The Future of Remote Sensing in US Crop Estimating Programs

Robert C. Hale
Mathematical Statistician
National Agricultural Statistics Service
U. S. Department of Agriculture
14th and Independence Ave. Room 4168
Washington, D.C. 20250
(202)-447-2206

Mickey B. Yost
Mathematical Statistician
National Agricultural Statistics Service
U. S. Department of Agriculture
14th and Independence Ave. Room 4168
Washington, D.C. 20250
(202)-447-2751

1988

ABSTRACT

The National Agricultural Statistics Service (NASS) is principally involved with crop acreage estimation in eight Midwestern states. To improve crop acreage estimates and to expand coverage to more states, NASS is developing new resource effective techniques to streamline processing. The use of desktop computers, more efficient data base management, and better use of field office personnel in digitization and processing are being considered. Data acquisition improvements and the use of TM data is also being considered. This paper will depict the current estimation procedure and the alternate techniques NASS is considering to improve this procedure.

SCOPE

MAT 1

<u>CURRENT PROGRAM:</u> The National Agricultural Statistics Service (NASS) is responsible for setting the official government estimates for agricultural items in the United States These items include crop acres planted and harvested, livestock numbers, and many economic items. In an effort to make these estimates as accurate as possible, NASS is using Landsat data as an auxiliary variable in a regression equation to estimate various crop items. Currently the Remote Sensing Section (RSS) of NASS uses Landsat data in Arkansas, Colorado, Illinois, Indiana, Iowa, Kansas, Missouri, and Oklahoma.

The RSS estimates winter wheat planted and harvested in Colorado, Kansas, Missouri, and Oklahoma. These states account for 43 percent of the winter wheat planted in the United States. Corn is estimated by the RSS in Illinois, Indiana, Iowa, and Missouri, which accounts for 41 percent of the nation's total. The RSS estimates soybeans in Arkansas, Illinois, Indiana, Iowa, and Missouri, which produce 51 percent of the total crop. The RSS also estimates cotton, rice, and sorghum in Arkansas and Missouri, which account for 7, 47, and 12 percent of these crops respectively. In Oklahoma, the winter wheat estimate is based on a multi-temporal classification using fall and spring scenes. The estimates in Arkansas and Missouri are also based on a multi-temporal classification but using spring and summer scenes.

FUTURE PROGRAM: At the present time there are no plans to expand the number of states in the program, especially complete states. The next expansion will probably be in a state or states where the crop of interest is concentrated in a specific area. This will allow for substantial improvements in the estimating program without the relative cost that would be incurred in processing an entire state.

SYSTEM LAYOUT

MAT 2

CURRENT PROGRAM: The RSS currently uses several types of computers to process the Landsat data. At the lowest level the analysts each use IBM PC's, primarily as terminals for communications to other computers. However, some steps are processed on the PC's. The RSS also uses a Forward Technology 3000 computer attached to digitizing tablets for segment calibration and scene registration.

The second level computer used is a DEC PDP-11/44 mini computer. The mini computer serves as a host computer for the video digitization system as well as a remote job entry (RJE) station for the main frame computer. In the role of a host computer, it stores the video images of the ground truth segment field boundaries. It also processes the video image into a segment mask, which relates the field boundary information to rows and columns in the Landsat scene. As a remote job entry station, it transfers various files to the main frame computer for use in other programs and receives output from jobs in the form of files and print documents.

The next level computer is an IBM main frame. The bulk of the data processing is accomplished on this computer. Both interactive and batch jobs are executed depending on the complexity and processing time necessary to accomplish the task. The main frame also provides access to two other computers. First, an on site IBM machine with a vector processor, which allows faster execution of sample level classification. Second is a Cray-XMP super computer where the full scene classifications are performed. Once the full scene classification is complete, the results are returned to the IBM main frame for further processing.

FUTURE PROGRAM: The RSS is converting some of the less computationally intensive main frame programs for processing on the PC's. This will allow the analyst the option to execute these programs off line if desired. In some cases, in an operational mode, these programs require so many input files that would have to be transferred to the PC that off line execution is not cost effective.

The Technology Research Section (TRS) of NASS will soon be testing a new Sun system. The system will consist of a server and two work stations networked together. One of the goals of the test is to convert all of the programs currently being used by the RSS for processing Landsat data to the Sun system, as well as to determine if a system like this is capable of handling all the processing needs of the RSS. Another goal is to determine how this system can use the classified Landsat scenes as an aid in the development of state area sampling frames.

ORDERING LANDSAT DATA

MAT 3

CURRENT PROGRAM: To maximize the probability of acquiring relatively cloud free Landsat data during the crop growing period, the RSS requests special acquisitions from Landsat 4. In the states where winter wheat estimates are made, the timing is from the middle of April to the end of May. In the states where corn and soybean estimates are made, the timing is from the middle of July to the end of August. Each of these time periods involves the special request of over 150 Landsat scenes.

To determine what Landsat data is available at any time, a staff member of the RSS twice weekly queries the EROS database located in Sioux Falls. As the database is queried, a log of the terminal session is stored on a PC. The file contains the path, row, scene id, cloud cover, band quality, and flight date. When the query is complete, this file is input to a program on the PC that reformats the information into a form that can be read into a database on the PC. Once the data is stored in the database several programs are used.

When an analyst wants to know what is available for a particular state, there is a program to query this database. Since the data stored here is a subset of what is stored on the EROS database, the search program is less complicated. When the analyst has decided the products to order for a scene, two more programs are run. One is to mark the scene with what products to order, such as 1:1,000,000 transparencies, photo products, or computer compatible tapes. The second program generates the actual order. The generated order shows the scene id, what products are requested, their cost, and the total cost of the order. A record of all products ordered is then stored in the database and at the end of each sesson a summary is run. The summary shows the total number of each product ordered and the total cost.

The products usually ordered are; 1:1,000,000 transparency of band 2, 1:250,000 photo product of band 2, and a band interleaved computer compatable tape. The transparency is oreded first to determine if any clouds that may be in the scene cover areas of the state that are being analized. If the scene is clear in the needed area then the photo product and the tape are ordered. The photo product is sent to the RSS where it is used to register the scene. The tape is sent to the super computer vendor where it will be read later.

FUTURE PROGRAM: No changes are expected to this system of ordering and record keeping.

GEOGRAPHIC REFERENCE

MAT 4

CURRENT PROGRAM: Classification of multi-channel Landsat data requires locating with precision the land area information (ground truth data) within the scene. Each pixel of information located in the scene must be tied to the same geographic base as the ground data. The process of relating Landsat row, column coordinates with ground truth latitude, longitude information requires a complicated process of registration, calibration, and digitization [Cook, 1982].

