N78-14538

TWO PHASE SAMPLING FOR WHEAT ACREAGE ESTIMATION[1]

Randall W. Thomas

Claire M. Hay

Remote Sensing Research Program
260 Space Sciences Laboratory
University of California
Berkeley, California
(415)642-2351

ABSTRACT

A two phase Landsat-based sample allocation
and wheat proportion estimation method was
developed. This technique employs manual,
Landsat full frame-based wheat or cultivated
land proportion estimates from a large number
of segments comprising a first sample phase
to optimally allocate a smaller phase two
sample of computer or manually processed seg-
ments. Application to the Kansas Southwest
CRD for 1974 produced a wheat acreage esti-
mate for that CRD within 2.42 percent of the
USDA SRS-based estimate using a lower CRD
inventory budget than for a simulated
reference LACIE system. Factor of 2 or
greater cost or precision improvements rela-
tive to the reference system were obtained.

1. INTRODUCTION

One of the most important aspects controlling the success of any inventory
system is the sampling agregation 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 inventory sample design options and improved estimate perfor-
mance possibilities. While progress has been made in this regard through the
Large Area Crop Inventory Experiment (LACIE) and through smaller projects,
current inventory performance capability falls significantly short of its
present potential.

2. STUDY OBJECTIVE

In order to provide a relatively simple demonstration of crop inventory
performance possibilities presently unexploited, a two phase Landsat-based
sample allocation and wheat proportion estimatation method was developed in
this study. A simulated second year LACIE inventory system was used as a
base for performance (precision, cost) comparison.

The two phase technique employs manual, Landsat full frame-based 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 are then linked by regression or probability proportional to estimated size (ppes) estimators to provide wheat proportion estimates and standard errors by reporting unit.

3. SAMPLING AND MEASUREMENT METHODS

3.1 Information Requirements and Performance Goals

The information target for the inventory was defined to be wheat acreage sown (1973-74) expressed as a proportion of total land area for 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 precision control was set to achieve a wheat acreage estimate within five percent of the corresponding USDA estimate, 95 times out of 100 at the Crop Reporting District level. Budget and inventory throughput rate constraints were selected to be similar to those of the reference LACIE year two system.

Two Kansas CRD's were chosen to demonstrate the Landsat two phase sample technique in the winter wheat region. The first of these, the Kansas, Southwest CRD (11,865 mi^2) occupies a predominantly semi-arid to sub-humid environment. The dominant small grain-related crop rotation in this water-limited area is summer fallow, wheat and sorghum. To provide a contrasting wheat distribution and appearance situation, the moister and more humid Central Crop Reporting District (8,968 mi^2) was selected as the second Kansas inventory test area. Here moisture is no longer the dominant limiting agent and double cropping sequences often result. Field size is generally smaller, wheat density lower, and noncultivated range-grassland interfringes more extensively with cultivated areas within the Central CRD.

Inventory data was 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 on-a-side segments, and ancillary crop calendar and cropping practice information. A more tailor-made domestic inventory system, not considered here, might also include aircraft and ground data for estimate and measurement calibration purposes.

3.2 Sample Design Specification

A stratified double sampling (i.e., two phase) design was selected to demonstrate the capability of remote sensing-aided systems to meet wheat proportion information requirements within the CRD performance constraints just described.

This design takes advantage of the relationship between a more expensive to measure variable Y (e.g., computer-based wheat proportion) and a corresponding less expensive to measure variable X (e.g., a rapid analyst estimate of sample segment wheat proportion). A relatively large first phase sample of observations on X may be used to efficiently allocate a much smaller sample of observations on Y. Similarly, the small sample of information on Y can be used to calibrate (to Y accuracy standards) the area-wide information on X. If the correlation between X and Y is sufficiently large, significant reductions in estimate (e.g., wheat proportion) variance and second phase (e.g., computer segment) sample size can result when compared with single phase sampling on Y alone.

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 (30 mi^2) 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 transparancies were used by an image analyst to produce rapid and inexpensive wheat proportion estimates (variable X) for all sample segments.*

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. Thus, after tabulating a list of phase 1 data, a small phase 2 sample can be allocated within the phase 1 strata with either equal or variable probability. Stratafied probability proportional to estimated size (of phase 1 wheat proportion) allocation was used to select sample phase 2 segments in this study. 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.

3.3 Determination of Optimal Phase 2 Sample Size

The optimal 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 detail in Thomas and Hay.[5] Optimal 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 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.

