T2

Lockheed Electronics Company, Inc.

A SUBSIDIARY OF LOCKHEED CORPORATION

1830 NASA Road 1, Houston, Texas 77058

Tel. 713-333-5411

JSC-14547

7.9 - 1 0 2 1 6

CR 160182

Ref: 642-6928
Contract NAS 9-15800
Job Order 73-783-02

TECHNICAL MEMORANDUM

A SIMULATION STUDY OF LARGE AREA CROP INVENTORY EXPERIMENT (LACIE) TECHNOLOGY

By

L. Ziegler and J. Potter

(E79-10216) A SIMULATION STUDY OF LARGE AREA CROP INVENTORY EXPERIMENT (LACIE) TECHNOLOGY (Lockheed Electronics Co.) 31 p HC A03/MF A01 CSCL 02C

N79-26443

Unclas

G3/43 00216

Approved By: Elmer M. Hsu

E. M. Hsu, Supervisor
Accuracy Assessment Section

May 1979

LEC-12180

ACKNOWLEDGMENTS

The studies reported here were carried out using a computer simulation program called the LACIE Performance Predictor. This program was developed by TRW, Inc., with the technical guidance of Drs. F. G. Hall and D. E. Pitts of the National Aeronautics and Space Administration and R. Chhikara of Lockheed Electronics Company, Inc. Also Dr. R. Chhikara provided valuable comments regarding the interpretation of the results described here.

CONTENTS

TABLE

FIGURES

1. INTRODUCTION

In order to assess the performance of the Large Area Crop Inventory Experiment (LACIE) system, several years of LACIE results are required. Two types of problems are thus presented: (1) It will be several years before these data are available; and (2) the LACIE system is evolving from year to year, so the results obtained over several years are actually representative of several different "LACIE systems."

The LACIE Performance Predictor (LPP) is a set of computer programs which simulate the performance of a given LACIE system (i.e., the system used in a given year or phase of LACIE). The LPP can be used to evaluate the system by simulating the input and thereby simulating the results that would be obtained in several years of operation of that system, and can also be used to study the effect of various error sources on the final LACIE estimates.

This study describes several runs that were made with the LPP, each of which simulated 15 years of LACIE Phase II operations. The runs correspond to different sets of assumptions about the basic error sources in the LACIE system.

2. THE LACIE PERFORMANCE PREDICTOR

The LPP simulates these major elements of the LACIE system:

Segment acquisition

Estimation of wheat proportion within the segment

Yield estimation

Area and production aggregation

The procedures used to perform these simulation tasks are described in following sections.

2.1 SEGMENT ACQUISITION

The first major task the LPP performs is to simulate the acquisition of sample segment data by the Landsat. The segments are located at the positions determined by the LACIE Phase II allocation. The LPP calculates the orbit of the Landsat and prepares a file which contains the dates on which data was acquired by Landsat for each segment.

Subsequently, an allowance is made for cloud cover as acquisitions with cloud cover above a given threshold are not used by the LACIE system. Historical cloud cover data from weather observations is used to simulate the cloud cover on each acquisition of each segment. This is done by randomly choosing a cloud cover in such a way that the probability of a given cloud cover percentage being selected is equal to the frequency it was observed in the past. If the simulated cloud cover is greater than the threshold value, the acquisition is rejected.

2.2 SIMULATION OF COUNTY PROPORTIONS

It is assumed that the county proportion P_i for the *ith* county is distributed according to a beta distribution; i.e.,

$$P_i \sim \beta(\mu_i, \xi_i)$$

2

where μ_i is the mean and ξ_i is the standard deviation of the distribution. The means are taken to be the 1975 proportions as determined by the Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA). For the *ith* county, the proportion is denoted by $P_{75,i}$. The standard deviations are calculated for each county by taking the standard deviation of the historical proportions for that county for the years 1965 through 1974. However, the LPP does not accept $P_{75,i}$ and ξ_i directly, but instead requires the following inputs:

$$CV_{1,i} = \xi_i / \overline{P}_{H,i}$$

$$\delta PW_i = (\mu_i - \overline{P}_{H,i}) / \overline{P}_{H,i}$$

where

$\overline{P}_{H,i}$ = average proportion for the *ith* county for the years 1965 through 1974

$\mu_i = P_{75,i}$

Both $CV_{1,i}$ and δPW_i are manually calculated from ξ_i, $\overline{P}_{H,i}$, and $P_{75,i}$ and used as inputs to the LPP which calculates μ_i and ξ_i for each county. A simulated "true" county proportion P_i is then calculated by the LPP for each county i by choosing a random number generated for the distribution $\beta(\mu_i, \xi_i)$.

2.3 SIMULATION OF "TRUE" PROPORTIONS FOR SEGMENTS

For segments in county i, it is assumed that the segment proportions X_i are distributed according to a beta distribution $\beta(P_i, \Theta_i)$, where P_i is the "true" county proportion described above and Θ_i^2 is the within-county variance of segment wheat proportions. In order to determine Θ_i^2, previous studies are used which provide an estimate of the within-county variance of small-grain proportions. It is assumed that this estimate is equal to Θ_i^2. The studies are based on LACIE analysts' interpretation of Landsat imagery of all the counties in the U.S. Great Plains (USGP). Each county was partitioned into segments, and estimates were made of the total agriculture (ag) proportion of each segment in each county. These estimates were used to calculate the average ag proportion $\overline{X}_{ag}$ and the ag variance $\Theta_{ag,i}^2$ for all of the counties in the USGP.

3

Another task consisted of doing the same type of analysis to produce estimates of average small-grain proportion $\overline{X}_{sg,i}$ and small-grain variance $\Theta^2_{sg,i}$. However, the results for small grains are limited to a subset of approximately 45 counties. A simple regression model based upon $\overline{X}_{ag,i}$, $\Theta^2_{ag,i}$, $\overline{X}_{sg,i}$, and $\Theta^2_{sg,i}$ in the subset is used to obtain values of $\Theta^2_{sg,i}$ for all of the counties. As stated previously, it is assumed that $\Theta_i = \Theta_{sg,i}$.

The LPP does not accept as input the variance Θ_i^2 itself but the coefficient of variation (CV) which is given by

$$CV_{2,i} = \Theta_i/P_i$$

These could be calculated if the P_i were known, but unfortunately the P_i, which are calculated by the LPP, are not available in advance to compute $CV_{2,i}$. Therefore, the following procedure is used to determine the $CV_{2,i}$. First, the following quantities are calculated:

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_{75,i}$$

$$\overline{\Theta^2} = \frac{1}{n}\sum_{i=1}^{n} \Theta_i^2$$

$$CV_2 = \frac{\sqrt{\overline{\Theta^2}}}{\overline{P}}$$

where n is the number of counties in a given state. The value obtained for CV_2 is then input to the LPP for each of the nine states. For the counties in each state, the LPP calculates the quantities:

$$\Theta_i' = (CV_2)P_i$$

These are taken as the estimates of the within-county variance of wheat proportion to be used in the model.

A "true" wheat proportion X_{ij} is then simulated for each segment j in the *ith* county by choosing a random number generated for the distribution $\beta(P_i, \Theta_i')$.

2.4 SIMULATION OF CLASSIFICATION AND MENSURATION SUBSYSTEM (CAMS) ESTIMATE

It is assumed that the CAMS estimates of wheat proportions $\hat{X}_{ij}$ for the *jth* segment in the *ith* county are distributed according to the beta distribution $\beta(X_{ij} + B_{ij}, \sigma_{ij})$, where B_{ij} is the CAMS bias for the *jth* segment in the *ith* county and σ_{ij} is the variance in the CAMS errors (i.e., in $\hat{X}_{ij} - X_{ij}$) for these segments. The values B_{ij} and σ_{ij} were estimated using blind site data and CAMS estimates as follows:

$$B_{ij} = \left(\frac{B}{\bar{X}}\right) X_{ij}$$

and

$$\sigma_{ij} = \left(\frac{\sigma}{\bar{X}}\right) X_{ij}$$

where

$$B = \frac{1}{N_B} \sum_{k=1}^{N_B} \left(\hat{X}_k - X_k\right)$$

$$\sigma^2 = \frac{1}{N_B - 1} \sum_{k=1}^{N_B} \left(\hat{X}_k - X_k - B\right)^2$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}$$

where

N_B is the number of blind sites in the state

N is the number of segments in the state

n is the number of counties in the state

n_i is the number of segments in the *ith* county

X_k is the measured ground-truth wheat proportion (the "true" wheat proportion)
$\hat{X}_k$ is the CAMS estimate of the wheat proportion for the *kth* blind site in the state.

