SMALL AREA ESTIMATORS: COUNTY CROP ACREAGE ESTIMATES USING LANDSAT DATA[1]

by Manuel Cárdenas
New Mexico State University

Michael E. Craig
U.S. Department of Agriculture

Mark Blanchard
U.S. Department of Agriculture

1979

ABSTRACT

This research considers several county estimators that incorporate LANDSAT satellite data with data obtained from U.S. Department of Agriculture operational June Enumerative Survey (JES). The radiometric satellite data are classified into the different crop types using a maximum likelihood discriminant function. The classified data are then used as the values of an auxiliary variable to JES questionnaire data. Approximate variance formulas for the proposed county estimators are presented.

BIOGRAPHIES

Dr. Cárdenas holds a B.S. and M.A. in Mathematics from Texas A&I and a Ph.D. in Statistics from Texas A&M. He is presently employed by the Department of Experimental Statistics at New Mexico State University. He was a 1977–1978 American Statistical Association Faculty Fellow with the Statistical Research Division, Washington, D.C.

Michael Craig is presently a Mathematical Statistician in the New Techniques Section, Statistical Research Division, ESCS, U.S. Department of Agriculture. He has been employed by USDA since September 1974 and is a member of the American Statistical Association. Mr. Craig received a B.S. in Mathematics Education in 1973 and an M.S. in Statistics in 1975, both degrees from Virginia Polytechnic Institute and State University in Blacksburg, Virginia.

Mark M. Blanchard holds an M.S. in Statistics from Oklahoma State University. Upon graduation he was employed by the Statistical Research Division, ESCS, USDA. He is presently a statistician for the Vermont Department of Health, Burlington, Vermont.

INTRODUCTION

This article is concerned with the estimation of small area characteristics from a sample designed for making large area estimates. In particular the interest is in making crop acreage estimates at the county level from data obtained in the June Enumerative Survey (JES), a survey conducted at the state and national levels.

The Economics, Statistics, and Cooperatives Service (ESCS) has been charged with making area estimates of crops based on the JES. County estimates are an integral part of the ESCS program of crop estimates. ESCS receives direct funding for making certain county estimates and has annual agreements with Agriculture Stabilization and Conservation Service (ASCS) and the Federal Crop Insurance Corporation to provide selected additional county data. State Statistical Offices (SSOs) are responsible for the preparation of county estimates. The county estimates are made by first allocating the official state estimate for a given crop proportionately among crop reporting districts (collections of contiguous counties) and then apportioning the estimates for these districts among the individual counties. Besides the information obtained from the JES the SSOs also use data derived from several other sources in their estimation procedures. Two such sources are (1) a mail survey, which may include 50–100 respondents, and (2) the agricultural census. The estimation procedure thus varies from state to state and from county to county depending upon the availability of data. No variance estimates are computed.

Since the advent of LANDSAT data, the New Techniques Section of the Statistical Research Division (SRD) of ESCS has focused its resources on the development of methodology that incorporates these data with that obtained from the JES for more efficient estimation. The potential for efficient estimation as well as a uniform

county estimation procedure using LANDSAT data has been recognized and is presently being investigated.

Actually, the small area estimation problem has attracted considerable attention in other governmental agencies as well. The National Center for Health Statistics (National Center for Health Statistics, 1977; Schaible et al., 1979) and the Department of Commerce (Gonzalez and Wakeberg, 1973), for example, are very involved in developing small area estimators for certain characteristics (e.g., unemployment rates, percent of population having completed college, percentage disabled by chronic conditions, population growth, etc.) from large area samples such as the Current Population Survey (CPS) and the Health Interview Survey (HIS).

DATA ACQUISITION

Before proceeding to the estimators, a brief discussion to acquaint the reader concerning the data acquisition seems imperative. A more detailed discussion can be found in several sources (e.g., see Enumerative & Multiple Frame Survey (1978) and Von Steen and Wigton (1976)).

