# Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames

Statistical Reporting Service

U.S. Department of Agriculture

Washington, D.C.

# PILOT STUDY OF THE POTENTIAL CONTRIBUTIONS OF LANDSAT DATA IN THE CONSTRUCTION OF AREA SAMPLING FRAMES


By

George A. Hanuschak

Kathleen M. Morrissey

Research and Development Branch
Research Division
Statistical Reporting Service
U. S. Department of Agriculture
Washington, D.C.


October 1977

Contents

## List of Tables

## List of Appendices

Acknowledgments

The strong support of several members from the following groups made this project a reality:

1. The Center for Advanced Computation (University of Illinois).

2. New Techniques Section, R&DB, RD, SRS.

3. California SSO and Enumerative Staff, SRS.

4. Methods Staff, ED, SRS.

5. Systems Branch, SD, SRS.

6. Data Collection Branch, SD, SRS.

7. Area Sampling Frame Section, SSRB, RD, SRS.

The authors wish to extend a special thanks to each participant for their hard work and contributions. Martin Ozga, CAC, deserves special recognition for his fine developmental programming efforts. Special thanks also goes to Barbara Holt and Kathy Wall for their patience and fine efforts in typing this report. Robert Slye, NASA-Ames, California, provided the DICOMED (color-coded) classified LANDSAT data print in Appendix C.

# PILOT STUDY OF THE POTENTIAL CONTRIBUTIONS OF LANDSAT DATA IN THE CONSTRUCTION OF AREA SAMPLING FRAMES

## I. PILOT STUDY OBJECTIVES

Two general topics were considered in the investigation of the potential contributions of LANDSAT data in the construction and utilization of area sampling frames. The first topic area investigated was the potential contribution of LANDSAT data in aiding current area frame construction methodology. Specific questions addressed were:

1. Can LANDSAT data replace aerial photography for land use stratification and frame unit construction in area sampling frame construction?
2. Can LANDSAT data be used together with conventional ASCS aerial photography in area frame construction?

The second topic investigated was the potential contribution of LANDSAT data in determining new area frame construction and utilization methodology. Specific questions addressed were:

1. Can LANDSAT data, grouped into crop and land use categories for each area frame unit, provide useful control data for area sampling frames?
2. If the answer to question 1 is yes, then what type of data processing system will be needed to incorporate promising techniques into the present operational system?

Since LANDSAT data was studied for its potential as control data for area frame units, county regression and ratio estimates for major crop acreages were also investigated.

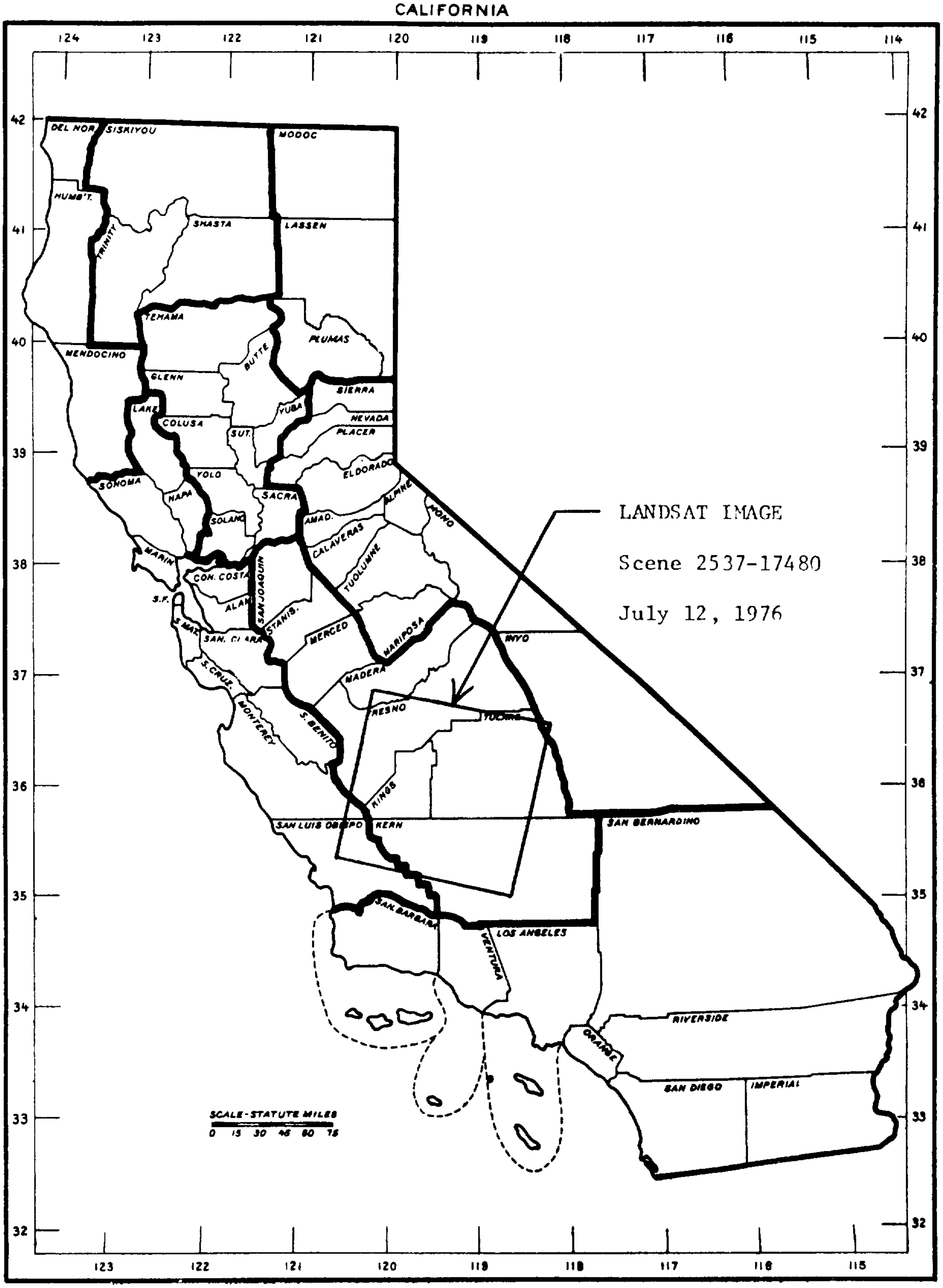
## II. ACQUISITION OF LANDSAT DATA

The study area in this report concerns one LANDSAT scene which was centered over the southern portion of the San Joaquin Valley in California. A cloud free image dated July 12, 1976 was available for analysis purposes. The LANDSAT scene 2537-17480 completely contained Kings County and the main agricultural areas of Tulare and Kern Counties as well as smaller portions of Fresno, Madera, San Luis Obispo and Monterey Counties. The geographic location of the scene on a California state map can be seen in Table 1 on page 3. The quality of the LANDSAT imagery was excellent and the image is displayed in Figure 4 in Appendix C.

The relative stage of maturity for the various crops was favorable at the time of the satellite pass for remote sensing purposes. Cotton was progressing well with some of the crop in the bloom stage. Orchards and vineyards had basically green covers while the non-irrigated pasture and rangeland were in critically dry condition. The corn crop was progressing with some tasseling and alfalfa cutting was active in the area. The winter wheat and barley crops were partially harvested across the Valley and required special analysis techniques which will be discussed in a later section entitled, "County Crop Acreage Estimation".

Table 1

Geographic Analysis Area



## III. COLLECTION AND USE OF 1976 CALIFORNIA JES SAMPLE SEGMENT DATA

Ground survey data for use in the LANDSAT analysis was collected during the 1976 June Enumerative Survey in California. A modified 1976 JES questionnaire (Part A) shown in Table 2 on page 8 was used for ground data collection. Data was recorded, keypunched and retained at the individual field level for all tracts and segments. In California, 20,749 fields were recorded in the JES segments.

Along with the preservation of the field level identification, new coded items were added to the Crop Section A questionnaire. Additional information recorded by the enumerator and retained in the keypunched record was:

| Item Number | Item |
|---|---|
| 1 | Total Acres in Field |
| 23 | Other Uses of Grain Planted |
| 38 | Field Appearance Code |

Unique codes were assigned for the intended crop utilization and field appearance items. Three digit codes for other intended usage of grains planted were designated for silage, hay, seed, pasture, abandoned and other. The field appearance item code was assigned a specific two digit value for the enumerator's description of the relative maturity or condition of the crop. The crop maturity definitions can be seen in Table 3 on page 7 . After JES processing was complete, the raw data including updates were transmitted via the INFONET system for special procedural editing and reformatting by Research and Development personnel. A strung record was created for each JES field using the Generalized Edit System. The strung record file was then inputted to the Statistical Analysis System for reformatting. Then the Generalized Edit System was used for updating the records for editing purposes. The final edited records were put on tape and sent to Bolt, Beranek and Newman, a data processing facility in Boston, for use in the LANDSAT data analysis.

Aerial photographs, produced by the Agricultural Stabilization and Conservation Service (ASCS) at a scale of 8 inches = 1 mile, were also a source of ground information for the project. Accurately located tract and field boundaries were essential to analyze the LANDSAT data. After field enumerators delineated the field boundaries to correspond with the recorded acreage information for the JES, the photographs were mailed to the State Statistical Office for review. For use in this project the photographic enlargements were reduced and copied at a scale of 4 inches = 1 mile. For all 1976 JES sample segments, tract boundaries and codes were outlined in blue ink and all field boundaries and numbers were in red ink.

In preparation for digitization [1/] and creation of the final ground observation file, a coordinated task of editing the photographs with the JES ground data file was performed. Field acreages were reviewed for consistency. That is, corresponding crop irrigation, appearance, and utilization codes were checked for a logical sequence. When harvesting of two crops was to occur during the year, the ground data was revised to be time-analogous with the July 12 LANDSAT imagery date covering the analysis area.

---

[1/] In this context digitization means the recording of segment, tract and field boundaries on an electronic X-Y coordinate system. With the use of several transformations, latitude and longitude map coordinates of field boundaries can be located on LANDSAT line and column coordinate systems.

For our research effort a subsample of 46 segments was drawn from the segments in Kings and Tulare Counties. This subsample was used in the training and testing data sets for the LANDSAT classification algorithms. When the editing process was completed, the final ground observation file contained data for 143 fields in Kings County, and 666 fields in Tulare County.

The next step was the digitization of the segments in Kings and Tulare Counties. The EDITOR software subsystem, an interactive data analysis system for processing LANDSAT data developed jointly by the Center for Advanced Computation at the University of Illinois and SRS, was utilized at this point as a means of recording latitude and longitude coordinates of segment boundaries. All tract and field boundaries within the segments were digitized. Plots of the segment, tract and field boundaries are produced at the scale of USGS quad maps, ASCS aerial photographs, and LANDSAT scales to aid in editing.

Registration procedures for locating the training segments on the LANDSAT data tapes was performed. Computing a third-order bivariate polynomial transformation between the LANDSAT coordinates and the USGS quad map coordinates, calibration errors were computed and found to be well within tolerance levels. Individual segment registration errors were in terms of a one pixel difference in extreme cases with the majority of the residuals less than one pixel for both lines and columns. Because these errors were within acceptable limits on the first attempt at registration, further refinements were not necessary. The maximum residuals using a third order polynomial transformation for the 84 control points located globally across the July 12th scene were .8 pixel for line and 1.4 pixels for column. This was the first successful "one-step registration" effort by SRS in locating segments on the LANDSAT data tapes.

A list of references regarding use of the EDITOR system is provided in Table 4 on page 9.