Registration of a Landsat scene is simply a mathematical process that translates the row and column coordinates of a pixel into latitude, longitude coordinates using a standard USGS map base. Calibration of a segment photo is similar to registration but instead locates the photograph of a segment on a USGS map using visual point definition. Roads, intersections, rivers with sharp bends, centers of lakes and other distinct points provide the basis for calibrating the segment photo to the USGS map. The USGS map then becomes the common link between the Landsat scene and the segment photograph.

New Landsat scenes must be registered to a USGS map each new crop year. New segments must also be calibrated to that same map base. However, only 20 percent of the sample needs to be calibrated, because segments remain in the sample for five years. The calibration marks drawn on the photo allow the analyst to specify exactly where the segment is located in terms of latitude and longitude, and therefore exactly where the segment is located in terms of the rows and columns of a Landsat scene.

<u>FUTURE PROGRAM:</u> The current program of registration and calibration is proving acceptable.

FIELD LEVEL DATA

MAT 5

CURRENT PROGRAM: Through a system of sample surveys, the USDA collects the ground data needed to classify Landsat information into groups of crop specific categories. The frame for these surveys consists of all the land area in the State of interest including intensively cultivated areas, cities, and towns. The land is divided into homogeneous areas, usually at the county level, and then subdivided into several blocks called primary sampling units (PSU's). PSU's range from 0.2 to 10 square miles depending on the intensity of land use. PSU's are then divided into segments of uniform size for sampling. Enumerators working through a cooperative agreement with the USDA personally visit selected segments of land and record, with the aid of the farm operator or other knowledgeable person, all information on land use in the segment. This field level information is recorded on survey questionnaires and drawn onto scaled and rectified photographs of the segment. These photographs are the basic source of field level information used in the classification process.

There are currently two methods of capturing field level information from the photographs. The first is a process of pinpointing with dots on an acetate tracing all intersecting lines, and changes in line direction, along with field, tract, and segment boundaries. These points are then manually digitized. The second process automates the hand dotting method by capturing boundary information using a television scanner. Both procedures, manual and video, use acetate tracings of field, tract, and segment boundaries taken from the segment photographs.

The files developed from these procedures then provide the necessary information to locate and use the Landsat information at the field level. For manually digitized segments, a single file called a segment network file is created. In this file resides all the information necessary to locate and identify the segment, tracts, and fields. It contains the location of the calibration marks that relate the segment to a map base, the tract and field labels that relate field identification with ground truth data, and finally the digital boundaries that locate the segment in the Landsat scene. For video digitized segments using the television scanning approach, three files are created: a calibration mark file, a label file, and a video scan mask file. These three files when processed together function identically to the manually digitized single segment network file.

The hardware and software systems used to handle the tasks of file creation and final mask creation are different for manual and video processing. Manually digitized segments utilize a hand held digitizer that relates the boundary dots to a map base. A North Star micro-computer handles the input data and writes one segment network file for each segment. These files are then transferred to the PDP-11/44 mini-computer located in Washington, DC using the Kermit file transfer protocol. From the PDP-11, the files are sent over high speed lines to an IBM main frame computer using the HASP protocol. Video digitized segments do not use the North Star mini-computer. All of the file creation is accomplished on the PDP-11. The PDP-11 creates a single file from the three input files (label, calibration mark, and scan mask). This file is then transferred to the IBM main frame computer using HASP.

As soon as the segment network file is created, whether from manual or video processing, the task of precision alignment of the segment and the Landsat scene must be done. The alignment process or segment shifting uses a one channel plotted representation of Landsat pixels called a greyscale, and the corresponding outline plot of the segment created from the segment network file. The outline plot is overlaid on the greyscale and shifted until the best match of segment field patterns and greyscales is achieved. The row, column shift to match greyscale shades with plot boundaries establishes the exact location of the segment in the Landsat scene.

FUTURE PROGRAM: New training and data collection techniques are being developed to increase the accuracy of ground truth data. The effects of new approaches to field level editing on non-sampling errors will be studied, as well as questionnaire design improvements.

The eventual elimination of manually digitized segments in favor of video processing will occur as the older North Star equipment begins to fail. Video processing improvements, especially the need for faster field labeling, are being considered. Improvements in file handling capabilities and the processing of scan mask, calibration, and label files into one file, as well as file transfers to the main frame are currently in the works.

An improved process of segment shifting is also being developed. Three channels of Landsat data can be displayed on an enhanced PC display screen with the segment outline plot superimposed over the image. The segment can then be shifted and the shift recorded in one step. The long process of printing greyscales, segment plots, and paper shifting will be eliminated.

LANDSAT DATA

MAT 6

<u>CURRENT PROGRAM:</u> Once the ground truth data has been collected and edited and the segment boundaries have been digitized, this information must be merged with the appropriate Landsat data. When the segment masks are generated, a corresponding file of row, column coordinates are also generated. The coordinates in this file define a rectangular block of Landsat pixels that completely cover the segment. The coordinate file is then transferred to the super computer where each of these blocks of Landsat data are read from tape. The file of these blocks of data is then transferred back to the main frame computer for further use.

To merge the data together, a program is run on the main frame computer. This program uses the segment mask file to get the field boundaries and labels, the ground truth file to get the cover of each field, and the Landsat data file to get the pixels and then write out several files. First is a file of all of the pixels within the boundary of the segment without regard to cover or internal field boundaries. This file will be used later for sample level classification.

From the segment mask all of the pixels that are on field boundaries can be determined and from the ground truth file the cover can be determined. The other files written consist of the pixels that are internal to the fields within the segment by cover. For example, one file would contain all of the pixels from all of the segments that were in soybean fields excluding any pixels that were on the borders of the soybean fields. This process continues until a file is written for every cover contained in each field across all segments in the sample.

When all of the cover specific files are generated, each is then clustered. The RSS uses two different clustering routines. The first is Classy clustering. This routine assumes an underlying Normal distribution and requires a fairly large number of pixels to be stable. The Classy clustering routine is generally executed on the super computer where the speed of that machine is advantageous. In this case the files of pixels are sent to the super computer and a file of individual cluster statistics is returned to the main frame computer. The statistics returned are the means, variances, and covariances of the each channel in the cluster. The other routine used is ordinary clustering. This routine also assumes an underlying Normal distribution, but it truncates the tails. This routine is used for smaller files (400 or less pixels) and is executed on the main frame computer. This routine also generates a file of cluster statistics.

When a statistics file is generated for each file of pixels, all of the files are then merged together into one file with each cluster being assigned a category number. In this format the various clusters from a cover can be compared. All of the means and variances are then printed and a measure of the distance between clusters is determined. The RSS use two different methods to determine the distance between clusters, Swain-Fu and transformed divergence. Both of these methods gives the analyst a way to determine which clusters are close (similar) and may be deleted. The final combined cluster statistics file is then used as the basis of the classification routine.

