Integrated Use of LANDSAT with Ground Data

Galen F. Hart, William H. Wigton
Michael E. Craig, George A. Hanuschak and Richard S. Sigman

AUG. 1978

Introduction

The Statistics Unit of the Economics, Statistics, and Cooperatives Service (ESCS), United States Department of Agriculture (USDA), is officially responsible for collecting and disseminating current crop and livestock statistics. To support this responsibility the Statistical Research Division (SRD), ESCS, continually seeks ways for improving the accuracy, timeliness, coverage and cost effectiveness of operating programs. Since the launch of LANDSAT I in July, 1972, SRD has conducted research investigations toward utilizing spectral reflectance data to improve crop area estimating ability. The interest in LANDSAT data stemmed from the potential for complete or census like coverage for large areas in a very short time span. The general objective for investigations is to develop methods for integrating the best features of an existing ground data collection system and LANDSAT digital data.

Ground Data Collection System

Since the general objective is to integrate or merge an existing system with a new system this paper first addresses the existing ground data collection system. The core of this system is the land area sampling frame and the underlying concepts of the land area sampling frame are basic to understanding how integration is possible.[1]

Galen F. Hart* is Chief, Research and Development Branch; William H. Wigton is Head and Michael E. Craig, George A. Hanuschak and Richard S. Sigman are Mathematical Statisticians, New Techniques Section, Research and Development Branch, Statistical Research Division, Economics Statistics and Cooperatives Service, United States Department of Agriculture, Washington, D.C. 20250

* PRESENTED AT XXIX CONGRESS INTERNATIONAL ASTRONAUTICAL FED., YUGOSLAVIA

The entire area of the United States is partitioned or stratified by agricultural land use. For a particular state (department, province), the number of partitions or strata may vary but for a typical state about ten uniquely separable strata of land use are delineated. The task of dividing the land area of a state into strata is accomplished by interpreting conventional low level black and white aerial photography.

Urban or non-agricultural lands are separated. By using a percent cultivation criteria degrees of cropping intensity are interpreted and delineated. Wood and grazing lands are separated. When the task of stratification is completed all land in a state has been uniquely assigned to a particular stratum. Within each stratum the total land area is subdivided into sampling units. A typical sampling unit in major crop producing areas is about 2.6 square kilometers. The collection of all sampling units for all strata is called the area sampling frame. A probability sample of units is selected from each stratum and each selected unit is delineated on a small scale aerial photograph (approximately 1:8000 or 12 centimeters to the kilometer).

A national survey of about 16,000 sampling units is conducted in late May of each year. This survey is known as the June Enumerative Survey (JES). About 1,600 part-time interviewers employed by ESCS obtain complete agricultural information for each of the selected sampling units during a two week interview period. Intense training of field supervisors and interviewers is conducted to minimize potential error. Each parcel of different land use is delineated on 1:8000 scale photography and land use and hectares are recorded on a questionnaire. The interviewer also obtains and records on the questionnaire information on crop utilization, grain storage, livestock inventory by various weight classes, and agricultural labor and economic items. The data collected during this Survey serves a wide range of current crop and livestock statistics programs. This same set of sampling units or sub-samples thereof are visited several other times during the year to obtain such information as yield and production of crops and to update information obtained at the time of the JES.

For major crops at the state level this survey provides estimates with relative sampling errors ranging from 2 to 8 percent. At the national level the relative sampling errors for major items are about I.5 to 3.5 percent. This system provides statistics for items relevant to current agricultural information needs at state and national levels that are timely (publication within three to four weeks after data collection), accurate and acquired at a reasonable cost. The total cost of annual surveys, including maintenance of the sampling frame, is about 3.5 million dollars.

Potential for Improvement Offered by Integrating LANDSAT Data

Probability sampling offers an apparent cost-effective means of collecting data at state and national levels. But it is an attribute of sampling theory that sample size is nearly independent of population size. For a typical state with about 350 sampling units, estimates for major crops have relative sampling errors of about 2 to 8 percent. The sample size required to achieve the same precision at the national level would be less than 1000 sampling units. However, in order to provide a target 2 to 8 percent sampling error at the county (parish, municipio) level, several hundred sampling units would be required and therefore, sampling is not considered to be a cost-effective method for providing small area statistics. LANDSAT, however, being a complete or census coverage method for collected data offers potential for small area land use statistics since energy readings in four spectral bands are acquired for each acre of land.

