SMALL AREA ESTIMATORS: COUNTY CROP ACREAGE ESTIMATES USING LANDSAT DATA

19-8

by

Manuel Cárdenas*
New Mexico State University

Michael E. Craig
U.S. Dept. of Agriculture

Mark Blanchard
U.S. Dept. of Agriculture

1979

This research considers several county estimators which incorporate LANDSAT satellite data with data obtained from USDA's 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 is then used as an auxiliary variable to JES questionnaire data. Approximate variance formulas for the proposed county estimators are presented.

Introduction

This paper 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

* Manuel Cárdenas is a 1977-78 ASA faculty fellow with the Statistical Research Division, Room 4844 South Building, Washington, D.C. 20050.

annual agreements with Agriculture Stabilization and Conservation Service (ASCS) and the Federal Crop Insurance Corporation to provide selected additional county data. State Statistical Offices (SSO's) 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 SSO's also use data derived from several other sources in their estimation procedures. Two such sources are (a) a mail survey which may include 50-100 respondents and (b) 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 but the coefficients of variation are believed to be on the order of 10 percent or more.

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 [5, 6] and the Department of Commerce [3], 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.

Data Acquisition

Before proceeding to the estimators a brief discussion to acquaint the reader on the data acquisition seems imperative. A more detailed discussion can be found in several sources (e.g. see [4] and [9]).

The JES is an annual agricultural 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 which 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 is 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.

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 is census data but of questionable reliability due to misclassification.

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 [7] using a jackknife method on 83 sampled segments tends to verify this.

In developing the estimates, the JES data which was taken at the segment level must be combined with the LANDSAT data which 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. To the extent that these two procedures of sampling are the same, it can be seen that taking a simple random sample

of n segments from a state is the same as the following two-stage sampling scheme: (a) a sample of n counties is taken with replacement and with probability proportional to size; (b) a simple random sample of t_i (t_i being the number of times county i appears in the sample) segments are 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 sub-sample 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 [8]. The estimators and variances presented in this paper 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 reference [1].

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

where $\sum_{h \in C_k}$ 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 is

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

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 segment in stratum h within county i ;

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

and 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 which 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,

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

where $\bar{X}_h$ = the mean number of pixels classified as the crop in question for stratum h in county k. If $\bar{X}_{kh}$ is greater (less) than the mean of stratum h for the given satellite pass, than 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 $y_{ijh} = A x_{ijh}$, where A is some constant, then using $B_h = \bar{\bar{Y}}^*_h / \bar{X}_h$ in equation (1) yields an unbiased estimator, $\hat{Y}_{rk}$, of Y_k. Other possible values which one might try for the B_h's would be the least squares-like estimates

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

These values of B_h substituted in (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

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

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

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

$$\hat{Y}_{ck} = \sum_{\ell \epsilon C_k} M_{k\ell} \sum_{h=1}^{L} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} w_{ih\ell(k)} t_{ih} \bar{Y}^*_{ih}\right]$$

with

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

and

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

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 $\tilde{Y}_k$ is derived in [8] and is given by

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

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

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 (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 (3) is

$$v(\tilde{Y}_k) = \sum_{h\epsilon C_k} M_{kh}^2 [n_h(n_h - 1)]^{-1} \left\{ \sum_{i=1}^{n_h} (w_{ih(k)} \bar{Y}^*_{ih} - \frac{1}{n_h} \sum_{i=1}^{n_h} w_{ih(k)} \bar{Y}^*_{ih})^2 \right.$$

$$\left. + s_{wh}^2 \left[\sum_{i=1}^{n_h} (1 - 1/t_{ih})\, w_{ih(k)}^2 - \frac{n_h - 1}{M_h} \sum_{i=1}^{n_h} w_{ih(k)}^2 \right] \right\} \quad (4)$$

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

is the pooled within county estimate 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 needed 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 post stratum. The county estimation is therefore made by post strata which relaxes the assumption from equal within county variances for the state to equal within county variance within each pass.

Conclusions

This estimation procedure was tried by the New Techniques Section of ESCS on 40% of the Kansas 1976 JES winter wheat data [2]. 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 when LANDSAT D data becomes available in 1981.

One could also consider other values for the B_h's. Also, a regular regression estimator could be developed. This approach would require "super-population" considerations.

References

[1] 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.

[2] Craig, M. E., Sigman, R. S., and Cárdenas, M., (1978); "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat," ESCS, USDA, Washington, D.C.

[3] 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.

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

[5] 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.

[6] Schaible, W. L.; Cassidy, R. J.; Schnack, G. A.; and Brock, D. B.; "Small Area Estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States," submitted for publication.

[7] Sigman, R. S.; Gleason, C. P.; Hanuschak, G. A.; and Starbuck, R. R.; (1977); "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, Indiana.

[8] Sukhatme, P. V. and Sukhatme, B. V., (1970); Sampling Theory of Surveys with Applications. Iowa State University Press, Ames, Iowa.

[9] Von Steen, D. H. and Wigton, W. H. (1976); "Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery," Statistical Reporting Service, USDA, Washington, D.C.