**Table 3**

**Field Appearance Codes**

| All Crop Types and Land Uses ("except" orchards and vineyards) | | Vineyards and Orchards | |
|---|---|---|---|
| Code | Field Appearance Definition | Code | Field Appearance Definition |
| 10 | Green Cover (not in planted crop) | 90 | New Planting and Row Space/Less Than 30 Feet |
| 20 | Prepared Land (worked land including planted but not emerged) | 91 | New Planting and Row Space/Larger Than 30 Feet |
| 30 | Emerged (Less than 50% of field covered with green foliage, but not mature) | 92 | Mature and Row Space/Less than 30 Feet |
| 40 | Green (50% or more of field covered with green foliage, but not mature) | 93 | Mature and Row Space/Larger Than 30 Feet |
| 50 | Mature (turning or ready for harvest) | | |
| 60 | Harvested Crop (but not worked or prepared) | | |
| 70 | Dried or Cut Vegetation (brown pasture, cut hay, etc.) | | |
| 80 | None of Above (water, F.S., waste, etc.) | | |

Table 2

Crop Section A Questionnaire

| FIELD NUMBER..... | | | 01 | 02 | 03 | 04 |
|---|---|---|---|---|---|---|
| 1. TOTAL ACRES IN FIELD | | | 828 | 828 | 828 | 828 |
| 2. CROP OR LAND USE *(Specify)* | | | | | | |
| 3. OCCUPIED FARMSTEAD OR DWELLING | | | 843 | 843 | 843 | 843 |
| 4. WOODS, WASTE, IDLE LAND, ROADS, DITCHES, ETC. | | | | | | |
| 5. TWO CROPS HARVESTED FROM THIS FIELD? | | | NO ☐ YES | NO ☐ YES | NO ☐ YES | NO ☐ YES |
| 6. ACRES LEFT TO BE PLANTED? | | | 61 _ | 61 _ | 61 _ | 61 _ |
| 7. ACRES IRRIGATED AND TO BE IRRIGATED? | | | 6 __ | 6 __ | 6 __ | 6 __ |
| 8. PASTURE | | | 842 | 842 | 842 | 842 |
| 10. DURUM WHEAT | | Planted and to be planted | 551 | 551 | 551 | 551 |
| 11. | WINTER WHEAT | Planted | 540 | 540 | 540 | 540 |
| 12. | | For Grain | 541 | 541 | 541 | 541 |
| 13. | RYE | Planted and to be planted | 547 | 547 | 547 | 547 |
| 14. | | For Grain | 548 | 548 | 548 | 548 |
| 15. | OATS | Planted and to be planted | 533 | 533 | 533 | 533 |
| 16. | | For Grain | 534 | 534 | 534 | 534 |
| 17. | BARLEY | Planted and to be planted | 535 | 535 | 535 | 535 |
| 18. | | For Grain | 536 | 536 | 536 | 536 |
| 19. | CORN | Planted and to be planted | 530 | 530 | 530 | 530 |
| 20. | | For Grain | 531 | 531 | 531 | 531 |
| 21. | SORGHUM | Planted and to be planted | 570 | 570 | 570 | 570 |
| 22. | *(Excl. crosses)* | For Grain | 571 | 571 | 571 | 571 |
| 23. OTHER USES OF GRAINS PLANTED -- | | Use | | | | |
| Acres abandoned, cut for hay, silage, etc. | | Acres | 82 _ | 82 _ | 82 _ | 82 _ |
| 24. | HAY Cut and to be cut | ALFALFA AND ALFALFA MIXTURES | 653 | 653 | 653 | 653 |
| 25. | | OTHER HAY Kind | | | | |
| | | Acres | 65 _ | 65 _ | 65 _ | 65 _ |
| 29. RICE | | Planted and to be planted | 605 | 605 | 605 | 605 |
| 30. | COTTON UPLAND | Planted and to be planted | 524 | 524 | 524 | 524 |
| 31. | | Abandoned | 523 | 523 | 523 | 523 |
| 32. DRY EDIBLE BEANS | | Planted and to be planted | 607 | 607 | 607 | 607 |
| 33. SUGAR BEETS | | Planted and to be planted | 691 | 691 | 691 | 691 |
| 35. IRISH POTATOES | | Planted and to be planted | 552 | 552 | 552 | 552 |
| 36. | OTHER CROPS | Name | | | | |
| | | Acres planted or in use | ___ | ___ | ___ | ___ |
| 37. SUMMER FALLOWED during 1976 | | | 847 | 847 | 847 | 847 |
| 38. FIELD APPEARANCE CODE (See Card) | | | 829 | 829 | 829 | 829 |

Table 4

List of EDITOR System References


1. Ozga, M.; Donovan, W.; Gleason, C., 'An Interactive System for Agricultural Acreage Estimates Using Landsat Data', Fourth Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, June 1977.

2. Sigman, R.S.; Gleason, C.P.; Hanuschak, G.A.; Starbuck, R.R., 'Stratified Acreage Estimates in the Illinois Crop Acreage Experiment', Fourth Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, June 1977.

3. Ozga, Martin, 'Crop Acreage Estimation in EDITOR', CAC Technical Memorandum No. 95, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, May 1977.

4. Starbuck, Robert R., 'Overview and Examples of the EDITOR System for Processing Landsat Data', Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., March 1977.

5. Ozga, Martin, 'Selection, Sampling, and Tabulation of Masked Files in EDITOR', CAC Technical Memo No. 79, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, December 1976.

6. Ozga, Martin; Donovan, Walter E.; Ray, Robert M.; Thomas, John D.; Graham, Marvin L., 'Data File Formats for Processing of Multispectral Image Data' CAC Technical Memorandum No. 19, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, October 1976 (revised).

7. Ray, Robert M.; Huddleston, Harold F., 'Illinois Crop-Acreage Experiment', Third Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, July 1976.

8. Donovan, Walter E.; Ozga, Martin, 'Retrieval of LANDSAT Image Samples by Digitized Polygonal Windows and Associated Ground Data Information', CAC Technical Memo No. 57, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, August 1975.

9. Ray, Robert M.; Ozga, Martin; Donovan, Walter E.; Thomas, John D.; Graham, Marvin L., "EDITOR An Interactive Interface to ILLIAC IV - ARPA Network Multispectral Image Processing Systems', CAC Technical Memo No. 114, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, June 1975.

10. Donovan, Walter E., 'Oblique Transformation of ERTS Images to Approximate North-South Orientation', CAC Technical Memo No. 38, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, November 1974.

11. Ray, Robert M.; Thomas, John D.; Donovan, Walter E.; Swain, Phillip H., 'Implementation of ILLIAC IV Algorithms for Multispectral Image Interpretation, Final Report', CAC Document No. 112, Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois, June 1974.

## IV. AREA FRAME CONSTRUCTION METHODOLOGY USING LANDSAT

### A. PHOTO INTERPRETATION OF LANDSAT IMAGERY

In determining the potential contribution of LANDSAT imagery in aiding current area frame methodology, several methods were investigated for photo interpretation of LANDSAT imagery to define stratification by broad land uses. All methods involved overlaying maps (county map or a USGS quad map 1:250,000 scale) onto the LANDSAT imagery for photo interpretation. Methods investigated were:

1. Tracing boundaries such as roads, railroads, and waterways, from a 1:250,000 scale USGS quad map onto clear acetate and then overlaying the acetate on the LANDSAT color composite imagery (1:250,000). The map features overlay quite well but there are not enough boundaries on a 1:250,000 quad map for area frame stratification or frame unit construction.
2. Enlarging the LANDSAT imagery (Black & White) to county map scale (1:126,720) and then transferring the county map to an acetate overlay. This proved to be successful for stratification and an aid in frame unit construction. There was information on the LANDSAT imagery for broad land use stratification using the following set of strata definitions:

| Stratum | Definition |
|---|---|
| 11 | Intensively cultivated land - 75+ percent of land cultivated. |
| 31 | Agricultural Urban - Residential mixed with agriculture. |
| 32 | Urban - Residential or Industrial. |

| Stratum | Definition |
| --- | --- |
| 40 | Rangeland - Less than 15% cultivated. |
| 50 | Non-Agricultural - National Parks, Military, Mountains, etc. |
| 60 | Water - Actual & Proposed. |

There was not enough land area in the 15% - 75% cultivation range to create additional strata. The detail in the California imagery (black & white) was sufficient for stratification. However, other geographic areas of the U.S. may require color LANDSAT imagery. The use of color LANDSAT imagery and county maps will be discussed next.

Strata boundaries were drawn on an acetate county map. The next objective in area frame construction is frame unit (count unit) construction. Initially in addressing the question of whether LANDSAT imagery can replace aerial photography in current area frame construction methodology, frame unit construction was attempted using only LANDSAT imagery and the county map. The following conventional frame unit (count unit) target sizes were used for the various strata.

| Stratum | Target Frame Unit Size (sq. miles) | Range (sq. miles) |
| --- | --- | --- |
| 11 | 10 | 2-18 |
| 31 | - | .2-4 |
| 32 | - | .2-3 |
| 40 | 45 | 5-120 |
| 50 | 45 | 5-120 |
| 60 | - | 1 up |

The conclusion was that not enough permanent boundaries could be recognized without introducing more variability in frame unit size. This may not be a serious restriction. However, some larger frame unit sizes could lead to more expense in segment sample selection.

If conventional frame unit target sizes are the objective, then it will be necessary to use ASCS photo index sheets for some of the permanent boundaries. As in conventional area frame construction, boundaries from the aerial photographs are sometimes used even if they do not appear on a county map. The basic situation where aerial photographs were needed instead of color LANDSAT imagery was the identification of narrow dirt roads that could be used as frame unit boundaries.

The identification of urban and agricultural urban stratum boundaries or frame unit boundaries using LANDSAT imagery and a county map is not acceptable. The best source of information for cities remains to be the most current aerial photography available. LANDSAT imagery can possibly provide good boundaries using rather expensive image enhancement techniques. Investigation of alternatives for using LANDSAT imagery for current urban and agricultural urban boundaries is recommended as a continuing research effort.

3. Reducing the county map acetate overlay to the scale of a 1:250,000 LANDSAT image was another method attempted. This method seemed to offer the best use of county map boundaries and the spectral information in LANDSAT imagery.

   This method has several advantages over method 2. More spectral information is retained for broad land use stratification on the false color composite LANDSAT image than from one band black and white imagery. If frame unit target sizes could be slightly altered without causing a significant increase in expense, then this method of using LANDSAT imagery for area frame construction could possibly stand alone with the exception of cities and agricultural urban areas.

4. Another method attempted was the use of a Baush & Lomb Zoom Transferscope for overlaying two products that have different scales. This method gives the best visual combination of the map and imagery overlayed but does not cover large enough areas at a usable resolution for broad land use stratification and frame unit construction.

   Further investigation of overlaying enhanced LANDSAT images with USGS 7 1/2' quad maps to outline city and agricultural urban boundaries seem warranted under this alternative.

5. A method that was not investigated but undoubtedly would improve the performance of methods (1-4) is the use of computer enhanced LANDSAT images. There is more information for photo interpretation purposes in enhanced images but the cost per image of $750 is presently prohibitive.

B. MACHINE ANALYSIS OF LANDSAT DATA FOR CONTROL DATA IN AREA SAMPLING FRAMES

SRS has considerable experience in the efficiency gains possible by using a sampling frame with control data for each unit as opposed to a frame without control data. For example, the use of a list frame with livestock control data for each farm is more efficient than the use of a list frame without livestock control data.

Thus, one of the desirable potential properties of LANDSAT data is to associate classified[2/] crop and land use data with each area frame unit. Only in the last two years has this capability been developed. The process of accurate registration of a map base area to the LANDSAT data with a root mean square error of approximately one-half pixel for lines and columns is a necessity in associating LANDSAT data with a relatively small area on a map base such as area frame units.

Thus, research was conducted to develop the software and investigate the feasibility of using categorized LANDSAT data as control data. Kings and Tulare Counties were analyzed for this purpose using the following procedures:

1. Photo interpreting false color LANDSAT imagery to construct stratum boundaries on a county map acetate overlay.
2. Using county map boundaries, and ASCS aerial photos when necessary, construct frame units for each stratum.
3. Digitize each area frame unit using the stratum and frame unit number for identification.

---

2/ A description of the process of classifying LANDSAT digital data into crop or land use types is provided in Appendix A.

4. Register the frame unit boundaries to the LANDSAT coordinate system.
5. Register the JES sample segment and all field boundaries to the LANDSAT coordinate system.
6. Extract LANDSAT digital data for each crop or land use type to compute the mean vector and covariance matrix for the classification algorithm.
7. Empirically, attempt to evaluate the optimum classification strategy and then use the selected strategy to categorize the LANDSAT data for the whole county.
8. Extract the classified LANDSAT data for each area frame unit.
9. Create an index of control data that is a function of the classified data for each frame unit. For example, a cultivated land index might be the sum of all crop pixels divided by the total number of pixels for each frame unit. Other types of indices could also easily be developed.
10. Investigate the use of a cultivated land index or crop index for stratification or sub-stratification.
11. Consider the potential uses of major crop control data for an area frame.

Results of the analysis are included in Tables 5-8. Table 5, page 18, shows an example of the cultivation index applied to Kings County frame units. In Table 6 on page 19, the frame units were ranked by the cultivated land index for Kings County across all photo interpreted land use strata. Misclassification of several units was obvious with some city and rangeland units with cultivated land indices larger than some intensive agricultural land frame units. In Table 7 on page 20, the frame units were ranked by the cultivated land index within each original photo interpreted stratum. The index within a stratum can be used for sub-stratification of frame units.

In Kings County, a single LANDSAT data classification algorithm was used for the entire county. Figure 5 in Appendix C shows the pictorial color display (DICOMED print) for the Kings County classification. A color code is assigned to each crop or land use type used in the classification. The color print does give a visual display of control data for the area frame.

Such an algorithm does not take into account any prior geographic knowledge such as broad land use stratification. The damaging effects of this algorithm can be seen in Table 6 where range frame units have cultivated land indices as large as .72. Thus, in Tulare County, a new algorithm was used. Different crop and land use categories were used for the different photo interpreted strata. For example, cropland would not be a valid category for Yosemite National Park. The terminology used for this procedure in the remote sensing scientific community is "masked classification." The masked classification algorithm takes prior geographic and land use information into account. As seen in Table 8 on page 21, masked classification provided cultivated land indices with value zero for all non-cultivated strata frame units. Another major use of the classified LANDSAT data is demonstrated in the Section "County Crop Acreage Estimation." The classified LANDSAT data for an area such as a county is a necessary ingredient for both regression or ratio estimates using LANDSAT data and JES segment data.