Based on a detailed cost analysis[5] it was determined that the cost ratio for unitemporal machine processing at phase 2 to analyst estimation at phase 1 was 170:1. If multidate manual classification of a small point sample was used instead at phase 2 then the cost ratio became 17:1.

A simulated LACIE system sample size was determined in order to define the total survey budgets available for the Kansas Southwest and Central CRD's. Crop year 1972-73 USDA statistics were used to give the proportion of wheat average sown, harvested, and produced in each CRD relative to the U.S. total.[2] Under an early LACIE assumption that 636 sample segments would be allocated to U.S. wheat regions, the total expected number of sample segments allocated to both CRD's was determined for each allocation factor.[6] Cost per unitemporarily processed computer segment was then multiplied times the sample size required under the acreage sown allocation assumption to give total available CRD survey budget. This budget represented that theoretically available to the reference LACIE system.

Given the crop reporting district budgets, phase 1 to 2 correlations and cost** ratios, and estimated phase 2 variances, optimal phase 2 sample sizes for the two phase sample with regression estimation were calculated. Sample selection was defined to be with replacement, ppes, by stratum.

*Since all phase 1 units are sampled, the sample design applied here 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.

**In order to be conservative relative to two phase sample system performance, a phase 2 to phase 1 ratio of 150:1 was assumed.

3.4   Specification of Measurement Procedures

Wheat or cultivated land percent estimates were obtained for phase 1 sample units by the first of two image analysis procedures developed in this study. The first image interpretation procedure allowed quick (approximately three minutes per segment including rest time) proportion estimates to be made from a base date Landsat full frame transparency. The base date was selected from a recent crop year date that gave maximum contrast between wheat versus other crop types. In the two Kansas CRD's examined, this base date occurred at or shortly after harvest. At this time, wheat fields appeared very white relative to all other cover categories.

When confusion situations were identified by reference to ancillary data concerning crop calendar and cropping practices as well as multidate interpretation of Landsat imagery, an additional one and rarely two dates of color infrared full frame data was referenced by the image analyst. Grain sorghum fields, not easily separable from wheat on the base date, represented an example of such a situation. Land use/soils association stratification on Landsat full frame data was found to provide a convenient means of coding circumstances in which wheat versus other confusion might occur.

A second image interpretation procedure served to provide phase 2 wheat proportion estimates. This technique was chosen to represent the best Landsat-based wheat proportion measurement capability available for phase 2 sample segments. Earlier tests had shown that this multitemporal image interpretation approach resulted in more accurate proportions than did corresponding unitemporal machine-aided classification. Ideally multitemporal machine processing should give results at least comparative to the manual method, and for this reason the machine cost figures were used for phase 2 sample size determination.[*]

The phase 2 wheat measurement procedure was to employ a systematic sample of 48 points over enlargements of phase 2 sample segments obtained from full frame transparencies. Enlargements were to CX120 "lantern slide" size representing a five to six times scale increase relative to the original 1:1,000,000 scale. Dates chosen for inclusion in this analysis included a representative having the least cloud cover, least noise, and most contrast between cover classes.

Wheat versus other classification were recorded on an acetate sheet covering a record photo for the given sample segment. In order to maximize wheat identification accuracy (correctly identifying wheat as wheat) and minimize commission error (classifying a sample point as wheat when it was not), other major non-wheat cover types and confusion crops were identified when possible. This additional identification task was designed to ensure a conscientious consideration of wheat alternatives by the photointerpreter.

3.5   Specification of Proportion Estimators

Two estimators were considered: stratified regression and stratified probability proportional to estimated size (ppes).[1,3,4] Generally the linear regression estimator is used when the relationship between X (phase 1 proportion) and Y (phase 2 proportion) can potentially move far from the origin and when the variance of Y about the regression line (σ^2_ϵ) remains approximately constant over the range of X. In this situation it is known as the best

_____

[*]Original unitemporal machine processing costs were retained as opposed to substituting higher multitemporal costs. Again this assumption is conservative relative to two phase sample system performance.

linearly unbiased estimator (BLUE). When the relationship between X and Y is thought to pass close to the origin and σ^2_ϵ increases proportionally to X then ppes estimators are termed BLUE. This latter situation may occur especially in areas with high wheat density variability. In addition, ppes allocation may be used to drive second phase sample unit selection towards a greater proportion of "higher value" areas and still maintain unbiased estimation. For example, it may be desired to force computer segment selection to units tending to have higher wheat density or higher wheat variety spectral class mixture representation in order to maximize signature extension success.

