MULTIPHASE CROP ACREAGE ESTIMATION INCORPORATING LANDSAT DATA

Randall W. Thomas, Remote Sensing Research Program
University of California at Berkeley

ABSTRACT

Landsat full frame data offers the possibility for relatively inexpensive auxiliary variate measurement over crop acreage sample frames. An example is developed in which rough estimates of sample unit wheat proportions obtained from full frame Landsat data are used to optimally allocate a small calibration sample. Crop area measurements were limited to intensive manual or machine analysis of Landsat data. Resulting stratified regression crop proportion estimates suggest that significant potential exists for precision or cost improvement over designs employing only historical county data for area stratification purposes. A link with conventional crop inventory designs is also described. This technique, currently being demonstrated for the California Department of Water Resources, employs a three phase regression sample of Landsat, aerial photography, and ground data to produce irrigated acreage estimates of extremely high precision at a state level.

I. INTRODUCTION

One of the most important aspects controlling the success of any inventory system is the sampling/aggregation plan utilized. Substantial differences in final estimate precision, bias, and cost can occur depending on which sample design is selected. Moreover, the number of parameters (e.g. different crop acreages or yields) that can be estimated and the reporting level at which they are available are similarly affected by the design.

The advent of timely and relatively inexpensive remote sensing data has fostered new agricultural inventory sample design options and improved estimate performance possibilities. Significant progress has been made in this regard through the Large Area Crop Inventory Experiment (LACIE) jointly sponsored by the National Aeronautics and Space Administration (NASA), the U.S. Department of Agriculture (USDA) and the National Oceanic and Atmospheric Administration (NOAA). The LACIE employed Landsat data and a stratified simple random sample design to develop wheat acreage and production estimates when little or no ground data was available (MacDonald et al. 1975). While this design enabled improved estimate precision in many foreign agricultural situations, the potential use of Landsat data for variance reduction in domestic inventory situations with significantly higher precision requirements has received much less attention.

This paper describes the development of two Landsat-aided inventory procedures useful in domestic agricultural survey applications. The first, a stratified two phase design, was intended to demonstrate increased precision capabilities available within the LACIE system itself at the same level of budget. The second, a stratified three phase design, has been developed for the California Department of Water Resources to demonstrate that Landsat data can be tied with aerial photography and conventional ground survey information to produce precise acreage estimates in less time and at lower cost than current survey techniques.

II. TWO PHASE SAMPLE FOR WHEAT ACREAGE ESTIMATION

The two phase technique employed manually processed Landsat full frame wheat or cultivated land proportion estimates from a large number of segments comprising a first sample phase to optimally allocate a small phase two sample of computer or manually processed segments. Proportion estimates from each phase were then linked by a regression estimator to provide wheat proportion estimates and standard errors by reporting unit.[1] A simulated second year LACIE inventory system was used as a base for performance (precision, cost) comparison.

A. Information Requirements

The information target for the inventory was defined to be wheat acreage sown (1973-74) expressed as a proportion of total land area by county and by U.S. Department of Agriculture (USDA) Crop Reporting District (CRD). Counties and CRD's were defined on a "pseudo" basis, meaning that their boundaries were slightly modified so as to avoid splitting inventory sample segments.

Inventory data were purposely limited to that available in the LACIE counterpart; namely Landsat full frame color infrared transparencies (not real-time), Landsat digital data for a small sample of five mile by six mile segments (30 sq. mi.) and ancillary crop calendar and cropping practice information.