The difficulty with the results in Tables 5 - 8 is that ground data were not available for entire frame units to evaluate the variability of misclassification between frame units. The only data available for evaluation was provided by a few individual farms. Presentation of the data could possibly divulge individual farm data and therefore will not be presented in any tables. This data, limited in volume, did however indicate a high degree of variability in misclassification between frame units. It is also unfortunate that the evaluation of the quality of area frame construction relies heavily on an operational sample to determine actual precision for various agricultural survey items.

Table 5
Cultivation Index
(Kings County Data)

| Area Frame Unit Frame Strata- Unit Number | Cotton Acres** | Wheat Acres** | Barley Acres** | . . . | Range or Waste Acres** | Total Acres** | Cultivation* Index (CI) |
|---|---|---|---|---|---|---|---|
| | Crop or land use types used in classification statistics file | | | | | | |
| 11-1 | 1621 | 692 | 381 | . . . | 1188 | 5215 | .7721 |
| 11-2 | 2148 | 237 | 88 | . . . | 1681 | 6149 | .7266 |
| . | . | . | . | | . | . | . |
| . | . | . | . | | . | . | . |
| 11-75 | 3082 | 2184 | 4968 | . . . | 3106 | 14918 | .7918 |
| 31-1 | 106 | 76 | 31 | . . . | 510 | 918 | .4439 |
| 31-2 | 404 | 92 | 46 | . . . | 925 | 1908 | .5153 |
| . | . | . | . | | . | . | . |
| . | . | . | . | | . | . | . |
| 31-13 | 3 | 59 | 122 | . . . | 269 | 183 | .4432 |
| 40-1 | 101 | 5452 | 20040 | . . . | 13904 | 42360 | .6717 |
| 40-2 | 2099 | 2298 | 8929 | . . . | 17146 | 36095 | .5327 |
| . | . | . | . | | . | . | . |
| . | . | . | . | | . | . | . |
| 40-9 | 2286 | 3994 | 3482 | . . . | 11934 | 24632 | .5155 |
| 50-1 | 24 | 301 | 292 | . . . | 1185 | 2715 | .5635 |

*Cultivation Index (CI) $= \dfrac{\text{COTTON+BARLEY+WHEAT+MINOR CROPS (ACRES**)}}{\text{TOTAL ACRES **}}$

where $0 \leq CI \leq 1$

More generally, an index (GCI), that is a function of the acres for the individual cover types and total acres for each frame unit, could be of use for special purpose surveys.

$GCI = f(C_1, C_2, \ldots C_n; T_A)$ where $n \leq p$ and p=number of crop or land use types categorized.

$T_A$ = total acres for frame unit

$C_i$ = total acres for i'th crop or land use type, $i=1, 2, \ldots n$

**All acres have been converted from pixels using a standard adjustment factor.

Table 6

Kings County - Area Frame Units Ranked Across Photo Interpreted Land Use Strata Using the Cultivation Index

| Rank | Cultivation Index | Stratum - Frame Unit | Rank | Cultivation Index | Stratum - Frame Unit |
|---|---|---|---|---|---|
| 1 | .9785 | 11 - 60 | 51 | .7578 | 11 - 17 |
| 2 | .9507 | 11 - 45 | 52 | .7561 | 11 - 3 |
| 3 | .9488 | 11 - 44 | 53 | .7521 | 11 - 27 |
| 4 | .9445 | 11 - 62 | 54 | .7409 | 11 - 18 |
| 5 | .9444 | 11 - 61 | 55 | .7377 | 11 - 16 |
| 6 | .9300 | 11 - 63 | 56 | .7327 | 11 - 21 |
| 7 | .9275 | 11 - 66 | 57 | .7266 | 11 - 2 |
| 8 | .9181 | 11 - 52 | 58 | .7248 | 11 - 19 |
| 9 | .9111 | 11 - 51 | 59 | .7232 | 11 - 9 |
| 10 | .9018 | 11 - 50 | 60 | .7230 | 40 - 4 |
| 11 | .8984 | 11 - 39 | 61 | .7203 | 11 - 54 |
| 12 | .8952 | 11 - 64 | 62 | .7195 | 11 - 57 |
| 13 | .8932 | 11 - 65 | 63 | .7171 | 11 - 32 |
| 14 | .8928 | 11 - 6 | 64 | .7130 | 11 - 30 |
| 15 | .8884 | 11 - 43 | 65 | .7099 | 31 - 12 |
| 16 | .8833 | 11 - 40 | 66 | .7083 | 11 - 15 |
| 17 | .8823 | 11 - 34 | 67 | .7034 | 11 - 70 |
| 18 | .8806 | 11 - 67 | 68 | .7029 | 11 - 37 |
| 19 | .8777 | 11 - 46 | 69 | .6997 | 11 - 68 |
| 20 | .8753 | 11 - 55 | 70 | .6932 | 11 - 71 |
| 21 | .8662 | 11 - 47 | 71 | .6717 | 31 - 1 |
| 22 | .8611 | 11 - 13 | 72 | .6684 | 11 - 36 |
| 23 | .8571 | 11 - 41 | 73 | .6632 | 11 - 73 |
| 24 | .8553 | 11 - 69 | 74 | .6627 | 11 - 33 |
| 25 | .8403 | 11 - 7 | 75 | .6624 | 11 - 23 |
| 26 | .8381 | 11 - 56 | 76 | .6561 | 11 - 35 |
| 27 | .8364 | 11 - 53 | 77 | .6537 | 11 - 72 |
| 28 | .8297 | 11 - 4 | 78 | .6201 | 40 - 5 |
| 29 | .8296 | 11 - 49 | 79 | .6087 | 11 - 29 |
| 30 | .8254 | 11 - 8 | 80 | .6016 | 31 - 5 |
| 31 | .8184 | 11 - 20 | 81 | .5972 | 40 - 3 |
| 32 | .8182 | 11 - 12 | 82 | .5635 | 50 - 1 |
| 33 | .8164 | 11 - 74 | 83 | .5590 | 40 - 6 |
| 34 | .8149 | 11 - 25 | 84 | .5451 | 31 - 7 |
| 35 | .8117 | 11 - 58 | 85 | .5327 | 40 - 2 |
| 36 | .8103 | 11 - 24 | 86 | .5280 | 40 - 8 |
| 37 | .8055 | 11 - 5 | 87 | .5227 | 31 - 9 |
| 38 | .8014 | 11 - 26 | 88 | .5172 | 31 - 3 |
| 39 | .7988 | 11 - 11 | 89 | .5155 | 40 - 9 |
| 40 | .7942 | 11 - 22 | 90 | .5153 | 31 - 2 |
| 41 | .7918 | 11 - 75 | 91 | .5029 | 31 - 4 |
| 42 | .7904 | 11 - 38 | 92 | .4866 | 40 - 7 |
| 43 | .7890 | 11 - 10 | 93 | .4609 | 31 - 6 |
| 44 | .7837 | 11 - 48 | 94 | .4439 | 31 - 1 |
| 45 | .7750 | 11 - 42 | 95 | .4432 | 31 - 13 |
| 46 | .7726 | 11 - 28 | 96 | .4380 | 31 - 8 |
| 47 | .7721 | 11 - 1 | 97 | .4174 | 31 - 11 |
| 48 | .7708 | 11 - 31 | 98 | .3333 | 31 - 10 |
| 49 | .7682 | 11 - 14 | | | |
| 50 | .7641 | 11 - 59 | | | |

After the frame units have been ranked by the cultivation index, the frame units can be grouped into a user supplied number of groups for stratification.

Table 7

Kings County - Area Frame Units Ranked Within Photo Interpreted Land Use Stratum Using the Cultivation Index

| Rank | CI | Stratum - Frame Unit | Rank | CI | Stratum - Frame Unit |
|---|---|---|---|---|---|
| 1 | .9785 | 11 - 60 | 51 | .7578 | 11 - 17 |
| 2 | .9507 | 11 - 45 | 52 | .7561 | 11 - 3 |
| 3 | .9488 | 11 - 44 | 53 | .7521 | 11 - 27 |
| 4 | .9445 | 11 - 62 | 54 | .7409 | 11 - 18 |
| 5 | .9444 | 11 - 61 | 55 | .7377 | 11 - 16 |
| 6 | .9300 | 11 - 63 | 56 | .7327 | 11 - 21 |
| 7 | .9275 | 11 - 66 | 57 | .7266 | 11 - 2 |
| 8 | .9181 | 11 - 52 | 58 | .7248 | 11 - 19 |
| 9 | .9111 | 11 - 51 | 59 | .7232 | 11 - 9 |
| 10 | .9016 | 11 - 50 | 60 | .7203 | 11 - 54 |
| 11 | .8984 | 11 - 39 | 61 | .7195 | 11 - 57 |
| 12 | .8952 | 11 - 64 | 62 | .7171 | 11 - 32 |
| 13 | .8932 | 11 - 65 | 63 | .7130 | 11 - 30 |
| 14 | .8928 | 11 - 6 | 64 | .7083 | 11 - 15 |
| 15 | .8884 | 11 - 43 | 65 | .7034 | 11 - 70 |
| 16 | .8833 | 11 - 40 | 66 | .7029 | 11 - 37 |
| 17 | .8823 | 11 - 34 | 67 | .6997 | 11 - 68 |
| 18 | .8806 | 11 - 67 | 68 | .6932 | 11 - 71 |
| 19 | .8777 | 11 - 46 | 69 | .6684 | 11 - 36 |
| 20 | .8753 | 11 - 55 | 70 | .6632 | 11 - 73 |
| 21 | .8662 | 11 - 47 | 71 | .6627 | 11 - 33 |
| 22 | .8611 | 11 - 13 | 72 | .6624 | 11 - 23 |
| 23 | .8571 | 11 - 41 | 73 | .6561 | 11 - 35 |
| 24 | .8553 | 11 - 69 | 74 | .6535 | 11 - 72 |
| 25 | .8405 | 11 - 7 | 75 | .6087 | 11 - 29 |
| 26 | .8381 | 11 - 56 | | | |
| 27 | .8364 | 11 - 53 | 1 | .7099 | 31 - 12 |
| 28 | .8297 | 11 - 4 | 2 | .6016 | 31 - 5 |
| 29 | .8296 | 11 - 49 | 3 | .5451 | 31 - 7 |
| 30 | .8254 | 11 - 8 | 4 | .5227 | 31 - 9 |
| 31 | .8184 | 11 - 20 | 5 | .5172 | 31 - 3 |
| 32 | .8182 | 11 - 12 | 6 | .5153 | 31 - 2 |
| 33 | .8164 | 11 - 74 | 7 | .5029 | 31 - 4 |
| 34 | .8149 | 11 - 25 | 8 | .4609 | 31 - 6 |
| 35 | .8117 | 11 - 58 | 9 | .4439 | 31 - 1 |
| 36 | .8103 | 11 - 24 | 10 | .4432 | 31 - 13 |
| 37 | .8055 | 11 - 5 | 11 | .4380 | 31 - 8 |
| 38 | .8014 | 11 - 26 | 12 | .4174 | 31 - 11 |
| 39 | .7988 | 11 - 11 | 13 | .3333 | 31 - 10 |
| 40 | .7942 | 11 - 22 | | | |
| 41 | .7918 | 11 - 75 | 1 | .7230 | 40 - 4 |
| 42 | .7904 | 11 - 38 | 2 | .6717 | 40 - 1 |
| 43 | .7890 | 11 - 10 | 3 | .6201 | 40 - 5 |
| 44 | .7837 | 11 - 48 | 4 | .5972 | 40 - 3 |
| 45 | .7750 | 11 - 42 | 5 | .5590 | 40 - 6 |
| 46 | .7726 | 11 - 28 | 6 | .5327 | 40 - 2 |
| 47 | .7721 | 11 - 1 | 7 | .5280 | 40 - 8 |
| 48 | .7708 | 11 - 31 | 8 | .5155 | 40 - 9 |
| 49 | .7682 | 11 - 14 | 9 | .4866 | 40 - 7 |
| 50 | .7641 | 11 - 59 | | | |
| | | | 1 | .5635 | 50 - 1 |

After the frame units have been ranked by the cultivation index within land use stratum, they can be grouped into a user supplied number of groups for sub-stratification (paper stratification). For example, if the index was defined to be GCI = Cotton Acres/Total Acres, then the frame units could be ranked within land use stratum according to the proportion of cotton in each frame unit. An efficient sub-stratification using the ranked data could then be performed. In essence, this is considerably more information for sub-stratification of area frame units than geographic sub-stratification.

