random variables.

2.2.3.2.2 <u>Ground Forest Type Classes</u>

The results for the forest types classification are shown in Table 16. The non-commercial proportion in Table 16 is the complement of the commercial forest category in Table 15. These figures total 99.32%, a good result considering that these tables were derived using separate processes and different groupings.

2.2.3.3 <u>CAI, Step 1</u>

The first step in arriving at estimates of continuous variables such as CAI is to select a suitable classification. This classification serves as a stratification for the second step. Either a secondary or an ultimate classification can be used.

To compare results, both types are tested. The forest types 3 classification (Section 2.2.3.1.3) and estimates in Table 10 are used for the secondary classification. The forest types of Section 2.2.3.2 (Table 14) and the estimates in Table 16 are used for the ultimate classification.

2.2.3.4 <u>CAI, Step 2</u>

The following kinds of estimates can be made for CAI (see Figure 2): by class, per acre; by class, total; by county, per acre; and by county, total. The per acre estimates by class are derived solely from the plot data. The "by class" estimates for the secondary classifications are presented in Table 17;

TABLE 17

CAI Estimates by
Forest Types 3 Class

(Cubic Feet)

Class	Per/Acre	Standard Error	Total	Standard Error	Percent
FCLS	30.0	0.0	60,135	160,539	266.
FCLB	21.4	7.3	285,862	156,683	54.8
FCHS	37.8	21.2	1,205,044	709,645	58.8
FCHB	93.9	10.2	11,322,034	1,350,470	11.9
FHLB	0.0	0.0	0	0	0
FHHS	0.0	0.0	0	0	0
FHHB	61.2	9.2	2,087,965	475,726	22.7
FMLS	0.0	0.0	0	0	0
FMLB	27.5	23.3	461,962	422,100	91.4
FMHS	0.0	0.0	0	0	0
FMHB	58.5	6.7	7,327,357	906,993	12.4
WL	60.1	15.3	1,733,242	548,222	31.6
DS	14.1	7.5	89,727	92,952	104.
OT	0.8	0.6	98,080	66,425	67.8
		COUNTY TOTAL	24,671,410	1,865,234	7.56

those for the ultimate classification, Table 18.

To ensure compatibility of the estimates, total estimates were made using the county acreage given in the U.S. Census Report of 1970. The same number was used by the Southeastern Experiment Station for its Forest Statistics of Kershaw County (Craver, 1978). A county area estimate was also obtained from the digitized boundary using program KCC. Another area figure was found in the Lockheed Ten-Ecosystem Study Final Report (Dillman, 1978). The different acreage figures are shown in Table 19.

TABLE 19

Kershaw County Acreage

Source	Acres
U. S. Bureau of the Census	499,840
Earth Satellite	501,283
Lockheed	503,100

Several other types of estimates for the total CAI were computed. All estimates are summarized in Table 20.

One estimate was made with $\hat{a}_{av}$ (no LANDSAT contribution) to obtain an idea of gain in efficiency due to the primary stage. This gain was estimated at (8.17 - 7.56)/8.17 x 100% = 7.5%.

The estimate obtained with the ultimate classification had a standard error approximately twice as large as the one computed with the secondary classification. This is due to the introduction of an additional set of random variables. Also, a stratification by species does not seem too meaninful when considering growth. A stratification by site class would be more

TABLE 18

CAI Estimates by
Ultimate Forest Type
Class

(Cubic Feet)

Class	Per/Acre	Standard Error	Total	Standard Error	Percent
LLP	22.1	11.2	269,645	185,421	68.7
SHP	98.7	18.7	5,449,095	1,648,118	30.2
LBH	94.4	8.9	9,834,346	1,823,546	18.5
SLP	81.0	22.0	880,818	457,549	51.9
PDP	94.0	27.0	392,686	295,973	75.4
OYP	40.1	7.3	2,101,352	613,955	29.2
OHI	46.7	7.6	2,582,602	720,295	27.9
SCO	11.2	2.3	215,796	83,500	37.7
OGC	45.0	8.6	1,350,130	495,670	36.7
EAC	73.0	14.0	1,373,063	563,622	41.0
NC	0.0	0.0	0	0	0
COUNTY TOTAL			24,449,534	3,858,145	15.78

TABLE 20

Total CAI Estimate Summary
Kershaw County

Type of Estimate	Total	Standard Error	Percent
Stratification with Secondary Forest Types 3 Classification using $\hat{a}_{pr}$	24,671,410	1,865,234	7.56
Same with $\hat{a}_{av}$ (without LANDSAT)	24,134,085	1,970,590	8.17
Stratification with Ultimate Forest Types using $\hat{a}_{pr}$, $\hat{g}$	24,449,534	3,858,145	15.78
Plot data used as Simple Random Sample	26,474,90	8,749,250	33.05
South Carolina '78 Forest Statistics	24,435,000	1,771,540	7.25

appropriate.