The quantities $B/\bar{X}$ and $\sigma/\bar{X}$ are input to the LPP, which performs a multiplication by X_{ij} to obtain B_{ij} and σ_{ij}. Subsequently, it generates the $\hat{X}_{ij}$ by choosing a random number generated for the distribution $\beta(X_{ij} + B_{ij}, \sigma_{ij})$.

Different values of B and σ^2 are computed for each of the four biowindows, and the appropriate values are used to simulate a value of $\hat{X}_{ij}$ corresponding to an acquisition in a given biowindow. In principle, one could also calculate values of B and σ^2 corresponding to various combinations of biowindows. However, this was not done for the runs described in this paper as not enough blind sites had the required combinations of acquisitions. All of the estimates of $\hat{X}_{ij}$ described here correspond to a single acquisition.

2.5 SIMULATION OF YIELD ESTIMATES

Yield estimates are simulated by the LPP for each Crop Reporting District (CRD) in the USGP. The "true" yield Y_i for the *ith* CRD is taken to be the 1975 yield estimates by the USDA/SRS. The final yield estimates corresponding to biowindow 4 for the *ith* CRD, Y_i, are assumed to be distributed normally; i.e.,

$$\hat{Y}_i \sim n(Y_i, \phi_{4i})$$

where the standard deviation ϕ_{4i} is determined from the results of the 10-year test made by the Center for Climatic and Environmental Assessment (CCEA) of the yield model used in LACIE.

Standard deviations of yield estimates for earlier biowindows have to reflect the increasingly unreliable nature of CCEA yield estimates made at earlier dates in the growing season. To do this, each standard deviation input for a CRD for a particular biowindow is assumed to be 4 percent larger than the

6

standard deviation input for the biowindow that followed in the season. Working backwards from harvest,

$$\phi_{3,i} = 1.04\phi_{4,i}$$

$$\phi_{2,i} = 1.04\phi_{3,i}$$

$$\phi_{1,i} = 1.04\phi_{2,i}$$

2.6 SIMULATION OF THE LACIE AGGREGATION PROCEDURE

The LPP simulates the LACIE aggregation procedure to produce estimates for each year of the harvested wheat area, the wheat yield, and the wheat production by CRD, state, region, and country. Estimates of the CV's (standard deviation divided by the "true" value) are also produced at these levels in a fashion identical to the way they are produced by the actual LACIE aggregation procedures.

Any LPP aggregation corresponds to a particular date, and the CAMS estimate $\hat{X}_{ij}$ is based on the latest acquisition prior to that date. Also, the time taken by the actual LACIE system to process an acquisition to the point where it is ready for aggregation is not considered in the LPP.

Two kinds of aggregations are performed corresponding to the kinds of error included in the aggregation estimates. They are:

Sampling error only, performed by aggregating the simulated "true" segment proportions X_{ij} with yield $\hat{Y}_i$ set equal to Y_i.

Sampling, classification, and yield errors, performed by aggregating the CAMS estimates $\hat{X}_{ij}$ to the CRD and multiplying by the yield estimates, $\hat{Y}_i$, for the CRD to obtain a production estimate for the CRD. The acreage and production estimates are then summed to obtain estimates for higher levels.

3. DESCRIPTION OF RUNS

In the evaluations described here, the outer loop shown in figure 3-1 is run four times, once corresponding to no clouds and three times using the regular cloud cover data, thereby producing four different sets of acquisition dates. By design, they each produce the same set of values for the "true" county proportion P_i. This is achieved by using the same random number seed for the generation of the P_i in all four runs. Each run of the outer loop produces a data tape containing the results of that run which is an input to the inner loop (fig. 3-1) which is a separate set of programs. In all, two separate runs can be made with the inner loop for each of the four runs made with the outer loop, as listed in table 3-1. Each of the eight runs could be made to simulate any desired number of "years" of the system.

The runs corresponding to S0, S1, S2, and S3 in table 3-1 were each made to simulate 15 separate "years" of LACIE operations. After each area estimate $\hat{A}_i$ for the nine-state USGP region was calculated, the CV for that and all the previous years was calculated, as shown in figure 3-2. It appeared that at 15 years CV's had converged sufficiently well to a constant value to stop the processing.

TABLE 3-1.— RUNS MADE WITH THE LPP

Outer-loop run	Inner-loop runs	
	Sampling error only	Sampling classification and yield errors
1	S0	SCY0
2	S1	SCY1
3	S2	SCY2
4	S3	SCY3

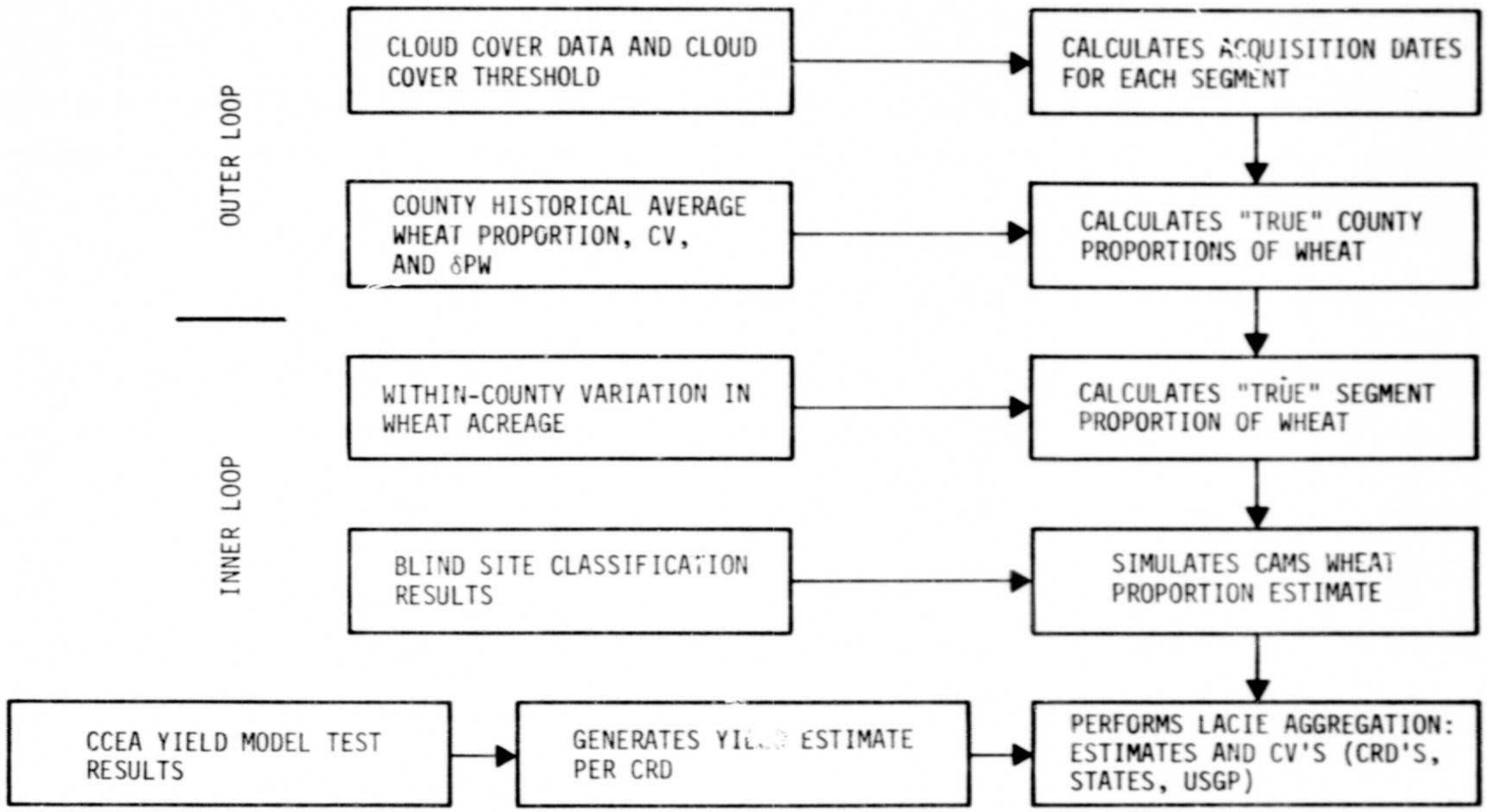

Figure 3-1.– LPP data flow.