Table 8

Masked Classification* and the Cultivation Index
(Partial Tulare County Data)

| Area Frame Unit: Stratum - Frame Unit Number | Crop or Land Use Types in First Classification File: Citrus Acres | Barley Acres | Wheat Acres | . . . | Range or Waste Land | Land Use Types Used in Second Classification File: Water Acres | Dense Woods Acres | Total Acres | Cultivation Index |
|---|---|---|---|---|---|---|---|---|---|
| 11 - 48 | 1444 | 515 | 199 | . . . | 180 | 0 | 0 | 4146 | .9206 |
| 11 - 49 | 2838 | 364 | 340 | . . . | 293 | 0 | 0 | 7208 | .9154 |
| 11 - 10 | 2820 | 376 | 349 | . . . | 262 | 0 | 0 | 5883 | .9131 |
| 11 - 49 | 3380 | 350 | 200 | . . . | 354 | | | 6421 | .9000 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 11 - 6 | 296 | 264 | 602 | . . . | 2570 | 0 | 0 | 4671 | .3555 |
| 31 - 1 | 0 | 0 | 0 | . . . | 461 | 0 | 57 | 518 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 31 - 10 | 0 | 0 | 0 | . . . | 651 | 1 | 110 | 762 | 0 |
| 40 - 1 | 0 | 0 | 0 | . . . | 40692 | 1 | 7992 | 48685 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 40 - 20 | 0 | 0 | 0 | . . . | 41069 | 21 | 15591 | 56681 | 0 |
| 50 - 1 | 0 | 0 | 0 | . . . | 5630 | 1 | 22723 | 28354 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 50 - 5 | 0 | 0 | 0 | . . . | 432 | 1 | 29211 | 29644 | 0 |

*Masked classification, in this context, refers to the use of two or more crop or land use statistics files when classifying an entire county. In this application, two different crop or land use statistics files were used for classification. Photo interpreted stratum 11 had a statistics file with categories (citrus, barley, wheat, alfalfa, cotton and grapes combined, pasture, and range). Photo interpreted strata 31, 40, and 50 had a statistics file with categories (range, water, and dense woods). This type of classification (masked) takes advantage of prior geographic knowledge in designing the classification algorithm.

## V. RECOMMENDATIONS FOR AREA FRAME METHODOLOGY USING LANDSAT AND ASSOCIATED COSTS

There are several levels of potential use of remote sensing data (including developed software and hardware) for area frame construction. Each level will be discussed. Probably, in actual application, only one selected level would be practical to incorporate into an operational system.

### A. POTENTIAL USES OF LANDSAT DATA IN AREA FRAME CONSTRUCTION

The following levels of the utilization of LANDSAT data are recommended for consideration by the Agency.

#### 1. The Digitization of An Area Sampling Frame for Storage on Computer Tapes

One of the short-term benefits that existing remote sensing techniques hold for the area frame construction process is that of digitizing the area sample frames. Utilizing a data tablet digitizer and a plotter, along with the interactive EDITOR software subsystem, it is possible to digitize and record all delineated area frame unit boundaries.

To digitize an already constructed area frame for any given state, the map materials that would be required are county maps for every county in the state, and necessary USGS quadrangle 7 1/2 minute maps for city areas. With or without an acetate overlay on the maps, the frame unit boundaries could be outlined, labeled and the vertices digitized.

The output of this digitization process can be a plot of all the frame unit boundaries at a user supplied scale. The paper product plot can be readily reproduced as the digitized information is stored permanently on computer tapes. This process would solve the problem of the replacement of the existing paper materials used in area frame construction due to loss, normal wear and tear on aged paper maps, and possibly fire and water damage.

Another advantage of digitizing frame units is that planimetering would not be necessary since acreage measurements are obtained in the digitization process for all enclosed areas. The digitized acreage readings are generally more accurate than planimetering. Also, the edit process is a simple and accurate one. If the plotted digitized frame units overlay on the county maps correctly, then the acreage of each frame unit is known to be accurate. An example of such a plot is presented in Table 9 on page 30.

2. Using the Photo Interpretation of LANDSAT Imagery as a Tool in the Updating of a Problem Stratum in an Area Sampling Frame

In the Western United States where pivotal irrigation is being developed in former dryland or rangeland areas, updating an area frame stratum by subdividing it using current LANDSAT imagery into k substrata seems to be a logical statistical alternative to waiting for construction of a new frame. Basically the problem is having a k modal population instead of a uni-modal population and increased sample size alone won't entirely solve the lack of precision problem.

An example that initially attracted attention to this problem was the monitoring of an area in Kansas using LANDSAT imagery. An area in Southwest Kansas along the Arkansas River of approximately 385 square miles was formerly dryland and classified as stratum 40 in the 1975 Kansas Area Sampling Frame (Figure 1 in Appendix C). However, by looking at two LANDSAT images of the area in Figures 2 and 3 in Appendix C, it becomes apparent that there has been a substantial increase in the amount of pivotal irrigation (approximately 105 square miles). If the current LANDSAT imagery was used to update stratum 40 for Kansas by subdividing it into two strata, then the frame would be more efficient for several crop items.

Presently, the estimate for a crop item is of the form:

$$\hat{Y} = \sum_h N_h \cdot \bar{y}_h$$

where h=11, 12, 20, 31, 32, 33, 40, 50, 61.

If stratum 40 was subdivided into two strata (41, 42) and resampled then the form of the estimate would be:

$$\hat{Y} = \sum_h N_h \cdot \bar{y}_h$$

where h=11, 12, 20, 31, 32, 33, 41, 42, 50, 61.

Also, variance calculations for the area frame are currently made by paper stratum (geographic substratum) within land use stratum. Changes in land use patterns for parts of several paper strata could result in a substantial increase in variation due to only a few segments containing large concentrations of new cropland. In a study conducted by the Sampling Studies Section, 53 percent of the sample segments in the rangeland stratum in Kansas violated the stratum percent cultivated land definition.[3/]

Perhaps, recommendations from the states about areas of rapidly changing agricultural land use could be monitored by LANDSAT imagery for different periods in time. Areas which have significant changes in land use could then be reviewed to see if the problem is confined to one or two strata. If the problem is limited in the number of strata, then only those strata need to be updated and not the entire frame.

3. Using the Photo Interpretation of LANDSAT Imagery as a Tool in New Area Frame Construction

As demonstrated by the Kansas situation there is potential in photo interpretation of LANDSAT images as a supplemental tool for constructing new area frames along with the traditional mosaics of the latest flown ASCS photographs, county highway maps and park maps, etc. The main advantage of LANDSAT imagery is that it is a current representation of the area while ASCS photos can be several years old. In areas of the country that have undergone major land use changes this can be a substantial benefit.

By utilizing a color (non-classified) LANDSAT image and acetate products of county highway maps it is possible to overlay the two sources at identical scales. A broad land use stratification can be done with the image and map. Frame units can be constructed with consideration of natural boundaries as delineated on the county

---

[3/] Ciancio, N.; Rockwell, D.; Tortora, R., "An Empirical Study of the Area Frame Stratification," U.S. Department of Agriculture, Statistical Reporting Service, Washington, D.C., July 1977.

highway maps. However, the ASCS photo index sheets will be needed to supplement the frame unit construction process and the sample selection process.

4. <u>Area Frame Construction Using Manual Interpretation of LANDSAT Imagery As An Aid and the Digitization of the Completed Area Frame</u>

We feel that remote sensing techniques can best be utilized in the area frame construction process using manual photo interpretation of the unclassified LANDSAT data in land use stratification along with the ASCS photo index sheets and also the digitization of frame units and permanent storage of the information on computer tape. This level of operation would incorporate the most recent techniques that have been developed to date. It warrants serious consideration for use in an operational test project for a state.

5. <u>Use of LANDSAT Digital Data Classified into Ground Cover Types as Control Data in an Area Sampling Frame</u>

The objective is to extract classified LANDSAT data for each area frame unit. Research has demonstrated that this can be done. Given that the LANDSAT classified data is reasonably accurate, the potential for more efficient area sampling frames is good. Potentially, timely control data (major crop acreages) could be used in more efficient sub-stratification, post stratification, or even regression or ratio estimation for the major crop acreage items. Accurate control data also opens the avenue for more efficient special purpose area frame surveys for major crop items. The potential for using control data for each frame unit is discussed by Houseman.[4/]

However, the control data supplied by LANDSAT is, at present, of questionable value since the classified LANDSAT data accuracy for a large area cannot be directly associated with the frame unit level. The variability of classification accuracy between frame units is not available but is suspected to be substantial.

Several issues need further investigation before any quasi-operational system should even be considered. These include: use of multitemporal data to increase classification accuracy, increased use of prior geographic knowledge (masked classification) to increase classification accuracy, and if necessary, investigation of future LANDSAT's C and D to significantly improve classification accuracy.

If categorized LANDSAT data were to be seriously considered as control data, then several methods presently used in frame construction and sample design might not be applicable. The first item that would require investigation is frame unit construction. Using LANDSAT data, what is the optimum method of frame unit construction concerning size and homogeneity? The second question would be: What is the optimum use of the LANDSAT control data and how can it be taken into account in frame construction and sample design? More specifically, paper stratification prior to sampling would undoubtedly complicate the use of LANDSAT control data for post-stratification, and regression or ratio estimation. Perhaps, paper stratification

---

[4/] Houseman, Earl E., "Area Frame Sampling in Agriculture," U.S. Department of Agriculture, Statistical Reporting Service.

after sampling should be used in a state when LANDSAT control data is seriously considered as an operational technique.

## B. RESEARCH COSTS FOR THE ALTERNATIVE RECOMMENDATIONS

### 1. Digitization Only

| | |
|---|---|
| Digitizer | $10,000 |
| Plotter | 9,000 |
| 2 Terminals | 3,600 |
| Processing & Storage (1 State) | 1,500 |
| TOTAL | $24,100 |

### 2. LANDSAT Imagery for Problem Areas

| | |
|---|---|
| Black and White State Mosaic (1:1,000,000) | $150 |
| (12,500 sq. miles per frame) 1:250,000 Color LANDSAT | $100 each |
| County Maps on Acetate | $ 20 each |
| Approximated Total of 10 Problem Counties | $1350 |

### 3. LANDSAT as an Auxiliary Photo Interpretation Tool

| | |
|---|---|
| 1:250,000 Color LANDSAT | $100 x 15 = $1,500 |
| County Maps on Acetate (1:250,000) | $20 x 100 = $2,000 |
| TOTAL | $3,500 |

### 4. Digitization of Frame & LANDSAT as a Photo Interpretation Aid

Cost = (Items 1 + 3) = $27,600

### 5. Machine Analysis to Use LANDSAT Data as Control Data

Ground Truth Follow-up Survey (Entire State Level)

| | |
|---|---|
| Enumeration | $10,000 |
| Data Processing | $ 1,000 |
| Personnel | 1 man month |
| Edit | 2 man months |
| Training | 1 man week |

Digitization of Segments - 2 man months

| | |
|---|---|
| LANDSAT Data | $250/image |
| State Level | ($3,750) |
| Multitemporal | ($7,500) |