To compare the estimates obtained with $\hat{a}_{av}$, g and $\hat{a}_{pr}$, which are based on the GIS technology, with a simple random sampling estimator, a fourth estimate was computed from the plot data by averaging the CAI's for each plot and multiplying this average with the county area. The standard error of this estimate is approximately four times the one obtained with the secondary forest types 3 stratification, thus dispelling any doubt that the employed technology does contribute to the sampling efficiency for estimating current annual increment.

The final estimate in Table 20 is the one reported in the publication, "Forest Statistics for the Northern Coastal Plain of South Carolina" (Craver, 1978): Table 8, column 6, Kershaw County. The corresponding standard error (7.25%) was obtained from the table on page 5 of the same publication. This figure supports the hypothesis that Phase IA estimates made with the prototype MAIS system are of the same quality as those obtained with current practices.

2.2.3.5 Soil Erosion, Step 1

The derivation of the secondary classification for erosion potential is given in Section 2.2.2.1.5. A sample set of 12 SSU's was created with seven erosion potential classes. This set was input to program ESTPRP, together with the 7 class spectral classification proportions. The estimated proportions in each erosion potential class are shown in Table 21.

The correlation coefficient is high, but the F statistic barely satisfies the four times critical value criterion of Draper and Smith (1966), at the 0.05 level (8.39 = 4.69 x 1.71). These effects are due to the small sample size. Given that enough coefficients are estimated in relation to the number

TABLE 21

Erosion Potential Class
Proportion Estimates
For Kershaw County

(Percent)

7 Spectral Classes				
Erosion Potential Class	Estimator: $\hat{a}_{pr}$		Estimator: $\hat{a}_{av}$	
	Proportion	Standard Error	Proportion	Standard Error
ER_1	54.95	4.33	58.43	5.47
ER_2	8.77	3.81	6.01	1.66
ER_3	24.99	3.96	27.30	5.33
ER_4	6.51	3.92	4.66	1.79
ER_5	4.25	3.92	3.27	1.87
ER_6	0.08	3.22	7	0.05
ER_7	0.38	3.22	0.28	0.11
R	0.9502			
$\tilde{F}$	8.39			
D.F.	(42,30)			

of observations, a correlation coefficient can always be forced to unity. As shall be seen in the next section, this situation is reflected in a large sampling error for the county total.

2.2.3.6 Soil Erosion, Step 2

The class midpoint was assigned as the continuous variable for each erosion potential class. These per acre estimates, as well as the total estimates by class and the total estimates for Kershaw County, are shown in Table 22. Both the estimators $\hat{a}_{av}$ and $\hat{a}_{pr}$ were tested.

The difference between the county total standard errors for the estimators using either $\hat{a}_{pr}$ or $\hat{a}_{av}$ (LANDSAT regression and area photo averages) is striking. The difference is due to the high standard errors of the high potential erosion classes for the $\hat{a}_{pr}$ estimator as indicated in Table 21. It is another sign of the marginal performance of regression techniques using an inadequate sample. In this situation, it seems that one is better off using the secondary sample proportions alone. It is hoped that the same test can be performed during Phase II with a larger sample size. It is also possible that $\hat{a}_{rs}$ may perform better than $\hat{a}_{pr}$ in the case of a limited number of of observations. The total county estimates for both techniques conform closely.

2.2.3.7 Map Legends

One of the unique aspects of the ICLS regression approach is that the estimated class transformation matrix can be used to construct a map legend of the primary map in terms of a secondary classification. Table 23 shows the P matrix to transform from seven spectral classes to ten secondary land use classes. Using this matrix and a convention of reporting the three major

TABLE 22

Erosion Potential Estimates
For Kershaw County

(Tons/Year)

