AREA ESTIMATES BY LANDSAT: KANSAS 1976 WINTER WHEAT

Michael E. Craig
Richard S. Sigman
Manuel Cárdenas

Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C.

AUGUST 1978

ABSTRACT

This paper describes the results of research done in Kansas in 1976 related to the estimation of area planted to winter wheat. LANDSAT multispectral scanner (MSS) data were used as the auxiliary variable and ground survey data as the primary variable in a regression estimator [1]. The main goal of the project was to improve the existing and operational ground survey estimation procedures at state, multi-county, and individual county levels. Eighteen of the 105 Kansas counties were not included due to cloud cover problems or lack of training data.

Achievement of the project goal was measured by the reduction in variance of the planted-area estimate computed using LANDSAT and ground data in comparison with the estimate computed using only ground data. The use of LANDSAT as as auxiliary variable was seen to reduce the variation for the multi-county areas (17 to 25 counties each) by 68 to 92 percent.

Several new concepts aided this project in achieving its goals. The major new concept was that of the combined regression, a statistical technique allowing estimation of certain parameters in areas that normally would not have enough samples. Two other new techniques used in this project were "masked" classification and psuedo-counties.

INTRODUCTION

The Economics, Statistics, and Cooperatives Service (ESCS) of the United States Department of Agriculture, is officially responsible for collecting and disseminating current crop and livestock statistics. The Statistical Research Division (SRD) conducts research investigations toward utilizing spectral reflectance data to improve crop area estimating ability. 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.

The following phases of the 1976 Kansas Winter Wheat Project are described in this report: ground data collection, LANDSAT data acquisitions and management, analysis procedures, and estimated areas of winter wheat.

GROUND DATA COLLECTION

As part of its operational program, ESCS conducts in late May an annual nationwide agricultural survey called the June Enumerative Survey (JES). The JES sample units, called segments, are well defined areas of land, typically one-square mile in size. Two levels of stratification are employed in the sample design. The first level strata are the individual states. Secondary strata are areas of land within a state which have similar patterns of land use. Defined in terms of land under cultivation, these secondary strata are determined by visual interpretation of aerial black and white photography. For Kansas, the annual JES allocation of samples is 435 segments. For this study a subsample of 156 segments was available. See Table 1 for the definitions of the agricultural strata. Land with under 15 percent cultivation was estimated using only JES data due to lack of training data.

For this project, the subsampled JES segments were also visited in late April or early May prior to the nationwide JES. Enumerators on both visits obtained complete cropland information including total field and crop area, crop or land use cover type, intended uses of crop fields, field appearance, and date of harvest. Field boundaries corresponding to this information were drawn onto black and white aerial photography (scale 8 inches to one mile) provided by ASCS.

To assist with the interpretation of ground information, low level color infrared (IR) aerial photography was taken of the subsample segments. This photography was prepared by the Remote Sensing Institute of the South Dakota State University at a scale of 5.25 inches to one mile. The color IR acquisition flights occurred during the period from May 1 to May 8, 1976.

Segment and wheat field boundaries were transferred from the black and white photos to the color IR, with all crops or cover types other than wheat being lumped together as "other".

Segment outlines were located and drawn onto USGS quadrangle maps (7½ to 15 minute scales). Field boundaries were then digitized and calibrated to the map base (latitude - longitude) using the color IR and the quadrangle maps. This process produces very precise area measurements for individual fields (called digitized size). Qualitative ground data from the enumerators questionnaires were coded and merged with the digitized size determinations to make field level records for both the April and June visits.

LANDSAT DATA ACQUISITION AND MANAGEMENT

In order to cover the state of Kansas with LANDSAT imagery, six satellite passes were required. Coverage was composed of five passes of three scenes each and another pass consisting of only one scene (to cover the southeast tip of the state).

It was felt that the separation of other land uses from winter wheat would be best in early spring imagery. Hence, the first criterion for selection of LANDSAT imagery was the optimum period for separation which was believed to be late April or May. The second criterion considered the machine quality of digital data over all four bands. Third, the presence or absence of clouds was considered in the selection.

Cloud cover presented a definite problem [2]. Four passes were available which were nearly cloud free. For another pass, two counties (two of the largest wheat producing counties) were lost due to a small cloud covered area. The remaining pass (over some of the best wheat area in central Kansas) had no cloud free imagery for the period required (either LANDSAT I or LANDSAT II). Two partially cloud covered scenes on one date were used to cover a seven county area found to be cloud free in this pass. Table 2 shows the imagery selected for use in this project.