B. Sample Design Specification

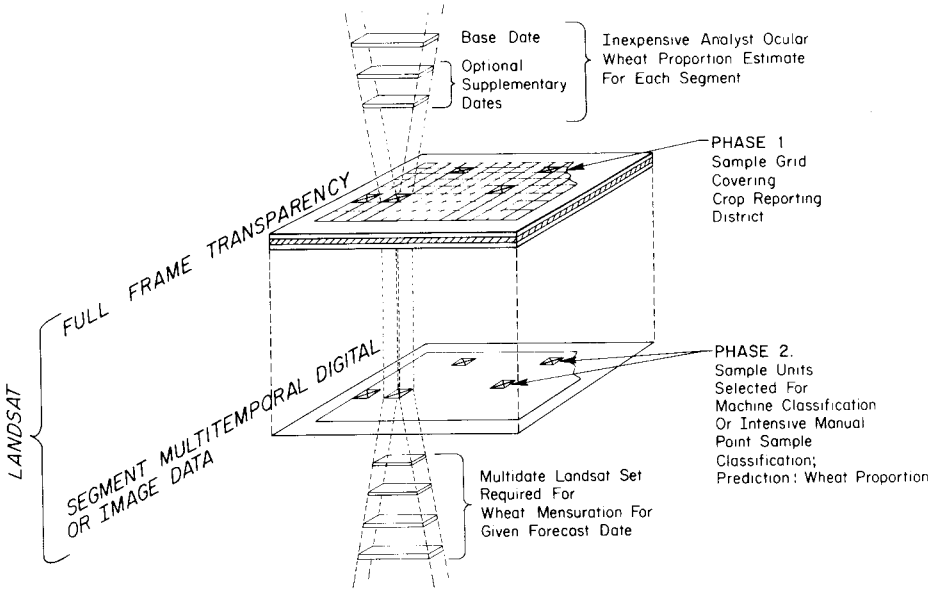
A stratified double sampling (i.e., two phase) design was selected to demonstrate the capability of remote sensing-aided systems to achieve an at-harvest CRD wheat acreage estimate within five percent of the corresponding USDA estimate 95 times out of 100.

Figure 1 illustrates the two phase sampling concept as applied to the wheat proportion estimation problem. The top layer in the figure was defined to represent a CRD-wide phase 1 sample frame composed of standard 5 x 6 mile sample segments. A "data sandwich" consisting of several previous-to-crop-year Landsat transparencies was associated with the phase 1 sample frame. These color infrared transparencies were used by an image analyst to produce rapid and inexpensive wheat proportion estimates (random variable X) for all sample segments.[2]

The resulting sample phase 1 proportion data were then used to minimize final crop estimate variance by stratifying the segment population into crop (in this case wheat or alternatively, cultivated land) density strata. After tabulating a list of phase 1 data, a small phase 2 sample could be allocated within the phase 1 strata

with either equal or variable probability. More accurate (Y variable) wheat proportion estimates were then made for each phase 2 segment selected by using multitemporal manual or machine-aided classification methods as illustrated by the lower layer in Figure 1.

Figure 1: TWO PHASE SAMPLE FRAME FOR WHEAT ACREAGE ESTIMATION



C. Determination of Phase 2 Sample Size

The second phase sample size, n, designed to minimize estimate variance for specified survey budget levels was determined via regression-based optimal sampling rate formulas. These are presented and discussed in Hay and Thomas (1976). Phase 2 sample size for each wheat density stratum is a function of the relative cost and correlation between phase 1 and phase 2 sample segment proportion measurements as well as the actual between sample segment variability represented by the variance of Y. The latter quantity was estimated by the variance obtained from phase 2 sample segment wheat proportion data. For purposes of sample size determination, correlation between phase 1 and phase 2 proportion estimates was assumed to be 0.8 on the basis of preliminary tests.

A detailed cost analysis was used (Thomas and Hay 1976) to determine between phase cost ratios as well as a total survey budget by CRD equivalent to the simulated year two LACIE system. The resulting phase 2 sample sizes were considered approximate since sample selection was defined to be with replacement, ppes[3], by stratum, while equal probability of selection was assumed in the sample size formulas.

D. Results and Discussion

CRD and County Wheat Proportion Estimates

Application of the two phase design to the Kansas Southwest CRD for 1974 produced a wheat acreage estimate for that CRD within 2.42 percent of the USDA SRS-based 1974 estimate using a lower CRD inventory budget than that for the assumed reference LACIE system. Table 1 presents the results using stratified regression (after O'Reagan and Boyd 1974 and Cochran 1963) and probability proportional to size (ppes) (Raj 1968) estimators for the Southwest CRD. Recall that both estimates are based on the same ppes draw of phase 2 sample segments. Consequently a comparison of the increased estimate precision available with ppes versus equal probability within stratum selection could not be made aside from that resulting from the formulas themselves.