A maximum likelihood routine is used in classification. Each pixel from the segments in the sample is compared to the statistics of each cluster and assigned to the category which it would most likely fit. The classification at the sample level is run on the main frame computer with the vector processor. This allows for a fast turn around for the job since no file transfers are required and because of the added speed that the vector processor provides. The output is a file which is very similar to the input file of pixels, except instead of four energy readings per pixel there is only the one category number to which the pixel was classified.

FUTURE PROGRAM: The only changes foreseen in this area of processing is the increased use of the PC's. Both the ordinary clustering routine and the statistics file editing programs have been converted to run on the PC's. Since ordinary clustering is usually performed on smaller files, the cost of transferring the data to and the results back to the main frame computer can be saved by the off line processing. The statistics file editing is highly interactive and involves a considerable amount of time connected to the main frame computer. These two factors made this program a good candidate for off line processing. Here again, after the various cluster statistics files are merged, there is only one file file to be transferred between computers.

CROP ESTIMATION

MAT 7

CURRENT PROGRAM: There are four stages of estimation involved in the Landsat estimates made by the RSS. The first stage is sample level estimation. The output from sample level classification is summed by category to the segment level and the ground truth data is summed by cover to the segment level. This leads to a natural regression relationship between the number of pixels classified to a specific crop and the ground truth data. On the main frame computer, a regression equation is estimated using the least squares approach, along with a measure of the goodness of fit. The analyst reviews the regression equation to determine if improvements can be made. If so the statistics file is re-edited, sample level classification, and estimation are rerun until the analyst is satisfied no further improvement can be made. At that time a parameter file is written giving the estimates of alpha and beta and variances as well as the average number of pixels classified to the crop and number of acres per segment in the sample.

The second level of estimation is at the population level. In this case the population is the one or more scenes being analyzed. First all of the pixels in the scene in the area of interest must be classified. This is accomplished on the super computer. The same statistics file that was used in sample level classification is transferred to the super computer where the same maximum likelihood routine is used to classify the pixels. The results of the classification are transferred back to the main frame computer for further use. To make the population level estimates the following are required; the population level classification, the sample level parameter file, and the size of the population from which the sample was drawn. Parts of each of these pieces are used to estimate the total number of acres of a cover in the scene or scenes using the equation:

$$\hat{Y} = N[(\bar{y} + b(\bar{X} - \bar{x})]$$

Where, $\bar{X}$ is the average number of pixels classified to the cover of interest per population unit and is calculated by dividing the total number of pixels classified to the crop by the total population, N. $\bar{y}$, $\bar{x}$, and b are read from the sample level parameter file. This process is repeated until a population level estimates made for all of the scenes in the state.

After the population estimates are made across the state, the state level estimate is made. This estimate consists of two parts, first the population level estimates made for each scene and a direct expansion estimate for the parts of the state that are not covered by Landsat data. The analyst runs a program on the main frame computer that combines all of the population level estimates, determines where the direct expansion estimate is needed and calculates it, and print a table showing the combined population estimate, direct expansion estimate, and the total for the state. This program also calculates the variances of the estimates and make a comparison to the NASS area frame estimate.

Finally, the RSS makes estimates at the county level. When the Landsat scenes were classified on the super computer, the classification information was retained by county for use in the county estimates programs. On the main frame computer, the analyst runs two programs to begin the county estimates process. First is a program to prorate the direct expansion estimate to the counties. This proration is based on the the population size in each county. Next is a program that utilizes the Landsat information. This program uses the Battese-Fuller estimator to estimate the cover acres in the counties in a scene [Battese-Fuller, 1981]. The output from these programs are transferred to the PC's where all of the pieces needed to completely cover a county are merged together. A database program on the PC totals the data, calculates county variances, and prints the county estimates. The program also write the data to a floppy disk, which is sent to the NASS office in the state for use by the state office personnel.

FUTURE PROGRAM: The main thrust for improvement in this area is to improve the flow from one program to the next. Each of these programs require or produce several data files. The RSS is working on methods for these programs to more automatically flow from one to the other. Efforts are also being made to improve the methods used to calculate the county estimates.

NEW DATA SOURCES

MAT 8

CURRENT PROGRAM: Currently the RSS uses Landsat 4 and 5 MSS data and will continue to do so as long as possible. However, the Landsat 6 satellite will not have the MSS sensor, so NASS is reviewing alternative data sources. The thermal mapper data that is currently available and will be available on Landsat 6 is of excellent quality. But, because of its smaller pixel size, it needs four times as many pixels to cover the same area. This makes the use of pure TM data on an operational basis prohibitively expensive. Therefore, NASS is testing alternative emulated MSS data. NASS is testing four types of emulated MSS data; 1) sampled TM bands 2, 3, 5, and 4 where the sampling scheme is every other row every other column, 2) averaged TM bands 2, 3, 5, and 4 with the average being from a block of four pixels, 3) sampled TM bands 1, 7, 6, and 4 with the same sampling and 4) averaged TM bands 1, 7, 6, and 4 with the same averaging. The test data is from an area in Missouri where several different crops are grown. A replicated sample with three replications was drawn in the area. Each of the emulations were used identically in each replication. Early indications are that the RSS will use the averaged 2, 3, 5, and 4 TM bands. Final results are still pending.

FUTURE PROGRAM: Presently the RSS has nine SPOT scenes from western Kansas. Research will begin soon using these new data. Comparisons will be made between these data and the results from the same area using MSS data. The use of SPOT data in the NASS operational program is still very much in the future. The cost of acquisition and processing the data are to great. However, NASS is keeping abreast of new advancements in remote sensing, and the use of these new data sources.

RESULTS

MAT 9

CURRENT PROGRAM: Since 1978 when NASS began making crop area estimates with the aid of Landsat data the goal has been to make quality estimates with smaller standard errors than the operational program. This is not an easy task, since the area frame procedures used by NASS generate quit precise estimates, especially at a regional level. However, in nearly every case the RSS estimate has had a smaller standard error. The incidences where the Landsat estimate had a standard error essentially the same as the area frame estimate were situations where very little cloud free data was available for a state.

The comparison between the area frame and Landsat coefficients of variation in 1986 shows the greatest reduction in variance in cotton and rice. This reduction is due to two factors; first these two crops had the possibility for the largest improvement and second both of these crops were estimated using multi-temporal procedures.

Although the reduction in the coefficients of variation in the other crops appears to be small, these improvements amount to quite a a reduction in terms of acres. For example, the .5 percent reduction in the coefficient of variation in the winter wheat estimate translates into a standard error that is 140,000 acres smaller than the estimate from the area frame sample.

