

R. S. Chhikara, Lockheed Electronics Company, Inc.*
 A. H. Feiveson, National Aeronautics and Space Administration

1. INTRODUCTION

1.1 Background

The National Aeronautics and Space Administration (NASA) land observatory satellite (Landsat) is equipped with a multispectral scanner (MSS) that measures the intensity of reflected electromagnetic energy in four different wavelength bands. When these measurements (spectral signatures) are correlated with the vegetation on the ground, the assessment of crop acreages by acquiring and processing MSS data for the area becomes feasible. To test the validity of this concept, NASA, in collaboration with the U.S. Department of Agriculture (USDA) and the National Oceanic and Atmospheric Administration (NOAA), started the Large Area Crop Inventory Experiment (LACIE) in 1974 to develop a large area wheat acreage estimation system by utilizing remote sensing technology.

The Landsat coverage of an area is in the form of a scene consisting of X scanlines with Y resolution elements per scanline, where X and Y are the number of scanlines and resolution elements, respectively. The size of a resolution element is approximately 1.1 acres. Thus, in order to evaluate a scene, each resolution element in the scene is classified according to its spectral measurement vector.

The spectral classes are identified through image analysis techniques. The MSS data for a scene are digitized and converted into color-infrared images, then photointerpreted to determine the classes of spectral data. In this effort the best guide is crop knowledge of the area to which the image corresponds and the analysts' experience with photointerpretation. Since Landsat multispectral data for an area are collected every 18 days and a crop can be distinguished from others by monitoring the temporal development of its fields from planting through harvest, the image analysts are able to label some spectral classes by crop types on the ground. Such labeling becomes the basis for estimating different crop acreages after discriminant analysis with respect to the spectral classes is performed.

1.2 LACIE Program

In LACIE it was envisioned that agricultural survey systems could be developed to do crop inventory globally, using satellite-acquired information. For demonstration purposes, wheat was chosen as the crop to be estimated due to the economic importance of estimating wheat production in various countries. Production was estimated by utilizing the direct observational

capabilities afforded by Landsat, together with estimates of weather variables. The geographic subregions of a particular country, which were selected to be relatively homogeneous with regard to wheat acreage and yield, were each monitored (1) to forecast the quantity of wheat acres available for harvest (both winter and spring, individually, in each subregion) and (2) to forecast the expected productivity, or yield, of the acres available for harvest for each subregion. The total wheat production for each subregion was then computed by multiplying the available acres for harvest by the yield for harvested acres. The production forecasts for all subregions were summed to obtain the country-level forecast. Acreage was estimated through a sample survey approach, while the yield predictions were obtained through models developed by regressing historical yields for an area on local weather variables. In this paper only large area acreage estimation is addressed. (Details and results of the LACIE program are being documented and will be presented in the forthcoming LACIE Symposium to be held at the Lyndon B. Johnson Space Center, Houston, Texas, in October 1978.)

Since LACIE was the first attempt to utilize Landsat MSS data for any large area crop acreage estimation and no large-scale usable satellite information previously existed, it was necessary to use historical data available by political subdivisions for developing the initial LACIE sampling design. Political subdivisions also formed the basis for stratification within countries. In the United States the 1969 Agricultural Census data at the county level were used to make the sample allocation since these data were more accurate and consistent at this level than the more recent estimates by the Statistical Reporting Service (SRS) of the USDA.

When data from the 1974 Agricultural Census and the processing of Landsat-acquired segments became available, both the stratification and sample allocation were updated for the 1976-77 crop season. Though consideration was also given to yield estimation in this updating effort, it had little effect on the sample allocation and there was no change in the basic sample design. Changes in sample size and sample allocation were primarily due to the development of a better sampling frame and the use of empirically assessed within-strata variances as compared to the variances previously computed, assuming these were proportional to those from a binomial distribution.

Finally, LACIE estimates of harvested wheat acreages for the U.S. Great Plains (USGP) during 1975, 1976, and 1977 were compared with the corresponding SRS estimates, and the coefficients of variation (CV's) of the LACIE estimates were estimated.

