1958

Statistical sampling for soil mapping surveys

Howard Lewis Taylor
Iowa State College

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Mathematics Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1679

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
STATISTICAL SAMPLING FOR SOIL MAPPING SURVEYS

by

Howard Lewis Taylor

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subjects: Statistics
               Economics

Approved:

Signature was redacted for privacy.

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Signature was redacted for privacy.

Head of Major Departments

Signature was redacted for privacy.

Dean of Graduate College

Iowa State College

1958
**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. REVIEW OF LITERATURE</td>
<td>7</td>
</tr>
<tr>
<td>III. TYPE OF SAMPLING UNIT</td>
<td>12</td>
</tr>
<tr>
<td>IV. OPTIMUM SIZE OF SAMPLING UNIT</td>
<td>15</td>
</tr>
<tr>
<td>A. A Variance Function</td>
<td>15</td>
</tr>
<tr>
<td>B. Homogeneity Considerations</td>
<td>20</td>
</tr>
<tr>
<td>C. A Cost Function</td>
<td>40</td>
</tr>
<tr>
<td>D. The Optimum Unit</td>
<td>51</td>
</tr>
<tr>
<td>V. SAMPLING PROCEDURES</td>
<td>68</td>
</tr>
<tr>
<td>VI. METHODS OF MEASUREMENT</td>
<td>77</td>
</tr>
<tr>
<td>VII. ESTIMATION PROCEDURES</td>
<td>79</td>
</tr>
<tr>
<td>A. Alternative Estimators</td>
<td>79</td>
</tr>
<tr>
<td>B. Pooling of Sample Information from Similar Areas</td>
<td>93</td>
</tr>
<tr>
<td>C. Utilization of Information from Outside the Sample</td>
<td>101</td>
</tr>
<tr>
<td>D. A Subsampling Problem</td>
<td>108</td>
</tr>
<tr>
<td>E. Estimating Present Land Use</td>
<td>112</td>
</tr>
<tr>
<td>VIII. ESTIMATION OF RELATIVE STANDARD ERRORS</td>
<td>116</td>
</tr>
<tr>
<td>IX. DISCUSSION</td>
<td>125</td>
</tr>
<tr>
<td>X. SUMMARY</td>
<td>127</td>
</tr>
<tr>
<td>XI. LITERATURE CITED</td>
<td>129</td>
</tr>
<tr>
<td>XII. ACKNOWLEDGMENTS</td>
<td>132</td>
</tr>
<tr>
<td>XIII. APPENDICES</td>
<td>133</td>
</tr>
<tr>
<td>A. Relationship of the Variance Function to the Cobb-Douglas Production Function</td>
<td>133</td>
</tr>
<tr>
<td>B. The Optimum Unit in Economic Terms</td>
<td>135</td>
</tr>
<tr>
<td>C. A Stochastic Treatment of the Optimum Unit</td>
<td>138</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

Various government agencies are concerned with conservation measures needed to adequately maintain and improve the nation's soil and water resources. Soil conservation activities are centered in the Department of Agriculture, in particular within the Soil Conservation Service, an organization created about twenty-five years ago specifically to develop an effective program of soil conservation.

While conservation measures have been established on millions of acres of land, leaders in the field realize that much remains to be done. They are constantly seeking ways to promote the use of sound conservation practices and better land management.

In the spring of 1956, Secretary of Agriculture Benson took a major step forward in the program of soil conservation. He directed personnel of the Soil Conservation Service and seven other departmental agencies to develop, within the three year period beginning January 1, 1957, a national inventory of soil and water conservation needs. The other agencies were ones whose work is closely related to conservation. They are Agricultural Conservation Program Service, Agricultural Marketing Service, Agricultural Research Service, Commodity Stabilization Service, Farmers Home Administration, Federal Extension Service, and Forest Service. Responsibility for leadership was assigned to the Soil Conservation Service.

The agencies were asked to develop the inventory for each county in the United States and certain subdivisions of the territories. They were also asked to keep the inventory current after its initial development.
In order to determine what conservation measures are needed, it is necessary to have considerable knowledge of the physical characteristics of the land. This concept is based upon the realization that soils are different, have different capabilities, and may need different conservation practices. For conservation use the Soil Conservation Service has developed a system of land classification known as "land use capability classification." The word "capability" refers to the land's capacity for permanent agricultural use considering the degree of hazards and limitations in managing the land. Eight classes have been established. These range from land requiring no special treatment for cultivation to land which is valueless for any use except wildlife or recreation.

Various physical factors are considered in determining the groupings used for the eight capability classes. The most important factors are soil type, per cent of slope, character and degree of erosion, and present land use. Also considered are the climate, wetness of the soil, stoniness, inorganic deposits, and any other significant characteristics. These necessary physical facts needed for classifying land may be obtained from a soil mapping survey made by a trained soil scientist.

Mapping of an area means preparation of a detailed map showing the distribution of soil types, slopes, and erosion conditions. Sometimes other soil conditions are also shown and, in most instances, present land use is indicated. Boundary lines are drawn on the map showing the extent of each significant variation in land. Symbols are used to distinguish the various characteristics and an accurate legend is kept showing the
meaning of all the symbols used. An example of a soil map is shown in Figure 1.

When the soil scientist goes to an area to map it, he ordinarily has an aerial photograph of the area and certain special tools and equipment. The soil scientist walks over the area to be surveyed, boring or digging holes at close enough intervals to be assured of the proper boundary for each mapping unit. He determines the depth of the soil as well as its color, texture, and permeability. He also estimates its moisture-holding capacity, inherent fertility, and organic-matter content. Any other soil characteristics that may relate to its classification and use capabilities are also observed.

The slope is measured with a hand level and the per cent of slope is shown on the map. From knowledge about the original depth of the topsoil at other points in the area, the soil scientist makes an estimate of the soil loss due to erosion. He also notes whether the erosion loss was caused by water or wind. Present use of the land is also noted.

After all the observed physical facts have been recorded on the aerial photograph, the technicians classify the land according to its capabilities. They are then in a position to make estimates of the conservation measures needed because the use and treatment required to keep the soil productive are determined by its capability. In general, certain recommendations are associated with each of the land capability classes.

The ultimate goal of the Soil Conservation Service and other agencies of the United States Department of Agriculture is that soil mapping surveys and estimates of conservation needs be made for all land in the country. In its farm planning work and standard soil survey work the Soil Con-
First symbol: soil name
84 Clyde silty clay loam
198 Floyd loam
395 Kenyon loam
397 Floyd loam, friable substratum variant
398 Clyde silty clay loam, till substratum variant

Second symbol: dominant percent of slope

Third symbol: degree of erosion or topsoil thickness
0 no apparent erosion, over 12 inches of A horizon present
1 slight erosion, 7 to 12 inches of A horizon present

Land use symbols:
P permanent pasture
L cropland
H farmsteads

Figure 1. Soil map for the northwest quarter-section of Section 34, Township 92 North, Range 16 West, Butler County, Iowa
socervation Service has, at one time or another, mapped millions of acres of land. But millions of other acres remain unmapped. Completion of this mapping during the three year period specified for making the conservation needs inventory would not be possible because of the shortage of funds and trained personnel necessary for a major project of this type.

The answer to the problem of obtaining satisfactory results in the specified time period lies in the use of probability sampling methods. If samples are drawn with known probability from a universe, it is possible to make estimates of universe characteristics and, further, to assess, on the basis of the sample, the precision of such estimates.

Recognizing the usefulness of probability sampling methods, the Soil Conservation Service made agreements with the Biometrics Unit at Cornell University and the Statistical Laboratory at Iowa State College to draw samples for all counties in the country; the former institution was designated to draw samples for the thirteen northeastern states and the latter for the remainder of the country, including Alaska, Hawaii, Puerto Rico, and the Virgin Islands.

The research reported in this study includes only that undertaken at the Statistical Laboratory for the specific purpose of determining the optimum sampling design to provide the basic information required for the national inventory of soil and water conservation needs. It is believed, however, that the results may also be useful in other situations requiring soil mapping surveys.

In the present study the principal factors in the optimum sampling design relate to the size of sampling unit and the method of estimation.
The decision regarding the most efficient size of sampling unit is based primarily on evidence about variability between units and information about the cost per unit. In this connection, variance functions, cost functions, and certain homogeneity considerations are examined. The type of sampling unit is considered briefly and a brief discussion on methods of measurement is included.

The present study also examines the problem of estimating population values from the sample data. Considered here are not only the form of the estimator but also the possible use of special information available from sources other than the sample. A short-cut method of estimating relative standard errors is also suggested.
II. REVIEW OF LITERATURE

In general, the many books and articles that have been written about soil conservation in the past few decades lie outside the scope of this study. A few of the books that are more general in scope will be mentioned here primarily for reference purposes.

A comprehensive treatment of soil conservation has been given by Bennett (3). In the first half of his volume Bennett went into considerable detail on the history and nature of the problem of soil erosion including the relationship of erosion to various soil factors and climatic elements. The second part of the book covered the many soil conservation practices which may be used to combat erosion. Special reference was made in this section to particular erosion problems of different parts of the United States.

More recently, Bennett (2) has written a shorter book. In this volume, also, emphasis was placed on the importance of the problem of soil wastage and the principal available methods of soil conservation.

A book by Archer (1), in addition to stressing the various methods of practical conservation, made a special effort to enumerate the public agencies and groups which serve the farmer in his conservation program and to discuss the nature of the financial and technical assistance offered by the agencies.

A book on soil conservation at a slightly more elementary level than the others mentioned so far was that by Butler (4). A special feature of this volume was the detailed treatment of conservation practices applied in each of ten different watersheds throughout the
United States.

Stallings (21) was the author of another recent book on soil conservation. His book emphasized the integration of modern conservation methods. An outstanding feature of his work was the comprehensive list of references accompanying each chapter.

A concise description of the major conservation measures which may be used to reduce erosion and improve soil fertility was given in a recent publication of the United States Department of Agriculture (28).

On the subject of soil surveys, which furnish the information necessary to determine needed conservation measures, by far the most valuable reference was also published by the Department of Agriculture (27). The publication was in the form of a manual describing in considerable detail each step involved in soil classification and mapping. It is generally considered to be the authoritative guide on questions of soil survey procedure.

Only one reference was found relating to the topic of sampling for soil mapping surveys and even this was on an aspect which will not be investigated in the present study. This work was that of Romero (19). He was essentially concerned with a comparison of different line units with one size of area unit while the present study considers only area units. The question of the type of unit will be considered further in Section III.

It should perhaps be made clear at this point that, while the literature abounds with articles on sampling soils to determine average water content, nutrient content, particle size, and similar character-
istics, the present study is not concerned with sampling for purposes such as these. It is concerned only with appropriate sampling methods for soil mapping surveys. Therefore, literature pertaining to other types of soil sampling will not be reviewed.

The literature does contain several studies based on samples drawn for soil mapping surveys. In each case, the author made use of information obtained from samples which were drawn at some previous time and consequently gave no consideration to units having a size other than that of the ones used in his particular case. Therefore, as far as the present work is concerned, these studies merely constitute examples of the use of sampling in connection with soil mapping surveys.

One such study was that of McCart (15), whose primary purpose was to determine whether or not farmers were following suggested practices for intertilled crops on specific soil groups in Tama County, Iowa. A similar study was that of Sutherland (24), who attempted to find out the intensity with which some of the important soils in Shelby County, Iowa, were being farmed and also the level of management practices applied. Prill (18) did research on the effects of present land use on land productivity, crop production and soil loss. His universe of study was Cherokee County, Iowa.

As previously pointed out, one of the major topics for consideration in the present study is the size of sampling unit. One of the important factors influencing the selection of the size of unit is the relationship between variance and unit size. This has been studied empirically by several authors but the first attempt appears to have
been due to Smith (20). The regression equation which he formulated will
be considered in Section IV, Part A.

Mahalanobis (13) independently developed essentially the same equa-
tion as Smith. In a later paper, Mahalanobis (12) expressed the equa-
tion in more general form. Another author to do work in this area was
Taylor (25). He derived an equation for estimating variance for any
size and shape of sampling unit. The equations given by Smith and
Mahalanobis did not take shape of unit into account.

Another equation which takes account of shape of unit was given by
Jessen (10). However, Jessen's equation is for estimating variance be-
tween units within plots rather than the variance between plots. It
should be noted that it is possible to relate these two variances since
the total variance is fixed in any given situation. Another empirical
formulation was given by Hansen et al. (6).

Relationships such as those described in the three previous para-
graphs have come to be generally known as "variance functions." The
origin of this usage is not definitely known, but it appears to be due
to Mahalanobis (13). The most comprehensive general discussion on var-
iance functions seems to be that of Cochran (5).

The second general class of considerations important in determining
the best size of sampling unit is that of cost functions. Smith (20),
Mahalanobis (13), and Jessen (10) apparently were also the first authors
to make use of cost information in determining an optimum size of unit.
The procedure followed by each of these writers was to minimize var-
iance subject to a fixed total cost having a specified functional form.
This procedure has also been employed by other authors. See, for example, King et al. (11).

A comprehensive treatment of the construction of cost functions was given by Hansen et al. (6). McCreary (16) examined in detail factors relating to field costs, usually an important component of cost functions. Additional references on the use of cost functions are Sukhatme (23) and Yates (30).

A reference for some of the estimation problems connected with the present study is the work of Williams (29), who presented a method for obtaining the variance of a post stratified estimator in any probability sampling design. Previous work relating to post stratification pertained only to the case of simple random sampling. See, for example, Cochran (5) and Stephan (22).

Also pertinent to the estimation problems of this study is an article by Mosteller (17). Of special interest is a criterion he presented which, in the present context, is helpful in deciding when to pool data from two similar strata in order to make better estimates for a characteristic in one of the strata. This criterion is a special case of results obtained later by Huntsberger (9) for a larger class of estimation problems.
III. TYPE OF SAMPLING UNIT

It was stated in Section I that one of the major goals of the present study is that of determining the most efficient size of sampling unit to provide the basic information required for the national inventory of soil and water conservation needs. The information desired is what is ordinarily obtained in making a soil mapping survey, namely, data on soil type, slope, erosion, and land use.

The idea of size of unit virtually implies an area unit. Actually, there are other types of units which might reasonably be considered in determining an optimum sampling unit. For example, a sampling unit might consist of a line or a point. Some literature pertaining to line and point units will now be mentioned.

The use of line samples was discussed briefly by Hubback (8) in a paper on estimating crop acreage. His description appears to have been the earliest reference on this type of unit. Some theoretical considerations pertaining to line samples were given later by Mahalanobis (12). Matérn (14) presented approximate formulas for calculating the standard error of a systematically arranged line survey and illustrated their use by numerical examples.

Line sampling was included by Yates (30) in his listing of the various types of samples. He suggested that line sampling might be useful in determining the proportions of a given area which are different types, a use which is closely related to the subject of the present study. Yates wrote only in general terms and did not provide
any numerical examples illustrating the use of line samples.

The use of line units in soil mapping surveys has been investigated by Romero (19). His study, done on a sampling basis, was confined to fifteen selected soils occurring in three Iowa counties. There appear to be some calculating errors in his work so that it is difficult to draw definite conclusions from the data he presented.

Romero (19) also did some work on point samples, considering them as subsamples of line units. He made no attempt to evaluate point samples by comparing them directly with line or area samples.

An example showing the relative precision of area samples and point samples for determining crop acreages was given by Yates (30). The same author also discussed other possible uses of point samples.

While the literature does contain the above references on line and point units, there is, unfortunately, no research available establishing the conditions under which it might be advantageous to use line or point units in soil mapping surveys. Some examination of this point might have been included in the present study but no attempt was made to do this.

The predisposition of the Soil Conservation Service to use an area unit was the principal reason for not considering line or point units in this study. This predisposition was based primarily on two factors. First, area samples had been drawn in Indiana, Iowa, Michigan, Minnesota, Missouri, Ohio, and Wisconsin several years previously for purposes practically identical to those of the present national project. It was felt that the utilization of the information
already obtained from the sample units selected in these states would constitute an economic gain of such a magnitude that it could not be discarded. Therefore, the use of other units in these states was not considered.

The decision to use the area units already sampled in the seven midwestern states established, in effect, a precedent for using area units elsewhere. The problem, then, was to determine the optimum size of unit, recognizing that the result might not be the same unit as the one used in the seven states sampled previously.

The second factor leading to the choice of an area unit is related to the mapping program of the Soil Conservation Service. One of the long-term goals of that agency is to complete a detailed mapping of all the land in the United States. In many states, large blocks of land have been mapped. In some states, even whole counties have been surveyed. In general, however, mapping is confined to scattered tracts of land. These tracts have been mapped at the request of farmers who sought assistance under the farm planning function of the Soil Conservation Service. Whatever the extent of the mapping, it is always done on an area basis; that is, the net result of any mapping effort is that some area, whether it be large or small, is mapped. Obviously, then, the use of an area unit for the present project would be entirely consistent with the regular mapping program of the Soil Conservation Service.
IV. OPTIMUM SIZE OF SAMPLING UNIT

A. A Variance Function

The optimum size of sampling unit has been defined by Cochran (5, p. 189) as "... that which gives the desired precision for the sample estimates at the smallest cost, or the greatest precision for fixed cost." In more general terms, the problem of choosing a sampling unit is one of striking the most effective balance between relative precision and relative cost.

The approach taken in this study regarding an optimum sampling unit is to establish a variance function showing the relationship between variance and unit size, a cost function indicating how costs change with size of unit, and finally, by means of these two functions, an expression indicating the optimum size of unit. The variance relationship is considered first.

Consider a population of \( A \) elements grouped into \( N_1 \) units of \( a_1 \) elements each. Let \( Y_i \) be the acreage of a characteristic of interest for the \( i \)-th unit, where \( i = 1, 2, \ldots, N_1 \). If the population total of the characteristic is denoted by \( Y \), then, by definition,

\[
Y = \sum_{i=1}^{N_1} Y_i \tag{1}
\]

and the mean per unit is given by

\[
\bar{Y}_1 = \frac{Y}{N_1} \tag{2}
\]

The population variance for the \( Y_i \) is defined to be
The population variance between elements is defined to be

\[ S^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (Y_i - \bar{Y})^2 \]  

(3)

where \( \bar{Y} = \frac{Y}{A} \) and the subscript \( i \) now refers to elements.

If a random sample of \( n_1 \) units is drawn from the \( N_1 \) units in the universe, an estimate of \( Y \) is given by

\[ \hat{Y}_1 = \frac{N_1}{n_1} \sum_{i=1}^{n_1} Y_i = N_1 \bar{Y}_1 \]  

(5)

where

\[ \bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i \]

The variance of \( \hat{Y}_1 \), ignoring the finite population correction, is

\[ V(\hat{Y}_1) = \frac{N_1}{n_1} S^2 \]  

(6)

\( S^2_1 \) may be estimated from the sample by

\[ S^2_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_i - \bar{Y}_1)^2 \]  

(7)

Therefore, an estimate of \( V(\hat{Y}_1) \) is

\[ V(\hat{Y}_1) = \frac{N_1}{n_1} S^2_1 \]  

(8)

Now suppose that the \( A \) elements were grouped into \( N_2 \) units of \( n_2 \) elements each. Assume an interest in the same characteristic as before. If \( Y_i \) is the acreage for the characteristic in the \( i \)-th unit
where now $i = 1, 2, \ldots, N_2$ in the population and $i = 1, 2, \ldots, n_2$ in the sample, then the pertinent formulas for estimates and variances are the same as those for units of size $a_1$ with the number 2 inserted in place of 1 wherever the latter occurs as a subscript. For example, the formula for the variance of the estimate of the population total would be $\frac{N_2^2 S_2^2}{n_2}$.

Combinations of the elements into still other sizes of units are possible, of course. What, in general, can be said about the precision expected from a proposed unit relative to the precision given by another size unit? Empirical relationships have been developed for such a situation by Mahalanobis (13), Jessen (10), Hansen et al. (6), and Smith (20).

The equations given by Smith and Mahalanobis were in terms of variance between units. Jessen's formulation was concerned with estimating the variance between elements within units. The suggestion made by Hansen et al. involved the intraclass correlation between elements within units. The various formulas can be related by means of the population analysis of variance.

As far as is known, there is no established superiority of one equation over another. Therefore, Smith's formula was chosen for the present study largely for the sake of convenience.

Smith proposed as a relationship between variance and unit size the regression equation

$$\log V_x = \log V_1 - b^3 \log x \quad (9)$$
where \( V_x \) is the variance of yield per element among units containing \( x \) elements, \( V_1 \) is the variance for elements, and \( b' \) is the regression coefficient. In non-logarithmic form the equation would be

\[
V_x = \frac{V_1}{x^{b'}}
\]  

(10)

The value of \( b' \) depends on the correlation between adjacent elements and lies, in general, between 0 and 1. The equation indicates that the variance of yield per element is ordinarily smaller, the larger the plot.

If \( b' \) is 0, there is perfect correlation between the elements. The equation reduces to \( V_x = V_1 \), indicating that no reduction in variance will be obtained by using larger units. A value of 1 for \( b' \) means that the elements within a plot are completely uncorrelated. Smith checked his relationship on thirty-eight sets of uniformity data and found that nearly all his values of \( b' \) were between 0.2 and 0.8.

In the notation used for the present study, Smith's relationship for units of size \( a_i \) would be

\[
\frac{S_i^2}{a_i^2} = \frac{s^2}{b'}
\]

(11)

or

\[
S_i^2 = S^2 a_i^{2-b'}
\]

(12)

where \( a_i, S^2, \) and \( S_i^2/a_i^2 \) correspond, respectively, to the quantities \( x, V_1, \) and \( V_x \) given by Smith.

As will be shown later, it is slightly more convenient to deal with the quantity \( 1-b' \) than with \( b' \). If \( g \) is substituted for \( 1-b' \), Equation
12 becomes

\[ S_1^2 = S_2^2 a_1^{1+g} \]  \hspace{1cm} (13)

which is the variance function for the present study.