Machine Analysis - 2 man months

Data Processing & Storage - $10,000

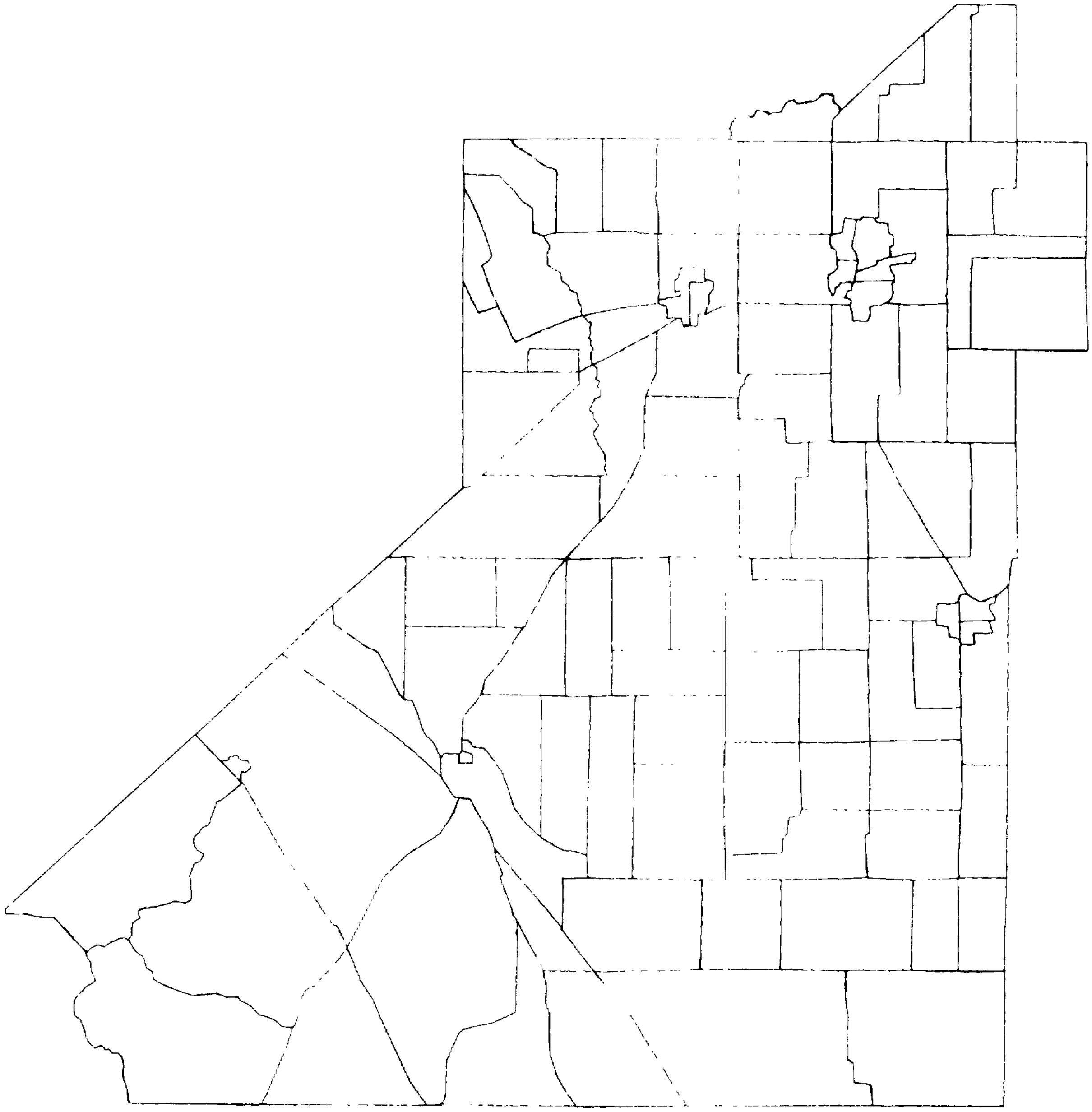
TOTAL ≈ $30,000 + 7.25 man months

C. SOFTWARE DEVELOPMENT, HARDWARE, AND PERSONNEL NEEDS

A revision is needed in the digitizing software to accomodate operational identifiers for the digitized frame units. This revision will be made by the Center for Advanced Computation at the University of Illinois. Adequate file transfer capabilities are required to use the information from the digitized area frame in operational sample selection programs at WCC. Possibly, the sample selection programs could be put into the EDITOR software at BBN in Boston, if this seemed to be a feasible alternative. Hardware requirements will probably include moving the SRS digitizer presently being operated at CAC. Personnel requirements for the photo interpretation uses of LANDSAT will involve training personnel in using LANDSAT data at different scales and for different spectral bands.

Table 9

Plot of Digitized Kings County Frame Units



## VI. COUNTY CROP ACREAGE ESTIMATION

While the primary goal of this study was to investigate the potential of utilizing LANDSAT data for construction of land area sampling frames, a useful by-product of the effort was crop acreage estimation for Kings and Tulare Counties. Direct expansion estimates using digitized JES field information for the different crops were calculated. Also regression and ratio estimates were computed using both ground information and classified LANDSAT data. For a detailed statistical explanation of the estimation procedures refer to Appendix B beginning on page 51.

Separate analyses were conducted using various classification procedures. See Appendix B on page 55 for a detailed description of the art of designing the classification algorithms. Initially each crop or land use type is clustered into distinct groups or categories and calculations made of the signature[5/] means and covariance matrix for the training set of labeled pixels (LANDSAT data resolution elements--slightly over one acre in size). These resulting statistics were then used to test the classification performance. Different clustering attempts for each crop or land use type were made until a set of statistics was obtained. Reference should be given to Appendix A beginning on page 43 for a further explanation of LANDSAT data, discriminant analysis, and clustering techniques.

---

[5/] Signature refers to the mean vector and covariance matrix for a specific crop or land use category and ideally is distinct or separable in the four dimensional LANDSAT scanner space from other categories.

Classification accuracy was also evaluated using different data sets for training and testing. For this study two methods, as described by Gray, were examined--Resubstitution and Holdout.[6/] Resubstitution is the method in which a training data set is also used as the testing data set. Results obtained in this manner tend to be overly optimistic as error rates are biased because the same data set is used for both training and testing. Where there was a large number of sample units the Holdout method was tried. This procedure uses a distinct sample of data to gather training statistics which are tested for a separate independent data set.

Because Kings County had only 14 segments and 143 fields, Resubstitution was used entirely in this county. However, with 32 segments and 666 fields in Tulare County, both procedures were used and evaluated. In Tulare County, the sample segments for the Holdout method were divided equally into two data sets from which one set was used for testing the classifier.

The use of different prior probabilities on classification performance was also evaluated. Table 14 on page 39 compares results of equal probabilities, identified as EP, and prior probabilities proportional to expanded reported acreage, identified as PER in the table, for the Resubstitution procedure on the Tulare County data set.

It was necessary to adjust the procedures of acreage estimation. Direct expansion estimates were based on only one stratum (intensively cultivated). Since size of segments in the rangeland stratum is so

---

[6/] Gray, H.L. and Schucany, W.R., The Generalized Jackknife Statistic, Dekker; New York, 1972.

variable the adjustments will not provide unbiased estimates. Sample units were pooled into one stratum for the regression estimates since the original area frame was not digitized. The original frame was not digitized because county crop acreage estimation was not the primary goal of the project.

Data analysis for Kings County as shown in Table 10 on page 37 was based on eight major cover categories. Using Resubstitution equal priors, the percent correct, that is, the percentage of the JES reported crop information that was classified correctly, ranged from 22 percent correct for alfalfa to 89 percent correct for safflower. The overall percent correct performance of the classifier for Kings County was 71 percent correct. The calculated r-squares for the major crops, with the exceptions of sorghum and alfalfa, were all quite encouraging - over .80. The two largest crops in the county, cotton and barley, had r-squares of .973 and .967 respectively.

Coefficients of variation for regression estimates for Kings County crops ranged from 7.5 percent for cotton to 48.3 percent for safflower. Relative efficiencies as defined in Appendix B of the regression estimator compared to thé direct expansion estimates were also quite significant for the two major crops. The relative efficiency for cotton was 34.5 and the relative efficiency for barley was 27.7. The results of the regression estimator were certainly affected by the fact that for each crop one or two segments with a high proportion of the crop influenced the strength of the linear relationship between categorized pixels and acres. A plot of the cotton data can be seen in Table 11 on page 36.

Comparisons of the direct expansion estimates, regression estimates, ratio estimates, and the county estimates published for Kings County are presented in Table 10 on page 37.

Table 12 on page 37 presents similar results for Tulare County using the Holdout training and testing procedure and equal priors. Sample estimation in this county was concentrated on eleven crops with cotton, alfalfa, grapes, citrus and other tree fruits comprising the major crops. Point estimates are not given for Tulare County. The reason is that the large size of Tulare County requires special software which is currently being developed.

Generally, results for Tulare County were not as favorable as in Kings County. There are various reasons which explain this fact. First, average field sizes in Tulare County tended to be much smaller than in Kings. Also more crops were introduced into the analysis which caused more difficult classification problems. In Tulare County at the time of the July 12th satellite pass, the spectral signatures of the LANDSAT data for cotton, alfalfa, and grapes were not highly separable. The signatures in two dimensions are displayed in Table 13 on page 38.

Percent correct for Tulare County ranged from 3 percent correct for tree fruit (except citrus) to 71 percent correct for rangeland. The overall percent correct for the county was 42 percent correct. Coefficients of determination (r-square) for several cover types were quite discouraging. Coefficients of determination for the five major crops ranged from a very low .143 for upland cotton to .761 for citrus. Only six cover types had r-squares above the .500 level. Regression estimate coefficients of variation for the Tulare County crops were also quite high - ranging from 30.5 percent to 69.1 percent.

Table 14 on page 39 compares the use of different prior strategies using Resubstitution in Tulare County. The use of prior probabilities proportional to the expanded reported acres generally resulted in higher r-squares. However, the changes were not significant.

Table 15 on page 39 shows the comparison of the r-squares for both Holdout and Resubstitution procedures in Tulare County. With the exception of one crop, grapes, the Holdout train/test procedure did not change the r-square values significantly. The r-square for grapes was the only result which showed a substantial difference between Resubstitution (.278) and the Holdout technique (.589). Only one Holdout sample was tested in the Tulare analysis although many more sampling combinations could have been randomly drawn. Because of the time factor and the fact that acreage estimation was not the primary objective in this project, all of the Holdout sampling combinations were not evaluated.

Finally, Table 16 on page 40 presents the average field sizes in the sampling JES segments for both Kings and Tulare Counties. For the three major crops in Kings County, cotton, barley, and winter wheat, fields averaged over 100.0 acres while the average field sizes of the sampling segments in Tulare County were considerably smaller.

Table 11

Plot of Digitized Cotton Acres vs. Cotton Pixels for Kings County Segments

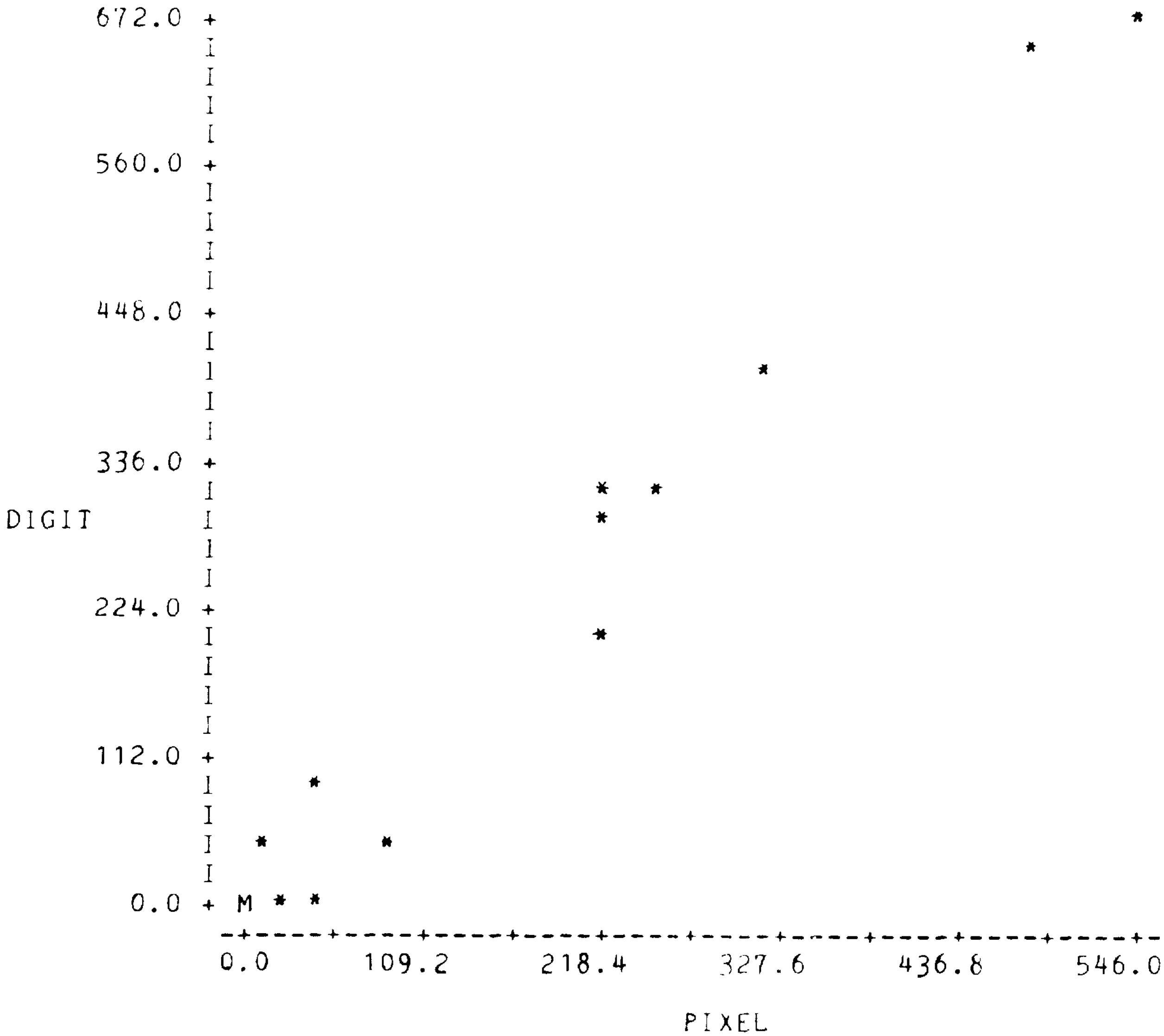


Table 10

Kings County Estimates (1976) - Resubstitution, Equal Priors

| Cover | Direct Expansion 4/ | | Regression Estimates 4/ | | | | | Ratio Estimate 4/ | | SSO County Estimate |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acres | C.V. | % Correct | $R^2$ | Acres | C.V. | R.E. | Acres | C.V. | Acres |
| Cotton (Upland) | 209,042 | 29.0 | 80.3 | .973 | 221,406 | 7.5 | 34.5 | 212,622.5 | 5.6 | 200,000 |
| Barley | 114,786 | 47.9 | 78.5 | .967 | 162,952 | 9.8 | 27.7 | 98,705.6 | 8.7 | 111,000 |
| Safflower | 49,313 | 99.3 | 89.3 | .996 | 10,793 | 48.3 | 218.8 | 17,009.2 | 12.3 | 18,000 1/ |
| Sorghum | 11,236 | 91.0 | 69.3 | .672 | 26,849 | 35.5 | 2.8 | 19,107.5 | 63.6 | 11,000 |
| Winter Wheat | 58,815 | 50.1 | 51.4 | .823 | 95,474 | 20.7 | 5.2 | 67,564.3 | 23.7 | 87,000 |
| Corn | 17,409 | 60.0 | 68.6 | .809 | 76,646 | 13.2 | 4.8 | 22,669.2 | 37.8 | 25,000 |
| Alfalfa | 27,327 | 52.5 | 21.6 | .668 | 28,399 | 47.3 | 2.8 | 23,226.8 | 29.7 | 56,000 1/ |
| Rangeland 3/ | 2/ | 2/ | NA | .908 | 217,153 | 10.6 | 10.0 | NA | NA | NA |
| OVERALL | | | 70.7 | | 839,672 | | | | | |

1/ County Estimates obtained from County Commissioner.