<u>FUTURE PROGRAM:</u> NASS is committed to the continued use of Landsat data as a supplement to it crop area estimation program. NASS feels that the use of Landsat data not only improves the quality of the estimates but keeps the agency abreast of new techniques that can be used. Research will continue on the use of new sensors as well as improving the use of current sensors. NASS will also look for new ways to use the data available for other parts of it's program. For example, improving the construction of area frames, crop production, or improving county estimates.

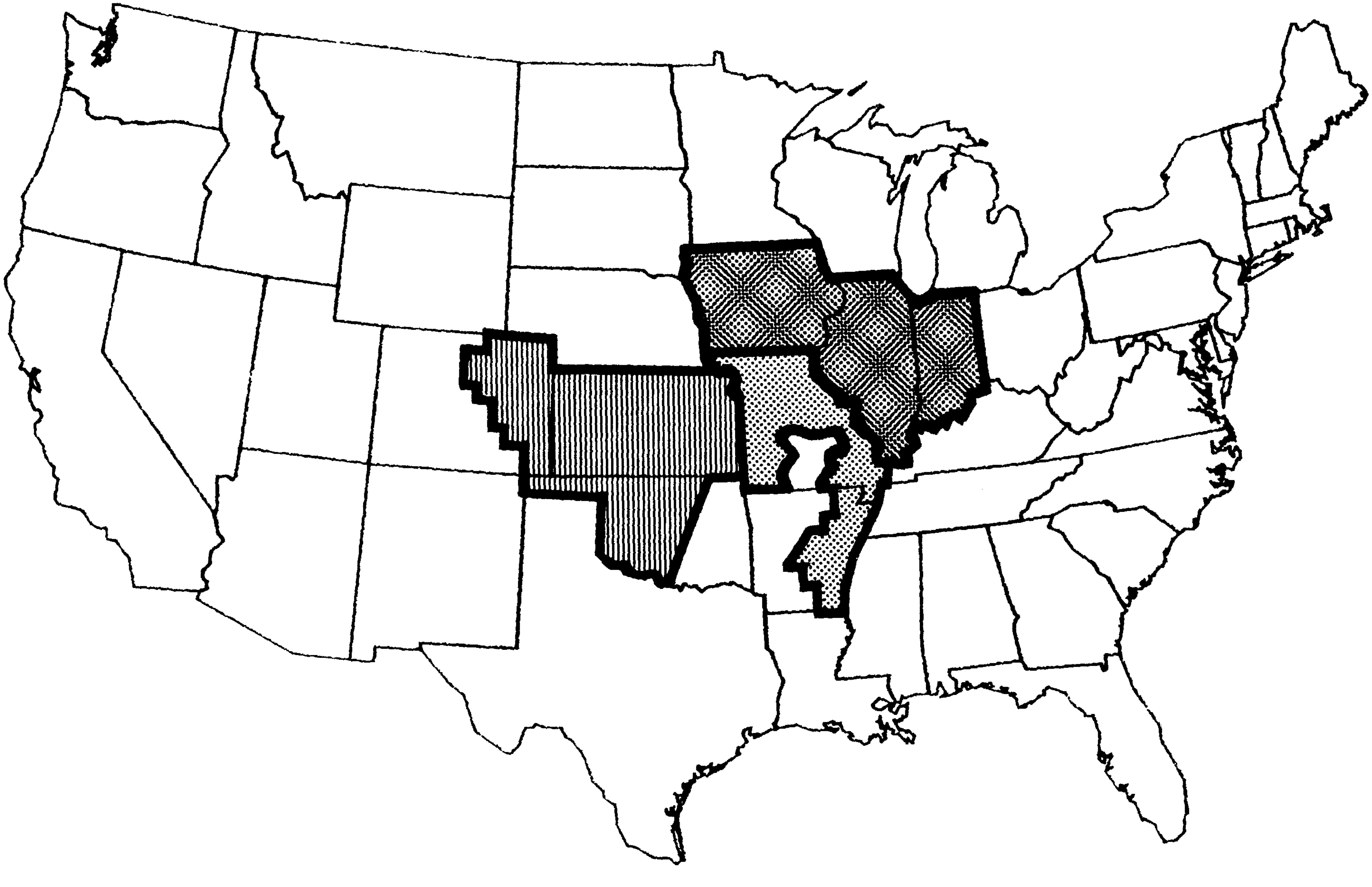
REFERENCES

Battese, G. E., W. A. Fuller. 1981. Prediction of County Crop Areas Using Survey and Satellite Data. Survey Section Proceedings, 1981 American Statistical Association Annual Meeting, Detroit, Michigan.

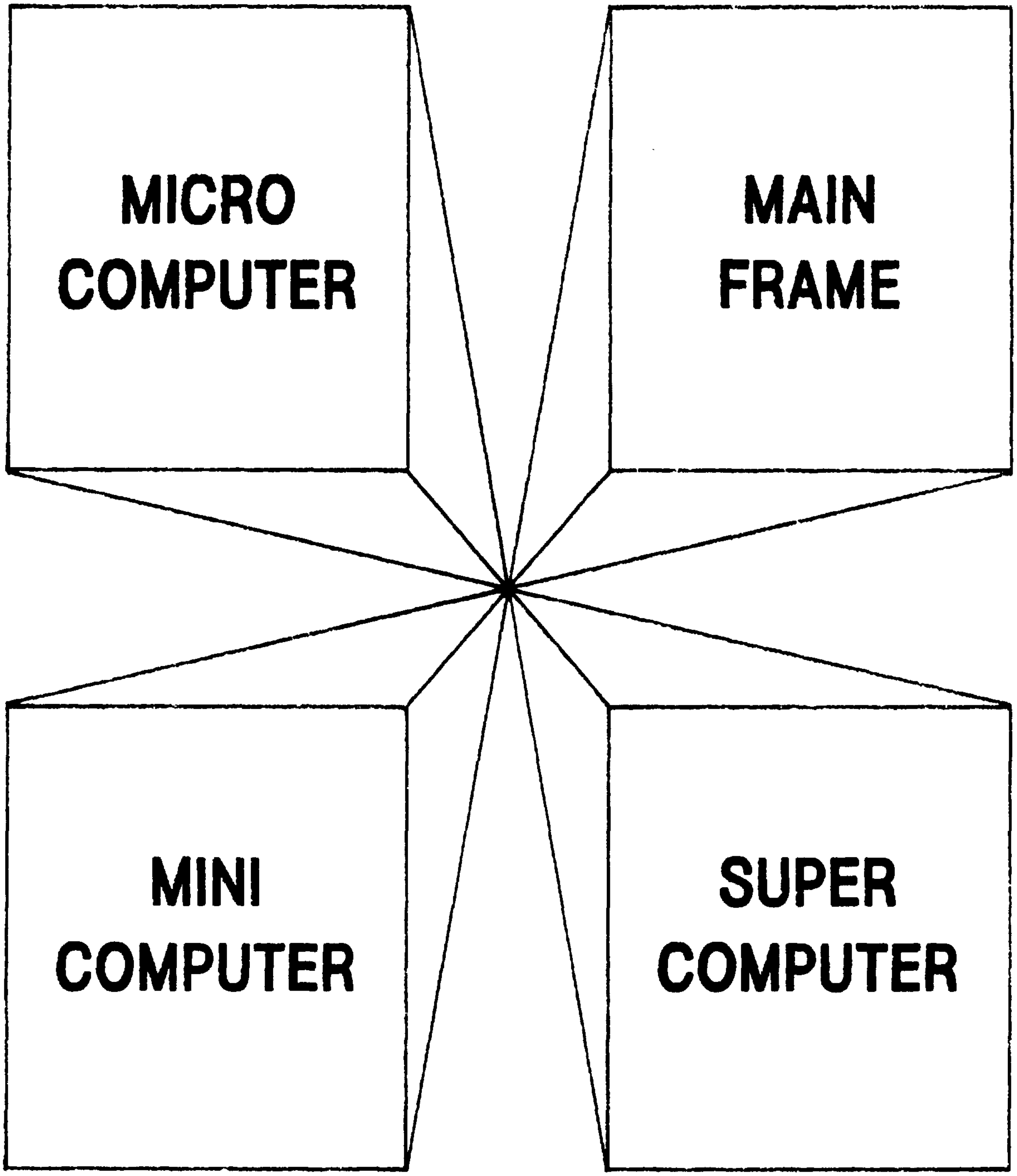
Cook, Paul W., Landsat Registration Methodology Used by U.S. Department of Agriculture's Statistical Reporting Service.