2. LACIE SAMPLING DESIGN

The sampling design was basically a classical survey, where the sampling unit was a 5 by 6

*This material was developed under NASA Contract NAS 9-15200 and prepared for the Earth Observations Division of the National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas.

nautical mile segment and Landsat data were used for the estimation. Classification accuracy and certain engineering constraints other than sampling errors were the primary considerations in deciding on a segment size of 5 by 6 nautical miles for the sampling unit.

2.1 Determination of Sampling Units and Frame

Due to various data base engineering constraints, a maximum of 4800 sample segments could be processed within a crop year, regardless of the size of the individual segment. Given this maximum number of sample segments, the physical dimensions of the sample segments were set at 5 by 6 nautical miles. This size was large enough for the Classification and Mensuration Subsystem (CAMS) analysts' use when obtaining wheat acreage estimates and small enough so the computer and manpower resources would not be taxed. Throughout this paper, the term "sample segment" refers to 5 by 6 nautical mile segments actually in the LACIE sample; the term "segment" refers to any 5 by 6 nautical mile area, whether or not it is in the sample.

A sampling frame was constructed by first overlaying a map of the wheat-growing regions of a country with a large grid of 5 by 6 nautical mile segments, and then excluding those segments which appeared to have less than 5-percent agriculture, as determined by an examination of the previous years' Landsat imagery. The remaining segments constituted the frame from which the actual sample segments were chosen.

2.2 Allocation of Samples to Countries

Initially it was decided that approximately the maximum of 4800 sample segments would be allocated to eight major wheat-producing countries, proportional to their most recent wheat-production statistics. Two types of sampling strategy were used in LACIE: one for countries with historical wheat data on a detailed level (level D) and one for countries with published historical data only for fairly large political subdivisions (level N). Table 1 lists the eight LACIE countries, their smallest political subdivision (SPD) for which published historical data exist, and the number of samples in the initial allocation.

2.3 Definition of Strata

In level N countries, sample segments were allocated at random within strata, which were approximately the intersection of SPD's with the sampling frame. More precisely, for a given SPD the corresponding stratum consisted of all 5 by 6 nautical mile segments that were in the sampling frame and whose center points lay in the SPD (see Fig. 1). As a consequence, the agricultural area in each SPD, and hence in the whole country, was approximated by the collection of 5 by 6 nautical mile segments from which the samples were drawn. At the country level, the error in this approximation was negligible; however, when the SPD's were small, adjustments had to be made to the wheat acreage estimate in order

for a stratum to obtain a more precise estimate for the corresponding SPD.

In level D countries, each SPD was also approximated by the collection of 5 by 6 nautical mile segments which lay in the sampling frame and whose center points lay in the SPD. However, the collection in this case was called a substratum rather than a stratum because in some cases no sample segment was selected from it. To distinguish between an SPD and its approximating collection of segments, the latter was called a pseudo SPD (PSPD).

A stratum in level D countries was defined to be the union of the PSPD's that correspond to the next higher political subdivision of the country. For example, in the United States the SPD was a county, and the next higher political subdivision was a crop-reporting district (CRD) within a state. Therefore, the stratum consisted of the collection of all pseudo counties whose corresponding counties lay within that CRD.

2.4 Allocation and Selection of Segments in Strata/Substrata

In the first two phases of LACIE, the sample sizes for individual countries were fixed as shown in Table 1. Since little or nothing was known about the accuracy of yield predictions at that time, it was decided to allocate the samples to strata (level N countries) or substrata (level D countries) in order to minimize the best *a priori* estimate of the variance of the country's wheat acreage estimate.