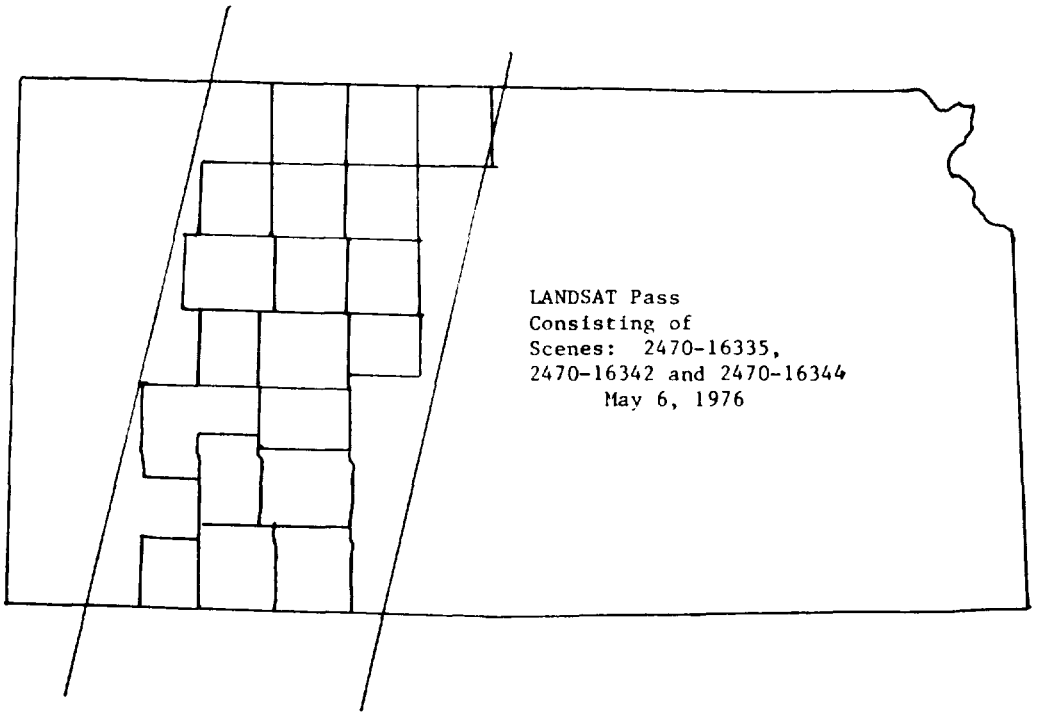
The JES is an annual argicultural survey conducted in late May. The sample for this survey employs two levels of stratification. The first level strata are the 50 individual states. The secondary strata are areas within a state that have similar patterns of land use as determined by photo-interpretation of aerial photography. The secondary strata are divided into primary sampling units, which can be further subdivided into sampling units. The sampling units chosen for the JES are called segments and are well-defined areas of land varying in size depending on the stratum in which they are located. Typically these segments are one square mile in size in the more cultivated strata. The acreage devoted to each crop or land-use are recorded for each field in each segment during the JES interviews.

The basic element of LANDSAT data is called a signature and is the set of measurements taken by the satellite's multispectral scanner (MSS) of an area of the earth's surface approximately one acre in size. The individual MSS resolution areas, are called pixels. The MSS measures the amount of radiant energy reflected and/or emitted from the earth's surface in various regions (bands) of the electromagnetic spectrum.

Presently satellite data are obtained from LANDSAT II and LANDSAT III. A given point on the earth's surface is imaged once every eighteen days by the same satellite and once every nine days by either of the two satellites. Each satellite pass covers an area 185 kilometers wide. Figure 1 shows one such pass over Kansas.

The satellite information used by ESCS is extracted from LANDSAT data by classifying individual signatures as to probable crop type. This classification is performed by a collection of discriminant functions. Therefore, LANDSAT data are census data, but of questionable reliability due to misclassification.

Figure 1. The Kansas study area consisting of the 19 countries wholly contained within the three scenes. 2470−16335, 2470−16342 and 2470−16344, May 6, 1976.