2/ Because of the variability of segment size for the rangeland stratum, an unbiased county estimate could not be obtained.

3/ Includes Wasteland.

4/ Rangeland stratum was not used in direct expansion and ratio estimation. Therefore, these estimates are not comparable to the regression estimates.

Table 12

Tulare County Estimates (1976) - Holdout, Equal Priors

| Cover | Direct Expansion | | Regression Estimates | | | | SSO County Estimate |
|---|---|---|---|---|---|---|---|
| | Acres | C.V. | % Correct | $R^2$ | C.V. | R.E. | Acres |
| Cotton, Upland | 100,870 | 34.3 | 36.6 | .143 | 33.5 | 1.09 | 138,000 |
| Alfalfa | 51,721 | 53.3 | 43.4 | .673 | 31.9 | 2.85 | 84,000 1/ |
| Corn | 36,379 | 66.4 | 24.9 | .001 | 69.1 | 0.93 | 52,000 |
| Wasteland | 87,825 | 47.0 | 33.4 | .617 | 30.5 | 2.44 | NA |
| Winter Wheat | 57,968 | 38.0 | 42.3 | .375 | 31.6 | 1.49 | 66,000 |
| Pasture | 57,049 | 59.1 | 4.6 | .001 | 61.7 | 0.93 | NA |
| Barley | 34,818 | 64.8 | 62.2 | .582 | 43.7 | 2.24 | 48,000 |
| Grapes | 146,527 | 50.2 | 35.4 | .589 | 33.7 | 2.27 | 2/ |
| Citrus | 52,640 | 68.0 | 55.3 | .761 | 34.6 | 3.91 | 2/ |
| Tree Fruit/except citrus | 54,018 | 52.8 | 3.3 | .447 | 41.1 | 1.69 | 2/ |
| Rangeland | 3/ | 3/ | 71.0 | .799 | 46.3 | 4.65 | NA |
| OVERALL | | | 42.2 | | | | |

1/ County Estimates obtained from County Commissioner.

2/ Current County Estimates have not been published to date.

3/ Because of the variability of segment size for the rangeland stratum, an unbiased county estimate could not be obtained.

Table 13

Plot Showing Lack of Signature Separability
for Alfalfa, Cotton, and Grapes in Tulare County

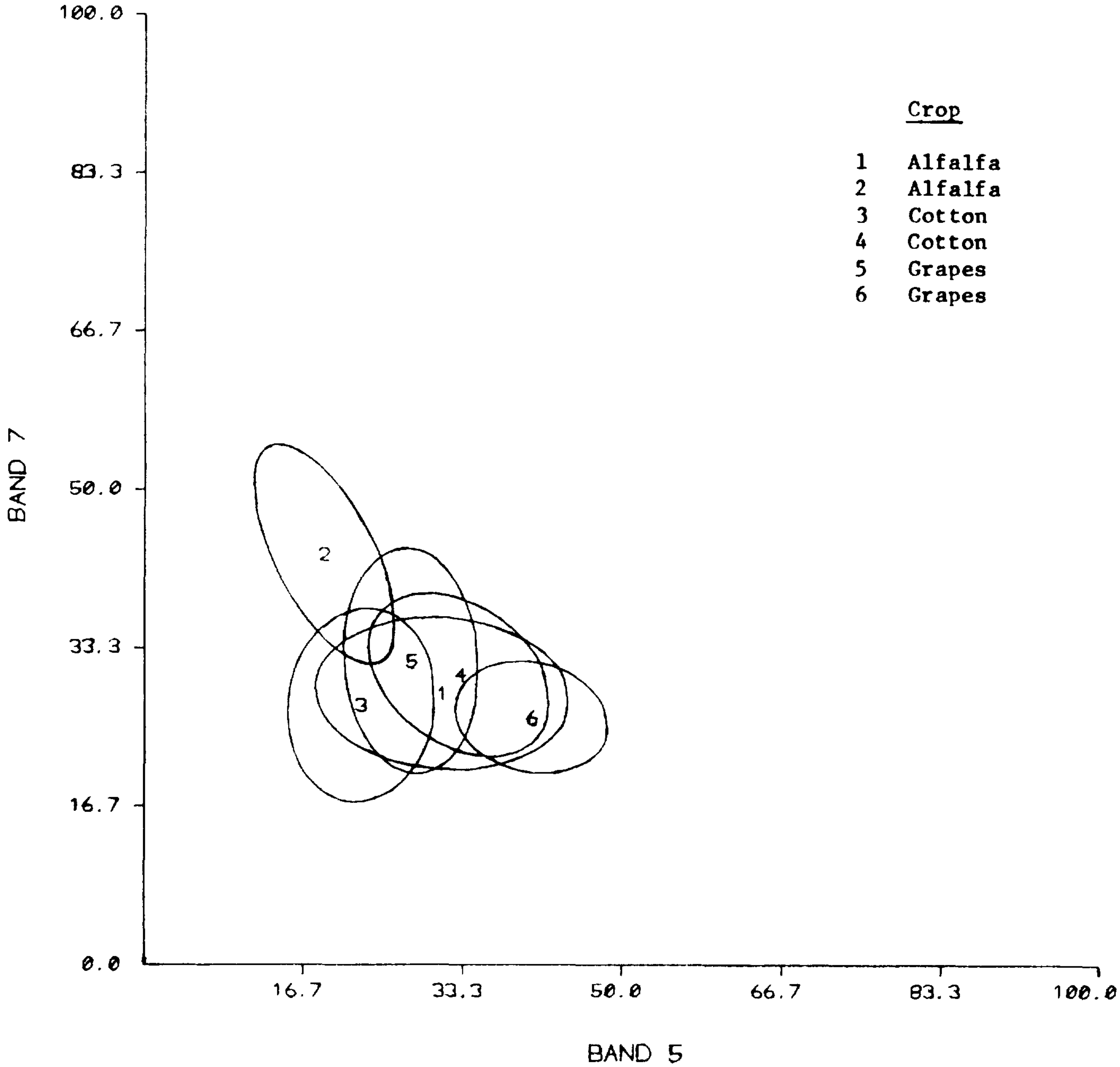


Table 14

Tulare County Comparisons of Prior Strategies Using Resubstitution

| | R-Squares | |
|---|---|---|
| | Priors | |
| Cover | EP | PER |
| Upland Cotton | .226 | .232 |
| Barley | .465 | .615 |
| Winter Wheat | .478 | .500 |
| Alfalfa | .671 | .653 |
| Corn | .017 | .069 |
| Grapes | .278 | .369 |
| Citrus | .719 | .724 |
| Tree Fruit other than Citrus | .558 | .596 |
| Permanent Pasture | .194 | .173 |
| Rangeland | .803 | .768 |
| Wasteland | .852 | .841 |

Table 15

Tulare County Comparisons of Resubstitution and Holdout Procedures with Equal Priors

| | R-Squares | |
|---|---|---|
| Cover | Resubstitution | Holdout |
| Cotton | .226 | .143 |
| Alfalfa | .671 | .673 |
| Corn | .017 | .001 |
| Winter Wheat | .478 | .375 |
| Barley | .465 | .582 |
| Grapes | .278 | .589 |
| Citrus | .719 | .761 |
| Tree Fruit other than Citrus | .558 | .447 |
| Rangeland | .803 | .799 |
| Pasture | .194 | .001 |
| Wasteland | .852 | .617 |

Table 16

Average **Field Size** of Sample Data

| | County | | | |
|---|---|---|---|---|
| | Kings | | Tulare | |
| | Number of Fields | Average Acres | Number of Fields | Average Acres |
| Cotton, Upland | 29 | 103.2 | 53 | 43.6 |
| Barley | 15 | 108.6 | 15 | 55.4 |
| Safflower | 1 | 710.0 | - | - |
| Winter Wheat | 8 | 104.6 | 25 | 43.3 |
| Alfalfa | 14 | 27.5 | 34 | 43.4 |
| Corn | 9 | 27.8 | 26 | 39.6 |
| Sorghum | 2 | 76.5 | 8 | 36.9 |
| Grapes | 2 | 4.0 | 34 | 62.7 |
| Tree Nuts | 2 | 5.8 | 29 | 27.2 |
| Citrus | - | - | 68 | 21.3 |
| **Tree Fruit/** except citrus | - | - | 65 | 16.1 |
| Permanent Pasture | 5 | 19.2 | 58 | 26.2 |
| Rangeland | 1 | 610.0 | 2 | 2579.3 |
| Wasteland | 20 | 40.4 | 60 | 16.0 |

## VII. CONCLUSION

We recommend an operational test effort using manual photo interpretation of LANDSAT imagery along with conventional tools to aid in the updating of an out-of-date land use frame or for current land use stratification for a new area frame. The resulting area frame should also be converted to computer tape for storage through the process of digitization. This level of effort will provide the training and systems necessary to use classified LANDSAT data as control data when it is appropriate. We feel it is too early to attempt using classified LANDSAT data as control data in area sampling frames. However, research on future LANDSAT's C and D and the use of multitemporal imagery to investigate the capabilities of classified LANDSAT data as control data is warranted.

## Appendix A

### Categorization or Classification Procedures

### A. Description of LANDSAT Data*

The satellite data used in this report is LANDSAT Multispectral Scanner (MSS) data and it is described in Section 3 of Data User's Handbook. 1/

The MSS is a passive electro-optical system that can record radiant energy from the scene being sensed. All energy coming to earth from the sun is either reflected, scattered, or absorbed, and subsequently, emitted by objects on earth. 2/ The total radiance from an object is composed of two components, reflected radiance and emitted radiance. In general, the reflected radiance forms a dominant portion of the total radiance from an object at shorter wavelengths of the electromagnetic spectrum, while the emissive radiance becomes greater at the longer wavelengths. The combination of these two sources of energy would represent the total spectral response of the object. This, then, is the "spectral signature" of an object and it is the differences between such signatures which allows the classification of objects using multivariate statistical techniques. This particular product in system corrected images refers to products that contain the radiometric and initial spatial corrections introduced during the film conversion. Every picture element (pixel) is recorded with 4 variables corresponding to one of the 4 MSS bands.

Sensor spectral band relationships.

| Sensor | Spectral Band Number | Wavelengths (micrometers) | Color | Band Code |
|---|---|---|---|---|
| MSS | 1 | .5 - .6 | Green | 4 |
| MSS | 2 | .6 - .7 | Red | 5 |
| MSS | 3 | .7 - .8 | Near Infrared | 6 |
| MSS | 4 | .8 - 1.1 | Infrared | 7 |

---

1/ Published by Goddard Space Flight Center.

2/ Baker, J.R. and E.M. Mikhail, Geometric Analysis and Restitution of Digital Multispectral Scanner Data Arrays. LARS information note 052875.

*Excerpted from Wigton, W. "The Technology of LANDSAT Imagery and Its Value in Crop Estimation for the U.S. Department of Agriculture." Statistical Reporting Service, March 1976.

B. Discriminant Analysis*

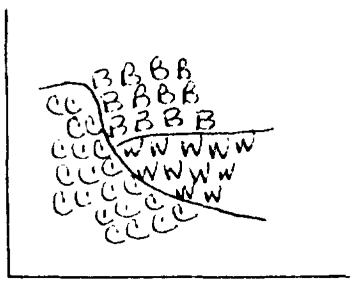
This background is intended to be general and enable the reader to understand the detailed computations and results in this report. Kendall and Stuart formulate Discriminant Analysis and Classification by stating . . .