It is well known (Cochran 1963) that if a population total is estimated by stratified sampling over L strata with a total sample size of n, the variance of the estimate (when the finite population correction is ignored) is minimized if n_k is proportional to $N_k \theta_k$, where n_k is the sample size for the k th stratum, N_k is the total number of segments in the k th stratum from which n_k samples were selected at random, and θ_k is the standard deviation of the segment characteristics (in this case, wheat acreages) within the k th stratum. This fact was used in LACIE to obtain allocations to strata in level N countries, where N_k was the number of segments comprising the k th stratum and θ_k^2 was assumed proportional to the binomial variance $p_k(1 - p_k)$, where p_k was the historical proportion of wheat in the SPD corresponding to the k th stratum. The optimal sample size for the k th stratum was given by

$$t_k = \frac{n N_k p_k^{1/2} (1 - p_k)^{1/2}}{\sum_{k'} N_{k'} p_{k'}^{1/2} (1 - p_{k'})^{1/2}} \quad (2.1)$$

except that, in general, t_k would not be an integer. For Phases I and II of LACIE, the n_k 's were taken to be the nearest integers to the t_k 's. In level N countries this rounding had little effect since the t_k 's tended to be rather large (between 10 and 50). Once n_k was computed, n_k sample segments were selected at random from the N_k segments comprising the k th stratum.

In level D countries an attempt to use this technique in substrata would result in many t_k 's

being less than 1 or between 1 and 2. As a result, the t_k 's, as computed in (2.1), were used to categorize substrata into three groups:

Group I: $t_k > 1.0$

Group II: $0.1 \leq t_k < 1.0$

Group III: $t_k < 0.1$

The substrata in Group I received n_k sample segments, selected at random, where n_k was t_k rounded to the nearest integer.

All Group II substrata within a stratum were called a Group II collection. Each entire collection received an allocation of segments equal to the rounded total of t_k within the collection. For example, in the United States, if there were three Group II pseudo counties (substrata) in a pseudo CRD (stratum) with respective t_k 's of 0.7, 0.6, and 0.5, then the collection of three pseudo counties would receive a total of two sample segments (the rounded value of $0.7 + 0.6 + 0.5$). Once the sample size, such as m , was determined for a Group II collection consisting of M substrata, the sample segments were chosen with a two-stage sampling scheme: in the first stage, m substrata were selected at random with probabilities proportional to their historical wheat acreage; then one sample segment was selected at random within each of the m chosen substrata. (Note that $m \leq M$.)

The Group III substrata were those that would hypothetically receive less than a tenth of a sample segment in the optimal allocation and were not sampled at all. Instead, their wheat acreage was estimated by first computing a historical ratio of their wheat acreage to that of the neighboring Group I and/or Group II substrata; then that ratio was applied to the current year's estimate for the neighboring Group I and Group II substrata. (See Section 3 for more detail.)

For Phase III of LACIE, some modifications were made to the allocation procedure. Instead of assuming that within-stratum wheat variances were proportional to the binomial $p(1-p)$, where p was the historical proportion of wheat in the stratum, it was decided that a better approximation could be reached by assuming that the wheat variance was proportional to the small grains variance. (The term "small grains" refers to the combined crops of wheat, barley, oats, rye, and flaxseed.) This could be directly estimated from a regression model using Landsat imagery for recent years. The advantage of the new procedure lies in the ability of analysts to examine a Landsat full-frame color image and to obtain crude estimates of small-grain proportions (but not of wheat alone) for all 5 by 6 nautical mile segments within the area covered by the image. [See Feiveson and Hallum (1978) for further details.]

For level D countries in Phase III, the definition of Group III was changed to the set (S) of all substrata, such that

- a. The total historical wheat acreage for substrata in S was approximately 2-1/2 percent of the country's historical wheat acreage.

- b. If S_1 and S_2 were substrata such that $S_1 \in S$ and $S_2 \notin S$, then S_2 had more wheat historically than S_1 .

The t_k 's in (2.1) were then computed only for the substrata remaining after the elimination of those designated as Group III.

Finally, rather than allocate the 4800 sample segments to the countries in proportion to their production, it was decided in Phase III that the total allocations in the United States and the U.S.S.R. would be revised by estimating the total sample sizes needed to satisfy given accuracy criteria and then using these sample sizes as long as the total was less than 4800. This was accomplished by specifying a desired CV for the production estimate of each country and then calculating the sample size necessary to achieve that CV, given the acreage sampling error as computed (Feiveson and Hallum 1978), the *a priori* classification error variances, and the yield prediction error variances. The resulting allocation for Phase III was a total of 601 segments in the United States and 947 segments in the U.S.S.R. The 601 segments in the United States were distributed among 288 Group I and 164 Group II counties.