TABLE 1:

RESULTING TWO PHASE KANSAS SOUTHWEST CRD WHEAT PROPORTION ESTIMATES (ACREAGE SOWN 1973-1974)

USDA-Based Estimate	Two Phase Regression			Two Phase PPES		
	Estimate	Std. Error	R.D. USDA VS. Two Phase	Estimate	Std. Error	R.D. USDA VS. Two Phase
27.63%	28.31%	1.68%	2.42%	28.30%	0.40%	2.42%

$$R.D. = \frac{\text{SAMPLE ESTIMATE} - \text{USDA ESTIMATE}}{\text{USDA Estimate}} \times 100$$

The regression estimator was used in a predictive manner to produce county estimates. County regression estimates for the Southwest CRD showed a greater range of departure from their corresponding USDA-based values than the CRD level estimates. This situation is expected when sample allocation

is optimized for the CRD as opposed to County level. Differences ranged from -6.66 percent in Stanton county to a low of 0.25 percent in Finney county to a 9.54 percent over-estimate in Ford county. The average difference, sign considered, was 0.18 percent (not statistically significant with the paired t-test). The average absolute difference, sign ignored, was 2.93 percent also found not to be statistically significant with the paired t-test.

The performance of the regression estimator in the Kansas Central CRD was below that obtained in the Southwest CRD. The regression estimate fell 3.50 percent absolute (or 10.94 percent relative) below the USDA-based proportion estimate. The regression estimate standard error was 1.67 times higher in the Central as opposed to the Southwest CRD.

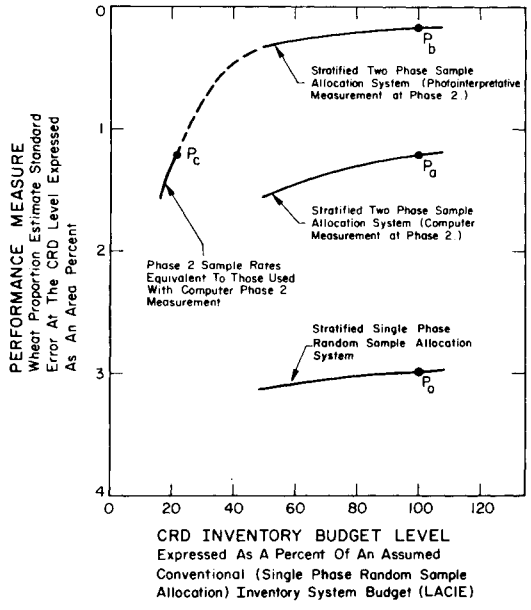
The less satisfactory performance in the Central Crop Report District resulted from a poor correlation between phase 1 and phase 2 proportion estimates. This low correlation was in turn traced to the fact that a significant amount of wheat had been plowed-down in some sample segments on the original phase 1 base date transparency. A test was run to determine if an earlier base date would produce correlations obtained (.8) in the Southwest CRD. This test was successful and suggested that inventory performance levels comparable to those achieved in Southwest should have been obtainable in the Central CRD.

An important shortcoming in the stratification scheme became apparent during the course of the study. Given the fixed survey budget, phase 2 sample sizes per stratum were too small for development of stable regression coefficient estimates. Instead a combined regression coefficient had to be estimated by pooling paired (X,Y) observations from all strata in a given CRD. Equal strata weighting of these paired observations gave the best characterization of the relationship thought to exist in the data. The question of bias in the strata-specific regression coefficients and hence bias in the strata wheat proportion estimates can thus be raised. Clearly sample size must be raised or number of strata reduced to minimize this problem. Given the fixed survey budget, reduction in number of strata from three to two would be recommended for future surveys of this type. Stratum-coefficient estimates, or stable, combined stratum specific size-weighted estimates of regression coefficients should then be available. As the number of years of available (X,Y) data increase, additional strata with stable regression coefficients can be added to increase the precision of the final Crop Reporting District estimates.

Cost-Effectiveness Comparison

A cost-effectiveness analytical framework was used to compare the relative precision and cost performance of (1) the reference LACIE sampling system with stratification based on historical agricultural wheat area statistics, (2) the two phase sample procedure with machine-aided wheat classification at the second phase, and (3) the two phase sample procedure with multitemporal manual processing at the second phase. Figure 2 illustrates the results of this analysis.

Figure 2: COST-CAPABILITY COMPARISON OF LANDSAT INVENTORY SYSTEMS USING TWO PHASE VERSUS SINGLE PHASE SAMPLE ALLOCATION STRATEGIES



Cost ratio, correlation, and phase 2 variance data obtained for the Kansas Southwest CRD was used to construct Figure 2. The LACIE reference system was defined to be a stratified random sample with sample unit allocation to wheat density strata proportional to area. This reference system was defined to represent as closely as possible the LACIE second year procedure. Stratification on historical county wheat data was assumed to give a 4 to 5 times reduction in variance relative to unstratified random sampling. The total CRD survey budget determined earlier for the LACIE reference system was defined as the 100 percent inventory level.

