Crop Reporting by Satellite

G. A. HANUSCHAK, G. F. HART, R. R. STARBUCK, R. S. SIGMAN

G. A. Hanuschak, Mathematical Statistician, USDA, SRS, RD, Washington, D. C.
G. F. Hart, Chief, R&DB, USDA, SRS, RD, Washington, D. C.
R. R. Starbuck, Mathematical Statistician, USDA, SRS, RD, Washington, D. C.
R. S. Sigman, Mathematical Statistician, USDA, SRS, RD, Washington, D. C.

AUGUST 1977

Abstract

The Statistical Reporting Service has been involved in LANDSAT data application research since the launch of LANDSAT I. Investigation has proceeded in two objective areas: (1) to improve the crop acreage estimating ability for small areas (counties and groups of counties), and (2) to develop broad land use classification for stratifying land area sampling frames.

The former area is exemplified by research results for the state of Illinois using 1975 LANDSAT digital data and ground collected data from a probability sample of 300 units selected from a stratified agricultural land use sampling frame. LANDSAT digital data are used as a secondary or ancillary data source with ground collected sample data as the primary data source. A statistical application of regression is used to combine ancillary LANDSAT digital data with ground collected sample data. Problems such as signature extension, classification bias and cloud cover bias are avoided by statistical inference. Entire scene classification, once considered to be a constraining limitation on the use of LANDSAT digital data is not a limiting constraint in this application. Digital analysis is performed in a network processing environment. All analysis software is documented and available on the ARPA network. Results indicate that greater estimating precision can be obtained by using ancillary LANDSAT data. Results also indicate limitations in the use of LANDSAT data including the variable amount of information contained in a particular LANDSAT pass and the requirement for a sufficient amount of ground data to optimize the methodology.

Relative to the second area of investigation, the lower San Joaquin Valley of California is used to exemplify the potential of LANDSAT data

for land use classification to construct agricultural land area sampling frames. Although this use of LANDSAT data is less dramatic than for crop acreage estimation, the immediate potential may be of greater importance. In this application, the timeliness and update potential of LANDSAT data are of greater importance than the information content of the data.

Introductory Note

At the outset it should be stated that much of the material presented in this paper concerning crop acreage estimates for small areas was presented in a paper given at the 1977 Machine Processing of Remotely Sensed Data Symposium, LARS, Purdue (4). The other topic, developing broad land use classification for stratifying land area sampling frames, has not previously been reported.

Introduction

The Statistical Reporting Service (SRS) of the U.S. Department of Agriculture uses as its primary data collection mechanism a stratified probability selected sample of about 16,000 map and aerial photograph defined sampling units from a frame containing all land area in the continental United States (3). A typical sample unit in major crop producing areas of the midwest is about one square mile. A staff of trained interviewers employed by SRS use aerial photographs, scaled at approximately 8 inches per square mile, to locate and account for every parcel of land in the selected sample units. A major survey is conducted in the latter part of May each year with followup surveys for more specialized or update information at several other times during the year. In addition to a complete accounting of land use, these surveys obtain information on livestock and poultry numbers and a variety of economic and cultural practice items from persons operating land in the selected units. Since a complete accounting of agricultural information is obtained for a probability sample of units, it is

possible to expand information from the sample units to state, regional, and national totals. The Agency's interest in LANDSAT data is in connection with support for this basic survey mechanism.

The complete or census like coverage of LANDSAT on a potentially near real time basis attracted our research interest. These attributes could be of value in two basic areas: (1) in the acreage estimation process itself, and (2) in stratified land area sampling frame construction and update.

The land area sample provides adequate current information on agricultural production items for major producing state and regional and national levels. However, land area sampling methodology is not cost effective in providing small area statistics, such as for counties or aggregations of counties or for minor production items at the state level.

The theory of sampling identifies a relatively powerful technique for improving sampling efficiency over straightforward simple random sampling. The theory addressed here is stratification, and in general, it specifies that if a population can be subdivided into unique groups of units that are relatively homogeneous and if simple random sampling is applied within each of these groups, one may expect improved sampling efficiency over using simple random sampling over the entire population. A simple example in agriculture would be to separate urbanized areas from the open country. Even though some agricultural activities are conducted in urban zones, the majority occur outside. Therefore, one

does not need a very large sample to cover the agricultural impact or portion of the total coming from urban zones. Our land area sampling frame methodology does in fact make this distinction as well as others through photographic interpretative processes.