Erosion Potential Class	Per/Acre	Estimator $\tilde{g}$, using $\hat{a}_{pr}$			Estimator: $\tilde{g}$, using $\hat{a}_{av}$		
		Total	Standard Error	Percent	Total	Standard Error	Percent
ER_1	1	274,686	21,631	8.0	292,035	27,342	9.0
ER_2	3	131,551	57,096	43.0	90,071	24,867	28.0
ER_3	6	749,552	118,725	16.0	818,637	160,329	20.0
ER_4	12	390,645	235,377	60.0	279,260	107,401	38.0
ER_5	24	510,059	470,754	92.0	392,374	223,972	57.0
ER_6	48	20,096	772,579	3844.	15,995	11,040	69.0
ER_7	96	180,215	1,545,158	857.	133,557	51,522	39.0
COUNTY TOTALS		2,256,805	1,486,338	66.0	2,021,931	337,965	17.0

TABLE 23

Class Transformation Matrix

	FC	FH	FM	AG	WL	WT	GR	UR	DS	BS
ALN	1.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B	0.5807	0.0000	0.2082	-0.0000	0.1888	0.0000	0.0000	0.0000	0.0224	-0.0000
C	-0.0000	0.1958	0.4050	0.2250	0.0426	0.0086	0.0889	0.0000	0.0322	0.0020
D	0.6834	0.0193	0.2907	0.0000	-0.0000	0.0000	0.0067	-0.0000	0.0000	0.0000
EGIJ	0.1107	-0.0000	0.1384	0.5049	0.0070	0.0000	0.1715	0.0669	-0.0000	0.0007
F	0.0000	0.2439	0.5939	-0.0000	0.0000	-0.0000	0.1013	0.0551	0.0059	-0.0000
HKM	-0.0000	0.0000	0.1079	0.0000	0.0000	0.8922	0.0000	0.0000	-0.0000	-0.0000

Cluster Composition for Spectral Classes:

ALN	1, 12, 14
B	2
C	3
D	4
EGIJ	5, 7, 9, 10
F	6
HKM	8, 11, 13

secondary classes for each primary class, one can assign a legend to each cluster color on the primary classification map, as shown in Table 24. This legend is for the LANDSAT classification map (also see Figure 7) submitted as a deliverable product for this project. To make this legend, a cluster-color assignment table was used. Twelve colors are shown on this map. Since only seven spectral classes are used in the analysis, a simplified map of seven colors could be made with a simpler legend. The legend of Table 24 is only a preliminary product and with a possibility for feedback, a much better map legend can be produced.

TABLE 24

Legend for Landsat
Classification Map

Color	Land Use Composition
Brown	100% Forest Conifer
Dark Green	58% Forest Conifer, 21% Forest Mixed, 19% Wetland
Light Green	40% Forest Mixed, 23% Agriculture, 20% Forest Hardwood
Orange	68% Forest Conifer, 29% Forest Mixed
Red	51% Agriculture, 17% Grass 14% Forest Mixed, 11% Forest Conifer
Purple	60% Forest Mixed, 24% Forest Hardwood, 10% Grass
White	See Red
Light Blue	89% Water
Black	See Red
Light Blue	See Red
Blue	89% Water
Yellow	See Brown
Blue	89% Water
Grey	See Brown

3.0 SUMMARY

A prototype system test was undertaken to asses the workability of the proposed MAIS design. The most important aspect in assembling a system from a set of components is the integration of the components into a workable entity. It was realized that for MAIS, the vital link in this process is the "estimation subsystem", and hence the evaluation effort was directed at trying the proposed techniques for this subsystem in a set of preliminary tests for one county.

The results seem to support the notion that the basic scheme works very well. Estimates which heretofore were impossible to make using conventional methods can be made using GIS and LANDSAT technology (erosion potential). Conventional estimates can be made with the same accuracy as current methods, hopefully at reduced cost. The proposed techniques seem to be both robust and flexible. High LANDSAT classification accuracies are by no means required, and using the class transformation concept, one can produce a wide variety of estimates to satisfy many needs. Valuable insight was gained into a possible structure for a permanent estimation subsystem. An automatic file handling system along the lines of the transaction concept outlined in the concept development document is highly recommended. Also, an automated report generator must be included in a future estimation subsystem.

In conclusion, it seems that the major reservations concerning the proposed techniques have been eliminated. Some problems remain due to time and resource constraints. It is hoped that they can be addressed in the Phase II effort.