Comparison of points P_o and P_a in Figure 2 indicates that the two phase sample with computer processing at phase 2 should give greater than a two fold increase in precision relative to the reference LACIE system. Alternatively, the same LACIE reference system standard error at point P_o should be obtainable with less than one half to one fifth the reference system cost by using the two phase sample approach. This cost relationship can be seen by projecting[4] the curve containing P_a to the level of P_o.

Similar comparison of P_o with P_b indicates a greater than 10 fold increase in precision relative to the LACIE reference system may be achievable with the two phase sample using manual wheat classification at phase 2.

Comparison of P_a and P_b shows a four fold increase in precision when two phase sampling with manual as opposed to machine-aided wheat classification is employed. A similar reduction in cost is indicated.

It should be emphasized that these results are limited to the Kansas data set examined and the particular sample design assumptions made. The author submits that the important information here is not the exact cost or precision

improvement values, but rather the relative performance relationship between the two phase and single phase (reference) sample system. Further, the county and CRD wheat proportion estimates presented in the previous section suggest that not only can foreign survey performance be improved by use of auxiliary information available in the full frame Landsat data, but that some domestic acreage estimation requirements may be rather inexpensively obtained by a two-phase Landsat sample.

III. THREE PHASE SAMPLE FOR IRRIGATED ACREAGE ESTIMATION

Landsat data, when linked via proper design with conventional ground and aircraft photography survey elements, can produce significant cost, bias, precision, and/or timeliness improvements in domestic agricultural survey systems. An example of such a survey design has been developed (Wall et al. 1977) for application by the California Department of Water Resources (DWR). The DWR is responsible for managing the state's water resources. Consequently up-to-date information is required concerning water use (demand) in order to plan for adequate water storage and delivery. In years of normal precipitation, approximately 85 percent of the total water used is consumed by agriculture. Thus a major DWR task is to inventory California's lands to determine the number of acres that are irrigated per year and the rate of water application. In the past, these data requirements have been met by land use surveys involving complete enumeration of agricultural areas on vertical 35mm aircraft photography, supplemented with field inspections. While this survey provides information for many DWR planning functions, it suffers from the standpoint of providing current estimates of agricultural water use (cost prohibits resurvey of a given county more than once every seven years on the average) and from the standpoint of detecting all crops grown in a given survey year (due to single date of aerial photography) that potentially require irrigation.

A. Inventory Objectives

In order to provide for more accurate and timely agricultural water use estimates, a survey design incorporating Landsat as well as conventional data has been developed[5] and tested to achieve the following objectives:

(1) Primary: provide acreage of land, by county, that is irrigated at least once during the calendar year; the technique should enable DWR to operationally inventory the entire state in one year at four year intervals. Further, estimates must be available for publication within six months following the calendar year of the inventory. Finally, the precision requirement for the irrigated estimate should be $\pm$ 3 percent sampling error at the 99 percent level of confidence as reported for the entire state.

(2) Secondary: provide estimates of acreage that supports more than one crop per year (multicropping); provide estimates of acreage of different crop types to enable the computation of different rates of water consumption; these estimates to be provided at the county and state levels.

B. Sample Design Specification

The initial design was developed and tested on ten counties in California covering the range of agricultural diversity in the state. Seven of the ten counties were located in the Central Valley of California. Others were located in mountain and coastal areas.

A three phase sample design was selected to take maximum advantage of the auxiliary variable data (spectral reflectance, field pattern) available on Landsat and aerial photography relating to irrigated acreage. Multiple dates of Landsat full frame color IR imagery served to provide relatively inexpensive, county-wide estimates of irrigated proportion and proportion of area multicropped. Vertical color aerial photography provided a cost-effective means to correct the Landsat estimates for bias. Finally, measurements made on a small sample of ground units were used in turn to calibrate the aerial photography estimates and provide the most accurate information on crop types present.

A rectangular sample frame of one by five mile sample units was defined to cover each county. The frame was aligned with a north-south/east-west retangular survey grid, the long axis of the sample units parallel to the east-west direction. Since no prior irrigated acreage variance versus sample unit dimension data was available, sample unit size and shape was chosen based on practical considerations. These considerations dealt with ease of data acquisition and measurement at each sample stage. In particular, the one mile strip width was determined as the area considered efficient for interpreting irrigated acreage data from 1:62,500 scale color 35mm photography; and the five mile length was easily located and flown over several dates. The one by five mile size was also considered workable for digitizing the location of irrigated fields on the Landsat data and obtaining ground information in the field.

