

Reprinted from

**Symposium on
Machine Processing of
Remotely Sensed Data
and
Soil Information Systems
and
Remote Sensing and Soil Survey**

June 3-6, 1980

Proceedings

The Laboratory for Applications of Remote Sensing

Purdue University
West Lafayette
Indiana 47907 USA

IEEE Catalog No.
80CH1533-9 MPRSD

Copyright © 1980 IEEE
The Institute of Electrical and Electronics Engineers, Inc.

Copyright © 2004 IEEE. This material is provided with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the products or services of the Purdue Research Foundation/University. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

BULK PROCESSING TECHNIQUES FOR VERY LARGE AREAS: LANDSAT CLASSIFICATION OF CALIFORNIA

WILLARD NEWLAND

Technicolor Graphic Services, Inc.

DAVID PETERSON AND SUSAN NORMAN

NASA/Ames Research Center

I. ABSTRACT

The passage of California forestry bill AB452 created a need to assess and analyze multiresource forest data on a statewide basis. The California Department of Forestry (CDF), in responding to part of this need, formed a cooperative effort with Ames Research Center (ARC) and the Jet Propulsion Laboratory (JPL) to prepare a land-cover map of the entire state, emphasizing forest types. This map was produced in less than 1 year, as required, by combining the latest techniques in digital image mosaicking at JPL with the high-speed processing capability available on the ILLIAC IV parallel processor and other computer systems at ARC. An operational and very responsive analysis method was developed at ARC that permitted on-time response to weekly workshops conducted with CDF field personnel to identify all 1,200 spectral classes and to produce final products. Over 100,000,000 acres were classified in the period between December 1, 1978, and April 15, 1979. All analyses were conducted using existing software.

II. INTRODUCTION

The continual improvement in classification capability and in the speed of processing, the emergence of planetary mosaicking technology applied to Landsat data, and the concurrent growth in need for statewide data have converged, permitting the California Department of Forestry (CDF) and Ames Research Center to cooperatively develop a land-cover classification of California (100 million acres). The law that mandated CDF to assess California's forest lands provided the impetus to complete digital mapping in a stipulated period that imposed severe but realistic time constraints on performance.

Techniques were constructed to manage and process an enormous data set in a bulk or "pipeline" environment. The combined requirements that the data be statewide and that the mapping be completed in less than 1 year provided the goals to develop responsive techniques from existing technology. As is true of most emerging technologies, perceived needs are far ahead of present capabilities. This

U.S. Government work not protected by U.S. copyright.

stress has helped to improve the convergent techniques and test them in a real case, utilizing user expertise and performing within tight time limits. The method was synthesized from the diverse computer capability at ARC that needed only the driving force of a large demanding effort to resolve the difficulties. This paper focuses on the bulk processing aspects of this project, which reveal a number of important characteristics of interest to users of large-scale programs.

III. BACKGROUND AND NEEDS OF CDF

California Law AB452¹ was passed in 1977 to provide a mandate for CDF to design and implement an information system to assess the forest land base for multiple uses and values. Policy-makers in California have had to rely, in the past, on data gathered primarily by federal agencies, mostly on commercial timberland, and have lacked an independent source to verify and complete the information on all forest lands. AB452 authorizes the implementation of a data storage and retrieval system, to be managed by CDF, to receive both cooperative federal data and other data available on lands not adequately assessed. CDF will conduct assessments of the supply and availability of the present and potential resource values associated with these opportunities: (1) productivity and reforestation; (2) wood fiber utilization and recycling; (3) salvaging of dead and diseased trees; (4) forest wildlife and wildlife habitat; (5) recreation; (6) forest range; and (7) wood fiber as a source of fuel.

Although numerous mapping efforts have been started in California since the 1880s, many were never finished.² For various reasons, a detailed forest map of the state was likely to be incomplete out-of-date, or otherwise unusable. The poorest information deals with "nonproductive" forest types, such as the extensive hardwood areas throughout the state and the nonmerchantable conifers. Even an up-to-date assessment of the condition of understocked or unstocked commercial forest land has not been adequate. CDF perceived the urgent need for a more detailed, even if general, map covering the entire state; the area comprises 100,000,000 acres,

of which some 40 million acres are forest land. Without the time or funding to produce such a map from aerial photography, CDF recognized the potential of combining the latest Landsat technology in classifying and mosaicking with the experience of CDF field personnel from throughout the state to produce a land-cover map of the state. The results of this effort would be used for planning, for developing an information base, and for conducting future detailed studies as part of an efficient assessment of the resource values.

IV. PROJECT PLANNING AND AVAILABLE RESOURCES

A. GENERAL DESCRIPTION OF THE PROCESSING PLAN

An extremely limited time period and the necessity to process data for the entire state were the basic guidelines that determined the approach to use, and the need to quickly establish a "pipeline" throughput from existing capabilities. The process consists of 14 general steps with many repetitious subelements that must be integrated carefully in time and maintained to avoid serious loss of time and conflicts. The flowchart (Fig. 1) details the subelements; general categories are outlined as dashed lines. These steps — data acquisition, mosaicking, digitizing, calibration, reformatting, data compression, clustering, editing, classification, identification, stratification, generation of map and acreage results, and verification — are briefly described in this paper. Important features are pointed out for each as they pertain to bulk processing. The classification was strictly from unsupervised clustering and was conducted separately for 32 major ecological regions in the state. Ames Research Center performed the classification work, and JPL provided the mosaicked data.

The schedule (Fig. 2) reflects primarily the pacing effect of the mosaicking. The mosaic was divided into five major parts. The portion within a particular 1°-latitude by 1°-longitude segment was created as one data tape and delivered from JPL to ARC in increments as shown.

The other critical pacing item was the workshops at ARC, attended by CDF field personnel, to identify the class assignment for each unsupervised spectral class. The steps from reformatting through classification had to be completed after receipt of raw data to maintain the workshop schedule. The workload peaked in April (workshops ending on April 20) leaving enough time to process final results before the May 15 CDF deadline. All summaries and the final report were completed by May 15; on June 15, a statewide map was delivered to CDF for their final report. Adherence to this extremely ambitious schedule illustrates the use of Landsat data in providing responsive and useful results over very large areas.

