81-10034
FC-LO-00428 CR-163554
JSC-16343
AUG 29 1980

AgRISTARS

A Joint Program for
Agriculture and
Resources Inventory
Surveys Through
Aerospace
Remote Sensing

1. Foreign Commodity
Production Forecasting

5. July 1980

ESTIMATION OF WITHIN-STRATUM VARIANCE FOR SAMPLE ALLOCATION

R.S. Chhikara
Lockheed Engineering and Management Services Company, Inc.

C.R. Perry, Jr.
Senior Resident Research Associate
National Research Council, NASA/JSC

(E81-10034) ESTIMATION OF WITHIN-STRATUM VARIANCE FOR SAMPLE ALLOCATION: FOREIGN COMMODITY PRODUCTION FORECASTING (Lockheed Engineering and Management) 49 p HC A03/MF A01 CSCL 02C G3/43

N81-12516

Unclas
00034

UNITED STATES DEPARTMENT OF AGRICULTURE

NASA

DEPARTMENT OF COMMERCE
UNITED STATES OF AMERICA

U.S. DEPARTMENT OF THE INTERIOR
MARCH 3, 1849

UNITED STATES OF AMERICA

Lyndon B. Johnson Space Center
Houston, Texas 77058

FC-L0-00428
JSC-16343

ESTIMATION OF WITHIN-STRATUM VARIANCE FOR SAMPLE ALLOCATION

Job Order 74-432

This report describes Sampling and Aggregation activities of the Foreign Commodity Production Forecasting Project of the AgRISTARS program.

PREPARED BY

R. S. Chhikara
Lockheed Engineering and Management Services Company, Inc.

C. R. Perry, Jr.
Senior Resident Research Associate
National Research Council, NASA/JSC

APPROVED BY

RM Bizzell
for R. O. Hill, Head
Inventory Technology Section
Crop Applications Branch
NASA/JSC

BL Carroll
B. L. Carroll, Manager
Commodity Forecasting Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.
Under Contract NAS 9-15800

For

Earth Observations Division
Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

July 1980

LEMSCO-14067

1. Report No.
FC-LO-00428, JSC-16343

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Estimation of Within-Stratum Variance for Sample Allocation

5. Report Date
July 1980

6. Performing Organization Code

7. Author(s)
R.S. Chhikara, Lockheed Engineering and Management Services Company, Inc.
C.R. Perry, Jr., NRC Senior Resident Research Associate, NASA/JSC

8. Performing Organization Report No.
LEMSCO-14067

9. Performing Organization Name and Address
Lockheed Engineering and Management Services Company, Inc.
1830 NASA Road 1
Houston, Texas 77058

10. Work Unit No.

11. Contract or Grant No.
NAS 9-15800

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas 77058, Technical Monitor: R. O. Hill

13. Type of Report and Period Covered
Technical Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The problem of determining stratum variances required for an optimum sample allocation for crop surveys by remote sensing is investigated considering an approach based on the concept of stratum variance as a function of the sampling unit size. A methodology using the existing and easily available information of historical crop statistics is developed for obtaining initial estimates of stratum variances. The procedure is applied to estimate stratum variances for wheat in the U.S. Great Plains and is evaluated based on the numerical results thus obtained. It is shown that the proposed technique is viable and performs satisfactorily with the use of a conservative value (smaller than the expected value) for the field size and the use of crop statistics from the small political subdivision level.

17. Key Words (Suggested by Author(s))
within-stratum variance
field size
sampling unit size
crop proportion
ratio of crop to agricultural acreages

18. Distribution Statement

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
49

22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161

PREFACE

The work documented herein was performed within the Earth Observations Division, Space and Life Sciences Directorate at the Lyndon B. Johnson Space Center (JSC), National Aeronautics and Space Administration (NASA). R. S. Chhikara, Lockheed Engineering and Management Services Company, Inc., and C. R. Perry, Jr., National Research Council (NRC), Senior Resident Research Associate, NASA/JSC are the principal scientists. C. R. Perry, Jr., is on leave from Texas Lutheran College, Seguin, Texas. Other scientists and personnel who assisted in this work are: A. H. Feiveson and C. R. Hallum of the NASA/JSC, Professor H. O. Hartley of Texas A&M University, and C. J. Liszcz of Lockheed. In addition, Liszcz developed part of the software required during this study.

The authors wish to express their appreciation to all of the above for their encouragement and helpful suggestions relating to this study.

CONTENTS

TABLES

Table Page

FIGURES

Figure Page

PRECEDING PAGE BLANK NOT FILMED

ACRONYMS

AgRISTARS	Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing
CAMS	Classification and Mensuration Subsystem
EOD	Earth Observations Division
JSC	Lyndon B. Johnson Space Center
LACIE	Large Area Crop Inventory Experiment
PIXEL	picture element
NASA	National Aeronautics and Space Administration
PC	pseudo count
TY	Transition Year
USGP	U.S. Great Plains

1. INTRODUCTION

In any cost-effective stratified sampling design, the optimal sample size and its allocation between the different strata depend on the within-stratum variances, the stratum size, and the precision required for the estimate. With the development of an area sampling frame, strata sizes are known in terms of the total number of sampling units per stratum. The precision goal is fixed in advance and hence known. However, prior to the survey, no direct knowledge of within-stratum variances is available; therefore, it is necessary to estimate them. Usually, a pilot survey is conducted and, subsequently, the information resulting from the pilot study is utilized in planning a full-scale sample survey. In this report, a methodology for indirectly estimating stratum variances using existing agricultural statistics and other ancillary information is proposed and evaluated for the U.S. Great Plains (USGP).

In most countries, crop statistics are computed annually either through complete enumeration or by employing sample survey methodology. However, the geographical level and the type of crop statistics reported vary considerably from one country to another. For example, reliable crop statistics for area, yield, and production are available in the United States at the county level. In contrast, crop statistics are not available for China at a political subdivision level lower than the country level. Canada, India, and several other countries provide fairly reliable annual crop statistics at a geographic level similar to the U.S. county. Yet, even among these countries, the type of crop statistics produced is varied; for example, in Australia, annual crop statistics contain no information on harvested acreage. Consequently, no fixed procedure can be applied to each and every country for determining the within-stratum variances.

Initially, in the Large Area Crop Inventory Experiment (LACIE), a proportional sample allocation based primarily on historical wheat production was employed. That is, a fixed total sample size was allocated to the different countries of interest and to the smaller political subdivisions within each country so as to be proportional to the historical wheat production of the different geographic subdivisions. In the later phases of LACIE, methods were devised to estimate the within-stratum variances by utilizing past Landsat imagery and other ancillary data. These estimates permitted a more nearly optimal sampling allocation to be employed during the final phases of LACIE.

During the first year of concentration in a crop/region, little to no previously analyzed Landsat data are available for making within-stratum variance estimates; this will be the case in many crop/regions of the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS) program. Thus, a technique is needed for making initial within-stratum variance estimates without the use of previously analyzed Landsat data. The description and the evaluation of such a technique are presented in this report. The technique is motivated by the empirical models employed by Perry and Hallum (ref. 1) in their study on sampling unit size. Also discussed in this context are the methodologies employed during the LACIE to estimate the within-stratum variances for sample allocation in the crop survey program of the Earth Observations Division (EOD), National Aeronautics and Space Administration (NASA), Lyndon B. Johnson Space Center (JSC). Other information included in this report are the following. The approaches adopted in LACIE Phases I, II, and III and in the Transition Year (TY) are described in section 2. Details of the proposed technique are given in section 3. Different variations of this procedure as applied to estimate refined-stratum variances for wheat in the USGP are given in section 4.1. [Refer to Chhikara (ref. 2) for details of the stratification considered in this study.] A discussion of the stratum-variance estimates obtained using the different methods is given in section 4.3. It is concluded in section 5 that if reliable historical crop acreages are available at a small political subdivision level (e.g., county in the U.S.), then fairly good stratum-variance estimates can be obtained using the proposed method.

The technique for making initial within-stratum variance estimates is designed to make optimal use of the available data (even if limited by its reliability) for estimating within-stratum variances on crop/regions that otherwise would not be estimated because previously analyzed Landsat data are not available.

2. PREVIOUS APPROACHES

2.1 LACIE PHASES I AND II

During Phases I and II of LACIE, the total sample size was determined primarily by engineering and resources constraints. However, sample survey methodology [the Neyman Optimum Allocation Formula (ref. 3)] shows that, if allocation of the total sample to the different strata were made proportional to the respective product of stratum size and within-stratum standard deviation, the resulting crop estimate should have a minimum variance for a fixed overall sample size. Thus, for a cost effective design, knowledge of within-stratum variances is required.

In order to estimate the within-stratum variances used as input into the Neyman allocation formula, the binomial model was assumed where the sampling unit had dimensions of 5- by 6-nautical miles (a segment). That is, if $\tilde{p}$ is the crop (wheat/small-grains) proportion for a stratum, then $\tilde{p}(1 - \tilde{p})$ is a rough estimate of the between-segment crop proportion variance for the stratum. That this model overestimates the within-stratum variance for all strata was recognized because the model assumes that every segment is entirely wheat or nonwheat, which is far from reality even in the new lands of the Union of Soviet Socialist Republics (U.S.S.R.). However, it was considered reasonable to assume that these estimates reflected the relative magnitudes among the true within-stratum variances. Hence, it was thought that the total sample was utilized in a cost-effective manner. It was recognized that an optimal overall sample size could not be determined using a binomial model because of considerable positive bias in the variance estimates produced by the model.