C. Sample Size Determination

In order to determine phase 1, 2, and 3 sample sizes by county (stratum) that would be expected to support the statewide $\pm$ 3%, 99% level of confidence irrigated acreage precision goal, a preliminary population model was constructed. Sample size (number of sample units) allocations were based on previously published estimates of of proportion of area irrigated by county, approximate between phase cost ratios and a non-linear programming algorithm which minimizes cost, subject to constraints on variance. California Experiment Station Bulletin 847 and 1974 County Agricultural Commission reports provided most of the numerical data on irrigated acreage. Samples were allocated with equal probability at each sample phase within each county. Sample units eligible for selection were confined to those selected for measurement at the previous phase.

D. Specification of Proportion Estimators

A three phase regression estimation system

was chosen to provide irrigated acreage (and crop type) proportion estimates. This model was thought to represent the between phase proportion relationships most correctly. The mean and variance estimators followed the treatment given by Tikkiwal (1955 and 1967). Basically, these estimators were iterative such that the phase 3 (ground) estimator used the phase 2 (photo) estimator which in turn used the phase 1 (Landsat) estimator. The parameters requiring estimation were the proportions of irrigated land within the sampling region of each county using all three phases together. In order to estimate these parameters it was necessary to obtain separate estimates for (1) irrigated proportion determined from phase 1, (Y*); (2) irrigated proportion determined from phase 1 and 2, (Y'), and (3) irrigated proportion determined from all three phases, (Y). Their corresponding estimators were denoted $\hat{Y}^*$, $\hat{Y}'$ and $\hat{Y}$. The last of these was the end result; $\hat{Y}^*$ and $\hat{Y}'$ were only used as needed to obtain $\hat{Y}$.

In that sample unit size varied slightly due to varying scales (nominally 1:154,000 for Landsat and 1:62,500 for the aerial photography) and due to varying length of units cut by county boundaries, the sample units were viewed as area clusters of unequal size. Consequently weighted means were used in the estimators as opposed to unweighted means that would increase the variance of the estimates.

E. Results

Multidate interpretation of both Landsat transparency data and corresponding color photography mosaics of individual sample units was used to identify parcels of land irrigated. These were then digitized by a graph pen device to compute measured proportions for each sample unit included in the sample. It was found that the measurement of the entire population of Landsat sample units was easier than locating individual units separately and performing measurement. Consequently, phase 1 irrigated proportion estimates were produced on a sample unit basis for only sample units also requiring phase 2 measurements. The variance of the phase 1 estimate was thus based on these units. All other phase 1 units were lumped into a large "pseudo" sample unit on which the proportion irrigated was determined by digitization. The pseudo sample unit was then incorporated in the proportion estimator by weighting its total irrigated area according to its actual size.

Of the total land area sampled, approximately 3 million acres or 21.6 percent was estimated to be irrigated. This value compared favorably with DWR's best available information.[6] In one county, Stanislaus, a direct, year-specific comparison was possible between the three phase sample and the conventional DWR inventory procedure. The three phase acreage estimate was within one-half of one percent of the corresponding DWR figure.

The relative sampling error[7] for the ten county area was 2.73 percent, or 7.04 percent expressed at the 99 percent level of confidence. Since the population sampled in this study represented less than half the agricultural land in Californtia, a sample of the larger area would be expected to produce precision performance approaching $\pm$ 3 percent at the 99 percent level requested by DWR for state-wide reporting.

Throughput rates for sample allocation, measurement, and processing in the ten county study indicated that the 18 month time constraint from inception of state-wide inventory to publication of results appeared feasible. This throughput performance represents the most important improvement in inventory performance over the conventional system as seen by DWR.

IV. SUMMARY AND CONCLUSIONS

The sampling and measurement methods described in this paper can be of practical utility in many domestic agricultural inventory situations. The Landsat two-phase system, while most useful in foreign survey problems where up-to-date ground or photo data may be unavailable, was demonstrated to achieve high precision in a Kansas environment. Estimates for area sown to wheat were within a few percent and not significantly different from corresponding USDA estimates at the county and Crop Reporting District levels. Landsat data was also shown to produce rapid, accurate irrigated acreage estimates of high precision when linked with conventional aerial photography and ground information in a three phase sample system.