It will often be of interest to compare the relative precision of one unit to another. Such a comparison can be made by means of the variance function. The relative precision (R.P.) of units of size \( a_1 \) to units of size \( a_2 \) is defined to be

\[ \text{R.P.} \left( \frac{a_1}{a_2} \right) = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1)} \]  \hspace{1cm} (14)

Substituting for \( S_1^2 \) and \( S_2^2 \) in the formulas for \( V(\hat{Y}) \) gives the following equations:

\[ V(\hat{Y}_1) = \frac{N_1^2 S_1^2 a_1^{1+g}}{n_1} \]  \hspace{1cm} (15)

and

\[ V(\hat{Y}_2) = \frac{N_2^2 S_2^2 a_2^{1+g}}{n_2} \]  \hspace{1cm} (16)

The relative precision of unit \( a_1 \) to unit \( a_2 \) may then be expressed as

\[ \text{R.P.} \left( \frac{a_1}{a_2} \right) = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1)} = \left( \frac{N_2}{N_1} \right)^2 \left( \frac{n_1}{n_2} \right) \left( \frac{a_2}{a_1} \right)^{1+g} \]  \hspace{1cm} (17)

Equation 17 may be further simplified by substituting \( \frac{A}{a_1} \) for \( N_1 \) and \( \frac{A}{a_2} \) for \( N_2 \) obtaining

\[ \text{R.P.} \left( \frac{a_1}{a_2} \right) = \left( \frac{n_1}{n_2} \right) \left( \frac{a_1}{a_2} \right)^{1-g} \]  \hspace{1cm} (18)
It is in this form that relative precision can be most easily determined.

If there is imposed the condition that the same amount of sampling is done in each case, that is, \( n_1 a_1 = n_2 a_2 \), then the efficiency of units with \( a_1 \) elements relative to units with \( a_2 \) elements is obtained. Specifically, the relative efficiency (R.E.) of \( a_1 \) to \( a_2 \) is given by

\[
R.E. \left( \frac{a_1}{a_2} \right) = \left( \frac{a_2}{a_1} \right)^g
\]

It will be noted that the variance function and the formulas for comparing the precision or efficiency of one unit relative to another all depend upon \( g \). Therefore, before any of these relationships can be utilized, it is necessary to have a value for \( g \). Determination of values of \( g \) will be the next topic considered.

B. Homogeneity Considerations

The empirical relationship between variance and plot size which was suggested by Smith (20) was given in Equation 9. It is

\[
\log V_x = \log V_1 - b' \log x
\]

where \( V_x \) is the variance of yield per element among units containing \( x \) elements, \( V_1 \) is the variance for elements, and \( b' \) is the slope or regression coefficient of \( \log V_x \) on \( \log x \) obtained by the method of least squares. As indicated previously, values of \( b' \) will ordinarily range from 0, when elements in a unit are perfectly correlated, to 1, when elements are completely uncorrelated. In rare instances there may be values of \( b' \) which are greater than 1. This will occur whenever there
is a negative correlation between elements in a unit, that is, when elements in different units tend to be more alike than elements in the same unit.

Smith's equation in terms of the notation of the present study was given in Equation 12. This equation in non-logarithmic form is

$$ S_i^2 = S^2 a_i^{2-b'} $$

With the additional substitution of $g$ for $1-b'$ the form of the relationship was that given by Equation 13,

$$ S_i^2 = S^2 a_i^{1+g} $$

Values of $g$ will ordinarily lie between 0 and 1, being 0 when elements within units are not correlated at all and 1 when elements are perfectly correlated. If there should be a negative correlation between elements, $g$ would have a negative value. Thus, $g$ is a measure of the homogeneity of adjacent elements.

The actual value of $g$ for a given characteristic is determined by the distribution of the characteristic over the universe of inquiry. In any practical situation this distribution will not be known in advance. Furthermore, each characteristic will have a different distribution and hence a different value of $g$. Since $g$ appears in the variance function which will be used to determine the optimum size of unit, it follows that each characteristic would also have its own optimum unit size which would be a function of $g$.

As is often the case in sample survey work, the present study is
also concerned with estimating many characteristics and hence is really a multipurpose survey. While information is desired on various characteristics, it would not be practical to secure the data from many different sampling schemes each involving a unit of different size. It is much more reasonable to employ one sampling plan using a size of unit selected so as to be optimum or near the optimum for a large number of the characteristics of interest. Such a selection would necessarily involve knowledge of values of \( g \).

In order to obtain information about values of \( g \), a study was made for eight widely dispersed counties whose soils had been completely mapped previously. The following counties were included in the study: Taylor, Iowa; Lawrence, Illinois; Forsyth, Georgia; St. Mary, Louisiana; Lynn, Texas; Kit Carson, Colorado; Contra Costa, California; Maricopa, Arizona (Queen Creek Soil Conservation District only). For each of these counties there was available a detailed map showing various soil characteristics. Most of the maps had information on soil types, slopes, and erosion conditions, and, on some of them, the land use and land capability classes were indicated. Refer to Figure 1 on page 4 for a typical soils mapping on an aerial photograph.

If an ideal study procedure could have been followed, each county map would have been divided into small units and the acreage determined in each unit for all characteristics occurring in the county. It would then have been possible to combine the small units into larger units of various sizes for the purpose of determining variability of differently sized units. This procedure could not be carried out because of the
limited time and resources available for this work. The procedure which was actually used was similar but involved a sample of characteristics and a sample of units.

Before the details of the study procedure employed are presented, it will be worthwhile to describe briefly the public land survey systems of the United States. There are essentially two systems in existence. One is known as the system of metes and bounds; the other is a system of rectangular surveys.

The system of metes and bounds is used for nearly all the land in the thirteen original states, Maine, Vermont, West Virginia, Kentucky, Tennessee, and Texas, and for some of the land in Ohio. It is an irregular system in the sense that each parcel of land is described independently and is not tied to any base line. The parcels vary considerably in size.

In all states other than those mentioned above, a rectangular survey system is ordinarily used. This system divides the land into equal-sized townships, sections, and portions thereof. Each township is six miles square and is related to an east and west base line and a north and south meridian. A township is divided into 36 square-mile sections of 640 acres each. A section is usually subdivided into sixteen tracts of 40 acres each. Sometimes a section is not exactly one square mile in area or a township is not exactly 36 square miles because of corrections for survey errors or corrections made necessary by the convergence of meridians as a result of the earth's curvature.

Townships are identified by reference to a base line and a
meridian. Within a township, sections are numbered from 1 to 36 beginning with the section in the northeast corner and following along rows in a serpentine manner. The section in the southeast corner of a township is thus given the number 36. In the present study, for states which were not surveyed in the rectangular system, a simulated structure of townships and sections was imposed.

In the study of the data from the eight counties having soils completely mapped, townships were divided into nine blocks of four sections, each block being two sections by two sections. Two blocks were first selected at random from the nine. Then, within each of the selected blocks two sections were randomly drawn and within each of the two sections two quarter-sections were selected randomly. Quarter-sections were drawn in this manner for every township.

Then, for an arbitrary selection of characteristics in each county, the acreage occurring in each sampled quarter-section was determined. The determination of acreages was facilitated by the use of transparent plastic sheets divided into small squares or grids and superimposed on the county maps. The acreage for each item for each unit was obtained from a count of the number of squares the item occupied in the unit.

In addition to the maps showing the distribution of the soil classifications, there were published county totals available for all characteristics delineated on the maps. It was possible from this published information to easily determine what per cent each item total was of the known total acreage for a county. Characteristics were then
purposely selected to represent varying rates of prevalence. That is, some items were selected from the relatively rare characteristics occupying about 1 per cent of the county; others were selected from a medium frequency class of about 5 per cent, while other characteristics were chosen from those with a relatively high prevalence rate of 10 per cent or larger. The purpose in choosing items in this manner was to provide a check of the validity of any general relationships when applied to the various frequency classes.

The hierarchical structure of the sampling permitted a straightforward analysis of variance for each characteristic. The form of such an analysis is indicated in Table 1. The quantity $s_q^2$ is the variance associated with quarter-sections, and the components of variance $s_s^2$, $s_b^2$, and $s_t^2$ are associated with sections, blocks, and townships, respectively.

Table 1. Sample analysis of variance for a county of 16 townships

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Expected mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between townships</td>
<td>15</td>
<td>$s_q^2 + 2s_s^2 + 4s_b^2 + 8s_t^2$</td>
</tr>
<tr>
<td>Between blocks within townships</td>
<td>16</td>
<td>$s_q^2 + 2s_s^2 + 4s_b^2$</td>
</tr>
<tr>
<td>Between sections within blocks</td>
<td>32</td>
<td>$s_q^2 + 2s_s^2$</td>
</tr>
<tr>
<td>Between quarter-sections within sections</td>
<td>64</td>
<td>$s_q^2$</td>
</tr>
<tr>
<td>Total</td>
<td>127</td>
<td></td>
</tr>
</tbody>
</table>
The number of quarter-sections sampled in each of the eight counties was never smaller than the number included in the analysis given in Table 1. This minimum rate of sampling of two blocks per township, two sections per block, and two quarter-sections per section was done for counties whose maps were available for study for only a limited period. For other counties whose maps were available more or less indefinitely, a larger number of quarter-sections was sampled by selecting additional blocks in each township and then two sections in each block and two quarter-sections in each section. The amount of sampling that was done in each of the eight counties and the total area of each county are given in Table 2.

Table 2. Amount of sampling in the eight study counties

<table>
<thead>
<tr>
<th>County and state</th>
<th>Total area</th>
<th>Number of quarter-sections sampled</th>
<th>Per cent of county sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Square miles</td>
<td>Quarter-sections</td>
<td></td>
</tr>
<tr>
<td>Taylor, Iowa</td>
<td>528</td>
<td>2,112</td>
<td>576</td>
</tr>
<tr>
<td>Lawrence, Illinois</td>
<td>374</td>
<td>1,496</td>
<td>172</td>
</tr>
<tr>
<td>Forsyth, Georgia</td>
<td>243</td>
<td>972</td>
<td>140</td>
</tr>
<tr>
<td>St. Mary, Louisiana</td>
<td>605</td>
<td>2,420</td>
<td>136</td>
</tr>
<tr>
<td>Lynn, Texas</td>
<td>.915</td>
<td>3,660</td>
<td>200</td>
</tr>
<tr>
<td>Kit Carson, Colorado</td>
<td>2,171</td>
<td>8,684</td>
<td>1,200</td>
</tr>
<tr>
<td>Contra Costa, California</td>
<td>276</td>
<td>1,104</td>
<td>290</td>
</tr>
<tr>
<td>Maricopa, Arizona</td>
<td>58</td>
<td>232</td>
<td>232</td>
</tr>
</tbody>
</table>

aData refer only to Contra Costa Soil Conservation District.

bData refer only to Queen Creek Soil Conservation District.
The plan of drawing two sections per selected block and two quarter-sections per section was followed in all counties studied with the exception of Maricopa County, Arizona. In this county the available mapping covered only the portion of the county contained in Queen Creek Soil Conservation District, slightly less than 60 square miles in area. Since the total area of interest was so small, the whole area was included in the study, using 40-acre units in this case.

An analysis of variance of the form shown in Table 1 was calculated in each county for each selected item. Data obtained from each analysis constituted the basis of calculations of a value of \( g \) for that particular characteristic.

The first step in the calculation of a \( g \)-value is the computation of estimates of the variance components \( S_q^2 \), \( S_s^2 \), \( S_b^2 \), and \( S_t^2 \). These estimates, which will be designated by \( s_q^2 \), \( s_s^2 \), \( s_b^2 \), and \( s_t^2 \) respectively, can be obtained by equating the mean squares from the sample data with their corresponding expected mean squares and solving the equations which result.

The estimated variance components may then be used to estimate mean squares between blocks, sections, and quarter-sections where each of these units is considered, in turn, as having been drawn at random within townships. A population analysis of variance is needed for the estimates to be unbiased. Table 3 illustrates this analysis for a county of 16 townships. The unbiased estimate for the mean square between blocks within townships is given by \( s_q^2 + b s_s^2 + 16 s_b^2 \) where values for the variance components are obtained from the analysis of the sample.
Table 3. Population analysis of variance for a county of 16 townships

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Expected mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between townships</td>
<td>15</td>
<td>( s^2_q + 4s^2_s + 16s^2_b + 144s^2_t )</td>
</tr>
<tr>
<td>Between blocks within townships</td>
<td>128</td>
<td>( s^2_q + 4s^2_s + 16s^2_b )</td>
</tr>
<tr>
<td>Between sections within blocks</td>
<td>432</td>
<td>( s^2_q + 4s^2_s )</td>
</tr>
<tr>
<td>Between quarter-sections within sections</td>
<td>1,728</td>
<td>( s^2_q )</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,303</strong></td>
<td></td>
</tr>
</tbody>
</table>

Data. An estimate of the mean square between sections within townships can be obtained by combining the sums of squares for blocks and sections in the population analysis of variance and dividing by the combined degrees of freedom obtaining \( \frac{1}{143} \left( 35s^2_q + 140s^2_s + 128s^2_b \right) \). The estimate for the mean square between quarter-sections within townships resulting from the division of the combined sums of squares for blocks, sections, and quarter-sections by the corresponding combined degrees of freedom is \( \frac{1}{143} \left( 143s^2_q + 140s^2_s + 128s^2_b \right) \).

In order that the estimates for mean squares of different-sized units can be utilized directly in the determination of a value of \( g \), a modification in the variance function given by Equation 13 will now be made. The modification will be made through the use of the equation

\[
S_i^2 = a_i \text{BMS}_i \tag{20}
\]

where \( S_i^2 \) is the variance between units of size \( a_i \) and \( \text{BMS}_i \) is the
"between" mean square for units of size $a_i$. It is here assumed that the analysis of variance has been performed on a per element basis such as that of Table 1.

Substituting $S_i^2$ from Equation 20 into Equation 13 gives the following result:

$$\text{EMSi} = S_i^2 a_i$$  \hspace{1cm} (21)

In logarithmic form Equation 21 is

$$\log \text{EMSi} = \log S_i^2 + g \log a_i$$  \hspace{1cm} (22)

If observations for $\log \text{EMSi}$ are plotted against $\log a_i$, an estimate of $g$ can be obtained by the method of least squares.

The procedure for calculating a value of $g$ will now be illustrated with data for Vona-Valentine loamy sand in Kit Carson County, Colorado. The county consists of 60 townships. By using the sampling plan summarized previously, five blocks were chosen from each township, two sections from each of the five blocks, and two quarter-sections from each section. The number of quarter-sections sampled was thus $(60)(5)(2)(2) = 1,200$. All draws were made randomly. The analysis of variance for the sample data is given in Table 4. The unit of observation was number of acres.

By equating mean squares to expected mean squares, estimates of the components of variance can be obtained. These estimates are $s_q^2 = 136.9$, $s_s^2 = 92.8$, $s_b^2 = 21.8$, and $s_t^2 = 174.6$. The next step is to calculate estimates of the mean squares between blocks, sections, and
Table 4. Analysis of variance for Vona-Valentine loamy sand in Kit Carson County, Colorado

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>Expected mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between townships</td>
<td>59</td>
<td>230,220.5</td>
<td>3,902.0</td>
<td>$s_q^2 + 2s_s^2 + 4s_b^2 + 20s_t^2$</td>
</tr>
<tr>
<td>Between blocks within townships</td>
<td>240</td>
<td>98,358.6</td>
<td>409.8</td>
<td>$s_q^2 + 2s_s^2 + 4s_b^2$</td>
</tr>
<tr>
<td>Between sections within blocks</td>
<td>300</td>
<td>96,766.7</td>
<td>322.6</td>
<td>$s_q^2 + 2s_s^2$</td>
</tr>
<tr>
<td>Between quarter-sections within sections</td>
<td>600</td>
<td>82,155.5</td>
<td>136.9</td>
<td>$s_q^2$</td>
</tr>
<tr>
<td>Total</td>
<td>1,199</td>
<td>507,501.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

quarter-sections as if these units had been randomly drawn within townships. The estimated mean square between blocks within townships is given by $s_q^2 + 4s_s^2 + 16s_b^2 = 136.9 + (4)(92.8) + (16)(21.8) = 857.2$. The estimate for the between sections within townships mean square is

\[
\frac{1}{35} (35s_q^2 + 140s_s^2 + 128s_b^2) = \frac{1}{35} \left[ (35)(136.9) + (140)(92.8) + (128)(21.8) \right] = 588.0.
\]

For quarter-sections within townships the estimated mean square is

\[
\frac{1}{143} (143s_q^2 + 140s_s^2 + 128s_b^2) = \frac{1}{143} \left[ (143)(136.9) + (140)(92.8) + (128)(21.8) \right] = 247.3.
\]

The data necessary for the computation of $g$ are now available. For convenience they are summarized in Table 5. If $Y_i = \log BMS_i$ and $X_i = \log a_i$, then the estimate of $g$ is given by

\[
g = \frac{\frac{1}{n} \sum x_i \bar{y}_i}{\frac{1}{n} \sum x_i^2}
\]
where

\[ \sum_{i=1}^{n} x_i y_i = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right) \]

and

\[ \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \]

For Vona-Valentine loamy sand in Kit Carson County, Colorado, \( n = 3 \),

\[ \sum_{i=1}^{n} y_i = 8.096, \quad \sum_{i=1}^{n} x_i = 8.418, \quad \sum_{i=1}^{n} x_i y_i = .325, \quad \sum_{i=1}^{n} x_i^2 = .725, \]

and \( g = .45 \).

Table 5. Data for \( g \)-value of Vona-Valentine loamy sand, Kit Carson County, Colorado

<table>
<thead>
<tr>
<th>BMS (_i)</th>
<th>( \log BMS (_i) (y(_i)) )</th>
<th>( a(_i) )</th>
<th>( \log a(_i) (x(_i)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>857.2</td>
<td>2.933</td>
<td>2,560</td>
<td>3.408</td>
</tr>
<tr>
<td>588.0</td>
<td>2.769</td>
<td>640</td>
<td>2.806</td>
</tr>
<tr>
<td>247.3</td>
<td>2.393</td>
<td>160</td>
<td>2.204</td>
</tr>
</tbody>
</table>

Values of \( g \) were computed for each selected characteristic in each of the eight counties available for study. Tables 6, 7, 8, 9, 10, 11, 12, and 13 indicate the computed values of \( g \) for the various counties. Also shown in the tables are the rates of prevalence of each characteristic, that is, the per cent of the county occupied by the characteristic. This information will be used later.
Table 6. Prevalence rates and values of g for characteristics in Taylor County, Iowa

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharpsburg silty clay loam, 2-6% slopes</td>
<td>16.6</td>
<td>.44</td>
</tr>
<tr>
<td>Olmitz-Wabash complex, 2-6% slopes</td>
<td>14.4</td>
<td>.35</td>
</tr>
<tr>
<td>Shelby silt loam, 7-11% slopes</td>
<td>7.1</td>
<td>.45</td>
</tr>
<tr>
<td>Lagonda-Clarinda complex, 7-11% slopes</td>
<td>6.3</td>
<td>.52</td>
</tr>
<tr>
<td>Clearfield silty clay loam, 2-6% slopes</td>
<td>2.6</td>
<td>.50</td>
</tr>
<tr>
<td>Winterset silty clay loam, 0-1% slopes</td>
<td>1.9</td>
<td>.70</td>
</tr>
</tbody>
</table>

As may be noted from Tables 6 to 13, values of g for the characteristics examined range from -.09 for soils having 1 to 3 per cent slope in Lynn County, Texas, to .97 for Buxin-Portland-Perry soils in St. Mary Parish, Louisiana.

As previously pointed out, it is necessary, as a practical matter, to use only one sampling scheme and one size of unit for a given universe even if there are many characteristics to be studied. In determining the optimum size of unit, one value of g must be assumed. Therefore, to provide rough measures of the homogeneity associated with each of the eight counties, the median of the values of g obtained for the various characteristics considered in each county was chosen. The
median, rather than the mean, was chosen so that extreme values would
not have a disproportionate effect. The medians are given in Table 14.
The values range from .38 for Lynn County, Texas, to .76 for St. Mary
Parish, Louisiana.

Table 7. Prevalence rates and values of g for characteristics
in Lawrence County, Illinois

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ava silt loam</td>
<td>10.5</td>
<td>.45</td>
</tr>
<tr>
<td>Carmi loam</td>
<td>5.4</td>
<td>.94</td>
</tr>
<tr>
<td>Erosion, silty deposition</td>
<td>1.6</td>
<td>.64</td>
</tr>
<tr>
<td>Erosion, none to slight</td>
<td>79.9</td>
<td>.46</td>
</tr>
<tr>
<td>Erosion, moderate</td>
<td>11.8</td>
<td>.47</td>
</tr>
<tr>
<td>Erosion, severe</td>
<td>6.7</td>
<td>.86</td>
</tr>
<tr>
<td>Slope, 0-1.5%</td>
<td>56.4</td>
<td>.62</td>
</tr>
<tr>
<td>Slope, 1.5-4%</td>
<td>23.1</td>
<td>.56</td>
</tr>
<tr>
<td>Slope, 4-7%</td>
<td>12.5</td>
<td>.31</td>
</tr>
<tr>
<td>Slope, 7-12%</td>
<td>6.1</td>
<td>.78</td>
</tr>
<tr>
<td>Slope, 12-18%</td>
<td>1.6</td>
<td>.63</td>
</tr>
<tr>
<td>Slope, 18-30%</td>
<td>0.3</td>
<td>.17</td>
</tr>
</tbody>
</table>
Table 8. Prevalence rates and values of g for characteristics in Forsyth County, Georgia

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alluvial soils undifferentiated (poorly drained)</td>
<td>1.4</td>
<td>.13</td>
</tr>
<tr>
<td>Cecil clay loam</td>
<td>13.8</td>
<td>.72</td>
</tr>
<tr>
<td>Cecil clay</td>
<td>1.3</td>
<td>.35</td>
</tr>
<tr>
<td>Appling sandy loam</td>
<td>4.5</td>
<td>.43</td>
</tr>
<tr>
<td>Hayesville fine sandy loam</td>
<td>10.6</td>
<td>.84</td>
</tr>
<tr>
<td>Slope, 14-25%</td>
<td>27.4</td>
<td>.41</td>
</tr>
</tbody>
</table>

Table 9. Prevalence rates and values of g for characteristics in St. Mary Parish, Louisiana

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maurepas peat, salt water marsh</td>
<td>10.0</td>
<td>.75</td>
</tr>
<tr>
<td>Alligator clay</td>
<td>7.2</td>
<td>.77</td>
</tr>
<tr>
<td>Baldwin silt loam</td>
<td>2.9</td>
<td>.68</td>
</tr>
<tr>
<td>Buxin-Portland-Perry soils, undifferentiated</td>
<td>5.6</td>
<td>.97</td>
</tr>
</tbody>
</table>
Table 10. Prevalence rates and values of \( g \) for characteristics in Lynn County, Texas

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of ( g )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amarillo fine sandy loam</td>
<td>26.8</td>
<td>.37</td>
</tr>
<tr>
<td>Brownfield fine sand (10 to 18 in. sand)</td>
<td>5.9</td>
<td>.67</td>
</tr>
<tr>
<td>Randall clay</td>
<td>1.3</td>
<td>.20</td>
</tr>
<tr>
<td>Cropland</td>
<td>77.3</td>
<td>.44</td>
</tr>
<tr>
<td>Slope, 0-1%</td>
<td>61.5</td>
<td>.24</td>
</tr>
<tr>
<td>Slope, 1-3%</td>
<td>33.6</td>
<td>-.09</td>
</tr>
<tr>
<td>Erosion, none to slight</td>
<td>83.8</td>
<td>.40</td>
</tr>
<tr>
<td>Erosion, moderate</td>
<td>11.4</td>
<td>.58</td>
</tr>
</tbody>
</table>

It was also deemed of interest to calculate the correlation between values of \( g \) and prevalence rates in order to determine the existence of any relationship such as a tendency for high values of \( g \) to be associated with rare characteristics. Data necessary for the calculations are given in Tables 6 to 13. The correlation coefficients resulting are recorded in Table 15. In no county was the correlation significant at the 5 per cent level of significance. In three counties there was a positive relationship between \( g \) values and prevalence rates while in the other five counties the correlation was negative. A slight over-all negative correlation was indicated.