2.2 LACIE PHASE III

For this period, greater emphasis was placed on achieving a more accurate crop acreage and production estimates. As a result, a decision was made to reallocate the sample segments in the USGP for LACIE Phase III. Among other factors, this decision was based on the desirability of having more reliable within-stratum variance estimates as input variables to the allocation formula than could be obtained from the binomial model. It was noted that the sample units were large and could be expected to contain some nonagricultural areas. Also, it was envisioned that if the segment crop area were related to the segment agricultural area, then this statistical relationship could be exploited to produce an improved within-stratum variance estimation procedure. This, in fact, proved to be the case. The resulting within-stratum variance estimation technique was derived using the following approach.

The crop proportion in a sample segment was expressed as

$$p = ra \tag{1}$$

where

p = the proportion of crop acreage in a segment

r = the ratio of crop acreage to agriculture acreage in a segment

a = the proportion of agricultural acreage in a segment

It was assumed that the ratio r did not depend on the proportion of agricultural acreage in a segment. Then, the variance of p was easily computed using the formula for the variance of the product of two independent random variables. For each stratum, this yielded the following formula.

$$\sigma_p^2 = \sigma_r^2 \left[E^2(a) + \sigma_a^2\right] + \sigma_a^2 E^2(r) \tag{2}$$

The mean and variance of the proportion of agricultural acreage, $E(a)$ and σ_a^2, respectively, were obtained directly from estimates of the proportion of agricultural land in each segment in a stratum. The available Landsat imagery was used for this determination. However, it was not feasible to obtain directly such information for the variable r. Instead, the mean and variance, $E(r)$ and σ_r^2, were estimated for each stratum as follows: $E(r)$ was estimated by

$$r_h = \frac{\text{Historical crop acreage for stratum } h}{\text{Landsat agricultural acreage for stratum } h} \tag{3}$$

and σ_r^2 by

$$\hat{\sigma}_r^2 = Kr_h(1 - r_h) \tag{4}$$

where K = 0.03

The value of K in equation (4) was based on an empirical study for small grains where the mean and variance of r were computed from segment data obtained from Landsat imagery for 40 counties in the USGP. These counties were considered as strata. Then the stratum variance was modeled by

$$\sigma_r^2 = Kr(1 - r) \tag{5}$$

where r was the mean ratio of crop acreage to agricultural acreage in the stratum. A least-squares fit of this model resulted in K = 0.03. The adjustment, K, to the binomial variance, $r_h(1 - r_h)$, reflected the departure from the assumption that the ratio of crop acreage to agricultural acreage in a segment was 0 or 1. Thus, the determination of $\hat{\sigma}_r^2$ from equation (4) could only be regarded as approximate and tenuous. Accordingly, the resulting stratum variance estimate was

$$S_p^2 = 0.03\ r_h(1 - r_h)\ (\bar{a}^2 + S_a^2) + S_a^2\ r_h^2 \tag{6}$$

where $\bar{a}$ and S_a^2 were the mean and variance of the proportion of agricultural acreage in a segment, respectively, and where this proportion was determined by using Landsat imagery for the stratum. The properties of S_p^2 could not be determined for several reasons. The most obvious reasons were the empirical nature of the derivation and the historical nature of the input data. Nevertheless, for initial within-stratum variance estimates, this model was expected to be an improvement over the binomial model considered in Phases I and II of LACIE.

2.3 TRANSITION YEAR (TY)

The method of computing initial stratum variance estimates for use in the TY project was influenced by two developments. First, a geographical stratification based on agrophysical characteristics had been developed for the TY sampling design (ref. 4). Second, sample data from LACIE Phase II in the form of segment wheat and small-grains proportion estimates were available for use in direct estimation of the stratum variances. Although these sample data did not constitute a random sample relative to the new stratification, it was generally assumed that estimates based on these data would be more reliable than those obtained by using the earlier indirect methods. However, for some strata, sufficient segment data needed for directly estimating the stratum variance were not available. When this occurred, the stratum variance was estimated indirectly by employing the approach used in LACIE Phase III. The nonrepresentative nature of the sample data used in the direct estimates and the use of two altogether different methods of estimation could have led to inconsistencies among the stratum variance estimates. If true, this would have adversely affected the associated sample allocation.

An evaluation of the TY sample allocation was performed using the LACIE Phase III sample segment estimates. Phase III segment estimates were used because they were available and were regarded as more reliable than those from Phase II. The evaluation indicated an underallocation of sample segments to some strata and an overallocation of sample segments to other strata. For further details, refer to Chhikara (ref. 2). However, in reference 2, the effect of the nonrepresentative nature of the LACIE Phase III segment data with respect to the TY strata was not considered.

For sample allocations in the future program of AgRISTARS, it would be ideal to have reliable and representative Landsat segment estimates in order to make direct initial estimates of the stratum variances. However, it is not expected that initially such data will be available for most countries of interest. Accordingly, some indirectly derived stratum-variance estimates will need to be determined for the purpose of making a sample allocation. The approach

6

used for LACIE Phase III seems reasonable and feasible except for the determination of the variance of the ratio of crop acreage to agricultural acreage. A new procedure for obtaining initial stratum crop proportion variances is offered and described in section 3. The procedure is equally applicable to estimating the stratum variance σ_r^2.

3. PRESENT METHODOLOGY

A procedure for indirectly estimating the stratum variances used in an initial allocation is presented. There are three basic underlying ideas. First, obtain estimates of the stratum variance for a set of sampling unit sizes including both large and small size sampling units; second, establish empirically a relationship between the sampling unit size and the stratum variance; and third use the empirical model to obtain an estimate of the stratum variance for the desired sampling unit size which is a segment.

In the context of crop estimation, Smith (ref. 5) and Mahalonobis (ref. 6), independently of each other, proposed that the stratum between-units variance could be modeled as a power function of the sampling unit size. Historically, a number of empirical studies [Smith, Mahalonobis, Jessen, Hansen et al., and Asthana (refs. 5, 6, 7, 8, and 9, respectively)] strongly indicate that the power function provides a simple, yet satisfactory, mathematical model for the functional dependence of the stratum between-units variance on the sampling unit size. The first application of this functional form specifically to the between-units crop proportion variance was made by P. C. Mahalonobis (ref. 6) in his 1938 study of jute production for Bengal (India). He considered the following function for the stratum between-units crop proportion variance.

$$\sigma_x^2 = \frac{\tilde{p}(1 - \tilde{p})}{(bx)^g} \qquad (7)$$

where $\tilde{p}$ is the stratum crop proportion and x is the sampling unit size. The sample sizes considered in this study were 1, 2.25, 4, 6.25, and 9 acres.

The rationale behind the variance formulation in equation (7) is as follows: when $x = 1/b$, the variance $\sigma_x^2 = \tilde{p}\ (1 - \tilde{p})$ and 1/b represents the largest area (e.g., crop field) for which the crop proportion is either 0 or 1. As x increases in size away from 1/b, the denominator in equation (7) increases and σ_x^2 decreases with $\tilde{p}\ (1 - \tilde{p})$ as an upper bound. If it is assumed that fields in a stratum are not mixed and all fields are approximately of equal size, the difference between the average field size and the sampling unit size being considered should be indicative of the decrease in σ_x^2 from $\tilde{P}\ (1 - \tilde{P})$; a smaller decrease in σ_x^2 is expected with a smaller difference between the sampling unit size and 1/b. Consequently, the bias in estimating σ_x^2 by $\tilde{P}\ (1 - \tilde{P})$ will be smaller for the smaller size sampling unit, and it is zero when the sampling unit size is less than or equal to 1/b.

This same model was employed by Perry and Hallum (ref. 1) in their sampling unit size study. Their study was based on the LACIE Phase III ground-truth data set and concluded that indeed the power function does provide a satisfactory model for the between-units wheat acreage (or proportion) variance for sampling unit sizes ranging from 171 to 25 426 acres. Several other studies, particularily those by Jessen (ref. 7) and Asthana (ref. 9), show this general relationship to hold reasonably well even for very large areal units, a county for example.

The relationship in equation (7) can be rewritten as

$$\sigma_x^2 = \alpha x^\beta \tag{8}$$

where

x = the sampling unit size
σ_x^2 = the stratum crop proportion variance corresponding to x

and α and β are parameters to be empirically determined for each stratum.

In developing this model for the different strata, it would be ideal to have knowledge of σ_x^2 over a wide range of sampling unit sizes, x. For most countries, this is not feasible because it would require expensive sampling or complete enumeration to be performed, thus defeating the purpose of employing the model in the first place. Therefore, one is led in least-squares estimation of the stratum parameters α and β to choose sampling unit sizes for which σ_x^2 can be estimated directly from existing agricultural statistics or can be mathematically modeled and then estimated from existing agricultural statistics.

In the U.S., crop statistics are available at the county level and a strataum normally consists of many counties. Thus, the between-counties variance can be easily computed and used as an estimate of stratum variance corresponding to a sampling unit approximately equal to the average county size. However, since the counties often vary considerably in size, the stratum variance should vary statistically as the sampling unit size varies from the smallest to the largest county. This statistical variability may be preserved by using a one-point estimate of σ_x^2 for each county in the stratum. The one-point estimates are obtained as follows. Consider the county as a sampling unit

where

x_i = the size of the i^{th} county in a stratum

p_i = the proportion of crop acreage for the i^{th} county in the stratum

$\tilde{p}$ = the proportion of crop acreage in the stratum

Then the squared deviation

$$S_{x_i}^2 = (p_i - \tilde{p})^2 \tag{9}$$

provides an estimate of $\sigma_{x_i}^2$ for the sampling unit size x_i. Although these county level estimates can be expected to provide guidance in estimating the stratum variance for a sampling unit approximately the size of a county, they alone can not be expected to be sufficient to predict the stratum variance for a sampling unit of the size of a LACIE segment since it will be outside the sampling unit size range for the counties.

