AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523

BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY

1. SUBJECT CLASSIFICATION

A. PRIMARY: Economics

B. SECONDARY: Agricultural Economics

2. TITLE AND SUBTITLE

Area Frame Sampling in Agriculture

3. AUTHOR(S)

Houseman, E.E.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1975	84 p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS

Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

8. SUPPLEMENTARY NOTES *(Sponsoring Organization, Publishers, Availability)*

(In U.S. Department of Agriculture Statistical Reporting Service, SRS No. 20)
(Available in English and Spanish; Spanish, 107 p.: PN-AAC-140)

9. ABSTRACT

An over-all view of area frame sampling (a means of collecting agricultural data), including the construction of area sampling frames and the selection of area samples. Concepts and general principles of area sampling, rather than specific applications, are discussed. Sampling frames should be constructed in recognition of the fact that agriculture is composed of numerous subpopulations that must be sampled. A sample designed efficiently for one subpopulation might be of little value for another. Thus several sampling frames might be required or if a single sampling frame is to be constructed, it probably should be multi-purpose. The development of improved sampling frames is called for by: 1) the trend toward larger, more specialized farms, 2) the general demand for more accurate statistics, and 3) the need to keep sample sizes and costs as low as possible. Also, to some degree, sample size is related inversely to capability of controlling non-sample error, which also is in favor of efficient sampling to keep sample sizes as small as possible. Survey plans should include provision for studies of sampling variance, response errors, coverage errors, and costs. Such studies should provide a continuing basis for adjusting the allocation of resources in an effort to achieve maximum accuracy at a given cost.

10. CONTROL NUMBER

PN-AAC-062 in English. PN-AAC-140 in Spanish

11. PRICE OF DOCUMENT

12. DESCRIPTORS

13. PROJECT NUMBER

14. CONTRACT NUMBER

RSSA/USDA 7-74 GTS

15. TYPE OF DOCUMENT

AID 590-1 (4-74)

AREA FRAME SAMPLING IN AGRICULTURE

by

Earl E. Houseman
Mathematical Statistician

Statistical Reporting Service
United States Department of Agriculture

Washington, D. C. 20250 November 1975

FOREWORD

This is the second document written by Earl E. Houseman under the auspices of AID, SRS, and the International Statistical Programs Center of the Bureau of the Census, with which SRS is cooperating. The first was "Expected Value of a Sample Estimate," published by SRS, September 1974. Mr. Houseman is among the first statisticians who worked on the application of area sampling in agriculture. He also draws on years of experience associated with the development and refinement of the area frame sampling methodology currently used by the Statistical Reporting Service.

This document was developed as part of a continuing effort to provide improved materials for teaching and reference in the area of agricultural statistics for foreign students and for development of staff working for these agencies.

WILLIAM E. KIBLER
Administrator

Area Frame Sampling in Agriculture

Preface

This publication presents an overall view of area frame sampling, including the construction of area sampling frames and the selection of area samples. Resources for the construction of area sampling frames and the conditions involved in the application differ widely. The objective is to present ideas about how to do area sampling and give emphasis to important factors that need to be considered. Concepts and general principles of area sampling, rather than specific applications, are discussed. Technically sound sampling concepts help form a solid foundation for any sample survey. If the concepts do not fit, the statistician should try to find more realistic technically sound concepts. Survey procedures evolve from concepts. Thus a full understanding of concepts provides a basis for decisions on many practical operational problems which help to assure good results. Tenure and patterns of agricultural production differ widely among countries and even regions within countries. This means that sampling plans must be tailored to individual situations and survey purposes. In other words, be cautious about copying the details of a plan that worked well in one situation and applying it to another without careful study.

In developing an overall view of area sampling it is necessary to include many general statements. The reader should be aware that some contradictions and exceptions can usually be found. Many statements will reflect goals, recognizing that resources or conditions are often such that very little can be done immediately toward achieving the ultimate goals. Expertise in sample design, familiarity with local conditions involved in the application of area sampling, survey experience, and the quality and detail of available maps regarding roads, landmarks, and land use are important factors in the development and effective use of area sampling.

The intended audience is students of sampling and persons who might be considering area sampling as a means of collecting agricultural data. It has been assumed that most readers will have at least an elementary knowledge of sampling theory and some experience in agriculture. However, interested readers without formal training in sampling methods should find this description of area sampling useful.

Contents

Page

1. Introduction 1

1.1 Definitions 1
1.2 Early Development of Area Sampling 3

2. Some Key Features of Area Sampling 4

2.1 Versatility 4
2.2 Coverage 4
2.3 Updating 5
2.4 Efficiency 5
2.5 Area Frames as a Complement to List Frames 6

2.5.1 List frame nearly adequate 6
2.5.2 List frame covers part of population 7
2.5.3 Adequate list frame not available 7

3. Size of Segment 7

3.1 Sampling Variance as a Function of Segment Size 7
3.2 Sampling Variance as a Function of Percentage Reporting 10
3.3 Defining Segments to Minimize Sampling Variance 11
3.4 Optimum Size of Segment 12

4. Definitions of Area Sampling Units 13

4.1 Introduction 13
4.2 The Closed-Segment Method 13
4.3 The Open-Segment Method 15

4.3.1 Farm-operator approach 16
4.3.2 Farm approach 17
4.3.3 Problems with establishing a definition of farm headquarters 18
4.3.4 Some general observations 20

4.4 The Weighted-Segment Method 21

4.4.1 Algebraic description of the weighted segment 22
4.4.2 Estimators and their variances 23
4.4.3 Ratio estimation 24
4.4.4 Unequal probabilities of selection 25
4.4.5 Domain estimation 26

5. Numerical Illustration 26

5.1 Domain Estimation and the Weighted Segment 30
5.2 Sampling Variance 32

Page

6. Discussion of the Three Definitions of Area Sampling Units..... 34

6.1 Closed Segment vs. Open or Weighted..... 35
6.2 Open- vs. Weighted-Segment Methods..... 36

6.2.1 Sampling variance and costs..... 37
6.2.2 Coverage error..... 39
6.2.3 Combination of methods..... 41

7. Construction of Area Sampling Frames..... 42

7.1 Background..... 42
7.2 Frame-Unit Specifications..... 43
7.3 Auxiliary Information and Its Use..... 45

7.3.1 Control of segment size..... 45
7.3.2 Stratification and the definition of frame units..... 46
7.3.3 Selection of auxiliary data about frame units..... 49

7.4 Maps for Frame Construction..... 50
7.5 Division of Frame Units into Segments..... 51

8. Frame Construction--Illustration No. 1..... 53

8.1 A Survey of Crop Acreages..... 54
8.2 A Survey for Economic Data..... 57
8.3 A Beef Cattle Survey..... 60

9. Frame Construction--Illustration No. 2..... 61

9.1 A Survey of Crop Acreages..... 62
9.2 A Survey of All Farms..... 64

10. Summary and a Brief Look Forward..... 64

AREA FRAME SAMPLING IN AGRICULTURE

1. Introduction

The concepts of area frame sampling are very simple: divide the total area to be surveyed into N small blocks, without any overlap or omission; select a random sample of n blocks; obtain the desired data for reporting units of the population that are in the sample blocks; and estimate population totals by multiplying the sample totals by $\frac{N}{n}$. The simplicity of the idea is in striking contrast to the complexity of successful application of the concepts. But a high proportion of the problems found in the application of area sampling ("area sampling" will be used as a shortened term instead of "area frame sampling") in agriculture are characteristic of the survey populations and therefore common to all survey methods, sampling or census. However, survey methods differ considerably with regard to effectiveness, or potential effectiveness, in coping with practical problems that exist.

The minimum requirement for the application of area sampling is maps for dividing the population into small area sampling units that have boundaries which can be accurately identified on site by an interviewer. There are three important conditions involved in the application: (1) The reporting units must be defined to serve the purpose of the survey, (2) there must be practical means of associating reporting units with the area sampling units, and (3) area sampling should compare favorably with alternative sample survey methods that are feasible.

1.1 Definitions

Before proceeding with the discussion, some concepts and definitions will be reviewed:

Reporting units are the individual elements or units that compose a population for data collection (reporting) purposes. There is no standard definition of a reporting unit. Typically, one questionnaire is filled out for each reporting unit. In the discussion that follows, the specific meaning of "reporting unit" will usually be a "tract," which is defined later, or a farm (holding).

Sampling units are the units that a survey population is divided into for sampling purposes. They are the units subject to random selection. Usually, each reporting unit in the population is associated with one and only one sampling unit. In area sampling, the number of reporting units in a sampling unit varies.

A sampling frame is a complete list (or specifications that would establish a complete list) of sampling units that cover a population. It provides access to a population in ways that enable probability sampling. If each reporting unit is associated with one and only one sampling unit and if there are M_i reporting units associated with the ith sampling unit, the population consists of

$M = \sum_{i}^{N} M_i$ reporting units, where N is the total number of sampling units in the population.

The term "sampling frame" suggests that a frame is used only for sampling purposes. Actually, a frame is also needed for a census, which involves collecting data for all units of the frame. For example, the equivalent of area sampling has been used for a long time in taking censuses--perhaps since the first censuses were taken. Enumeration districts are defined and one or more field investigators enumerate each district. The list of ED's (enumeration districts) is the area frame for taking a census. Incidentally, there are sample surveys and census surveys, the only difference being that a census survey is an attempt to enumerate completely the frame, rather than a sample selected from the frame.

A segment is a piece of land with boundaries delineated on a map. In area sampling, the total area for the population to be sampled is divided into segments. In addition to meaning a piece of land, "segment" is used in sampling terminology instead of "area sampling unit". "Segment," meaning area sampling unit, refers to the aggregate of the reporting units that compose an area sampling unit. Whether "segment" refers to a piece of land delineated on a map or to an area sampling unit (group of reporting units) should be clear from the context.

Sampling efficiency refers to the sampling variance for one plan (that is, a specific method of sampling and estimation) in comparison with the sampling variance for another. Sampling variances are usually compared under an assumption of equal sampling fractions or of equal costs. Unless otherwise specified, "sampling efficiency" will refer to comparison of alternatives under an assumption of equal sampling fractions.

Cluster sampling is the general term for sampling plans wherein the sampling units are groups (clusters) of reporting units. An area sampling unit is a "cluster" of reporting units associated with a segment. In other words, area sampling is a form of cluster sampling and the theory of cluster sampling applies.

A survey population is the population actually sampled (or completely enumerated). It is defined by the sampling frame and the procedures for using it. Sometimes a distinction is needed between the "survey population" and a "target population."

A target population is the population which, given full freedom of choice, one might wish to survey; but, for various practical reasons, the population actually sampled could be different from the target population. For example, one might prefer to estimate the total production of a crop, but decide to omit some regions where the amounts produced are very small.

In theory, estimates (statistical inference) from the sample pertain to the survey population, not the target population. For an excellent discussion of sampling frames and populations, and for an overall view of sampling and of inference from samples, the reader is referred to the first four chapters of

Deming's book.[1/] The introductory chapters of other books on sampling also discuss general principles of sampling and estimation.

Sampling variance is the variance of an estimate from a sample.