B. COMPUTER SYSTEMS USED

An eclectic design that could access the most efficient and reliable components of eight computer

software systems available at ARC was developed. Some of these systems provided redundancy to overcome the inevitable downtime. No one system at ARC could have maintained the schedule. The time demands did not permit major software development, recoding, or restructuring of hardware. From a diverse system capability, a design was drawn from existing techniques and algorithms, coupled through a single data link with batch support and peripheral equipment as needed. The backbone of the design is the EDITOR software³ implemented by Ray from original LARS algorithms, modified to be executed on the ILLIAC IV via the ARPA network.⁴ Although some ARC systems have sufficient analysis routines, it was the judicious combination of high-speed processing with other capabilities that was able to maintain the pace. Table 1 lists the computer systems used and the major functions performed by each. From this project, at least a preliminary framework has been formed to evaluate the appropriateness and utility of other systems or developments.

Several advantages accrue from a multiple-system environment. The memory and access characteristics of overlapping systems provide redundancy for emergency backup and efficiency when linked. The ARPANET links the ILLIAC IV, BBN TENEX, ARC TENEX, and IBM 360/67. Simple file transfer through the net provides an efficient internal structure. Some of the hardware restrictions imposed by the ARC TENEX — a highly variable memory ranging from 128 to 256 K, no direct access to tape drives, 2 to 3 hr delays in tape reading, downtime, and other characteristics — were overcome by use of the BBN TENEX. Memory boxes were frequently pulled due to hardware failures on the ARC TENEX, limiting memory; BBN TENEX has ample memory as well as direct remote access to the tape drives in a real-time mode. The trade off is the distance between BBN at Boston and the local Ames facility. Large amounts of raw data cannot be efficiently transferred through the net. Tapes had to be reformatted to BBN format and were delivered overnight by express air freight. Only small files, which remained after BBN TENEX processing, could be file-transferred.

There was a 4-week shutdown of the ILLIAC IV beginning in late December. A 2-week downtime for maintenance was anticipated, but the longer time necessitated the use of the CDC 7600 as backup. Although the CDC 7600 cannot classify using mask files, classifications could still be performed to provide data for the workshops and to maintain their schedule. Any slippage in early identification workshops due to lack of data would have had a severe effect on the anticipated peak period in March and April. This limited redundancy resulted in three or four separate classifications for some quads. Fortunately, this batch system was only needed a time or two. The ILLIAC IV, with its extremely fast processing, handled all of the clustering and most of the classification, including remakes of the 7600 runs. The ARC TENEX was still used to set up for the ILLIAC IV.

An efficient master run that could perform all of the final processing steps in one batch run (except when terrain data were used) was developed for the IBM 360/67. This significantly reduced repetition of many individual steps and required a very limited set of instructions for each quad. Nearly all of the final products were produced in about 6 weeks.

Although the IDIMS/HP 3000 system at ARC is a stand-alone system and has most of the software required, it is limited in processing speed and is usually in heavy demand.⁵ Its major significant feature was the interactive color display that was used to display classified data retrieved from the ILLIAC. CDF field people identified each class using this CRT display and color infrared U-2 photography.

V. DEVELOPMENT AND ORGANIZATION OF THE DATA SET

A. MOSAICKING

The experience of the Jet Propulsion Laboratory at Pasadena with planetary mapping missions was extended to Earth applications, using Landsat with the addition of known geographic reference for registration and reprojection to a desired map projection. The statewide digital image was created from 32 individual Landsat frames, as shown in Fig. 3. All the frames were obtained during August 1976, during which time there was virtually no cloud cover. Two additional cloud-free frames were required (May 1977) because of inland fog. The statewide mosaic is too large to process in one piece, and six pieces were ultimately used as large segments of the whole. The $1^\circ \times 1^\circ$ quadrangles separated from each piece are shown in Fig. 4.

The statewide mosaic was an extension of work already in progress at JPL for the Bureau of Land Management, including a large piece (Ia) over the California desert. Two decisions had already been made on that project and were accepted statewide: (1) that the image data be resampled to 80×80 -m pixels, and (2) the images be reprojected to a Lambert conformal conic map base. The details of digital image mosaicking will not be discussed here; they are available elsewhere.⁶ A brief summary of the major steps follows to describe the data set.

A gridded map base, consisting of 80×80 -m pixels projected to the Lambert conformal conic and aligned at 11° to the 118° meridian, was constructed. This map base covers the entire state with about 14,000 samples by 12,000 lines. A projection algorithm was developed that related each pixel to geographic coordinates. The goal of the mosaicking is to determine the correct grey level for each pixel using geometric control between Landsat and the map base and accounting for radiometric differences between scenes, using a continuous radiometric surface defined by the spectral data itself. Registration is performed using two sets of control points. Typical ground control points identifiable

on imagery and maps are used, as are points found through autocorrelation in the overlap areas of the frames, called edge tiepoints. All the tiepoints are used to form the corners of quadrilaterals for a "rubbersheet" transformation and the spectral values determined using bilinear interpolation. The spectral value of the edge tiepoints are also used to develop the radiometrically smooth and continuous surface for adjusting brightness for all pixels.

One large section of the state was completed at a time. The data was then subdivided by 1° latitude by 1° longitude to form data tapes. Each $1^\circ \times 1^\circ$ quad is half of a 1:250,000 scale USGS topographic map. Digital terrain data from the Defense Mapping Agency terrain tapes (DMA) were also registered to the map base and quantized into 40-ft elevation intervals. Using a moving 3×3 -pixel window the local slope gradient and aspect were calculated for each pixel. Thus, one data tape consists of four spectral bands and three terrain values registered pixel for pixel. There are 57 such quads for California; they are indicated on Fig. 4 as the straight orthogonal lines. All mosaic functions were performed using VICAR software on JPL's IBM 360/65.

Geometric infidelities and distortions were discovered in the mosaicked data, but there was not time to correct them. A full discussion of the implications and nature of these problems is beyond the scope of this paper. The mosaicking techniques used here appear to be weak in two areas. The transformation process itself may introduce errors, especially in concert with the geometric control, the second problem area. The radiometric smoothing techniques were successful for this data set, which was confined over a short time period; thus, it is not a universal conclusion for areas plagued by cloud cover. Methods of establishing geometric control capable of achieving rms errors of about 1 pixel are firmly established, but were not used on this project. Both near-term and long-range uses of the data set should be considered in planning. These errors are very difficult to remove after the fact. Many applications are severely compromised on a local level although the effects statewide are not dramatic. A full and complete examination of mosaicking technology is recommended before additional large-scale projects are attempted. In particular, all aspects and attributes of mosaic techniques should be carefully weighed and agreed upon in advance to secure full future usefulness of the data.