3. ACREAGE ESTIMATION

Nonresponse due to cloud cover often reduced the number of sample segments acquired from a stratum. Because the strata were generally large in level N countries, it was felt that unless three or more segment acreage estimates were available in a stratum, no acreage estimate for the stratum should be made. In the case of level D countries, the strata were considerably smaller in size; therefore, this requirement was waived at the stratum level, but was imposed at the higher level (zone). (The term "zone" is defined as the political subdivision next higher to the stratum.)

For level N countries, the wheat acreage estimate of a stratum is given by

$$A_j = \frac{N_j R_j}{n_j} \sum_{k=1}^{n_j} A_{jk} \quad (3.1)$$

where A_{jk} is the wheat acreage estimate for the k th segment in the j th stratum, n_j is the number of sample segments from the j th stratum for which estimates are available, N_j is the number of segments in the j th stratum, and R_j is the ratio of the actual area to the gross pseudo area for the j th stratum. (The term "gross pseudo" refers to the case where nonagricultural segments are not excluded from the accounting of the PSPD.) The use of the R_j in (3.1) enables an estimate to be applicable to a political subdivision area.

In level D countries, the acreage estimate of a stratum may consist of the Group I, II, and III component estimates. A Group I substratum and/or the collection of Group II substrata is treated as Group III substrata if no acreage estimate is available from at least one sample segment in each case. The Group I substrata are treated as

strata, and a stratified random sampling estimator is employed for estimating their wheat acreages. On the other hand, a probability-proportional-to-size (PPS) estimator is used to estimate the wheat acreage for the entire Group II collection of substrata in a stratum. The wheat acreage for the Group III collection of substrata in a stratum is estimated using a ratio estimator.

Let A_{1j} , A_{2j} , and A_{3j} denote the Group I, II, and III component acreage estimates, respectively, for the j th stratum. Then

$$A_{1j} = \sum_{k=1}^{L_{1j}} \frac{N_{1jk} R_{1jk}}{M_{1jk}} \sum_{i=1}^{M_{1jk}} A_{1jki} \quad (3.2)$$

where N_{1jk} is the number of segments in the k th substratum (PSPD) of the j th stratum, M_{1jk} is the number of sample segments for which estimates are available in the k th substratum of the j th stratum, R_{1jk} is the ratio of the true k th substratum area to its gross pseudo substratum area, A_{1jki} is the estimated wheat area for the i th sample segment in the k th substratum of the j th stratum, and L_{1j} is the number of Group I substrata in the j th stratum.

$$A_{2j} = \sum_{k=1}^{M_{2j}} R_{2jk} \frac{A_{2jk}}{\pi_{2jk}} N_{2jk} \quad (3.3)$$

where A_{2jk} is the wheat area estimate of the sample segment belonging to the k th substratum in the j th stratum (Only one segment was allocated in each selected Group II substratum.), M_{2j} is the number of sample segments for which acreage estimates are available in the Group II substrata of the j th stratum, N_{2jk} is the number of segments in the k th Group II substratum of the j th stratum, R_{2jk} is the ratio of the true k th Group II substratum area to its gross pseudo substratum area, and π_{2jk} is the probability of selection for the k th Group II substratum of the j th stratum, which is given by

$$\pi_{2jk} = \frac{M_{2j} W_{2jk}}{W_{2j}} \quad (3.4)$$

where W_{2jk} is the harvested wheat area during the primary epoch year in the k th Group II substratum of the j th stratum and

$$W_{2j} = \sum_{k=1}^{L_{2j}} W_{2jk}$$

where L_{2j} is the number of Group II substrata in the j th stratum.