The next three estimates are developed for use with small sampling unit sizes. Any one of these estimates along with the one-point variance estimates from equation (9) is used for the least-squares estimation of the parameters α and β. The resulting regression curve is evaluated for the sampling unit size of interest (segment) to obtain the corresponding stratum variance estimate. Later, it will be observed empirically that the last two relationships provide fairly reliable stratum variance estimates.

First, suppose that all fields are of the same size and shape and the sampling unit is randomly placed with the exception that it intersects only one field. Then the stratum variance corresponding to the field size, x_0, is given by the binomial variance

$$\sigma_{x_0}^2 = \pi(1 - \pi) \tag{10}$$

where π is the proportion of the fields belonging to the crop type of interest. For a fixed crop proportion $\tilde{p}$ and a fixed sampling unit size, the between-units variance is maximized when the sampling unit proportions are all either 0 or 1. Thus, equation (10) provides an upper bound of $\tilde{p}(1 - \tilde{p})$ for the stratum variance regardless of the sampling unit size. This feature and the method, in general, are illustrated in figure 3.1.

Second, in a Lansat type sampling process, the sampling unit is randomly located and is expected to intersect more than one field. Thus, a closer approximation to $\sigma_{x_0}^2$ than that given in equation (10) is desirable. An exact determination of the variance $\sigma_{x_0}^2$ is not feasible. However, a realistic approximation is developed in appendix A under the following assumptions: (1) all fields are square and equal in size to the sampling unit size, x_0, (2) the contents of any four adjacent fields are uncorrelated with respect to the crop of interest, and (3) the sampling unit is randomly placed with the exception that its sides are parallel to the field boundaries. The resulting estimate is given by

$$\sigma_{x_0}^2 = \frac{4}{9}\tilde{p}(1 - \tilde{p}) \tag{11}$$

where $\tilde{p}$ is the stratum crop proportion.

Third, when the sampling unit size x_0 is small relative to the size of the fields, then it is possible to derive the variance in a somewhat exact form as described in appendix B. In this case, the estimate corresponding to the

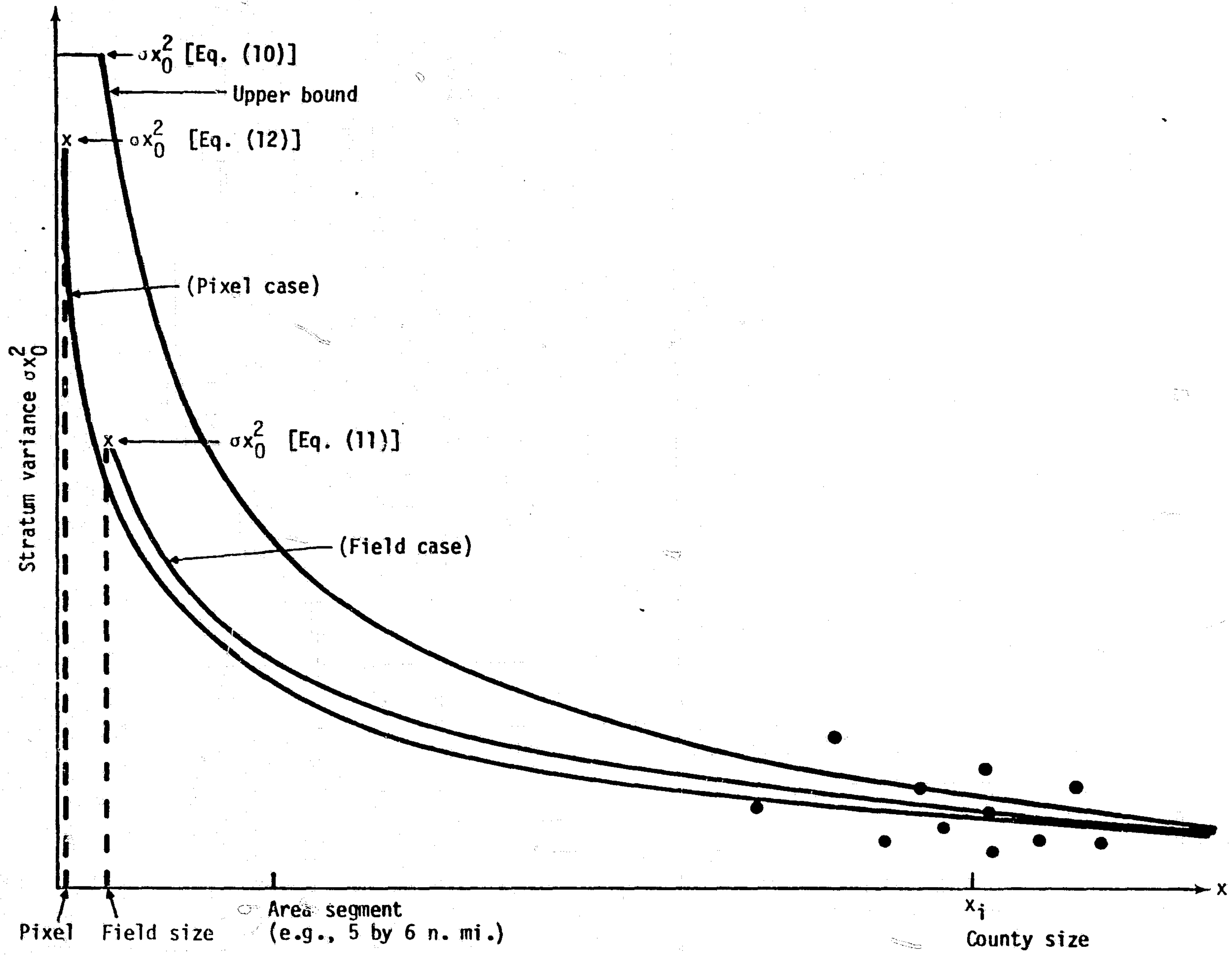


Figure 3-1.- An illustration of the fitted model.

small sampling unit x_0, referred to as a pixel, is approximated by the equation

$$\sigma_{x_0}^2 = \alpha_1(1 - \tilde{p}) + \alpha_2\tilde{p}^2 + \alpha_3(0.3682 - \tilde{p} + \tilde{p}^2) \qquad (12)$$

where α_1, α_2, and α_3 are defined and evaluated in terms of the crop proportion and the field size distribution.

As outlined earlier, equation (9) combined with any one of the equations (10), (11), or (12) provide stratum-variance estimates over widely separated sampling unit sizes from which the parameters α and β can be determined using a least-squares fit. An estimate of the stratum variance corresponding to a specified sample unit size, x, is then obtained by evaluating along the fitted curve

$$\hat{\sigma}_x^2 = AX^B \qquad (13)$$

where A and B are the least-squares estimates of the parameters α and β.

It will be seen from the numerical results that use of both equations (11) and (12) lead to fairly reliable segment level variance estimates. Yet, equation (11) is probably preferable if accurate determination of the field sizes can be made or if the field sizes are large. Otherwise, it is probably better to use equation (12) since it should be less sensitive to error in the field size measurements.

Other estimates of the within-stratum variances can be developed by, first, using one of the above methods to estimate σ_r^2 followed by the application of equation (2) to estimate σ_p^2. However, this type of substitution will likely result in less reliable estimates unless the proposed method estimates σ_r^2 significantly better than σ_p^2.

4. VARIANCE ESTIMATION FOR WHEAT IN THE USGP

4.1 WITHIN-STRATUM VARIANCE ESTIMATION METHODS

Described in this section and evaluated in section 4.3 are the within-stratum variance estimation methods derived from the methodology discussed in section 3. Different methods are created not only by combining the county size units with the field or smaller size units but also by combining the type of least-squares fit used with either a direct estimation of σ_p^2 or an indirect estimation of σ_p^2 by way of σ_r^2 . The three combinations of the sampling unit sizes for the stratum variance estimation are considered in the evaluation: field, equations (9) and (10); field, equations (9) and (11); pixel, equations (9) and (12). The least-squares fit is approached in three different ways: (1) transform the data into logarithmic scale and then minimize the sum of squared deviations; (2) minimize the absolute difference between the aggregated variance resulting from the use of the model equation and the aggregated squared deviations obtained using equation (9); and (3) minimize the sum of squared deviations of variances given by the model from those resulting from the use of equation (9). In each case, the curve $\hat{\sigma}_x^2 = Ax^B$ is passed through the point $(x_0, \sigma_{x_0}^2)$. The different criteria are listed in table 4-1 where, of course, A is replaced by $\sigma_{x_0}^2/x_0^B$ and the summation $\sum_i$ is understood to be taken over all the counties in a stratum.

There are 2 x 3 x 3 = 18 combinations between the type of variance σ_p^2 or σ_r^2, the type of small sampling unit [equations (10), (11), or (12)], and the type of estimation criterion that can be tried for empirical model development. As the computations were made and as the results were evaluated, it was discovered that the introduction of variable r led to less accurate variance estimates than when only the variable p was used. In addition, criterion C-3 in table 4-1 appeared to yield more accurate estimates than the other two criteria. Consequently, no further combinations involving the variances σ_r^2 or the criterion C-1 or C-2 were given consideration. This action resulted in

TABLE 4-1.— MODEL PARAMETER ESTIMATION CRITERIA

Criterion	Approach
C-1	$\text{Min} \sum_i \left(\log S_{x_i}^2 - \log A - B \log x_i\right)^2$
C-2	$\text{Min} \left\| \sum_i \left(Ax_i^B - \frac{n}{n-1} \sum_i S_{x_i}^2\right) \right\|$
C-3	$\text{Min} \sum_i \left(Ax_i^B - S_{x_i}^2\right)^2$

only 8 of the 18 combinations actually being studied. Each of these combinations is designated as a variance estimation method and is listed in table 4-2.