APPENDIX A

ES 1211

NATIONWIDE FORESTRY APPLICATIONS PROGRAM

MULTIRESOURCE INVENTORY METHODS
PILOT TEST, Phase I

EVALUATION OF MAIS PROCESSING COMPONENTS

(TYPE II REPORT)

April 18, 1980

Earth Satellite Corporation
Forestry Division
Berkeley, California 94704

Prepared for:

USDA-FOREST SERVICE
LYNDON B. JOHNSON SPACE CENTER
Houston, Texas 77058

EARTH SATELLITE CORPORATION *(EarthSat)*
7222 47th Street, N. W., Washington, D. C. 20015
2150 Shattuc' ..enue, Berkeley, California 94704
(301) 652-7130 *(415) 845-5140*

MULTIRESOURCE INVENTORY METHODS PILOT TEST - EVALUATION OF MAIS PROCESSING COMPONENTS

Table of Contents

1.0 GENERAL APPROACH

In Volume I (Multiresource Inventory Design and Sampling Network) and in Volume II (MAIS Concept Development) a number of techniques have been proposed for use on the Pilot Test Implementation. These techniques are at many different stages of development. Some have been widely applied in the past, so their value and limitations are well-documented in at least some geographic areas and application areas. Others are proposed that have not, to our knowledge, been applied before in the context of a multiresource inventory.

In this volume, we address two questions:

- What is the level of maturity of each technique that will be employed in the Pilot Test Implementation? This includes experience with the method, the way that prior experience can be related to the proposed use on the Pilot Test, and the way that the technique has been embodied in available software and hardware.
- How far does the state-of-the-art in each component used on the Pilot Test match the perceived needs of the project? This implies a comparison of requirements against capabilities, and this comparison must be at least partially a subjective one.

Following a component-by-component analysis that examines the two questions given above, research and development needs can be stated. When the state-of-the-art falls short of the needs, the work required to improve that state-of-the-art can in many cases be identified. Thus, component evaluation leads naturally to suggestions for areas of development that should be pursued by research activities in support of performance of the Pilot Test. These research activities are themselves outside the scope of the Pilot Test Implementation; realistically, their results may not be available in time to be useful in the South Carolina Test. There is a possibility that new techniques can be developed to apply to the Western State Test. In addition to identifying research areas and giving subjective assessments of the magnitude of the effort needed, we will also attempt to scope the length of time that may be needed to convert research work to a useful new tool for real inventories.

2.0 SYSTEM REQUIREMENTS

Two sets of elements must work together if the Pilot Test is to be performed successfully. These consist of the data sources and the processing components that manipulate them, where the latter will embody the mathematical models. The processing components must contain a capability to handle all major data types. These data types consist of:

- Landsat data, in image and CCT format
- NCIC digital terrain data in tape format
- Collateral data in map format
- Aerial photography, in resource photography or optical bar format
- Attribute (point) data

The processing components that manipulate these data types consist of the following:

2.1 Upper Level GIS Components

INPUTS	COMPONENT	OUTPUTS
A. Old Format Landsat Data	Preprocessor	Band Sequential, Decalibrated, Synthetic Pixels Deleted, Data Drops Filled
B. Output of A	Scan Line Suppression	Same Format, Scan Line Removed
C. New Format Landsat P Tape, Band Sequential	Preprocessor	Band Sequential Landsat, Data Drops Filled

INPUTS	COMPONENT	OUTPUTS
D. Output of B or C User Input UTM Base Maps	GCP Location	List of Control Points
E. Outputs of B & D	Landsat Specific Geometric Correction	Landsat Band Sequential in 50 Meter UTM
F. NCIC Digital Terrain File	Preprocessor	Elevation File, Map Control Points
G. Output of C & D Or F	General Map Transform	Input File at UTM 50 Metre
H. Elevation File From G	Slope Calculation	Slope Class File
I. Elevation File From G	Aspect Calculation	Aspect File
J. Map Input	Digitizer Interface	Line Segment Coordinate Chains
K. Output of J User Inputs Polygon Attributes, Control Points.	Arc-Cell Convert	50 Metre UTM Cell File of Map Input
L. Outputs From E, G, H, I, K	Map Unit Extraction	Map Unit Files in Data Base Format
M. Multiple Data Layers # of Classes Desired	Cluster	Class Map Class Means Class Variance
N. Multiple Data Layers Training Areas	Unsupervised Classification	Class Map

	INPUTS	COMPONENT	OUTPUTS
O.	Multiple Data Layers	Boolean Combination	Resultant Data Layer
P.	Multiple Data Layers, Combination Coefficients	Linear Combination	Resultant Data Layers
Q.	Single Layer, Or Single Layer Plus Boundary Layer	Area Tabulation	Acreage Totals for Classes for Entire Layer or Within Selected Boundaries
R.	Multiple Layers	Statistical Processor	Statistics (Mean, Variance, Covariance Cross-Correlation)
S.	Single Layer, Desired Scale Color Assignments	Map Processor	Chloropleth Map of Data Layer