Design efficiency, sometimes called "design effect," refers to the sampling variance corresponding to any particular sample design and estimator in comparison with the sampling variance corresponding to some other sample design or estimator. Simple random sampling is often used as the base of comparison. In the discussion that follows, "sampling efficiency" will sometimes be used instead of "design efficiency."

Coverage error refers to omission and duplication of reporting units, including incorrect determination of the land area that composes a reporting unit.

Response error refers to accuracy of data for any particular reporting unit.

Some statisticians would define coverage and response error somewhat differently but these definitions are convenient when discussing area sampling.

1.2 Early Development of Area Sampling

The first ideas of area sampling in the United States appear to have been in the context of purposive sampling. A selection of areas about the size of MCD's (minor civil divisions) or ED's (census enumeration districts) was sought which would be a permanent sample that would permit accurate measurement of year-to-year changes. MCD's and ED's were recognized units that had been defined on maps. Unpublished data about each MCD from previous censuses were available for sampling purposes. Results from investigation of the MCD or the ED as a sampling unit were not encouraging. The size of sample required for acceptable levels of sampling variance was regarded as much too large. At that time very little was known about the relation between the size of sampling units and sampling efficiency, but early investigations indicated that sampling units probably should be much smaller than MCD's.

We now know that, in general, a sampling unit as large as an ED (75 to 100 farms or more) is simply very inefficient. The degree of inefficiency is related to the size of the sampling unit (the number of reporting units in the sampling unit) and the extent to which adjacent or neighboring farms (reporting units) tend to be alike. Since agricultural resources and environment tend to be similar in a small locality, characteristics of farms within a locality have generally exhibited a strong tendency to be alike. This indicates why, for example, a 2-percent sample of large area sampling units generally has much larger sampling variances than a 2-percent sample of small sampling units that are much more widely distributed. That is, sample data in a sample of 2,500 farms, for example, would come from only 25 locations if each area sampling unit contains 100 farms; but, if each sampling unit is composed of 5 farms, that would be 500 locations where data would be collected and the sampling variances would be much lower.

1/ Deming, W. Edwards, "Sample Design in Business Research," John Wiley and Sons, 1960.

For agricultural surveys, the first significant test of probability area sampling in the United States, using small areas as sampling units, occurred in Iowa.2/ Two surveys, one at the end of 1938 and the other at the end of 1939, were conducted, using quarter sections as area sampling units. (Quarter sections are approximately square, 1/2 mile on a side, and contain approximately 160 acres.) At that time, the average number of farms per quarter section was about 0.9. The sample for each survey represented the entire State and was a widely dispersed, geographically stratified random sample of about 900 quarter sections. The sampling fraction was less than 1/2 of 1 percent.

Considering the small size of the sample, the survey results were very encouraging. The relative standard error (coefficients of variation) of estimates for important farm characteristics were generally less than 4 percent. Also, it was possible to compare estimates from the area samples with other sources of information, including a farm census conducted each year by the State of Iowa, and the Federal census of agriculture that related to 1939. Three things, (1) the information obtained about random sampling error, (2) the experience in the field regarding sources of error that were not related to sampling, and (3) comparisons of the sample estimates with other sources of information, strongly suggested at that time that much attention must be directed in the future to minimizing error from sources other than sampling. From this and other experiences with probability sampling, a new perspective of the total error in estimates from surveys started to develop.

One outgrowth of this test of area sampling was the development, by 1945, of an area sampling frame for all States.3/

2. Some Key Features of Area Sampling

2.1 Versatility

Possible uses of area sampling are unlimited. The survey population could be composed of reporting units that are households, persons, farms, plants, animals, cotton gins, suppliers of agricultural inputs, tractors, tracts of land, grain storage facilities, processors of agricultural products, or any other definable reporting units that can be uniquely associated with segments. Adaptability to particular uses, and versatility, are strong attributes of area sampling. Many needs for information have been filled where area sampling was the only means available for selecting a probability sample.

2.2 Coverage

Conceptually, an area sampling frame is always current and complete with regard to any definition of a reporting unit. For example, an area sample of farms is a sample of farms as they are defined and exist at the time of the survey. In other words, if a random sample of 1/5 of all segments in the

2/ Jessen, Raymond J., "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," Iowa State University, Research Bulletin 304, June 1942, Ames, Iowa.

3/ King, A.J. and Jessen, R.J., "Master Sample of Agriculture," Journal of the American Statistical Association, Volume 40:38-46, 1945.

population is selected, the sample of segments is "expected" to contain 1/5 of the reporting units in the population regardless of how the reporting units are defined. (The word "expected" is used in the sense of mathematical expectation.) To further clarify the point, consider the estimator $\frac{N}{n}\Sigma x$. The number of segments, N, in the population and the number, n, in the sample are known. The sample total, Σx, is the total of characteristic X for all reporting units associated with the sample of n segments. Hence, the sample can be expanded regardless of how a reporting unit is defined. Notice that one does not need to know the number of reporting units in the population in order to apply area sampling. In fact, from an area sample, one can estimate the number of reporting units in the population. One estimator is $\frac{N}{n}(r)$, where r is the number of reporting units found in the sample of n segments.

The preceding paragraph pointed out that an area sampling frame is conceptually complete. The term "conceptually complete" needs to be stressed because, in practice, coverage error is a major problem. If one selects an area sample and expects to use $\frac{N}{n}$ as an expansion factor, the fieldwork of identifying and associating reporting units with each segment in the sample must be performed with great care. If the association of farms with segments is incomplete, or is not done correctly, the actual sampling fraction with regard to the number of farms in the sample in relation to the population total will not be $\frac{n}{N}$. Therefore, $\frac{N}{n}\Sigma x$ will not be an unbiased estimate of the population total.

2.3 Updating

An area frame does not become out-of-date in terms of coverage of a population, unless the population extends into areas not covered by the frame. Changes in land use, or number and location of reporting units, have a bearing on the sampling variance but do not introduce bias. Some boundaries of sampling units will lose identity as time passes, which could increase the potential for bias as a result of greater ambiguity about boundary locations. There are two possible reasons for updating an area frame: (1) To maintain or achieve improvements in sampling efficiency, or (2) to introduce updated or new maps to achieve better boundaries of sampling units. Parts can be updated as needed.

2.4 Efficiency

The characteristics of a sampling frame have an important bearing on the quality of results from a survey. Serious biases, low sampling efficiency, or both might be the result of deficiencies in the sampling frame. For minimum coverage error, statisticians would like to have an up-to-date list of all farms (complete and without duplication) for sampling purposes. But agricultural characteristics vary widely among farms. Consequently, to enable the design of efficient samples for a wide range of purposes, it is important to have some information about each farm on the list. For example, it is generally very helpful to have farms classified by: (1) Type (for example, whether the farm is a livestock farm, a fruit farm, etc., or perhaps whether some specified commodities are produced on the farm), and (2) size (preferably a relevant measure of size

corresponding to each type of farm). Obtaining and maintaining a complete and up-to-date list of farms, classified by type and size, is a major undertaking that might be regarded as a goal to be achieved to the extent feasible.

The attributes of a list frame (list of farm operators) that make it most effective for sampling purposes also apply to an area sampling frame. That is, for designing area samples, one would like to have information on the type and size of each segment (sampling unit) in the population. But, construction of a sampling frame (list or area) that will enable a high level of sampling efficiency could require a major investment, unless relevant information exists which can be easily incorporated in the sampling frame. Technical analyses and considerations of costs, variances, and biases can be very helpful in determining the merits of alternative, feasible specifications for a sampling frame. If a good background of experience does not exist, there should be adequate testing of feasible alternatives before setting final specifications and undertaking the entire job of constructing a sampling frame. In fact, some testing is generally advisable even though there has been much experience to build on.

2.5 Area Frames as a Complement to List Frames

A complete up-to-date list of farms, including relevant information about the farms, is highly desirable for sampling purposes and has strong advantages with regard to sampling efficiency and cost. But, the coverage of list frames rapidly becomes out of date. Moreover, area sampling is often needed because of deficiencies in, or absence of, list frames. As pointed out above, an area frame is always conceptually complete. There are three general situations pertaining to the application of area sampling:

2.5.1 List frame nearly adequate. Suppose a list of farms exists or there is a means of developing a list that defines a survey population that is nearly the same as the target population. In this case, the survey population defined by the list might be accepted and a sample selected from the list would be used for the survey. As a means of checking on the adequacy and completeness of the list, an area sample might be used. This would involve matching the list with reporting units found in the area sample. If the list is complete, all reporting units in the area sample should be on the list. But matching involves many problems, because a reporting unit is not always defined and identified in the same way. Discussion of matching problems is outside the scope of this publication.

Consideration of costs, sampling efficiency, and innumerable technical factors could lead to a decision to use a list frame for sampling even though the list frame defines a survey population that differs somewhat from the target population. For example, consider a survey of wheat producers. Suppose a list of wheat producers exists which is believed to be adequate, but an investigation of its coverage would be appropriate. Area sampling could be used, but it would involve contacting all farmers in the area sampling units to find those who are producing wheat. If the production of wheat is widely scattered and the proportion of farmers producing wheat is small, economics strongly suggest sampling from the list. In this case, the survey might be based on a sample from the list and an area sample could be used to obtain information about the adequacy or quality of the list.

2.5.2 List frame covers part of population. A list frame might be very good but cover only a part of the population to be surveyed. If the list frame covers a major or important part of the population and is satisfactory, except for incompleteness, a sample from it might be selected. To get representation of the part of the population not included on the list an area sample could be used. This is an example of multiple-frame sampling, which is concurrent use of two or more sampling frames. For some surveys multiple-frame sampling has important advantages, but those advantages are often very difficult to realize when estimating population totals, owing to practical difficulties of accurately determining which reporting units in the area sample are also in the list frame.

2.5.3 Adequate list frame not available. A list frame might not exist and it might not be feasible to create one that provides a satisfactory sampling frame for even a part of the population. In this case, area sampling is the only possibility for selecting a probability sample.

In the first two situations (2.5.1 and 2.5.2), reporting units enumerated in the area sample must be matched with reporting units in the list frame. Such uses of area sampling are appropriately discussed under multiple-frame sampling which is outside the scope of this publication. Discussion will be limited to the third situation.

3. Size of Segment

3.1 Sampling Variance as a Function of Segment Size

"Size of segment" is a general term. It might refer, for example, to the land area of a segment, to the number of farm operators living in a segment, to the number of dwelling units in a segment, to the amount of irrigated land, or to the amount of land under fruit trees. However, in this section, "size of segment" will be discussed in terms of the number of farms "in" a segment. A farm is "in" a segment if its headquarters is within the boundaries of the segment. This will be discussed in Section 4.3, The Open-Segment Method.

Factors to consider when defining segments include: Sampling variance, costs, problems associated with segment boundaries, topographic detail on available mapping materials, and the method of associating farms with segments. Cost considerations have often given rise to strong intuitive impressions that favor sampling units that are larger than they should be. This evidently comes from the fact that, for a given cost, more farms can be included in the sample when the sampling units are large. Optimum size of segment will be discussed after a brief review of the situation regarding the relation between sampling variance and size of segment.