There are some characteristics of LANDSAT, when viewed as a data collection device, that must be recognized and accommodated to achieve a successful integration. Resolution could be described as "course" — 4800 sq. meters is not a highly descriminating unit of observation for detailed land cover information. The quality of data is quite variable ranging from no information content, caused by cloud cover; to low information content, due primarily to atmospheric variation such as haze; to high information content, from clear atmosphere at an optimal time for land use or crop discrimination.[2, 3] Extracting land cover information from reflectance data is therefore a difficult task. A way of describing this statistically is that we have information concerning a population o interest, land cover, imbedded in data from a population that we know to be different. There is no direct way to "scale" the covered population to the population of interest.

It has been demonstrated through many experiments that information extraction from LANDSAT digital data is very directly associated with the amount of ground data available to convert spectral data to land cover information.[2, 3] Ground data are needed to obtain the "signatures" of spectral data. Selected sampling units from the land area sampling frame provide a substantial amount of ground data to "train" a computer to classify land cover from spectral data. Also since the area sample is a probability sample the data collected can be "expanded" independently to area totals (local, state, and national levels) and transfer inference power to the process of combining this data with spectral data. The statistical procedure of combining these two data sources is known as "double sampling" and estimation is performed by the "regression estimator". LANDSAT data are appropriate to use as auxiliary data to apply the theory. If the correlation between ground and spectral data for a particular land cover is sufficiently high then, since LANDSAT data is without sampling error, estimates resulting from the combination would have a lower overall net sampling error. In fact, if there was a perfect one-to-one relationship between ground and spectral data, estimates resulting from the combination would be without sampling error.

Procedure for Integration

The task of developing a method for combining spectral and ground data has been completed and is in a computer network (ARPA) environment with software developed cooperatively between ESCS and the Center for Advance Computation, University of Illinois.[4] The network permits communication between researchers and an efficient computer processor known as ILLIAC IV at the NASA, Ames Laboratory, Moffett Field, California. Some of the features of this software, known as EDITOR, include: interactive digitizing, storage, and retrieval of ground data; extracting LANDSAT reflectance data for sampling units where ground data is collected; computing statistics necessary for establishing a relationship between spectral and land cover data; classifying each pixel into a land cover type; and generating combined estimates and sampling variances. Methods have also been developed to handle the situation when cloud cover or lack of LANDSAT coverage allows for only the use of ground data.[5] EDITOR software is not proprietary and is documented and available to anyone interested in knowing details of the system.

The first step of integration, called registration, is to use mathematical equations to achieve a best fit of LANDSAT scene data to a map base. The preferred registration is to U.S. Geological Survey maps of either 1:24,000 or 1:62,500 scale. For a LANDSAT scene about 50 points are selected throughout the scene that can be located in the map base and LANDSAT data.

Next, each enumerated sampling unit is calibrated or locally registered to about 1/2 pixel accuracy in the registered LANDSAT scene. This accuracy is required so that individual pixels can be identified and labeled with the actual crop cover for classifier training. The process of sampling unit calibration is accomplished by using computer generated "gray scale" printouts of LANDSAT bands and field boundary plots as recorded from interviewer outlined field boundaries recorded on 1:8000 scale aerial photography. Agreement is achieved by making the appropriate row and column shifts in LANDSAT data necessary to align field boundaries and pixel printed output. An alternative to using direct reflectance band gray scale printouts is to conduct a preliminary clustering or classification using all four LANDSAT bands to better distinguish field patterns. This is more time consuming but does result in improved field boundary identification.

After registration, each pixel in each sampling unit within a scene or scenes in the same LANDSAT pass is labeled with the actual crop cover. Using these labels pixels from the major cover types are put into separate files. Next, various pixel clustering alternatives are attempted on the separate files to arrive at the "best" or optimal set of descriptive statistics for the major land cover categories. Discriminant functions in a four dimensional measurement space are used to determine this optimal clustering. Number of categories of land cover vary from about 5 to 15; the number depending upon the primary land cover types within a scene, quality of LANDSAT data, amount of sampling unit data available within a scene or pass, etc.

A statistics file containing a mean vector and covariance matrix for each category of land cover is then created. Usually, data is assumed tc be multivariate normal because of simplicity of calculations and so long as this assumption is not seriously violated, the errors induced are not substantial. The entire measurement space, all pixels in a LANDSAT scene, not just the subset covered by sample units, is then classified using the descriptive statistics to assign every pixel value to a land cover class.