> "We shall be concerned with problems of differentiating between two or more populations on the basis of multivariate measurements . . . We are given the existence of two or more populations and a sample of individuals from each. The problem is to set up a rule, based on measurements from these individuals, which will enable us to allot some new individual to the correct population when we do not know from which it emanates."

For example, the land population of interest was a portion of San Joaquin Valley in California. Cotton, wheat, and barley are the major crop populations of interest. From every acre in the San Joaquin Valley we have light intensity readings for green light, red light, and two infrared wavelengths. These light intensities are multivariate measurements that will be used to allot or classify each data point into a crop type such as cotton, wheat, or barley.

A sample of fields from each crop type is selected and their respective light intensities obtained. These sample points are plotted on a two-dimensional graph showing relative positions of each crop in the Measurement Space (MS). The problem is to partition the measurement space in some optimal fashion so that points are allotted as nearly correct as possible.

Figure A. Two-dimensional Measurement Space



There are many ways to partition a measurement space. We have done a simple non-statistical partition above, merely by drawing lines. Visually partitioning the measurement space may work when it is one or two dimensional, but for more than two dimensional measurement spaces, a visual partition is not possible. For most LANDSAT and aerial photography classification studies a four dimensional measurement space has been used.

---

*Excerpted from Wigton, W. "The Technology of LANDSAT Imagery and Its Value in Crop Estimation for the U.S. Department of Agriculture." Statistical Reporting Service, March 1976.

The method used in this report was that of constructing contour "surfaces" in the MS. These dividing surfaces were constructed so that points falling on the dividing surface have equal probabilities of being in either group on each side. Those points not on the dividing surface always have a greater probability of being classified into the crop for which the point is interior to the contour surface. If prior knowledge of the population density function indicates that the density is multivariate normal, then a multivariate normal density distribution will be estimated for each crop. It is hoped that the data is approximately multivariate normal since only the mean vector and covariance matrix is required to estimate a discriminant function. Usually small departures from normality will not invalidate the procedure, but certain types of departures (for example, bimodal data) may be very detrimental to the statistical technique. However, the error rate and estimator properties are dependent on the assumptions of the distributions and prior information.

For example, in this study a multivariate normal density was assumed so it becomes quite simple to estimate the density functions and the discriminant scores which in turn determine boundaries.

The discriminant score for ith population is:

$$P_i \, (2\pi)^{-\frac{q}{2}} \, |\Sigma_i|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\chi-\mu_i)' \Sigma_i^{-1} (\chi-\mu_i)}$$

where $P_i$ is the prior probability for the ith crop

$\Sigma_i$ is the covariance matrix (qxq) for the ith crop

$\mu_i$ is the mean vector (q length) for the ith crop

$\chi$ is a set of measurements of an individual from the ith population.

or its equivalent discriminant score the $\log_{(e)}$ of

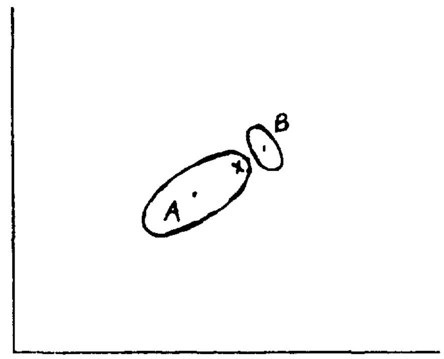
$$D_i = \log_e (P_i) - 1/2 \log_e |\Sigma_i| - 1/2 \, (\chi-\mu_i)' \, \Sigma_i^{-1} \, (\chi-\mu_i)$$

The boundary between two populations is quadratic (curved) and the point $\chi$ that falls in the boundary has an equal probability of being in either population.

When an unknown land point is classified, its measurement vector is compared to the mean vector for each crop represented. The point is assigned to the crop whose mean point is "nearest" from a statistical point.

The procedure used for finding the "nearest" mean uses the Mahalanobis measure of distance, not the Euclidean. This is illustrated in Figure B.

Figure B. Measurement Space Showing Two Crop Density Functions and An Unknown Point ($\chi$).




The point is actually closest (Euclidean distance) to the mean vector (center point) of B. However, when one takes into account the variance and covariances, $\chi$ is found to be closest to Group A based on a probability concept and an outlier of Group B. Therefore, the point would be classified into Group A, because the probability that the point ($\chi$) is a member of Group A is much greater than for Group B.

So the partitioning of the MS is done by computing the means for each crop type and using the Mahalanobis distances from this mean. This distance depends on the covariance matrix and is a measure of probability. The discriminant functions without prior probabilities are:

(1) $(X - \bar{X}_i)' S_i^{-1}(X - \bar{X}_i)$, which is a sample estimate of

$(X - \mu_i)' \Sigma_i^{-1}(X - \mu_i)$ if linear discriminant functions are used,

and

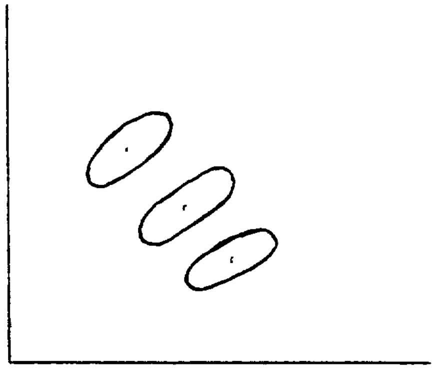
(2) $-1/2 \log_e |S_i| - 1/2 (X - \bar{X}_i)' S_i^{-1}(X - \bar{X}_i)$ if quadratic discriminant functions are used. These functions involve the exponent of the density formula of the multivariate normal distribution

$$C_{exp} = -1/2 \ (X - \mu_i)' \Sigma_i^{-1} (X - \mu_i)$$

of the i'th crop. If $\Sigma_i = \Sigma_j$ for all $i \neq j$ linear discriminant functions are used.

It is worth pointing out that if linear discriminant functions are used, one assumes (1) that $\Sigma_i = \Sigma_j$ and (2) that for all crops in the MS the major and minor axes are equal, and (3) the sample data for each crop has the same slope. Such an event in two-space is shown in Figure C.

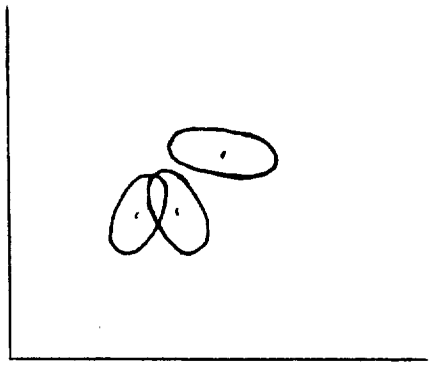
Figure C. Measurement Space Where Crop Types Have Same Covariance Matrix and Slope



This space can be partitioned effectively with straight lines. Thus, we can use linear discriminant functions.

Figure D shows a MS where covariance matrices are not equal, and therefore, linear discriminant functions are not appropriate. In either case, the Mahalanobis distance is used.

Figure D. Measurement Space When Crops Have Different Covariance Matrices



In Figure C, even though a common center point is not present, a common covariance (ellipse) matrix would be computed. In Figure D, a different covariance matrix will be needed for each crop type. When the off-diagonal elements in the covariance matrix are unequal, the slopes of the data are different and linear discriminant functions are <u>not</u> appropriate.

The above techniques follow from our first assumption that the data is normally distributed in the MS. In practice, however, one does not decide what the distribution of the population density is in the MS and program the correct procedure. One uses the available procedures for analyzing data. Most available programs assume multivariate normal data because the program and the calculations are greatly simplified.

In order to explain better how a parametric procedure can reduce the work load, consider that the first step in the discriminant analysis (DA) is to estimate the population density function in the MS, with a sample of points from each crop. Once these population density functions have been estimated, then partitioning the space is extremely simple.

To estimate a multivariate population density in MS for cotton where we have no prior information except sample data on cotton is extremely difficult. If a sample of 1000 points were available, each of these 1000 data points would need to be stored in the computer. On the other hand, if we are working with a multidimensional normal distribution, theory tells us that the sufficient statistics are computed (mean vector, and covariance matrix) and stored in the computer.

The individual data points could be discarded because no additional information about the population distribution in the MS is available in these points. (There would be information about how well the data fits the normal distribution in these 1000 data points).

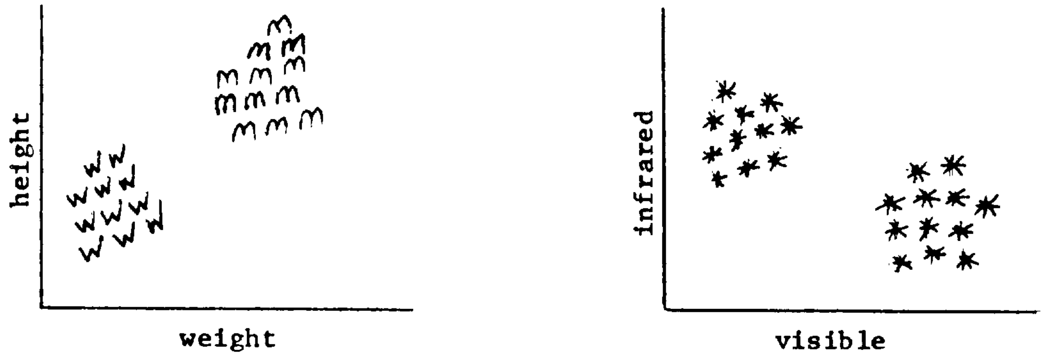
Another consideration is that all the techniques we have described require independent random samples from each crop in order to estimate the population density in the MS (training data). This point is mentioned because most remote sensing analysts do not work with randomly selected points. In this study, we have tried to work with randomly selected fields. However, the points within these fields are not a random sample of all possible points in a given crop, but the data are nested within fields. Consequently, the random selection is restricted to the selection of fields within the randomly selected segments.

One type of prior information that can be used in the classification procedure is the relative frequency or occurrence (prior probabilities) for each of the K populations in the total land population. For example, if 1/3 of all land is cotton, and 1/4 is barley, this information would be used and it would effect the partitioning of the measurement space accordingly. If a crop has a high chance of selection, then the area in the MS would be increased. Conversely, if a certain crop has a very low change of occurrence, then the area in MS would be adjusted **downwards**.

C. Clustering*

Clustering is a data analysis technique by which one attempts to determine the natural or "inherent" relationships in a set of observations or data points. To get an intuitive idea of what is meant by natural or inherent relationships in a set of data, consider the examples in Figure E. If one were to plot height versus weight for a random sample of students, without regard to sex, on a college campus, it is likely that two relatively distinct clusters of observations would result, one corresponding to the men in the sample (heavier and taller) and another corresponding to the women (lighter and shorter). Similarly, if the spectral reflectance of vegetation in a visible wave band, were plotted against reflectance in an infrared wave band, dry vegetation and green vegetation could be expected to form discernible clusters.

Figure E. Clustering Patterns




If the data of interest never involved more than two attributes (measurements or dimensions), cluster analysis might always be performed by visual evaluation of two-dimensional plots such as those in Figure E. But beyond two or possibly three dimensions, visual analysis is impossible. For such cases it is desirable to have a computer perform the cluster analysis and report the results in a useful fashion.

In regards to the application of clustering to remote sensing research, the greatest use of cluster analysis has been for the purpose of assuring that the data used to characterize the crop or land use classes do not seriously violate the assumption of Gaussian statistics. In general it may be expected that each distinct cluster center will correspond to a mode in the distribution of the data. Therefore, with the objective of defining a crop or land use subclass for each cluster center, the possibility of multimodal (and hence definitely non-Gaussian) crop or land use distributions is essentially eliminated.

A more detailed report on the technical development of several clustering algorithms, is provided by Swain.

---

*Excerpted from Swain, P.H., Pattern Recognition: A Basis for Remote Sensing Data Analysis. LARS information Note 111572.

# Appendix B

## Crop Acreage Estimation Procedures and Classifier Design Methods

### A. Direct Expansion Estimation (Ground Data Only)*

Aerial photography obtained from the Agricultural Stabilization and Conservation Service is photo-interpreted using the percent of cultivated land to define broad land-use strata. Within each stratum, the total area is divided into $N_h$ area frame units. This collection of area frame units** for all strata is called an area sampling frame. A simple random sample of $n_h$ units is drawn within each stratum. The Statistical Reporting Service then conducts a survey in late May, known as the June Enumerative Survey (JES). In this general purpose survey, acres devoted to each crop or land use are recorded for each field in the sampled area frame units. Intensive training of field statisticians and interviewers is conducted providing rigid controls to minimize non-sampling errors.

The scope of information collected on this survey is much broader than crop acreage alone. Items estimated from this survey include crop acres by intended utilization, grain storage on farms, livestock inventory by various weight categories, and agricultural labor and farm economic data.

Let h = 1, 2, . . . , L be the land-use strata. For a specific crop (corn, for example) the estimate of total crop acreage for all purposes and the estimated variance of the total are as follows.

Let $Y$ = Total corn acres for a state (Illinois, for example).