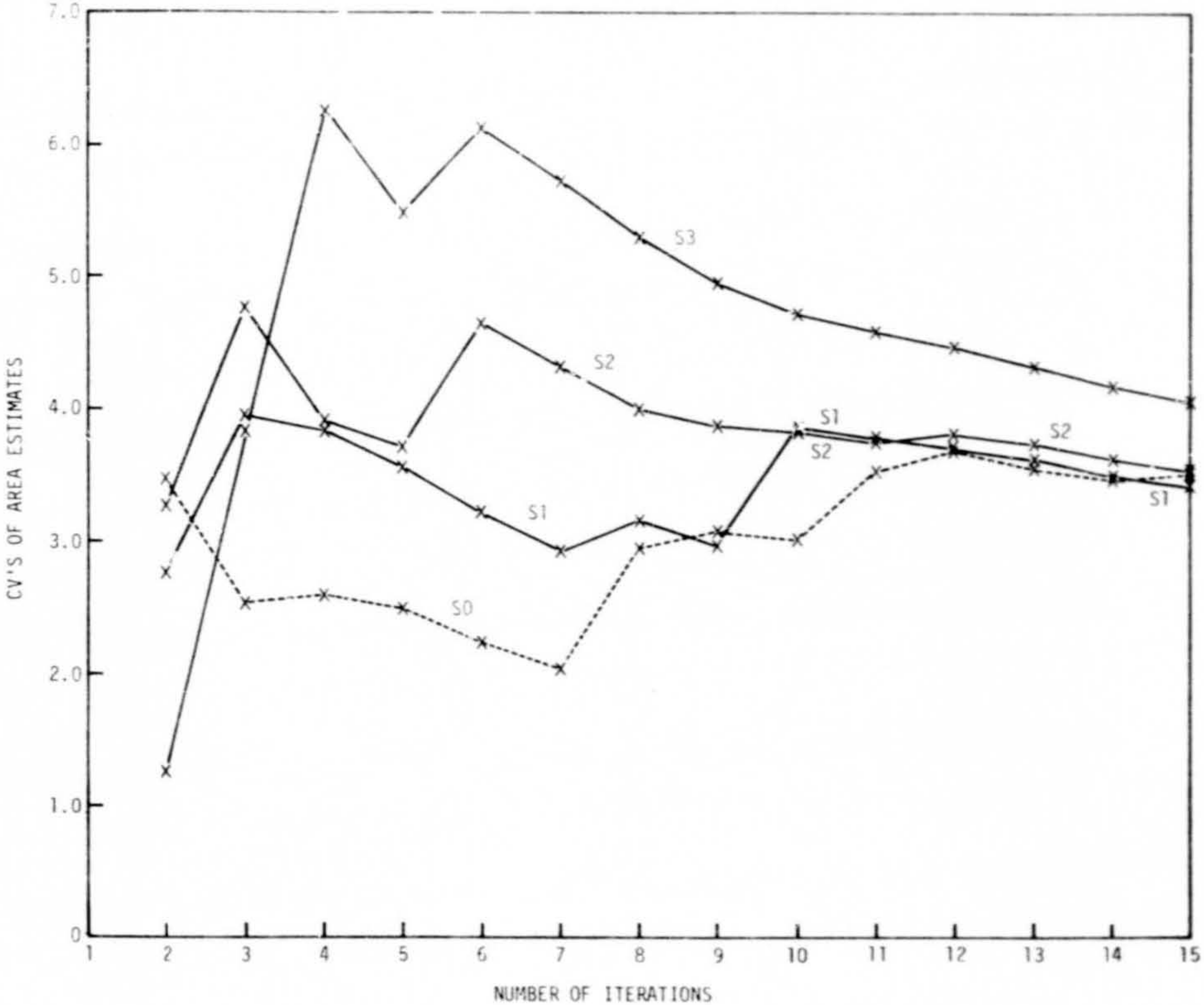

Figure 3-2.— CV's for area estimates for the USGP as a function of the number of iterations.

4. RESULTS

4.1 SEGMENT ACQUISITION

The results of these runs were used to make a study of the acquisition simulation part of the LPP. The fraction of the sample segments having at least one acquisition as determined by the LPP was plotted as a function of time and compared with the number actually obtained in LACIE. The results are shown in figure 4-1. The curve labeled A is the LPP results in the case where zero cloud cover was assumed; i.e., the cloud cover simulator was programmed to always produce a cloud cover of zero. By December 1, all of the winter wheat segments had been acquired and the curve is flat until April 1, when the acquisition of spring wheat sites begins. All sites had been acquired at least once by July 1.

The three curves labeled B correspond to simulations of three different "years" of LACIE operations where the only factor which varied was the cloud cover (i.e., S1, S2, and S3). A threshold of 50 percent was used and was chosen to obtain approximately the same total number of acquisitions over the year as was obtained in LACIE Phase II.[1] The three curves are quite close together, indicating only a small effect of cloud cover on acquisition history. This is probably due to the fact that in the LPP it is assumed that cloud cover at each segment is independent of the cloud cover at all other segments, whereas in fact there is probably a high degree of correlation between the amounts of cloud cover over segments that are reasonably close together.

Curve C is the actual LACIE Phase II acquisition history. It is lower than the curves produced by the LPP for all dates, partly because the cloud cover threshold of 50 percent was too high. Also, the discrepancy is quite large early in the year. The reason for this is not known.

[1]Actually the threshold was too high, and the three curves labeled B correspond to about 15 percent more acquisitions than the 2249 acquisitions in LACIE Phase II.

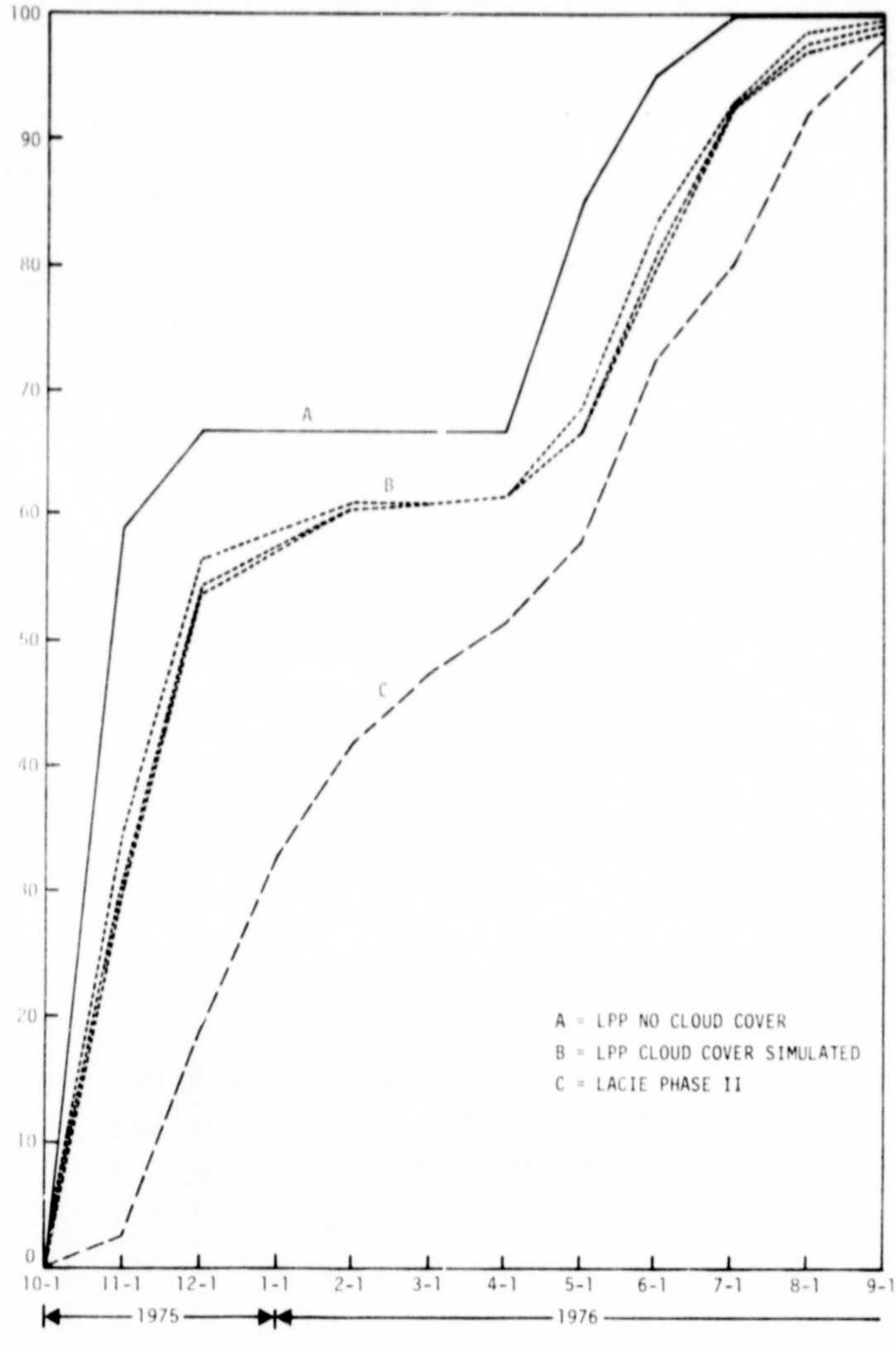

Figure 4-1.— Percent of sites acquired as a function of date.