An inspection of the paper products for one pass showed a visible edge separating light pixels in the middle (3M) and dark pixels in the southernmost scene (3S). This edge (or front) ran in a diagonal fashion across the two scenes and was believed to be caused by wet versus dry soil. A possible explantion for this difference was a large rain front over the wet-looking area a day or so before the imagery. Whatever the reason for it, this difference tended to confuse the classification of wheat and other between the two areas within the same day's pass. Healthy wheat fields in the "dry" area looked similar to abandoned wheat or waste fields in the "wet" area.

Once the best available imagery for each pass was determined, the computer compatible tapes were ordered from the EROS Data Center. Also ordered were LANDSAT black and white paper products of each scene on 1:500,000 scale to be used for registration of the digital data to a USGS map base. Once this registration was accomplished, local movements of the predicted segment areas to more exact locations were done.

ANALYSIS PROCEDURES

A. Definition of Analysis Districts

One characteristic of LANDSAT data is that it does not consider political boundaries upon acquisition. Thus, the state was divided into analysis districts which were determined by LANDSAT boundaries and geopolitical boundaries, usually county lines. These analysis districts are not comparable to ESCS's Crop Reporting Districts. An analysis district is a group of counties or parts of counties that is wholly contained in a LANDSAT pass. Estimates for these multi-county areas were made and then individual county estimates were derived from them. See figure 1 for analysis districts.

B. Split Counties

In addition to the 18 counties lost due to clouds or no training data, 14 counties were found to be split across scene boundaries. In earlier experiments (see [1]) when this situation was encountered, psuedo-frames were constructed by putting together the bottom part of the northern scene and the top section from the southern scene thus creating a new frame. This method was only valid when the scenes were from the same date and when counties were cut north or south by LANDSAT frame boundary lines and not east or west by LANDSAT column boundaries. Another drawback of the psuedo-frame approach was that it requires registration of the new psuedo-frames.

A new, quicker method was developed to handle these split counties whether they were divided by lines or columns. The new approach, called the psuedo-county approach, was to digitize a figure that divided a county into two (or more) parts, or sub-counties, that were each completely within one LANDSAT scene, utilizing the fact that scenes partially overlap. This figure was then used to cut up the original digitized county file into parts called psuedo-counties. Each psuedo-county was distinct from all others and was estimated as were the non-split counties. In Kansas, only one county was split across analysis districts and it happened to lie partly in the one-scene pass (Pass-7) that did not have enough training data and so it was not used. The other counties were split across scenes within the same pass and thus the only adjustment to the estimation process was to sum the wheat pixels by strata for each county's parts. For estimation when the psuedo-counties for a given county are in different analysis districts, each psuedo-county would be considered a separate county all the way through the actual estimation and would require adjustment in the number of area frame units for each analysis district. This did not occur in the Kansas study.

C. Clustering and Classification

A separate analysis was conducted for each analysis district using various clustering and classification techniques. For this project, the pixel classifier for each pass was based only on training data from segments interior to the wholly contained counties. This training set of labelled pixels was defined by the digitized segment and field boundaries. Both the wheat and "other" cover types were clustered independently and the resulting signatures combined into statistics files usable for creating a set of discriminant functions for a maximum likelihood classification.

This classification was done at two levels; small scale and large scale. Segment level (small scale) classifications were used to test the performance of the classifier and for estimation of regression parameters. Large scale (or full scene) classifications were used as the independent variable for the actual regression estimate of winter wheat area at analysis district and county levels. Best classifier performance for wheat was achieved with five categories while the "other" cover type comprised from four to seven categories for classification.

Each pass had only one statistics file, with the exception of Pass-3 which had two because of the "wet and dry" areas visible in the image. After examining the visible differences in this pass, it was decided that another level of classification was needed to allow more than one classifier per scene. This classifier, named a "masked" classifier, would take into account completely different signatures for various cover types as a function of location in the scene. New software was written and used in Pass-3 to allow different signatures in the light and dark areas.

Some idea of the performance of the classifier may be obtained from the percent correct, that is, the percentage of the digitized segment pixels that were classified correctly. Since the classifier was trained and tested on the same data (called resubstitution) the numbers may be somewhat optimistic. Table 3 shows the percent correct determinations found using the final statistics files. The final criterion for picking a statistics file will be discussed later.

D. Statistical Methodology

The usual ESCS area estimator, called the direct expansion (DE) estimator, uses only the

ground survey data from the JES. For major crops, such as winter wheat in Kansas, the JES provides state level estimates with relative sampling errors on the order of three to eight percent.