2.2 Lower Level GIS Components

	INPUT	PROCESS	OUTPUT
A.	Arc Extraction		
	Maps, Aerial Photos (Optical Bar)	1. Thinning	Digitized Arcs in Standard Format
		2. Conversion to UTM Coordinates	
		3. Conversion to Standard Format	
		4. Scale Change and Adjustment	
		5. Photogrammetric Adjustment	

INPUT	PROCESS	OUTPUT
B. Arc Polygon Conversion		
Arcs in Standard Format	1. Converting Arcs-Polygons 2. Area Calculation	Polygons and Areas in Standard Format
C. Attribute Data Entry		
Attribute Data From Forms and Maps	1. Interactive Attribute Entry 2. Bit Packing of Data Item Values	Attribute Records in Standard Format
D. Slope/Aspect Generation		
Contour Line Areas In Standard Format	1. Generation of Slope and Aspect Maps	Slope and Aspect Polygons in Standard Format
E. Data Base Insertion		
Arc and Polygon Records in Standing	1. Insert Ready Images Into Data Base	Data Base
F. Data Base Retrieval		
Data Base	1. Obtain SSU Images From Data Base	SSU Image Records In Standard Format
G. Command Language		
Analyses and Retrieval Problem Formulation	1. Compile Routine Executable Code	Executable Code
H. Attribute Search		
SSU Image Records	Filtering of SSU Image Subareas for Selected Attributes	SSU Image Records With Desired Attributes

INPUT	PROCESS	OUTPUT
I. Zone Generation		
SSU Image Records With Desired Attributes	Generation of Zones Around SSU Sub-Units	SSU Image Zone Records
J. Overlay		
SSU Image Records	Boolean Combination of SSU Layer Images	SSU Image Records, Derived Layers
K. Report Generation		
SSU Image Records	Retrieval and Sorting of SSU Sub-Unit Attributes	Reports, Attributes for Correlation with Upper Level GIS Information
L. Map Generation		
SSU Image Records	1. Spatial Display and Assembly of SSU Images	Maps for Visual Inspection Screen Displays
M. Updating		
SSU Images Data Base	Process SSU Changes Against Existing SSU Images to Obtain SSU Images	Updated Data Base SSU Images

2.3 Additional Project Components

A. Sample Assignment

Provides the distribution of data samples needed for estimation of resource variables.

B. Map Regression Relationships

Develops the correlations between aerial photo and Landsat data for both continuous and non-continuous variables.

C. Landsat Mapping Extension

Permits the use of Landsat data as a mapping tool that can extend resource information to areas not covered by aerial photos and ground samples.

D. Photointerpretive Keys

Provides the set of interpretive keys that allow Landsat scenes to be classified to multiple resource classes via manual photographic analyses.

3.0 WORKING STATUS OF GIS COMPONENTS

Not all components listed in the previous section have reached the same level of maturity in either concepts or in practical implementation. In this section, a subjective evaluation is offered of the . of development of each component. In Section 4 the useful research activities that may apply to certain selected system components are described.

3.1 Upper Level GIS Components

3.1.1 Working versions of components A through G are known to be available, although the capability to fill in dropped data lines from Landsat data have not been added to them. (Programs exist for DEC PDP computers at Goddard Space Flight Center, and for IBM and PRIME computers at Earth Satellite Corporation; other versions also undoubtedly exist at ERIM.)

3.1.2 Working versions of the slope and aspect calculations (H and I) exist, though not in the format conceived for use in the Pilot Test. Minor modifications of the presently working software will take care of this (programs exist on the AOIPS system at Goddard Space Flight Center).

3.1.3 Working versions of components J and K exist for the PRIME computer, and also in the case of component K for the IBM 360 computers (see CACM, 22, 518; 1979).

3.1.4 Working versions of component L exist on several different computers, including the IBM 360/158 and sister machines (Purdue University, Johnson Space Center).

3.1.5 Working versions of components M and N exist at a wide variety of installations, and with a variety of options (Bayes estimation, maximum likelihood, binary classifiers, etc.). However, there is a definite lack of information about reliability of classification results in mixed-resource environments. This is particularly true if collateral variables are included in the classification process.