In connection with the determination of the values of \( g \) there are several other points which should be mentioned. First, it is realized
that values of \( g \) determined in the manner described are only estimates of the true values of \( g \). One reason for this is that only three sizes of unit were included in the analyses. This means that a value of \( g \), the slope of the regression of the logarithm of variance on the logarithm of the size of unit, was calculated from only three points. Also pertinent here is the fact that there may be considerable sampling fluctuation associated with the variance components entering into the estimates of the variance between units since estimates are based on a relatively small sample in most cases.

Table 11. Prevalence rates and values of \( g \) for characteristics in Kit Carson County, Colorado

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of ( g )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land-capability class II</td>
<td>5.0</td>
<td>.67</td>
</tr>
<tr>
<td>Land-capability class VI</td>
<td>14.8</td>
<td>.49</td>
</tr>
<tr>
<td>Pasture and range</td>
<td>50.9</td>
<td>.25</td>
</tr>
<tr>
<td>Ascalon sandy clay loam, thin surface soil phase</td>
<td>1.0</td>
<td>.31</td>
</tr>
<tr>
<td>Vona-Valentine loamy sand</td>
<td>3.1</td>
<td>.45</td>
</tr>
<tr>
<td>Weld silt loam</td>
<td>24.4</td>
<td>.61</td>
</tr>
</tbody>
</table>

A second additional point to be mentioned concerns more refined methods of estimating \( g \). Smith (20), in estimating \( b' \), which is equal to \( 1-g \), suggested that, since the variances were based on different numbers of units, each point should be weighted inversely as its variance. In the regression actually fitted, the dependent variable
is the logarithm of a variance. For the variance of the logarithm of a variance, Smith gives, as a first approximation, the quantity $2/n$ where $n$ is the number of degrees of freedom upon which the estimate of variance is based. According to Smith, then, each observation should be weighted by its corresponding degrees of freedom.

Table 12. Prevalence rates and values of $g$ for characteristics in Contra Costa Soil Conservation District, Contra Costa County, California

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of $g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rincon clay</td>
<td>1.0</td>
<td>.72</td>
</tr>
<tr>
<td>Los Osos loam</td>
<td>0.9</td>
<td>.49</td>
</tr>
<tr>
<td>Altamont clay</td>
<td>10.1</td>
<td>.14</td>
</tr>
<tr>
<td>Clear Lake clay</td>
<td>5.1</td>
<td>.49</td>
</tr>
<tr>
<td>Slope, 25-40%</td>
<td>21.1</td>
<td>.50</td>
</tr>
<tr>
<td>Erosion, moderately severe</td>
<td>14.1</td>
<td>.73</td>
</tr>
</tbody>
</table>

Due to a misunderstanding regarding symbols, a small amount of land having 15-25% slope was inadvertently included in the analysis of this item.

A further refinement is given in a recent paper by Hatheway and Williams (7). The authors point out that variance estimates for different-sized units are frequently highly correlated, since they are based on common components. They present a method which takes this correlation into account by means of weights. The method leads to unbiased estimates having asymptotically minimum variance.

Neither of these methods was seriously considered for use in the
present study because it was believed that efforts expended in this manner would not be warranted in view of the sampling errors referred to above and the fact that the g values obtained were expected to be only rough indications of values that might prevail generally.

Table 13. Prevalence rates and values of g for characteristics in Queen Creek Soil Conservation District, Maricopa County, Arizona

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Prevalence rate (%)</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land-capability class I</td>
<td>39.4</td>
<td>.70</td>
</tr>
<tr>
<td>Land-capability class II</td>
<td>32.8</td>
<td>.68</td>
</tr>
<tr>
<td>Land-capability class III</td>
<td>18.5</td>
<td>.72</td>
</tr>
<tr>
<td>Land-capability class IV</td>
<td>6.9</td>
<td>.68</td>
</tr>
<tr>
<td>Land-capability class VII</td>
<td>1.6</td>
<td>.44</td>
</tr>
<tr>
<td>Land-capability class VIII</td>
<td>0.8</td>
<td>.37</td>
</tr>
<tr>
<td>Glendale silty clay loam</td>
<td>4.3</td>
<td>.48</td>
</tr>
<tr>
<td>Gila sandy loam, moderately deep, over loam and</td>
<td>15.2</td>
<td>.67</td>
</tr>
<tr>
<td>Anthony sand loam, shallow, over loam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazito loam</td>
<td>1.1</td>
<td>.42</td>
</tr>
<tr>
<td>Anthony sandy loam, deep, over Mohave soil material</td>
<td>9.0</td>
<td>.83</td>
</tr>
</tbody>
</table>
### Table 14. Median values of g for the eight study counties

<table>
<thead>
<tr>
<th>County and state</th>
<th>Value of g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taylor, Iowa</td>
<td>.48</td>
</tr>
<tr>
<td>Lawrence, Illinois</td>
<td>.59</td>
</tr>
<tr>
<td>Forsyth, Georgia</td>
<td>.42</td>
</tr>
<tr>
<td>St. Mary, Louisiana</td>
<td>.76</td>
</tr>
<tr>
<td>Lynn, Texas</td>
<td>.38</td>
</tr>
<tr>
<td>Kit Carson, Colorado</td>
<td>.47</td>
</tr>
<tr>
<td>Contra Costa, California</td>
<td>.50</td>
</tr>
<tr>
<td>Maricopa, Arizona</td>
<td>.68</td>
</tr>
</tbody>
</table>

### Table 15. Coefficients of correlation between values of g and prevalence rates for the eight study counties

<table>
<thead>
<tr>
<th>County and state</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taylor, Iowa</td>
<td>-.75</td>
</tr>
<tr>
<td>Lawrence, Illinois</td>
<td>-.12</td>
</tr>
<tr>
<td>Forsyth, Georgia</td>
<td>.34</td>
</tr>
<tr>
<td>St. Mary, Louisiana</td>
<td>.07</td>
</tr>
<tr>
<td>Lynn, Texas</td>
<td>-.11</td>
</tr>
<tr>
<td>Kit Carson, Colorado</td>
<td>-.39</td>
</tr>
<tr>
<td>Contra Costa, California</td>
<td>-.10</td>
</tr>
<tr>
<td>Maricopa, Arizona</td>
<td>.59</td>
</tr>
</tbody>
</table>
Concerning his empirical relationship, Smith (20, p. 21) makes this comment, which is also applicable to the present study: "Since it implies that adjacent areas are equally correlated irrespective of their size and this condition must sooner or later break down, the relationship cannot be extended indefinitely." It does appear to be quite useful, however, for units of a size normally considered.

C. A Cost Function

Ordinarily, the use of a large sampling unit will result in less precise estimates than the use of a small unit. This is indicated, for example, by Equation 19, which gives the relative efficiency of a unit of size $a_1$ compared to a unit of size $a_2$,

$$R.E. \left(\frac{a_1}{a_2}\right) = \left(\frac{a_2}{a_1}\right)^g$$

Suppose $g$ has the value .5 and $a_1$ is four times as large as $a_2$. Then,

$$R.E. \left(\frac{a_1}{a_2}\right) = \left(\frac{a_2}{4a_2}\right)^{.5} = \left(\frac{1}{4}\right)^{.5} = \frac{1}{2}$$

Thus $a_1$ would be one half as efficient as $a_2$.

In the example just considered, the total amount of sampling is assumed to be the same in the two cases, that is, $n_1a_1 = n_2a_2$, where $n_1$ is the number of units of size $a_1$, and $n_2$ is the number of units of size $a_2$. While the sampling of $n_1$ units of size $a_1$ is not expected to yield as precise estimates as the sampling of $n_2$ units of size $a_2$, there is another factor which must be considered in choosing between the two
units. This factor is the cost associated with each unit. It is quite likely that the sampling of $n_2$ units of size $a_2$ will be more costly than the sampling of $n_1$ units of size $a_1$. If the difference in costs is quite large, perhaps the unit of size $a_1$ would be preferred after all. Clearly, then, both a variance function and a cost function must be taken into account in choosing a unit size. It is the purpose of this section to consider a cost function. The section following will then deal with the combination of the variance function and the cost function for the purpose of obtaining the optimum size of unit.

For purposes of this study, a cost function may be defined as a relationship between costs per unit and size of unit. It would seem that a reasonable source of data for the function might be actual cost records of the Soil Conservation Service for mapping operations. However, much of the mapping done by that agency is in a continuous manner across a county. Information from this type of operation is helpful only as a guide; it furnishes no real knowledge on costs per unit. In another type of soil survey, the mapping is done on scattered farms, but in this case the areas mapped vary from a few acres up to a few square miles in extent. Again no data are available for different-sized units.

Another possibility might have been to undertake field operations for the specific purpose of obtaining cost data for units of various sizes. To obtain reliable data using this plan would have been prohibitive in terms of both time and money. Therefore, this approach was not taken.

The plan followed to obtain cost data was to contact soil
scientists and other qualified personnel of the Soil Conservation Service for their best estimates on various cost factors. The most important information which they supplied was for estimated average rates of mapping for different-sized units, considering the type of land to be mapped as well as the size of unit. They also furnished information on time required for traveling, locating specific units, explaining the procedures to farmers, and measuring the acreage of mapping units. Each component of cost will now be discussed briefly.

Data on rates of mapping were obtained in the form of acres per eight-hour day for units of 40 acres, 160 acres, and 640 acres, each for four types of land. Each technician's estimates were independent of estimates made by the others. As might be expected, there was some variation in the answers received for each of the twelve items. For purposes of this study, composite figures were used. The rate used for each item was the median of all estimates for the item. It is interesting to note that for seven of the twelve items the median figure was also the mode, indicating that the distribution of replies was rather symmetric. The median figures are presented in Table 16.

Table 16. Estimated average mapping rates (acres per eight-hour day)

<table>
<thead>
<tr>
<th>Type of land</th>
<th>40 acres</th>
<th>160 acres</th>
<th>640 acres</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigated</td>
<td>80</td>
<td>120</td>
<td>160</td>
</tr>
<tr>
<td>General farming</td>
<td>160</td>
<td>320</td>
<td>600</td>
</tr>
<tr>
<td>Dry farming</td>
<td>200</td>
<td>480</td>
<td>800</td>
</tr>
<tr>
<td>Range</td>
<td>240</td>
<td>640</td>
<td>1,280</td>
</tr>
</tbody>
</table>
It will be noted in the table that for each type of land the mapping rate increases as the size of unit increases. In making their estimates, the soil scientists assumed that they had only a general knowledge of the soil of the area. In each new unit to be mapped, regardless of its size, a certain amount of time must be spent in determining the general soil patterns of the area, prior to the detailed mapping. The fact that this initial period takes proportionately less time as the size of unit increases is reflected in the data. Differences in mapping rates between types of land for a given size of unit are largely due to different intensities of mapping, a standard practice in soil survey work.

Time required for contacting a farmer or rancher and explaining to him the procedures involved constituted the cost of orientation. The average time estimated for orientation was 20 minutes per farm. For units of 40 acres and 160 acres it was assumed that only one farmer would be contacted per unit. For the 640-acre unit the same assumption of one farm per unit was made for the dry farming and range classifications, which occur predominantly in the mountain and western states where farm sizes are generally larger than in other sections of the country. However, for the irrigated and general farming classifications it was assumed that four farmers would need to be contacted for each unit of 640 acres. Thus, for these classifications the orientation time per unit was considered to be 80 minutes.

It was assumed that determining the acreages of the mapping units could be done at a rate of 275 acres per hour. This operation, commonly
known as measuring, is ordinarily done in the office rather than in the field. It is included here because it is an item which will vary according to the size of unit.

There are two kinds of costs related to travel. One is for the time actually spent traveling and the other is a mileage allowance for car use. In this study the latter will be called transportation costs and the former will be designated travel costs.

Travel costs include the time required for two types of travel, daily travel from the home office to the field and return and travel between units. An estimate of distance traveled will be obtained first, and then this quantity will be divided by an assumed rate of travel in order to obtain an estimate of time. The formula used for distance is derived from one given by McCreary (16). It is

\[ d = \sqrt{\frac{L}{k}} \left[ 1 + \frac{k - 1}{\sqrt{n}} \right] \]  

(24)

where \( d \) is the number of miles traveled per sampling unit, \( L \) is the universe area in square miles, \( k \) is the number of units that can be completed in one day, and \( n \) is the total number of units to be done in the county. It is assumed here that \( n \geq k \) and that the second term in the brackets is zero when \( k \leq 1 \).

Some justification for Equation 24 seems appropriate. This will be indicated for the special case of a square county having a systematic arrangement of points or units. It will also be assumed that headquarters are in the center of the county and that there is return to headquarters at the conclusion of each day's work. Slight departures
from these assumptions do not vitiate the result.

Consider a county of $\sqrt{L}$ miles by $\sqrt{L}$ miles. Suppose it is divided into $\frac{n}{k}$ rectangles of equal area. It may be shown that, if travel is always in a "grid" fashion (on a map, movement would always be either in a horizontal or vertical direction, never diagonal), the average distance from the center of the county to the center of a rectangle and return is $\sqrt{L}$ miles. Since there are $n$ units in the county and $k$ may be done in one day, $\sqrt{L}$ represents the average distance out to a day's work and back.

Inside one of the rectangles $k$ points are to be visited in one day. The area of a rectangle is $Lk/n$. The minimum grid distance between $k$ points in a rectangle of area $Lk/n$ may be shown to be $k-1 \sqrt{\frac{L}{n}}$, there being $k-1$ paths each of length $\sqrt{\frac{L}{n}}$. Total travel in a day is thus given by $\sqrt{L} + (k-1) \sqrt{\frac{L}{n}}$. This expression divided by $k$ results in Equation 24.

It may be argued that in going to the center of a rectangle and returning, part of the distance between units will have been covered. This is true. On the other hand, the expression for the minimum distance between points does not allow for return travel. The supposition here is that these two distances will approximately balance each other.

Using Equation 24 a value of $d$ was calculated for each of the twelve combinations of type of land and size of unit. The values assumed for $L$, $k$, and $n$ are given in Table 17, along with the resulting values of $d$. The value of 576 square miles for $L$ is the area of a county of
16 townships and was chosen as being representative of the area of counties falling into the irrigated and general farming classifications. Wheeler County, Nebraska, is an example of a county having this area. For the dry land and range portion of the country \( L \) was assumed to be 2,304 square miles, equivalent to a county of 64 townships. Stutsman County, North Dakota, is such a county.

Table 17. Calculated values of distance traveled per unit (d) and assumed values of quantities needed for the calculation as indicated in Equation 24

<table>
<thead>
<tr>
<th>Type of land</th>
<th>Unit size (acres)</th>
<th>( L )</th>
<th>( k )</th>
<th>( n )</th>
<th>( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigated</td>
<td>40</td>
<td>576</td>
<td>2.00</td>
<td>96</td>
<td>13.22</td>
</tr>
<tr>
<td>Irrigated</td>
<td>160</td>
<td>576</td>
<td>0.75</td>
<td>48</td>
<td>32.00</td>
</tr>
<tr>
<td>Irrigated</td>
<td>640</td>
<td>576</td>
<td>0.25</td>
<td>24</td>
<td>96.00</td>
</tr>
<tr>
<td>General farming</td>
<td>40</td>
<td>576</td>
<td>4.00</td>
<td>96</td>
<td>7.84</td>
</tr>
<tr>
<td>General farming</td>
<td>160</td>
<td>576</td>
<td>2.00</td>
<td>48</td>
<td>13.73</td>
</tr>
<tr>
<td>General farming</td>
<td>640</td>
<td>576</td>
<td>0.94</td>
<td>24</td>
<td>25.53</td>
</tr>
<tr>
<td>Dry farming</td>
<td>40</td>
<td>2,304</td>
<td>5.00</td>
<td>96</td>
<td>13.52</td>
</tr>
<tr>
<td>Dry farming</td>
<td>160</td>
<td>2,304</td>
<td>3.00</td>
<td>48</td>
<td>20.62</td>
</tr>
<tr>
<td>Dry farming</td>
<td>640</td>
<td>2,304</td>
<td>1.25</td>
<td>24</td>
<td>40.36</td>
</tr>
<tr>
<td>Range</td>
<td>40</td>
<td>2,304</td>
<td>6.00</td>
<td>96</td>
<td>12.08</td>
</tr>
<tr>
<td>Range</td>
<td>160</td>
<td>2,304</td>
<td>4.00</td>
<td>48</td>
<td>17.20</td>
</tr>
<tr>
<td>Range</td>
<td>640</td>
<td>2,304</td>
<td>2.00</td>
<td>24</td>
<td>28.90</td>
</tr>
</tbody>
</table>
Values of k were obtained by dividing each number in Table 16 by its corresponding unit size. Values for n were assumed to be 96, 48, and 24 for units of 40, 160, and 640 acres, respectively. It will be observed that these are in the ratio of 4:2:1. This ratio was chosen because some preliminary cost data indicated that unit costs for the three sizes were approximately in the ratio 1:2:4 so that reversing the ratio for the sample size would assure that total costs would be about the same for each of the possible plans. Given this ratio, values of n for the other unit sizes are determined by the choice for one size. The number 48 was selected for 160-acre units based on the sampling referred to in Section III for seven midwestern states.

The components of cost may now be conveniently summarized in tabular form. Tables 18, 19, 20, and 21 present this information in terms of dollars per unit for the four types of land. Values in the tables follow directly from the data already given and a few additional assumptions. One other assumption is that the rate of travel is 35 miles per hour in irrigated and general farming areas, 25 miles per hour in dry farming counties, and 15 miles per hour in the range country. A rate of pay of $2.60 per hour and a mileage payment of 7 cents per mile are also assumed.

Tables 18, 19, 20, and 21 each furnish three points for estimating a cost function indicating how costs change with unit size. A very good linear relationship was found for each of the land types. If \( c_i \) indicates the total cost per sampling unit of size \( a_i \), then each linear relationship will be of the form
\[ c_i = d_o + d_1 a_i \]  

(25)

where \( d_o \) and \( d_1 \) are constants which may be determined by the technique of least squares. Since the cost figures are different for each type of land, the values of each of the fitted constants would be expected to be different in the four equations. The values obtained for \( d_o \) and \( d_1 \) are given in Table 22.

<table>
<thead>
<tr>
<th>Component of cost</th>
<th>Size of sampling unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 acres</td>
</tr>
<tr>
<td>Mapping</td>
<td>$10.40</td>
</tr>
<tr>
<td>Orientation</td>
<td>.87</td>
</tr>
<tr>
<td>Travel</td>
<td>.98</td>
</tr>
<tr>
<td>Measuring</td>
<td>.38</td>
</tr>
<tr>
<td>Transportation</td>
<td>.93</td>
</tr>
<tr>
<td>Total cost</td>
<td>$13.56</td>
</tr>
</tbody>
</table>

As a check on the fit of each linear regression equation the correlation between \( a_i \) and \( c_i \) values was computed. The correlations turned out to be .9996, .9981, .9997, and .9989 for irrigated, general farming, dry farming, and range lands, respectively. All of these values are significant at the 5 per cent level of significance. It was therefore decided that the linear functions adequately represented the data. It may be noted that with a linear function a test of the
Table 19. Costs per sampling unit for general farming land

<table>
<thead>
<tr>
<th>Size of sampling unit</th>
<th>Component of cost</th>
<th>40 acres</th>
<th>160 acres</th>
<th>640 acres</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping</td>
<td>Mapping</td>
<td>$5.20</td>
<td>$10.40</td>
<td>$22.19</td>
</tr>
<tr>
<td>Orientation</td>
<td>Orientation</td>
<td>.87</td>
<td>.87</td>
<td>3.47</td>
</tr>
<tr>
<td>Travel</td>
<td>Travel</td>
<td>.58</td>
<td>1.02</td>
<td>1.90</td>
</tr>
<tr>
<td>Measuring</td>
<td>Measuring</td>
<td>.38</td>
<td>1.51</td>
<td>6.05</td>
</tr>
<tr>
<td>Transportation</td>
<td>Transportation</td>
<td>.55</td>
<td>.96</td>
<td>1.79</td>
</tr>
<tr>
<td>Total cost</td>
<td>Total cost</td>
<td>$7.58</td>
<td>$14.76</td>
<td>$35.40</td>
</tr>
</tbody>
</table>

correlation coefficient is equivalent to a test of the sum of squares due to regression in an analysis of variance. That is, the regression explains a significant part of the total sum of squares.