Depending upon the number of segments in a stratum for which data are available, three categories of Group III acreage estimates are possible. Categories 1, 2, and 3 correspond, respectively, to three or more segments, one or two segments, and no segments having data available in the stratum. The ratio used for the Group III estimator is the ratio of historical

wheat acreages for Group III substrata to Group I and II substrata.

For category 1 estimates (three or more usable segments in the stratum), the ratio is based solely on historical acreages within the stratum. The acreage estimate of the Group III substrata in the j th stratum is given by

$$A_{3j} = \left(\frac{A_{1j} + A_{2j}}{W_{1j} + W_{2j}} \right) W_{3j} \quad (3.5)$$

where A_{1j} and A_{2j} are given by (3.2) and (3.3) and W_{1j} , W_{2j} , and W_{3j} are the historical wheat acreages for the Group I, II, and III substrata in the stratum, respectively.

For the category 2 and 3 estimates (less than three usable segments in the stratum), the ratio is based on acreages in the zone containing the stratum for which the estimate is being made. The acreage estimate of the Group III substrata in the j th stratum is obtained by

$$A_{3j} = \left(\frac{A_{1\cdot} + A_{2\cdot}}{W_{1\cdot} + W_{2\cdot}} \right) W_{3j} \quad (3.6)$$

where a dot (\cdot) in a subscript denotes the summation over all the Group I or Group II substrata in the zone, whichever applies. The reason for differentiating between categories 2 and 3 is to facilitate the stratum variance estimation.

Thus, the wheat acreage estimate for the j th stratum of a level D country is

$$A_j = A_{1j} + A_{2j} + A_{3j} \quad (3.7)$$

Wheat area estimates for zone and for higher levels (e.g., region and county) are obtained by adding estimates for the strata included in the zone, region, and country.

In level D countries, the problem of acreage variance estimation involves several complexities resulting from the use of a two-stage PPS sampling scheme for the Group II substrata and the availability of only one sample segment per substratum in most cases. The variance estimation procedure in such countries consists of a series of steps. On the other hand, the estimation of the variance in the case of level N countries is fairly straightforward. For these countries, no variance estimates are attempted for strata containing less than three available segments, and all strata belong to the Group I category. [For details see Chhikara and Feiveson (1978).]

4. LACIE ESTIMATES FOR THE U.S. GREAT PLAINS

Starting in 1975, LACIE estimates were made of wheat acreages for the nine states in the USGP: Colorado, Kansas, Minnesota, Montana, Nebraska, North Dakota, Oklahoma, South Dakota, and Texas. During the first year of LACIE (Phase I), the intention was to develop a system capable of processing Landsat data on a large scale and to adapt the pattern recognition and sample survey methodologies for remote sensing

applications. From the total of 411 allocated segments, usable data were obtained on 272. These data were used to estimate harvested acreages by state; separate estimates were made for winter and spring wheat regions of the USGP.

The LACIE estimate for winter wheat relative to the corresponding SRS estimate differed by only a fraction of a percent, but it was almost 40 percent lower than the SRS estimate for spring wheat. For the total wheat acreage in the USGP, the LACIE estimate was about 10 percent below the SRS estimate. The estimated CV was 6 percent for winter wheat and 9 percent for spring wheat. This indicates that the winter wheat estimates were quite reliable and accurate, but that the spring wheat was significantly underestimated.

The tendency to underestimate spring wheat acreage was in part a result of the inability to discriminate wheat from the other spring small grains (e.g., barley, oats, and flaxseed) using Landsat data alone. Spectrally these crops are similar, as are their growth cycles. Therefore, estimates of total small grains were made for segments in spring wheat areas, and historic ratios of these acreages (i.e., the ratio of wheat to small grain acreages) were used to reduce these Landsat estimates of small grains to estimates of wheat acreage for the segments. Another contributing factor was the practice of strip-fallow farming in spring wheat regions. Strip-fallow fields, which are small compared to the Landsat resolution, are difficult to detect and measure from the Landsat imagery.