12

4.2 "TRUE" PROPORTIONS FOR SEGMENTS

The "true" wheat proportions for the blind site segments generated by the LPP in run number S1 are compared with the actual blind site proportions in figure 4-2. The LPP produced more segments with 0 to 4 percent wheat and more segments with a high proportion of wheat (greater than 55 percent). A Kolmogorov-Smirnov test was performed and showed that there was no significant difference between the two distributions.

4.3 CAMS PROPORTION ESTIMATES

Originally it was planned to also make all the runs corresponding to column 2 in table 3-1. However, as each run took about 3 hours of computer time, it was decided to drop the unrealistic case SCY0; as there was little difference between the results of the runs of S1, S2, and S3, it was decided to drop SCY2 and SCY3. Thus, the only run made which included more than sampling error alone was SCY1, which used the same "true" county proportions as S1.

Figure 4-3 shows a histogram of the LACIE errors ($\hat{X}_k - X_k$) for all of the blind sites in the USGP region. Figure 4-4 shows a histogram of the errors simulated by the LPP in the run SCY1; i.e., $\hat{X}_{ij} - X_{ij}$ for all blind sites in the USGP. These histograms should be similar if the LPP is correctly simulating the results of the CAMS classification procedures.

Leptokurtosis (peakedness in the center of distribution) is evident in the LPP results. A comparison of the two distributions (figs. 4-3 and 4-4) using Kolmogorov-Smirnov statistics showed that the distributions were significantly different.

4.4 ACREAGE AND PRODUCTION ESTIMATES FOR THE USGP

The 15 different acreage and production estimates by the LPP for the USGP final prediction date (September 1) of the SCY1 run are plotted in figure 4-5. The abscissa and ordinate are respectively the relative differences of the production and acreage estimates relative to the "true" values used by the LPP for the 15 years. The "true" state wheat acreage value is obtained simply

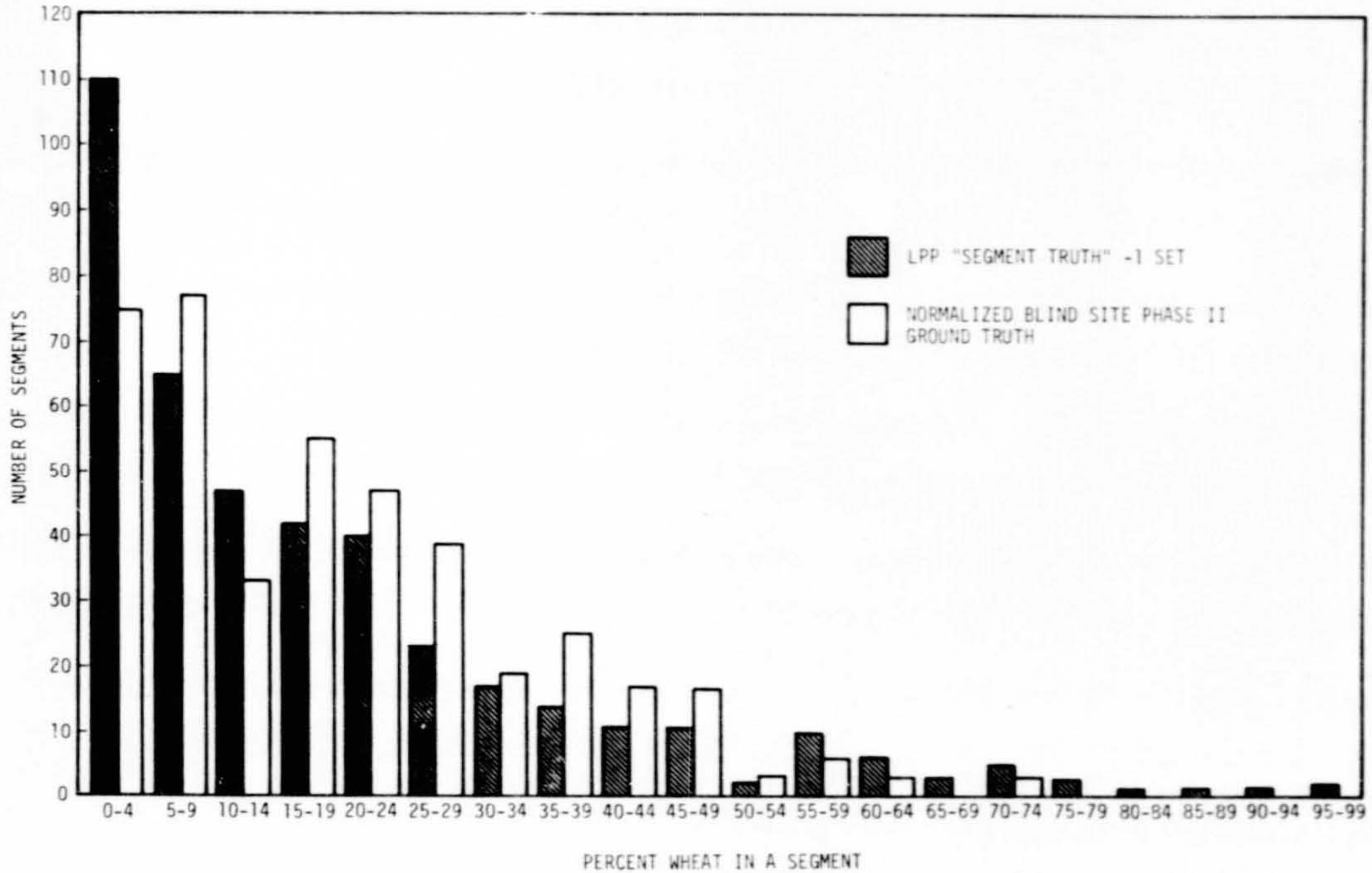

Figure 4-2.— Comparison of model-generated sample-segment wheat proportions with LACIE Phase II ground truth.