In both designs discussed above, the Landsat imagery provided an inexpensive source of auxiliary information correlated with a crop parameter of interest. Significant reduction in estimate variance, survey time, or cost resulted. The repetitive coverage characteristic of Landsat combined with extensive spatial information relating to ground cover type, may be even more important in addressing multipurpose survey problems. These can arise when information is desired simultaneously on one or more parameters (e.g.acreage, yield, water consumption) for several crops. On-going work, for example, suggests that significant growth-related auxillary information exists in the Landsat digital data in addition to that concerning crop-specific areal extent. Moreover, land use and land form information, readily available from interpretation of multidate Landsat transparencies, may prove particularly effective in developing multivariate stratification schemes designed to simultaneously control variance on several parameters.

The author wishes to stress that Landsat in and of itself is not a sampling panacea. However, when included in designs incorporating calibrating subsample information, Landsat can provide important opportunities for variance or cost reduction as well as for increased survey speed.

ACKNOWLEDGEMENTS

The author would like to acknowledge Claire M. Hay for her assistance in development of the Landsat image interpretation procedures used in the two-phase design for wheat proportion estimation. Special thanks are also due to Sharon L. Wall for permission to publish results relating to the Irrigated Lands Study and for her assistance during preparation of the

manuscript. Finally, S.J. Titus deserves special mention as the statistician primarily responsible for developing the design and sample allocation procedures used in the Irrigated Lands study.

LITERATURE CITED

Cochran, W.G. 1963. Sampling Techniques (Second Edition). John Wiley & Sons Inc., New York. 413 pp.

MacDonald, R.B., R.B. Erb, and F.G. Hall. 1975. The use of Landsat data in a Large Area Crop Inventory Experiment (LACIE). In: Proceedings of the 1975 Symposium on Machine Processing of Remotely Sensed Data. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. 23 pp.

NASA-Johnson Space Center. 1975. LACIE operations plan Phase III; Level III baseline. NASA Johnson Space Center, Houston, LACIE-C00606, JSC-09855. September.

O'Reagan, W.G. and R.W. Boyd. 1974. Regression sampling: some results for resource managers and researchers. USDA Forest Service Research Note PSW-286. U.S. Forest Service Pacific Southwest Forest and Range Experiment Station, Berkeley. May.

Raj, Des. 1968. Sampling theory. McGraw-Hill Book Company, San Francisco, 302 pp.

Thomas, R.W. and C.M. Hay. 1976. Variable probability sampling for acreage estimation. In: Application of Photointerpretative Techniques to Wheat Identification, Signature Extension, and Sampling Strategy. NAS 9-14565, Principal Investigator: R.N. Colwell, Space Sciences Laboratory, Series 17, Issue 33, University of California, Berkeley. May.

Tikkiwall, B.D. 1955. Multiphase sampling on succesive occasions. Ph.D. Thesis, North Carolina State College, Raleigh. 90 pp.

Tikkiwall, B.D. 1967. Theory of multiphase sampling from a finite or infinite population on successive occasions 1,2. In: Review of the International Statistical Institute, Vol. 35:3.

U.S. Department of Agriculture. 1974. Agricultural statistics 1974. U.S.D.A. Statistical Reporting Service, Washington, D.C.

Wall, S.L., D.K. Noren, J.M. Sharp, and S.J. Titus. 1977. An inventory of irrigated lands for selected counties within the State of California based on Landsat and supporting aircraft data. NASA Contract No. NAS 5-20969, Principal Investigator: R.N. Colwell. Space Sciences Laboratory, Series 18, Issue 50, University of California, Berkeley, January. 52 pp.

NOTES

1. Work supported by NASA Contract NAS 9-14565.

2. Since all phase 1 units were sampled, the sample design applied in this example becomes regression sampling. However, the more general technique developed in this study can be applied when sampling less than the population size at phase 1.

3. Probability proportional to estimated size selection used to evaluate differences in mean and variance estimates based on regression versus ppes formulas.

4. Using the shape relationship of the curve containing P_b. The shape relationships are approximately equivalent.

5. Work supported by NASA (Contract No. NAS 5-20969 and NSG 2207) and performed by the Remote Sensing Research Program, University of California at Berkeley in conjunction with the DWR.

6. Based on Census of Agriculture and State Crop Report data.

7. Assuming the acreage measurements were without error.