PRELIMINARY DISCUSSION

The county estimation procedure presented here makes the assumption that the mean number of pixels per segment in stratum h within county i classified as the crop in question, $\bar{X}_{ih}$, is fixed with respect to the JES sample. With the present procedure of sampling and classification this assumption is not satisfied. However, with a large enough sample the variability of these values should be negligible in comparison with the variability of the y_{ijh} values (i.e., the reported acreage of the crop in question in the j^{-th} segment of the h^{-th} stratum within the i^{-th} county). A recent study (Sigman et al., 1977), using a jackknife method, on 83 sampled segments tends to verify this.

In developing the estimates, the JES data that were taken at the segment level must be combined with the LANDSAT data that can be taken at the county level. This is done by noting that whenever a segment is chosen the county in which that segment is contained is automatically selected also. Moreover, taking a small sample without replacement from a large population is practically equivalent to taking the sample with replacement from that population in that the probability of that particular sample being chosen is about the same for both procedures. To the extent that these two procedures of sampling yield the same results, it can be seen that taking a simple random sample of n segments from a state is equivalent to the following two-stage sampling scheme: (1) a sample of n counties is taken with replacement and with probability proportional to size; (2) a simple random sample of t_i (t_i being the number of times county i appears in the sample) segments is taken from each of the distinct counties in the sample. This two-stage sampling procedure was first proposed, in a more general form (i.e., a subsample of size $m_i t_i$ rather than t_i is taken from the i^{-th} primary unit in the sample), by Sukhatme and Sukhatme (1970, p. 328). The estimators and variances presented in this article are based on this two-stage sampling scheme. The derivations of variances and their estimators follow the logic used by Sukhatme and Sukhatme and are found in a previous paper (Cárdenas et al., 1978).

COUNTY ESTIMATORS

If the assumption were made that the mean per segment in land-use stratum h of the crop in question for each county were equal to the mean of the populations $\bar{\bar{Y}}_h$, the total for a particular county, say county k, would be

$$Y_k = \sum_{h \,\varepsilon\, C_k} M_{kh} \, \bar{\bar{Y}}_h$$

where $\sum_{h \,\varepsilon\, C_k}$ denotes the summation over all strata in county k, and M_{kh} = total number of segments in the h-th stratum within the k-th county. An unbiased estimate of Y_k under this assumption is

$$\hat{Y}_k = \sum_{h \,\varepsilon\, C_k} M_{kh} \, \bar{\bar{Y}}^*_h$$

where

$\bar{\bar{Y}}^*_h = \frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih} \, \bar{Y}^*_{ih}$ an unbiased estimate of $\bar{\bar{Y}}_h$;

$\bar{Y}^*_{ih} = \sum_{j=1}^{t_{ih}} y_{ijh} / t_{ih}$ The sample mean of the acreage per stratum h within county i;

n_h = number of counties (distinct or otherwise) in the sample of the h-th stratum,

N_h = number of counties containing any part of the h-th stratum.

Recognizing that the above assumption is not satisfied in general, we then search for supplementary information that indicates deviation of a particular county mean from the population mean. This information is found in the form of classified pixels in each county. Using these auxiliary data we define the family of estimators:

$$\hat{Y}_{Bk} = \sum_{h \,\varepsilon\, C_k} M_{kh} \, [\bar{\bar{Y}}^*_h + B_h \, (\bar{X}_{kh} - \bar{X}_h)] \quad (1)$$

where $\bar{X}_h$ = the mean number of pixels classified as the crop in question for stratum h. If $\bar{X}_{kh}$ is greater (less) than the mean of stratum h for the given satellite pass, then the mean area estimate should be increased (decreased) by an amount proportional to this difference. It follows that the B_h's should be positive.