3.1.6 Working versions of O. exist for the IBM, DEC, Honeywell, and UNIVAC equipment.

3.1.7 Working versions of P through S exist on the PRIME, DEC, and IBM 360 computer systems.

SUMMARY: The functions of the upper level GIS present no real problems for implementation of Phase II of the Pilot Test. All components have been developed already. The major uncertainties arise from the question of accuracy of results obtained in some components.

3.2 Lower Level GIS Components

All components of the lower level GIS exist in FORTRAN in working form on at least one computer system (PRIME).

3.3 Additional Project Components

3.3.1 A. Sample Assignment

This does not currently exist in the form that will be needed for use on the Pilot Test program. Methods are clear, but code must be developed.

3.3.2 B. Map Regression Relationships

A good deal of development work is still needed, for both the details of the methods and for the programming of the meth.ds. This will require some added thought and development during Phase II of the project.

3.3.3 C. Landsat Mapping Extension

This also needs development. The proposed method for use on Phase II of the Pilot Test has not been used before. The programs must be developed and applied in the course of the Phase II implementation.

3.3.4 D. Photointerpretive Keys

The procedure for PI Key development is well known, but its application to the specific environment of South Carolina has not been made in the context of multiresource survey. Most of the tools required for the application of these keys to Landsat data in South Carolina will be provided by the results of Phase I of the project, but there will still be elements to be looked at further during Phase II implementation.

4.0 RESEARCH AND DEVELOPMENT CONSIDERATIONS

This section presents a partial list of research topics that could provide a significant contribution to the effectiveness of the MAIS. They are in no particular order of priority, because it is difficult to determine in advance just how valuable they will be if successfully carried out and incorporated into the MAIS.

These activities are not appropriate for inclusion in the test itself, since there is no way of guaranteeing that results useful to test performance can be derived in time. Note also that no allowance for the cost of these items is given in Volume IV, nor is any calendar time devoted to them. Task 19 of Volume IV serves a different function. It is intended exclusively to provide the effort needed to monitor and remain aware of on-going activities of possible use to Pilot Test activities.

4.1 Computerized Classification Algorithms

Supervised and unsupervised algorithms for multispectral data have been applied predominantly in agricultural experiments, and have proved most successful in dealing with large, regular, single crop areas, with gentle terrain. They have been much less successful in dealing with agriculture where small fields are common, and least successful in dealing with cases where there is considerable variation in altitude and aspect within a scene.

Useful accuracy figures for computerized multispectral classification methods are hard to come by. There appears to be no specific experience that will be directly useful in telling how effective the classification algorithms will be in the South Carolina Pilot Test.

This is certainly a case where application of some research and development effort is worthwhile, to provide some quantitative measures of accuracy, and to show the best way to combine Landsat data and collateral data for that purpose. In particular, the sensitivity of classification accuracies to misregistration of multiple coverage needs to be looked at in some systematic way for application to forested scenes.

4.2 Photogrammetric Reduction of Aerial Photography

This has been done for many years with conventional resource photography. The high resolution color I/R photography obtained from the optical bar camera, with its wide angle and significant variation in aspect angle across the image, is another matter. There is little or no experience that shows how effective this source will be in the multiresource mapping and inventory problem.

Development effort for assessment of the use of optical bar data would appear to be well worthwhile.

4.3 Regression Relationships Between Landsat Data, Aerial Photography, and Ground Samples

The use of regression relations of this type is not new (see Volume I), but the way in which it is proposed to apply the method to the Pilot Test does appear to be novel and so far untried. A key factor will be the correlation potential of Landsat with the interpreted aerial photography. Unless that correlation is reasonable, the method for using Landsat in either the resource estimation or the mapping mode is of doubtful value.

Although this component has not been previously evaluated, it is difficult to see how any independent evaluation outside the bounds of the Pilot Test would be of much use in contributing to the Pilot Test. Therefore, although this is an untried area, the recommendation here is to apply the technique in the Pilot Test and monitor closely the results. Although this is certainly a research area, the research will be conducted in performance of this project.

4.4 The Use of Landsat Data in a Mapping Mode

The use of Landsat in the resource inventory mode can be thought of as a form of data averaging, since in that case the objective is aggregate figures for different resources. The mapping mode, however, makes much greater demands on the data. In this mode, extrapolation from areas of known resources to areas of unknown resources is attempted via the use of Landsat. This process is more sensitive than an averaging procedure, and is more likely to lead to errors of both omission and commission. Additional research on the proposed method is definitely needed.