SCOPE OF COVERAGE

1

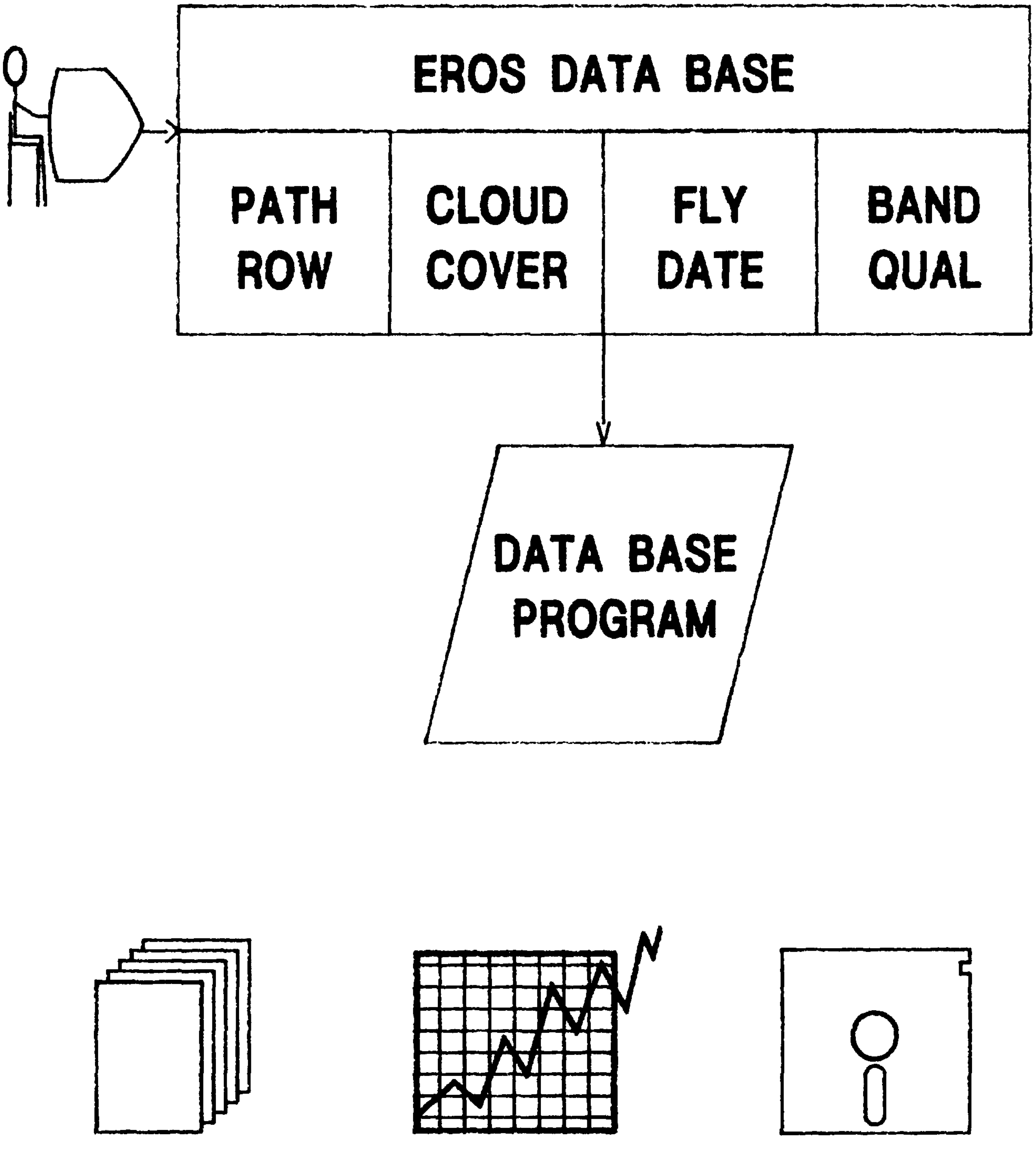


System Layout



Order Landsat Data

3



EROS DATA BASE

PATH ROW

CLOUD COVER

FLY DATE

BAND QUAL

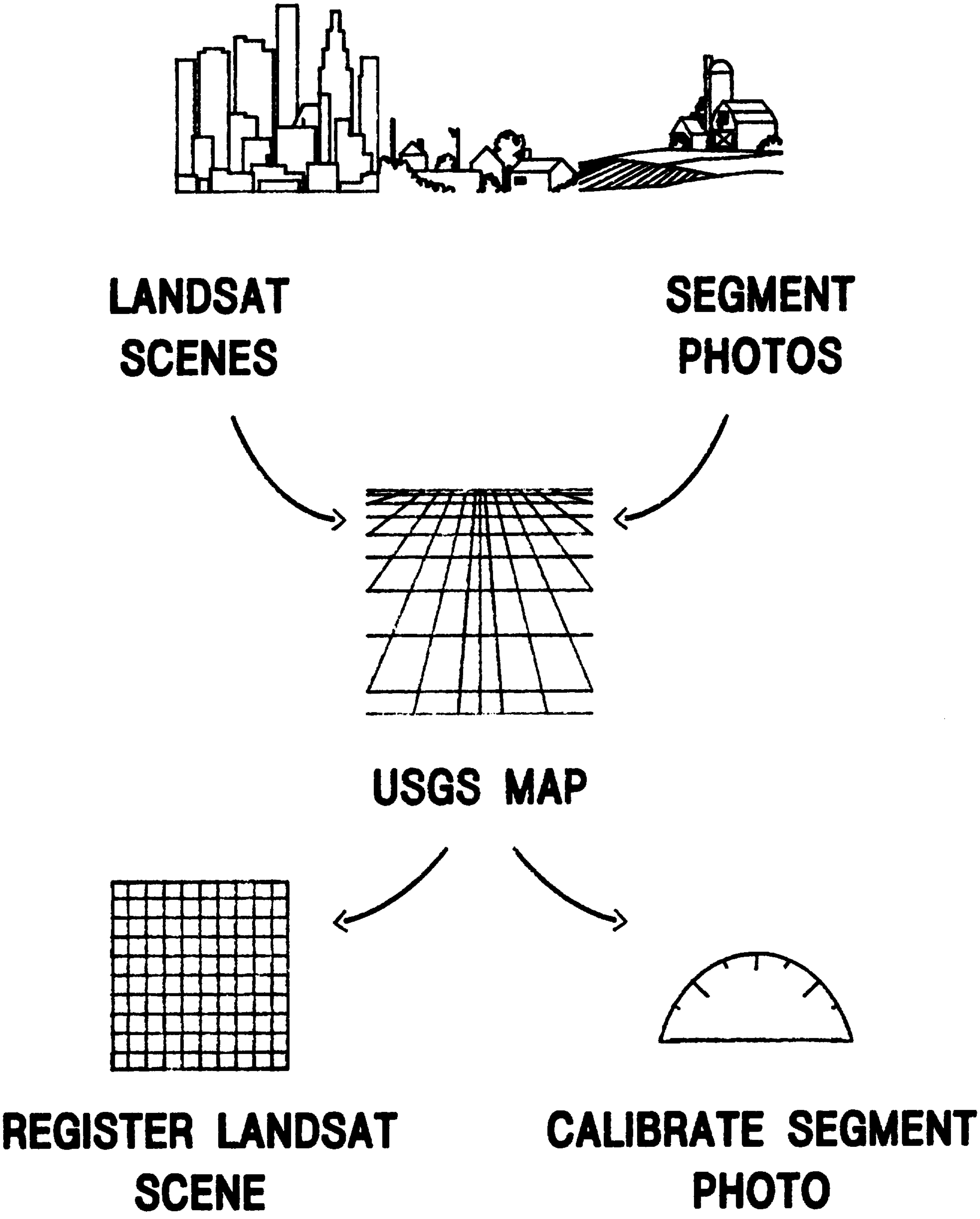
DATA BASE PROGRAM

ORDERS

COSTS

RECORDS

Geographic Reference



LANDSAT SCENES

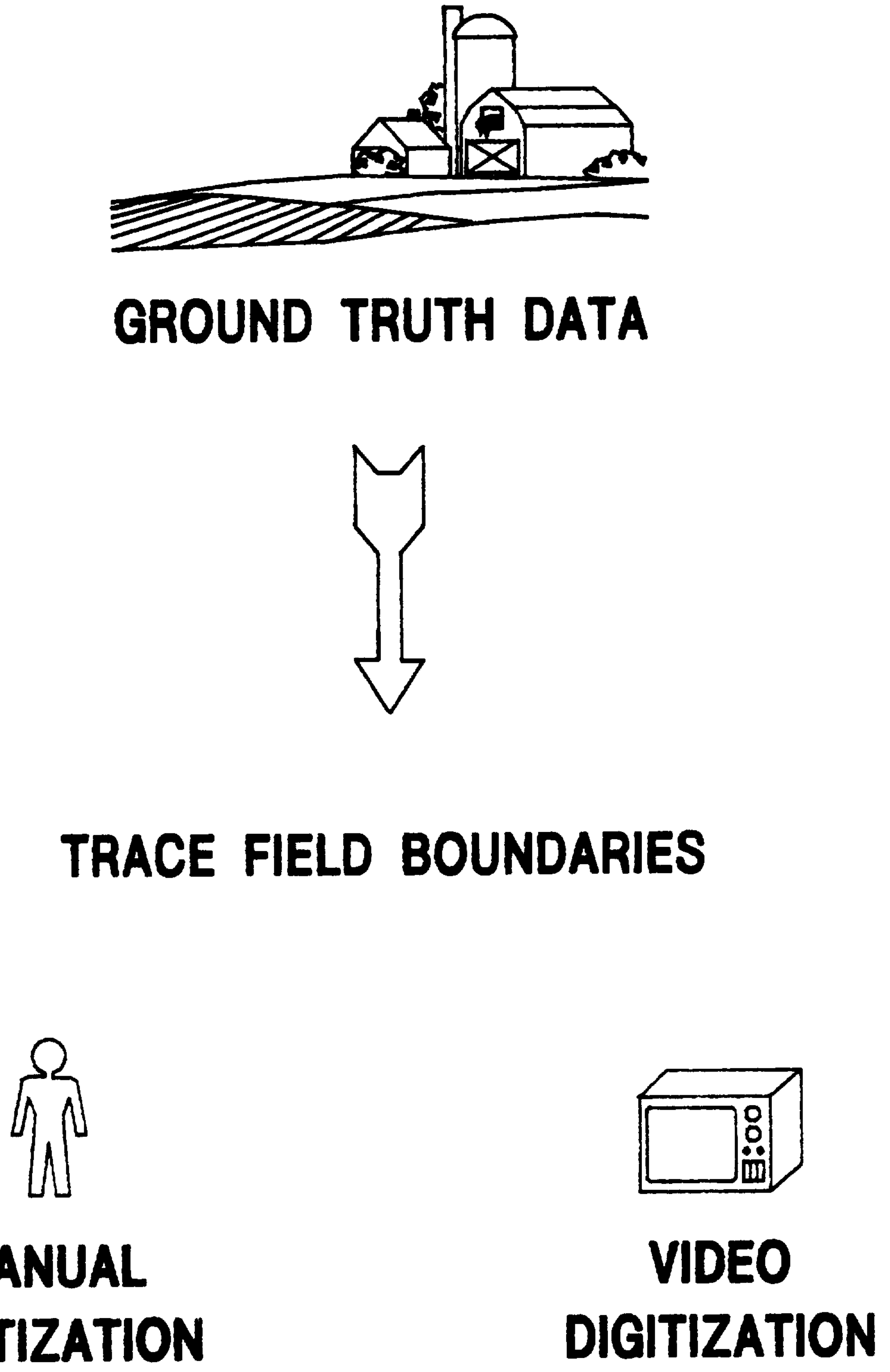