B. DATA SET MANAGEMENT

A large number of files, records, and tapes were generated for each quad. A simple coding system was used to manage recordkeeping. A four-digit number in the lower right-hand corner of a quad designated latitude and longitude (e.g., 3419 is 39° latitude and 119° longitude); these numbers were used with either prefixes or suffixes to identify the files or tapes, such as CL3419 being a classified tape.

Use of the two TENEX systems and the ARPANET permitted maintenance of a single directory of files archived at BBN and ARC. A complete record of all major processing operations was maintained. The only record from the IDIMS workshops, a list of identifications, was kept in statistic file and quad folders.

C. STRATIFICATION FOR CLUSTERING - ECOZONES

A continuous data set makes it possible to cross scene boundaries with natural boundaries for assembling data for cluster analysis. CDF modified Kuchler's map of the natural vegetation regions of California to develop an ecozone map of the state with 32 major natural and important urban regions (Fig. 5). This stratification achieves two purposes: it reduces the confusion caused by very similar spectral signatures derived from vastly different sources, such as desert alkali flats, high-mountain bare rock, and heavy commercial urban types. It also confines the spectral signatures within an ecozone to a much narrower range of resource types, for example, between the conifer types of the South Sierras with the hardwood-conifer types of the foothills. Although there may be difficulties at borders between regions in a transitional area, the goal is to make each ecozone more characteristic of that region as a whole.

The list of classes selected for the classification scheme are presented in Table 2. The major emphasis is on forested types, with the agricultural, urban, and desert areas having numerous spectral classes lumped into a few types. The use of a discrete definition can present difficulties in interpretation when many classes really represent a continuum in composition.

D. DIGITIZATION AND GEOMETRIC CALIBRATION

To make stratification of the digital data possible, the ecozone boundaries were delineated on 1:250,000 scale maps, digitized, and formed into masks. A mask is a mapping of digitized polygonal data converted to raster format and used to extract data from a registered image using Boolean logic. The ecozone components and the county boundaries of each 1° x 1° quad were separately digitized and plotted to scale to verify their accuracy. The file consists of nodes and vectors defining each area, an area identifier, and a geographic calibration, all stored on the TENEX systems.

To generate a mask, a calibration between the map coordinates of the digitization (latitude and longitude) and the digital image data were created. Fourteen control points for each quad, supplied by JPL, were used to generate a second-order least-squares transformation equation. In some cases, points were edited to reduce the rms error. The digitized lines were processed using each equation to form a compressed raster-type mask to overlay either the raw or classified data to extract information using logical functions.

E. REFORMATTING

Because most of the systems used were developed independently, different formats are specified by each. Incompatibility between systems begins with the format. Data received from JPL is in VICAR format (one or more 360-byte headers) in band-by-band format on 1,600 bpi tape, and record sizes of 1450. The EDITOR functions required pixel interleaved data. The four bands were read onto the CDC 7600 in parallel, the information from the VICAR header on each file placed in the appropriate field of the EDITOR (1024 byte) header and the stretched (0-255 dn value) VICAR pixel values halved as they were interleaved. The 800-bpi output tape is compatible with the ARC TENEX. More reformatting is required for the ILLIAC IV. The data were reblocked to 8192 bytes/record and output to 1600-bpi tape. The BBN tape requires a 664 header with TENEX blocking and 1600-bpi tape. Some economy was achieved when four quad data sets could be placed on one 1600-bpi tape for BBN and ILLIAC. Reformatting was a potential bottleneck if not performed efficiently. Because the JPL deliveries peaked late in the project, all reformatting had to be quickly accomplished (in batches of about 10 quads) so that processing could proceed on schedule.

VI. BULK PROCESSING TECHNIQUES

Bulk-processing, the repetitive routines used to classify a large data set broken into small processing units, consists of data compression, cluster analysis, classification, and generation of final products. The approach taken for classification was unsupervised; this was necessary due to the magnitude of the project and a lack of ground-truth to train the classifier. As described earlier, efficiency is gained by access to several computer systems linked by a network, in this case the ARPANET. All of the analysis was performed at a single network tip. Although there is overlap in capability of these software systems, we were able to take advantage of the strong points of each, build a smoothly operating design, and maintain the pace of a gruelling set of workshops. The design eventually achieved a high throughput, using existing routines capable of processing about 10 quads about 20 million pixels) from date of receipt to classification in less than 10 days. Most of the bulk processing was performed by a single senior analyst.

A. DATA COMPRESSION

Data compression is a means of packing and weighting four-band data into compact files for clustering. The data to be compressed were defined by the ecozone masks; however, the ecozone boundaries cross quad lines, as shown in Fig. 6. Although the preferred method is to reassemble the parts of an ecozone from the appropriate quads, in the initial stages of the project each ecozone part of a quad was separately compressed and analyzed

through classification. Limitations in disk space and poor access to tape drives on the ARC TENEX dictated this approach. The method produces an unmanageable number of statistics sets, however, and the problem of combining and editing statistics sets after identification of classes quad-by-quad to form a single statistics file for each ecozone becomes unwieldy. The problem is worse for ecozones like the South Sierra, which extend into 8 quads. Nonetheless, during Phase I several ecozones were completed this way; they required almost 3 months to complete.

A dramatic improvement in efficiency and probably the key to maintaining the schedule was the addition of the BBN TENEX through contract. Greater disk space, larger memory, and direct access from the link to the tape drives were the characteristics that dictated this choice. Data compression could be achieved from an entire ecozone at once, resulting in a single statistics file. All quads containing a portion of a specified ecozone were read to disk, the mask files were used to extract the appropriate data, and were then combined in a single multiwindow file. The windows were packed to fill longer record lengths and the spatial relationship was abandoned.