4.5 Determination of Resource Variables - Current Annual Increment

This variable is not estimated by any of the conventional techniques of remote sensing. However, by regarding timber as a crop which responds to the same environmental variables as any other crop, there is a potential for modeling the "yield" of timber year by year in just the same way as crop yields can be estimated. This approach is the subject of the document "MAIS - Multi-Resource Analysis and Information System Research and Development Component Requirements Discussion for Dynamic Factors" (Earl S. Merritt, April 1980) where several dynamic aspects of the forest environment and resources are discussed. These include:

- Mean annual increment
- Erosion
- Forest fire prevention and management
- Range stocking
- Watershed factors
- Disease and insect vector propagation.

The use of such modeling methods needs to be looked at in much more detail before the potential can be evaluated. It is a suitable subject for research and development in connection with the MAIS. If successful, it would require the addition of certain dynamic components to the presently conceived MAIS.

4.6 Change Detection, Identification and Measurement Using Landsat Data

Experiments in the Pacific Northwest show that change detection, particularly for clear-cut areas, is very feasible using Landsat as a data source. The minimum size of areas that can be monitored in this way, however, is of the order of 10 acres (though others have reported limited success at change detection of as little as a single Landsat pixel). Change identification is more difficult, and more work is needed on it. Measurement is relatively straightforward if the changed categories are not subject to confusion, but here also there is need for research work to determine the practical limits of what can be done with Landsat. No experience base exists for change detection, identification and measurement using Landsat in the South Carolina environment.

4.7 Multiple Resource Keys

The use of PI Keys for forestry using remotely sensed data is well established. Less well-established is the development of multiple resource keys, where several resources may be contained in a single location, and where each resource may call for a completely separate resource map (for example, the geographic species distribution of the understory may be quite different from the species distribution of the overstory). Work on these multiple resource keys is going on in Asheville, and additional efforts to provide good keys in the Southeast may be appropriate as part of the Pilot Test associated research efforts.

APPENDIX B

Photo Interpretation (PI) Classes and Codes for Multiresource Methods Pilot Test, Phase 1A

Class Name	EarthSat PI Code
Forest	
1. Conifer	F1 i j k
2. Hardwood	F2 i j k
3. Mixed	F3 i j k

Where: i is an index of size of the dominant trees,
- 1 = sappling size and smaller
- 2 = pole size
- 3 = sawtimber

j is an index for tree density (% crown cover),
- 1 = 0 - 5%
- 2 = 6 - 15%
- 3 = 16 - 35%
- 4 = 36 - 65%
- 5 = 66 - 100%

k is a background component, an explanation for the remaining ground cover,
- 1 = pine understory
- 2 = hardwood understory
- 3 = mixed pine-hardwood understory
- 4 = grass or herbaceous understory

Class Name	EarthSat PI Code
Agriculture	
1. Idle farmland	A1
2. Irrigated cropland	A2
3. Non-irrigated cropland	A3
4. Other farmland	A4
Wetland	
1. Permanent, trees	S1
2. Permanent, other cover	S2
3. Intermittent, trees	S3
4. Intermittent, other cover	S4
Water	
1. Flowing, census	W1
2. Flowing, non-census	W2
3. Contained, census	W3
4. Contained, non-census	W4
Grass	
1. Natural rangeland	G1
2. Improved pasture	G2
Urban (developed or industrial)	U
Disturbed	
1. Regenerative	D1
2. Non-generative	D2
Brush	B
Rock	R
Other	T

REFERENCES

1. Chayes, F. and W. Kruskal. "An approximate Statistical Test for Correlation Between Proportions," Journal of Geology, Vol. 74, No. 5 (1966), pp. 692-702.

2. Cochran, W. G. Sampling Techniques. New York: John Wiley and Sons, 1963.

3. Colwell, R. N. "Remote Sensing of Natural Resources--Its Basic Concepts in Retrospect and Prospect." Proceedings of Remote Sensing for Natural Resources, Moscow, Idaho, 1979.

4. Dantzig, G. B. and R. W. Cottle. "Complementary Proof of Mathematical Programming," Studies in Optimization, Vol. 10. Washington: Mathematical Association of America, 1974.

5. ______________________________. "Positive (Semi-)Definite Matrices and Mathematical Programming," Nonlinear Programming, North Holland Publishing Company, 1967.

6. Dillman, R. D. "Ten-Ecosystem Study (TES), Site V, Kershaw County, South Carolina." Nationwide Forestry Application Program, Lockheed Electronics Co., Inc., System and Service Division, Prepared for Earth Observations Division, NASA, LBJ Space Center, Houston, Texas, 1978.