The present method of updating or renewing the land area sampling frames requires interpretation of the aerial photography to identify and delineate different classes of land use--stratification. Photographic interpretation is time consuming and expensive and photograph coverage is in many instances out-of-date. Stratified land area sampling frames become outdated and therefore inefficient, because of urbanization, irrigation, change in physical land features, such as roads and drainage ways, and shifts over a period of time in agricultural production patterns. Recently, we have found, in developing new stratified land area sampling frames to replace those that have become inefficient, that the aerial photography available for new frame construction is, in many instances, over five years old. Therefore, inefficiencies in land use stratification are being built into new frame construction at the very outset.

Improving Crop Acreage Estimating Ability for Small Land Unit Areas

Background

The entire state of Illinois was the test area for this research activity. The objective was to estimate the acreage of major spring planted crops at the county level.

It should be noted that Illinois has a relatively new sampling frame. In 1975, an old sampling frame with 350 selected sample units was replaced by a new land use sampling frame with 300 selected sample units (Table 1). Even though the sample size from the new frame was smaller, the relative sampling efficiency improved. In other words, the precision of estimates from the new sampling frame with only 300 selected sample units was greater than that of the old frame with 350 selected sample units. This indicates the efficiencies to be gained by new stratified frame construction after a fifteen year period of frame use.

The state of Illinois was chosen for investigation in part because the state is a major corn and soybean as well as generalized farming state. Field sizes are somewhat typical of the central and eastern corn belt posing a realistic test of the resolution of LANDSAT data. One would expect optimum results from LANDSAT data investigation to occur in states where there are large field sizes and relatively few major crop types. The western corn belt or great plains wheat states would be more ideally suited for LANDSAT investigation, however, would not provide such a rigorous test of the ability of LANDSAT to accommodate a major segment of U.S. agricultural production; spring planted row crops.

The research question is then, can sample data with nearly perfect information content be combined with spectral data from LANDSAT with relatively low information content to improve crop acreage estimates? LANDSAT data, with relatively low information content, has the desirable attribute of not having sampling error; data are obtained for every acre

Table 1. Illinois Stratum Numbers and Definitions

STRATUM

No.	Description
11	75%+ cultivated
12	50% - 75% cultivated
20	15% - 49% cultivated
31	Urban
32	
33	
40	Range Land
61	Proposed Water
62	Water

of land use with the exception of when cloud cover and atmospheric disturbance cause either nonresponse or low information response.

Statistical Methodology

The procedure of combining ground enumerated sample and LANDSAT data involves a statistical application of correlation and regression. If pixel data from a LANDSAT multispectral scanner are sufficiently correlated with acreage data from ground enumeration, then a regression estimator, taking advantage of the correlation, can improve efficiency over what could be obtained from ground enumerated sample data alone.

Data collected from the sample units are summarized within each land use stratum (4). Let h = 1, 2,..., L be the L 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 corn acres in j^{th} sample unit in the h^{th} stratum

Then

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

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 $\hat{Y}$ is commonly called a direct expansion estimate, and is redefined as $\hat{Y}_{DE}$.

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

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

r.s.e. = relative sampling error = $100 \cdot \sqrt{v(\hat{Y})} / \hat{Y}_{DE} = 2.4\%$

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

r.s.e. = relative sampling error = $100 \cdot \sqrt{v(\hat{Y})} / \hat{Y}_{DE} = 2.9\%$

Keeping in mind that LANDSAT offers complete coverage, including coverage for each acre of each enumerated land area sample unit, the task is to determine the relationship between crop acres and reflectance values for sample units and employ this relationship through a regression model using all LANDSAT data. The estimates and variances thus generated can be compared with direct expansion estimates and variances from the enumerated land area sample units. If we sampled LANDSAT data, as has been done in some investigations, then our chance of improving estimating ability would be reduced because then there would be sampling error associated with LANDSAT data.