Classification can be accomplished by a variety of ways utilizing the statistics from discriminant analysis. Choice of the "best" way depends on the type of result desired. An optimum classifier for the purpose of estimation is not necessarily the optimum for maximizing classification accuracy. An example of tnis is--if twice as many pixels were identified as corn as actually are corn and this relationship is consistent across sample units, then the classification accuracy would be poor, but estimation accuracy using regression would be perfect, two-to-one. In other words, if classification bias is consistent it can be removed by regression.

It should be noted at this point that we have not been successful in combining data from different passes. Atmospheric and other changes destroy uniqueness of signatures. We have been successful in combining data from adjacent scenes in the same pass but north-south differences limit even this extension. The key is to have enough probability data within a single scene.

When training and classification procedures have been completed, the remaining task is to utilize a statistical application of correlation and regression to integrate ground collected data with LANDSAT data. If pixel data from LANDSAT are sufficiently correlated with sample unit data, then a regression estimator taking advantage of the correlation can improve efficiency over what could be obtained from ground enumerated sample unit data alone. Data collected from sample units are summarized within each land use stratum. Let h = 1,2,..., L be the land use strata. For a specific crop the estimate of total area and the estimated variance of the total are as follows:

Let Y = Total corn acres for a state.

$\hat{Y}$ = Estimated total area of a specific land cover for a state.

y_{hj} = Total area of a specific land cover in the j^{th} sample unit in the h^{th} stratum.

Then,

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

This estimator is commonly called a direct expansion estimate, and we will denote this by Y_{DE}.

The estimated variance of the total is:

$$v(Y_{DE}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h (n_h - 1)} \quad \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - \overline{y}_h)^2$$

In order to summarize an estimate and its variance a ratio is formed. This ratio, the relative sampling error, expressed in percent is:

$$\text{r.s.e.} = 100 \cdot \sqrt{v(Y_{DE})}/Y_{DE}$$

Note that we have not yet made use of an auxiliary variable—classified LANDSAT pixels.

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

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

where $$\overline{y}_{h(reg)} = \overline{y}_h + \hat{b}_h (X_h - \overline{x}_h)$$

and $\overline{y}_h$ = the average area of a specific land cover 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 reported area 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 specific land cover pixels per frame unit for all frame units in the h^{th} land use stratum. Thus, whole 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 specific land cover pixels classified in the i^{th} area frame unit for the h^{th} stratum.

$\bar{x}_h$ = the average number of specific land cover pixels 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 a specific land cover in the j^{th} sample unit in the h^{th} stratum.

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

$$V(\hat{Y}_R) = \sum_{h=1}^{L} \frac{N_h^2}{n_h} \; \frac{N_h - n_h}{N_h} \cdot \sum_{j=1}^{n_h} (y_{hj} - y_h)^2 \cdot \frac{1 - r_h^2}{n_h - 2}$$

where

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

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

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 $(\hat{Y}_R) = 0$ as $r_h^2 \to 1$ 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 sampling error is: $r.s.e. = 100 \cdot \sqrt{v(\hat{Y}_R)}/\hat{Y}_R$

Research Results

Two examples of summary results will be presented from 1975 Illinois and 1976 Kansas data. In both cases the objective was to classify every pixel in the state to estimate major crops area. For Illinois the two major crops were corn and soybeans and for Kansas winter wheat was the major crop. Results are not to be compared directly since both projects were large scale research investigations and alternative analytical methods were being tried. The basic procedures as outlined earlier in the paper were followed. Within study project comparisons can be made and general conclusions can be drawn from results.