4.2 DATA INPUT

The wheat acreages given in the 1974 Agricultural Census Reports were used in computing the crop proportion data and in computing the ratios of crop acreage to agricultural acreages for both counties and refined strata. The agricultural acreages utilized in the computations came from a complete enumeration of the 5- by 6-nautical-mile segments in the USGP. In this enumeration, Landsat full-frame imagery was used to classify each segment as either 0- to 5-, 5- to 10-,•••, or 95- to 100-percent agricultural land. The segments with 5-percent or more agricultural land were designated as agricultural segments and were used in the computation of county and stratum sizes. The number of agricultural segments in a region is called its pseudo count (PC) and was taken from the LACIE sampling frame.

The average field size (more precisely the distribution of field size) varies from strata to strata and was difficult to determine. The following technique, employing 1974 Agriculture Census Reports data, was used to estimate the average field size for a given stratum. Suppose N_i and A_i, respectively, are the number of operators and the 1974 crop acreage for the i^{th} crop in a stratum. Then, average field size, f_0, for the stratum is estimated by

$$\hat{f}_0 = \left[\sum_{i=1}^{k} A_i \Big/ \sum_{i=1}^{k} N_i\right] \qquad (14)$$

where k is the number of major crops in the stratum. The field size estimates resulting from this computation are listed in column 7 of table 4-3.

4.3 EVALUATION OF VARIANCE ESTIMATES

The stratum variances were estimated for the USGP by each method listed in table 4-2, and the results were compared with estimates based on the TY sample

TABLE 4-2.— VARIANCE ESTIMATION METHODS

Method	Variable	Sampling unit combination	Minimization criterion
1	r	County and field, equation (10)	C-1
2	r	County and field, equation (10)	C-2
3	r	County and pixel, equation (12)	C-3
4	P	County and field, equation (10)	C-1
5	P	County and field, equation (10)	C-2
6	P	County and field, equation (10)	C-3
7	P	County and pixel, equation (12)	C-3
8	P	County and field, equation (11)	C-3

17

TABLE 4-3.— REFINED STRATA DATA INPUT FOR VARIANCE ESTIMATION FOR WHEAT IN THE USGP

State	State code	Refined stratum	Number of counties	Number of all segments	Number of agricultural segments	Average field size in acres	Proportion of wheat acreage	Between-county standard deviation	Mean proportion of agriculture	Standard deviation for agriculture
Colorado	8	9	3	162	150	450	0.16	0.020	0.63	0.1017
		10	20	816	558	345	.13	.088	.57	.1202
		101	21	1075	227	126	.03	.031	.34	.0764
Kansas	20	7	10	229	226	276	.39	.121	[illegible]	.0732
		8	8	179	179	288	.30	.061	.90	.0401
		9	13	258	258	460	.25	.049	.63	.0622
		11	18	410	409	239	.21	.040	.77	.0693
		12	17	317	311	152	.22	.107	.78	.0620
		13	18	271	271	57	.07	.032	.86	.0479
		14	11	161	161	52	.07	.033	.86	.0424
		15	2	37	37	173	.29	.120	.91	.0158
		60	3	78	75	390	.20	.033	.51	.1074
		102	4	84	74	73	.04	.007	.55	.0839
Minnesota	27	15	15	254	238	34	.02	.019	.89	.0475
		19	16	351	317	60	.06	.053	.77	.1012
		20	13	321	308	189	.23	.090	.66	.0624
Montana	30	21	3	141	141	502	.23	.045	.79	.0579
		22	6	280	212	363	.11	.035	.53	.0915
		23	13	1013	662	490	.15	.067	.59	.1019
		104	32	1603	503	213	.04	.030	.30	.0500
Nebraska	31	10	9	234	203	340	.18	.118	.79	.0759
		11	15	315	297	131	.09	.042	.77	.0852
		14	9	137	137	47	.08	.029	.96	.0194
		15	44	672	651	56	.04	.051	.81	.0819
		16	4	120	114	64	.00	.002	.67	.1057
		17	3	121	89	189	.09	.067	.63	.0979
		103	7	275	0	83	.00	.001	.00	.0000
North Dakota	38	19	20	599	582	292	.28	.055	.35	.0537
		20	7	215	214	268	.34	.041	.74	.0321
		21	24	904	831	259	.19	.069	.73	.0895
		22	2	52	30	263	.14	.097	.47	.1153
Oklahoma	40	3	5	88	42	93	.06	.041	.39	.0645
		7	22	516	401	232	.37	.151	.60	.0998
		9	2	96	84	380	.19	.063	.62	.0964
		13	3	49	23	69	.07	.058	.40	.0988
		60	11	285	219	250	.22	.058	.50	.0944
		102	26	578	131	75	.02	.021	.29	.0556
South Dakota	46	15	7	99	99	44	.01	.007	.87	.0393
		16	22	451	441	186	.06	.058	.89	.0444
		17	10	355	358	352	.07	.037	.49	.1211
		18	5	278	204	249	.05	.014	.44	.0902
		19	12	286	283	139	.14	.060	.90	.0343
		21	6	212	197	208	.09	.030	.77	.0917
		104	5	238	89	179	.03	.012	.44	.1128
Texas	48	2	13	307	230	84	.03	.032	.47	.0715
		3	28	598	458	105	.04	.035	.53	.0847
		4	23	556	525	170	.06	.066	.79	.0855
		5	12	276	153	201	.12	.088	.46	.0857
		9	7	192	161	476	.18	.087	.71	.0992
		60	5	130	55	385	.15	.074	.41	.1054
		61	13	290	219	216	.07	.079	.49	.0882
		101	28	673	228	89	.01	.009	.35	.0538
		102	26	499	290	76	.01	.013	.49	.1000

segment data. Comparisons were made not only against stratum variance estimates computed from the Classification and Mensuration Subsystem (CAMS) segment wheat proportion estimates but also against estimates computed from actual segment wheat proportions for the blind sites. Listed in table 4-4 are these two sets of TY stratum variance estimates. Only refined strata with two or more available CAMS segment proportion estimates are listed. Not listed are eight strata, three of which had one segment.

Suppose S_{jk} is the estimated standard deviation for the jth stratum using the kth method, and σ_j is the TY standard deviation estimate for the jth stratum. Consider the two cases for σ_j (either CAMS or blind sites) and compute the set of differences, $\{(S_{jk} - \sigma_j)\}$, for each method and both cases. The mean and variance of each set of differences are then easily computed. Assuming the difference to be an estimate of the error in estimating the within-stratum variance by a method, then they (i.e., mean and variance for the difference) provide an estimate of the possible bias and the variance expected in estimating a stratum variance using this method. Listed in table 4-5 are the estimated bias and variance for each method as measured against both CAMS and blind site standard deviations. In both cases, bias estimates are consistently positive for all methods. Except for method 7, these estimates are significantly different from zero; with the possible exception of method 7, this approach is likely to overestimate the stratum variance.

Both the bias and the variance estimates are consistently higher for variable r than for the variable p as observed by a comparison of methods 1, 2, and 3 with methods 4, 5, and 7, respectively. As a result, no further consideration of computing stratum variances was given to combinations involving the variable r. For example, combinations of the sampling unit and minimization criterion corresponding to methods 6 and 8 were not tried for the variable r. Next, parameter estimation criterion C-1 (method 4) resulted in higher mean square error estimates than criterion C-3 (method 6). Although criteria C-2 and C-3 competed well in this respect (e.g., the mean square error for method 5 versus that for method 6), it is preferable to choose criterion C-3 rather than C-2 because C-3 gives consideration to the variation in county sizes

19

TABLE 4-4.— REFINED STRATUM VARIANCE ESTIMATES USING TY DATA

State code	Refined stratum	CAMS segment estimates			Ground-truth proportions for blind sites		
		Number of segments	Average wheat proportion	Standard deviation	Number of blind sites	Average wheat proportion	Standard deviation
8	9	3	0.143	0.090	1		
	10	21	.140	.138	6	0.095	0.064
20	7	10	.351	.131	4	.333	.074
	8	7	.302	.044	3	.339	.080
	9	10	.294	.105	3	.355	.040
	11	23	.213	.075	7	.232	.078
	12	21	.255	.105	6	.297	.157
	13	7	.035	.034	2	.028	.001
	14	11	.040	.054	3	.051	.055
	15	3	.284	.121	2	.338	.127
	60	2	.300	.113	0		
	102	7	.026	.038	3	.026	.014
27	15	7	.031	.019	1		
	19	8	.120	.052	3	.097	.064
	20	7	.211	.082	2	.159	.060
30	21	6	.273	.110	2	.259	.108
	22	7	.129	.104	4	.105	.059
	23	6	.245	.078	2	.159	.068
	104	14	.056	.071	2	.063	.024
31	10	4	.305	.194	2	.195	.272
	11	5	.084	.083	3	.091	.040
	14	2	.085	.007	0		
	15	15	.063	.079	4	.051	.073
	16	2	.000	.000	0		
	103	2	.020	.028	1		
38	18	30	.226	.102	9	.257	.089
	20	12	.288	.079	3	.308	.046
	21	34	.156	.098	11	.185	.111
40	3	9	.037	.040	3	.052	.073
	7	25	.365	.160	7	.339	.167
	9	4	.304	.173	1		
	60	7	.167	.095	3	.184	.033
	102	10	.018	.019	3	.022	.022
46	15	3	.021	.012	1		
	16	9	.067	.048	2	.014	.004
	17	4	.049	.086	2	.082	.094
	18	3	.004	.004	1		
	19	5	.070	.069	0		
	21	4	.082	.061	1		
48	2	9	.076	.070	3	.023	.014
	3	8	.043	.042	3	.032	.055
	4	8	.034	.041	3	.051	.044
	5	7	.061	.073	1		
	61	3	.017	.029	1		
	102	5	.038	.050	1		