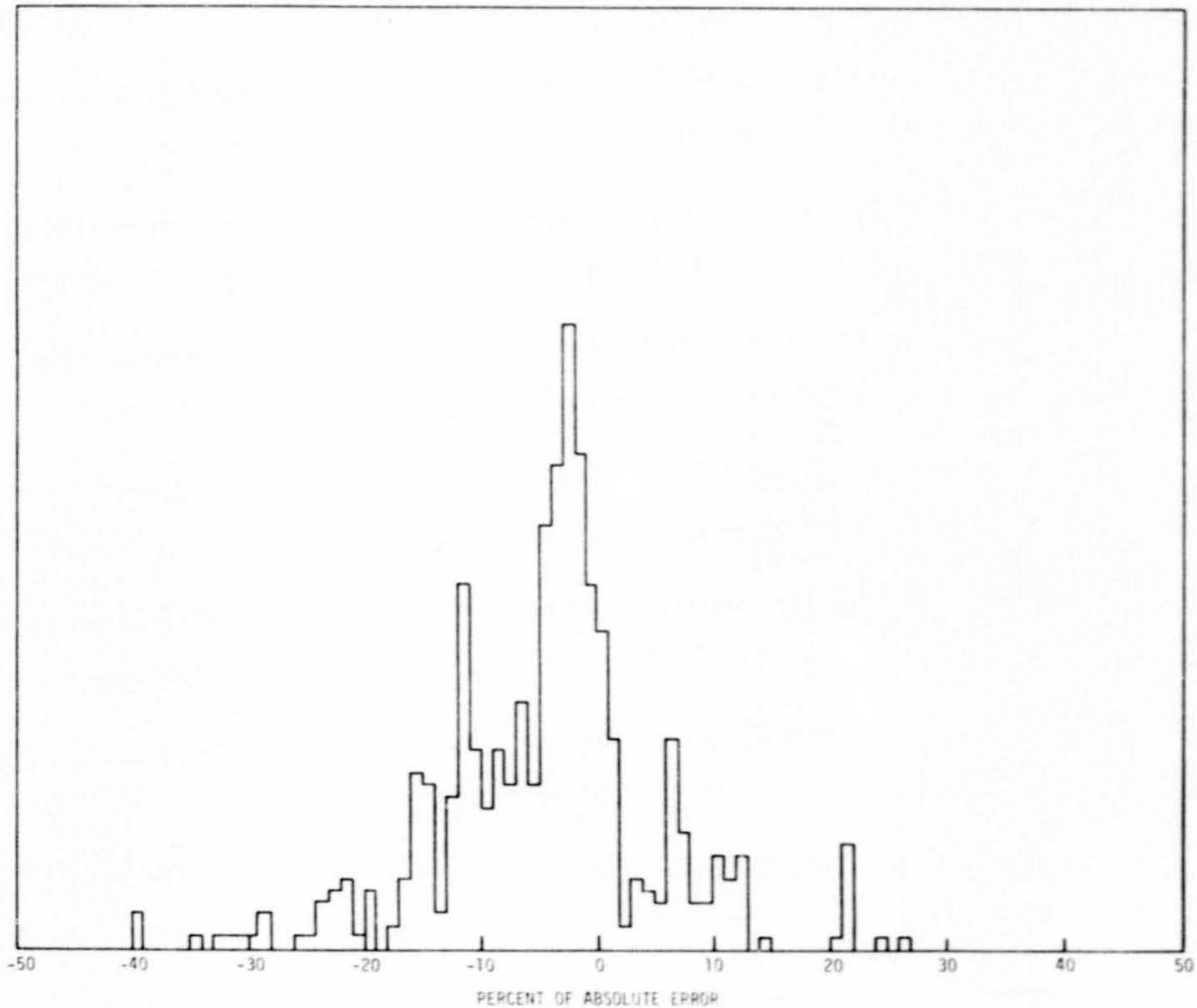

Figure 4-3.— LACIE Phase II blind site wheat proportion estimation errors.

15

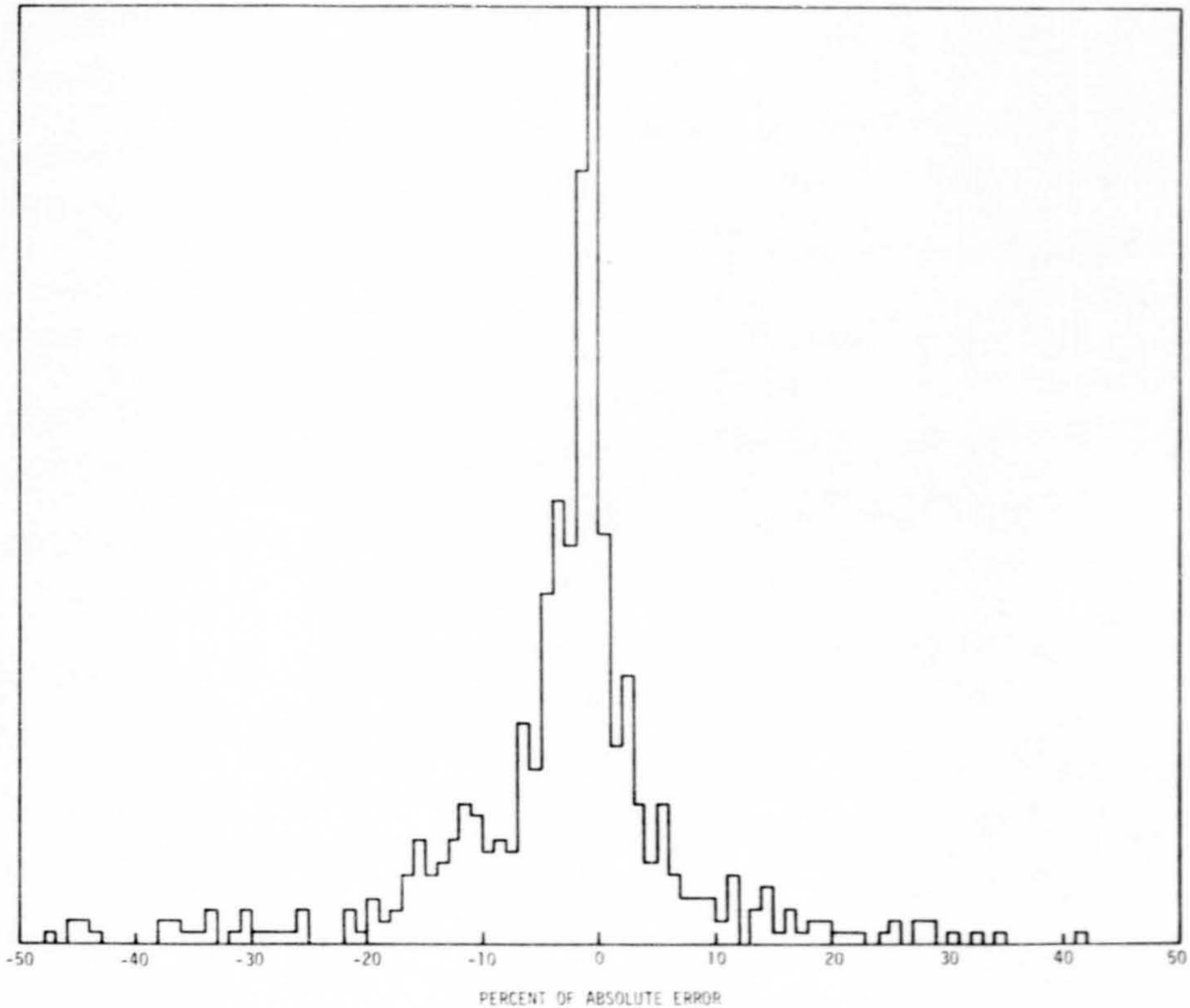

Figure 4-4.– LPP simulation of segment wheat proportion estimation errors.

by adding up all the "true" county acreages (which are simulated by the LPP). The "true" state production value is obtained as follows:

a. The "true" acreage for each CRD is determined by summing the "true" county acreages for each county in the CRD.

b. The "true" acreage is multiplied by the "true" yield (input to the LPP at the CRD level) to get "true" production for the CRD.

c. A sum is performed over all the CRD's in the state.

The LACIE Phase II result is also plotted in figure 4-5 except that it is expressed as the percent differences from the last USDA/SRS figures.

The LACIE result is very close to the mode of the LPP values. A normal curve with the mean and standard deviation of the production data is shown in figure 4-6. The LACIE result is also shown. The relative bias in the LPP production estimate is the value corresponding to the peak in this distribution, -8.7 percent. At the 5-percent level of significance, the value is not significantly different from the LACIE relative bias of -11.0 percent. However, -8.7 percent is significantly different from zero, which indicates that, if the assumptions made in the LPP are correct, the LACIE technology will, on the average, produce an underestimate of wheat production. This could be caused by (1) low segment proportion estimates, (2) low yield estimates, or (3) a bias in the aggregation system. It has been shown (reference) that the aggregation system was not biased, and the fact that the actual LACIE aggregated acreage estimate has a larger (negative) relative difference than the actual LACIE production estimate indicates that low acreage estimates are the cause of the low production estimates. Finally, it should be noted that a relative bias of -8.7 is too large to satisfy the 90/90 criterion.