If classification is such that $y_{ijh} = Ax_{ijh}$, where A is some constant, then using $B_h = \bar{Y}^*_h / \bar{X}_h$ in eq. (1) yields an unbiased estimator, $\hat{Y}_{rk}$, of Y_k. Other possible values that one might try for the B_h's would be the least squares-like estimates,

$$B_h = \frac{M_h \sum_{i=1}^{N_h} t_{ih} \, (\bar{X}_{ih} - \bar{X}_h) \, \bar{Y}^*_{ih}}{n_h \sum_{i=1}^{N_h} M_{ih} \, (\bar{X}_{ih} - \bar{X}_h)^2} .$$

These values of B_h substituted into eq. (1) yield unbiased estimates, $\hat{Y}_{sk}$, of Y_k when $y_{ijh} = a + b_h x_{ijh}$, where a and b_h are constants. Actually, in this case B_h is an unbiased estimate of Cov $(\bar{X}_{ih}, \bar{Y}_{ih})$ /V $(\bar{X}_{ih})$ for all h. If b_h

= b for all h, then we can use the combined data for all strata to estimate b. In this case substitution of

$$B_h = \frac{\sum_{h=1}^{L} \frac{M_h^2}{n_h} \sum_{i=1}^{N_h} t_{ih} (\bar{X}_{ih} - \bar{X}_h) \bar{Y}^*_{ih}}{\sum_{h=1}^{L} M_h \sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_h)^2},$$

where L is the number of strata, gives unbiased estimates, $\hat{Y}_{ck}$, of Y_k. The sum over k for all three of the estimators, $\hat{Y}_{rk}$, $\hat{Y}_{sk}$, and $\hat{Y}_{ck}$, is unbiased for the population total.

The estimators, $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ can be written as

$$\hat{Y}_k = \sum_{h \in C_k} M_{kh} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} w_{ih(k)} t_{ih} \bar{Y}^*_{ih}\right] \quad (2a)$$

where

$$w_{ih(k)} = \begin{cases} \bar{X}_{kh}/\bar{X}_h, & \text{for } \hat{Y}_{rk} \\ 1 + M_h \dfrac{(\bar{X}_{ih} - \bar{X}_h)(\bar{X}_{kh} - \bar{X}_h)}{\sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_h)^2}, & \text{for } \hat{Y}_{sk} \end{cases}$$

The estimator, $\hat{Y}_{ck}$, can be written as

$$\hat{Y}_{ck} = \sum_{l \in C_k} M_{kl} \sum_{h=1}^{L} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} w_{ihl(k)} t_{ih} \bar{Y}^*_{ih}\right] \quad (2b)$$

with

$$w_{ihl(k)} = \delta_{lh} + \frac{M_h^2 (\bar{X}_{ih} - \bar{X}_h)(\bar{X}_{kl} - \bar{X}_l)}{\sum_{h=1}^{L} M_h \sum_{i=1}^{N_h} M_{ih} (\bar{X}_{ih} - \bar{X}_h)^2}$$

and

$$\delta_{lh} = \begin{cases} 1 \text{ if } l = h \\ 0 \text{ otherwise} \end{cases}$$

This estimator will not be discussed further, since its variance should be at best as large as the variance of $\hat{Y}_{sk}$.

The variance for $\hat{Y}_k$ is derived in Cárdenas et al. (1978) and is given by

$$V(\hat{Y}_k) = \sum_{h \in C_k} M_{kh}^2 \left\{ \frac{1}{n_h} \sum_{i=1}^{N_h} (M_{ih}/M_h) [w_{ih(k)} \bar{Y}^*_{ih} - \sum_{i=1}^{N_h} \frac{M_{ih}}{M_h} w_{ih(k)} \bar{Y}^*_{ih}]^2 \right.$$

$$\left. + \frac{1}{n_h M_h} \sum_{i=1}^{N_h} (M_{ih} - 1) w_{ih(k)}^2 S_{ih}^2 - \frac{n_h - 1}{n_h M_h^2} \sum_{i=1}^{N_h} M_{ih} w_{ih(k)}^2 S_{ih}^2 \right\} \quad (3)$$

where

$$S_{ih}^2 = \frac{\sum_{j=1}^{M_{ih}} (y_{ijh} - \bar{Y}_{ih})^2}{M_{ih} - 1}$$

and

$$\bar{Y}_{ih} = \sum_{j=1}^{M_{ih}} y_{ijh}/M_{ih}.$$

The variance for $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ are obtained from eq. (3) by the appropriate substitution for $w_{ih(k)}$.