TABLE 4-5.- THE ESTIMATED BIAS AND VARIANCES IN ESTIMATING STRATA VARIANCES

Method	Blind site ground truth		CAMS segment estimates	
	Bias estimate	Variance estimate	Bias estimate	Variance estimate
1	0.0379	0.00337	0.0274	0.00148
2	.0585	.00397	.0477	.00204
3	.0307	.00278	.0195	.00140
4	.0432	.00256	.0359	.00253
5	.0348	.00295	.0215	.00162
6	.0494	.00219	.0350	.00150
7	.0134 N*	.00200	.0013 N*	.00123
8	.0239	.00200	.0110	.00109

Symbol definition:

CAMS = Classification and Mensuration Subsystem

N* = Insignificant bias when the 5-percent significance t-test is used

21

that is ignored in C-2. Thus, the crop proportion, p, is the variable of choice, and the minimization criterion is C-3.

It should be noted that bias and variance estimates were consistently higher for blind site data than for CAMS data. For variance estimates, this was perhaps due to a much smaller number of blind sites than the number of acquired segments for which CAMS estimates were available. However, higher numbers for the bias estimates reflect that stratum variance estimates were on the average closer to those obtained from the CAMS segment estimates than to those using ground-truth proportions. This implies that the proposed approach is more likely to estimate the total error (i.e., sampling and classification combined) variance than the sampling error variance. Though desirable, this result is somewhat intriguing since no consideration was given to the classification variance while developing this methodology.

The stratum variance estimates produced by this methodology are further influenced by the sampling unit size, x_0, (either field or pixel) used in developing the modeled variance $\sigma^2_{x_0}$. The situation is graphically illustrated in figure 3-1 in section 3. A comparison of the numerical results for methods 6, 7, and 8 shows that the most accurate variance estimates are obtained using the pixel variance model [i.e., equation (12) for $\sigma^2_{x_0}$]. This result was somewhat surprising since better variance estimates were expected from the use of field variance model [i.e., equation (11) for $\sigma^2_{x_0}$] and it may have been due to the sensitivity of method 8 to the poor field size estimates used in the evaluation. The field size estimates computed from the ratio of crop acreages to farm operators were on the average four times larger than field size estimates computed from a limited set of ground truth given by Pitts and Badhwar (ref. 10). Note that a farm operator (accounted for by crop type) may have more than one field of a given crop type, hence, the average field size can be expected to be smaller than the value estimated using equation (14). The numerical results tend to confirm this. Regardless of the method used, the stratum field sizes must be determined and the best possible information should be used for the evaluation. If data on crop statistics and cropping practices from which the field size, f_0, can be estimated is unavailable, then Landsat imagery can be employed to obtain an estimate of average field size for a stratum.

To examine the effect of field size on the stratum variance estimates, similar computations were made using method 6 corresponding to reduced field sizes of $0.5f_0$, $0.25f_0$, $0.1f_0$, $0.05f_0$, and the average field size from Pitts-Badhwar data. The estimated bias and variance resulting from these calculations are listed in table 4-6. From the table, it is noted that bias estimates decreased by two and one-half times as the field was reduced to 5 percent of its original size. Yet, variance estimates show no major change. The case of Pitts-Badhwar corresponds to using a constant value of $0.25f_0$ for the field size in all strata. The reduction in bias associated with field size reduction can be taken as numerical confirmation of the fact that the actual size of sample units having crop proportions either 0 or 1 is substantially smaller than the stratum field size, f_0.

From the derivation of equation (12) given in appendix B, it is observed that an adjustment is made to the variance $\sigma^2_{x_0}$ for the proportions of small squares (pixels) in the strata that are mixed. And, the proportion of mixed squares is a function not only of the stratum crop proportion but also of the stratum field size. Yet, when a field size of $0.25f_0$ was substituted for f_0 in method 7, no change in the variance from the value reported in table 4-5 was observed although a slight reduction in the bias was observed, 0.0009 versus 0.00013. Similarly, the relationship of equation (10) to equation (11) is that of making an adjustment to the variance $\sigma^2_{x_0}$ for a sampling unit equal to the size of an average field to account for the fact that such a sampling unit is expected to contain both crop and noncrop acreage. Since the adjustment factor from equation (10) to equation (11) is a constant multiplier of 4/9, the primary improvement of equation (11) over equation (10) is to reduce the bias. Note in table 4-5 that the bias is considerably less for method 8 in both cases although the reduction in variance is only from 0.00150 to 0.00109 in the case of the CAMS comparison and from 0.00219 to 0.00200 in the case of the ground-truth comparison.

23

TABLE 4-6.- ESTIMATED[a] BIAS AND VARIANCE FOR REDUCED FIELD SIZE FOR METHOD 6

Field size	Bias estimate	Variance estimate
x_0	0.0350	0.00150
$0.5x_0$	.0334	.00192
$0.25x_0$	.0231	.00137
$0.10x_0$	.0176	.00133
$0.05x_0$	.0143	.00131
Pitts-Badhwar (Average field size)	.0231	.00142

[a] Computed in the case of TY CAMS segment estimates.

Listed in table 4-7 are individual stratum standard deviation estimates obtained for methods 7 and 8. The coefficient values of A and B are also given. The comparison between the two sets of estimates shows that, with only four exceptions, the method 8 stratum variance estimates are larger. This result is expected of the methodology, as discussed previously. In addition, an examination of A and B values across the strata suggests that A is significantly influenced by the stratum crop proportion and B is highly dependent upon the between-county variance. (See table 4-3 for information on the stratum crop proportion and the between-county variance.) This indicates that there is a positive correlation between the crop proportion and the value of A, as well as between the value of B and the between-county variance. The correlation is exhibited more in the case of method 7 than in the other method.

It should be noted that the parameter B takes on values between -1 and 0 . When the largest area with crop proportion near 0 or 1 is considered for the sampling unit, the intraclass correlation is near 1 and the stratum variance is close to the binomial form and almost equal to A; therefore, $B \doteq 0$. On the other hand, if the sampling unit is chosen to be a large cluster made of randomly selected elements, the interclass correlation is zero and the stratum variance is equal to A/x, where x is the sampling unit size; therefore, $B \doteq -1$. An intuitive understanding of the observed dependence of B on the between-county variance component is given as follows. Since a smaller between-county variance component is indicative of a possible larger within-county variance component and thus a lower intraclass correlation, it follows that a smaller value for B may be expected when the between-county variance is small.

TABLE 4-7.— WITHIN-STRATUM VARIANCE ESTIMATES FOR METHODS 7 AND 8

State code	Refined stratum	Method 7			Method 8		
		A	B	Standard deviation estimate	A	B	Standard deviation estimate
8	9	0.127	-0.447	0.038	1.716	-0.572	0.074
	10	.108	-.204	.118	.242	-.269	.127
	101	.023	-.273	.039	.058	-.355	.041
20	7	.221	-.215	.160	.289	-.182	.216
	8	.197	-.313	.092	1.124	-.447	.113
	9	.182	-.337	.078	1.825	-.512	.103
	11	.157	-.353	.068	.888	-.456	.095
	12	.162	-.210	.141	.273	-.211	.164
	13	.058	-.320	.048	.109	-.343	.059
	14	.061	-.328	.048	.124	-.381	.052
	15	.189	-.253	.122	.684	-.403	.109
	60	.155	-.408	.051	1.881	-.563	.081
	102	.034	-.527	.013	.204	-.620	.020
27	15	.022	-.332	.028	.035	-.371	.029
	19	.054	-.233	.073	.082	-.293	.066
	20	.166	-.239	.122	.375	-.306	.132
30	21	.172	-.351	.071	2.485	-.565	.093
	22	.098	-.335	.058	.994	-.533	.069
	23	.125	-.248	.102	.532	-.365	.117
	104	.034	-.287	.044	.125	-.397	.048
31	10	.144	-.187	.148	.230	-.221	.158
	11	.076	-.297	.062	.133	-.344	.076
	14	.068	-.362	.042	.179	-.454	.043
	15	.038	-.213	.067	.043	-.225	.067
	16	.003	-.473	.005	.016	-.623	.005
	17	.079	-.242	.083	.220	-.344	.084
	103	.001	-.614	.001	.018	-.865	.002
38	19	.190	-.313	.090	.777	-.389	.125
	20	.210	-.373	.070	1.238	-.459	.111
	21	.147	-.258	.105	.402	-.328	.122
	22	.112	-.248	.096	.285	-.306	.115
40	3	.057	-.321	.047	.166	-.427	.048
	7	.216	-.178	.191	.325	-.216	.193
	9	.150	-.312	.081	.702	-.392	.117
	13	.057	-.270	.062	.084	-.291	.067
	60	.162	-.307	.086	.647	-.389	.114
	102	.022	-.343	.026	.073	-.478	.024
46	15	.009	-.436	.011	.024	-.481	.014
	16	.058	-.199	.089	.097	-.254	.087
	17	.060	-.296	.056	.370	-.453	.063
	18	.042	-.420	.025	.441	-.578	.036
	19	.115	-.270	.087	.258	-.324	.100
	21	.080	-.340	.051	.380	-.426	.073
	104	.031	-.468	.017	.430	-.679	.022
48	2	.028	-.261	.045	.054	-.327	.045
	3	.033	-.264	.048	.058	-.291	.056
	4	.055	-.196	.088	.071	-.203	.096
	5	.101	-.219	.106	.191	-.275	.110
	9	.140	-.237	.113	.321	-.269	.147
	60	.121	-.272	.089	.558	-.396	.102
	61	.060	-.183	.098	.068	-.143	.127
	101	.007	.380	.013	.030	-.484	.015
	102	.011	-.345	.019	.029	-.414	.021

5. CONCLUSION AND SUMMARY

The present study considers several stratum-variance estimation techniques and proposes a new method to obtain initial variance estimates for sample allocations in designing crop surveys. The approach is to develop empirically a relationship between the stratum variance and the sampling unit size.