The spectral data were compressed from millions of entries to thousands by a process of weighting. This algorithm simply counts the number of pixel occurrences of each unique combination of dn values from the four MSS bands. The file is thresholded at 2^{15} values for ILLIAC clustering. On some ecozones, the threshold was exceeded. This was corrected by deleting all single value occurrences (up to 2,000 were deleted in each delete cycle until the maximum allowed was achieved). If still in excess, the last bit of one or more bands was zeroed, rounding down to an even integer. This reduced the number of unique values and had the effect of dropping cluster means by a half dn value. These compression methods allowed the cluster analysis to consider each pixel in the data set without sampling.

An example of the benefits of data compression is illustrated by the Central Foothills ecozone. This zone contains about 1.25 million pixels, far too many for conventional nonsampled clustering algorithms. After weighting, the file is reduced to 42,598 dn values. This is acceptable to the ILLIAC weighted-clustering routines, which can accept up to 66,000 unique values or 100,000 pixels of unweighted data. In this case, a complete reenumeration is being processed when conventional algorithms would require a one-sixteenth sample.

B. CLUSTERING

To prepare the weighted file above for clustering, the analyst must specify the parameters of the process. Using a histogram of the raw data, the analyst estimated the expected number of clusters. By transferring the weighted file via the ARPANET to the ILLIAC, a batch run could be

prepared using the ARC TENEX for overnight processing. Although the maximum CPU time for any ecozone was less than 1,000 sec with 60 classes, the evening runs were used because of the very limited daytime access and the priority given extremely large users at ARC. At times, four to five clustering runs were executed in 1 day. Although the ILLIAC, until recently, has not been available to nonfederal users, it illustrates the kind of processing capability one must consider to be efficient in large-scale programs that do not have the luxury of dedicated large computers or that have to compete for computer resources. It is clear that bulk processing steps should not be attempted on smaller systems, but should be sent to either mainframe or array processors, either dedicated or on contract, so that the special benefits of the small systems (editing, management, display, etc.) can be efficiently utilized.

The output of clustering is the typical spectral class statistics set: means, variances, separabilities, and covariance matrix. The analyst edited these sets to reduce confusion by deleting or pooling classes. The entire clustering process lasted about 1 day for each ecozone. The final statistics file was then inverted for classification. An example of the effect of ecozone stratification is shown in Figs. 7 and 8. Two similar ecozones, the Central and South Sierras, which differ mainly in elevation, latitude range, and orientation to the weather patterns modulated by the Golden Gate at San Francisco, are compared. Although the spectral range is very similar, the class assignments are quite different. A different example is afforded by a comparison of the South Sierra and the South Foothills (Figs. 8, 9). The major differences in these ecozones are in elevation and soils. The transition from Sierra conifers to interior valley hardwoods is revealed by the assignments. Once again, the spectral data overlap significantly. It is impossible to account for the subtle variations in composition, as the transition occurs, using discrete classes and discrete ecozone boundaries. The main goal is to minimize confusion and improve the results overall.

C. CLASSIFICATION

Classification was run in parallel with clustering when possible. Quads, which were classified on an individual basis, could not be completed until all the ecozones within them had been clustered. For this reason, some quads were postponed or partially classified until JPL supplied more pieces into which an ecozone could be extended. Once statistics had been obtained for each zone within the quad, classification was completed. The mask file used to extract the ecozones from the raw data was again used to perform the masked classification; this is shown in Fig. 6. The mask files and the appropriate statistics files were combined through the ARPANET with the raw data to classify using maximum likelihood decisions on ILLIAC. These runs never exceeded 300 sec of CPU time and were usually performed during the daytime. In some cases, a classified tape might have up to 250 classes.

The output of the classifier was read to disk; it contained a header and two bytes per pixel, the assigned class and the maximum likelihood probability. The header and probabilities were stripped to generate a categorical tape for IDIMS. All values were retained but blocked to a 8192 record length for a second file to be used on ILLIAC IV.

VII. GENERATION OF FINAL RESULTS

Unsupervised clustering makes a trade off in effort with supervised methods in that all of the spectral classes must be assigned a category label after classification through the laborious review of each class on a quad-by-quad basis. This process could take up to several weeks on the IDIMS display for each ecozone. Thus, the workshops conducted to identify the classes become a fundamental scheduling guideline and consumed the greatest amount of analyst time. The state was divided into large areas roughly conforming to the mosaicking pieces, and one analyst had responsibility for each area (about 12 quads) to process through to final products. The tedious and numerous repetitions required to complete the final processes were further improved by the development of a common batch run with minimum specifications for generation of final products. This team approach, with minimum overlap of responsibilities, worked reasonably well.

A. IDENTIFICATION OF SPECTRAL CLASSES

The task of attaching a category to each class required the combined efforts of a field person from CDF with an analyst at ARC. Six CDF foresters selected the ecozones with which they were most familiar. For periods of 1 to 3 weeks, beginning in December, a forester spent 4 to 6 hr/day with an analyst on the IDIMS CRT display; by comparing CRT displays with aerial photographs, the forester then assigned the best possible label. U-2 CIR photos at scales from 1:300,000 to 1:120,000 were used exclusively. During April, two workshops were conducted in the same week, back to back. Because none of the foresters had experience with U-2 photography or digital data, several training sessions were conducted to familiarize them with the data. Their field experience was essential to understanding the scope of California's diverse resources and is recommended for any large-scale effort.

Each spectral class was displayed individually on the CRT display. By enlarging specific areas and correlating features to the photographs, a label could be chosen. For some ecozones, up to 8 quads had to be checked before a label was confidently assigned. For a number of classes, an ambiguity was readily recognized for some spectral classes. For example, in the South Sierra, brush at lower elevations was confused with some conifers at higher elevations. Anywhere the ambiguity could be resolved by terrain data alone, the pertinent elevation contour was noted. The result was a listing by class of the assignments to be used by each analyst to generate final products.

B. STRATIFICATION WITH TERRAIN DATA

The ambiguities discussed above, usually arising in areas of high physical relief, were corrected through stratification. The elevation data provided in the original data set was quantized into 40-ft intervals from sea level to 10,200 ft. These values were further reduced to two or more contours by mapping all values below a critical elevation into one value, and similarly, into another value above the elevation. The resulting images were converted into masks on EDITOR when USDA programmers modified an existing routine to accept the data. The masks were overlaid on the classified data and the affected classes divided as indicated, with changed pixels being relabeled as one of similar categorical makeup. This technique was used for only four ecozones.