Equation 25 gives the cost \( c_1 \) per sampling unit of size \( a_1 \). If \( E \) is assumed to represent the total funds available for field work in a county and \( n_1 \) is the number of units of size \( a_1 \) which are sampled, then the complete cost function, apart from overhead costs, may be written as

\[
E = c_1 n_1 \tag{26}
\]

It is here assumed that overhead costs are constant. Under this assumption, \( E \) may be considered to be total expenditure less overhead.

An additional comment about Equation 24 may be helpful. In the development presented here only two values were considered for \( L \). Of
course, any particular value of \( L \) may be used in Equation 24 if more refined cost estimates are desired. Actually, counties vary considerably in their shape as well as their area. Equation 24 works best for square-shaped counties, but it was found to hold quite well for moderate departures from square. If a county is decidedly rectangular, then an improved formula for distance is given by

\[
d = \frac{p + q}{2k} \left[ 1 + \frac{k - 1}{\sqrt{n}} \right]
\]  

(27)

where \( p \) and \( q \) are the length and width of the rectangle and the other quantities are defined as before.

Table 20. Costs per sampling unit for dry farming land

<table>
<thead>
<tr>
<th>Component of cost</th>
<th>Size of sampling unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 acres</td>
</tr>
<tr>
<td>Mapping</td>
<td>$ 4.16</td>
</tr>
<tr>
<td>Orientation</td>
<td>.87</td>
</tr>
<tr>
<td>Travel</td>
<td>1.41</td>
</tr>
<tr>
<td>Measuring</td>
<td>.38</td>
</tr>
<tr>
<td>Transportation</td>
<td>.95</td>
</tr>
<tr>
<td>Total cost</td>
<td>$ 7.77</td>
</tr>
</tbody>
</table>

It should be pointed out that costs of drawing the sample and costs of tabulation have not been included among the components considered here. It has been estimated that the cost per unit of the two items together would constitute only about 2 per cent of the other
Table 21. Costs per sampling unit for range land

<table>
<thead>
<tr>
<th>Component of cost</th>
<th>Size of sampling unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 acres</td>
</tr>
<tr>
<td>Mapping</td>
<td>$3.47</td>
</tr>
<tr>
<td>Orientation</td>
<td>.87</td>
</tr>
<tr>
<td>Travel</td>
<td>2.09</td>
</tr>
<tr>
<td>Measuring</td>
<td>.38</td>
</tr>
<tr>
<td>Transportation</td>
<td>.85</td>
</tr>
<tr>
<td>Total cost</td>
<td>$7.66</td>
</tr>
</tbody>
</table>

Table 22. Values of \(d_0\) and \(d_1\) for cost relationships

<table>
<thead>
<tr>
<th>Type of land</th>
<th>(d_0)</th>
<th>(d_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigated</td>
<td>8.6434</td>
<td>.15349</td>
</tr>
<tr>
<td>General farming</td>
<td>6.5333</td>
<td>.04540</td>
</tr>
<tr>
<td>Dry farming</td>
<td>6.5267</td>
<td>.03770</td>
</tr>
<tr>
<td>Range</td>
<td>6.9284</td>
<td>.02736</td>
</tr>
</tbody>
</table>

costs. Therefore, it is believed that their inclusion would have only a negligible effect. For purposes of this study, they may be considered as part of the constant overhead costs.

D. The Optimum Unit

The approach taken here regarding an optimum unit will be to determine the size of unit which gives estimates with minimum variance
for a certain fixed expenditure. The optimum will first be worked out in general terms. Later, specific values from the cost relationships will be considered.

If the empirical relationship given by Equation 13 is assumed to hold, the variance of the estimate of the population total of a characteristic y is, ignoring the finite population correction, given by Equation 15 for units of size $a_i$. If i replaces l wherever the latter occurs as a subscript, the formula gives the variance for units of size $a_i$. If the additional substitution of $\frac{A}{a_i}$ for $N_i$ is made, the variance takes a form convenient for minimizing, namely,

$$V(\hat{y}_i) = \frac{A^2 S^2}{n_i a_i}$$ \hspace{1cm} (28)

It is now desired to minimize Equation 28, subject to the restrictions given by Equations 25 and 26,

$$c_i = d_o + d_i a_i$$

and

$$E = c_i n_i$$

The minimum may be determined by the use of Lagrangian multipliers. The problem is to minimize the function

$$\phi = \frac{A^2 S^2}{n_i a_i} + \lambda(n_i c_i - E) + \mu(c_i - d_o - d_i a_i)$$ \hspace{1cm} (29)

where $\lambda$ and $\mu$ are Lagrangian undetermined constants and the other
quantities are as defined previously. The solution follows in a straightforward manner if partial derivatives of $\phi$ with respect to $n_i$, $a_i$, $c_i$, $\lambda$, and $\mu$ are taken and the resulting equations are solved for the unknowns. Details will not be presented here. The solution for $a_i$ is

$$a_i = \left(\frac{d_c}{d_i}\right) \left(\frac{1-g}{g}\right)$$

An alternative method of obtaining the general expression for the optimum unit is to solve Equations 25 and 26 for $n_i$, substitute for $n_i$ in Equation 28, and then find the $a_i$ which minimizes the resulting equation. Solving for $n_i$ gives

$$n_i = \frac{E}{d_o + d_1a_i}$$

Substituting for $n_i$ in Equation 28 gives the following equation to be minimized:

$$V(\hat{y}_i) = \frac{A^2S^2(d_o + d_1a_i)}{Ea_i^{1-g}}$$

The value of $a_i$ which makes Equation 32 a minimum can be found by taking the first derivative of $V$ with respect to $a_i$, equating the result to zero, and solving for $a_i$. That is, the solution can be obtained from

$$\frac{dV}{da_i} = \frac{A^2S^2d_o(g-1)}{Ea_i^{2-g}} + \frac{A^2S^2d_1g}{Ea_i^{1-g}} = 0$$

The resulting expression for $a_i$ is the same as Equation 30. Equation 32 can be shown to be a minimum at this point by showing that the second derivative is positive when evaluated at the point. The second
The right-hand side of Equation 34 can be shown to be positive for the optimum $a_1$ for $0 < g < 1$. Details will not be given here.

Equation 30 indicates that the optimum size of unit is a function of $g$ and the two parameters of the cost equation. A similar result was found by Smith (20).

It will be recalled that there is a different cost relationship for each of the land types considered. Hence, each land type will have a different optimum unit depending on the values of $d_0$ and $d_1$ for that land type. As indicated by Equation 30, some value of $g$ must also be selected before an actual optimum unit is obtained. First, however, it will be of interest to determine the optimum unit for each land type assuming various values of $g$. Optimum sizes are given in Table 23 for $g$ values at intervals of .1.

From Table 23 it may be observed that for each of the land types there is an association of small units with large values of $g$ and large units with small values of $g$. For example, the optimum size of unit for a $g$ of .1 is approximately 80 times the size of unit for a $g$ value of .9. It is obvious that with such a wide range in the sizes of unit the further consideration of values of $g$ is an important matter. However, certain other points regarding an optimum unit need to be considered at this time. These will, in fact, be helpful in regard to the question of values of $g$. 

\[
\frac{d^2y}{da_1^2} = \frac{A^2S^2d_0(g-1)(g-2)}{E a_1^{3-g}} + \frac{A^2S^2d_1g(g-1)}{E a_1^{2-g}}
\] (34)
Equation 30 is an expression for \( a_1 \) based on what might be considered purely theoretical considerations, the mathematical minimization of a variance function subject to a cost function. Resulting values are continuous, ranging from 0 to \( \infty \) as \( g \) goes from 1 to 0. It is obvious that, as a practical matter, it would not be reasonable to use either an extremely small or large size which might be indicated by the equation. For example, on practical grounds one might reasonably reject values indicated by Equation 30 for \( g \leq 0.1 \) or \( g \geq 0.9 \).

<table>
<thead>
<tr>
<th>Value of ( g )</th>
<th>Irrigated</th>
<th>General farming</th>
<th>Dry farming</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>393</td>
<td>1,295</td>
<td>1,558</td>
<td>2,279</td>
</tr>
<tr>
<td>0.2</td>
<td>175</td>
<td>576</td>
<td>692</td>
<td>1,013</td>
</tr>
<tr>
<td>0.3</td>
<td>102</td>
<td>336</td>
<td>404</td>
<td>591</td>
</tr>
<tr>
<td>0.4</td>
<td>66</td>
<td>216</td>
<td>260</td>
<td>380</td>
</tr>
<tr>
<td>0.5</td>
<td>44</td>
<td>144</td>
<td>173</td>
<td>253</td>
</tr>
<tr>
<td>0.6</td>
<td>29</td>
<td>96</td>
<td>115</td>
<td>169</td>
</tr>
<tr>
<td>0.7</td>
<td>19</td>
<td>62</td>
<td>74</td>
<td>109</td>
</tr>
<tr>
<td>0.8</td>
<td>11</td>
<td>36</td>
<td>43</td>
<td>63</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>16</td>
<td>19</td>
<td>28</td>
</tr>
</tbody>
</table>

In addition to restricting the utility of the function at both ends of its range, there is another practical limitation that is even more important. This concerns the delineation of the sample units and their location in the field. It will be recalled that information on costs
was originally obtained for only three sizes of unit, 40 acres, 160 acres, and 640 acres. The three sizes were intentionally chosen, because of their relationship to the public land survey system in existence in most of the United States. In this system, described previously in Section IV, Part B, townships are divided into sections of 640 acres each. Section boundaries are nearly always shown on detailed county highway maps and make a rather ideal frame of reference for drawing the sample in the office. Also, in a large portion of the country, section boundaries may be quite easily identified in the field. Appropriate subdivisions of 640-acre units, such as units of 160 acres or 40 acres, may also be handled in both the office and the field in a fairly routine manner. On the other hand, units of 29, 109, or 336 acres, for example, would not be appropriate from the standpoint of either office or field work. This would be true for nearly any size of unit indicated by Equation 30.

For practical reasons, then, the possibilities for an optimum unit were restricted to the three sizes of 40, 160, or 640 acres. Intermediate sizes of 80- or 320-acre units, also easily drawn in the office and identified in the field, were excluded to avoid the question of orientation, a problem which does not exist with the three square units.

It is now obvious that establishing a cost function, while useful in providing a theoretical background for the problem, would not have been necessary in this case in order to decide on an optimum unit. Since the only sizes of unit actually considered are those which were points of observation for the cost data, a decision could have been
reached by direct comparison of the precision obtained from each of the three sizes. It should be noted, however, that the technique of forming a function is required in situations where the practical units which present themselves as alternatives do not correspond to units for which sample data are available.

A comparison of the precision expected from the three sizes of unit is given for the four types of land in Tables 24, 25, 26, and 27, where a unit of 160 acres is considered as the standard of comparison. The equation used for computing the relative precision follows from Equation 18 by substituting for the sample size from the cost relationship of Equation 26. That is, the substitution of \( \frac{E}{c_1} \) for \( n_1 \) and \( \frac{E}{c_2} \) for \( n_2 \) results, for units of size \( a_1 \) and \( a_2 \), in the expression

\[
R.P. = \left( \frac{a_1}{a_2} \right) \left( \frac{c_2}{c_1} \right) \left( \frac{a_1}{s_2} \right)^{1-g}
\]  

(35)

Table 24. Per cent precision of three sizes of unit relative to 160-acre units for irrigated land

<table>
<thead>
<tr>
<th>Value of g</th>
<th>Size of sampling unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>.2</td>
<td>100</td>
</tr>
<tr>
<td>.3</td>
<td>100</td>
</tr>
<tr>
<td>.4</td>
<td>100</td>
</tr>
<tr>
<td>.5</td>
<td>100</td>
</tr>
<tr>
<td>.6</td>
<td>100</td>
</tr>
<tr>
<td>.7</td>
<td>100</td>
</tr>
<tr>
<td>.8</td>
<td>100</td>
</tr>
<tr>
<td>.9</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 25. Per cent precision of three sizes of unit relative to 160-acre units for general farming land

<table>
<thead>
<tr>
<th>Value of $g$</th>
<th>40</th>
<th>160</th>
<th>640</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>56</td>
<td>100</td>
<td>145</td>
</tr>
<tr>
<td>.2</td>
<td>64</td>
<td>100</td>
<td>126</td>
</tr>
<tr>
<td>.3</td>
<td>74</td>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>.4</td>
<td>85</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>.5</td>
<td>97</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>.6</td>
<td>112</td>
<td>100</td>
<td>73</td>
</tr>
<tr>
<td>.7</td>
<td>128</td>
<td>100</td>
<td>63</td>
</tr>
<tr>
<td>.8</td>
<td>148</td>
<td>100</td>
<td>55</td>
</tr>
<tr>
<td>.9</td>
<td>170</td>
<td>100</td>
<td>48</td>
</tr>
</tbody>
</table>

It may be observed from Tables 24 to 27 that with the restriction of possible units to three, the decision about a value of $g$ is now not so critical. That is, the same size of unit may offer the most precision for several values of $g$. In Table 26, for example, a unit of 160 acres is shown to be best for dry farming land if the value of $g$ is .4, .5, or .6. Thus, in general, it is necessary to have only a rough indication of the value of $g$ in order to be fairly sure which unit is optimum.

By substituting a relative precision of 100 per cent in Equation 35 and solving the resulting expression for $g$, it is possible to determine the value of $g$ for which the precision of one unit is equal to that of another. The results of such calculations are shown in Table 28, which may be used to quickly indicate for each land type the preferred choice of unit, given a value of $g$. For dry farming land, for example, a unit
Table 26. Per cent precision of three sizes of unit relative to 160-acre units for dry farming land

<table>
<thead>
<tr>
<th>Value of g</th>
<th>Size of sampling unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td>.1</td>
<td>48</td>
</tr>
<tr>
<td>.2</td>
<td>55</td>
</tr>
<tr>
<td>.3</td>
<td>63</td>
</tr>
<tr>
<td>.4</td>
<td>72</td>
</tr>
<tr>
<td>.5</td>
<td>85</td>
</tr>
<tr>
<td>.6</td>
<td>95</td>
</tr>
<tr>
<td>.7</td>
<td>109</td>
</tr>
<tr>
<td>.8</td>
<td>126</td>
</tr>
<tr>
<td>.9</td>
<td>144</td>
</tr>
</tbody>
</table>

of 40 acres would be preferred if the value of g were greater than .635. For a value less than .377, a 640-acre unit would be preferred. The preferred unit for dry farming land is 160 acres if the g value is between .377 and .635.

Final recommendations on an optimum size of sampling unit can be made if certain values of g are now associated with the four types of land. The best data available for this purpose are the median g values presented in Table 14 for the eight study counties. The only one of the counties which may be classed in the irrigated category is Maricopa County, Arizona, for which the median g value is .68. In checking Table 13 for the g values associated with each characteristic examined, it is observed that none of the ten values is as low as .322, the point
Table 27. Per cent precision of three sizes of unit relative to 160-acre units for range land

<table>
<thead>
<tr>
<th>Value of g</th>
<th>40</th>
<th>160</th>
<th>640</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>44</td>
<td>100</td>
<td>168</td>
</tr>
<tr>
<td>.2</td>
<td>51</td>
<td>100</td>
<td>146</td>
</tr>
<tr>
<td>.3</td>
<td>58</td>
<td>100</td>
<td>127</td>
</tr>
<tr>
<td>.4</td>
<td>66</td>
<td>100</td>
<td>111</td>
</tr>
<tr>
<td>.5</td>
<td>77</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>.6</td>
<td>88</td>
<td>100</td>
<td>84</td>
</tr>
<tr>
<td>.7</td>
<td>101</td>
<td>100</td>
<td>73</td>
</tr>
<tr>
<td>.8</td>
<td>116</td>
<td>100</td>
<td>64</td>
</tr>
<tr>
<td>.9</td>
<td>134</td>
<td>100</td>
<td>55</td>
</tr>
</tbody>
</table>

where a 40-acre and a 160-acre unit would be expected to give the same precision. It seems clear, then, that a unit of 40 acres can be safely recommended as being optimum for irrigated land.

Five counties that were studied fall within the classification of general farming or humid agriculture. These counties are Taylor, Iowa; Lawrence, Illinois; Forsyth, Georgia; St. Mary, Louisiana; and Contra Costa, California, having median g values of .48, .59, .42, .76, and .50, respectively. In this case the question of an optimum unit is not so clear-cut as for irrigated land. Two of the values of g are greater and three values lower than .519, at which point there would presumably be no choice between a unit of 40 acres and one of 160 acres. The three values lower than .519 are relatively closer to it than are the two
Table 28. Values of $g$ for which precision of two units is the same

<table>
<thead>
<tr>
<th>Type of land</th>
<th>Units compared</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 and 160</td>
<td>160 and 640</td>
<td></td>
</tr>
<tr>
<td>Irrigated</td>
<td>.322</td>
<td>.191</td>
<td></td>
</tr>
<tr>
<td>General farming</td>
<td>.519</td>
<td>.369</td>
<td></td>
</tr>
<tr>
<td>Dry farming</td>
<td>.635</td>
<td>.377</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>.691</td>
<td>.475</td>
<td></td>
</tr>
</tbody>
</table>

greater values. However, the median $g$ value for all the thirty-four characteristics studied in the five counties is .51. If there is considered the additional fact that, as shown in Table 2, only 5.6 per cent of St. Mary Parish and 11.5 per cent of Lawrence County were sampled, compared to 14.4 per cent of Forsyth County, 26.3 per cent of Contra Costa County, and 27.3 per cent of Taylor County, and if corresponding median values of $g$ are weighted by these per cents, the resulting average is again .51. The justification for such a weighting is that $g$ values would be expected to be more accurate for counties sampled at the higher rates. While the decision is admittedly a close one, the evaluation of the available evidence seems to indicate a preference for a unit of 160 acres in general farming counties.

Two of the counties studied were in the dry farming classification. They are Lynn, Texas, and Kit Carson, Colorado. Median $g$ values for these two counties are .38 and .47, respectively. Both these values lie in the interval where a unit of 160 acres is preferred. If the two values are weighted by 5.5 and 13.8, respectively, the per
cent of each county sampled as indicated in Table 2, the resulting weighted average is .44. It seems clear that a unit of 160 acres is the optimum for the dry farming category.

Unfortunately, no range county was available for study, so there is no real evidence on g values for this classification. However, in noting the trend in g values from .68 for irrigated land to .51 general farming land to .44 for dry farming land, it seems reasonable to conjecture that the value for range land will be smaller, perhaps about .35. In any event, it seems quite likely that the value of g for range country will be below .475, the point indicated in Table 28 where 160-acre units and 640-acre units would give equal precision. Thus, a unit of 640 acres is recommended for range land.

In order to provide a convenient summary and reference, the results of the investigation on an optimum size of sampling unit are given in Table 29. Also listed there for ready reference are the values of g used in making the final decision regarding an optimum unit.

Table 29. Optimum size of unit and value of g by type of land

<table>
<thead>
<tr>
<th>Type of land</th>
<th>Value of g</th>
<th>Size of unit (acres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigated</td>
<td>.68</td>
<td>40</td>
</tr>
<tr>
<td>General farming</td>
<td>.51</td>
<td>160</td>
</tr>
<tr>
<td>Dry farming</td>
<td>.44</td>
<td>160</td>
</tr>
<tr>
<td>Range</td>
<td>.35</td>
<td>640</td>
</tr>
</tbody>
</table>
The expression given in Equation 35 for the relative precision of two different-sized units can be used to determine the relative precision that would be expected from the use of a non-optimum unit as compared to the optimum. Calculated relative precisions for units of various sizes are given for the four types of land in Tables 30, 31, 32, and 33, assuming values of g given in Table 29. The optimum size of unit as determined from Equation 30 is included in each of the tables and has the relative precision of 100 per cent.

Table 30. Precision of sampling units of selected sizes for irrigated land, relative to the optimum-sized unit when g = .68

<table>
<thead>
<tr>
<th>Size of unit (acres)</th>
<th>Relative precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>91.4</td>
</tr>
<tr>
<td>20</td>
<td>99.2</td>
</tr>
<tr>
<td>26</td>
<td>100.0</td>
</tr>
<tr>
<td>40</td>
<td>98.1</td>
</tr>
<tr>
<td>160</td>
<td>68.1</td>
</tr>
<tr>
<td>640</td>
<td>32.9</td>
</tr>
</tbody>
</table>

As may be seen in Tables 30 to 33, the variance is not highly sensitive to changes in sizes of unit; it ordinarily takes a considerable departure from the optimum size of unit before there is a large loss in precision.

If it is desired to determine the proportional increase in variance
expected from the use of a non-optimum unit, the following equation can be used:

\[
\frac{V(\hat{Y}_1) - V(\hat{Y}_1^*)}{V(\hat{Y}_1^*)} = (g)\left(\frac{a_1}{\bar{a}_1}\right)^{g-1} + (1-g)\left(\frac{a_1}{\bar{a}_1}\right)^g - 1
\] (36)

where the symbol * has been attached to the optimum \(a_1\) and its associated variance and where \(0 < g < 1\). Equation 36 can be derived directly from Equation 35.