To emphasize the difference in sampling variance for large segments in comparison with small ones, some results from an unpublished analysis of data from a farm census in the State of Wisconsin are presented in table 1. In this census, farms were enumerated by townships. ("Township" is the name for the smallest political subdivision in the State). Thus it was possible to compute sampling variances for area sampling when sampling units are townships and to compare the results with variances when individual farms are the sampling units.

Table 1.--Relative variance of townships as sampling units compared with individual farms 1/

Item	Percentage of farms reporting 2/	Average number of farms reporting per township 3/	Relative variance among all farms 4/	Relative variance among farms reporting 5/	Variance among townships relative to variance among farms 6/	
					Method 1	Method 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Farmland	100.0	69.5	0.73	0.73	---	---
Alfalfa	70.2	48.9	1.28	0.59	53.7	19.4
Corn	66.5	46.2	3.33	1.88	26.0	14.2
Pasture	61.4	42.7	3.04	1.49	40.2	29.5
Milk cows	58.7	40.8	1.34	0.37	43.6	10.9
Beef cattle	26.4	18.4	16.1	3.50	9.0	7.2
Hay for silage	15.9	11.1	13.2	1.25	8.1	4.9
Cattle marketed	7.5	5.2	163.9	11.36	3.7	3.5
Soybeans	4.1	2.8	73.4	1.99	9.5	9.3
Peas	3.1	2.2	134.2	3.24	5.0	4.6
Sheep	2.7	1.9	138.2	2.76	2.1	1.9
Spring wheat	1.2	.8	488	4.74	4.1	4.1
Potatoes	0.7	.5	789	4.76	3.2	3.2
Snap beans	0.2	.2	1,501	2.45	3.9	3.9

1/ This table was compiled from an unpublished analysis of 1970 data from the State farm census in Wisconsin.

2/ Percentage of farms for which $X_i > 0$, where X_i is the value of characteristic X for i^{th} farm.

3/ Number of farms in the State for which $X_i > 0$, divided by the total number of townships in the State.

4/ The relative variance among all farms is $\frac{\Sigma^{F}(X_i-\overline{X})^2}{\overline{X}^2(F-1)}$, where F is the total number of farms in the State.

5/ Relative variance among farms reporting is the relative variance of X among the farms reporting the item; that is, farms for which $X_i = 0$ are not included in the calculation of $\overline{X}$ or in the sum of squares $\Sigma(X_i-\overline{X})^2$.

6/ See text.

The average number of farms per township was 69.5, and there was a total of nearly 102,000 farms in the State. Columns (2), (3), and (4) of table 1 are explained in the footnotes to the table. Column (5) was included to emphasize an important point that will be discussed later. Columns (6) and (7) are discussed in the following paragraphs. They show the ratios of sampling variances for townships to sampling variances for farms.

To compare the sampling variances for townships with the sampling variances for farms, simple random sampling was assumed. For townships, variances for two different estimators were computed. The first was a mean per township estimator:

$$x_1' = \frac{T}{t} \Sigma^{t} x_i$$

where T is the number of townships in the State, t is the number of townships in the sample, and x_i is the total of characteristic X for the ith township in the sample. The second estimator is a ratio estimator:

$$x_2' = F \frac{\Sigma^{t} x_i}{\Sigma^{t} f_i}$$

where F is the total number of farms in the State, and f_i is the number of farms in the ith township in the sample. The ratio estimator, x_2', was included because it removes from the sampling variance at least part of the variation among townships that is correlated with variation in size (number of farms) of the townships.

The estimator for a simple random sample of farms was:

$$x_3' = \frac{F}{f} \Sigma^{f} x_j$$

where f is the number of farms in the sample and x_j is the value of characteristic X for the jth farm in the sample.

We want to compare the sampling variances for townships and farms, assuming the sampling fractions are the same; that is, when f = 69.5t. Thus, column (6) is the variance of x_1' divided by the variance of x_3', assuming f = 69.5t. Similarly, column (7) is the variance of x_2' divided by the variance of x_3'.

The first entry in column (6), for example, means that for alfalfa the sampling variance for townships using the first estimator, x_1', is 53.7 times larger than the sampling variance for farms. Columns (6) and (7) may also be interpreted in terms of sample sizes needed for equal precision (that is, equal sampling error). Taking the first estimator and alfalfa as an example, a simple

random sample of 100 farms has the same precision as a sample of 5,370 farms when townships are the sampling units. It would take a sample of approximately 77 townships to get a sample of 5,370 farms. The difference is much less for other characteristics.

Notice that the sampling variance for townships relative to the sampling variance for individual farms is related to the proportion of farms reporting the commodity (compare columns (6) and (7) with column (2)). For some commodities there is an average of less than one farm reporting per township. (See column (3)). If size of township is measured by number of farms reporting, then a township is a "small" sampling unit for some commodities, namely the commodities at the bottom of the list. The production of these commodities is widely scattered. For such commodities the township as a sampling unit has less loss of efficiency, as shown in the last two columns of table 1. The results clearly indicate a very large loss in sampling efficiency when area sampling units have large numbers of farms reporting, but other things need to be considered.

3.2 Sampling Variance as a Function of Percentage Reporting

Columns (4) and (5) of table 1 were included because they reflect an important general situation that needs to be recognized in sampling. Based on simple random sampling of all farms, column (4) shows that the relative variance of various items is closely related to the proportion of farms reporting the item. (For a definition of relative variance see footnote 4/, table 1.) Column (5), as explained in the footnote, shows the relative variance when all values of $X_i = 0$ are eliminated from the variance calculations. It is the relative variance among farms reporting the item. There is little or no relation between the variances in column (5) and the percentage reporting, column (2).

The relation between the relative variance of all values of X including zeros and proportion reporting has been shown in sampling theory 4/. In fact, the relation between columns (4) and (5) is as follows:

$$V_4^2 = \frac{V_5^2+(1-P)}{P} \qquad (1)$$

where V_4^2 is the relative variance among all farms, column (4), V_5^2 is the relative variance among all farms reporting, column (5), and P is the proportion of farms reporting, that is, column (2) expressed as a decimal fraction rather than as a percentage.

Suppose a simple random sample of f farms is selected and that x_3 is the estimator of the population total. The relative variance of x_3' is

$$\frac{V_4^2}{f} = \frac{V_5^2+(1-P)}{fP}$$

4/ Hansen, Hurwitz and Madow, "Sample Survey Methods and Theory," Vol. 1, p. 122, John Wiley & Sons, 1953.

assuming that the correction for finite population, $(\frac{F-f}{F})$, is small enough to be ignored. We have noted that V_5^2 varies by a relatively small amount from one commodity to another. Hence, the value of P is a major important factor in determining the relative variance of x_3', the estimate from a sample.

Equation (1) also applies to area sampling, assuming a simple random sample of segments. Suppose there are N segments in the population and that N′ is the number of segments in the population for which X_i is greater than zero, where X_i is the total of X for the ith segment in the population. Then $P = \frac{N'}{N}$, and V_5^2 is the relative variance of X_i among the N′ segments for which X_i is greater than zero. Suppose that a simple random sample of n segments is selected. The relative variance of the estimated total,

$$\frac{N}{n} \sum^{n} x_i, \text{ is } \frac{V_5^2+(1-P)}{nP},$$

assuming that the correction for finite population is small enough to be ignored. Without getting involved in a full explanation, this indicates that it would be undesirable to define a population of segments wherein the proportion of "zero segments" (segments that do not possess the characteristics being measured) is more than a small percentage of all segments.

Many commodities are produced on less than 20 percent of the farms and equation (1) indicates high sampling variance when the percentage is low. This points to the recognized need for what is often called special-purpose sampling; that is, developing sampling frames and designing samples that are efficient with regard to particular commodities or purposes. It is not possible in this publication to pursue various implications of this with regard to sampling agricultural populations. Briefly, it indicates including, to the extent feasible, information in sampling frames about who is producing various commodities or detailed information on where the commodities are produced.

3.3 Defining Segments to Minimize Sampling Variance.

Sampling variance is a function of the variation among segments. Therefore, one objective in defining segments should be to make the variation among segments as small as possible. It is well known, as indicated in section 3.1, that sampling variance is related to the average size of segment and to variation in the size of segment. With regard to variation in size of segment, the objective is to make the segments nearly equal in "size", where the measure of size is a variable closely related to the variables to be measured in the survey. If it is not feasible to equalize the size of segments, but a relevant measure of size is available, ratio estimation might be a possibility for reducing sampling variance that is associated with variation in the size of segments.

With regard to average size of segment, and considering <u>only</u> sampling variance, the objective would generally be to define segments so there is one reporting unit in each. For example, if the proposed survey involves only

livestock farms, the objective would be to have segments defined so there is one livestock farm in each. But available information for defining segments is usually very limited. Therefore, the degree of realization of the objective of segments of equal "size" is limited by the nature of any relevant information that might exist.

3.4 Optimum Size of Segment

A random sample of 500 segments with four farms each can be enumerated at less cost than a random sample of 2,000 segments with one farm in each. The latter will have a smaller sampling error. The optimum size of segment might be about two or three farms, depending on variance and cost functions. Accumulated experience points to very small segments; that is, small in terms of number of reporting units as defined for the survey. Optimum size is difficult to define and determine in practice, especially when estimates are calculated for many characteristics and for several domains as well as for the whole population. A difference of one or two reporting units in the average size of segments might be difficult to assess. Nevertheless, assuming that the survey cost is held constant, as segment size increases, a point is reached where the sampling variance increases rapidly. That is, small departures from optimum might be negligible but large departures could result in a serious loss of sampling efficiency. Therefore, as an objective, try to specify a segment size that is in the vicinity of optimum, unless topographic detail for delineating segments dictates otherwise. In the United States, considering variance and cost, the experience has been that the "optimum" size of segment, for many purposes, is less than the practical minimum dictated by problems associated with segment boundaries and limitations of topographic detail on maps[5/].

Optimum size of segment, as discussed in the preceding paragraph, referred to sampling variance, not to mean square error, which is a combination of sampling variance and bias. This brings us to matters of bias associated with segment boundaries. The ratio of the perimeter of a segment to its area is a function of its size and shape. The ratio is greater for small segments than large ones, hence one expects the impact of any biases associated with ambiguity about segment boundaries to be relatively greater for small segments. Also, as the size of segment decreases, topographic features suitable for use as segment boundaries become less prevalent. Therefore, in terms of mean square error, the optimum size of segment could be larger than the optimum based only on sampling variance. There is very little, if any, quantitative information available on this point. But experience strongly indicates that high priority must be given to delineating segments that have boundaries which can be positively identified by interviewers in the field. The question of average size of segment often resolves into a matter of determining the smallest average size that is practical with regard to topographic detail.

5/ Houseman, Earl E. and Trelogan, Harry C., "Progress Toward Optimizing Agricultural Area Sampling." Proceedings of the 36th Session of the International Statistical Institute, Sydney, 1967.

4. Definitions of Area Sampling Units

4.1 Introduction

It is not possible to delineate segments so that no farms will overlap segment boundaries. This is the root of many practical operating problems of associating farms with segments. In coping with such problems, three primary methods of using area sampling have evolved: Closed segment, open segment, and weighted segment. These three methods refer to three different ways of defining an area sampling unit. However, before discussing these methods we need to define "tract," which plays an important role in all three methods.