As mentioned earlier, ESCS researchers use LANDSAT data as an auxiliary variable in a regression procedure. In past estimation projects where more segments per analysis district were available, a separate regression equation was estimated for each land use stratum. The technique of pooling strata was used in the Illinois project [1] to alleviate the problem of small sample sizes within an individual stratum. For the Kansas project, a combined regression estimator was developed. Theoretical considerations and formulae for the direct expansion and regression estimator have been described in another paper at this conference by Hanuschak et al [3].

A combined regression estimator is useful where you may not have enough segments to develop a separate regression per stratum for one or more of the strata concerned. This method assumes that the regression coefficient of the estimator is the same for all strata but the intercepts are obtained from the stratum means. Using the small-scale classifications of the sampled segments the regression coefficients for each stratum in each analysis district were computed. These values (as shown in Table 4) were apparently estimates of a common value within each analysis district.

E. Selection of a Classifier for Estimation

Besides estimating the regression coefficients the small scale classification and estimation provide a measure of the performance of the classifier with respect to the variances of the estimates. The various types of regressions are separate, pooling, and the combined strata approach.

In the "separate" regression, the sample coefficients of determination (r-squared) between digitized wheat acres and classified wheat pixels are determined for each stratum. As shown in [3], maximizing the r-square values minimizes the variance of the regression estimates resulting from a classification. Thus, one criterion used to compare classifier performances on the same strata was the respective r-squares. These values were calculated in all Kansas analysis except for strata 12 and 20 in Pass-4 and stratum 11 in Pass-6. In some analysis districts, however, the small sample sizes per stratum make this estimator somewhat unreliable. The pooling of data to derive the classifier relationships in this case assumed only a single stratum was sampled. All segment data from strata 11, 12, and 20 were pooled together and regression was calculated as for an unstratified population. The various r-squared values for the Kansas analysis districts are shown in Table 9 for both equal prior probabilities (EP) and unequal priors computed using proportional to expanded digitized area (PED).

Since the major objective of this project is estimation of winter wheat acreages with reduced variances, maximization of percent correct or reduction of the classification error was not considered in the choice of classifiers. Maximization of the r-squared values was the final criterion used for selection of a statistics file to use for large scale classification in a given analysis district. The equal priors (EP) file was selected in all analysis districts except Pass-6. In this analysis district, the EP classifier tends to classify a large portion of "other" pixels into the wheat categories. Wheat in this area was not a very large crop in terms of total cropland and thus the application of PED priors with small probabilities for wheat tended to give a more reasonable classification and thus better r-squares.

Although in most analysis districts the unequal priors classifier was not chosen for full frame classification, the r-squares found using the PED priors are very close to the corresponding equal priors (EP) values (except in Pass-6). Thus, if the objective of the study was yield computation or some type of stratification based on classified pixels, and not estimation of acreage, the better classifier would be the unequal priors classifier.

WINTER WHEAT AREA ESTIMATES

Multi-county regression estimates for winter wheat area planted were calculated for the various analysis districts. The regression estimates were compared to estimates calculated by direct expansion of the subsample segments, direct expansion of the total 435 JES segments,

and to estimates obtained from the summation of final 1976 county estimates published by the Kansas SSO. The final SSO estimates in Kansas are predominately based on the Kansas State Farm Census. Note that the SSO estimates do not have a calculable variance associated with them because they are based on several non-probability indications, not just the JES direct expansion.

For the multi-strata and multi-county analysis districts, performance of the combined regression estimator was compared to the direct expansion estimator in terms of the relative efficiencies (denoted RE) of the resulting estimates. RE measures the gain, in terms of increased precision, of the combined regression estimate over the respective JES or subsample direct expansion estimate. The equation for calculating the RE follows:

$$RE = \frac{\text{Var (direct expansion)}}{\text{Var (combined regression)}}$$

Table 6 gives the estimated wheat area, coefficients of variation (CV's), and relative efficiencies for all passes with LANDSAT classifications available. Note that the direct expansion estimates shown are based on the subsample chosen for the LANDSAT project.

A swiss cheese estimate consists of regression estimates on the strata included in the classification analysis and prorating the direct expansion estimates of the whole state with respect to area frame units on the strata excluded from the classification analysis. Since some strata were deleted from the classification analysis, "swiss cheese" estimates were computed in order to compare regression estimates with the summations of SSO published county estimates. Table 7 gives the pass-level 'swiss-cheesed' estimates for regression along with the summation of SSO county estimates. The prorated estimate for strata 31, 32, 33, 40, and 50 range from 2.9 percent of the total for Pass-2 to 11.3 percent for Pass-6. The state level estimate uses a direct expansion for Pass-4C.