During 1976 and 1977 (LACIE Phases II and III), estimates were made each month from February until November, when the final estimates were made, for the five USGP winter wheat States of Colorado, Kansas, Nebraska, Oklahoma, and Texas. Estimates were made from June to November for the four spring or mixed wheat States of Minnesota, Montana, North Dakota, and South Dakota in the U.S. northern Great Plains. These estimates are compared with the corresponding estimates released by the SRS in Figure 2 for 1976 and in Figure 3 for 1977. Again, separate comparisons are made for winter, spring, and total wheat.

As in 1975, the LACIE estimates were in good agreement with the SRS estimates for winter wheat and significantly different for spring wheat. For total wheat in 1977 the two estimates were almost equal. The estimated CV's and the number of segments used are listed in Table 2. The smaller CV estimate in 1977, as compared to that in 1976, was partly due to a larger number of available segments and partly due to a better allocation of sample segments.

5. SUMMARY

The results indicate that fairly accurate and reliable estimates can be obtained using Landsat data in conjunction with good ancillary information on crops in the region. In a technological sense, remote sensing shows great potential for surveying the totality of crops with a common growth pattern (e.g., spring wheat, barley, and oats in the U.S. northern Great Plains); however, difficulty in distinguishing such crops from each other was experienced in LACIE.

Both a measurement error and nonresponse problems were encountered. The measurement error was due to a fallible classification method, in which an image analyst labeled the spectral class and then performed the statistical discriminant analysis for the segment. The main thrust of the LACIE research and development (R&D) program has been to develop classification techniques that would minimize this error. The second problem arose when segment data were lost because of segment nonacquisition and/or inability in obtaining its estimate. However, the bias caused by nonresponse in LACIE estimates was assessed to be much smaller than the bias due to the measurement error.

ACKNOWLEDGEMENTS

The authors would like to thank Professor H. O. Hartley of Texas A&M University, who initially suggested the LACIE sampling design and was consulted throughout the reported experimental work.

REFERENCES

- Chhikara, R. S., and Chang, J. (1976), "LACIE Area Variance Estimate in the U. S.," Technical Memorandum LEC-8056, Lockheed Electronics Company, Inc., Houston.
- , and Chang, J. (1976), "An Empirical Study of Variance Estimation with One Unit per Stratum," paper presented at the 136th annual meeting of the American Statistical Association in Boston.
- , and Feiveson, A. H. (1978), "LACIE Large Area Acreage Estimation," *Proceedings of the LACIE Symposium* to be held in October 1978 in Houston.
- Cochran, W. G. (1963), *Sampling Techniques*, New York: John Wiley and Sons, Inc.
- Feiveson, A. H., and Hallum, C. (1978), "LACIE Sampling Design," *Proceedings of the LACIE Symposium* to be held in October 1978 in Houston.
- Hartley, H. O., and Rao, J. N. K. (1962), "Sampling with Unequal Probabilities and Without Replacement," *Annals of Mathematical Statistics*, 33, 350-374.
- , Rao, J. N. K., and Kiefer, Grace (1969), "Variance Estimation with One Unit per Stratum," *Journal of American Statistical Association*, 64, 841.
- Hensen, M. H., Hurwitz, W. N., and Madow, A. G. (1953), *Sample Survey Methods and Theory*, New York: John Wiley and Sons, Inc.
- Liszczy, C. J. (1978), "LACIE Area Sampling Frame and Sample Selection," *Proceedings of the LACIE Symposium* to be held in October 1978 in Houston.

MacDonald, R. B., and Hall, F. G. (1977), "LACIE: A Look to the Future," *Proceedings of the Eleventh International Symposium on Remote Sensing of Environment* held in

Ann Arbor, Michigan.

Seth, G. R. (1966), "On Collapsing of Strata," *Journal of the Indian Society of Agricultural Statistics*, 18, 1-3.