A tract is a portion or subdivision of a segment that is under one management. It is either an entire farm, part(s) of a farm, or a nonfarm area of land. That is, a tract is determined by the definition of a farm and by the boundaries of a segment. A farm is composed of one or more tracts.

With one exception, which will be discussed later, rigorous application of area sampling requires that each sample segment be divided into tracts and that all land within the segment be carefully accounted for as illustrated in figure 1. This is necessary to minimize coverage error. The description of the seven tracts in figure 1 is not intended as an illustration of the information that would need to be obtained in an actual survey. The information to be recorded and procedural detail vary with the method of applying area sampling. As references to figure 1 will be made in the following discussion, it is suggested that readers become familiar with it at this point.

Early uses of area sampling employed the open segment, but practical difficulties led to use of the closed segment whenever it was not necessary for the reporting units to be farms. For surveys in which the reporting units must be farms, only the open segment and the weighted segment are applicable.

4.2 The Closed-Segment Method

A strong virtue of the closed-segment method is its simplicity. The idea is to collect data on specific items or activities within the boundaries of the sample segments. For example, if information on land use is required, data are collected on the use of all land within the boundaries of each sample segment. Or, if information about cattle is wanted, the goal is to get information about all cattle within the boundaries of the segment at the time of the interview. Tracts as defined above are the reporting units unless some other definition of a reporting unit is more appropriate. With reference to figure 1, the "closed segment" (meaning the closed-segment method of defining the area sampling unit) is composed of all tracts A thru G. If no information about nonfarm tracts is to be collected, one could say that the closed segment is composed of six tracts: A, B, D, E, F, and G. Tract D is composed of two parts.

Where applicable, the closed segment has a major advantage, compared with the open- and weighted-segment methods, because ambiguity is eliminated about what a farm is--ambiguity that has the affect of causing coverage error due to

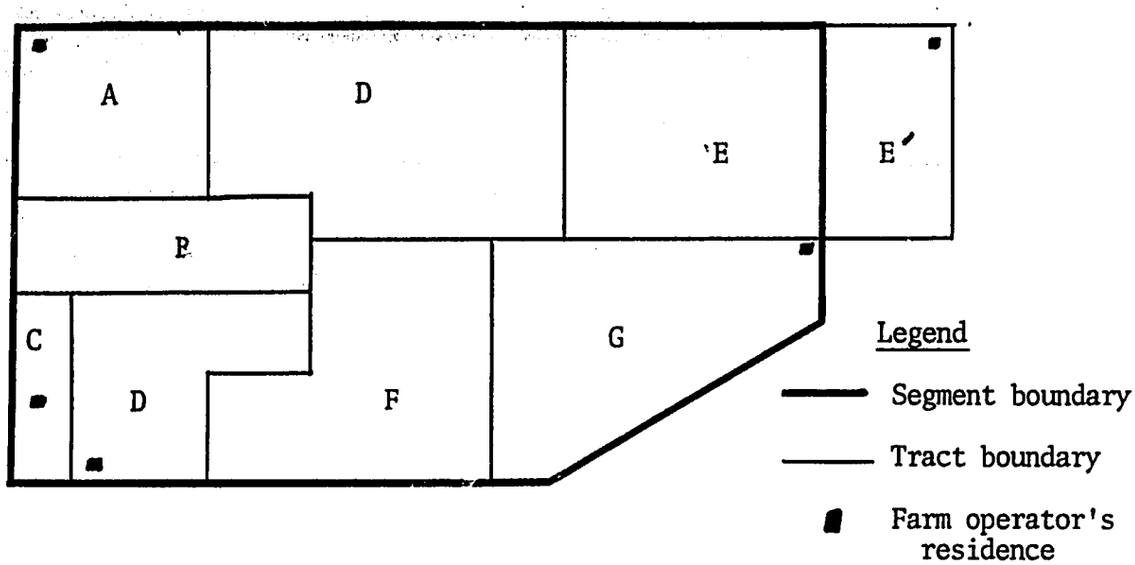


Figure 1.--Division of a segment into tracts

Description of figure 1:

Tract	Farm	Description
A	1	Tract A is an entire farm. The operator lives on his farm.
B	2	Tract B is a farm but the operator does not live on his farm or inside the segment.
C	3	Tract C is a nonfarm tract. That is, no agricultural operations are performed within it. However, one of two brothers who operate a farm lives on this tract. No part of their farm is located in this segment. But according to previously defined rules that designate one person as the "operator" of a farm, the brother living in tract C is the operator of farm number 3, rather than the brother who helps operate the farm and lives on the farm in another segment.
D	4	Tract D is composed of parcels of land at two locations within the segment. It is operated by one person who lives in the segment and has no land outside the segment.
E	5	Tracts E and E′ compose farm number 5. This is an example of a segment boundary crossing a farm and dividing the farm into two tracts. The operator lives in tract E′.
F	6	Tract F is part of farm number 6. The remainder of the farm is a tract located a few miles away from this segment. The operator lives outside the segment.
G	7	Tract G is part of farm number 7. The operator lives in the segment and on his farm.

duplication or omission of parts of farms or of whole farms. For land use, including crop acreages, the closed segment has proven generally to be much superior to the open- and weighted-segment methods, particularly if photographs are available as an aid to identifying tract boundaries. Nearly all farm operators in the United States know the acreages of their fields and, therefore, are generally able to report accurately the acreages of fields within a segment. If the operator of a tract within a segment is not available for an interview, the crops in the tract can be identified and acreages might be estimated from photographs or by other means. Therefore, response error and coverage error are relatively low. Also, the sampling variance for the closed segment is generally much lower than the sampling variance for the open segment.

Unfortunately, for many characteristics farmers are not in a position to provide accurate data pertaining to parts of their farms; that is, for tracts within segments as required by the closed-segment method. For example, an operator would probably know the man-hours of hired labor used on his farm and how much he paid for hired labor. But, if his farm overlaps a segment boundary he might have to make an inaccurate guess as to how much hired labor was used on a tract within a segment. The problem which an operator has of reporting for a tract within a segment, rather than for his entire farm, varies from virtually no difficulty in the case of crop acreages to being impracticable for most economic data such as purchases of inputs or sales of agricultural products.

Segment boundaries should follow permanent landmarks, but that is not always possible, and some landmarks change. An interviewer will occasionally find instances where a portion of a segment boundary cuts across a field. Such cases might be handled in one of two ways: (a) Have the interviewer obtain information for the entire field; then, in the office a random determination could be made to drop the entire field from the segment or to include the entire field in the segment; or, (b) if a sufficient basis exists, a preferred method is to estimate the proportion of the field that is in the segment and multiply the field total by that proportion. The interviewers might be given instructions for making such determinations, but that is usually less desirable than having them supply the necessary facts so that the disposition of such cases can be handled in the office. Office staff should be trained so they are less inclined than interviewers to introduce bias when discretion is exercised.

Since livestock can roam, some problems occur that are peculiar to livestock. For example, even though the boundary between tracts E and E′ in figure 1 is a visible landmark, it might be possible for the farmer's livestock to move between the two tracts. In that case, the operator might not know at the time of an interview exactly where all of his livestock are located with regard to segment boundaries. This case could be dealt with by using techniques like those suggested in the preceding paragraph. The open- and weighted-segment methods discussed later are also possibilities.

4.3 The Open-Segment Method

The general idea of the open-segment method is to formulate practical rules that associate every farm in the population with one and only one segment. To

do this, a unique reference point, called "headquarters," is defined and located for each farm. A farm then belongs to the segment in which its headquarters is located. Conceptually, the probability of a farm's being in the sample is clear. It is the same as the probability of selecting the segment in which its headquarters is located.

There have been two general approaches to identifying and delimiting a farm: The farm-operator approach, and the farm approach.

4.3.1 Farm-operator approach. This approach involves canvassing each sample segment for farm operators. A farm operator's residence is, by definition, the farm headquarters. Each residence (dwelling unit) within a sample segment should be visited and appropriate questions asked to determine whether anyone living in the residence is a farm operator. A questionnaire for the farm of each operator living in the segment is filled out regardless of where the farm is located. With reference to figure 1, farms numbered 1, 3, 4, and 7 are in the sample because the residences of the operators of these farms are within the boundaries of the segment. No information would be collected about the other farms.

The application of the farm-operator approach requires formulating rules that create, by definition, a one-to-one correspondence between farm operators and farms. This is needed because it is possible for more than one person to be accepted as the farm operator of a particular farm. A good example of this is a farm operated jointly by two brothers who live in different houses. Under the farm-operator approach the farm could easily be counted twice (or have a double chance of being in the sample) unless some rules that define one of the two brothers as the operator are strictly applied. For example, with reference to figure 1, two brothers operate farm number 3. One of the brothers lives outside the segment and one lives on tract C within the segment. By definition, the brother living in tract C is the farm operator. Therefore, farm number 3 is "in" the segment in the figure rather than "in" the segment where the other brother lives.

Because there are many cases where more than one person or household might be involved in the operation of a farm, a short questionnaire should be developed for use at each dwelling unit within a segment. The questions must be carefully worded and designed to ascertain whether anyone living in the dwelling unit is a farm operator in accordance with the prescribed definition of a farm and of a farm operator that establishes a one-to-one correspondence between farms and farm operators.

In addition to the opportunities for omission and duplication arising from ambiguity about the correspondence between farm operators and farms, another important practical problem is often encountered with the farm-operator approach. It is the problem of finding all farm operators in segments containing many nonfarm dwellings (dwellings not occupied by farm operators, as in urban areas). Since it is a major undertaking to visit all dwelling units in a segment containing many nonfarm dwellings, special procedures might be needed. There are at least two possibilities:

(1) Let the interviewers visit dwelling units more or less at their discretion in an effort to find all farm operators. That is, at dwelling units

which they visit, inquiries would be made to discover farm operators living in neighboring dwellings as well as in the one visited. This possibility is not regarded by the writer as satisfactory, because operators are likely to be overlooked.

(2) Another possibility is to work out a plan for selecting a random subsample of dwelling units to be canvassed in the segment. For example, the segment might be divided into smaller segments and one of the smaller segments selected at random for the sample. Do not overlook the need for adjusting (or weighting) the data because of the subsampling. A preferred method might be to use smaller segments, initially, in residential areas and also to use smaller sampling fractions in such areas. Remember, the case under discussion is an area where the proportion of nonfarm dwelling units is high. Villages where most of the dwelling units are occupied by farm operators pose a different problem.

The difficulty of achieving complete identification of operators living within sample segments in densely populated areas, where the proportion of farm operator dwellings is low, and the difficulty of applying rules to establish a one-to-one correspondence between farm operators and farms have often led survey statisticians to adopt the farm approach discussed in the next section. The farm-operator approach does not require dividing each segment into tracts, whereas the farm approach does.

4.3.2 Farm approach. This approach involves identifying a farm and its land area and determining the operator or a suitable respondent who can give accurate information about the farm. The difference between the farm-operator and the farm approaches is mostly a matter of procedure--whether one looks for farm operators and the identity of their farms or for farms and then the operators. Even though the definition of a farm is the same, the coverage error might be quite different because the survey procedures are different. Also, the choice of approach might have an important bearing on how segments are defined. This will be discussed under frame construction.