C. FINAL PRODUCTS

The final color-coded product consists of a film recorder displaying the classified data grouped into 16 classes, overlaid with county boundaries with a title, and an annotated color bar. The process to generate a tape to be used on the film recorder was developed into a single program requiring only the listing of grouped categories, the county boundary mask file, and the identification of the title to be drawn from a file. All other instructions, such as the pseudo-color table, the color bar, and annotations, are repetitive; they were common to all. This listing is hardly trivial since some quads had over 200 classes arranged in an order dictated by the ecozone masking and was easily subject to clerical error. The program also recorded acreage tabulations by county and quad. All analysts with an area responsibility used this batch technique, which became very efficient.

The output tape was used to generate a 4 x 5-in. Polaroid proof print and two negatives on a DICOMED film recorder. These 57 negatives were enlarged and printed photographically at scales of 1:1,000,000 and 1:250,000 and color matched. The Polaroid proofs and the 1:1,000,000 scale prints have been manually mosaicked to create a statewide image map. A tabular listing of the acreage totals by county have been collated.

The final processing for all 57 quads was done in about 1 month by four analysts working part-time. The film recorder results and the acreage tabulations were completed by May 15 and sent to CDF as part of the final report on May 15. The statewide mosaic was completed June 15 and the 1:1,000,000 scale mosaic in the spring of 1980.

VIII. CONCLUSIONS AND RECOMMENDATIONS

A. SOFTWARE DEVELOPMENT

The limited time available and the large scope of this project required the use of existing software capability. Two problem areas were soon discerned. First, most software routines had not

been designed to handle both the very large throughput and the variance of spectral values inherent in the data. This stretched most system capabilities to their limits, but the process motivated improvements and upgrades through minor modifications. In some cases, software from two or more computers had to be used to achieve results. Two key examples were the packing programs to accept larger data sets and the mask program to accept terrain data for stratification. Secondly, the various systems had never been previously integrated to respond to a heavy throughput. The major bottleneck was reformatting. Modifications to existing routines provided the reformatting but the problem of large throughput remained. At least 1 week was required for reformatting during each of the 6 phases.

B. COMPARISON WITH OTHER SYSTEMS

In view of the limited time and the magnitude of the project, a single-image processing system probably could not have completed the work on schedule. The system design required a comprehensive software system with access to a high-speed main frame for bulk processing. At ARC, the required software routines are resident on a number of computer systems. For these reasons, the major asset to completing this project on schedule was the access to a computer network system such as the ARPANET. Another asset was the availability of off-line minicomputers used in displaying data and other tasks, such as line printer maps and tape copy. The multimachine design is not essential to the analysis, but the analysis could not have been completed on schedule without it.

Projects such as this can now be handled by a number of image-processing systems. The issue becomes one of the speed and responsiveness pertinent to use in real-world applications. This project illustrates at least these two points: (1) the value of computer networks for bulk processing, since no one system contains all the required software or speed; and (2) the required usage of array or parallel processors for processing large areas.

C. COST OF PROJECT

The total cost of this project, including salaries at industry equivalents, was less than \$300,000, or 0.3 cents per acre.

D. CONCLUDING REMARKS

The results have been exceptionally well received by both the CDF and other state users. The data are presently being used by numerous federal, state, and county users throughout California in both raw and classified form. CDF is using the data and refining it through supervised techniques as part of their design as described in the early part of this report.

The techniques described in this paper illustrate a workable and efficient system that can

achieve desired results within strict time constraints.

IX. ACKNOWLEDGMENTS

We especially recognize the efforts of Dale Wierman, the project manager of the California Department of Forestry (CDF), and Nancy Tosta-Mille: CDF, technical coordinator, for their exceptionally fine work and cooperation. EDITOR programming support, provided by Martin Ozga from the U.S. Department of Agriculture, ESCS Remote Sensing Group, made it possible to handle many of the large data sets. Claudia Bertrand, Claire Bowen, Barbara Kimmy, and others at the Advanced Computation Laboratory were invaluable for their coordination of ILLIAC and TENEX operations and their continual responsiveness. We also extend our thanks to Dr. Charles Poulton, of Airview Specialists, for the excellent training courses conducted on the project. Finally, Ron McLeod, Nevin Bryant and others from the Jet Propulsion Laboratory are recognized for the large workload completed on schedule to create the statewide mosaic.

X. REFERENCES

¹Assembly Bill 452, Keene: Forest Resources Assessment and Policy Act of 1977, Chapter 12 (Secs. 4800 to 4807) of Part 2 of Division 4 of the Public Resources Code, State of California, Sacramento, Calif.

²1977, ed., The Map of the Natural Vegetation of California (Appendix). Barbour, Michael G.; Major, Jack, ed., Terrestrial Vegetation of California: New York, John Wiley & Sons, p. 909-915.

³Ray, Robert M.; Ozga, Martin; Donovan, Walter E.; Thomas, John D.; and Graham, Marvin L.: EDITOR: An Interactive Interface to ILLIAC IV - ARPA Network Multispectral Image Processing Systems. Center for Advanced Computation, CAC Document No. 114, University of Illinois, Urbana, Ill., 1974.

⁴Ray, Robert M.: Implementation of ILLIAC IV Algorithms for Multispectral Image Interpretation. Center for Advanced Computation, University of Illinois, Urbana, Ill., 1974.

⁵Interactive Digital Image Manipulation System (IDIMS) User Manual. Electromagnet Systems Laboratories, Sunnyvale, Calif., ESL, Inc., Technical Memorandum ESL-TM 711, 1976.

⁶Zobrist, Albert L.: Multiple-Frame, Full Resolution Landsat to Standard Map Projections. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, Calif.