In Illinois 300 selected sampling units, the entire area frame sample for that year, were used in analysis with one exception—there were two counties in the center of the state that were not covered by a cloud free LANDSAT scene and these two counties were not included in the analysis. In Kansas there were 435 selected sampling units available from the area frame. At the outset of the project it was decided to use 174 of the sampling units (40% of the total) for LANDSAT analysis to reduce the impact of LANDSAT research on the operating program survey. In these projects, LANDSAT data for a specific analysis area were taken from one date only. Multi-temporal studies have been conducted but results are not presented in this paper. In Kansas it was believed that the optimal time for spectral separation of winter wheat from other cover types was April or May and for corn and soybeans in Illinois, late July to early September. In Kansas it was possible to obtain estimates for only 87 of the 105 counties during the given time period due to cloud cover or lack of sufficient training data. Several of the counties not estimated were important wheat counties. In Kansas 6 satellite passes were required to acquire total area, single time coverage and in Illinois 5 passes were required. Cloud cover continues to be a major problem and even though it is possible to handle this problem without bias, significant improvements in estimation accuracy can not be achieved without a large amount of high quality imagery.[5]

Each state was divided into analysis areas or districts. These districts were the aggregation levels for utilizing ground data for training and classification. The evaluation criteria for success was reduction in the relative sampling error (r.s.e.). Analysis areas or districts are shown in Figures 1 and 2. Tables 1 and 3 show both estimates and relative sampling errors for the regression and direct expansion estimators. A third data set, identified as SSO estimates, was used for comparison of results. Both States have post growing season accountings that approach a census or complete coverage. However, since coverage is not complete adjustments are made for consistency. These data provide independent comparisons.

Significant reductions in the relative sampling errors were achieved by the integrated use of LANDSAT data with ground data through the regression estimator. The reduction for Kansas is more significant than in Illinois keeping in mind that only 40% of the total possible ground data were used. Tables 2 and 4 are example results at the county level. Relative sampling errors are unacceptably high by the standards ESCS normally places on estimates. However these projects provided estimates with a measure of precision which would not have been possible without the integrated use of LANDSAT with ground data.

Remarks

Since the technical feasibility of using LANDSAT digital data to generate improved estimates of major land covers at the county or multi-county levels seems assured the question then becomes "Can the integrated use of LANDSAT with ground data be cost justified in operating programs based on improved information value?" The question is relevant since no national level agricultural operating program utilizes LANDSAT digital data as an integral part of operations.

At present USDA is exploring to determine if it might be possible to generate a series of outputs, including statistics, image products, and special overlays, that would benefit a number of different national and state program needs.

Within USDA, several agencies have program responsibilities that require land use inputs. Many state governments are also involved in land use planning and are seeing a greater need to monitor changes in land use.

If a "core" processing system similar to that discussed in this paper could be modified so that many users could obtain their products as marginal outputs then the cost of core processing could be distributed over a number of "benefited" programs. This approach assumes the "core" processing is the major cost of a particular product output. In other words the "marginal" cost of generating the user specified product is small as compared to the cost of the user independently generating the product. Also, the distributed core cost plus the product marginal cost must be favorable as compared to the value of program improvement as a result of including the product. Candidate program areas for product utilization are: (1) forest and range inventory and monitoring, (2) inventoring and monitoring of irrigated croplands for planning agricultural water supply demands, (3) integrating land cover and topographic data for erosion potential and water quality management, and (4) monitoring of urban development patterns as they relate to important or prime agricultural lands.

If it is possible to satisfy some of these needs then the next likely occurrence would be to create a land use data base or information system in a geographic format. This would include point data or summary data for geographic areas that could be digitized in a common map base. The creation and utilization of such a base could considerably expand present land use analysis capability. The next five years should be an exciting period in the utilization of space technology for current, very down to earth, problems relating to land use and change.

References

1. Statistical Reporting Service, U.S. Department of Agriculture, 1975, "Scope and Methods of the Statistical Reporting Service," Misc. Pub. No. 1308, Washington, D.C.

2. Gleason, C. P., Starbuck, R. R., Sigman, R. S., Hanuschak, G. A., Craig, M. E., Cook, P. W. and Allen, R. D., 1977, "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment," SRS-21, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

3. Craig, M. E., Sigman, R. S. and Cardenas, M., 1978, "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat," Statistical Research Division, Economics, Statistics and Cooperatives Service, U.S. Department of Agriculture, Washington, D.C.

4. Ozga, M., Donovan, W. E. and Gleason, C. P., 1977, "An Interactive System for Agricultural Acreage Estimates Using LANDSAT Data," LARS Symposium Proceedings, Purdue University, West Lafayette, Ind.

5. Hanuschak, G. A., 1976, "LANDSAT Estimation with Cloud Cover," IEEE Catalog No. 76CH1103-1 MPRSD, LARS Symposium Proceedings, Purdue University, West Lafayette, Ind.