REFERENCES

1. Gleason, Chapman; Starbuck, Robert R.; Sigman, Richard S.; Hanuschak, George A.; Craig, Michael E.; Cook, Paul W.; and Allen Richard D.; "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., October 1977.

2. Hanuschak, George A.; "The Effect of the LANDSAT Cloud Cover Domain on Winter Wheat Acreage Estimation in Kansas During 1976," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

3. Hanuschak, George A.; "Precision of Crop Area Estimates," Proceedings of the Thirteenth International Symposium on Remote Sensing of Environment, April 1979, Environmental Research Institute of Michigan, Ann Arbor, Michigan.

4. Cárdenas, Manuel; Blanchard, Mark M.; Craig, Michael E.; "On the Development of Small Area Estimators Using LANDSAT Data as Auxiliary Information," August 1978; ESCS Report; USDA: Washington, D.C.

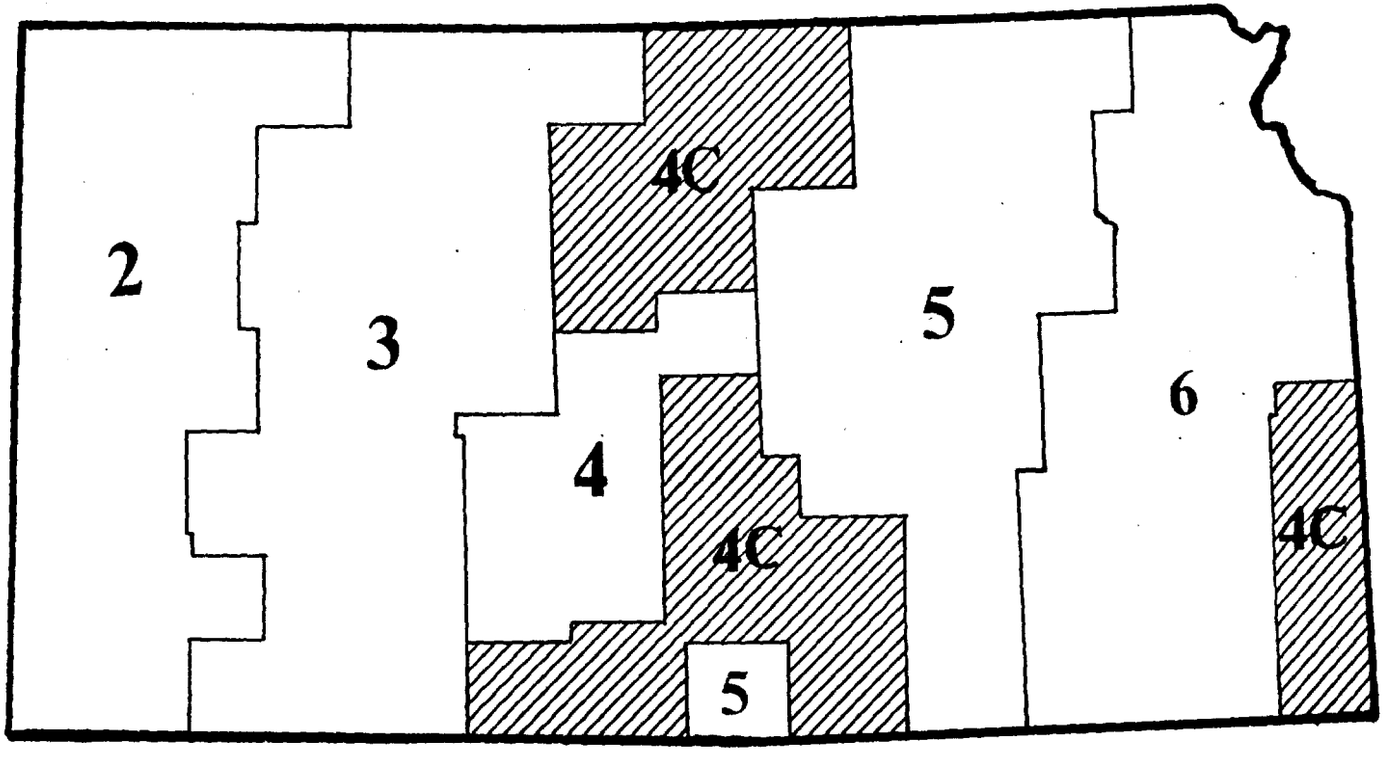

FIGURE 1: FINAL ANALYSIS DISTRICTS BY PASS NUMBER

TABLE 1: KANSAS AGRICULTURE STRATA ALLOCATIONS

Strata	Percent Land Cultivated	Population Number of Segments	LANDSAT Project Sample
11	80-100	25028	68
12	50-79	21704	48
20	15-49	21286	40