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

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

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 h^{th} land use stratum

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

$\hat{b}_h$ = the estimated regression coefficient for the h^{th} land use stratum when regressing ground enumerated acres on classified pixels for the n_h sample units

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

$\bar{X}_h$ = the average number of pixels of corn per sampling unit for <u>all</u> sampling units in the h^{th} 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^{th} area sampling unit of the h^{th} land use stratum

$\bar{x}_h$ = the average number of pixels of corn per sample unit in the h^{th} land use stratum

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

x_{hj} = number of pixels classified as corn in the j^{th} sample unit in the h^{th} land use stratum

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

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

where

r_h^2 = sample coefficient of determination between reported corn acres and classified corn pixels in the h^{th} land use stratum

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

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 $\lim v(Y_R) = 0$ as $r_h^2 \to 1$ for fixed n_h. Thus a lower variance is attained if the coefficient of determination is large for most strata.

r.s.e. = relative sampling error = $100 \cdot \sqrt{v(\hat{Y}_R)} \,/\, \hat{Y}_R$

Results

Now for some results for the state of Illinois in 1975. We have worked out procedures whereby it is not necessary to have cloud free imagery in order to utilize the regression estimator procedure (1). However, in this, our first large scale effort, we were able to obtain cloud free imagery by utilizing several LANDSAT passes and scenes to cover the state of Illinois. There is one exception to complete coverage but this had nothing to do with cloud cover. There are two counties in the center of the state that were not contained in a LANDSAT scene. That is, even with overlap of LANDSAT coverage on different passes, these counties were not included in a single LANDSAT pass. This does not make analysis impossible but it does require special techniques and it was decided not to attempt classification for these two counties--all other counties in the state were classified.

Due to the different LANDSAT scenes and passes, the state was divided into analysis areas. Seven such areas were defined for our study (Fig. 1).

Our criteria for evaluation success was reduction of the relative sampling error (r.s.e.). Both estimation procedures, direct expansion of data from the enumerated sample units and regression using both enumerated sample unit and LANDSAT data provide estimates of r.s.e. that can be compared directly. A third data set was also used in comparison--the Illinois State Farm Census. This is a post growing season accounting of specified crop and livestock items obtained as an adjunct to a state tax accounting. The census is not a controlled accounting and adjustments are made for consistency. But these data provide independent comparisons.

The use of LANDSAT data did result in significant reductions in the relative sampling errors from the use of enumerated sample unit data alone for the analysis areas but the reductions are not overwhelming (Table 2). The estimates themselves are within sampling error. County estimates were also made but the relative sampling errors were unacceptably high by the standards the Agency normally places on estimates (Table 3). We are, however, reviewing the regression estimator and the associated variance formulation. It appears that we may have overstated the variance and if we can be statisfied that a modified procedure is theoretically sound, resummarization will be performed. We have not completed our analysis and if an adjustment is in order, it will be published.

Conclusion

We have shown that a statistically sound procedure can be used to incorporate LANDSAT data with probability ground collected data to improve estimating ability. Further research into the use of multitemporal data, soil background information, and other analysis techniques is continuing.

We have developed a complete software system for full frame LANDSAT analysis that is available via network processing (2). The software is available on the ARPA network. Full frame analysis is conducted on the ILLIAC IV computer. We can provide documentation to anyone interested in investigating our procedures for possible application in other areas.

We have not shown our methodology to be cost effective for small area estimation. However, during the past two years of research investigation, we have experienced a continuing reduction in analysis costs.

Table 2. Estimated Acres of Corn and Soybeans for Wholly Contained Counties in Each Analysis Area.

Analysis Area	No. of Counties Wholly Contained In the Analysis Area	Estimator	Corn		Soybeans	
			Acres	r.s.e.	Acres	r.s.e.
W123[a]	29	Direct Expansion	4,110,150	3.6%	1,539,200	7.7%
		Regression	4,125,400	2.5%	1,681,800	5.2%
		SSO	3,682,300		1,657,800	
C1A	7	Direct Expansion	1,191,400	7.1%	532,700	13.9%
		Regression	1,180,500	2.9%	523,200	8.2%
		SSO	1,196,900		502,900	
C12	20	Direct Expansion	2,907,700	4.5%	2,217,200	5.5%
		Regression	2,945,100	4.3%	2,127,200	5.1%
		SSO	2,939,700		1,990,400	
C33+	16	Direct Expansion	1,158,000	9.5%	1,675,100	8.6%
		Regression	1,077,000	8.6%	1,540,000	6.8%
		SSO	1,233,000		1,246,000	
E12	12	Direct Expansion	1,781,300	5.6%	1,439,500	6.3%
		Regression	1,577,300	4.1%	1,290,700	6.5%
		SSO	1,792,000		1,383,000	
E23+	32	Direct Expansion	1,669,500	7.5%	2,431,950	5.2%
		Regression	1,615,000	6.9%	2,357,850	3.8%
		SSO	1,767,000		2,045,000	
West CRD[b]	9	Direct Expansion	1,316,000	8.5%	562,000	13.1%
		Regression	1,269,000	4.6%	574,100	10.6%
		SSO	1,125,000		680,000	