Table 31. Precision of sampling units of selected sizes for general farming land, relative to the optimum-sized unit when \(g = .51\)

<table>
<thead>
<tr>
<th>Size of unit (acres)</th>
<th>Relative precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>83.6</td>
</tr>
<tr>
<td>120</td>
<td>99.8</td>
</tr>
<tr>
<td>138</td>
<td>100.0</td>
</tr>
<tr>
<td>150</td>
<td>99.9</td>
</tr>
<tr>
<td>160</td>
<td>99.7</td>
</tr>
<tr>
<td>200</td>
<td>98.3</td>
</tr>
<tr>
<td>400</td>
<td>87.3</td>
</tr>
<tr>
<td>640</td>
<td>76.2</td>
</tr>
</tbody>
</table>
Table 32. Precision of sampling units of selected sizes for dry farming land, relative to the optimum-sized unit when $g = .44$

<table>
<thead>
<tr>
<th>Size of unit (acres)</th>
<th>Relative precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>71.0</td>
</tr>
<tr>
<td>100</td>
<td>92.6</td>
</tr>
<tr>
<td>160</td>
<td>98.7</td>
</tr>
<tr>
<td>200</td>
<td>99.9</td>
</tr>
<tr>
<td>220</td>
<td>100.0</td>
</tr>
<tr>
<td>250</td>
<td>99.8</td>
</tr>
<tr>
<td>300</td>
<td>98.9</td>
</tr>
<tr>
<td>400</td>
<td>95.8</td>
</tr>
<tr>
<td>640</td>
<td>87.9</td>
</tr>
</tbody>
</table>

The "flatness" of the variance function can also be demonstrated using Equation 36. The proportional increase in variance for various ratios of $a_i$ to $a_i^*$ is given in Table 34 for $g = .5$. For this value of $g$, Equation 36 takes the special form

$$\frac{V(\hat{Y}_i) - V(\hat{Y}_i^*)}{V(\hat{Y}_i^*)} = \frac{z + 1}{2\sqrt{z}} - 1$$

(37)

where $z = \frac{a_i}{a_i^*}$. As the table indicates, the proportional increase in variance is small unless the departure from optimum is quite large. For example, the increase in variance is smaller than 10 per cent for
Table 33. Precision of sampling units of selected sizes for range land, relative to the optimum-sized unit when $g = .35$

<table>
<thead>
<tr>
<th>Size of unit (acres)</th>
<th>Relative precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>49.7</td>
</tr>
<tr>
<td>100</td>
<td>74.9</td>
</tr>
<tr>
<td>160</td>
<td>86.9</td>
</tr>
<tr>
<td>400</td>
<td>99.7</td>
</tr>
<tr>
<td>470</td>
<td>100.0</td>
</tr>
<tr>
<td>640</td>
<td>99.0</td>
</tr>
<tr>
<td>800</td>
<td>97.0</td>
</tr>
<tr>
<td>1,000</td>
<td>94.3</td>
</tr>
<tr>
<td>2,000</td>
<td>82.3</td>
</tr>
</tbody>
</table>

\[ 0.4a_i^* < a_i < 2.5a_i^*. \]

The determination of an optimum unit rests upon the validity of the variance function, the cost function, and the estimated degree of homogeneity. The possibility that some of these quantities may be inaccurate and hence may produce an incorrect optimum was discussed by Hansen et al. (6, p. 288) as follows:

\[ \ldots \text{we never know our costs exactly, or the variances or other measures of homogeneity exactly, in advance of drawing the sample. If we are in roughly the right neighborhood in our advance estimates of these values, then we shall achieve results reasonably close to the optimum and will lose very little efficiency by moderate departures from the optimum.} \]

The results given in Tables 30 to 34 seem to substantiate this remark.
Table 34. Proportional increase in variance for selected ratios of $a_i$ to $a_i^*$ when $g = .5$

<table>
<thead>
<tr>
<th>$z = \frac{a_i}{a_i^*}$</th>
<th>Increase in variance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/10 or 10</td>
<td>73.9</td>
</tr>
<tr>
<td>1/5 or 5</td>
<td>34.2</td>
</tr>
<tr>
<td>3/10 or 10/3</td>
<td>18.7</td>
</tr>
<tr>
<td>2/5 or 5/2</td>
<td>10.7</td>
</tr>
<tr>
<td>1/2 or 2</td>
<td>6.1</td>
</tr>
<tr>
<td>3/5 or 5/3</td>
<td>3.3</td>
</tr>
<tr>
<td>7/10 or 10/7</td>
<td>1.6</td>
</tr>
<tr>
<td>4/5 or 5/4</td>
<td>0.6</td>
</tr>
<tr>
<td>9/10 or 10/9</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
</tr>
</tbody>
</table>
In order to provide a background for some of the estimation problems to be considered in Section VII, it seems worth-while to describe the plan actually used in drawing samples for the conservation needs inventory. The description will be given first for a "standard-sized" county of 576 square miles consisting of sixteen townships. Modifications for counties of different size will be indicated later. In states that were surveyed according to the system of metes and bounds, a simulated structure of townships and sections was imposed.

The system of stratification will be considered first. Each regular township of thirty-six sections was divided into three blocks of twelve sections each. Each block consisted of two rows of sections. Thus, one block was composed of sections 1 to 12, another contained sections 13 to 24, while the third block was made up of sections 25 to 36. Each of the blocks so formed was considered a stratum. When the sample units were drawn for a county, they were drawn in such a manner that at least one unit fell in each stratum. This system of stratification thus assured that samples would be distributed over the whole county.

An alternative method of stratification would have been to use as strata irregular areas of land known as land-resource units. It was decided that, in order to maintain low costs and simple procedures for the sample draw, the geographical stratification already described would be employed and that the possible utilization of information contained in the comparison of units occurring in each of the land-
resource units would be investigated at the time of estimation to obtain more precise county estimates if possible. This possibility will be examined in Section VII, Part A.

For this survey the basic rate of sampling chosen for standard-sized counties in the general farming area was one 160-acre unit in forty-eight, or approximately 2 per cent. Since a stratum consisted of twelve sections or forty-eight quarter-sections, the 2 per cent rate was easily achieved by drawing one unit at random from the forty-eight. Thus, there was one unit in each stratum.

The choice of the 2 per cent rate was somewhat arbitrary. According to the methods of Section IV, the number of units and hence the sampling rate are determined by the funds available for the work and the choice of size of unit. Within fairly narrow limits this was the case. Actually, however, no fixed amount of money was allocated to this project and a sampling rate slightly more or less than 2 per cent might have been used.

The final decision on the sampling rate was made after an examination of the amount of error that might be expected from different rates. It was the judgment of personnel of the Soil Conservation Service that, among the sampling rates consistent with general budgetary allowances, the error rates expected from a 2 per cent sample were reasonable for the soil classifications of primary importance in this survey. More information on the subject of errors will be presented in Section VIII.

The basic 2 per cent sample was delineated in red on two copies of each county map. An additional 2 per cent sample, in blue, was also
delineated for each county, so that, if desired, estimates having a higher degree of accuracy could be obtained by mapping units of both colors. One copy of each county highway map was returned to the state concerned, and the other was retained in the office of the survey section of the Statistical Laboratory.

Township boundaries were outlined in green and each was identified by a row number and a column number appearing in the margins. Within each township, sample units were numbered serially within each color classification. The identification 3-5-2R, for example, indicates that the unit is in the third row, the fifth column, and is the second red unit in that township.

Many counties are not regular in shape so that around their borders there often were partial townships, strata, and sections. In such a county, all complete strata were first sampled in the manner already described. Then, for the remainder of the county, strata were formed for every twelve sections, including both full sections and partial sections.

The plan of draw was the same as for the regular strata. In this case, however, the quarter-section drawn, if it happened to fall in one of the partial sections, might not be a full-sized unit or it might be outside the county altogether. These results were accepted in order that the method of sampling be unbiased, giving every point of land in the county the same chance of coming into the sample. Units falling outside the county as a result of this procedure were ignored. The two main consequences of the procedure were that partial units were de-
lineated and that the sampling rate, while still averaging 2 per cent for all counties, varied slightly from 2 per cent for any given county.

An example of a 2 per cent sample showing the numbering system and partial units is given in Figure 2. The second sample of 2 per cent has not been shown in this case.

The actual selection of units to be included in the sample was facilitated by a specially prepared table of random numbers. Each number consisted of two parts, the first part indicating the location of a section and the second part referring to a particular quarter-section. To aid in the delineation of units on the maps, a number of special devices were made. These included special rubber stamps and various kinds of templates.

Up to this point, this section has been concerned with a "standard-sized" general farming county of 576 square miles with sampling units of 160 acres. It is recognized, of course, that many counties are either considerably smaller or larger in size. If a 2 per cent sample of 160-acre units were drawn in a very small county, there would obviously not be as many units for making estimates; hence the estimates would be subject to a somewhat larger error than in a "standard-sized" county. Also, the total expenditure in the county would be less than in the larger counties. On the other hand, a 2 per cent rate applied to a large county would be expected to yield more precise estimates than in a standard-sized county since there would be more units. In this case, however, there would probably be more units than it would be possible to map for the present survey, particularly in states having nearly
Figure 2. Location of sampling units for 2 per cent sample in Bon Homme County, South Dakota.
all large counties, and the total costs would be greater than in a standard-sized county.

The obvious answer to the issues just raised was to adjust the sampling rate according to the size of the county so that every county had approximately the same number of units delineated. Total costs would then also be about the same in each county, under the assumption of constant costs per unit. In fact, the relative standard error of an estimated total, insofar as it depends on sample size, would also be expected to be of the same order of magnitude, regardless of size of county.

A county of 1,152 square miles, if sampled at a 1 per cent rate, would have the same number of units as a county of 576 square miles sampled at 2 per cent. In this case, instead of there being one unit every stratum or every twelve sections, there would be one unit every twenty-four sections. Other rates would require similar adjustments. With a rate of 1/2 per cent there would be one red unit every forty-eight sections. The stratum size thus changed along with the sampling rate.

In order to keep the rates to a relatively small number for convenience in drawing the sample, ten size-classes were established covering all counties with a particular rate of sampling associated with each size-class. Any county with area from 384 to 767 square miles was sampled at 2 per cent. The next larger interval, requiring a 1 per cent sample, was 768 to 1,535 square miles. The complete list of size-classes is given in Table 35. A county whose area came at the
Table 35. Sampling rates (%) which provide standard relative precision for 10 size-classes and 3 sizes of unit

<table>
<thead>
<tr>
<th>Size-class (square miles)</th>
<th>Size of unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 acres</td>
</tr>
<tr>
<td>1 47 and less</td>
<td>16</td>
</tr>
<tr>
<td>2 48 - 95</td>
<td>8</td>
</tr>
<tr>
<td>3 96 - 191</td>
<td>4</td>
</tr>
<tr>
<td>4 192 - 383</td>
<td>2</td>
</tr>
<tr>
<td>5 384 - 767</td>
<td>1</td>
</tr>
<tr>
<td>6 768 - 1,535</td>
<td>1/2</td>
</tr>
<tr>
<td>7 1,536 - 3,071</td>
<td>1/4</td>
</tr>
<tr>
<td>8 3,072 - 6,143</td>
<td>1/8</td>
</tr>
<tr>
<td>9 6,144 - 12,287</td>
<td>1/16</td>
</tr>
<tr>
<td>10 12,288 and over</td>
<td>1/32</td>
</tr>
</tbody>
</table>

midpoint of an interval would receive forty-eight units, while one at the lower end of an interval received thirty-two units, and one at the upper end received sixty-four. Thus, the effect of establishing the ten size-classes was to vary the possible number of units by 50 per cent in either direction from the average number of forty-eight. The plan was quite efficient as far as office procedures were concerned. A clerk, given a new county to sample, had merely to look up the area of the county to know what rate of sampling to use. The exact procedure to follow was clearly specified for each sampling rate.

With reference to Equation 18, which gives the relative precision of units of size $a_1$ to units of size $a_2$, it is easily verified that for
the case of \( g = 0.5 \) equal precision is obtained, with units of 40 acres, from a rate of sampling half as large as that used with 160-acre units. The same statement holds with reference to 160-acre and 640-acre units. It is then possible, if the relative precision obtained from a 2 percent sample of a standard-sized county is regarded as "standard," to indicate the rate of sampling required for each combination of size-class and size of unit to provide standard relative precision. This information is presented in Table 35. In addition to the assumption of a \( g \) value of 0.5, the formulation given in the table assumes that the variance between units is independent of size of county. The effect of the finite population correction is also ignored.

There were instances when, for some logical reason, a size of unit would be preferred which was different from the optimum recommended in Section IV. The proper rate of sampling for such cases was easily obtained from the table. It should be mentioned that in indicating that the sampling rate for each combination of size-class and size of unit will provide estimates with standard relative precision, the assumption of a fixed total cost has to be relaxed slightly. For any size-class the number of units associated with the three rates and corresponding sizes of unit has the ratio of 4 to 2 to 1. Total costs will be constant, or very nearly so, only as costs per unit are then in a ratio of 1 to 2 to 4. The data of Tables 18 to 21 indicate that most of the cost ratios are close to this proportion.

As mentioned previously, the basic universe for sampling was the county. However, in some cases, estimates with "standard" accuracy
were also desired for subuniverses, such as land-resource units, water-sheds, and soil conservation districts. The usual method of handling these cases was to draw supplementary samples, that is, samples in addition to those falling in the area as a part of the standard sampling for the county.

A special problem arose when it was decided that units falling on federally-owned land would not be mapped at this time. In order to compensate for the loss in effective sample size that would have resulted, it was necessary, to achieve a level of "standard" accuracy, to draw a larger sample for the non-federal land. This problem was usually handled by considering only the area of the non-federal land in initially determining the rate of sampling applicable to that county.
VI. METHODS OF MEASUREMENT

This section contains brief descriptions of the different methods now in use for measuring or determining acreage of mapping units. No attempt will be made to evaluate or compare the various methods.

The problem is, given a detailed mapping such as was shown in Figure 1, to determine the acreage of each separation or mapping unit. In general, of course, mapping units are quite irregular in shape; thus areas ordinarily cannot be determined by using formulas applicable to well-known geometric configurations.

One method of dealing with irregular map areas is to use a planimeter, an instrument especially designed for measuring areas. A pointer of this instrument is passed around the boundary of the region, and from the readings on the dials the area of the region may be determined. The method has been found to be fairly accurate but is rather slow and costly.

Another method is known as the weight apportioning method, or the method of "cutting and weighing." So that the original field sheet may be retained intact, a photostatic copy is made to use in the determination of the acreages. This is first trimmed to include only the area that has been surveyed, that is, the aggregate of all mapping units. This trimmed sheet is then weighed, the total weight thus corresponding to the total area. The sheet is then cut up into individual mapping units. Each of these may be weighed if desired, but ordinarily all units having like symbols are weighed together since it is the total area of each classification which usually is of interest. In any
event, the area associated with each weighing is, of course, obtained by multiplying the weight by a conversion factor based on the total area and the total weight. The method has been widely used.

The counting of "grids" is another method. In this method, a plastic sheet containing small grids or squares is placed over the map of the region. The area of any unit is obtained from the count of the number of grids that cover the unit.

A method quite similar to the counting of "grids" is the counting of "dots." Instead of squares on the plastic sheet there are dots and the area of any mapping unit is approximately determined from a count of the dots falling in the unit. The methods of grid count and dot count are both quite popular.

A relatively new device for measuring irregular areas is the electric area calculator. A transparent grid is placed over the map of the region and the whole area is traversed with a marking stylus along parallel guide lines. The resulting area appears in a counter. The instrument is claimed to be both accurate and fast.

One of the methods described was used for determining the acreage of each mapping unit in each sampling unit. This work, like the soil surveys, was done by personnel of the Soil Conservation Service. Information on each mapping unit was entered on one line of a form designed for recording the data from each sampling unit. The aggregate of forms for a county contained the data for making estimates of totals of various characteristics of interest.
VII. ESTIMATION PROCEDURES

A. Alternative Estimators

A summary of the sample design was given in Section V. The procedure that was employed for drawing the sample is known as stratified random sampling. That is, each county was divided into strata and a sample was drawn from each stratum independently of the draws in all other strata.

If there are $M$ strata and the number of units in the $j$-th stratum is $N_j$ in the population and $n_j$ in the sample, an estimate of $Y$ the population total for a characteristic $y$ is given by

$$\hat{Y} = \sum_{j=1}^{M} N_j \bar{y}_j$$

where $\bar{y}_j$ is the mean of the $n_j$ observations in the $j$-th stratum. The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{j=1}^{M} \frac{N_j^2}{n_j} s_j^2 \left( \frac{N_j - n_j}{N_j} \right)$$

where $s_j^2$ is the variance between the $N_j$ units in the $j$-th stratum. $V(\hat{Y})$ is estimated by

$$V(\hat{Y}) = \sum_{j=1}^{M} \frac{N_j^2}{n_j} s_j^2 \left( \frac{N_j - n_j}{N_j} \right)$$

where $s_j^2$ is the variance between the $n_j$ sample units in the $j$-th stratum.

When the sampling rate is small, the finite population corrections can be ignored. The formula for the estimate of variance then becomes
Since the sampling rate was the same in all strata, the survey design falls in the class of designs known as stratification with proportional allocation. With proportional allocation, Equation 41 can be expressed in the simple form

\[ v(\hat{Y}) = \frac{N}{n} \sum_{j=1}^{M} \frac{N_j s_j^2}{n_j} \quad (41) \]

Actually, in this survey the allocation was equal because the strata were equal in size \((N_j = \frac{N}{M})\). In this case, the equation for the estimate of variance is

\[ v(\hat{Y}) = \left( \frac{N^2}{n} \right) \left( \frac{1}{M} \right) \sum_{j=1}^{M} s_j^2 \quad (42) \]

Whether Equation 40, 41, 42, or 43 is used for estimating variance, it is necessary that there be at least two units per stratum. Specifically, \(s_j^2\) cannot be calculated unless \(n_j \geq 2\). Therefore, in counties having \(n_j = 1\), none of these equations can be used to estimate the variance.

One possibility leading to an approximation of the variance would be to assume that units in a county constitute a simple random sample. This would be equivalent to ignoring the stratification. Under this plan, \(Y\) could be estimated by

\[ \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i \quad (44) \]
and the estimated variance of $\hat{Y}$ would be

$$ v(\hat{Y}) = \frac{N^2 s^2}{n} \left( \frac{N-n}{N} \right) \quad (45) $$

where $s^2$ is the variance between the $n$ sample units in the county.

Usually the finite population correction can be ignored and Equation 45 becomes

$$ v(\hat{Y}) = \frac{N^2 s^2}{n} \quad (46) $$

Before concluding that this method of estimating would be generally appropriate, it seemed reasonable to determine, if possible, the magnitude of the gain in precision expected from stratification relative to simple random sampling. If the gain were small, the assumption of simple random sampling would seem justified under the circumstances. If a large gain were indicated, then it would be desirable to find a method which did not ignore the gain, but instead took it into account.

To study this question, two counties were selected from among the few having more than one unit per stratum. The counties were Johnson, Nebraska, and Trousdale, Tennessee. In these counties, the characteristics of slope, land use, and land capability class were considered and in Trousdale County erosion was also analyzed. The initial step was to calculate for each item in each county an analysis of variance similar to the one given in Table 36 for Slope 1 in Johnson County. Data for each unit were in acres.

Now with proportional allocation and with the additional assumption that the variance between units in a stratum is the same for all strata,
it is possible to use information from the analysis of variance to estimate the gain in precision from stratification as compared to simple random sampling. For the exact method of doing this, see Cochran (5, p. 102). To a very close approximation the estimated precision of stratified sampling relative to random sampling is given by the ratio of the total mean square in the analysis of variance to the mean square for between units within strata, the latter mean square being an estimate of the assumed constant value of $S_j^2$. In Table 36 this would be $1,876.91$ divided by $1,761.47$ or $1.066$. That is, for Slope 1 in Johnson County, Nebraska, there was a 6.6 per cent gain due to stratification.

Table 36. Analysis of variance for Slope 1, Johnson County, Nebraska

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between strata</td>
<td>31</td>
<td>61,762.82</td>
<td>1,992.35</td>
</tr>
<tr>
<td>Between units within strata</td>
<td>31</td>
<td>54,605.50</td>
<td>1,761.47</td>
</tr>
<tr>
<td>Total</td>
<td>62</td>
<td>116,368.32</td>
<td>1,876.91</td>
</tr>
</tbody>
</table>

In Johnson and Trousdale counties an analysis of variance and the estimated gain due to stratification were calculated for 49 characteristics. The per cent gain ranged from a loss of 10.7 per cent for Woodland in Johnson County to a gain of 130.7 per cent for Slope F in Trousdale County. The over-all average gain for the 49 characteristics was 12.9 per cent. The fact that the amount of gain varies considerably from item to item is a common thing, often encountered in sample
surveys. The computed values are, of course, only estimates of the true gains in precision. The true gains would not, in general, vary to the extent indicated by the sample data. In particular, there is seldom a loss in true precision as a result of stratification.

While the average gain from stratification is rather moderate in magnitude, it is large enough that, for the case $n_j=1$, some modification of the formulas for stratified sampling would generally be preferred over using formulas based on simple random sampling. One possibility would be to regard the townships as strata, with the assumption that the three units falling in each township had been drawn randomly from the whole township. A between unit within township estimate of variance would then be possible.

To check on the possibility of this scheme, the data from Johnson County, Nebraska, were analyzed further. For each item, the between strata variation was broken into two parts, one part associated with townships and one with strata within townships. Table 37 gives this revised analysis of variance for Slope 1.

The combination of the between strata and between units mean squares, weighted by their respective degrees of freedom, results in a mean square between units within townships. If there were evidence that the mean square between units within townships differed but little from the mean square between units within strata, then the use of the former in counties having $n_j=1$ would seem justified.

For each item in Johnson County, the between units within townships mean square, obtained in the manner indicated above, was compared with
the mean square between units within strata. For Slope 1, for example, 
the mean square between units within townships was 1,768.30 which is 0.4 
per cent larger than 1,761.47, the mean square between units within 
strata. For all characteristics, the former mean square was 2.3 per 
cent larger, on the average, than the latter mean square.

Table 37. Revised analysis of variance for Slope 1, Johnson County, 
Nebraska

| Source of variation          | Degrees of freedom | Sum of squares | Mean  
square |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between townships</td>
<td>10</td>
<td>24,416.49</td>
<td>2,441.65</td>
</tr>
<tr>
<td>Between strata within townships</td>
<td>21</td>
<td>37,346.53</td>
<td>1,778.40</td>
</tr>
<tr>
<td>Between units within strata</td>
<td>31</td>
<td>54,605.50</td>
<td>1,761.47</td>
</tr>
<tr>
<td>Total</td>
<td>62</td>
<td>116,368.32</td>
<td>1,876.91</td>
</tr>
</tbody>
</table>

The evidence from Johnson County thus indicates that the within 
strata mean square can be approximated quite satisfactorily by the with­
in township mean square. For this county it provided a slight over­
estimate, a result which would be expected to hold generally.