Under the farm approach, the task is to identify farms with headquarters within the sample segments and to fill out questionnaires for such farms. Giving interviewers a sample of segments delineated on maps and telling them to fill out questionnaires for farms with headquarters within the sample segments is generally inadequate, even though complete definitions of farms and headquarters are provided. Experience has shown that success with the farm approach requires doing a thorough, rigorous job of identifying all farms that have any land within the segment and then of determining which of these farms have headquarters located within the segment. As a minimum, it seems necessary to have interviewers follow a three-step process with the aid of a specially designed form:

Step 1--Account for all land in each sample segment by dividing each segment into tracts and describing each tract as illustrated in figure 1.

Step 2--On a special form list each farm that corresponds to a tract identified in step "1" and obtain answers to questions on this form which will establish the land area of each farm.

The idea is to obtain answers to questions that will clearly establish the boundaries, area, and identity of each farm uniquely.

Step 3--Determine the location of the headquarters of each farm. Questions that need to be included on the form will depend on the definition of headquarters.

4.3.3 Problems with establishing a definition of farm headquarters. Operational specifications of a headquarters must be formulated so each farm has one and only one point called a headquarters. Examples of headquarter locations that might be considered are the farm operator's dwelling, the northeast corner of the farm, the place where farm records are kept, the place where farm machinery is kept, and the main entrance to the farm. There is some ambiguity in the application of any definition of a headquarters. A dwelling unit and its location in relation to a segment boundary are quite distinctive, but the degree of success using the operator's dwelling as the headquarters depends, among other things, on obtaining of a one-to-one correspondence between farm operators and farms. The northeast corner often lacks uniqueness in application because the geometrical configuration of farms varies widely. Machinery might be kept at more than one location and the main entrance is not always distinctive. Thus, lack of simplicity and uniqueness in operational specifications of a headquarters is a key problem with the open-segment method.

Under the operator approach (section 4.3.1), the farm operator's residence is the logical point to define as the farm headquarters. As indicated in the preceding paragraph the major practical problem with the operator approach relates to farm tenure and who is the operator of a farm. If farm (or land) tenure is such that simple rules will fully specify a particular person as the unique farm operator, then the operator approach (and use of the operator's residence as the farm headquarters) could be the best survey technique. However, if matters of tenure or farm organization are complex, or if a large amount of screening is required to identify farm operators in densely populated areas, some other technique might be more effective.

With the farm approach (section 4.3.2), the operator's residence could also be defined as the farm headquarters. In this case, the questions asked in step 3 would be for the purpose of determining, uniquely, the farm operator. Then the location of each operator's residence would be ascertained to determine whether the farm is "in" the segment. However, operational procedure must be developed and tested in detail.

Farm number 3 in figure 1 provides an example of the kind of detail that must be considered in the process of formulating specifications and instructions for interviewers to follow. Suppose the farm approach is used and that farm headquarters is defined as the operator's residence. According to the specifications, farm number 3 is "in" the segment shown in figure 1 because the headquarters (place where the operator lives) is in this segment. But will the open segment, farm-approach field procedures correctly include this farm in the sample, if the segment shown in figure 1 happens to be selected for the sample? Remember, tract C was described as a nonfarm tract. If only farm tracts are included on the listing called for by step 2 (see page 17), farm

number 3 would be omitted when it should be included. Farm number 3 illustrates a problem that is peculiar to the farm approach but not the farm operator approach. The problem is how to account for farms where the operator does not live on his farm and his residence is by definition the headquarters of the farm.

One solution is to always include the operator's residence (the land on which it is located) as a part of the farm. This would call for procedures for dividing segments into tracts so tract C (or a small lot on which the operator residence was located) would be identified as a part of farm number 3. To be sure that an operator's residence is always included as part of a farm, it would be necessary to visit all dwellings within a sample segment to identify all operator dwellings and include them in farms. That takes us back to the farm-operator approach.

An alternative solution requires formulating rules that enable a clear determination of whether an operator is living on his farm or is not living on his farm. Operators living on their farms have sometimes been referred to as resident operators. Those not living on their farms would be called nonresident operators. Briefly, the plan is as follows: For farms with resident operators define the operator's residence as the headquarters. For farms with nonresident operators, some point other than the operator's residence would be defined as the headquarters. This plan has been used in many surveys; but, with the farm approach, a generally best or accepted way of defining farm headquarters has not emerged. The search for a satisfactory operational definition continues and will probably continue whenever the open-segment method is used.

The following definition of headquarters is one illustration of some of the efforts that have been made. It represents an early effort to establish an operational definition of headquarters for an area where a high proportion of the operators lived on their farms. It assumes the farm approach, and in areas having many nonfarm dwellings it requires looking for farms rather than operators. Also, its application requires operational specifications (not included herein) for determining whether an operator lives on his farm. Such specifications need to include a definition of a farm operator that establishes a one-to-one correspondence between farm operators and farms. The following definition of headquarters is not necessarily recommended. It is presented as an illustration of criteria that might be used in an operational definition:

(1) If the operator of the farm lives on the farm, his residence is the headquarters.

(2) If the operator does not live on the farm but there is one and only one occupied dwelling on the farm, that dwelling is the headquarters.

(3) If the operator does not live on the farm and there are two or more occupied dwellings on the farm, the occupied dwelling of greatest value is the headquarters.

(4) If there are no occupied dwellings on the farm but other buildings are present, the building of greatest value is the headquarters.

(5) If there are no buildings on the farm, the "main entrance" to the farm is the headquarters.

(6) If no point can be identified as the main entrance the corner of the farm farthest west and farthest north (in that order) is the headquarters.

As an alternative one could combine parts (2), (3), and (4) and parts (5) and (6) as follows:

If the operator does not live on his farm and there is one or more buildings on his farm, the most valuable building is the headquarters.

If there are no buildings on the farm, the corner of the farm farthest west and farthest north (in that order) is the headquarters.

With reference to figure 1, sufficient information was not given to illustrate application of the above definition. However, it gives some indication of how complex the definition could be. One should look for a simple definition that is easy to apply and is as free from error as possible.

In practice, any definition must be interpreted with regard to the many situations that will be encountered. What does "on the farm" mean? What is a building? What is a farm? Who is the operator? Fortunately, for most farms the answers to such questions are quite clear, but there are many cases where ambiguity gives rise to coverage errors. Much experience is required to develop complete, well-adapted definitions and instructions and to develop training programs and procedures for supervising fieldwork that lead to results of high quality. It is the detail necessary for dealing with all of the numerous situations that is onerous. Do not overlook the need for balance. For example, one can focus so much attention on completeness of instructions that emphasis on the most important points is lost.

4.3.4 Some general observations. General survey experience with the open segment reveals a strong tendency toward undercoverage. For example, assume a 5-percent area sample. The number of farms identified and surveyed by interviewers as being in the sample tends to be less than 5 percent. Even with experience and much emphasis on getting all farms correctly defined and associated with segments, it is difficult to reduce coverage error to a level that is negligible. Incidentally, coverage error varies from one characteristic to another in the same survey. For example, there are many small farming operations that present problems of ambiguity about whether they qualify as a farm. Whether one of these small farms gets counted has a greater impact, for example, on an estimate of the number of farms than on an estimate of acres in farmland.

In summary, ambiguity about farm headquarters and ambiguity about whether a farm operation satisfies the definition of a farm are both major sources of coverage error. They can be avoided by using the closed segment where applicable. However, when a farm must be the reporting unit, there are two possible survey methods that do not involve headquarters:

(i) The first is to have a questionnaire filled out for every farm that is within, or partly within, each sample segment (refer to

step 2 on page 17). This possibility is called the "weighted" segment because the data need to be weighted. It will be discussed in detail in the next section.

(ii) The other possible way of avoiding the headquarters problem is not generally feasible. Give each farm listed in step 2 a conditional probability of being in the sample that is equal to the proportion of the farm that is within the sample segment, without acquiring detail about the operator. It is not feasible, in the writer's opinion, to have interviewers perform the probability determinations. It would be necessary to have the step 2 listings sent to the office for random determinations. The need to send the step 2 listings and information to the office adds to cost and time required to do the survey, as compared with letting the interviewers proceed with step 3 and the necessary interviewing. Moreover, the sampling variance would be very large.

4.4 The Weighted-Segment Method

The weighted-segment method calls for collecting data from every farm that is within, or partly within, a sample segment. The data for each farm are then weighted by the proportion of the entire farm that is within the segment.

Initial reactions to the weighted segment have often been unfavorable for various reasons. One is the fact that the data for individual farms need to be weighted. Another is that only about half of the farms listed in step 2 on page 17 will have headquarters within the sample segments. Therefore, for a given number of sample segments, the weighted segment requires interviews for twice as many farms as the open segment. An initial impression of sampling variance, assuming a fixed number of farms in the sample, might also be unfavorable compared with that of other methods. Moreover, the ambiguities about what constitutes a farm are not avoided. However, the weighted segment has some important desirable characteristics and it should be fully investigated. Compared with the open-segment, the weighted-segment method avoids the problems associated with establishing farm headquarters; and it appears to have a better potential for minimizing coverage error. Also, as we shall see later, it has a much lower sampling variance per segment than the open segment. These points will become more apparent as the weighted-segment method is discussed.

The weighted-segment method is better understood by thinking about a whole population of segments rather than a sample of segments. In effect, each farm in the population gets prorated among all segments in which it is located. That is, with reference to a particular segment, the data for each farm that is within, or partly within, the segment get multiplied by the proportion of the farm in the segment. Therefore, when the prorated data for each segment are summed over all segments in the population, each farm is accounted for in such a way that the total for all segments is the correct population total. This will be shown in a numerical illustration presented later. Turn to the numercal illustration on page 26, if you encounter difficulty with the following algebraic formulation. Corresponding mathematical descriptions for the closed- and open-segment methods are not included because the theory of cluster sampling, discussed in sampling textbooks, is sufficient.

4.4.1 Algebraic description of the weighted segment. Suppose A_j is the amount of farm land in the jth farm in the population where $j = 1,\ldots,F$ and F is the number of farms in the population. Let A_{ij} be the amount of farmland in the jth farm that is within the ith segment of the population where $i = 1,\ldots,N$. Then $P_{ij} = \frac{A_{ij}}{A_j}$ is the proportion of the jth farm that is in the ith segment. If all of the jth farm is in the ith segment, $P_{ij} = 1$. If none of the jth farm is within the ith segment, $P_{ij} = 0$. Also,

$$\sum_i^N P_{ij} = \sum_i \frac{A_{ij}}{A_j} = 1, \text{ and } \sum_j\sum_i^{FN} P_{ij} = F.$$

Remember, P_{ij} is a proportion, not a probability.

Suppose X_j is the value of some characteristic X for the jth farm. Then, $\sum_j^F X_j$ is the total of X for the population. The total of X for the ith segment is defined as

$$X_i = \sum_{j=1}^{F} P_{ij} X_j \qquad (2)$$

Excluding the possibility of reporting errors, X_i is a unique value for the ith segment. When summed over all segments of the population, the values of X_i add to the population total. Thus

$$\sum_i^N X_i = \sum_i\sum_j^{NF} P_{ij} X_j = \sum_j\sum_i^{FN} P_{ij} X_j$$

Observe that

$$\sum_i^N P_{ij} X_j = X_j \text{ because } \sum_i^N P_{ij} = 1.$$

Therefore, it follows that $\sum_i^N X_i = \sum_j^F X_j$, which shows that $\sum_i^N X_i$ is the correct total.