TABLE 2: LANDSAT II DATA, KANSAS 1976

Scene	Date	LANDSAT ID-Number	Comments
2N	4/1/76	2435-16404	Clear
2M	4/1/76	2435-16410	Clear
2S	4/1/76	2435-16413	Clear
3N	5/6/76	2470-16335	Clear
3M	5/6/76	2470-16342	Clear
3S	5/6/76	2470-16344	Clear
4M	4/17/76	2451-16291	Heavy Clouds
4S	4/17/76	2451-16293	Heavy Clouds
5N	4/16/76	2450-16230	Clear
5M	4/16/76	2450-16232	Clear
5S	4/16/76	2450-16235	Some Clouds
6N	5/3/76	2467-16165	Clear
6M	5/3/76	2467-16171	Clear
6S	5/3/76	2467-16174	Clear
7S	5/20/76	2484-16113	Clear

TABLE 3: PERCENT CORRECT PIXEL CLASSIFICATIONS FOR SEGMENT DATA

Analysis District	Wheat	Other	Overall
Pass-2	86.83	89.38	88.73
Pass-3	71.73	87.07	82.81
Pass-4	77.28	73.43	75.10
Pass-5	70.07	84.70	80.10
Pass-6	43.48	97.17	92.79

TABLE 4: ESTIMATED REGRESSION COEFFICIENTS FOR EACH STRATUM

Stratum	ANALYSIS DISTRICT				
	PASS-2	PASS-3	PASS-4	PASS-5	PASS-6
11	1.1738	1.1785	1.0435	0.9155	1.2140
12	1.1973	1.1132	-	0.6962	1.6004
20	0.9618	1.0929	-	0.3788	1.6604
Combined	1.0648	1.1206	1.0117*	0.7909	1.6206

*Only two data values existed in Stratum 12 and none in Stratum 20.

TABLE 5: R-SQUARE VALUES BY ANALYSIS DISTRICT AND PRIORS

TYPE & STRATA	PASS 2		PASS 4**		PASS 5		PASS 6		PASS 3
	EP	PED	EP	PED	EP	PED	EP	PED	EP
Separate-11	.8516	.7762	.6161	.6398	.8522	.8361	*	*	.6719
Separate-12	.9953	.9920	*	*	.4785	.3883	.1454	.9836	.9430
Separate-20	.9965	.9950	*	*	.3962	.5098	.0832	.7429	.7100
Pooling	.8818	.8215	.5975	.5614	.7450	.7181	.1911	.7659	.8073

*Not calculated due to lack of data
**Pass-4 pooling includes strata 11 and 12 only

TABLE 6: PLANTED AREA ESTIMATES OF WINTER WHEAT FOR STRATA 11, 12, and 20.

Analysis District	Number of Segments	Number of Counties	Estimator	Estimate (hectares)	CV	RE
Pass-2	29	17	Regression	886500	4.9	13.1
			Direct Expansion	876300	18.1	
Pass-3	35	19	Regression	946900	6.7	4.8
			Direct Expansion	1114400	12.5	
Pass-4*	11	7	Regression	382800	7.8	1.3
			Direct Expansion	459300	7.3	
Pass-5	31	19	Regression	876700	5.5	3.2
			Direct Expansion	889800	9.8	
Pass-6	16	25	Regression	358900	4.8	10.6
			Direct Expansion	258500	21.7	
Overall**	122	87	Regression	3488600	2.8	

*Strata 11 and 12 only
**Stratum 20 estimate was prorated from state estimate in Pass-4.

TABLE 7: PLANTED AREA ESTIMATES OF WINTER WHEAT FOR ALL STRATA

Analysis District	Estimator	Estimate (Hectares)	CV
Pass-2	Regression	912900	4.8
	SSO Sum	1035600	-
Pass-3	Regression	984200	6.5
	SSO Sum	1106400	-
Pass-4	Regression	431300	6.9
	SSO Sum	494500	-
Pass-5	Regression	947500	5.3
	SSO Sum	945800	-
Pass-6	Regression	404700	4.7
	SSO Sum	382400	-
State	Regression	5141900	2.7
	SSO Sum	5220400	-

*Regression and direct expansion estimators are based on the 'swiss cheese' technique and use only the sub-sample segment data for strata 11, 12, and 20.