In order to obtain evidence on the gain due to stratification by 
township, beyond that obtainable from the data for Johnson County, data 
from five counties having \( n_j = 1 \) were analyzed. The characteristics con­
sidered were slope, erosion, land use, and land capability class for the 
following counties: Foster, North Dakota; Harvey, Kansas; Guilford, 
North Carolina; and Coffee, Alabama. In addition, in the first two 
counties and in Jefferson County, Iowa, soil type was analyzed.
An analysis of variance, similar to that shown in Table 38 for land-capability class VI in Foster County, North Dakota, was computed for each item. In each case an estimated gain from stratification was also obtained. The values ranged from a loss of 21.8 per cent for Erosion 1 in Guilford County, North Carolina, to a gain of 249.9 per cent for Clinton silt loam in Jefferson County, Iowa. The average gain for all 212 items analyzed in these counties was 13.2 per cent.

Table 38. Analysis of variance for Land-capability class VI, Foster County, North Dakota

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between townships</td>
<td>17</td>
<td>11,970.76</td>
<td>704.16</td>
</tr>
<tr>
<td>Between units within townships</td>
<td>36</td>
<td>16,883.33</td>
<td>468.98</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>28,854.09</td>
<td>544.42</td>
</tr>
</tbody>
</table>

In these five counties, as well as in Johnson County, Nebraska, and Trousdale County, Tennessee, the moderate gains from stratification were usually for items with moderate or large prevalence rates. The large gains, most of the small gains, and most of the losses were for items with small prevalence rates.

The available evidence seems to indicate that, if there is only one sampling unit per stratum, a fairly good estimate of variance, if desired, can be obtained by grouping the actual strata to form larger strata and then computing the variance as if the sample units had been drawn randomly in each of the large strata. In counties where the rate
of sampling was 2 per cent, the township seems satisfactory as the large
stratum. For smaller rates of sampling an area larger than a township
will be needed. With a 1 per cent sampling rate, for example, two
townships would provide three units. For sampling rates larger than 2
per cent there will be at least two units per stratum and a grouping of
strata will not be necessary.

Equation 43 should be a satisfactory one to use in making estimates
of variance. If no grouping of strata is necessary, M in Equation 43
represents the number of actual strata; when the original strata are
grouped, M indicates the number of strata after grouping.

If only a rough estimate of variance is desired, stratification can
be ignored completely and the calculation can be done according to Equa-
tion 46, the variance formula for simple random sampling. In general,
the computation of a variance can be done somewhat faster using Equation
46 than using Equation 43. If many calculations of variance are re-
quired, it may be advantageous to use Equation 46 and then, if desired,
take account of the empirical evidence that variances calculated in this
manner are from 10 to 15 per cent too large, on the average. For many
purposes, the expected difference of 10 to 15 per cent between the two
methods may not matter and, in this event, the simpler method can be
used.

The preceding discussion has dealt with the variance for a strati-
fied sample compared with the variance of a simple random sample. It
should be pointed out that with proportional allocation there is no
difference between the estimators themselves. In this special type
of stratification the estimator given by Equation 38 can be shown to be equivalent to the estimator given by Equation 44. The latter will usually be easier to compute.

Another estimator was also considered. As described in Section IV, Part B, there was some variation in the size of the units of the public land survey with the result that sampling units, drawn as some specified portion of the survey unit, also varied slightly in their size. Also, it will be recalled that in Section V it was mentioned that partial sampling units were occasionally delineated. In order to take account of this variability in the total size of units, a ratio estimator with total acreage as the auxiliary variate was examined.

If \( Y_i \) is the acreage in the \( i \)-th unit for a characteristic \( y \), \( X_i \) is the total acreage of the \( i \)-th unit, and \( X = \sum_{i=1}^{N} X_i \) is the total acreage for the universe under consideration, then an estimate of \( Y \), the population total for the \( Y_i \)'s, is given by

\[
\hat{Y} = \hat{R} X \quad (47)
\]

where

\[
\hat{R} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} \quad (48)
\]

Simple random sampling is assumed, with \( n \) sampling units being drawn randomly from a total of \( N \). \( R \) is an estimate of \( \frac{Y}{X} \).

An approximate formula for the variance of \( \hat{Y} \) is given by

\[
V(\hat{Y}) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (Y_i - \hat{R} X_i)^2 \quad (49)
\]
The variance may be estimated from the sample data by

$$v(\hat{Y}) = \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (Y_i - \hat{R} X_i)^2$$  \hspace{1cm} (50)

When the finite population correction is ignored, Equation 50 becomes

$$v(\hat{Y}) = \frac{N^2}{n(n-1)} \sum_{i=1}^{n} (Y_i - \hat{R} X_i)^2$$  \hspace{1cm} (51)

The ratio estimator and its variance will be discussed in more detail in Section VII, Part E.

The ratio estimator of Equation 47 was used to estimate the county totals for thirty-six soil types in Jefferson County, Iowa. The variances were computed according to Equation 51 and compared with those obtained from using Equation 46. The relative precision of the ratio estimator compared to the simple expansion estimator was calculated for each item. These varied from 98.1 per cent to 101.0 per cent and averaged 99.9 per cent. Since variance estimates were, on the average, practically identical for the two methods, there was clearly no reason to use the ratio estimator for achieving a gain in precision.

It should be mentioned that of the thirty-six sample units delineated in Jefferson County, Iowa, only five had a total acreage different from 160 acres and their acreages differed from 160 only by a few acres. The correlation between the $Y_i$ and $X_i$ values was therefore small and in such a case there is usually no advantage to using the ratio estimator.

To study the question further, however, it was thought desirable to check the gains from the ratio estimator in a county having a
larger number of units different from 160 acres; Caldwell County, Kentucky, having fourteen such units out of sixty-three, was selected. A comparison similar to that described for Jefferson County, Iowa, was made for estimates of the acreage in the eight land-capability classes. The relative precision of the ratio estimator compared to simple expansion ranged from 99.4 per cent to 114.4 per cent. The average was 105.5 per cent. Even in this county, selected because it was believed that the gain from the ratio estimator might be at or near its maximum, the gain was not large.

Equation 51 is considerably more tedious to compute than Equation 46. Considering this and the available evidence regarding the results from the two equations, it is doubtful if variance calculation by Equation 51 is ever warranted for this survey except, perhaps, in a few unusual cases.

On the other hand, the estimator itself, given by Equation 47, was found to be quite useful. Its results were about the same as those from Equation 44 and, after \( \sum_{i=1}^{n} X_i \) was obtained, it was just as easy to compute. Its chief advantage, however, stemmed from the fact that the estimates it produced for any given class of characteristics added to the known universe total. For example, there are eight land-capability classes and each mapping unit can be classified as belonging to one of the eight classes. If Equation 47 is used to estimate a county total for each of the eight classes, the estimates will then add to \( X \), the total universe acreage. This was a very desirable feature for this survey.
One other estimator was also examined. It was a post stratified estimator with land-resource units as the post strata. A land-resource unit is a geographic area of land, at least several thousand acres in extent, characterized by a particular combination or pattern of soils (including slope and erosion), climate, water resources, land use, and types of farming. The boundaries of these areas were known or could be determined so that each sampling unit could be classified according to the land-resource unit to which it belonged. The total acreage in each land-resource unit was also known.

If it is assumed that the sample as originally drawn is a simple random sample, then the form of the post stratified estimator is the same as for stratification in advance of sampling. This estimator was given in Equation 38. However, the variance is slightly larger under post stratification, since there is a contribution from the variability in sample size for each of the post strata.

It can be shown that the equation for the estimated variance of the post stratified estimator \( \hat{Y} \), ignoring finite population corrections, is approximately

\[
v(\hat{Y}) = \frac{N}{n} \sum_j N_j \sigma_j^2 + \frac{N^2}{n^2} \sum_j s_j^2 \left[ \frac{N-N_j}{N} \right]
\]  

(52)

For the basic work leading to Equation 52, see Stephan (22).

Post stratified estimates and their estimated variances were computed for characteristics in Blount County, Alabama. These were compared with the results from Equations 44 and 46. In this county
there were three land-resource units. There were 56 sample units in the county, 18 in one land-resource unit and 19 in each of the other two. The characteristics examined first were 11 major soil types and 16 land-capability units, the latter being subclassifications of the land-capability classes described previously. The relative precision of the post stratified estimator to the simple expansion estimator averaged 97.5 per cent for the 27 characteristics. The magnitude of the relative precision ranged from 86.3 per cent to 118.2 per cent.

To obtain additional information on the merits of the post stratified estimator, computations were next made for the cross-classification of land-capability unit and land use. In all, there were 127 items in the cross-classification. The average relative precision for the 127 items was 100.3 per cent. The largest was 127.0 per cent and the smallest was 79.0 per cent. The large losses and large gains were usually associated with relatively rare items.

If there is to be anything more than negligible gain from stratification or post stratification relative to simple random sampling, the effect of the stratification must be to divide the population into groups that are as different from each other as possible, but are still homogeneous internally. Post stratification by land-resource units apparently fails to do this except occasionally for rare characteristics. Stated in another way, there appears to be little difference between the land-resource units, at least in regard to characteristics of interest in this survey. The large number of sample units with observations equal to zero for many of the characteristics is undoubtedly
one of the principal reasons for there being such a small difference between the land-resource units.

The geographic stratification actually used in drawing the sample has been ignored in the treatment of post stratification given here, that is, a simple random sample was assumed. Analytically, post stratification of a simple random sample or of a stratified random sample can be studied by the methods worked out by Williams (29). However, the evidence indicates that stratification by geographic area and post stratification by land-resource units are both often rather inefficient. Hence, there would have to be an unusual "interaction" between the two classes of stratification before the precision from considering them both would be much better than from the geographic stratification alone.

The conclusions in regard to the post stratification by land-resource units are therefore negative. The computations are somewhat tedious and the expected gain in precision, if any, is small.

Thus, the estimator suggested for general use in this survey is the ratio estimator of Equation \(47\). Evidence indicated that its variance was practically equivalent to that given by Equation \(46\), which is the recommended formula if rough estimates of variances are considered satisfactory. In some cases, however, Equation \(43\), based on a grouping of the actual strata, may need to be computed.
B. Pooling of Sample Information from Similar Areas

As was indicated previously, the county was the basic universe in determining the sample design to provide information for the conservation needs inventory. However, the data obtained could also be used to provide estimates for other universes of interest, since each sample unit was also classified according to soil conservation district, watershed, and land-resource unit. The problem now to be considered arose in making estimates for a portion of a land-resource unit falling in one county. The solution is believed to also have application to other problems of a similar nature.

Suppose that land-resource unit A lies within two counties, County 1 and County 2. Assume that \( N_1 \) and \( N_2 \) are the number of population units in A in Counties 1 and 2, respectively, and that the population variance between the \( N_1 + N_2 \) units in A is \( S^2 \). Also assume that in A there are \( n_1 \) sample units in County 1 and \( n_2 \) sample units in County 2. It is desired to make estimates for the portion of A which is in County 1. The problem has particular importance when A occupies only a small part of County 1.

Let \( \overline{Y}_1 \) and \( \overline{Y}_2 \) denote the means of some characteristic \( y \) for the portion of A in County 1 and County 2, respectively. Under an assumption of simple random sampling, the usual estimators for \( \overline{Y}_1 \) and \( \overline{Y}_2 \) would be \( \overline{y}_1 \) and \( \overline{y}_2 \), respectively, where

\[
\overline{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i
\]  

(53)
and
\[ \bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \quad (54) \]

However, since all the observations are from the same land-resource unit, it seems reasonable that under some circumstances the \( n_2 \) observations in County 2 could be used in addition to the \( n_1 \) observations in County 1 to produce a better estimate of \( \bar{Y}_1 \). The question considered here is:

When would it be desirable to pool information in this manner?

Let
\[ \bar{y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{n_1 + n_2} \quad (55) \]

The variance of \( \bar{y} \) is then
\[ V(\bar{y}) = \frac{s^2}{n_1 + n_2} \quad (56) \]

The bias of \( \bar{Y} \) in estimating \( \bar{Y}_1 \) is \( (\bar{Y}_1 - \bar{Y}) \), and the mean square error (MSE) of \( \bar{y} \) as an estimator of \( \bar{Y}_1 \) is given by the variance plus the square of the bias, that is,
\[ \text{MSE}(\bar{y}) = \frac{s^2}{n_1 + n_2} + (\bar{Y}_1 - \bar{Y})^2 \quad (57) \]

where
\[ \bar{Y} = \frac{N_1\bar{Y}_1 + N_2\bar{Y}_2}{N_1 + N_2} \quad (58) \]

Pooling would be favored whenever
\[ \text{MSE}(\bar{Y}) < \text{MSE}(\bar{Y}_1) \]

Now, \( \text{MSE}(\bar{Y}_1) \) is simply \( \frac{s^2}{n_1} \), and \( \text{MSE}(\bar{y}) \) is given in Equation 57, which may
also be written as

\[
\text{MSE}(\bar{y}) = \frac{s^2}{n_1 + n_2} + \frac{N_2^2(\bar{Y}_1 - \bar{Y}_2)^2}{(N_1 + N_2)^2} \tag{59}
\]

since

\[
\bar{Y}_1 - \bar{y} = \bar{Y}_1 - \frac{N_1 \bar{Y}_1 + N_2 \bar{Y}_2}{N_1 + N_2} = \frac{(N_1 + N_2) \bar{Y}_1 - N_1 \bar{Y}_1 - N_2 \bar{Y}_2}{N_1 + N_2} = \frac{N_2 (\bar{Y}_1 - \bar{Y}_2)}{N_1 + N_2} \tag{60}
\]

Hence, \( \text{MSE}(\bar{y}) < \text{MSE}(\bar{Y}_1) \)

when

\[
\frac{s^2}{n_1 + n_2} + \frac{N_2^2(\bar{Y}_1 - \bar{Y}_2)^2}{(N_1 + N_2)^2} < \frac{s^2}{n_1}
\]

or

\[
s^2 \left[ \frac{n_2}{n_1(n_1 + n_2)} \right] > \frac{N_2^2(\bar{Y}_1 - \bar{Y}_2)^2}{(N_1 + N_2)^2} \tag{61}
\]

If the sampling is proportional, that is,

\[
\frac{n_2}{n_1 + n_2} = \frac{N_2}{N_1 + N_2} \tag{62}
\]

then Equation 61 may be written as
The inequality given in Equation 64 may be applied to a practical situation if estimates of the parameters are substituted in the inequality. \( \bar{Y}_1 - \bar{Y}_2 \) may be estimated by \( \bar{y}_1 - \bar{y}_2 \); and an estimate of \( S \) is \( s \), the standard deviation of the characteristic \( y \) for the \( n_1 + n_2 \) values in the sample. The resulting rule is to pool if

\[
\frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < 1
\]  

(65)

It is interesting to note that the criterion for pooling is equivalent to "Student's" t-test of the hypothesis of no difference between the means of two normal populations assumed to have a common variance. The value of 1 in the criterion is, in the test of hypothesis, the value of t for one degree of freedom at the 50 per cent probability level.

It should be mentioned that the pooling criterion given in Equation 65 is closely related to one given by Mosteller (17) in studying the problem of pooling means from two normal populations to estimate the mean of one of the populations.

Equation 65 could also be applied to problems similar to the one
considered here. For example, if it were desired to make estimates for the portion of a watershed in a certain county, would it be advantageous to also use information from the part of the watershed outside the county? The criterion suggested here would be applicable to this case.

One other type of pooling problem will also be considered. It is the pooling of information from several counties in order to produce more precise estimates in one of the counties. The procedure will be described in regard to the estimation of the acreage in a certain land-capability unit within a specified land use classification. The method can be applied when it is possible to form a group of counties such that each county in the group has, with respect to a given land use, roughly the same per cent distribution of acreage by land-capability units.

Let $Z$ be the acreage of soil in County A which has Land-capability unit IIIe4, for example, and is devoted to cropland. Suppose there are $n$ sample units in County A and suppose that in the homogeneous group of counties to which County A belongs there are a total of $m$ sample units. It is here assumed that all the counties in the group have been sampled at the same rate. If $Y_i$ is the acreage of cropland in the $i$-th unit, $X_i$ is the total acreage of the $i$-th unit, $X$ is the total acreage of County A, and $Z_i$ is the acreage of Land-capability unit IIIe4 and cropland for the $i$-th unit, then $Z$ can be estimated by

$$ \hat{Z} = \hat{Y} \hat{P} $$

(66)

where
\[ \hat{Y} = \left[ \frac{\sum_{i=1}^{n} Y_i}{n} \right] X \]  
\[ \hat{P} = \frac{\sum_{i=1}^{m} Z_i}{m} \]  

\( \hat{Y} \) is based on the \( n \) sample units in County A and estimates \( Y \), the total acreage of cropland in the county. \( \hat{P} \), which is calculated from the \( m \) sample units in the group of counties, estimates \( P \), the proportion of IIIe received occurring on cropland in the group of counties. In the estimator \( \hat{Z} \), this same proportion is applied to the individual county estimate of cropland in order to estimate the county total of IIIe on cropland.

The estimator \( \hat{Z} \) is biased and its mean square error (MSE) is, approximately,

\[ \text{MSE}(\hat{Z}) = P^2 V(\hat{Y}) + Y^2 V(\hat{P}) + V(\hat{Y}) V(\hat{P}) + Y^2 \left[ \frac{Z}{Y} \cdot \frac{Z_G}{Y_G} \right]^2 \]  

where \( Y_G \) is the total cropland in the group of counties and \( Z_G \) is the total of Land-capability unit IIIe occurring on the cropland. The first three terms on the right-hand side of Equation 69 constitute the variance of the product \( \hat{Y} \hat{P} \) if the correlation between \( \hat{Y} \) and \( \hat{P} \) is ignored. This correlation, caused by the \( n \) units employed in common in both the estimators \( \hat{Y} \) and \( \hat{P} \), would be expected to be small and to have little effect on the formula if \( m \) is large relative to \( n \). The last term is the square of the bias in using \( \hat{Z} \) as an estimate of \( Z \).
The usual ratio estimator of $Z$, using only information from County A, is

$$Z' = \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i}{X_i} \right] X \tag{70}$$

Its variance, given by the right-hand side of Equation 49 when $Y_i$ is replaced by $Z'_i$, is

$$V(Z') = \frac{N(N - n)}{n(N - 1)} \sum_{i=1}^{N} (Z_i - RX_i)^2 \tag{71}$$

where $R = \frac{Z}{X}$. $V(Z')$ will now be compared with $\text{MSE}(\hat{Z})$.

Since the acreage proportion of total cropland is larger than the acreage proportion for IIIa4 in cropland, the inequality

$$\frac{V(\hat{Y})}{Y^2} < \frac{V(Z')}{Z^2} \tag{72}$$

would be expected to hold. See Figure 3 in Section VIII for supporting evidence. Equation 72 can also be written as

$$P^2 V(\hat{Y}) < V(Z') \tag{73}$$

since $P^2$ is approximately $Z^2/Y^2$.

As an example, consider the case where the proportion of total cropland is .25 and there are ten land-capability units each with proportion .025. Then from Figure 3 the relative standard errors for a land-capability unit and for total cropland would be, respectively, approximately in the ratio of 2.50 to .70 or about 3.5 to 1. The ratio of the variances would then be about 12 to 1. Thus in Equation 73 the right-hand side of the inequality would be about 12 times as large as
the left-hand side, which is the first term in Equation 69, the expression for MSE(\( \hat{Z} \)).

The second and third terms in MSE(\( Z \)) can be written as \( Y^2 V(\hat{P}) \left[ 1 + \frac{V(\hat{Y})}{Y^2} \right] \). First, consider \( \frac{V(\hat{Y})}{Y^2} \) inside the bracket. If \( Y \) is again about 25 per cent of the total land in the county, then from Figure 3, \( \frac{S.E.(\hat{Y})}{Y} \), based on a sample of size 1 is approximately .7 and its square would be about .5. With a 2 per cent sample \( n \) would be about 48 so that \( \frac{V(\hat{Y})}{Y^2} \) for a sample of size 48 would be about \( \frac{50}{48} \), or approximately .01, which is negligible compared to 1.

\[ Y^2 \frac{V(\hat{P})}{Y^2} \] compared to \( V(Z') \) will be approximately in the ratio of \( n \) to \( m \) since the former is based on \( m \) units and the latter on \( n \) units. If there were, say, twelve counties in the group, then \( V(Z') \) would be approximately twelve times as large as \( Y^2 \frac{V(\hat{P})}{Y^2} \).

When the results from the first term are also considered, it is seen that in the present example, \( V(Z') \) is about six times as large as the first three terms in MSE(\( \hat{Z} \)). The success of the estimator \( \hat{Z} \) is thus largely dependent on the remaining term, the square of the bias. If the bias is small as would be the case if a truly homogeneous group of counties is formed, then \( \hat{Z} \) would be considerably more precise than \( Z' \) as an estimator for \( Z \).

A special technique, much like estimating variance components in an analysis of variance, could be used for estimating the bias for a practical situation. Unfortunately, actual data are not available at this time for studying this problem further and the estimation technique is therefore not given here.
C. Utilization of Information from Outside the Sample

An important estimation problem arose in counties where land had been mapped previously. For some counties in particular, a substantial body of information was available. It seemed unreasonable to ignore this previous work and rely only on the sample data. The question then was: Could this information be utilized either to reduce sample size for the present study or to produce better estimates of county totals? It should be noted that in most cases the areas mapped had not been selected in a random manner.

A rather obvious method of dealing with this question was to consider a county as composed of two parts or strata, one part consisting of all the land mapped previously and the other part comprising the unmapped land. If it is assumed that there is complete knowledge about the mapped land, then sampling is required only on the unmapped land. Estimates for the unmapped land could be made in the usual way from the sample data. Then, by adding to them the information for the land previously mapped, an estimate for the whole county could be obtained. Some features of this procedure will now be examined assuming simple random sampling.