4.5 OVERALL VARIABILITY OF AREA AND PRODUCTION ESTIMATES

Figure 4-7 shows (1) histograms of the estimated standard deviations σ_i calculated for each of the 15 iterations, (2) the standard deviations σ of the area estimates produced by the 15 iterations, and (3) the estimated standard

17

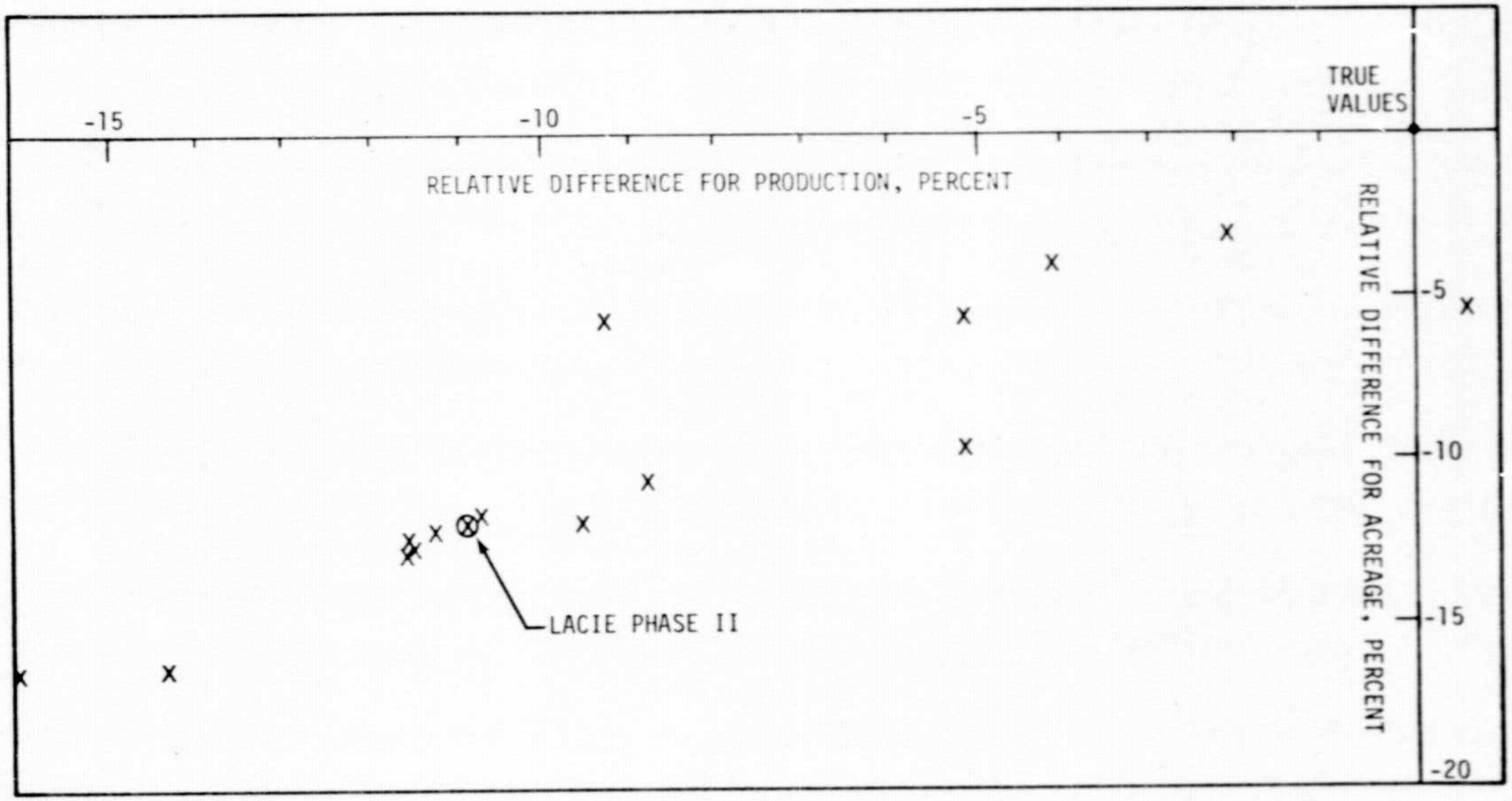

Figure 4-5.— LPP estimates for acreage and production for the USGP.

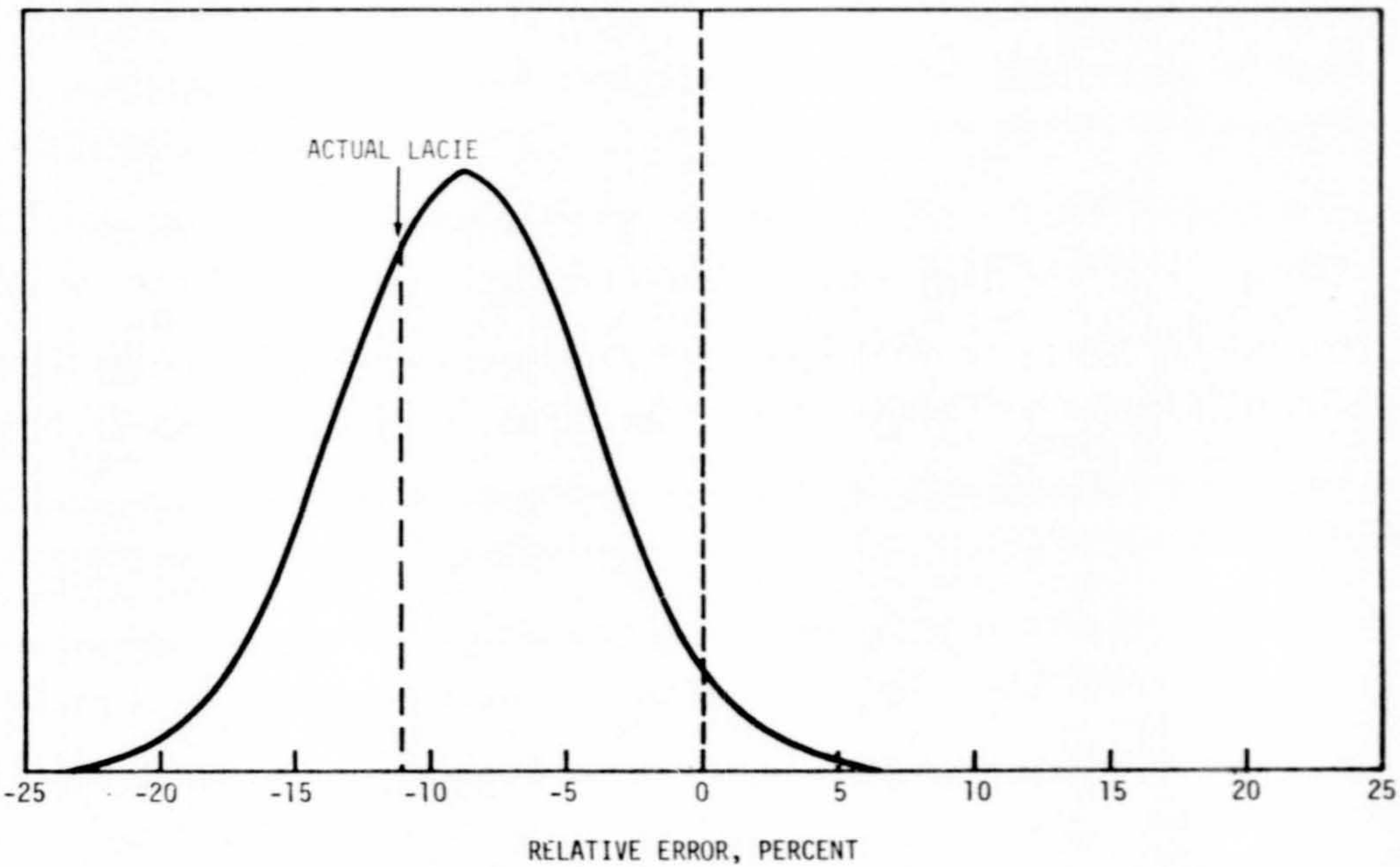

Figure 4-6.— Normal distribution approximating 15 iterations of the LPP.