[a] W1 and W2 (Fig. 1) were analyzed individually and joined with W3 (not shown on Fig. 1 but follows W2) to form W123

[b] Wholly contained within W2

Table 3. Regression Estimates for Corn and Soybeans in Individual Counties in Western Pass

County	Corn		Soybeans	
	Acres	r.s.e.	Acres	r.s.e.
Adams	166,600	24.0%	83,600	35.3%
Brown	53,700	33.4	24,300	50.7
Bureau	254,000	18.7	110,600	33.4
Calhoun	56,700	25.1	23,300	39.9
Carroll	126,500	17.5	57,200	29.6
Cass	91,700	20.3	54,100	25.5
Fulton	172,100	29.0	91,400	37.8
Greene	136,800	19.2	76,000	24.8
Hancock	190,500	19.3	74,800	36.2
Henderson	104,000	17.3	37,100	36.4
Henry	276,800	17.2	79,400	46.6
Jersey	85,700	21.6	48,900	27.0
Jodaviess	108,300	34.1	27,100	94.2
Knox	174,100	19.5	79,600	31.6
Mason	129,100	21.3	76,100	27.9
McDonough	162,500	17.4	82,500	26.3
Mercer	139,800	18.7	43,900	43.4
Morgan	147,200	17.6	93,700	20.9
Ogle	223,000	19.0	51,500	64.2
Peoria	124,000	24.0	65,300	32.6
Pike	160,100	25.7	78,300	37.3
Rock Island	107,000	18.7	27,500	52.7
Schuyler	84,000	29.0	36,650	46.2
Scott	61,100	19.9	31,500	28.6
Stark	92,000	18.2	40,600	32.1
Stephenson	172,100	18.6	30,600	81.8
Warren	161,800	16.5	64,100	32.2
Whiteside	242,800	16.2	62,400	49.0
Winnebago	121,500	21.5	29,600	68.0

Land Use Classification For Stratification

Background

A second area of investigation involves the use of LANDSAT data for broad land use classification to stratify for land area sampling frame construction.

The primary supplier of aerial photography for land area sampling frame construction has been the Agricultural Stabilization and Conservation Service (ASCS) of the U.S. Department of Agriculture. Some coverage is also obtained from the Soil Conservation Service, Forest Service, and U.S. Geological Service, Department of Interior. The primary supplier, ASCS, due to program constraints in recent years, has not been able to support the level of repeat coverage obtained in previous years. This is particularly true in the less important field crop production areas of the west, east, and south. Although relatively less important to the total agricultural production of the United States, these areas do provide a significant amount of the agricultural production and require up-to-date land area sampling frames. The current complete coverage attribute of LANDSAT lead us to investigate the potential of this data to replace or supplement conventional aerial photograph coverage. Our site for this investigation is in the lower San Joaquin Valley of California. Procedures and techniques developed in this test area will be applied to a stratification program for an entire state.

Different alternatives are being tried but they have one operation in common, a process called digitization. This involves delineating boundaries, such as roads and streams that appear on a photograph or map, in a high

density numeric form. The numeric form is a stream of coordinates in a common map reference base such as latitude and longitude. A digitizer uses a grided tablet and computer software or preprogrammed hardware to generate coordinate values at the rate of several hundred per inch. The machine operator moves a cursor over the map surface following a boundary. This boundary is transformed into a digital coordinate format that goes to tape or disk. Digitizing permits transferring boundary data from one format to another--an example would be converting digitized boundaries from a map base to LANDSAT computer compatible tape line and column.

Land use can be determined in a number of ways. One is by photographic interpretation. Personnel are trained to recognize certain photograph features distinguishing them from others. Our current procedure for land area stratification involves distinguishing different cultivation intensities, urban areas, grazing and forested land and enclosing like or homogeneous areas with definable boundaries (roads, streams, etc.). All like areas, no matter where they are located in a state, form a stratum. Stratum definitions for most states are similar to those for the state of Illinois (Table 1).