SEGMENT PHOTOS

USGS MAP

REGISTER LANDSAT SCENE

CALIBRATE SEGMENT PHOTO

Field Level Data



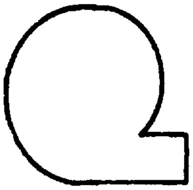
GROUND TRUTH DATA

TRACE FIELD BOUNDARIES

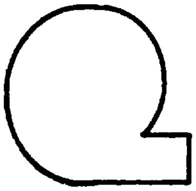
MANUAL DIGITIZATION

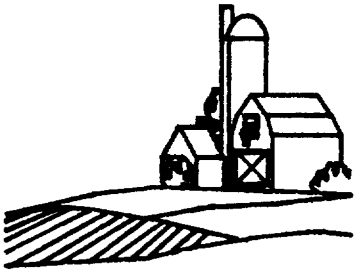
VIDEO DIGITIZATION

Landsat Data

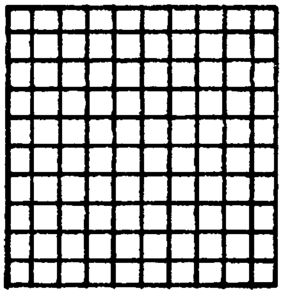


READ LANDSAT DATA
AND
GROUND TRUTH DATA



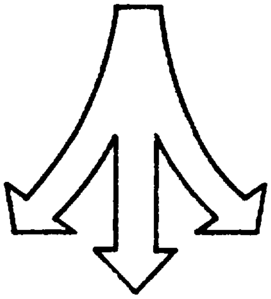


COMBINE



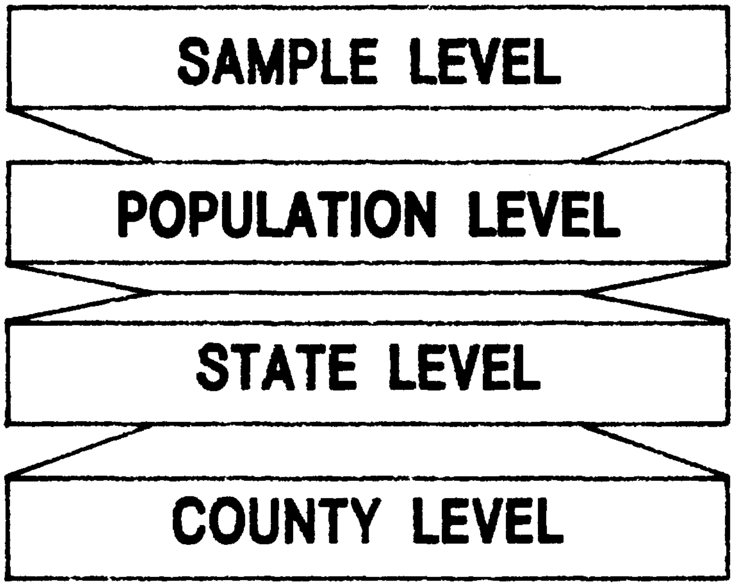
GROUND
TRUTH
DATA

LANDSAT
DATA



CLUSTER STATISTICS CLASSIFICATION

Crop Estimation



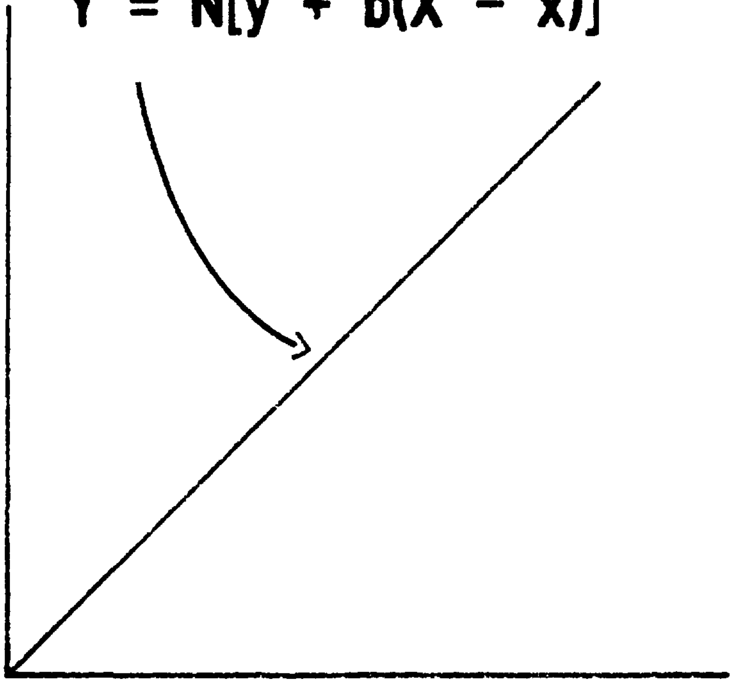
SAMPLE LEVEL