If the assumption that the within-county variance is equal for all counties is made, then an unbiased estimate of the variance formula given by eq. (3) is

$$v(\hat{Y}_k) = \sum_{h \in C_k} M_{kh}^2 [n_h (n_h - 1)]^{-1}$$

$$\cdot \left\{ \sum_{i=1}^{n_h} (w_{ih(k)} \bar{Y}^*_{ih} - \frac{1}{n_h} \sum_{i=1}^{n_h} w_{ih(k)} \bar{Y}^*_{ih})^2 + s_{wh}^2 \right.$$

$$\left. [\sum_{i=1}^{n_h} (1 - 1/t_{ih}) w_{ih(k)}^2 - \frac{n_h - 1}{M_h} \sum_{i=1}^{n_h} w_{ih(k)}^2] \right\}$$

where

$$s_{wh}^2 = \frac{\sum_{i=1}^{n'_h} \sum_{j=1}^{t_{ih}} (y_{ijh} - \bar{Y}^*_{ih})^2}{n_h - n'_h}$$

is the pooled within-county variance and n'_h = the number of distinct counties in the sample within the h-th stratum. Again, estimated variances for $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ are obtained by the appropriate substitution for $w_{ih(k)}$. The assumption of equal within-county variances is used because some counties have only one observation in some

strata. Actually, in most cases it takes more than one pass of the satellite to completely cover a state. Since these passes occur at different dates and since signatures for the same crop differ from pass to pass, each pass is used as a poststratum. The county estimation is therefore made by poststrata, which relaxes the assumption from equal within-county variances for the state to equal within-county variance within each pass.

AN APPLICATION

The estimators were used on actual data taken from a 19 county area of Kansas (Fig. 1). For the purpose of this article the three most cultivated strata (80+% cultivated, 50−80% cultivated, and 14−49% cultivated) in the particular 19 counties in Kansas were considered as the complete population, since the contribution from the other strata was considered negligible in this area.

As was stated previously the estimators developed in this article are unbiased under certain linear relationships between the reported acreage of the crop in question and the number of classified pixels of this crop per segment within each stratum. Unfortunately the sample size used in this study was too small to permit any real analysis to determine which particular estimator to use. However it is known that the relationship between reported acreage and classified pixels is not strictly linear and therefore some bias is bound to exist. The amount of bias is dependent on the degree to which the relationship between the two variables fails to be linear. This is a topic for future research.

Because there was no agricultural census in 1976 with which to assess the bias, the SSO estimates were used for comparison. Table 1 shows the estimates obtained for the study area in Kansas by the SSO, $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ estimators. Actually, the $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ estimators are unbiased under different conditions and both should not be used simultaneously. However, since the sample was small, it was felt that the existant conditions could not be evaluated accurately and so estimates using both estimators were presented.

As can be seen some of the differences between the SSO estimates and the estimates calculated with either $\hat{Y}_{rk}$ or $\hat{Y}_{sk}$ are fairly substantial. These differences can be attributed to the fact that the SSO estimates are themselves biased county estimates, partly, if not wholly, based on the JES, which is designed for making state and national estimates. Moreover, the $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ estimators are based on a random sample of only 40% of the JES data available for the 19 county area in question, along with the LANDSAT data. The small number of segments (11, 12, 12, respectively for the three most cultivated strata in the 19 county area) subsampled for the LANDSAT project was determined to reduce the impact of the LANDSAT research on the 1976 JES data collection effort.