After stratification, the next step is to create units that are subdivisions of strata--these units are called count units. Count units are land definable units that will be further subdivided into from about two to twenty sampling units. An attempt is made to make count units as heterogeneous as possible--the same being true for the further subdivision of count units into sampling units.

The Research Approach

One procedure utilizing LANDSAT data is to construct count units first. Hereafter in this paper, count units will be referred to as frame

units. This can be done without knowledge of land use using maps where roads, streams, and other boundaries are readily identifiable. Once this task is completed and the boundaries digitized, then the digitized boundaries are transformed to line and column in LANDSAT computer compatible tapes. Next, supervised or unsupervised classification is performed to identify the agricultural characteristics of frame units. Finally, frame units are ranked by agricultural characteristics grouping similar frame units into strata. Frame units could then be rearranged or sorted into different strata for particular survey purposes. That is, the stratification or grouping of frame units for a crops survey would be different than for a livestock survey. Carrying this one step further, for a crops survey the arrangement might be different for cotton than for corn. For whatever the purpose of the survey then, sample units would be selected from or identified with the appropriately stratified grouping of frame units. Note, the use of the frame unit has changed--it is now a stratum building block rather than a subdivision of a larger stratum unit. Frame unit data could be updated with new LANDSAT data as significant changes in land use occur. Research results to date, however, indicate the current level of information in LANDSAT data is not acceptable for this approach.

Another approach is to use LANDSAT imagery in much the same way as we use aerial photography. That is, photographically interpret unclassified or classified LANDSAT image data to determine the intensity of cultivation, delineate areas of like intensity, aggregate like areas into strata and subdivide these areas into frame units. Then after the sample selection process, further subdivide frame units for sample unit selection. The final

step is to digitize all stratum and frame unit boundaries so that at a future point in time LANDSAT data could be used to update the frame to reflect changes in land use--such as urbanization or irrigation.

Potential Use

The potential for effective use of LANDSAT data in land area sampling frame construction may be more important in developing countries than in the United States. In many developing countries, aerial photography, even old aerial photography, may not be available and the use of LANDSAT data could be a reasonable alternative in constructing land area sampling frames. SRS is actively involved in a program of land area sampling frame construction in many developing countries and our research in this country, if successful, may be applied in a cooperating developing country.

Utilizing LANDSAT data to update land area sampling frames could avoid some fairly major problems that have occurred in the past. Some areas of the western great plains have come under pivotal irrigation quite rapidly. In many cases, the frame material utilized to construct the original frame was several years old and after a few years we found nearly complete cropping in areas where there was grazing land a few years earlier. This makes the frame inefficient and sampling errors increase substantially as a result. Urbanization has also caused difficulty in many states. Suburbs extending into agricultural areas create enumeration problems, increase survey costs and to some extent increase sampling error.

Conclusion

As mentioned at the outset, SRS LANDSAT research is oriented toward improving the basic data collection mechanism of the Agency--the land area

sampling frame. Neither of the two areas of research covered are operational. However, the utilization of LANDSAT data for land area sampling frame construction may be used in an operational test within the next year. The potential benefit of LANDSAT for this use does not appear to be as great as it is for improving small area estimation. However, we do not know that there is a significant economic benefit to having precise estimates for small areas. The greatest need is for accurate national level estimates and secondly, for adequate state level estimates. Presently, we do not know how much users would value accurate acreage information for areas the size of a county or production aggregation areas of similar or larger size.

We do know that satellite technology is improving. LANDSAT C and LANDSAT D offer improvements in sensing that should result in improved information content. What we can say right now is that we are conducting very interesting research.