7. Dissmeyer, G. E. and R. F. Stump. "Predicted Erosion Rates for Forest Management Activities and Conditions Sampled in the Southeast." U. S. Department of Agriculture, Forest Service, April, 1978.

8. Gentle, T. E. and W. T. Kennedy. "Algorithms for Linear Regression with Linear Restriction." Proceedings of Computer Science and Statistics: 12th Annual Symposium on the Interface, University of Waterloo, Waterloo, Ontario, Canada, 1979.

9. Goodman, L. A. "A Further Note on Miller's 'Finite Markov Processes in Psychology'," Psychometrika, Vol. 18, pp. 245-248.

10. Graves, G. C. "Forest Statistics for the Northern Coastal Plain of South Carolina." U. S. Department of Agriculture, Forest Service Resource Bulletin SE-47, 1978.

11. Hildebrandt, G. "Remote Sensing in German Forestry: Practical Applications and New Research Results." Proceedings of Remote Sensing for Natural Resources, Moscow, Idaho, 1979.

12. Isaacson, D. L., C. J. Alexander, B. J. Schrumpf and R. Murray. "Analysis of Association of LANDSAT Spectral Classes with Ground Cover Classes in Wildland Inventories." Proceedings of Remote Sensing for Natural Resources, Moscow, Idaho, 1979.

13. Judge, G. G. and T. Takayama. "Inequality Restrictions in Regression Analysis," Journal of American Statistical Association, Vol. 61 (1966), pp. 166-181.

14. Langley, P. G., R. C. Aldrich and R. C. Heller. "Multi-Stage Sampling of Forest Resources by Using Space Photography--An Apollo 9 Case Study." Proceedings of the Second Annual Earth Resources Aircraft Program, Volume 2: Agricultural, Forestry and Sensor Studies, 1969.

15. Lemke, C. E. "A Method of Solution for Quadratic Programs," Management Science, Vol. 8, No. 4, 1962.

16. Liew, C. K. "Inequality Constrained Least Squares," Journal of the American Statistical Association, Vol. 71, No. 355, 1976.

17. _________ and J. K. Shim. "A Computer Program for Inequality Constrained Least Squares Estimation," Econometrica, Vol. 46, No. 1, 1978.

18. Mayer, K. E., L. Fox III, and J. L. Webster. "Forest Condition Mapping of the Hooper Valley Indian Reservation Using LANDSAT Data." Proceedings of Remote Sensing for Natural Resources, Moscow, Idaho, 1979.

19. Murthy, M. N. "Sampling Theory and Methods." Calcutta: Statistical Publishing House, 1967.

20. O'Reagan, W., Personal communication, 1980.

21. Pielou, E. C. An Introduction to Mathematical Ecology. New York: John Wiley and Sons, 1969, pp. 32-36.

22. Ritter, P. and A. Kaugars. CLUSTER, User's Guide, Version 1.0. Remote Sensing Research Program, Department of Forestry and Conservation, University of California, Berkeley, California, 1978.

23. Sader, S. A. "Methods of Obtaining Multiresource Information from Remote Sensing and Auxiliary Data Sources for Resources Assessment in Costa Rica." Proceedings of Remote Sensing for Natural Resources, Moscow, Idaho, 1979.

24. Telser, L. G. "Least Squares Estimation of Transition Probabilities," Measurement in Economics. Stanford: Stanford University Press, 1963.

25. Todd, W. J., D. G. Gehring, and J. F. Haman. "LANDSAT Wildland Mapping Accuracy." Photogrammetric Engineering, Vol. 66, No. 4, 1980.

26. "Universal Soil Loss Equation." Technical Notes, Resource Conservation Planning-WY-1, U. S. Department of Agriculture, Soil Conservation Service, Casper, Wyoming, January 1, 1978.

27. van Roessel, J. W. "Machine Estimation of Timber Volumes for Use in Sampling Surveys--A Method for High Flight and Space Imagery, Interface Considerations and Results." Proceedings. Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, 1976.

28. ________________. Unpublished notes. Earth Satellite Corporation, Berkeley, California, 1974.

29. Zarkovic, S. S. "On the Efficiency of Sampling with Varying Probabilities and the Selection of Units with Replacement," Biometrica, Vol. 3, 1964.

30. Zelner, A. Introduction to Bayesian Inference in Econometrics. New York: John Wiley and Sons, 1971.