4. SYSTEM EVALUATION: COST-EFFECTIVENESS ANALYSIS

A portion of the analysis involved a precision versus cost performance comparison between the double sampling system described in this study and the reference LACIE sampling system. This analysis was done to demonstrate the relative amount of improvement to be expected with inclusion of the full frame Landsat data in the system. The form of cost-effectiveness analysis used is known as a "system comparison study". It helps a decision-maker answer questions about how to achieve a given set of objectives at the least cost, or conversely, how to obtain the most effectiveness from a given set of resources.

5. RESULTS AND DISCUSSION

5.1 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 for the assumed referenced LACIE system. Table 1 presents the results for regression and probability proportioned to size (ppes) estimation 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 random within stratum selection could not be made aside from that resulting from the formulas themselves.

The regression estimator was used in a predictive manner to produce county estimates (see Table 2). County regression estimates for the Southwest CRD show 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 range 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 both the regression and ppes estimators in the Kansas Central CRD was below that obtained in the Southwest CRD. The regression estimate fell 3.50 percent absolute below the USDA-based proportion estimate while the ppes estimate was found to be 6.09 percent low. These same departure percentages represent 10.94 and 19.04 percent of the USDA-based estimate, respectively. Resulting estimate standard errors were 1.67 times higher for regression and 1.53 times higher for ppes 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.

913

Use of correct base date transparencies for phase 1 wheat estimation resulted in phase 1 to phase 2 correlations of .82 and .79 for the Southwest and Central CRD's respectively. These correlations were achieved when strata were pooled. Within stratum correlations varied from .54 to .83. The generally lower stratum-specific correlations suggest that some strata should be grouped or phase 2 sample sizes increased somewhat so as to allow a more accurate representation of the stratum phase 1 to phase 2 relation.

Interestingly, phase 1 cultivated land proportion estimates gave a phase 1 to phase 2 (wheat) correlation of .89 in the Southwest CRD. The corresponding value for the Central District, however, dropped to .68. Dominance of the wheat crop in Southwest CRD may explain the former result, while the more complex multicrop patterns in the Central may be responsible for the latter result. In any event, the importance of inexpensive phase 1 cultivated land estimates, easily obtained in most agricultural situations, should not be overlooked as an inventory performance improvement option.

5.2 Cost-Effectiveness Comparison

The cost-effectiveness 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 multi-temporal manual processing at the second phase. Figure 2 illustrates the results of this analysis.

Cost ratio, correlation, and phase 2 variance data obtained for the Kansas Southwest CRD was used to construct the Figure. The LACIE reference system was defined to be a stratified random sample with phase 2 sample 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_0 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_0 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* the curve containing P_a to the level of P_0.

Similar comparison of P_0 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 authors submit 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.

6. CONCLUSIONS

The sampling and measurement methods described in this study are of practical utility in many agriculture inventory situations. Optimum allocation of

---

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

914

sample units to control precision of acreage estimation is a common sampling concern. The spatial information provided on the full-frame Landsat imagery can, as demonstrated in this study, be used to cost-effectively stratify a population of segments so as to minimize final estimate variance. For the Kansas test areas examined in this study, it appears that remote sensing-aided inventory systems can perform with high precision and accuracy at the Crop Reporting District level.

7. LITERATURE CITED

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

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

3. O'Regan, W.G. and R.W. Boyd. 1974. Regression sampling: some results for 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. 7pp.

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

5. 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.

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

TABLE 1:   RESULTING TWO PHASE KANSAS SOUTHWEST CRD WHEAT PROPORTION ESTIMATES

(ACREAGE SOWN 1973 - 1974)

USDA-Based	Two Phase Regression			Two Phase ppes		
Estimate	Estimate	Std. Error	R.D. USDA vs. Two Phase	Estimate	Std. Error	R.E. USDA vs. Two Phase
27.63	28.31%	1.68%	2.42%	28.30%	0.40%	2.42%

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

TABLE 2.   COUNTY TWO PHASE RESULTS FOR THE KANSAS SOUTHWEST CRD
(1973 - 1974)

COUNTY	WHEAT PROPORTION ESTIMATE DIFFERENCE (Two Phase - USDA Based)
Hamilton	-1.95%
Kearny	-5.95%
Finney	0.25%
Hodgeman	5.33%
Stanton	-6.66%
Grant	0.27%
Haskell	-0.91%
Gray	2.08%
Ford	9.54%
Morton	-0.74%
Stevens	-3.48%
Seward	-0.19%
Meade	1.74%
Clark	2.56%