Consider a county composed of $N$ units. Suppose there are $N_1$ units of mapped land and $N_2$ units of unmapped land in the county, where $N_1 + N_2 = N$. It is assumed that information concerning which land is already mapped and which is not is generally not available to the person drawing the sample. Therefore, a sample of $n$ units is drawn for the whole county. It turns out later that $n_1$ units fall in
the mapped portion of the county and \( n_2 \) units in the unmapped part \((n_1 + n_2 = n)\). The \( n_1 \) units will not need to be sampled, since information already exists for them, as well as for the other \( N_1 - n_1 \) units in this stratum.

For a characteristic \( y \), an estimate of its total for the unmapped stratum is given by

\[
\hat{Y}_2 = N_2 \bar{y}_2
\]  

where

\[
\bar{y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i
\]

If \( Y_1 \) is the total for \( y \) in the mapped portion of the county, then an estimate of \( Y \), the total for the whole county, is

\[
\hat{Y} = Y_1 + N_2 \bar{y}_2
\]

The variance of \( \hat{Y} \), ignoring the finite population correction, is given by

\[
V(\hat{Y}) = \frac{N_2^2 S^2}{n_2}
\]

where \( S^2 \) is the variance between population units, that is,

\[
S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2
\]

\( \bar{Y} \) is here defined to be \( \frac{Y}{N} \).

It should be noted that the problem considered here is a special case of post stratification since the sample size \( n_2 \) would vary upon repeated sampling. Therefore, the formula for variance given in Equation 77 is only approximately correct.
In the usual case where \( n \) units are mapped in the county the variance of the total is given by

\[
V(\hat{Y}) = \frac{N^2 S^2}{n} \tag{79}
\]

The relative precision of the scheme using the previous mapping (Scheme 2) to the usual scheme (Scheme 1) is given by

\[
\text{R.P.} \left( \frac{\text{Scheme 2}}{\text{Scheme 1}} \right) = \frac{N^2 S^2}{n} / \frac{N_2^2 S^2}{n_2} = \frac{N^2 n_2}{N_2^2 n} \tag{80}
\]

On the average, the sampling will be proportional, that is, \( \frac{n_2}{n} = \frac{N_2}{N} \), so that Equation 80 can be written as

\[
\text{R.P.} \left( \frac{\text{Scheme 2}}{\text{Scheme 1}} \right) = \frac{N}{N_2} \tag{81}
\]

Equation 81 indicates what might be expected -- that the gain in precision is directly dependent on how much of the land was previously mapped. In the case of no land being mapped prior to sampling (\( N_2 = N \)) there is no gain in precision. Of course, in this case the two schemes become identical and would be expected to yield the same precision.

For the other extreme, that is, when \( N_2 = 0 \), the expression for relative precision becomes infinite. This might be anticipated also. In this case the county is completely mapped, and to make use of this information is infinitely better than any plan of sampling.

The cases of practical interest are the ones lying between the two extremes, when a portion of the county has been mapped. The precision of considering the prior information relative to the usual case
is given by the reciprocal of the portion of land unmapped. Thus, the relative precision when $1/3$ of the county is mapped and $2/3$ is unmapped is the reciprocal of $2/3$ or $3/2$. When the mapping fraction is $2/3$ and the proportion not mapped is $1/3$, the relative precision is $3$. For the gain in precision to be as much as 10 per cent, the fraction of land already mapped must be $1/11$. In most counties the fraction of land already mapped is less than $1/11$, so that the gains from using previous mapping are usually less than 10 per cent.

In the treatment just given it was shown that by considering the land as being divided into two parts and by sampling only the land not previously mapped, it was possible both to reduce sample size and to increase precision. The relative precision was shown to be $\frac{N}{N_2}$, and, at the same time, the effective sample size was reduced from $n$ to $n_2$. It is possible, of course, to increase precision still more if it is desired to sample $n$ units in the unmapped area, instead of $n_2$. Such a plan might be followed in a state where funds had definitely been allocated for the sampling of $n$ units in each county, although in practice it might not be possible to predict costs closely enough to make such a plan feasible. The expected precision of sampling $n$ units in this case relative to sampling $n_2$ is simply $\frac{n}{n_2}$. To delineate $n$ units exactly it would be necessary to know the boundaries of the mapped area. If the boundaries were unknown to the person drawing the sample, then a sample of $\frac{Nn}{N_2}$ units should be drawn in the whole county to assure that, on the average, $n$ units would fall on the unmapped land.

If it were desired, on the other hand, to utilize the known mapping
information and still maintain precision at the level expected for a county having no previous mapping, it would be possible to reduce the size of the sample in the unmapped portion to a number smaller than \( n \). This number would be equal to \( \frac{n}{n^2} \). In this case, also, it would be necessary to know which land was mapped and which was unmapped in order to delineate exactly \( \frac{n}{n^2} \) units for the unmapped portion.

The problem of dealing with the situation where some land is already mapped is related to the pooling problem considered in Section VII, Part B. However, it is believed that a pooling procedure will probably not be applicable because usually in such situations there is reason to suspect that the unmapped portion definitely differs in certain characteristics from the unmapped portion. The sample falling in the mapped portion will therefore generally be wasted from the point of view of improving the estimation of the unmapped portion.

The above discussion on utilizing previous mapping assumes that the information obtained sometime in the past still holds for the present. This would ordinarily be true for soil type, per cent of slope, and degree of erosion. Data on land use, however, would not necessarily be the same as in a former period; hence, this characteristic cannot be treated in the manner suggested here. A method for using prior information on land use will be given in Section VII, Part E.

One other plan of using information from outside the sample was investigated. This was the utilization of the land use data available from the U. S. Census Bureau’s 1954 Census of Agriculture, the
latest in the series of quinquennial national agricultural censuses.

The idea to be examined here is whether this information can be used to produce more precise estimates of acreages in the various land-capability units within each land use category. The suggested method is to use a ratio estimator of the form given by Equation 47 where, in this instance, for the i-th unit, $X_i$ is the total acreage in a particular land use class and $Y_i$ is the acreage having a particular land-capability unit within the specified land use class. For the county, the total acreage in the land use class is then given by $X$, obtained from the Census Bureau data.

The estimator and its estimated variance, given by Equation 51, were calculated for each of the land-capability units in two land uses, cropland and pasture, using the data from Miami County, Kansas. Simple random sampling was assumed. The results from the ratio estimator were then compared with those obtained by using Equations 44 and 46.

The relative precision of the ratio estimator compared to simple expansion ranged from 94.7 per cent for Land-capability unit IIW4 in cropland to 153.8 per cent for Land-capability unit IIIe4 in pasture. The calculation of relative precision was made for a total of 48 such items. The average relative precision was 104.8 per cent.

In general, the largest gains were for the items with the largest prevalence rates. For example, the seven items with the largest prevalence rates had the following relative precisions, in per cent: 116.3, 113.3, 120.5, 98.7, 153.8, 141.3, and 132.0. Only one of the seven was less than 100 per cent and the others were all considerably above the
average for the 48 items.

However, a matter which has been ignored up to this point appears to impose serious limitations on the usefulness of the method. This matter relates to the differences between the land use definitions used in this survey and those given in the publication by the Census Bureau (26).

One major difference is that all land except federally-owned non-cropland is included in the survey while land use data were collected by the Census Bureau only for land used for agricultural purposes. Considerable acreages of woodland and wasteland were thereby excluded from the census figures. For the land use classifications, forest and other, of interest in the survey, the method of using data from the Census Bureau thus has little value.

Since most cropland and pasture would be included in the Census Bureau's figures, the method would seemingly be advantageous for these two uses. However, it turns out that there is not exact correspondence between the two definitions for either of the uses. For example, the Census Bureau apparently includes under cropland some permanent open pasture which in the survey would be classed as pasture. Another discrepancy occurs for the classification woodland pastured. The Census Bureau has a separate category for this land, which includes all woodland that was used for pasture or grazing. In the survey, such land is classed as pasture if the canopy of the trees occupies less than 10 per cent of the area; otherwise, it is classed as forest or woodland.

If the ratio estimator is used in the manner considered here,
that is, the sample value of $\hat{R}$, based on definitions appropriate for
the survey, is multiplied by $X$, based on definitions used by the Census
Bureau, the resulting figures will not be unbiased estimates for the
survey classifications. The best that can be said for the method, then,
is that it appears to be a reasonable one to use for items that are not
rare if there is evidence to indicate that the difference in defini-
tions would have little or no consequence. Such evidence would likely
be largely subjective in nature; hence the value of the method for
estimating totals is very much in doubt.

D. A Subsampling Problem

Another estimation problem arose when it developed that in some
counties all the sample units originally delineated could not be mapped
within the period designated for completing the inventory. A shortage
of both funds and personnel was the cause of this situation. To provide
the required estimates some plan of subsampling seemed to be indicated.

In counties where none of the sample units had yet been mapped at
the time of the decision to subsample, there was no special problem in
regard to the selection of a new sample. In these counties it was a
routine matter to draw a new sample which was some specified fraction
of the old one. However, in counties where some of the units original-
ly delineated were already mapped, there were questions regarding the
procedure to use in subsampling.

Sample units already mapped had usually not been done in any par-
ticular order. Most often these units were concentrated in one part of
the county and could not be considered a random subsample of the original random sample. The approach taken on this problem is related to the method described in Section VII, Part C in that a county was imagined to be composed of two parts, one part with \( N_1 \) units represented proportionally by \( n_1 \) units already mapped and the other part with \( N_2 \) units represented proportionally by \( n_2 \) unmapped units so that \( n_1/N_1 = n_2/N_2 \). It is assumed that \( N_1 + N_2 = N \) and that \( n_1 + n_2 = n \). A common variance \( S^2 \) between the \( N \) units is also assumed.

Given the above situation, now suppose it was decided that a subsampling scheme be used. In counties having no units mapped assume it was decided that \( k \) of the \( n \) units would be mapped. In counties having \( n_1 \) units already mapped, the decision was to map from the remaining \( n_2 \) units a number \( k_2 \), smaller than \( k \), which, together with the \( n_1 \) units already mapped, would be expected to give estimates with the same precision as \( k \) units out of the original \( n \) would have given. In the latter scheme the \( k_2 \) units can be regarded as a random sample drawn from the \( N_2 \) units in the unmapped portion of the county.

In the first case, with \( k \) units, an estimate of \( Y \), the county total for a characteristic \( y \), would be

\[
\hat{Y} = \frac{N}{k} \sum_{i=1}^{k} y_i
\]  

and its variance, ignoring the finite population correction, would be

\[
V(\hat{Y}) = \frac{N^2 S^2}{k}
\]

In determining \( k_2 \) for counties already having \( n_1 \) units mapped it
will be helpful to appeal to formulas for stratified sampling. Suppose \(k_i\) units were sampled in each stratum. Then \(Y\) would be estimated by

\[
\hat{Y} = \sum_i \frac{N_i}{k_i} \bar{y}_i
\]  

(84)

where \(N_i\) is the number of population units in the \(i\)-th stratum, \(\bar{y}_i\) is the mean of the \(k_i\) sample units in the \(i\)-th stratum, and the summation is taken over all strata. The variance of \(\hat{Y}\) in the usual case is given by

\[
V(\hat{Y}) = \sum_i \frac{N_i^2 S_i^2}{k_i}
\]  

(85)

where \(S_i^2\) is the variance between all units in the \(i\)-th stratum and the summation is again over all strata.

In the situation considered here, there were two strata and hence two terms in the summation. If it is also assumed that \(S_1^2 = S_2^2\), then Equation 85 can be written as

\[
V(\hat{Y}) = S^2 \left( \frac{N_1^2}{k_1} + \frac{N_2^2}{k_2} \right)
\]  

(86)

But since sampling was assumed to be proportional so that

\[
\frac{n_1}{N_1} = \frac{n}{N}
\]  

(87)

and

\[
\frac{n_2}{N_2} = \frac{n}{N}
\]  

(88)

Equation 86 can also be written as

\[
V(Y) = S^2 \left[ \frac{N_1^2}{n_1^2 k_1} + \frac{N_2^2}{n_2^2 k_2} \right] = \frac{N^2 S^2}{n^2} \left[ \frac{n_1^2}{k_1} + \frac{n_2^2}{k_2} \right]
\]  

(89)
It is desired to determine the value of \( k_2 \) which, when \( k_1 = n_1 \), will make this variance equal to that for \( k \) units sampled at random from the original \( n \). This is easily obtained by equating Equation 83 and Equation 89, substituting \( n_1 \) for \( k_1 \), and solving for \( k_2 \). The result is

\[
\frac{n_2^2}{\frac{n}{k} - n_1}
\]

It should be mentioned that the problem on subsampling and the solution presented here fit into the framework of what Hansen et al. (6) have called "double sampling with stratification." In this method a large sample is drawn initially but only enough information is obtained to classify units into two categories. In one category, detailed information is then obtained for all units, while units in the other category are subsampled. In the notation used here, detailed information is obtained for the \( n_1 \) units already mapped, while \( k_2 \) units are subsampled from the \( n_2 \) units originally drawn and not mapped.

From double sampling theory the estimator of \( Y \) is

\[
\hat{Y} = \frac{n}{n} \left[ \frac{n_1}{\frac{n}{k} \sum_{i=1}^n Y_1} + \frac{n_2}{k_2} \sum_{i=1}^{k_2} Y_1 \right]
\]

Under the assumption of proportional allocation, Equation 84 and Equation 91 can be shown to be equivalent. If the finite population correction is ignored and if a common variance \( s^2 \) is assumed, the variance of \( \hat{Y} \) in the double sampling framework can be expressed as

\[
V(\hat{Y}) = \frac{s^2}{n} \left[ 1 + \frac{n_2^2}{n} \left( 1 - \frac{k_2}{n_2} \right) \frac{1}{k_2} \right]
\]
Equation 92 can be shown to be equivalent to the variance given by Equation 89 when $k_1 = n_1$.

The double sampling approach is not dependent on the imaginary subdivision of the county into two strata of size $N_1$ and $N_2$ and the assumption that sampling is proportional in the two strata.

E. Estimating Present Land Use

As indicated previously, sample units were drawn for all counties whether or not any mapping had already been done. For characteristics not expected to change through time, a remapping of areas previously mapped would provide little information of value. However, for estimating present land use, a characteristic which might change from year to year, it was necessary to have the data from the sample units. A method which utilizes past data as well as the sample data will now be given for estimating land use. It is assumed that complete data on land use are available from some previous year.

The method suggested here for making estimates of land use is to employ a ratio estimator. This type of estimator was considered in Section VII, Part A. In that section, where the auxiliary variate used was total acreage, it was found that the ratio estimator offered practically the same precision as merely multiplying the sample total by the inverse of the sampling rate. For estimating land use it is proposed to use former land use as the auxiliary variate. Since a high correlation between former and present land use would be expected for most counties, estimates made from this method should be consider-
ably more precise than estimates which ignore the existing data. Simple random sampling is assumed.

Consider a particular land use. Let its acreage in the i-th sample unit for the past and the present be \( X_i \) and \( Y_i \), respectively. Let the county total of the past data for the land use be \( X = \frac{\sum X_i}{N} \). Assume that a random sample of \( n \) units has been drawn from the \( N \) in the county. What is desired is to estimate \( Y \), the county total for present land use. Appropriate formulas for estimation and variance calculation were given in Section VII, Part A. They are repeated below for convenience.

The suggested estimate of \( Y \) is

\[
\hat{Y} = \hat{R} X
\]

where

\[
\hat{R} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}
\]

It may be seen that \( \hat{R} \) estimates \( R = \frac{Y}{X} \), the relative change that has occurred since the previous mapping. \( \hat{R} \) is then multiplied by the known total \( X \) to estimate the present land use \( Y \).

The variance of \( \hat{Y} \) is given by

\[
V(\hat{Y}) = \frac{N(N - n)}{n(N - 1)} \sum_{i=1}^{N} (Y_i - RX_i)^2
\]

Equation 95 is an approximate formula. For a discussion of this approximation as well as the bias in \( \hat{Y} \) see Cochran (5). The variance may be estimated from the sample data by

\[
V(\hat{Y}) = \frac{N(N - n)}{n(n - 1)} \sum_{i=1}^{n} (Y_i - \hat{R}X_i)^2
\]
For computing, a more convenient form of Equation 96 is

\[ v(\hat{Y}) = \frac{N(N - n)}{n(n - 1)} \left[ \sum_{i=1}^{n} \frac{Y_i^2}{n} - 2 \sum_{i=1}^{n} \frac{Y_i X_i}{n} + \sum_{i=1}^{n} \frac{X_i^2}{n} \right] \] (97)

It may be shown that the ratio estimator of Equation 93 will have a smaller variance than the estimator of Equation whenever the correlation between the values of \( Y_i \) and \( X_i \) is greater than the relative standard error of \( X_i \) divided by twice the relative standard error of \( Y_i \). The proof of this is straightforward. See, for example, Cochran (5, p. 122). Whenever the two relative standard errors are approximately the same, there would be a gain in precision by using the ratio estimator if the correlation between past and present values were larger than 0.5. In land use data brought to the author's attention the correlation coefficients were larger than 0.5, indicating that there will be a gain in precision from using the past data on land use. Unfortunately, no data were immediately available for further examination of this point.

Achieving the gain in precision expected from the ratio estimator will, however, usually entail some additional work. The reason for this is that, while \( X \) will generally be known, the individual \( X_i \) will not be known immediately. That is, the usual situation is that the total land use for the county has been measured from the mapping but not by individual units. The values for \( X_i \) may, of course, be obtained by locating and measuring each unit on the mapping for the county. In cases where there is no particular interest in having more precise estimates the expenditure of obtaining the information for the ratio
estimator will not be justified. In this event, the estimate obtained by simple expansion may serve the purpose.

In so far as the ratio estimator suggested here provides estimates with higher precision for land use classes, these estimates could be substituted in place of the Census Bureau’s land use figures in the estimation scheme presented in Section VII, Part C. The following estimator would be the result:

\[
\hat{Y} = \left[ \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \right] X
\]  

(98)

where \( Y_i \) is the present acreage in the \( i \)-th unit having a particular land-capability unit and a specified land use, \( X_i \) is the past acreage in the \( i \)-th unit for the same land use, and \( X \) is the county total of the land use based on the past data. The variance formulas already given for the ratio estimator also apply in this case.
VIII. ESTIMATION OF RELATIVE STANDARD ERRORS

The primary object of the tabulation phase of the project was to obtain the various estimates required to prepare the conservation needs inventory. As indicated by some of the estimation formulas given in Section VII, the estimates themselves can be calculated fairly easily. The calculation of variances is, however, usually relatively tedious. For this reason and because they were not specifically needed for the inventory, variance estimates were not furnished for each characteristic, although, since necessary formulas are given in Section VII, these may be calculated if needed.

Since individual estimates of variance were generally not computed, it was felt that there might be some value in presenting a graphic method of estimating the relative standard error for any item. This section is concerned with that topic.

It will be recalled that the sample design was that known as stratification with proportional allocation. It will be assumed here that townships constituted the strata and that units were selected randomly in each township.

Formulas appropriate to this design were given in Section VII, Part A. For example, Equation 44,

\[ \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i = N \bar{y} \]

can be used as the estimator of \( Y \), the population total of a characteristic \( y \). The estimated variance of \( \hat{Y} \), ignoring the finite population correction, was given by Equation 43,
In this section it will be assumed that \( \frac{1}{M} \sum_{j=1}^{M} s_j^2 \) is satisfactorily given by the mean square between units within townships as calculated in an analysis of variance. This procedure is exact when the township sizes are exactly equal; it gives a close approximation when slight differences occur. For convenience of notation this mean square will be designated \( s^2 \), the symbol formerly used for the estimated variance between units in a simple random sample. Therefore, the estimated variance of \( \hat{Y} \) will be taken to be

\[
v(\hat{Y}) = \frac{N^2}{n} s^2
\]

(99)

where \( s^2 \) is obtained in the manner indicated above.

The square root of \( v(\hat{Y}) \) is the estimated standard error of \( Y \), namely,

\[
S.E.(\hat{Y}) = \frac{Ns}{\sqrt{n}}
\]

(100)

The relative standard error of \( \hat{Y} \) is then

\[
R.S.E.(\hat{Y}) = \frac{S.E.(\hat{Y})}{\overline{Y}} = \frac{s}{\overline{Y} \sqrt{n}}
\]

(101)

Values of \( s, \overline{Y}, \) and \( \frac{s}{\overline{Y}} \) were calculated from sample data for many characteristics in various counties. The value of \( s \) for a characteristic was the square root of the mean square between units within townships in an analysis of variance of the acreages per unit for the characteristic. Corresponding values of \( \overline{Y} \) were obtained from \( \overline{Y} = 160 \overline{R} \).
where \( \hat{R} \) is the estimated fraction of land occupied by the characteristic \( y \) and 160 is the number of acres per unit. \( \hat{R} \) was defined in mathematical terms in Equation 48. It is

\[
\hat{R} = \frac{\sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i}
\]

where \( X_i \) is the total acreage for the \( i \)-th unit and \( Y_i \) is the acreage possessing the characteristic \( y \). \( \hat{R} \) is an estimate of the prevalence rate mentioned in Section IV. Calculating \( \overline{y} \) in this manner gave practically identical results with those obtained from \( \frac{1}{n} \sum_{i=1}^{n} Y_i \) since nearly all sample units in a county were of the same size. The former method of computation was also more convenient since \( \hat{R} \) was calculated as part of the recommended procedure of estimation. See Section VII, Part A.

It will be noted that \( s/\overline{y} \) can be considered to be the relative standard error for a sample of size 1. Estimates of the relative standard error for a sample of size \( n \) can then be obtained by dividing \( s/\overline{y} \) by \( \sqrt{n} \).

Values of \( s/\overline{y} \) and \( \hat{R} \) from sample data are plotted in Figure 3. The data are from the following counties: Brown, Kansas; Coffee, Alabama; Foster, North Dakota; Guilford, North Carolina; Harvey, Kansas; Jefferson, Iowa; Johnson, Nebraska; Pawnee, Kansas; and Trousdale, Tennessee. Characteristics examined were soil type, slope, erosion, land use, and land-capability class. Units of 160 acres were used in all nine counties. As might be expected, \( s/\overline{y} \) is large for small values of \( \hat{R} \) and is small when \( \hat{R} \) is large. The plotting of points also shows
Figure 3. Relationship of $s/y$ to prevalence rate $R$, based on sample data from nine selected counties

\[ \frac{s}{y} = \frac{0.413}{1-R} \]
that there is considerable variation in values of $s/y$ for any given $\hat{R}$. Still, there is apparent a definite curvilinear relationship. A curve of the form

$$\frac{s}{y} = \frac{c\sqrt{1 - \hat{R}}}{\sqrt{\hat{R}}}$$  \hspace{1cm} (102)

was fitted to the data by a special least squares technique. The curve is also shown in Figure 3. The value of $c$ turned out to be .413.