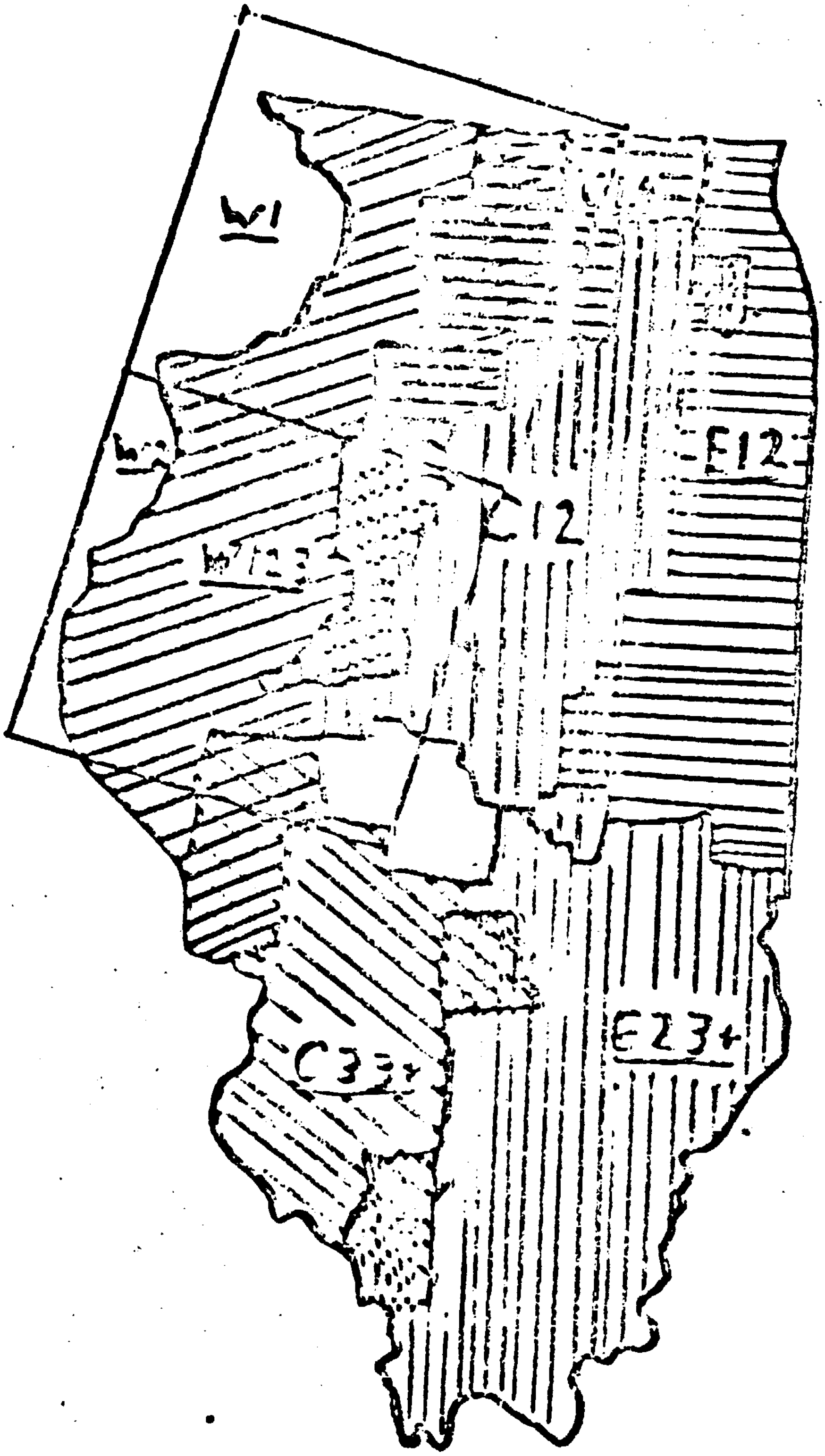

Figure 1. Analysis Areas Illinois 1975

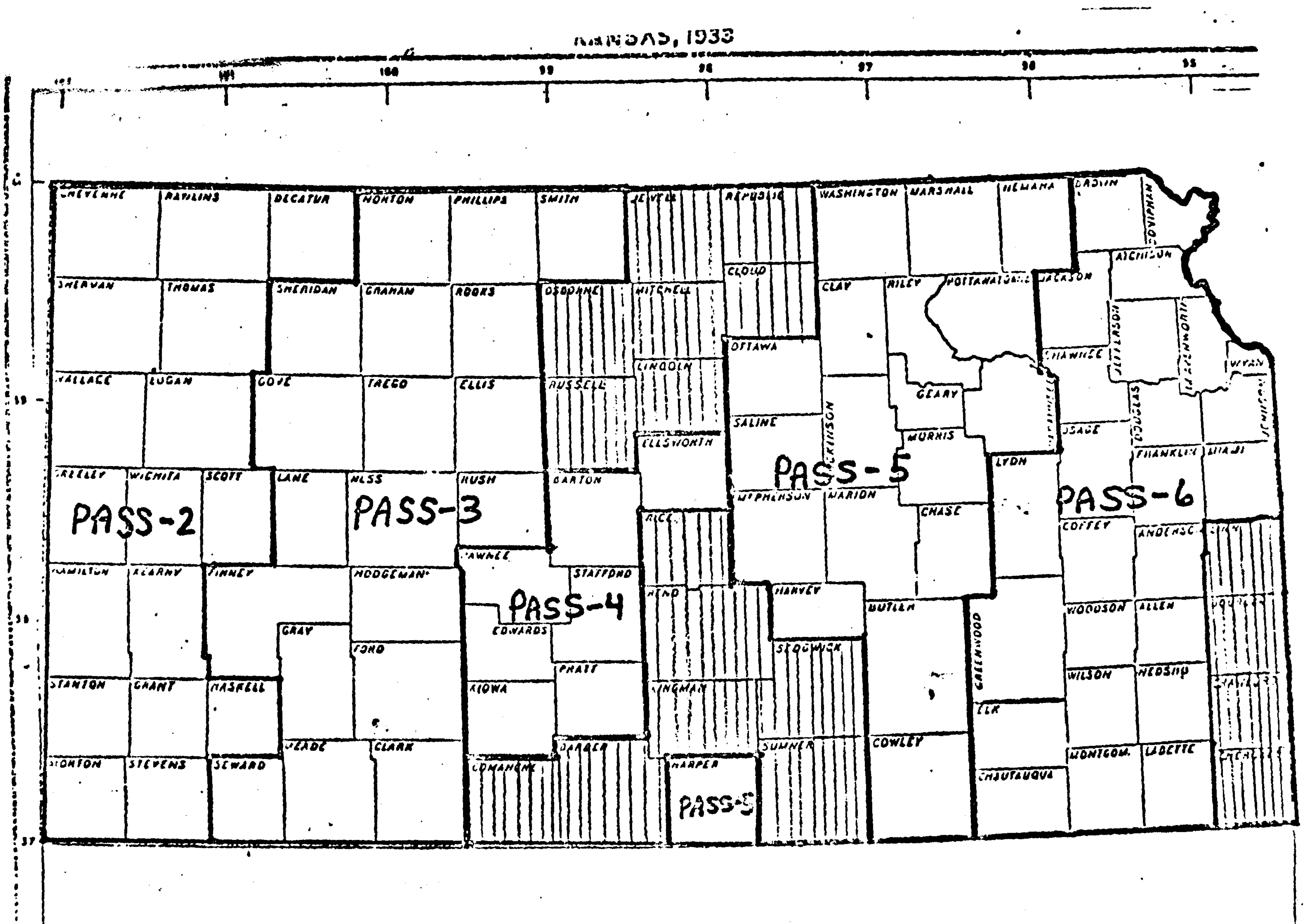

Figure 2: Analysis Districts - Kansas 1976

Table 1. Estimated Hectares of Corn and Soybeans in Illinois for Wholly Contained Counties in Each Analysis Area.