Table 1. Winter Wheat Estimates[a] and Coefficients of Variation for the 19 County Area in Kansas

	SSO[b]	Y_{rk}		Y_{sk}	
County	est	est	CV	est	CV
Clark	42	54.3	16.6	79.3	31.0
Ellis	58	38.0	13.1	18.9	74.5
Finney	80	62.2	14.4	50.5	41.3
Ford	94	106.1	13.5	113.3	32.0
Gove	53	56.0	12.9	33.6	46.3
Graham	46	57.4	15.0	52.7	14.6
Gray	68	52.1	15.8	47.7	27.2
Lane	52	47.9	14.8	49.2	18.7
Meade	73	43.9	13.3	37.6	23.7
Ness	87	66.2	13.4	58.7	13.8
Phillips	39	69.7	16.8	92.4	29.6
Rooks	56	56.5	14.1	58.8	18.1
Rush	76	69.9	12.8	83.2	24.7
Seward	34	25.9	15.5	16.1	60.1
Sheridan	47	50.0	13.1	45.5	34.4
Smith	49	84.1	13.3	108.5	25.3
Trego	53	46.9	13.8	33.3	28.3
Hodgeman	57	56.1	13.0	48.1	17.3
Norton	46	71.2	13.7	86.6	21.3

[a] The estimates are given in thousand hectares and except for the SSO are based on only 3 strata.

[b] SSO is the estimate derived by the State Statistical Offices and is based on all the strata.

CONCLUSIONS

This estimation procedure was tried also by the New Techniques Section of ESCS on 40% of all of the Kansas 1976 JES winter wheat data (Craig et al., 1978). The results seem promising, but unfortunately they can only be compared to the SSO estimates, which are of unknown reliability. Presently the procedures are being tried on the 1978 JES data for Iowa.

As was mentioned in the text, the estimators suggested are unbiased under certain linear conditions. However, the classification is not strictly linear. The classification and therefore the estimation is expected to get better due to a smaller area of resolution when LANDSAT D data becomes available in 1981.

The work done here represents a first step towards developing a uniform scheme of county estimators for crop acreage estimation employing LANDSAT data. Further study is of value to better determine the quality of these estimators and to examine their feasibility from an operational standpoint.

ACKNOWLEDGMENTS

The first author wishes to thank the Statistical Survey Institute for its support. Ms. Margaret Land from New Mexico State University and the referees also deserve the authors' gratitude for their reviews and suggestions, which helped strengthen this article.

FOOTNOTES AND REFERENCES

1. This material is a revision of an earlier work, which was presented at the annual statistical meetings at San Diego, California, August 15, 1979, and appears in a booklet published by the U.S. Department of Commerce, Bureau of the Census entitled *Methodology and Use of Small-Area Statistics in Decisionmaking and 1977 Economic Censuses and their Use in Private and Public Sectors*.

Cárdenas, M., Blanchard, M. and Craig, M. E.
1978 "On the development of small area estimators using LANDSAT data as auxiliary information." ESCS, USDA, Washington, D.C.

Craig, M. E., Sigman, R. S., and Cárdenas, M.
1978 "Area estimates by LANDSAT: Kansas 1976 winter wheat." ESCS, USDA, Washington, D.C.

Gonzalez, Maria Elena, and Wakeberg, Joseph
1973 "Estimation of the error of synthetic estimates," unpublished paper presented at the First Meeting of the International Association of Survey Statisticians, Vienna, Austria.

Enumerative & Multiple Frame Survey
1978 Interviewer's Manual. ESCS, USDA, Washington, D.C., June.

National Center for Health Statistics
1977 "State estimates of disability and utilization of medical services: United States, 1969–71." DHEW Publication No. (HRA) 77–1241, Washington, D.C.

Schaible, W. L., Cassidy, R. J., Schnack, G. A., and Brock, D. B.
1979 "Small area estimation: An empirical comparison of conventional and synthetic estimators for states." submitted for publication.

Sigman, R. S., Gleason, C. P., Hanuschak, G. A., and Starbuck, R. R.
1977 "Stratified acreage estimates in the Illinois Crop-Acreage experiment." in Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data. W. Lafayette, Indiana: Purdue University.

Sukhatme, P. V. and Sukhatme, B. V.
1970 *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, p. 328.

Von Steen, D. H. and Wigton, W. H.
1976 "Crop identification and acreage measurement utilizing LANDSAT imagery." Statistical Reporting Service, USDA, Washington, D.C.