A procedure is devised that uses existing and easily available information of historical crop statistics in developing this relationship. Consideration is given to the field size in order to effect a modification in stratum variance that is necessary for small sampling unit sizes.

Variance estimation is approached in two ways: (1) estimate the stratum variance for crop proportion directly by developing the empirical model, and (2) first, estimate the stratum variance for the crop to agricultural acreage ratio by developing the empirical model, and then combine this variance estimate with the stratum mean and variance for the agricultural acreage.

The numerical results indicated that the first approach should be preferred because it led to more accurate estimates (when compared with variance estimates obtained from segment data for wheat in USGP) than did the second approach.

In addition, the numerical results tend to show that methods 7 and 8 perform about equally well and that either method produces realistic stratum variance estimates, given reliable input data. However, method 8 is probably more sensitive to the field size variable and should be used if accurate field size determinations can be made. Otherwise method 7 is preferable.

In summary, the study suggests that (1) the technique is viable, (2) care should be exercised to insure the reliability of the input data, and (3) the field sizes must be realistically estimated either from historical statistics or Landsat imagery.

6. REFERENCES

1. Perry, C. R. and Hallum, C. R.: Sampling Unit Size Considerations in Large Area Crop Inventory, Using Satellite-Based Data. NASA/EOD Technical Report, JSC-13767, 1979.

2. Chhikara, R. S.: An Evaluation of Natural Stratification and Sample Allocation Used in Transition Year for the U.S. Great Plains. Lockheed Electronics Company, Inc., (Houston, Texas) Technical Memorandum, LEC-13079, January 1979.

3. Feiveson, A. H; Chhikara, R. S.; and Hallum, C. R.: LACIE Sampling Design. Proceedings of the LACIE Symposium, JSC-16015, vol. 1, July 1979, 6pp.

4. Hallum, C. R.; and Basu. J. P.: Natural Sampling Strategy. Proceedings of the LACIE Symposium, JSC-16015, Appendix B, July 1979, pp. 1010-1013.

5. Smith, H. F.: An Empirical Law Describing Heterogeneity in the Yields of Agriculture Crops. Journal of Agricultural Science, vol. 28, 1938, pp. 1-23.

6. Mahalanobis, P. C.: A Sample Survey of the Acreage Under Jute in Bengal. Sankhya (New Delhi, India), vol. 4, 1940, pp. 511-530.

7. Jessen, R. J.: Statistical Investigation of a Sample Survey for Obtaining Farm Facts. Iowa Agricultural Experimental Station, Research Bulletin 304, 1942.

8. Hansen, M. H. and Hurwitz, W. N.: Relative efficiencies of Various Sampling Units in Population Inquiries. Journal of American Statistics, no. 37, 1942, pp. 89-94.

9. Asthana, R. S. The Size of sub-Sampling Unit in Area Estimation. Indian Council of Agricultural Research (New Delhi, India), 1950, (unpublished thesis).

10. Pitts, D. E. and Badhwar, Gautam: Field Size, Length, and Width Distributions Based on LACIE Ground-Truth Data. Submitted to Remote Sensing of Environment, August, 1979.

APPENDIX A
WITHIN-STRATUM VARIANCE FOR FIELD SIZE SAMPLING UNIT

APPENDIX A

WITHIN-STRATUM VARIANCE FOR FIELD SIZE SAMPLING UNIT

Let f_0 be the acreage field size. Suppose a stratum is divided into square units, each equal to the average field size. In general, a randomly placed sample element consist of areas from four different square units as shown in figure A-1. When the field boundaries are aligned with the grid coordinates and the units are assumed to be independent for the crop of interest, the field crop acreage is given by

$$A = \sum_{i=1}^{4} a_i A_i$$

where

$A_1 = XY$

$A_2 = (1 - X)Y$

$A_3 = (1 - X)(1 - Y)$

$A_4 = X(1 - Y)$

$X \sim u(0,1)$ and $Y \sim u(0,1)$ are two stochastically independent uniform random variables, and the random variables a_i are defined by

$$a_i = \begin{cases} 1, & \text{Prob}[a_i = 1] = P \\ 0, & \text{Prob}[a_i = 0] = 1 - P \end{cases}$$

Then

$$E(A) = \sum_{i=1}^{4} E(a_i A_i)$$

$$= \sum_{i=1}^{4} E(a_i)E(A_i)$$

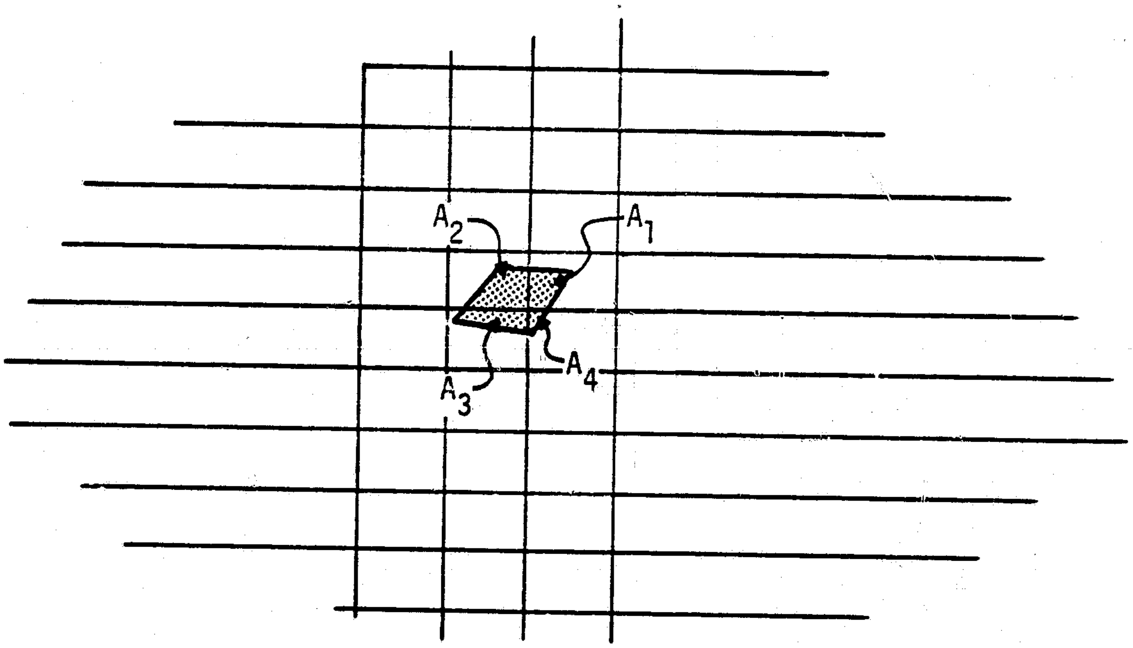


Figure A-1.- Four different square units from a randomly placed sample element.

31

$$= P \sum_{i=1}^{4} E(A_i)$$

$$= PE\left(\sum_{i=1}^{4} A_i\right)$$

$$= P$$

$$\mathrm{Var}(A) = E\left[\mathrm{Var}\left(\sum_i a_i A_i \mid A_i\text{'s}\right)\right] + \mathrm{Var}\left[E\left(\sum a_i A_i \mid A_i\text{'s}\right)\right]$$

$$= E\left[\sum_i A_i^2 \ \mathrm{Var}(a_i)\right] + \mathrm{Var}\left[\sum A_i E(a_i)\right]$$

$$= P(1 - P) \sum_{i=1}^{4} E(A_i^2) + P \ \mathrm{Var}\left(\sum A_i\right)$$

$$= 4P(1 - P)E(A_1^2) + 0$$

since $\sum_{i=1}^{4} E(A_i^2) = 4E(A_1^2)$ due to symmetry.

Next

$$E(A_1^2) = E(X^2Y^2)$$

$$= [E(X^2)][E(Y^2)]$$

$$= \left(\frac{1}{3}\right)^2$$

$$= \frac{1}{9}$$

Thus

$$\mathrm{Var}(A) = \frac{4}{9} P(1 - P)$$

APPENDIX B

WITHIN-STRATUM VARIANCE FOR A VERY SMALL SAMPLING UNIT (PIXEL)

APPENDIX B

WITHIN-STRATUM VARIANCE FOR A VERY SMALL SAMPLING UNIT (PIXEL)

Developed in this appendix is a statistical model for the within-stratum variance for sampling units, which are very small relative to the field size of the crop of interest. Crop X will refer to the crop of interest. The model is developed using the definitions and assumptions in the following conceptual experiment.