Equation (2) may be written in another form that is more convenient when working with sample data. Let $k = 1,\ldots,f_i$ be the index for farms associated with the ith segment. "Associated with" refers to all farms that are entirely in or partly in the segment. Let X_{ik} be the value of X for the kth farm in the ith segment, and P_{ik} be the proportion of the kth farm that is within the ith

segment. Then, X_i can be written as follows:

$$X_i = \sum_k^{f_i} P_{ik} X_{ik} \qquad (3)$$

It seemed somewhat easier to use equation (2) than equation (3) to show that the X_i's added to the correct population total.

4.4.2 Estimators and their variances. Since there is a unique value, X_i, for every segment in the population, sampling theory for cluster sampling applies in developing a sampling design. Any suitable probability sampling plan may be used to select a sample of segments. However, for simplicity and to illustrate how estimates from a sample could be made, assume a simple random sample of n segments. Let x_{ik} be the value of X for the kth farm associated with the ith segment in the sample. The questionnaire must provide a numerical value of A_{ik} and A_k so $p_{ik} = \frac{A_{ik}}{A_k}$ can be calculated, where p_{ik} is the proportion of the kth farm that is within the ith segment. Incidentally, "A" was defined above as farmland. Other possible measures of the proportion of a farm that is within a segment need to be explored. Estimators of interest include:

Estimator of the population total of X:

$$\hat{X} = \frac{N}{n} \sum_{ik}^{nf_i} p_{ik} x_{ik} \qquad (4)$$

Estimator of the total number of farms, which is obtained by letting $x_{ik} = 1$:

$$\hat{F} = \frac{N}{n} \sum_{ik}^{nf_i} p_{ik} \qquad (5)$$

Estimator of the average value of X per farm:

$$\frac{\hat{X}}{\hat{F}} = \frac{\Sigma\Sigma p_{ik} x_{ik}}{\Sigma\Sigma p_{ik}} \qquad (6)$$

The notation in the estimators could be simplified by using one index of farms in the sample, but subtotals by segments are needed for estimating sampling error.

Let $x_i = \sum_k^{f_i} p_{ik} x_{ik}$ and $p_i = \sum_k^{f_i} p_{ik}$. Then, assuming simple random sampling, formulas for estimating the variance of the estimates may be written as follows:

$$\mathrm{var}(\hat{X}) = \frac{N(N-n)}{n} \frac{\sum_i^n (x_i-\bar{x})^2}{n-1} \tag{7}$$

$$\mathrm{var}(\hat{F}) = \frac{N(N-n)}{n} \frac{\sum_i^n (p_i-\bar{p})^2}{n-1} \tag{8}$$

$$\mathrm{var}\left(\frac{\hat{X}}{\hat{F}}\right) = \left(\frac{\hat{X}}{\hat{F}}\right)^2 [\mathrm{var}(\hat{X})+\mathrm{var}(\hat{F})-2\mathrm{cov}(\hat{X},\hat{F})] \tag{9}$$

where

$$\bar{x} = \frac{\sum_i^n x_i}{n}, \quad \bar{p} = \frac{\sum_i^n p_i}{n},$$

and

$$\mathrm{cov}(\hat{X},\hat{F}) = \frac{N(N-n)}{n} \frac{\sum_i^n (x_i-\bar{x})(p_i-\bar{p})}{n-1}$$

Even though a part of the same farm might be found in more than one segment in the sample, the above formulas apply; that is, a weighted part, $p_{ik}x_{ik}$, of the farm is included in each segment in which it is found.

4.4.3 Ratio estimation. If a measure of the size of each segment is available, ratio estimation might be used. For example, the total land area of the population might be known and it might be feasible to obtain the land area, y_i, for each segment in the sample. If the segments vary considerably in size and X_i is correlated with Y_i, a ratio estimator of the total of X might have a lower variance. The estimator, $\hat{X}_1$, would be

$$\hat{X}_1 = \left(\sum_i^N Y_i\right)\frac{\hat{X}}{\hat{Y}} \tag{10}$$

where $\sum_i^N Y_i$ is the total land area of all N segments, $\hat{X}$ is given by equation (4),

and

$$\hat{Y} = \frac{N}{n}\sum_i^n y_i$$

The estimated variance of $\hat{X}_1$ is

$$\mathrm{var}(\hat{X}_1) = Y^2\left(\frac{\hat{X}}{\hat{Y}}\right)^2 [\mathrm{var}(\hat{X})+\mathrm{var}(\hat{Y})-2\ \mathrm{cov}(\hat{X},\hat{Y})]$$

where $$Y = \sum_i^N Y_i$$

$$\mathrm{var}(\hat{X}) = \frac{N(N-n)}{n} \frac{\sum_i^n (x_i-\bar{x})^2}{n-1},$$

$$\mathrm{var}(\hat{Y}) = \frac{N(N-n)}{n} \frac{\sum_i^n (y_i-\bar{y})^2}{n-1},$$

and

$$\mathrm{cov}(\hat{X},\hat{Y}) = \frac{N(N-n)}{n} \frac{\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

With appropriate modifications a ratio estimator like equation (10) might also be used with the closed segment. With the open segment, if ratio estimation is used it probably would not involve land area of the segments. Before deciding to use a ratio estimator, it is important to consider the conditions under which it will be better than the estimator specified by equation (4). Moreover, with reference to equation (10), do not overlook the fact that the conditions should be such that the <u>expected value</u> of $\hat{Y}$ is very close to $\sum_i^N Y_i$. Otherwise, there is a bias in the expansion of the sample. To illustrate, suppose that the total land area used in equation (10) to expand the ratio, $(\frac{\hat{X}}{\hat{Y}})$, comes from a geodetic survey of the whole area. The total land area determined by the geodetic survey might not be the same as $\sum_i^N Y_i$, which is the expected value of $\hat{Y}$, because the geodetic survey did not obtain the total land area by summing measurements of the land areas of each segment in the population. In fact, experience shows that different methods of measuring the same thing generally do not give identical results and the difference is often large enough to be important. This does not mean that Y_i must be a measurement that has no error. There could be considerable error in the values of Y_i. The two important things are that the expected value of $\hat{Y}$ be close to $\sum_i^N Y_i$ and that Y_i be related to X_i in a way that will reduce sampling variance. (See ratio estimation in the textbooks on sampling.)

4.4.4 <u>Unequal probabilities of selection</u>. The weighted segment method is not limited to sampling segments with equal probabilities. With unequal probabilities of selection the estimators, equations (4) and (5) would become:

$$\hat{X} = \sum_i^n R_i \sum_k^{f_i} p_{ik} x_{ik} \qquad (11)$$

and

$$\hat{F} = \sum_{i}^{n} R_i \sum_{k}^{f_i} p_{ik} \qquad (12)$$

where R_i is the reciprocal of the probability which the ith segment had of being in the sample. However, the variance estimators (7) and (8) no longer apply. Variance formulas for the particular design of the sample should be used.

4.4.5 Domain estimation. In many surveys, estimates by domains are desired. "Domain" is a general expression that refers to a part of the population, for example, a class of farms such as livestock farms or farms with more than 500 acres of farmland. The estimation and variance formulas in section 4.4.2 are still applicable if we make the following modification. Simply let $x'_{ik} = x_{ik}$ and $p'_{ik} = p_{ik}$ if a farm belongs to the domain and let $x'_{ik} = 0$ and $p'_{ik} = 0$ if the farm does not belong to the domain. Substitute x'_{ik} and p'_{ik} for x_{ik} and p_{ik} in equation (4), (5), and (6). Equation (4) is then an estimator of the total for the domain, equation (5) provides an estimate of the number of farms in the domain, and equation (6) gives an estimate of the average per farm in the domain. The use of $x'_i = \sum_{k}^{f_i} p'_{ik} x'_{ik}$ and $p'_i = \sum_{k}^{f_i} p'_{ik}$ instead of x_i and p_i in equations (7), (8), and (9) provides estimates of the sampling variances of the domain estimates.

5. Numerical Illustration

To illustrate and compare the three methods of applying area sampling, a small hypothetical population composed of 25 segments, 47 tracts, and 30 farms was formulated. Most of the data for this illustration were copied from a listing of tract and farm data from an area sample in an area where cattle-feeding farms were concentrated. A disproportionately large number of farms with cattle and corn were selected for this illustration.

Table 2 shows farm and tract data by segments. In the first column, the number to the left of the decimal identifies the segment, and the number on the right side of the decimal identifies tracts within segments (see section 4.1 for a definition of a tract). Tracts having the same farm number (see column 5) compose a farm. An asterisk affixed to a farm number signifies the tract in which the farm headquarters is located. For example, farm number 3 is composed of tracts 2.2 and 3.1 and its headquarters is in tract 2.2.

To summarize briefly, the three methods of defining area sampling units call for data collection as follows:

Closed segment. In a survey using the closed segment, data for tracts within the sample segments would be collected.

Open segment. If the open segment is used, farm data would be collected for all farms with headquarters within the sample segments.

Weighted segment. In a survey employing the weighted segment, farm data would be collected for every farm that is in or partly in a sample segment.

As a specific example of the data that would be collected under each of the three methods, suppose segments numbered 5, 7, and 19 have been selected for a sample. Depending on which method is used, one of the following sets of data (refer to table 2) would be collected.

Closed Segment

Segment number	Tract number	Tract data		
		Farmland	Cattle	Corn
5	--	--	--	--
7	1	630	0	0
7	2	120	0	116
19	1	160	0	0
19	2	160	28	0
19	3	80	201	19

Open Segment

Segment number	Tract number	Farm data		
		Farmland	Cattle	Corn
5	--	--	--	--
7	10	120	0	116
19	24	160	28	0
19	25	300	201	118

Weighted Segment

Segment number	Farm number	Farmland in segment	Farm data		
			Farmland	Cattle	Corn
5	--	--	--	--	--
7	2	630	1,260	246	203
7	10	120	120	0	116
19	23	160	640	0	116
19	24	160	160	28	0
19	25	80	300	201	118

Since each of the 47 tracts in the population is associated with one and only one segment, it is clear that the closed-segment totals, when summed over all segments in the population, must add to the correct population totals.