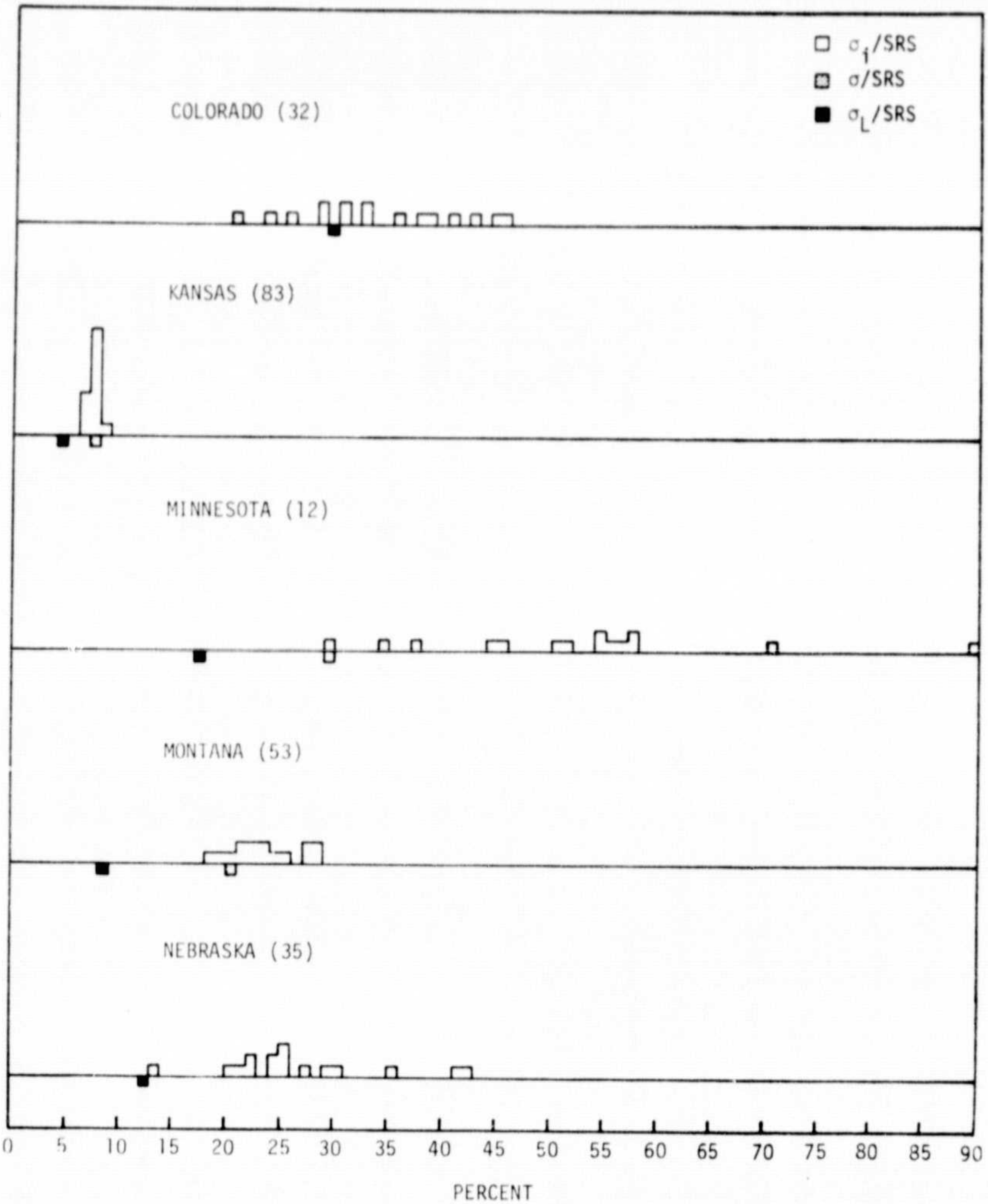

Figure 4-7.— CV's of acreage estimates by state.

20

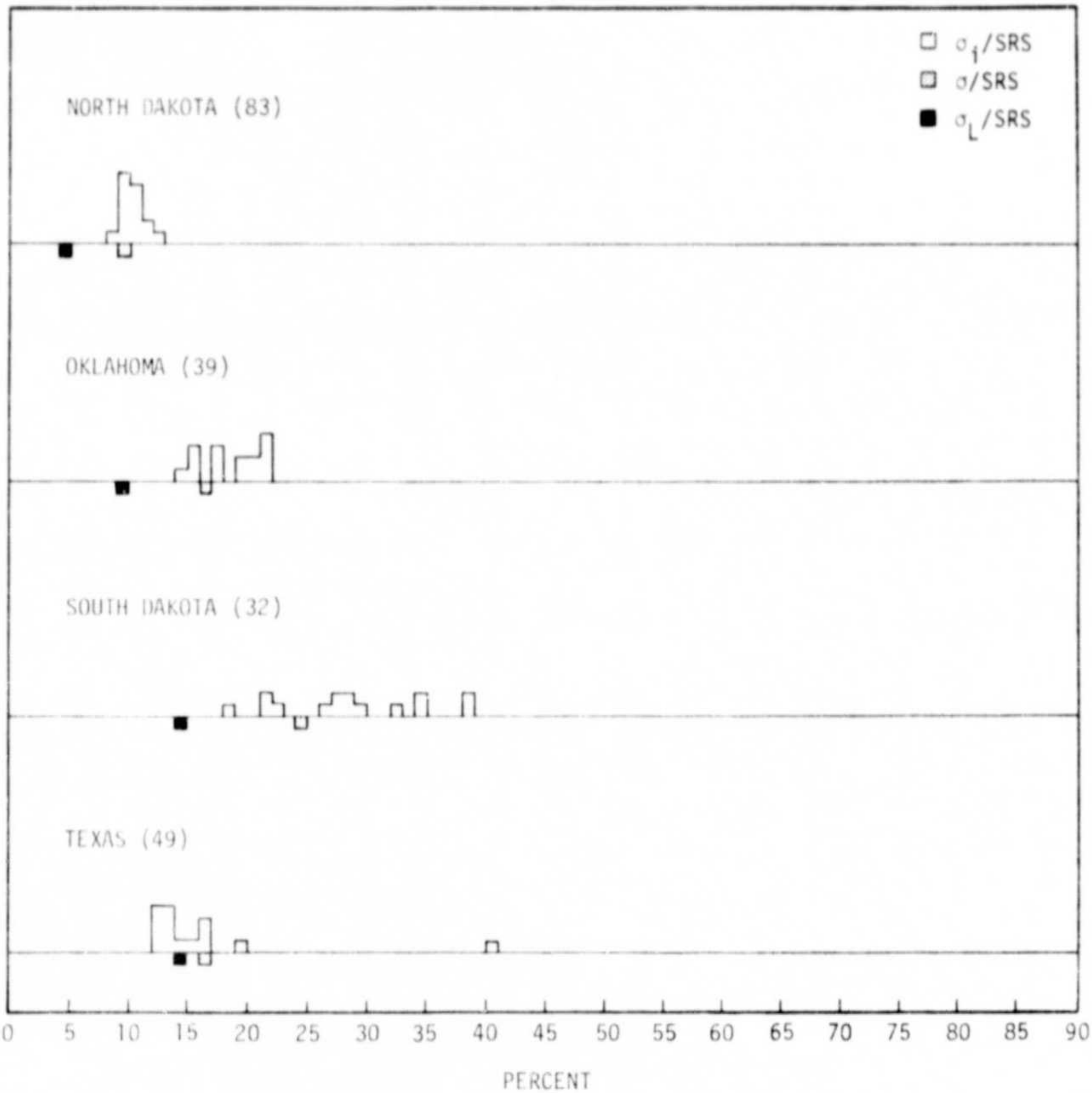

Figure 4-7.– Concluded.

deviations σ_L of the LACIE estimate.[2] Each of these is divided by the SRS acreage estimate. The number after the name of each state is the number of segments "acquired" by the LPP for that state. For most states there is a rather broad distribution of the σ_i for the 15 years. The width of the distribution is generally smaller when the number of segments acquired in a state is larger. In particular, the states with the largest number of acquisitions, Kansas and North Dakota, have quite narrow distributions.

There are two important observations to be made concerning these results:

a. For every state except Texas and Colorado, σ_L is smaller than all of the σ_i. This is partly caused by the tendency of CAMS to overestimate the wheat in segments with low wheat proportions and underestimate the wheat in segments with high wheat proportions, which reduces the variance in the CAMS estimates. This phenomena is apparent in the significant difference between the distributions of proportion estimation errors shown in figures 4-3 and 4-4.

b. With the exception of Kansas, South Dakota, and Texas, σ falls near the lower end of the distribution of the σ_i, as expected, because the formulas for calculating σ_i were designed to give a conservative estimate (i.e., on upper bound). This result is important since it implies that σ_L as calculated by the CAS is also likely to be an overestimate of the true LACIE standard deviation.

Figure 4-8 shows similar results for the standard deviations for production. These histograms are very similar to those in figure 4-7 and the same observations apply.

[2] σ_i is calculated by the LPP in the same manner that σ_L is calculated by the Crop Assessment Subsystem (CAS).