$\hat{R}$ is always in the range $0 \leq R \leq 1$. A value of 0 for $\hat{R}$ means, of course, that there was none of the characteristic occurring in the sample. The value of $s/y$ corresponding to $\hat{R} = 0$ is infinity. When $\hat{R} = 1$, $s/y$ becomes 0. This is what would be expected since, when $\hat{R} = 1$, the characteristic $y$ occupies the whole universe and the variance between sampling units would be 0.

Equation 102 is closely related to the formula for the coefficient of variation of the estimated total number of units in one of the classes of a binomial population. In the notation used here, this coefficient of variation is $\sqrt{1-R}/\sqrt{nR}$ for a sample of size $n$ if the finite population correction is ignored. $R$ is the population proportion of units in the class. Apart from the difference between $R$ and its estimator $\hat{R}$, the ratio $\sqrt{1-R}/\sqrt{\hat{R}}$, which might be considered the coefficient of variation for a sample of size 1, differs only by the constant $c$ from the right-hand side of Equation 102, which is also in terms of a sample of size 1.

To estimate the relative standard error (R.S.E.) of a characteristic $y$ from a sample of size $n$ by the short-cut method suggested here,
it is necessary to determine \( \hat{R} \) from the sample, obtain \( \frac{s}{\sqrt{y}} \) from the relationship

\[
\frac{s}{\sqrt{y}} = \frac{4.13\sqrt{1 - \hat{R}}}{\sqrt{R}}
\]

and divide \( \frac{s}{\sqrt{y}} \) by \( \sqrt{n} \).

It is obvious that the value of the R.S.E. for any given characteristic might be more or less than the value resulting from the use of Figure 3 or Equation 103. In cases where it is desired to know the R.S.E. more exactly than can be obtained by the short-cut method given here, it is suggested that the necessary computations from the sample data actually be carried out. Figure 3 is intended for use only in cases where the actual calculations are not completed but some idea of the magnitude of the error is still wanted.

As an example of the use of the method, consider estimating the R.S.E. for Slope D in Trousdale County, Tennessee. \( \hat{R} \) for this characteristic was .236. From Equation 103 or Figure 3, \( s/\sqrt{y} \) is estimated to be .74. For this county \( n \) was 47, so \( \sqrt{n} \) was 6.86. The estimated R.S.E. is, therefore, .108 or 10.8 per cent. The actual value calculated according to Equation 101 turned out to be .113 or 11.3 per cent.

It will be recalled that in "standard-sized" counties in the general farming classification a 2 per cent sample of 160-acre units was drawn. Since this sampling plan was used in more counties than any other plan, it is of interest to prepare a table indicating the relative standard error that may be expected for various values of \( \hat{R} \). This is given in Table 39 for \( n = 48 \). The procedure used was that outlined
above. For each assumed value of \( \hat{R} \), \( s/\sqrt{N} \) was calculated from Equation 105 and the result was then divided by \( \sqrt{48} \) to obtain the estimated R.S.E. associated with \( \hat{R} \). Similar calculations were made for other values of \( n \), ranging from 6 to 192. These results are also included in Table 39. The same method may be used to determine the estimated R.S.E. for any combination of \( n \) and \( \hat{R} \) that is of interest. Values in the table have been rounded to the nearest whole per cent.

Table 39. Estimated per cent relative standard error for soil characteristics by size of sample and prevalence rate

<table>
<thead>
<tr>
<th>Prevalence rate (( R ))</th>
<th>Number of 160-acre sampling units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td>.01</td>
<td>168</td>
</tr>
<tr>
<td>.03</td>
<td>96</td>
</tr>
<tr>
<td>.05</td>
<td>73</td>
</tr>
<tr>
<td>.10</td>
<td>51</td>
</tr>
<tr>
<td>.15</td>
<td>40</td>
</tr>
<tr>
<td>.25</td>
<td>29</td>
</tr>
<tr>
<td>.50</td>
<td>17</td>
</tr>
<tr>
<td>.75</td>
<td>10</td>
</tr>
<tr>
<td>1.00</td>
<td>0</td>
</tr>
</tbody>
</table>

As an example of the use of Table 39 consider the estimation of the total acreage of cropland in Brown County, Kansas. \( \hat{Y} \) was 265,603 acres and \( X \), the total area of the county, was 369,860 acres. The estimate of \( \hat{R} \) for this characteristic was, therefore, .718. The sample size was 48. From Table 39, the estimated R.S.E. of \( \hat{Y} \) is slightly
greater than 5 per cent.

If the assumption is made that the normal approximation applies, it is possible to use the R.S.E. to make confidence interval statements about \( \hat{Y} \). See Cochran (5) for a discussion of the validity of the normal approximation. The simplest statement to make is the following: The chances are about 2 in 3 that the interval from \( \hat{Y}(1-R.S.E.) \) to \( \hat{Y}(1+R.S.E.) \) contains the true value \( Y \). Of course, statements for other levels of probability can be made. For example, the 95 per cent confidence limits are \( \hat{Y}(1-2 R.S.E.) \) and \( \hat{Y}(1+2 R.S.E.) \). In general, the limits are \( \hat{Y}(1-t_{\alpha,n-1} R.S.E.) \) and \( \hat{Y}(1+t_{\alpha,n-1} R.S.E.) \) where \( t \) is "Student's" \( t \) for the probability level \( \alpha \) and \( n-1 \) degrees of freedom.

A more exact confidence interval statement can be made. For example, if a probability level of two thirds is always used for computing the upper and lower confidence limits and the statement is made that the interval from the lower limit to the upper limit covers the true value, such statements will be correct two times out of three, on the average. The statements in this paragraph and the preceding one do not, of course, take into account any "non-sampling" errors such as errors of measurement, etc.

Table 39 was based on units of 160 acres. However, as indicated by Table 35 and the discussion in Section V, it may be easily converted so as to be applicable to 40-acre and 640-acre units. For use with 40-acre units the column headings on number of units should be doubled. With units of 640 acres the column headings should be halved. Thus, the expected error of 18 per cent for a characteristic
with a 10 per cent prevalence rate, for example, is applicable to samples of 48 units of 160 acres, 96 units of 40 acres, or 24 units of 640 acres. Comments made in conjunction with Table 35 also apply here. In particular, a $g$ value of .5 is assumed and the finite population correction is ignored.

A brief discussion on the finite population correction seems appropriate at this point. The effect of ignoring the correction is to overestimate the variance. In cases where the rate of sampling is small, less than 5 per cent, say, leaving out the correction has little effect on the results. However, for larger rates of sampling, such as indicated in Table 35 for size-classes 1, 2, and 3, the variance may be overestimated considerably by not including the correction. The form of the finite population correction applicable to variance is $\frac{N - n}{N}$ where $n$ is the number of units sampled from $N$ in the universe. For standard errors the correction is $\sqrt{\frac{N - n}{N}}$. Thus, if it is desired to take the correction into account for a relative standard error appearing in Table 39, the figure given there should be multiplied by $\sqrt{\frac{N - n}{N}}$. 
IX. DISCUSSION

As the data for the conservation needs inventory are obtained and tabulated, two points related to the present study should in the author's opinion be considered further.

One point is concerned with cost information. In Section IV, Part C, where a cost function was established, it was necessary to use estimates obtained from soil scientists for some of the cost components since the cost records for actual mapping operations contained no information on costs per unit. The opinions of these trained scientists are based on years of experience so it is quite likely that their composite ideas on costs are close to the actual figures. Still, the mapping of the sample units that were drawn for the inventory affords an ideal opportunity for checking on many of the cost components. The suggestion advanced here is that, as the mapping of the sample units is done, costs for the various cost factors be recorded. Records on mapping time per unit would be of particular importance since the component for mapping is by far the largest in the total cost per unit. From the viewpoint of possible future research the value of information of this type should more than compensate for any inconvenience in obtaining it.

The second point is concerned with the estimator given by Equation 66 in Section VII, Part C. This estimator uses data from a group of homogeneous counties in an attempt to make more precise estimates in one of the counties. The theoretical consideration of the estimator indicated that, of the possibilities examined, it is
probably the most promising for making more precise estimates of the total acreage of characteristics of major interest in the inventory.

It would have been desirable, therefore, to include in the present study some results of the application of this estimator to actual data. However, as explained previously, this was not possible. It is suggested that this step be taken when the data become available. As soon as the general magnitude of the bias can be determined from the data, the merits of the estimator can be assessed more precisely. It is expected that it will turn out to be more precise in numerous situations.

Finally, three special aspects of this study will be treated in appendices. Two of these are related to certain economic concepts. For example, the form of the variance function used in this study is analogous to the well-known Cobb-Douglas production function. This analogy will be developed further in Appendix A. Also, the determination of an optimum size of unit can be related to the standard marginal analysis given in economic theory. This relationship will be set forth in Appendix B. In this study, the optimum size of sampling unit was estimated using information on cost and variability which, in turn, is subject to variation. It is of interest, therefore, to estimate the effect of such variability on the determination of the optimum unit. Such a stochastic theory is developed in Appendix C.
The problem considered in this study was the determination of an optimum sampling design to provide the basic information required for a national inventory of soil and water conservation needs. In particular, the optimum size of unit was determined and methods were developed for estimating acreages of various soil classifications for each county.

The decision regarding the optimum size of sampling unit was based primarily on evidence about the variability of different-sized units and information about the cost per unit. The approach taken in regard to an optimum sampling unit was to establish a variance function showing the relationship between variance and size of unit and a cost function indicating how costs change with size of unit. Based on these two functions, an expression was obtained for the optimum size of unit. The procedure used was to determine the size of unit which gave estimates with minimum variance for a certain fixed expenditure. In this connection, certain homogeneity considerations were also examined.

The recommended size of sampling unit was different for different types of land. For irrigated land a 40-acre unit was the optimum. For general farming and dry farming land a unit of 160 acres was preferred. A 640-acre unit was recommended for range land.

Several different estimators were considered for possible use in the estimation and tabulation phase of the project. The estimator suggested for general use was a ratio estimator having total acreage
as the auxiliary variate. Various other estimators were considered for special problems.

Two types of problems involving the pooling of data from similar areas were considered. One estimator based on a pooling procedure appeared to offer considerable reduction in the variance of county estimates.

Other estimation problems which were examined included the utilization of information from outside the sample, the estimation of the acreage of present land use when complete data from a previous mapping were available, and a special problem regarding subsampling.

Although complete variance formulas for the estimators were developed and evaluated in key examples, a short-cut method in graphical form, giving approximate values for the relative standard error, was presented for general use.
XI. LITERATURE CITED


XII. ACKNOWLEDGMENTS

For valuable assistance with many of the early problems encountered in this research, the author wishes to acknowledge his debt to R. J. Jessen. The approaches to several of the problems were, to a large extent, due directly to him. In later stages of the work, Professor H. O. Hartley was particularly helpful in giving advice and assisting with topics requiring additional work. The author also wishes to thank Professors Edna Douglas, Gerhard Tintner, W. D. Shrader, and N. V. Strand for reading a draft of the manuscript and suggesting improvements.

The interest and support of the Soil Conservation Service of the United States Department of Agriculture in this project is also gratefully acknowledged. In that agency, John Barnard and Henry Adams, in particular, were helpful in supplying information and materials needed in the research.
A. Relationship of the Variance Function to the
Cobb-Douglas Production Function

The variance function used in this study was given in Equation 13. It is

\[ S_i^2 = S_i^2 g_i \]

where \( a_i \) is the size of unit, \( S_i^2 \) is the variance between units of size \( a_i \), \( S^2 \) is the variance between elements, and \( g \) is a measure of the correlation between adjacent elements.

The variance function is obviously of the same mathematical form generally given for a Cobb-Douglas production function with a single input variable. The latter can be expressed as

\[ Y = AX^e \] (104)

where \( Y \) refers to output, \( X \) refers to a factor input, \( e \) is the elasticity of production, and \( A \) is a constant.

The elasticity of production, \( e \), is given by the relationship

\[ \frac{dY}{dX} = \frac{eY}{X} \] (105)

or

\[ e = \frac{dY}{Y} / \frac{dX}{X} \] (106)

The economic interpretation of \( e \) is that it indicates the approximate per cent increase in output \( Y \) for a 1 per cent increase in input \( X \).
It can also be used as an indication of scale returns. If $e$ is 1, then constant returns to scale are said to exist; that is, a 1 per cent increase in input will add 1 per cent to output. If $e$ is less than 1, decreasing returns to scale exist; for a 1 per cent increase in input there results a less than 1 per cent increase in output. If $e$ is greater than 1, then increasing returns to scale exist; output increases by a greater per cent than input.

Corresponding to the value of $e = 1$ in the production function is the value $g = 0$ in the variance function. When $g = 0$, an increase of 1 per cent in $a_1$, the size of unit, will increase the variance between units by 1 per cent. There are thus constant returns to scale in this case. When $g$ is greater than 0, there are increasing returns to scale; a 1 per cent increase in $a_1$ will result in an increase in $s_1^2$ which is greater than 1 per cent. Decreasing returns to scale exist if $g$ is negative; in this case, a 1 per cent increase in $a_1$ brings about a less than 1 per cent increase in $s_1^2$. Since $g$ is usually positive, the variance function will usually provide increasing returns to scale.

It will be recalled that in Section IV, Part B, the unknown parameter $g$ was estimated essentially by applying least squares to $\log a_1$. This is indeed the method usually used by economists in conjunction with the Cobb-Douglas function. It should be remembered, of course, that such a least square method assumes independent, constant variance residuals $Z_1$ in the logarithmic equation

$$\log s_1^2 = \log s^2 + (g + 1) \log a_1 + Z_1 \quad (107)$$
This assumption is approximately equivalent to assuming that

\[ s_i^2 = s^2 a_i^q q_i \]  \hspace{1cm} (108)

where

\[ Z_i = \log q_i \]  \hspace{1cm} (109)

Thus, in the original law a constant "proportional" error rather than a constant additive error is assumed. This assumption does not appear to be unreasonable for the present variance function. Discussions by economists of corresponding situations arising with the Cobb-Douglas law are helpful.

B. The Optimum Unit in Economic Terms

In Section IV, Part A, two definitions of an optimum unit were given. These were taken from Cochran (5, p. 189). In one case the optimum size of unit is the one that gives the desired precision at smallest cost. In the other case, the optimum unit gives the greatest precision for some fixed cost. Each of these definitions is analogous to a special case in economics of determining a firm's equilibrium position of maximum profits.

In general, profits are at a maximum when marginal revenue and marginal cost are equal or when the difference between total revenue and total cost is greatest. Now consider a special case in which the number of units produced is a function of their quality such that total revenue is held constant. That is, resources can be used to
produce either a few high-quality units each of which has a high selling price or many low-quality units each having a low price. In this case where total revenue is constant, maximum profit is clearly obtained at the output for which total cost is a minimum. The analogy of this special case to the sampling problem of determining the size of unit which gives desired precision for the smallest cost is as follows: the constant revenue is the specified amount of information (the latter being defined as the reciprocal of variance), the number of units produced is the number of units sampled, the quality of units produced is the size of the sampling units, and the varying cost of producing a unit of different quality is the varying cost of sampling a unit of different size.

A similar analogy exists between the economic problem of maximizing profit when total production costs are constant and the sampling problem of minimizing variance (or maximizing information) when total sampling costs are constant. In this case, the revenue which is to be maximized is analogous to information which is to be maximized or variance which is to be minimized. The analogy for each of the other concepts is the same as in the first definition.

An approach will now be given which is more general in nature than the two cases just outlined. The variance of the estimate of the population total was given in Equation 28. It is

$$V(\hat{Y}_i) = \frac{A^2 S^2}{n_i a_i^{1.5}}$$

If the number of units of information is given by the reciprocal of
variance and if each unit of information is valued at an amount $d_2$, then the total value of information is

$$I(Y_i) = \frac{d_2 n_i a_i^{1-g}}{A^2 S^2}$$  \hspace{1cm} (110)

The total cost of obtaining this information was given by Equations 25 and 26 and is

$$E = d_0 n_1 + d_1 n_1 a_1$$  \hspace{1cm} (111)

It is now proposed to find the optimum size of unit under the assumption that the total area sampled is a constant, that is, $n_1 a_1 = k$. Since the product of $n_1$ and $a_1$ is constant, the determination of one of these variables exactly determines the other. For convenience, Equations 110 and 111 will be put in terms of $n_1$ to form the following function to be maximized:

$$\phi = \frac{d_2 k^{1-g} n_1^{g}}{A^2 S^2} - d_0 n_1 - d_1 k$$  \hspace{1cm} (112)

The maximum of $\phi$ is at the point where

$$\frac{d\phi}{dn_1} = \frac{d_2 k^{1-g} n_1^{g-1}}{A^2 S^2} - d_0 = 0$$  \hspace{1cm} (113)

or where marginal information is equal to marginal cost, the latter being constant in this case.

Equation 113 can be solved for $n_1$ and then for $a_1$ by means of the relationship $n_1 a_1 = k$. Details will be omitted. The solution for $a_1$ is, say, $a_1^*$, given by
The answer for the optimum sampling unit is independent of \( d_{\perp} \). This is due to the fact that \( d_{\perp} n_{1} a_{i} = d_{\perp} k \) is a constant cost for all competitive sizes of unit. The solution is also independent of the value of \( k \), the total area sampled. The latter independence occurs only because the variance formula does not include a finite population correction.

From the point of view of sample surveys this problem is somewhat unusual in two respects. First, it is not usually regarded necessary to keep the total sampled area at a constant value. Second, it will usually be found difficult to attach a monetary value to the amount of information (reciprocal of variance).

C. A Stochastic Treatment of the Optimum Unit

The optimum unit as used in this study was given in Equation 30 and will be denoted here by

\[
\begin{align*}
    a_{i}^{*} &= \left( \frac{d_{o} A^{2} S^{2}}{gd_{2}} \right) \frac{1}{1-g} \\
    \text{Equation (114)}
\end{align*}
\]

The values of \( d_{o}, d_{\perp}, \) and \( g \) were used as if they were known. Actually, each was obtained as an estimate of an unknown parameter of a linear regression equation. The fact that these should not be considered as fixed will now be taken into account through an examination of the variance of \( a_{i}^{*} \).
Equation 30 can be written as

\[ a_i^* = \left( \frac{d_o}{d_1} \right) \left( \frac{1}{g} - 1 \right) \]  

(115)

Since values of \( g \) and values of \( d_o \) and \( d_1 \) were estimated from different sources of information, they are independent. The variance of \( a_i^* \) will thus be the variance of the product of two independent variables. The formula for the estimated variance of \( a_i^* \) is the following:

\[
\nu(a_i^*) = \left( \frac{d_o}{d_1} \right)^2 \nu \left( \frac{1}{g} \right) + \left( \frac{1}{g} \right)^2 \nu \left( \frac{d_o}{d_1} \right) + \nu \left( \frac{1}{g} \right) \nu \left( \frac{d_o}{d_1} \right)
\]  

(116)

Equation 116 would give the exact variance of \( a_i^* \) if true variances were used for the \( \nu \) symbols and expected values were used for the two squared quantities. Only the estimated variance of \( a_i^* \) will be dealt with here.

Consider first the estimated variance of \( \frac{1}{g} \). It is given approximately by the expression

\[
\nu \left( \frac{1}{g} \right) = \frac{s_E^2}{\sum_{i=1}^{n} (\log a_i - \overline{\log a})^2}
\]  

(117)

where

\[
s_E^2 = \sum_{i=1}^{n} (\log EMS_i - \log EMS_i')^2
\]  

(118)

and

\[
\overline{\log a} = \frac{1}{n} \sum_{i=1}^{n} \log a_i
\]  

(119)

In Equation 118, the quantity \( \log EMS_i - \log EMS_i' \) represents the difference between an observed value of the dependent variable and its corresponding estimated value from the regression equation whose form is given in Equation 22.

The estimated variance of \( \frac{d_o}{d_1} \) is given approximately by
The quantities $c_i$ and $a_i$ are those used in fitting Equation 25, for which the $c_i$ are the resulting estimated values. The independent variable is usually regarded as fixed in regression analyses. In the present instance the $a_i$ values were purposely chosen and hence were, in fact, fixed. The estimated variance of $a_i^*$ is then given by

\[
v(a_i^*) = \left( \frac{d_0}{d_1} \right)^2 \frac{s_E^2}{s_{\text{log} a_i - \text{log} a}^2} \\
+ \left[ \left( \frac{1}{g} \right)^2 + \frac{s_E^2}{g^{\frac{1}{2}} i^{\frac{1}{2}}} (\text{log } a_i - \text{log } a)^2 \right] \left( \frac{s_E^2}{d_1^2} \left( \frac{1}{n} + \frac{c^2}{a_i^*} \frac{n}{i^{\frac{1}{2}}} (a_i - a)^2 \right) \right)
\]  

(124)

As an example of the use of Equation 124 consider the characteristic 1.5 - 4% slope in Lawrence County, Illinois. The value of $g$ was .56 and from Table 22 the values of $d_0$ and $d_1$ were 6.5355 and .04540, respectively. From Equation 30 the optimum value of $a_i$ is 113 acres. The variance was calculated from Equation 124 and was found to be 4,742.8. The standard error is thus 68.9 and the relative standard error of the optimum value of $a_i$ is 61 per cent.
It should be pointed out that Equation 124 provides a variance for the optimum size of unit which depends on the particular method by which estimates of both $g$ and the cost coefficients, $d_o$ and $d_1$, were estimated. Since only very small sample sizes were available, it is not surprising that large estimated variances result. However, the formula will be useful in situations in which larger sample sizes are available. The possibility cannot be ruled out that, in other situations, the data available for estimation of $g$, $d_o$, and $d_1$ might not permit the two regression analyses here employed. In the latter case, Equation 124 would require modification.