A square area unit with diagonal 2d is randomly selected from the area of a stratum having a proportion $\tilde{p}$ for crop X. A random variable P is defined over the sample space of the experiment as follows. P has value p if the randomly selected square has proportion p for crop X. Probabilities α_1, α_2, and α_3 are associated, respectively, with the following events: the square selected is pure and contains only crop X; the square selected is pure and does not contain crop X; and the square selected is mixed. With this notation, it is observed that

$$\alpha_1 = \text{Prob}(P = 1)$$

$$\alpha_2 = \text{Prob}(P = 0)$$

$$\alpha_3 = \text{Prob}(0 < P < 1)$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

$$E(P) = \tilde{p}$$

$$\text{Var}(P) = \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3 E_{P|0<P<1}(P - \tilde{p})^2$$

where the expectation in the last equation is understood to be taken over the collection corresponding to the mixed squares. Tractable analytic expressions for the probabilities α_1, α_2, and α_3 and the expected value $E_{P|0<P<1}(P - \tilde{p})^2$ in terms of the stratum-field-size distribution and the crop proportion, $\tilde{p}$, for crop X will be derived first.

Assume that the stratum has area A and the crop X fields of length l_i and w_i have relative frequencies f_i, i = 1, 2, ···, N. A typical field of crop X is displayed in figure B-1, where b is the expected "width" of a square falling on the field boundary (mixed square). It will be shown later that the average value of 2d cos θ over $0 \leq \theta \leq \pi/4$ gives a reasonable value for b. Since the model derived is for sampling units that are small relative to crop X field sizes, assume that $b << l_i$ and $b << w_i$ for all i and the distance between any two fields of crop X is greater than or equal to b.

To determine the probabilities α_1, α_2, and α_3, first note that the pure crop area and the mixed area associated with a field of length l_i and width w_i are given, respectively, by

$$(l_i - b)(w_i - b)$$

and

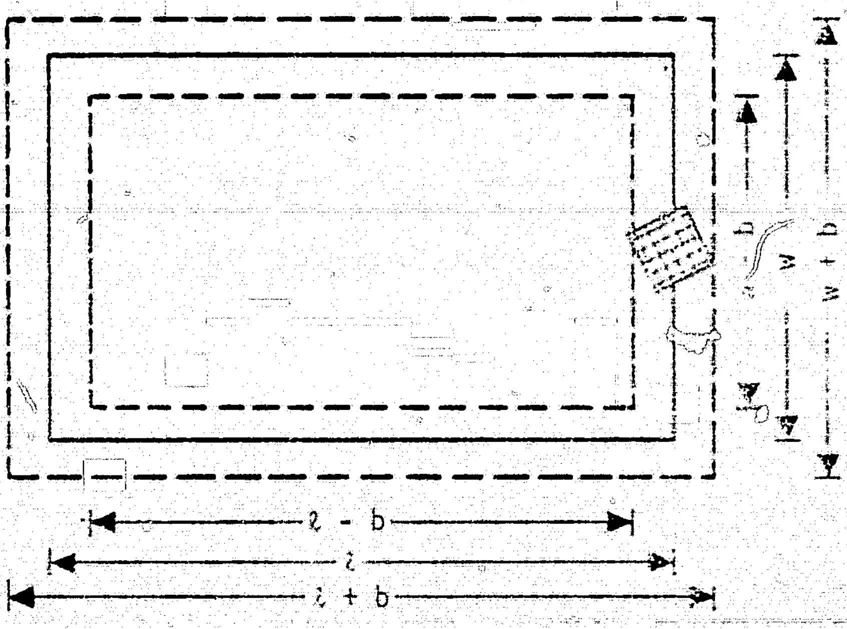
$$(l_i + b)(w_i + b) - (l_i - b)(w_i - b) \tag{B-1}$$

Next note that the total number of fields of length l_i and w_i is given by

$$f_i\left(\frac{\tilde{p}A}{l_i w_i}\right) \tag{B-2}$$

From these equations and the definition of α_1, α_2, and α_3, it follows that

$$\begin{aligned}
\alpha_1 &= \frac{1}{A}\left[\sum_{i=1}^{N}\left(\frac{f_i \tilde{p} A}{l_i w_i}\right)(l_i - b)(w_i - b)\right] \\
&= \tilde{p} \sum_{i=1}^{N} f_i \frac{(l_i - b)(w_i - b)}{l_i w_i} \\
\alpha_3 &= \frac{1}{A}\left\{\sum_{i=1}^{N}\left(\frac{f_i \tilde{p} A}{l_i w_i}\right)[(l_i + b)(w_i + b) - (l_i - b)(w_i - b)]\right\} \\
&= \tilde{p} \sum_{i=1}^{N} \frac{2bf_i(w_i + l_i)}{w_i l_i}
\end{aligned}$$



Legend:

ℓ = length
w = width
b = the expected width of a
square falling on the
field boundary

Figure B-1.- Typical field [illegible]

and

$$\alpha_2 = 1 - \alpha_1 - \alpha_3 \qquad \text{(B-3)}$$

To facilitate the evaluation of $E_{P|0<P<1}(P - \tilde{p})^2$, assume that a square falling on a field boundary is configured as in figure B-2. The directed distance from the center of the square to the field boundary is denoted by x, where x is taken to be positive if the center of the square is not in the field, and x is taken to be negative if the center of the square is in the field. The smallest angle that a diagonal makes with the horizontal is denoted by θ. Now it is easy to see that $|x| \leq d \cos \theta$ and $0 \leq \theta \leq \pi/4$.

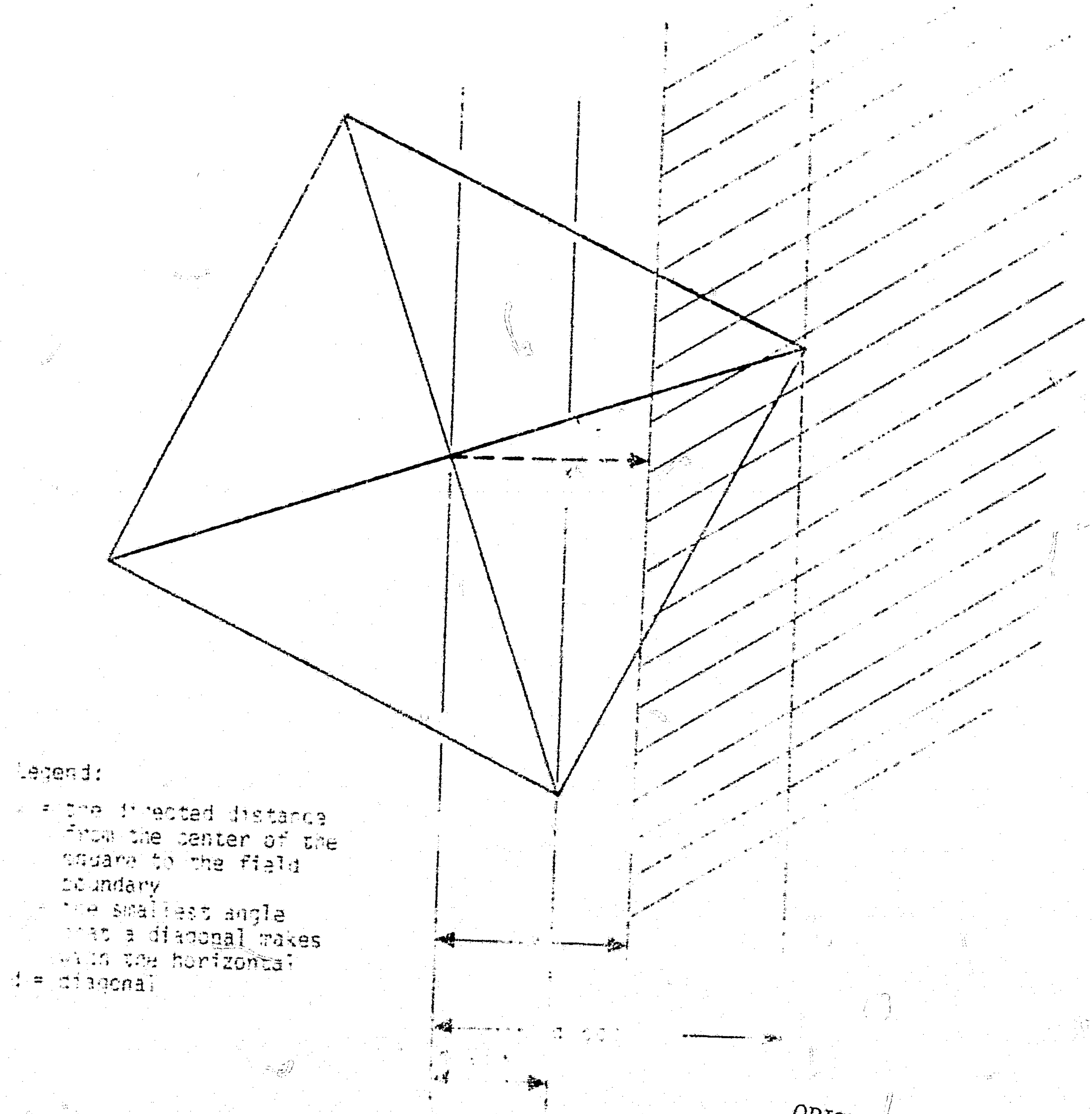
The area of the square contained within the crop field can be expressed as a function of x and θ for $0 \leq \theta \leq \pi/4$ and $0 \leq x \leq d \cos \theta$ using simple geometric observations as follows.

$$A(\theta,x) = \begin{cases} (d \cos \theta - d \sin \theta)[\tan(\pi/4 - \theta) + \tan(\pi/4 + \theta)]\left(\dfrac{d \cos \theta + d \sin \theta}{2} - x\right) \\ \text{for } 0 \leq x \leq d \sin \theta \\ 1/2\ (d \cos \theta - x)^2[\tan(\pi/4 - \theta) + \tan(\pi/4 + \theta)] \\ \text{for } d \sin \theta \leq x \leq d \cos \theta \end{cases} \qquad \text{(B-4)}$$

This formula is readily extended to negative values of x and then adjusted for the total area of the square, A_0, to obtain the following expression for the proportion of the square contained within the crop field.

$$p(\theta,x) = \begin{cases} \dfrac{A(\theta,x)}{A_0} \ ,. \ \text{for } 0 \leq x \leq d \cos \theta \\ 1 - \dfrac{A(\theta,-x)}{A_0} \ , \ \text{for } -d \cos \theta \leq x \leq 0 \end{cases} \qquad \text{(B-5)}$$

Observe that any angle $0 \leq \theta \leq \pi/4$ corresponds to two positions of the square: one where the angle is measured below the horizontal and the other where the angle is measured above the horizontal. Thus, it follows that the first and second moments of P, given $0 < P < 1$, are obtained by the following.