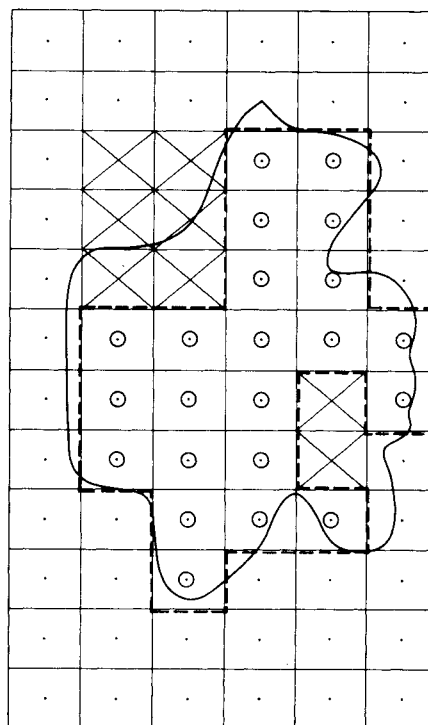
TABLE 1. LACIE COUNTRIES AND THEIR SAMPLE SEGMENT ALLOCATIONS

Country	SPD with published wheat data	Number of segments in initial allocation
United States	County (D)	637
U.S.S.R.	Oblast (N)	1949
China	Province (N)	810
Canada	Crop subdistrict (D)	283
India	State (N)	626
Australia	Shire (D)	257
Argentina	Partido (D)	165
Brazil	State (N)	47
Total		4774

TABLE 2. NUMBER OF SAMPLE SEGMENTS AND ESTIMATED CV'S FOR 1976 AND 1977

Wheat type	1976		1977	
	Number of segments	CV	Number of segments	CV
Winter	278	5.0	298	3.2
Spring	129	6.0	178	3.5
Total	407	4.0	476	2.4

FIGURE 1. SPD AND CORRESPONDING STRATUM IN LEVEL N COUNTRIES



Legend:





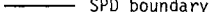
-  5 by 6 nautical mile segment with at least 5-percent agriculture on stratum
-  5 by 6 nautical mile segment with at least 5-percent agriculture outside stratum
-  5 by 6 nautical mile segment with less than 5-percent agriculture
- center point
-  stratum boundary
-  SPD boundary

FIGURE 2. MONTHLY COMPARISON OF LACIE AND SRS ACREAGE ESTIMATES FOR USGP IN 1976

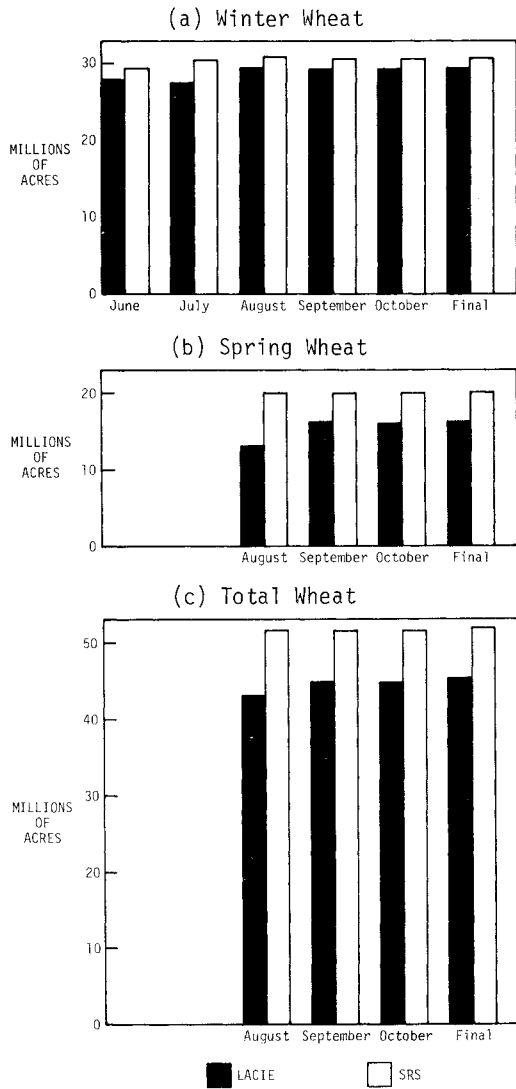


FIGURE 3. MONTHLY COMPARISON OF LACIE AND SRS ACREAGE ESTIMATES FOR USGP IN 1977