Number of Counties	Estimator	Corn		Soybeans	
		Hectares (000)	r.s.e. (%)	Hectares (000)	r.s.e. (%)
29	Direct Expansion	1,663.4	3.6	622.9	7.7
	Regression	1,669.5	2.5	680.6	5.2
	SSO	1,490.2	-	670.9	-
7	Direct Expansion	482.2	7.1	215.6	13.9
	Regression	477.7	2.9	211.7	8.2
	SSO	484.4	-	203.5	-
20	Direct Expansion	1,176.7	4.5	897.3	5.5
	Regression	1,191.9	4.3	860.9	5.1
	SSO	1,189.7	-	805.5	-
16	Direct Expansion	468.6	9.5	677.9	8.6
	Regression	435.9	8.6	623.2	6.8
	SSO	499.0	-	504.3	-
12	Direct Expansion	720.9	5.6	582.6	6.3
	Regression	638.3	4.1	522.3	6.5
	SSO	725.2	-	559.7	-
32	Direct Expansion	675.6	7.5	984.2	5.2
	Regression	653.6	6.9	954.2	3.8
	SSO	715.1	-	827.6	-
9	Direct Expansion	532.6	8.5	227.4	13.1
	Regression	513.6	4.6	232.3	10.6
	SSO	455.3	-	275.2	-

(Fig.1) were analyzed individually and joined with W3 (not shown on Fig. 1 s W2) to form W123

tained within W2

able 2. Regression Estimates for Corn and Soybeans in Illinois for Individual Counties in Western Pass W 123 (Fig.1)

	Corn		Soybeans	
ounty	Hectares (000)	r.s.e. (%)	Hectares (000)	r.s.e. (%)
dams	67.4	24.0	33.8	35.3
rown	21.7	33.4	9.8	50.7
ureau	102.8	18.7	44.8	33.4
alhoun	22.9	25.1	9.4	39.9
arroll	51.2	17.5	23.1	29.6
ass	37.1	20.3	21.9	25.5
ulton	69.6	29.0	37.0	37.8
reene	55.4	19.2	30.8	24.8
ancock	77.1	19.3	30.3	36.2
enderson	42.1	17.3	15.0	36.4
enry	112.0	17.2	32.1	46.6
ersey	34.7	21.6	19.8	27.0
odaviess	43.8	34.1	11.0	94.2
nox	70.5	19.5	32.2	31.6
ason	52.2	21.3	30.8	27.9
cDonough	65.8	17.4	33.4	26.3
ercer	56.6	18.7	17.8	43.4
organ	59.6	17.6	37.9	20.9
gle	90.2	19.0	20.8	64.2
eoria	50.2	24.0	26.4	32.6
ike	64.8	25.7	31.7	37.3
ock Island	43.3	18.7	11.1	52.7
chuyler	34.0	29.0	14.8	46.2
cott	24.7	19.9	12.7	28.6
tark	37.2	18.2	16.4	32.1
tephenson	69.6	18.6	12.4	81.8
arren	65.5	16.5	25.9	32.2
iteside	98.3	16.2	25.3	49.0
innebago	49.2	21.5	12.0	68.0

Table 3. Planted Area Estimates of Winter Wheat in Kansas for Counties in Each Analysis District.

Analysis District	Number of Counties	Estimator	Hectares (000)	r.s.e. (%)
Pass-2	17	Direct Expansion	902.7	17.6
		Regression	912.9	4.8
		SSO	1,035.6	-
Pass-3	19	Direct Expansion	1,151.7	12.1
		Regression	984.2	6.5
		SSO	1,106.4	-
Pass-4	7	Direct Expansion	431.3	6.9
		Regression	507.7	6.7
		SSO	494.5	-
Pass-5	19	Direct Expansion	960.6	9.1
		Regression	947.5	5.3
		SSO	945.8	-
Pass-6	25	Direct Expansion	304.3	18.6
		Regression	404.7	4.7
		SSO	382.4	-

Table 4. Regression Estimates for Wheat in Kansas for Individual Counties in Pass-3 and Pass-4 (Fig.2)

County	Hectares (000)	r.s.e. (%)
Clark	58.8	17.3
Ellis	37.3	25.8
Finney	64.5	31.6
Ford	99.4	19.5
Gove	46.3	36.0
Graham	47.6	31.4
Grey	50.2	35.6
Hodgman	50.3	20.5
Lane	40.2	21.7
Meade	47.7	26.4
Ness	57.1	24.7
Norton	88.2	14.4
Phillips	61.2	23.4
Rooks	50.7	20.7
Rush	61.4	17.1
Seward	30.6	28.5
Sheridan	45.1	30.3
Smith	72.2	17.3
Trego	40.7	27.3
Barton	92.9	13.0
Edwards	42.5	20.2
Ellsworth	50.3	8.1
Kowa	36.1	19.1
Pawnee	68.2	23.7
Partt	68.7	14.9
Stafford	72.4	17.8