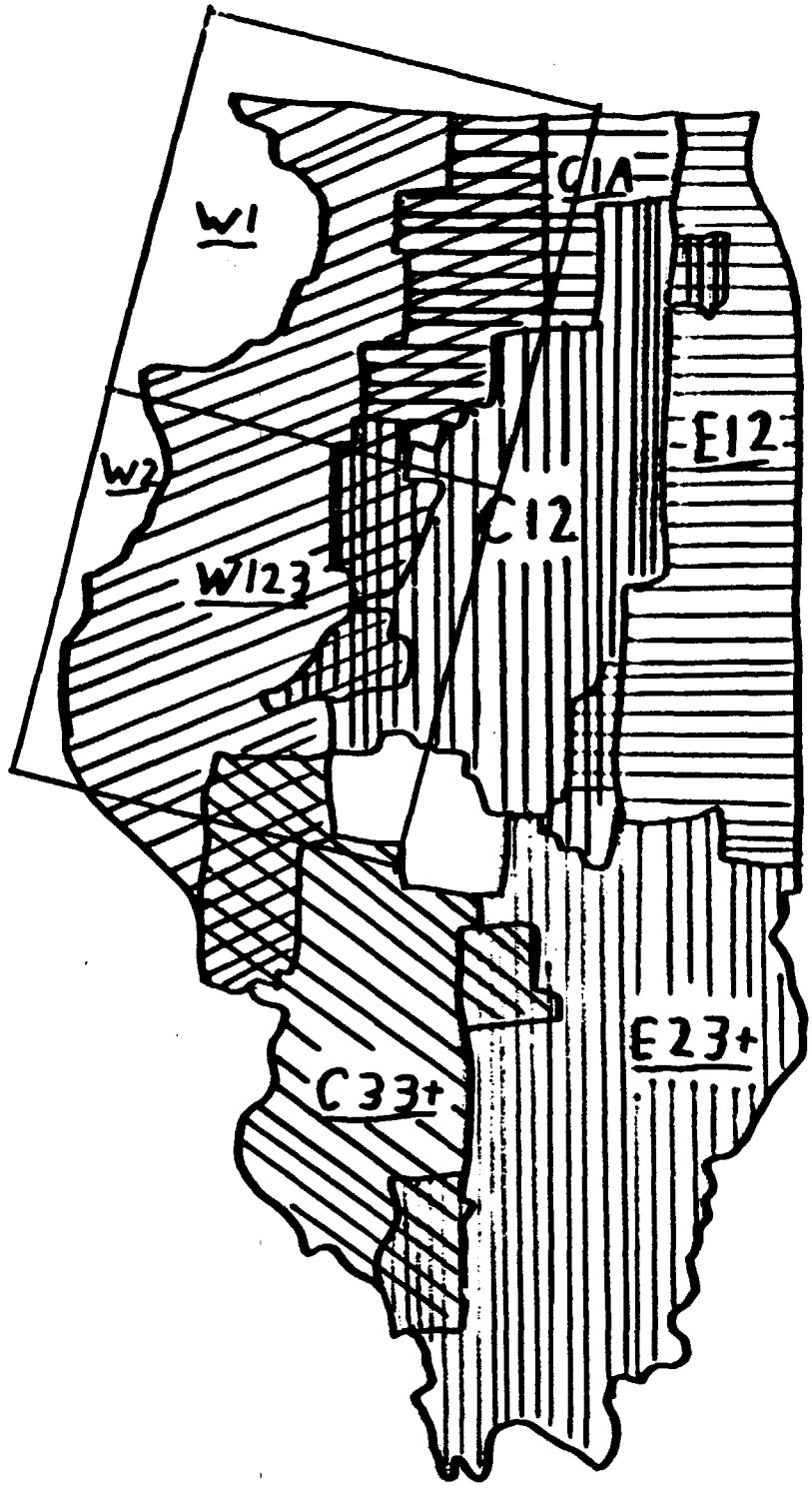

Figure 1. Analysis Areas for 1975 Illinois Acreage Estimation Project.

1. Hanuschak, G. . 1976. "LANDSAT Estimation With Cloud Cover" . Purdue: LARS Symposium Proceedings, Machine Processing of Remotely Sensed Data. IEEE Catalog No. 76CH1103-1 MPRSD. Purdue University, West Lafayette, Indiana.

2. Ozga, M., Donovan, W. E., and Gleason, C. P. . 1977. "An Interactive System for Agricultural Acreage Estimates Using LANDSAT Data" . Purdue: LARS Symposium Proceedings, Machine Processing of Remotely Sensed Data. Purdue University, West Lafayette, Indiana.

3. "Scope and Methods of the Statistical Reporting Service" . 1975. Miscellaneous Publications No. 1308. USDA, SRS, Washington, D.C. .

4. Sigman, R., Gleason, C. P., Hanuschak, G. A., and Starbuck, R. S. . 1977. "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment" . Purdue: LARS Symposium Proceedings, Machine Processing of Remotely Sensed Data. Purdue University, West Lafayette, Indiana.