POPULATION LEVEL

STATE LEVEL

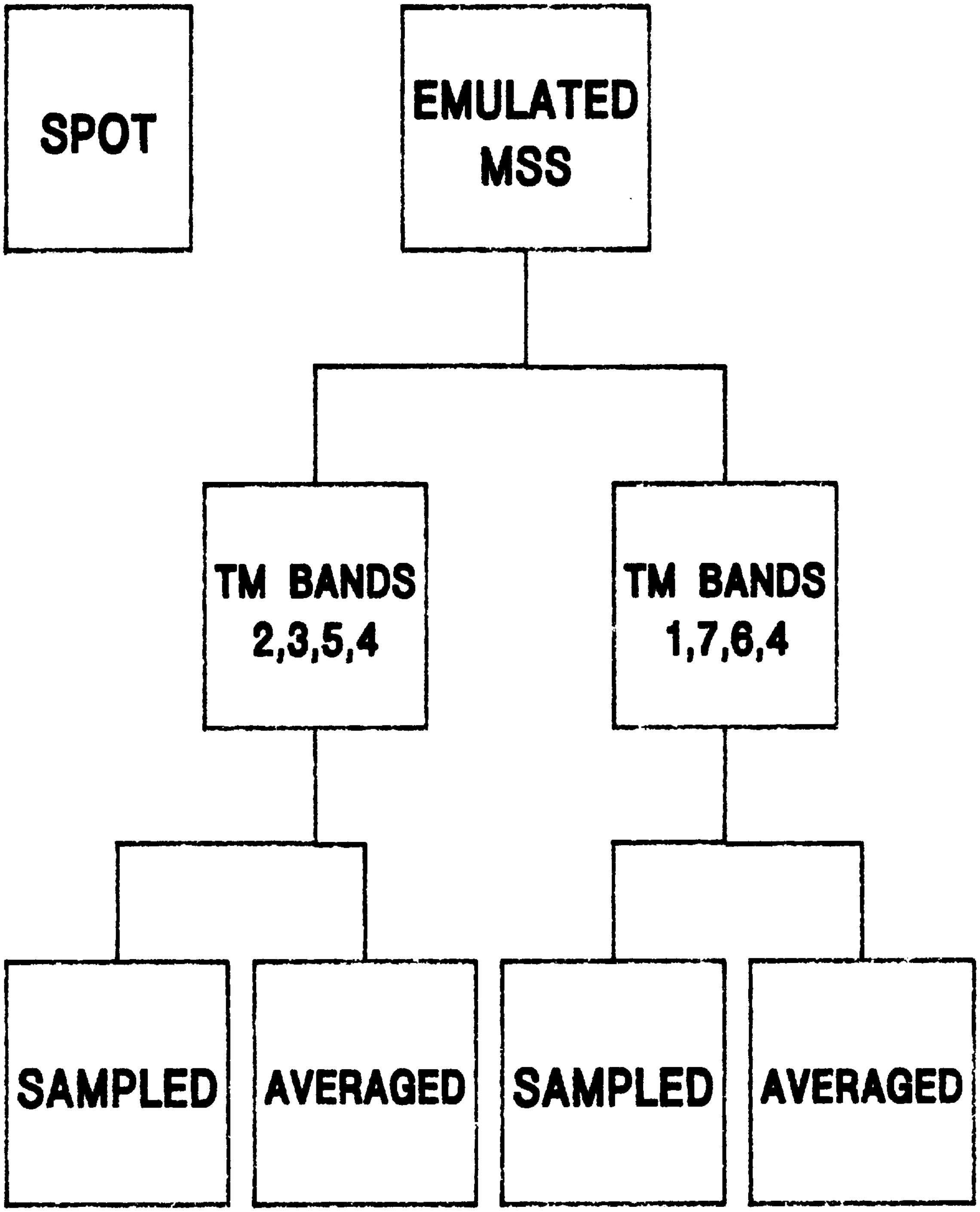
COUNTY LEVEL

$$\hat{Y} = N[\bar{y} + b(\bar{X} - \bar{x})]$$



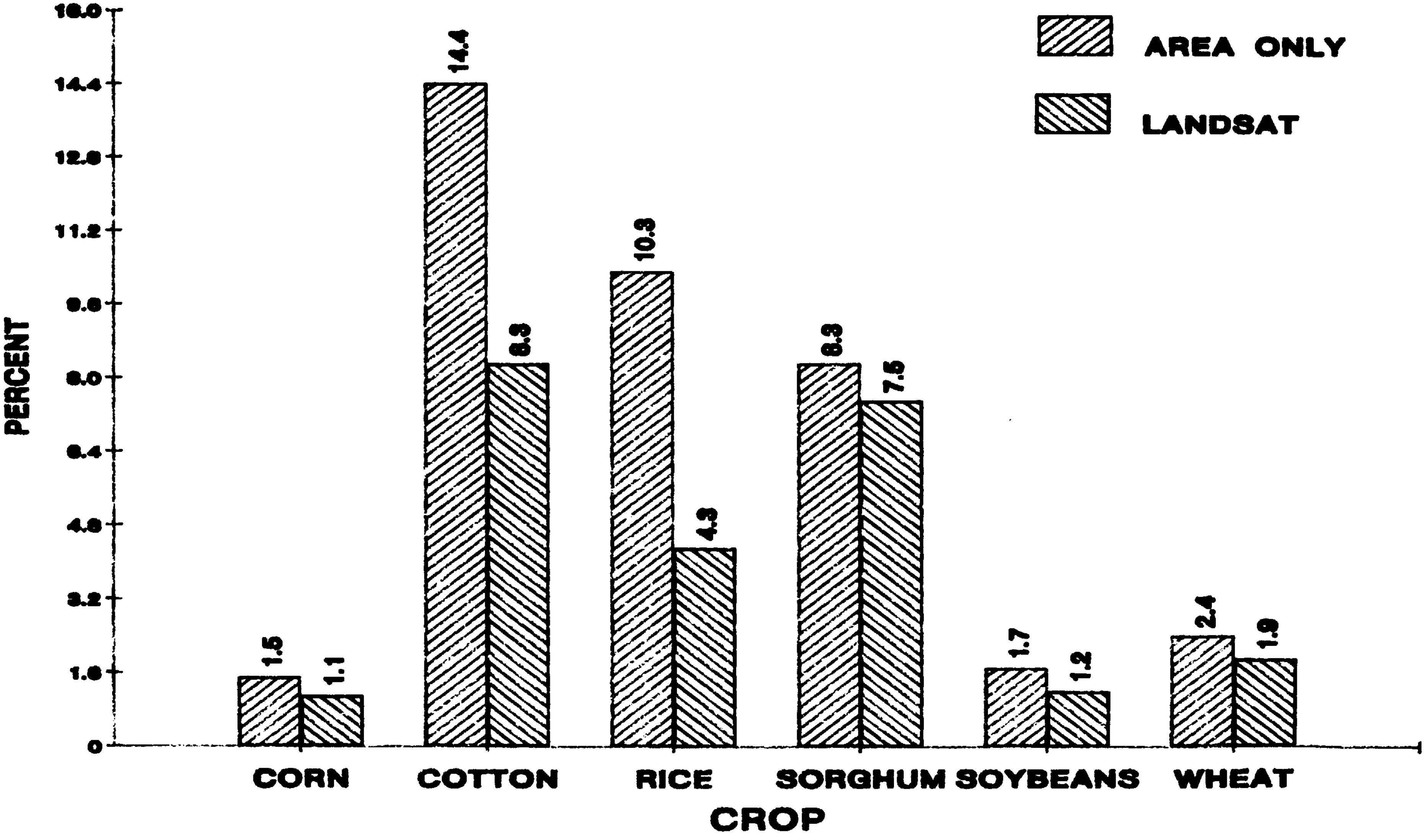
New Data Sources

8



Results

COMPARISON OF COEFFICIENTS OF VARIATION 1986


PERCENT
AREA ONLY
LANDSAT
CORN
COTTON
RICE
SORGHUM
SOYBEANS
WHEAT
CROP