$\hat{Y}$ = Estimated total of corn acres for a state.

$y_{hj}$ = Total acres in the j<u>th</u> sample unit in the h<u>th</u> stratum.

Then,

$$\hat{Y} = \sum_{h=1}^{L} N_h \left( \sum_{j=1}^{n_h} y_{hj} \right) / n_h$$

---

*Excerpted from Sigman, Richard R.; Gleason, Chapman P.; Hanuschak, George A.; and Starbuck, Robert S.; "Stratified Acreage Estimation in the Illinois Crop-Acreage Experiment", <u>Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data</u>, Purdue University, West Lafayette, Indiana.

**In this context, all area frame units mean all the segments in the population and is not the same concept of area frame unit (count unit) used in the body of this report.

The estimated variance of the total is:

$$v(\hat{Y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h(n_h - 1)} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

Note that we have not yet made use of an auxiliary variable such as classified LANDSAT pixels. The estimator is commonly called a direct expansion estimate, and we will denote this by $\hat{Y}_{DE}$.

As an example, for the state of Illinois in 1975, the direct expansion estimates were:

Corn $\hat{Y}_{DE}$ = 11,408,070 Acres

Relative Sampling Error = 2.4% $= \sqrt{v(\hat{Y})} / \hat{Y}$

Soybeans $\hat{Y}_{DE}$ = 8,569,209

Relative Sampling Error = 2.9% $= \sqrt{v(\hat{Y})} / \hat{Y}$

B. Regression Estimation (Ground Data and Classified LANDSAT Data)

The regression estimator utilizes both ground data and classified LANDSAT pixels. The estimate of the total Y using this estimator is:

$$\hat{Y}_R = \sum_{h=1}^{L} N_h \cdot \bar{y}_{h(reg)}$$

where

$$\bar{y}_{h(reg)} = \bar{y}_h + \hat{b}_h(\bar{X}_h - \bar{x}_h)$$

and $\bar{y}_h$ = the average corn acres per sample unit from the ground survey for the hth land-use stratum

$$= \sum_{j=1}^{n_h} y_{hj} / n_h$$

$\hat{b}_h$ = the estimated regression coefficient for the hth land-use stratum when regressing ground-reported acres on classified pixels for the $n_h$ sample units.

$$\hat{b}_h = \frac{\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)(y_{hj} - \bar{y}_h)}{\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2}$$

$\bar{X}_h$ = the average number of pixels of corn per frame unit for <u>all</u> frame units in the h<u>th</u> land-use stratum. Thus <u>whole</u> LANDSAT frames must be classified to calculate $\bar{X}_h$. Note that this is the mean for the population and not the sample.

$$= \sum_{i=1}^{N_h} X_{hi}/N_h$$

$X_{hi}$ = number of pixels classified as corn in the i<u>th</u> area frame unit of the h<u>th</u> stratum.

$\bar{x}_h$ = the average number of pixels of corn per sample unit in the h<u>th</u> land-use stratum

$$= \sum_{j=1}^{n_h} x_{hj}/n_h$$

$x_{hj}$ = number of pixels classified as corn in the j<u>th</u> sample unit in the h<u>th</u> stratum.

The estimated (large sample) variance for the regression estimator is:

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{N_h^2}{n_h} \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 \cdot \frac{1 - r_h^2}{n_h - 2}$$

where

$r_h^2$ = sample coefficient of determination between reported corn acres and classified corn pixels in the h<u>th</u> land-use stratum.

$$r_h^2 = \frac{\left[\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)\ (x_{hj} - \bar{x}_h)\right]^2}{\left[\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2\right]\ \left[\sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2\right]}$$

Note that,

$$v(\hat{Y}_R) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h - 2}\ (1 - r_h^2)\ v(\hat{Y})$$

and so lim v $(\hat{Y}_R) = 0$ as $r_h^2 \rightarrow 1$ for fixed $n_h$. Thus a gain in lower variance properties is substantial if the coefficient of determination is large for most strata.

The relative efficiency of the regression estimator compared to the direct expansion estimator will be defined as the ratio of the respective variances:

$$\text{R.E.} = v(\hat{Y}_{DE})\ /\ v(\hat{Y}_R)$$

## C. Ratio Estimation*

A ratio estimate of the total Y for a particular cover type is:

$$\hat{Y}_{RATIO} = \sum_{h=1}^{L} (\bar{y}_h/\bar{x}_h)X_h$$

$$= \sum_{h=1}^{L} r_h X_h, \text{ where } r_h = \bar{y}_h/\bar{x}_h$$

---

*Excerpted from Ozga, Martin; Donovan, Walter E.; and Gleason, Chapman P.; "An Interactive System for Agricultural Acreage Estimates Using LANDSAT Data", Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

The variance of the ratio estimate is:

$$v(\hat{Y}_{RATIO}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} \left(S^2_{h,y} + r^2_h S^2_{h,x} - 2\, r_h\, \rho_h\, S_{h,y}\, S_{h,x}\right)$$

where,

$\rho_h$ = sample correlation coefficient between x and y for the h-th stratum

$S^2_{h,y}$ = sample variance for the h-th stratum for the y variate

$S^2_{h,x}$ is similarly defined.

## D. Designing a Classifier

The pixel classifier is a set of discriminant functions corresponding one-to-one with a set of classification categories. Each discriminant function consists of the category's likelihood probability multiplied by the category's prior probability. If the prior probabilities used are correct for the population of pixels being classified, then the resulting Bayes classifier minimizes the posterior probability of misclassifying a pixel for a 0-1 loss function.

In crop-acreage estimation, however, the objective is to minimize the variance of resulting acreage estimates. Since minimizing the posterior probability of misclassification does not necessarily achieve this objective, optimum acreage estimation may require the use of prior probabilities different than the optimum Bayes set.

For the case of multivariate normal signatures, the category likelihood functions are completely specified by the population means and covariances of the category signatures. Thus, the calculation of category discriminant functions involves the estimation of signature means and covariances and category prior probabilities.

Designing the classifier for this experiment consisted of the following steps:

1. Identification of classification categories.

2. Calculation of signature means and covariances and category prior probabilities from a training set of labeled pixels (called "training the classifier").

3. Measurement of classifier performance on a test set of labeled pixels (called "testing the classifier").

4. Heuristic optimistic of the classifier by repeating steps 1 through 3 for different numbers of categories and/or different prior probabilities, and then proceeding to step 5 for the "optimized" classifier.

5. Estimation of classifier performance in classifying the entire pixel population.

Because of the availability of ground data, which supplied the location and cover type of agricultural fields, supervised identification of classification categories was possible. A classification category was created for each cover type in which the number of training pixels exceeded a specified threshold, usually 100 pixels. In addition, a classification category for surface water was created using pixels from rivers, lakes, and ponds.

Appendix C

| Figure Number | Description |
|---|---|
| 1 | 1975 Kansas Area Sampling Frame |
| 2 | LANDSAT Image 1025-16565 - Kansas<br>August 17, 1972<br>Black and White - Band 5 |
| 3 | LANDSAT Image 2201-16451 - Kansas<br>August 11, 1975<br>Black and White - Band 5 |
| 4 | LANDSAT Image 2537-17480 - California<br>July 12, 1976<br>Color - Bands 4, 5, 7 |
| 5 | Classified LANDSAT Data<br>Kings County |

Figure 1

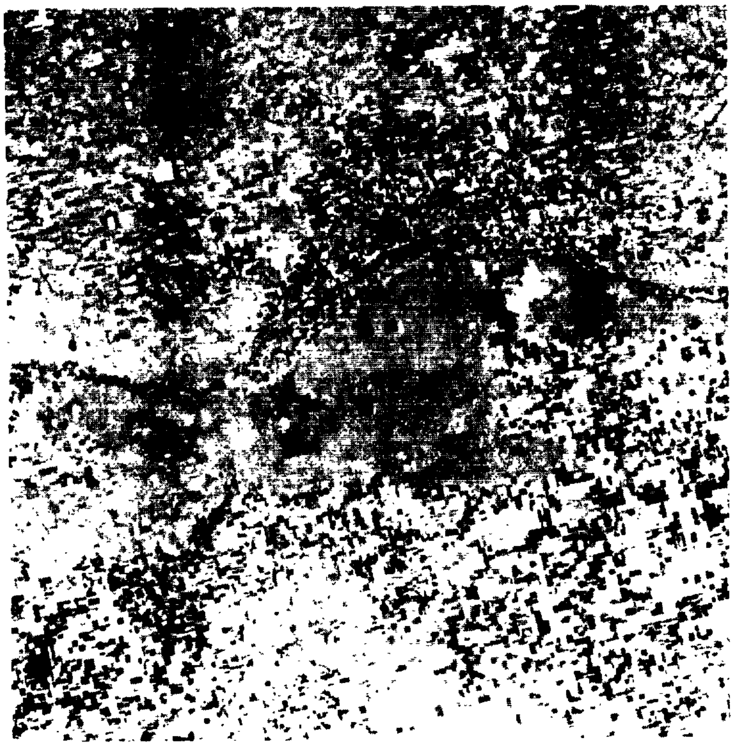
1975 Kansas Area Sampling Frame
(Photo on Next Page)

| Land Use | Stratum | Color |
|---|---|---|
| Intensive Cultivation (76% - 100%) | 11 | Pink |
| Intensive Cultivation (50% - 75%) | 12 | Pink |
| Extensive Cultivation (15% - 49%) | 20 | Light Blue |
| Agricultural Urban | 31 | Green |
| Urban | 32 | Green |
| Resort | 33 | Green |
| Rangeland, Forest | 40 | Orange |
| Non-Agricultural | 50 | Brown |
| Water | 62 | Dark Blue |

The picture of the broad land use stratification can be seen on the following page. The area enclosed in the black rectangle along the Arkansas River is the area of interest shown in Figures 2 and 3. This area is classified as rangeland in the 1975 Kansas Area Frame.

Figure 2

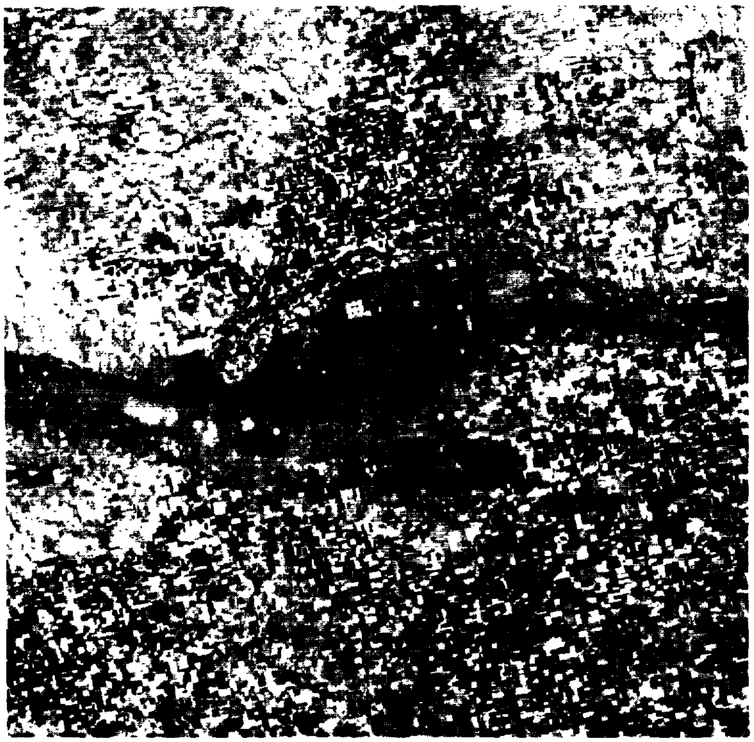
LANDSAT IMAGE 1025-16565
August 17, 1972
Black and White - Band 5



Area shown above is along the Arkansas River in the Garden City, Kansas Area. The picture clearly shows some pivotal irrigation fields on August 17, 1972. The same area can be seen in Figure 3 on August 11, 1975.

Figure 3

LANDSAT IMAGE 2201-16451
August 11, 1975
Black and White - Band 5



Area shown above is the same area as Figure 2 three years later. A substantial increase can be seen in the number of pivotal irrigation fields since the 1972 image.

Figure 4

LANDSAT IMAGE 2537-17480
San Joaquin Valley, California
July 12, 1976
False Color Composite
Bands (4, 5, 7)
(Photo on Next Page)

Figure 5

Classified LANDSAT Data
Kings County, California
(Photo on Next Page)

Each acre of land was computer classified into one of the following crop or land use types. Using the information from the classification, the color coded (picture-like) product on the next page is formed and is called a DICOMED **print**. Cities and Non-Agricultural Land were broken out and color coded prior to classification.

| Crop or Land Use | Color |
|---|---|
| Cotton | Red |
| Barley | Green |
| Cities | Orange |
| Range or Waste | Yellow |
| Winter Wheat | Brown |
| Other Crops or Forest | Dark Blue |
| Non-Agricultural | Purple |