Legend:

[illegible] = the directed distance from the center of the square to the field boundary

[illegible] the smallest angle that a diagonal makes with the horizontal

d = diagonal

[illegible] - Con[illegible]

[illegible]

ORIGINAL PAGE
OF POOR QUALITY

$$E_{P|0<P<1}(P) = 4/\pi \int_0^{\pi/4} \left[\frac{1}{2d \cos \theta} \int_{-d \cos \theta}^{d \cos \theta} P(\theta,x)dx \right] d\theta \qquad \text{(B-6)}$$

$$E_{P|0<P<1}(P^2) = 4/\pi \int_0^{\pi/4} \left\{ \frac{1}{2d \cos \theta} \int_{-d \cos \theta}^{d \cos \theta} [p(\theta,x)]^2 dx \right\} d\theta \qquad \text{(B-7)}$$

The first integral is readily evaluated as follows.

$$E_{P|0<P<1}(P) = 4/\pi \int_0^{\pi/4} \frac{1}{2d \cos \theta} \left\{ \int_{-d \cos \theta}^{0} \left[1 - \frac{A(\theta,-x)}{A_0} \right] dx + \int_0^{d \cos \theta} \frac{A(\theta,x)}{A_0} dx \right\} d\theta$$

$$= 4/\pi \int_0^{\pi/4} \frac{1}{2d \cos \theta} \int_0^{d \cos \theta} dx d\theta$$

$$= 1/2 \qquad \text{(B-8)}$$

Evaluation of the second integral is considerably more involved, requiring several steps. By using elemetry properties of integration and the definition of $p(\theta,x)$, the second integral can be written as follows.

$$E_{P|0<P<1}(P^2) = 4/\pi \int_0^{\pi/4} \frac{1}{2d \cos \theta} \left\{ \int_0^{d \cos \theta} dx - \frac{2}{A_0} \int_0^{d \cos \theta} A(\theta,x)dx + \frac{2}{A_0^2} \int_0^{d \cos \theta} [A(\theta,x)]^2 dx \right\} d\theta \qquad \text{(B-9)}$$

where

$$\int_0^{d \cos \theta} A(\theta,x)dx = d^3(\cos \theta - \sin \theta)[\tan(\pi/4 - \theta) + \tan(\pi/4 + \theta)](1/6 + 1/6 \cos \theta \sin \theta)$$

and

$$\int_0^{d \cos \theta} [A(\theta, x)]^2 dx = \frac{d^5}{12} (\cos \theta - \sin \theta)^2 [\tan(\pi/4 - \theta)$$

$$+ \tan(\pi/4 + \theta)]^2 (3 \cos^2 \theta \sin \theta + \sin^3 \theta)$$

$$+ \frac{d^5}{20} [\tan(\pi/4 - \theta) + \tan(\pi/4 + \theta)]^2 (\cos \theta - \sin \theta)^5$$

Combining these last three equations and then simplfying reduces equation (B-7) for $E_{P|0<P<1}(P^2)$ to the following.

$$E_{P|0<P<1}(P) = 4/\pi \left\{ \pi/8 - \frac{d^2}{3A_0} \left[\int_0^{\pi/4} \frac{d\theta}{\cos \theta(\cos \theta + \sin \theta)} + \int_0^{\pi/4} \frac{(\sin \theta)\ d\theta}{\cos \theta + \sin \theta} \right] \right.$$

$$+ \frac{d^4}{A_0^2} \left[\int_0^{\pi/4} \frac{(\cos \theta + \sin \theta)\ d\theta}{(\cos \theta + \sin \theta)^2} + \frac{1}{3} \int_0^{\pi/4} \frac{(\sin^3 \theta)\ d\theta}{\cos \theta(\cos \theta + \sin \theta)^2} \right]$$

$$\left. + \frac{d^4}{5A_0^2} \int_0^{\pi/4} \frac{(\cos \theta - \sin \theta)^3\ d\theta}{\cos \theta(\cos \theta + \sin \theta)^2} \right\} \qquad \text{(B-10)}$$

Each of the integrals in equation (B-10) can be evaluated by making the substitution θ = Arctan x and then using partial fraction techniques. This yields

$$E_{P|0<P<1}(P^2) = 4/\pi \left\{ \pi/8 - \frac{d^2}{3A_0} \left[(\ln 2) + \left(\pi/8 - \frac{\ln 2}{4} \right) \right] \right.$$

$$+ \frac{d^4}{A_0^2} \left[\left(\pi/8 - \frac{1}{4} \right) + \frac{1}{3} \left(\ln 2 - \frac{1}{4} - \pi/8 \right) \right]$$

$$\left. + \frac{d^4}{5A_0^2} \left[2 - \pi/2 - \frac{3 \ln 2}{2} \right] \right\} \qquad \text{(B-11)}$$

Taking the sampling unit to be one unit square ($A_0 = 1$ and $d = \sqrt{2}/2$ gives the approximation $E_{P|0<P<1}(P^2) \doteq 0.3682$. Using this approximation for $E_{P|0<P<1}(P^2)$ and the expression derived earlier for $E_{P|0<P<1}(P)$ yields the following approximation for Var(P).

$$\mathrm{Var}(P) \doteq \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3(0.3682 - \tilde{p} + \tilde{p}^2) \qquad \text{(B-12)}$$

Taking the width of the band of mixed squares on field boundaries to be the average "width" of a mixed square (fig. B-2) implies that

$$\begin{aligned} b &= 4/\pi \int_0^{\pi/2} 2d \cos \theta d\theta \\ &= \frac{4d\sqrt{2}}{\pi} \\ &= 1.2732 \qquad \text{(B-13)} \end{aligned}$$

This completes the formulas for the probabilities α_1, α_2, and α_3, and hence, the derivation of Var(P).

In summary, for the derivation of Var(P), it has been assumed that the square did not fall on a field corner. This, of course, introduces a slight error. To estimate the magnitude of this error, first note that the probability of a square falling on a corner is given by

$$\alpha_4' = \tilde{p}\sum_{i=1}^{N} \frac{4bf_i}{l_i w_i} \qquad \text{(B-14)}$$

and the probability of a square falling on a field boundary and not on a corner is given by

$$\alpha_3' = \alpha_3 - \alpha_4' \qquad \text{(B-15)}$$

Hence, a more precise equation for Var(P) is

$$\mathrm{Var}(P) = \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha_3'(0.3682 - \tilde{p} + \tilde{p}^2) + \alpha_4' E_c(P - \tilde{p})^2 \qquad \text{(B-16)}$$

where the expectation E_c is understood to be taken over the collection corresponding to the mixed squares that intersect a corner.

It would be very laborous to derive an analytic expression for $E_c(P - p)^2$. However, if θ is assumed to be $\pi/4$ (the case when the sides of the field are parallel to the sides of the square), then it is easy to show that

$$\begin{aligned} E_c(P - \tilde{p})^2 &= E_c(P^2) - 2\tilde{p}E(P) + \tilde{p}^2 \\ &= \frac{1}{9} - \frac{\tilde{p}}{2} + \tilde{p}^2 \qquad \text{(B-17)} \end{aligned}$$

Hence,

$$\mathrm{Var}(P) \doteq \alpha_1(1 - \tilde{p})^2 + \alpha_2\tilde{p}^2 + \alpha'_3(0.3682 - \tilde{p} + \tilde{p}^2) + \alpha'_4\left(\frac{1}{9} - \frac{\tilde{p}}{2} + \tilde{p}^2\right) \qquad \text{(B-18)}$$

For the field sizes and proportion $\tilde{p}$ encountered in this study, equation (B-18) yields values that are within a few percentage points of the values obtained using equation (B-12) for Var(P) derived earlier. Table B-1 gives the relative change encountered using equation (B-18) for Var(P) for the selected proportions $\tilde{p}$ and field sizes S in acres.

TABLE B-1.- VARIANCE OF P FOR SOME COMBINATIONS OF $\tilde{p}$ AND S

S	$\tilde{p}$, 0.01	$\tilde{p}$, 0.10	$\tilde{p}$, 0.20	$\tilde{p}$, 0.30	$\tilde{p}$, 0.40	$\tilde{p}$, 0.50	$\tilde{p}$, 0.60
	Percent						
25	5.0	4.8	3.7	2.6	1.0	-1.7	-7.3
50	2.5	2.2	1.8	1.2	0.5	-0.7	-2.9
100	1.2	1.1	0.9	0.6	0.2	-0.3	-1.3