Ave. Difference sign considered = 0.18%

Ave. Difference sign ignored   = 2.93%

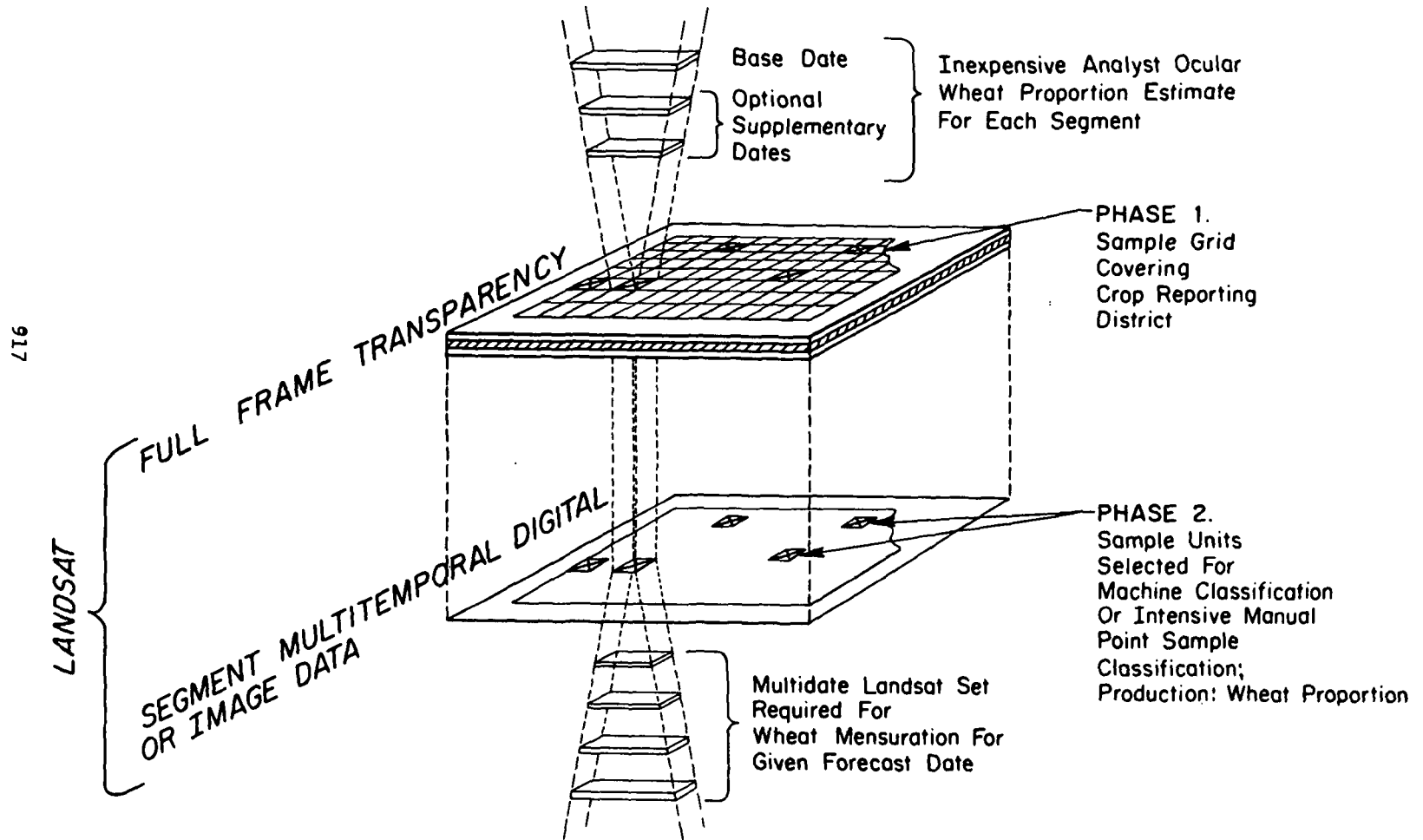Figure 1: TWO PHASE SAMPLE FRAME FOR WHEAT ACREAGE ESTIMATION

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

PERFORMANCE MEASURE
Wheat Proportion Estimate Standard Error At The CRD Level Expressed As An Area Percent

Stratified Two Phase Sample Allocation System (Photointerpretative Measurement at Phase 2.)

Stratified Two Phase Sample Allocation System (Computer Measurement at Phase 2.)

Phase 2 Sample Rates Equivalent To Those Used With Computer Phase 2 Measurement

Stratified Single Phase Random Sample Allocation System

CRD INVENTORY BUDGET LEVEL
Expressed As A Percent Of An Assumed Conventional (Single Phase Random Sample Allocation) Inventory System Budget (LACIE)