Table 2.--Tract and farm data by segments

Segment and tract number	Tract data			Farm data					
	Farmland	Cattle	Corn	Farm number	Farmland	Cattle	Corn	Other tracts in farm	Proportion of farm in tract
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1.1	160	37	64	1*	160	37	64	None	1.000
2.1	150	246	43	2*	1,260	246	203	3.3,7.1,9.1	.119
2.2	312	26	122	3*	576	26	262	3.1	.542
2.3	18	6	0	4*	18	6	0	None	1.000
3.1	264	0	140	3	576	26	262	2.2	.458
3.2	80	24	32	5	400	91	90	6.2	.200
3.3	320	0	160	2	1,260	246	203	2.1,7.1,9.1	.254
4.1	237	93	114	6*	400	93	159	9.2	.592
4.2	90	0	0	7*	90	0	0	None	1.000
5.				no farm tracts in segment no. 5					
6.1	160	23	43	8*	160	23	43	None	1.000
6.2	320	67	58	5*	400	91	90	3.2	.800
6.3	4	0	0	9*	4	0	0	None	1.000
7.1	630	0	0	2	1,260	246	203	2.1,3.3,9.1	.500
7.2	120	0	116	10*	120	0	116	None	1.000
8.1	159	27	25	11	320	82	25	14.2	.497
8.2	236	82	104	12*	236	82	104	None	1.000
9.1	160	0	0	2	1,260	246	203	2.1,3.3,7.1	.127
9.2	163	0	45	6	400	93	159	4.1	.408
9.3	80	0	0	13	4,400	777	320	10.1,11.1,12.1,13.1	.018
10.1	630	340	160	13	4,400	777	320	9.3,11.1,12.1,13.1	.143
11.1	1,275	437	160	13*	4,400	777	320	9.3,10.1,12.1,13.1	.290
12.1	1,800	0	0	13	4,400	777	320	9.3,10.1,11.1,13.1	.409
13.1	615	0	0	13	4,400	777	320	9.3,10.1,11.1,12.1	.140
13.2	140	26	39	14*	140	26	39	None	1.000
14.1	55	7	11	15*	55	7	11	None	1.000
14.2	161	55	0	11*	320	82	25	8.1	.503
14.3	160	0	120	16*	160	0	120	None	1.000
14.4	86	6	16	17	160	25	56	15.1	.538

Continued...

Table 2.--Tract and farm data by segments--Continued

Segment and tract number	Tract data			Farm data					
	Farmland	Cattle	Corn	Farm number	Farmland	Cattle	Corn	Other tracts in farm	Proportion of farm in tract
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
15.1	74	19	40	17*	160	25	56	14.4	.462
16.				no farm tracts in segment no. 16					
17.1	360	100	170	18	366	100	170	18.1	.984
17.2	2	2	0	19*	2	2	0	None	1.000
17.3	81	0	23	20*	81	0	23	None	1.000
17.4	74	0	0	21*	74	0	0	None	1.000
17.5	320	145	120	22*	320	145	120	None	1.000
18.1	6	0	0	18*	366	100	170	17.1	.016
18.2	480	0	116	23*	640	0	116	19.1	.750
19.1	160	0	0	23	640	0	116	18.2	.250
19.2	160	28	0	24*	160	28	0	None	1.000
19.3	80	201	19	25*	300	201	118	20.1	.267
20.1	220	0	99	25	300	201	118	19.3	.733
21.1	320	0	0	26*	320	0	0	None	1.000
22.1	160	19	86	27	360	63	186	23.1,24.2	.445
22.2	120	0	60	28*	120	0	60	None	1.000
23.1	80	44	0	27*	360	63	186	22.1,24.2	.222
24.1	280	46	80	29*	280	46	80	None	1.000
24.2	120	0	100	27	360	63	186	22.1,23.1	.333
25.1	400	0	160	30*	400	0	160	None	1.000
Total	12,082	2,106	2,645						30.000

*There is one asterisk for each farm number that indicates the tract in which the farm headquarters is located.

Likewise, with the open segment, since each of the 25 farms is associated with one and only one segment, the open-segment totals must add to the correct population totals. It is less obvious, but the weighted-segment totals (after the data are "weighted") must also add to the correct totals. Consider segment no. 19. Three farms, 23, 24, and 25, are within or partly within the segment. The proportions of these farms that are within the segment are:

Farm	Proportion
23	$\frac{160}{640} = .250$
24	$\frac{160}{160} = 1.000$
25	$\frac{80}{300} = .267$

These proportions are values of P_{ik} that appear in equation (3) and in the estimators, equations (4), (5), and (6), for the weighted-segment method. The last column of table 2 contains the values of P_{ik}. Notice that the values of P_{ik} add to 1 for each farm. Using segment 19 as an example, the weighted-segment totals are:

Cattle	(.250) (0) + (1.000) (28) + (.267) (201) = 81.7
Corn	(.250) (116) + 1.000 (0) + (.267) (118) = 60.5
Farmland	(.250) (640) + (1.000) (160) + (.267) (300) = 400
Number of farms	(.250) + (1.000) + (.267) = 1.517

These totals and corresponding weighted-segment totals for all other segments are recorded in table 3. Segment totals for the closed- and open-segment methods are also shown. Notice that the weighted-segment totals for farmland (400 for segment no. 19) are the same as the closed-segment totals. Hence, the weighted-segment totals for farmland are not shown in table 3.

5.1 Domain Estimation and the Weighted Segment

Some analysts have sought reassurance regarding the applicability of the weighted segment for analytical studies. Since the value of X for a farm is multiplied by the proportion of the farm that is in the segment, it might seem, at first, that one is dealing with fractions of farms rather than whole farms. But that is not actually the case. The situation is similar to weighting sample data when several sampling rates are involved. This point was considered briefly in section 4.4.5. The technique that was outlined is commonly used by statisticians as a short, general means of specifying a procedure for making estimates by domains as well as for the whole population.

To illustrate, suppose farms numbered 2, 7, 12, 17, and 22 compose a domain and that one wishes to make estimates for this domain. From table 2 the totals and averages for the 5 farms in this domain can be obtained. The results are:

Table 3.--Segment totals--closed, open, and weighted

Segment number	Farmland		Number of farms		Cattle			Corn		
	Closed	Open	Open	Weighted	Closed	Open	Weighted	Closed	Open	Weighted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	160	160	1	1.000	37	37	37.0	64	64	64.0
2	480	1,854	3	1.661	278	278	49.4	165	465	166.2
3	664	0	0	.912	24	0	92.6	332	0	189.5
4	327	490	2	1.592	93	93	55.1	114	159	94.1
5	0	0	0	0	0	0	0	0	0	0
6	484	564	3	2.800	90	114	95.8	101	133	115.0
7	750	120	1	1.500	0	0	123.0	116	116	217.5
8	395	236	1	1.497	109	82	122.7	129	104	116.4
9	403	0	0	.553	0	0	83.2	45	0	96.4
10	630	0	0	.143	340	0	111.1	160	0	45.8
11	1,275	4,400	1	.290	437	777	225.3	160	320	92.8
12	1,800	0	0	.409	0	0	317.8	0	0	130.9
13	755	140	1	1.140	26	26	134.8	39	39	83.8
14	462	535	3	3.041	68	89	61.7	147	156	173.7
15	74	160	1	.462	19	25	11.5	40	56	25.9
16	0	0	0	0	0	0	0	0	0	0
17	837	477	4	4.984	247	147	245.4	313	143	310.3
18	486	1,006	2	.766	0	100	1.6	116	286	89.7
19	400	460	2	1.517	229	229	81.7	19	118	60.5
20	220	0	0	.733	0	0	147.3	99	0	86.5
21	320	320	1	1.000	0	0	0	0	0	0
22	280	120	1	1.445	19	0	28.0	146	60	142.8
23	80	360	1	.222	44	63	14.0	0	186	41.3
24	400	280	1	1.333	46	46	67.0	180	80	141.9
25	400	400	1	1.000	0	0	0	160	160	160.0
Total	12,082	12,082	30	30.000	2,106	2,106	2,106.0	2,645	2,645	2,645.0

Item	Total	Average
Farmland	2,066	413.2
Cattle	498	99.6
Corn	483	96.6

A reader may verify that the estimators, equations (4), (5), and (6), and the procedure outlined in section 4.4.5 are appropriate for estimating these totals and averages. Treat the 25 segments as a sample. That is, make the calculations as though the 25 segments were a sample from a larger population. Taking the 25 segments as a sample, eliminates random sampling error and the results should agree exactly with the above totals and averages for the 5 farms.

5.2 Sampling Variance

Since the sampling variance is a function of variation among segment totals, it is important to study table 3 and its derivation from table 2. Examine the variation among segments with regard to the three methods. For crop and other items that are limited by amount of land, the closed-segment method imposes a maximum on the segment total. Obviously, the acreage under corn, for example, cannot be greater than the amount of farmland within the segment. But with the open segment, the maximum amount of corn that could be "in" a segment can be at least as much as the amount for the farm in the population that is growing the largest amount of corn.

Observe, in table 3, the variation among segments in the amount of farmland and compare the open and closed segments. For characteristics that are highly correlated with amount of farmland, the closed segment will have much lower sampling variances than the open segment, assuming the amount of land in segments can be effectively controlled in the process of delineating segments. One might expect the differences in variances between open and closed segments to be less for livestock than for crops, because the number of livestock is limited to a lesser degree by the amount of land in a segment.

For characteristics correlated with amount of farmland the weighted-segment method, like the closed segment, imposes some control on the maximum values of totals for segments. For example, the acreage of corn for a segment after the data are weighted cannot exceed the amount of farmland in the segment. That is, with reference to equation (3), if X is the acreage in any given crop, the weighted-segment total, X_i cannot exceed the land area of the segment. Remember, the sampling variance for the weighted segment involves variance among the X_i.

As another example of how the weighted and open segments differ with regard to sampling variance, refer to table 2 and farm no. 13. Parts of this farm are in five segments. It has 4,400 acres of farmland and 777 cattle. The open-segment method assigns all 777 cattle, regardless of where they are located, to segment number 11. This one farm has a major impact on the sampling variance for the open segment. The weighted-segment method reduces, in this case, the sampling variance by "dividing" the farm into parts. Regardless of where the

cattle are located, the weighting involved in the weighted-segment method has the effect of distributing the 777 cattle among the five segments as follows:

Segment	Cattle
9	14
10	111
11	225
12	318
13	109
TOTAL	777

Notice that the more segments that a farm is located in, the greater its chance of being in the sample.

Table 4 shows the relative variance among segments for each of the three methods. The variances were computed from the data shown in table 3. Although this numerical illustration does not provide a basis for generalization, the results in table 4 are not contrary to general experience. As one would expect from the above discussion, and as found in various studies, the open segment has much larger variances than the closed segment.

Table 4.--Relative variance among segment totals

Item	Relative variance 1/		
	Closed	Open	Weighted
Farmland.....	0.68	3.55	0.68
Number of farms...	xxx	0.87	0.84
Cattle.....	2.12	3.71	0.97
Corn.....	0.73	1.21	0.48

1/ $\frac{\Sigma(X_i-X)^2}{\bar{X}^2(N-1)}$ where X_i is a segment total in table 3.

Since a farm is equal to or larger than a tract, a sample of n segments using the weighted segment gets data for a larger proportion of the population than the closed segment does. But, after weighting the data, the "size" of the weighted segment with regard to acres of farmland is the same as the "size" of the closed segment. Hence, the part of the variance among segments (sampling variance) that can be associated with the variation in size of segments appears to be approximately the same for weighted and closed segments. Moreover, the weighting of the weighted-segment data has an averaging effect. Therefore, it is reasonable to expect the sampling variances for the weighted segment to be generally somewhat less than the sampling variances for the closed segment. However, costs must be taken into account.