Table 1. Systems Used

Computer	Program	Functions
ILLIAC IV (ARC)	EDITOR IV functions	High speed bulk processing Weighted clustering Masked classification
TENEX (BBN) ^a	EDITOR	Packing and weighting by ecozone Digitization
TENEX (ARC)	EDITOR	Digitization, mask generation preparation for ILLIAC IV runs, file management, editing, aggregations
IBM 360/65 (JPL) IBM 360/67 (ARC)	VICAR/IBIS (JPL) (ARC)	Mosaicking procedures (JPL) Reformatting (ARC)
HP 3000	IDIMS	Identification of classes Selection of color code, color bar, and annotations
IBM 360/67	Image processing routines	Generation of final products Aggregations Reformatting Overlay county boundaries
CDC 7600	Image processing routines	Reformatting Interim classifications
SEL 32	ERL software	Line printer maps Tape copies

^aBBN: Bolt, Beranek and Newman at Boston

Table 2. Classification^a

Alkali flats - Areas essentially devoid of vegetation; highly alkali
Bare rock
Barren - Areas with less than 5% vegetation; predominantly bare soil
Water
Other - Ice or snow, clouds
Grassland - Areas where the cover is predominantly grass; vegetative cover is greater than 5%; less than 10% tree canopy; there may be intermixed herbaceous species
Open shrub - 5 to 24% of vegetative cover is shrub species; less than 10% tree canopy; substratum dominates the landscape; shrubs are often xeric
Brush - 25% or more of vegetation is shrub species; less than 10% tree canopy; may have herbaceous understory; includes riparian vegetation
Conifer - Conifers create 25% or more of the tree canopy closure; hardwoods comprise less than 20% of the tree species present
Conifer-hardwood - More than 25% tree canopy closure, with conifers comprising more than 50% but less than 80% of the stand; hardwood species comprise 20 to 49% of the stand
Hardwood - More than 25% canopy closure of hardwood species; less than 20% conifer species present
Hardwood-conifer - More than 25% tree canopy closure, with hardwoods comprising greater than 50% but less than 80% of the stand; conifers comprise 20 to 49% of the stand
Conifer-woodland - Conifers create 10 to 25% tree canopy closure; hardwood species comprise less than 50% of the species present; herbaceous or brush understory may be present
Hardwood-woodland - Hardwoods create 10 to 25% tree canopy closure; conifer species comprise less than 50% of tree species present. Herbaceous or brush understory may be present
Agriculture - Crops, irrigated fields, orchards, etc.
Urban - Concentration of buildings, structures, roads, and other man-made items; may be residential or commercial

^aDefinitions: (1) grass or herbaceous - vegetation with no significant woody structure in the stem; (2) brush - woody vegetation less than 5-m tall; (3) tree - woody, nonclimbing vegetation, \geq 5-m tall; (4) hardwood - broad-leafed, most species deciduous; (5) conifer - needle-leafed, most species evergreen.

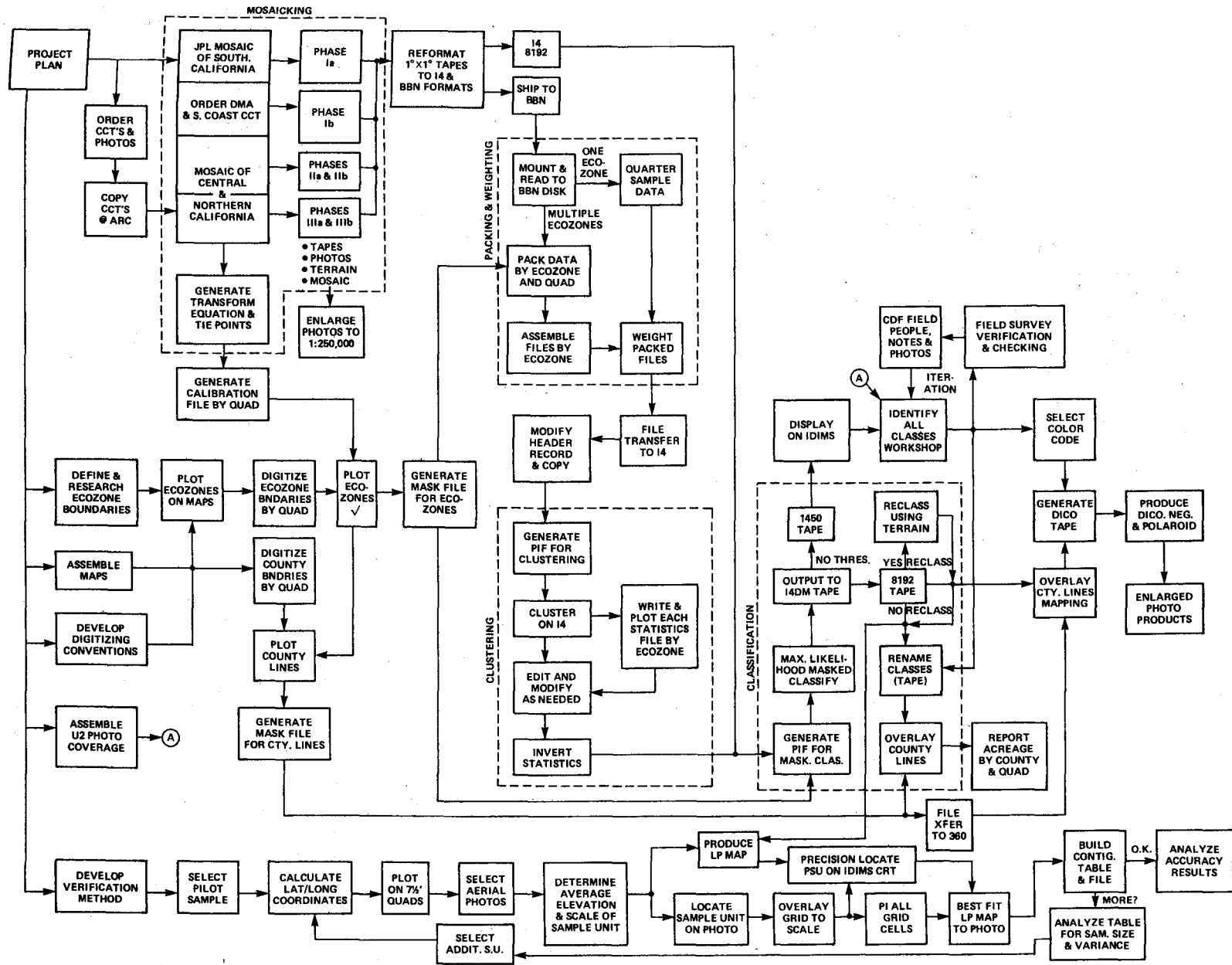


Figure 1. Flowchart of Analysis.