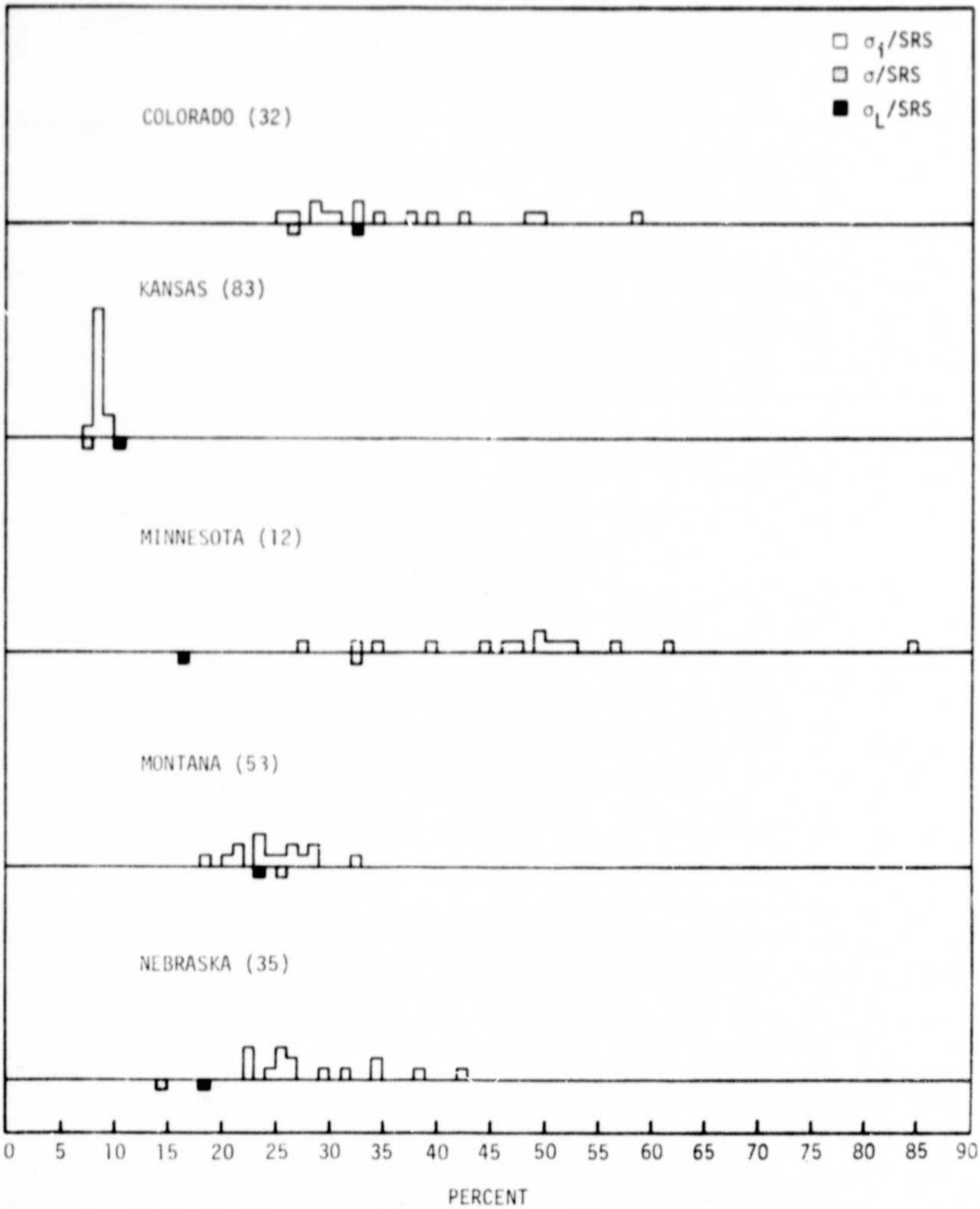

Figure 4-8.— CV's of production estimates by state.

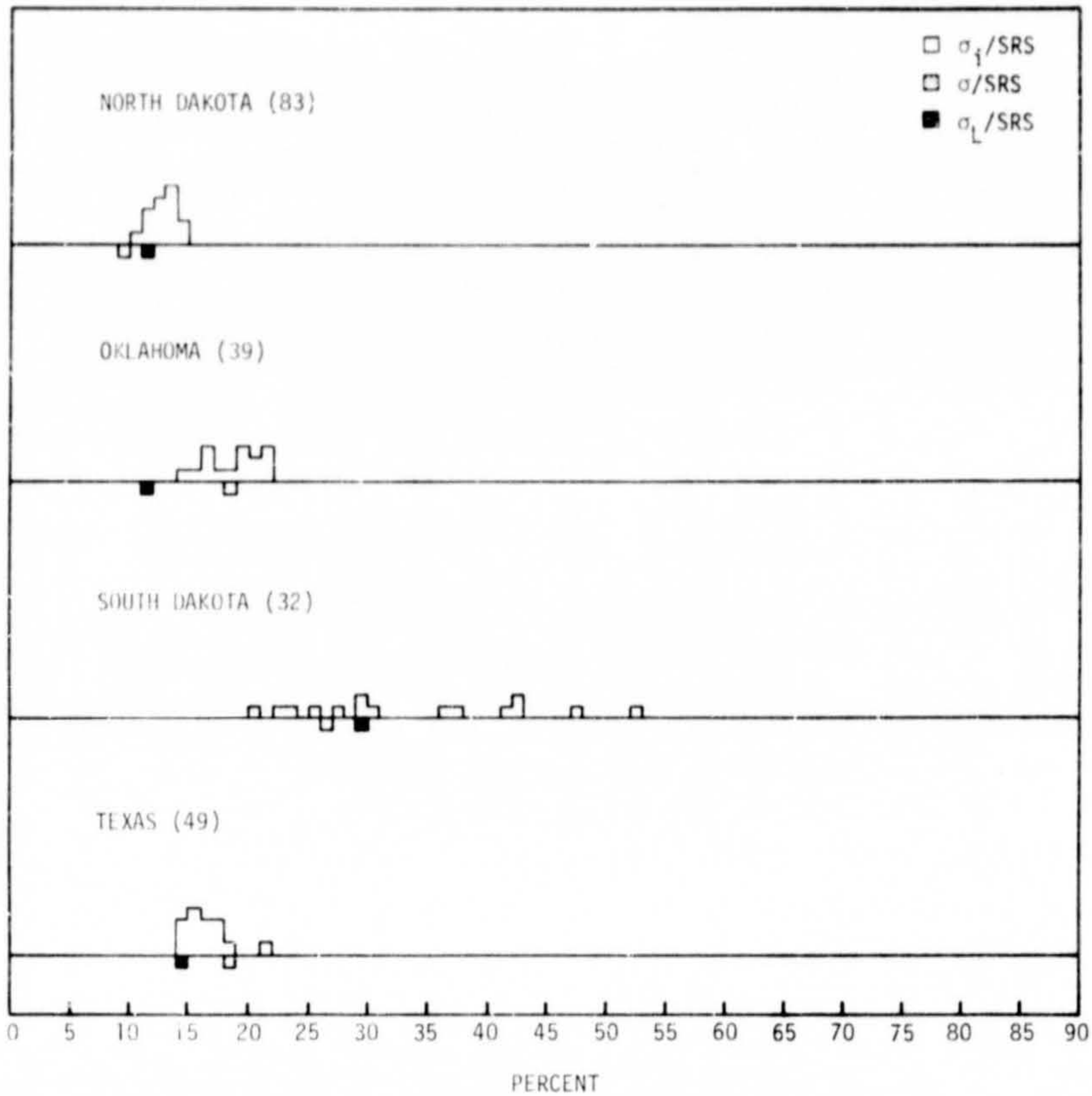

Figure 4-8.— Concluded.

5. CONCLUSIONS

The LPP has been used to replicate LACIE Phase II for a 15-year period using accuracy assessment results for Phase II error components. The results indicate that the LPP simulates the LACIE Phase II procedures reasonably well. For the 15-year simulation, only 7 of the 15 production estimates were within 10 percent of the true production. Further, the simulations indicate that the acreage estimator, based on CAMS Phase II procedures, has a negative bias. This bias is too large to support the 90/90 criterion with the CV observed and simulated for the Phase II production estimator. The results of this simulation study validate the theory that the acreage variance estimator in LACIE is conservative. The simulated results also indicate that the estimated variance for the production estimator is conservative; that is, it tends to be larger than the true variance of the production estimator. Hence, more bias can be tolerated than indicated by the estimated CV. However, even with a reduction in the estimated CV to account for this overestimation, the bias indicated by the simulations is still too large to support the 90/90 accuracy goal.

25

6. REFERENCE

Houston, A. G.: Effects of Non-Response Including Cloud Cover on Aggregation of Wheat Area in the U.S. Great Plains. LACIE-00441 (JSC-11672), Dec. 1976.