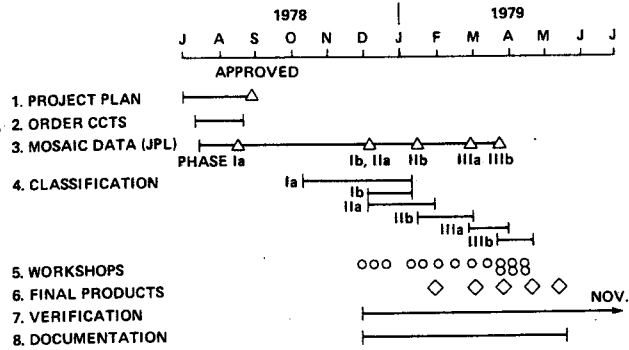


Figure 2. Schedule of Classification and Mosaicking Processes.

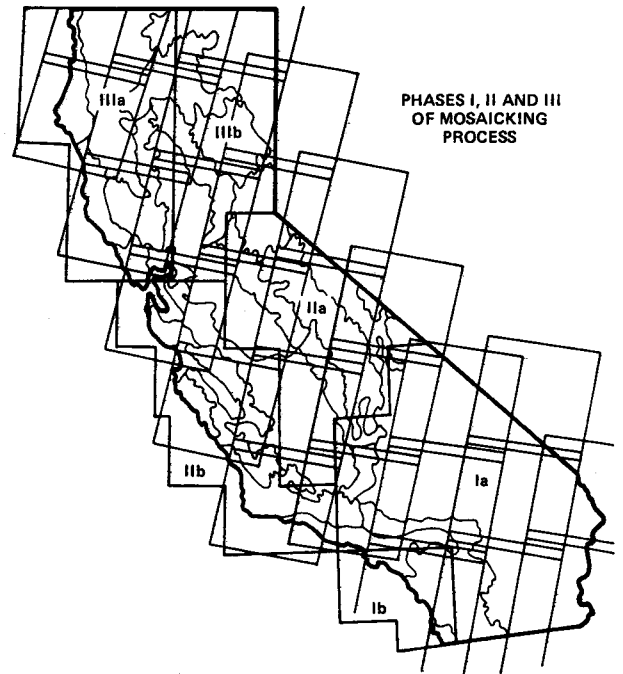


Figure 4. 1° x 1° Quads and 6 Mosaic Phases.

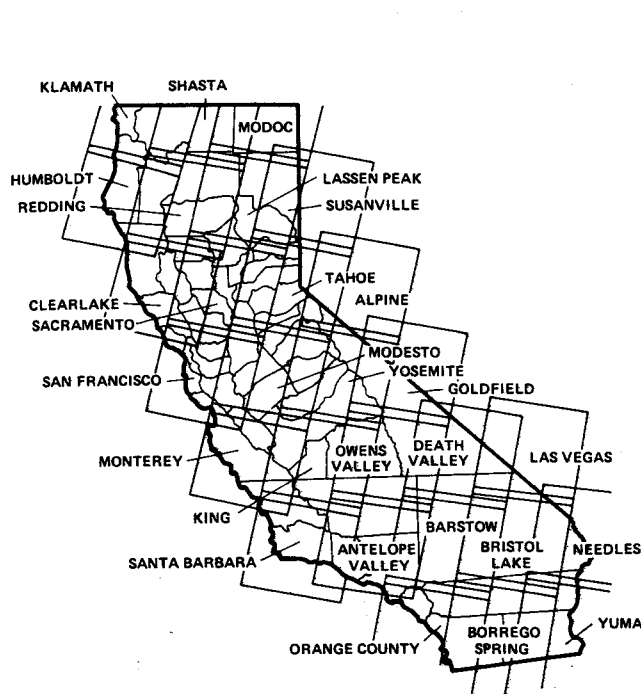


Figure 3. Landsat Scenes of California.

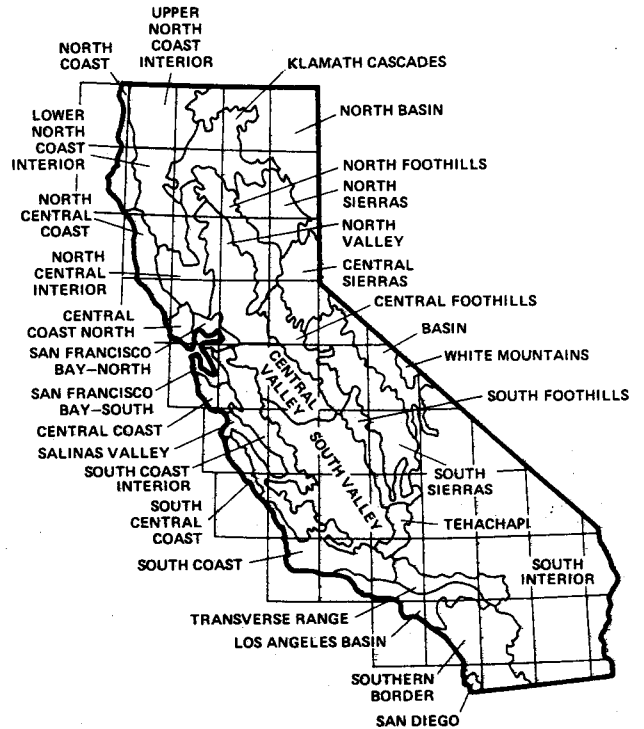


Figure 5. Ecozones of California.

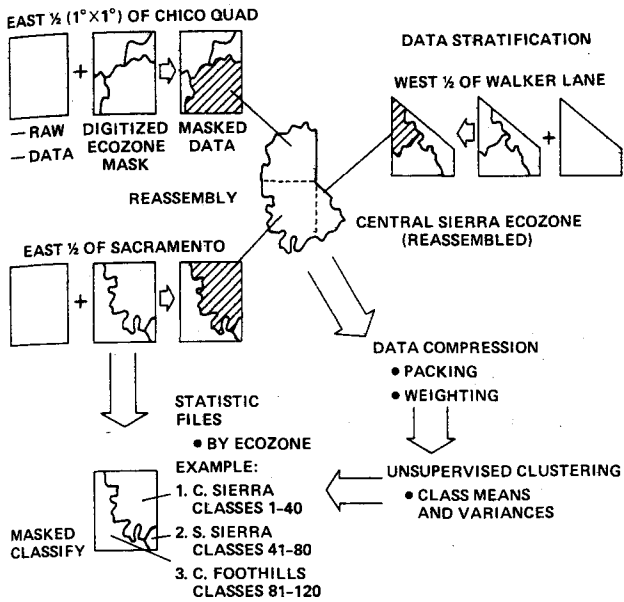


Figure 6. Classification Processing.

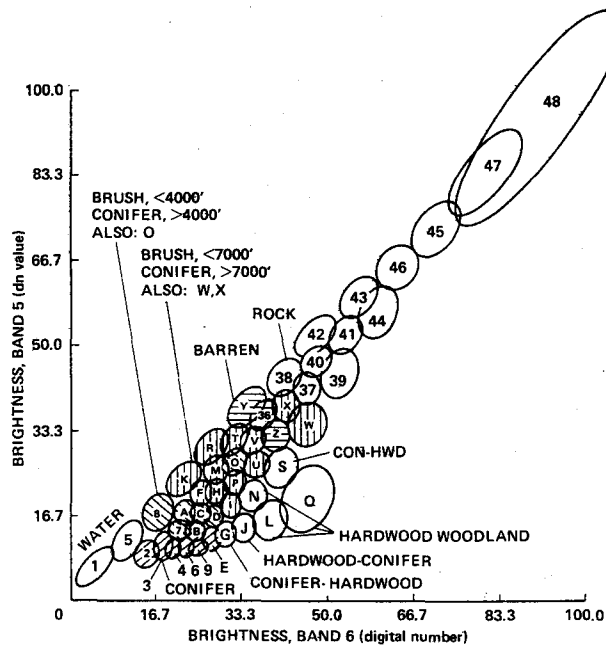


Figure 8. Two Dimensional Cluster Plot.

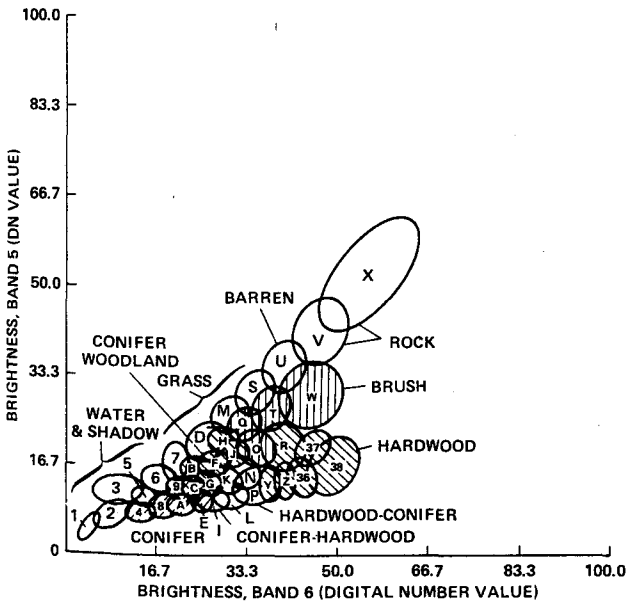


Figure 7. Two Dimensional Cluster Plot.

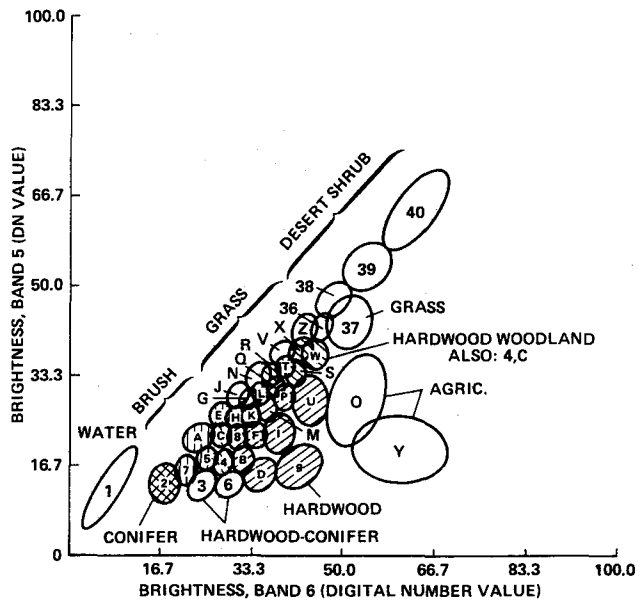


Figure 9. Two Dimensional Cluster Plot.

Willard L. Newland was born in Clinton, Indiana, on December 21, 1940. He attended San Jose State University and received a bachelor of arts degree in the field of Geography in 1974. In 1980, he will receive his master's of arts degree in the same field, from San Jose State as well. Since 1974, he has worked at the NASA-Ames Research Center under different federal agencies. He has worked for the Department of Interior-United States Geological Survey at NASA-Ames as well as under contract to NASA. Currently, he is employed by Technicolor Graphics, Inc., at NASA-Ames as Senior Data Analyst.

Susan D. Norman was born on June 4, 1943, in Detroit, Michigan. She attended and graduated from the University of Michigan with a bachelor of science degree in Mathematics in 1965. She later continued on with her education and received her master of science degree in Computer Science from Stanford University in 1969. Norman came to NASA-Ames in 1965 and first worked for the Mission Analysis Division. She stayed at Ames until 1974 when she left to work for Sikorsky Helicopters in Connecticut for a year. She then returned to NASA-Ames and in 1977 became a member of the Technology Applications Branch, which oversees the Western Regional Applications Program (WRAP). Currently, Norman is the Regional Manager for all Southern WRAP states.

David L. Peterson was born on July 26, 1943, in Brigham City, Utah. He graduated from the University of Utah with a bachelor of science degree in Mechanical Engineering, 1966, and then continued on to Stanford University to receive his master of science degree in the same field, in 1969. Soon after getting his bachelor's, he came to work for NASA-Ames Research Center at Moffett Field, California. For 10 years, he worked in the Thermodynamics and Gas Dynamics Division; he is now a member of the Technology Applications Branch. His current job includes being a technical manager for the Western Regional Applications Program (WRAP) as well as head of forestry projects for WRAP.