It is of interest to compare the relative variances among the 30 farms in the numerical example with the relative variances among segments. The relative variances among the 30 farms are presented in the last column of table 5. For purposes of comparison, the relative variances among segments need to be converted to the equivalent of one farm. The open segment has an average of $\frac{30}{25} = 1.2$ farms per segment and for the weighted segment the average number of farms (unweighted) was $\frac{47}{25} = 1.88$. To convert the variances in table 4 to the equivalent of one farm, multiply the open-segment variances by 1.2 and the weighted-segment variances by 1.88. This gives the results for the open and weighted segments shown in table 5.

Table 5.--Relative variance per farm

Item	Relative variance among segments on a per farm basis		Relative variance among farms
	Open	Weighted	
Farmland	4.26	1.28	3.89
Number of farms	1.04	1.58	xxx
Cattle	4.45	1.82	4.40
Corn	1.45	0.90	0.90

As expected, owing to within-segment correlation, the variances among open segments, table 5, are greater than the variances among individual farms. With reference to the weighted segment, the impact of within-segment correlation was more than offset by the fact that the weighted segment had the effect of dividing large farms into smaller units. Therefore, as shown in table 5, the net result was that (even on a per farm basis) the variance for the weighted segment was less than the variance among individual farms. This numerical illustration does not provide a basis for generalization; however, the results are not contrary to what one might expect.

6. Discussion of the Three Definitions of Area Sampling Units

The magnitude of differences among the three methods of defining area sampling units depends on local conditions. At one extreme the three methods could be identical. For example, assume a situation where every farm operator lives on his farm and where every farm is a small, continuous piece of land. If none of the farms overlaps segment boundaries, the closed-, open-, and weighted-segment methods would be identical. But farms vary widely in size and type. Some farms are composed of more than one tract, and managerial and tenure arrangements give rise to ambiguity about what constitutes a farm and who is the operator. It appears that one method is not universally better than another.

When comparing the three methods we need to consider the character of the population to be sampled, the kind of data to be collected, the applicability of the concepts on which each method is based, sampling variance, coverage error, response error, and costs. Much additional experience is needed as a

basis for practical judgments on the choice of methods. In this publication it is not feasible to go much beyond a brief discussion of concepts and some indication of the circumstances where one method would be expected to work better than another. Documented studies of comparisons of alternative methods and procedures for applying area sampling are very limited.

6.1 Closed Segment vs Open or Weighted

Since the closed segment is limited to surveys where tracts are suitable reporting units, a comparison of the closed segment with the open or weighted must be limited to such surveys.

Initially, at least in the United States, the open-segment method was used. But, problems of coverage error, particularly problems of identifying farms and of associating farms with segments, led statisticians to search for a better alternative. The closed segment was tried and it proved, where applicable, to be far superior to the open segment with regard to sampling variance and coverage error, particularly if photographs are utilized in the enumeration of segments. As a result a strong tendency developed to use the closed segment to the fullest extent. Although coverage error for the closed segment is relatively low, response error is one factor that limits its applicability. Response error varies from being nil in the case of crop acreages, to a problem of some magnitude in the case of livestock inventories, to being impracticable for characteristics where a farmer is not in position to report for a tract. For example, it is generally not practical to collect data by tracts on characteristics such as costs of production or sales of agricultural products. Such data are often referred to as economic data and are usually associated with a farm as a business enterprise and not with a tract.

Hendricks, Searls, and Horvitz have compared the closed, open, and weighted segments when sampling for crop acreages[6/]. Their results, as well as many unpublished sampling variances computed by the Statistical Reporting Service, show that sampling variances are definitely smaller with the closed segment than with the open segment. The results reported by Hendricks et al. also showed that the weighted-segment variances range from about the same to moderately lower than the closed-segment variances. Comparisons might be quite different for other kinds of data.

The average field cost per closed segment depends heavily on whether it is necessary to contact the operators of all tracts in the segments. For some tracts and kinds of data it might not be necessary to interview the operators of all tracts. For example, in a survey to collect data on crop acreages it might not be necessary to contact operators of tracts that are covered by trees. However, if we assume that the operators of all tracts are to be interviewed, the closed-segment field cost could be nearly as much as the field cost for the weighted segment. That statement is based on an assumption that the questionnaire is the same except that in one case it pertains to a tract and in the other to a farm. For the weighted segment the average interview time would probably be somewhat longer, although in many cases a farm operator can respond

6/ Hendricks, W.A., Searls, D.T., and Horvitz, D.C. Chapter 11 of "Estimation of Areas in Agricultural Statistics", Food and Agriculture Organization, Rome. 1965.

more readily for his farm than for a tract. However, the cost of dividing segments into tracts and of contacting operators for personal interviews is a substantial part of the total cost. Perhaps, for some surveys, the difference in average cost per segment would be as low as 10 percent. Thus there are circumstances where the closed- and weighted-segment methods appear to be competitive (or nearly so) in terms of sampling variance per dollar. Therefore, since coverage and response error tend to be major sources of error, there is a strong indication that for some surveys the most important criterion in making a choice between the closed and weighted segment is the question of which method involves the least coverage and response error.

A similar comparison between the closed and open segment is more difficult to make because they have less in common. However, at this point in the discussion, the question seems to resolve into a matter of how the open- and weighted-segment methods compare. That is, when the closed segment is not applicable, which alternative, open or weighted, is better? In practice, there has been a trend to use of the closed segment to the fullest extent possible and to use the open segment only when the closed is not applicable; but the weighted segment is beginning to attract more attention.

As pointed out earlier, the closed segment is not applicable when (1) survey requirements dictate that farms must be the reporting units or (2) response errors preclude use of tracts as reporting units. In some surveys it is feasible to collect only part of the required data by the closed-segment method. Therefore, to take advantage of the closed segment, a combination of two methods (either closed and open or closed and weighted) has been used simultaneously in the same survey and sample of segments. Which combination is better? Since the answer depends partly on how the open and weighted segments compare, discussion of this question will be deferred to a later section.

6.2 Open- vs Weighted-Segment Methods

The open- and weighted-segment methods are applicable when farms are used as the reporting units.

With the open segment, the choice between the farm-operator and the farm approaches as discussed in 4.3.1 and 4.3.2 is an important consideration. The weighted segment entails only the farm approach; that is, the concepts of the weighted segment and the farm-operator approach are not compatible. Hence, in the discussion of the open vs weighted segment that follows, the farm approach is assumed. But first let us review the conditions that are favorable to the farm-operator approach and the open segment.

You will recall that with the farm-operator approach the objective is to find, within the boundaries of each sample segment, all residences of farm operators. The farms corresponding to farm operators who have a residence (dwelling unit) in a sample segment are in the sample. (Note: Surveys in which farm households are the appropriate reporting units are not included in this dicussion.)

The farm-operator approach will have minimal coverage error when (1) simple rules establishing a one-to-one correspondence between operators and farms can be formulated and applied with very little ambiguity, (2) every operator has

only one residence, and (3) most residences within the sample segments are occupied by farm operators. Under these conditions the task of screening for farm operators is not a costly factor and tendency to overlook any farm-operator residences should be minimal. If, in addition, it is possible to design the sample so there is approximately the same number of farm operators in each segment, the conditions are generally favorable to the open segment (using the farm-operator approach) with regard to coverage error and sampling variance.

As pointed out previously, reasons for considering the farm approach as an alternative to the farm-operator approach are (1) the problems of screening for farm operators in segments where many nonfarm families live, and (2) the problems of matching farms and operators. Conceptually, for any given sample of segments the two approaches give identically the same sample of farms unless there is a difference in the definition of farm headquarters. There is a wide difference in procedures for applying the two approaches. In either case, the major challenge is to achieve complete and accurate identification of all farms with headquarters in the sample of segments. Omission is usually greater than duplication. The percentage of incompleteness can vary from perhaps nil to several percent, depending on survey materials and procedural details and whether such details are in accord with sound concepts. The experience of the survey organization and the amount of emphasis on training and supervising interviewers are also important factors that contribute to achievement of complete and accurate coverage. There has been much experience with the open-segment method and many different procedures have been tried. However, better solutions to the problems of coverage error are needed, which is an important reason for directing more attention to the weighted-segment method.

The main purpose of the next two sections is to indicate that the weighted-segment method has much merit and that it should be thoroughly tested as an alternative that might be much superior to the open segment, at least under some circumstances.

6.2.1 Sampling variance and costs. To review briefly, the weighted-segment method requires dividing each sample segment into tracts and interviewing the operator (or some other appropriate respondent) of each farm that is within, or partly within, the boundaries of the segment. The data collected pertain to farms, not tracts. The open segment (farm approach) also requires dividing each segment into tracts. Farms with headquarters within the sample segments are in the sample and the operators of such farms are interviewed. Assume that headquarters is defined so it is always a unique point within the boundaries of the farm. Then, for any given sample of segments, farms in the sample using the open segment are a subset of farms that would be in the sample if the weighted-segment method was used.

As an aid to discussion, very simple variance and cost models will be helpful. Assume a stratified random sample of segments, using a constant sampling fraction. Ignoring the correction factors for finite population, the variances of the sample means per segment can be written as follows:

$$V(\bar{x}_o) = \frac{V_o}{n_o} \quad \text{and}$$

$$V(\bar{x}_w) = \frac{V_w}{n_w}$$

where $V(\bar{x}_o)$ is the variance of $\bar{x}_o$, the sample mean per segment for the open-segment method,

n_o is the number of segments in the open segment sample,

V_o is the variance among open segments within strata, and

$V(\bar{x}_w)$; n_w, and V_w are similarly defined for the weighted-segment method. For cost models assume:

$$C = C_f + n_o C_o$$

$$C = C_f + n_w C_w$$

where C is the total cost of the survey (it is the same for both methods),

C_f is the fixed part of the total cost that is not related to the number of segments in the sample,

C_o is the average cost per segment with the open segment, and

C_w is the average cost per segment with the weighted segment.

Assuming the total cost is fixed, the sample sizes n_o and n_w are determined from the cost models. It can be shown that the variance, $V(\bar{x}_w)$, with the weighted segment will be less than the variance, $V(\bar{x}_o)$, with the open segment if the following inequality holds

$$\frac{V_w}{V_o} < \frac{C_o}{C_w}$$

It appears, in general, that V_w is much less than V_o. As pointed out previously, there is good reason to believe that the sampling variance for the weighted segment is about equal to or less than the sampling variance for the closed segment; and it is well established that, in general, the sampling variance for the closed segment is (at least for crop acreages) much less than the sampling variance for the open segment. Incidentally, the results published by Hendricks et al. showed that for the acreages of seven crops the variance with the weighted segment averaged about 25 percent less than the variance with the open segment. For estimates of the difference between two years, using a matched sample of segments, their analyses showed that the variances with the weighted segment were less than half of the variances with the open segment.

To look at comparative costs, consider the cost of the weighted segment and the savings that would occur if the open-segment method were used instead. With the weighted segment, the first two steps at the end of 4.3.2 would be carried

out and a questionnaire filled in for every farm listed in step 2 as having some land within a sample segment.

Next, assume that the field procedures thru step 2 in 4.3.2 are the same for both the open- and weighted-segment methods. In the United States roughly one-half of the farms listed in step 2 as having some land within a segment also have headquarters inside the boundaries of the segment. Such farms are included in both the open and weighted segments. The costs of acquiring data for the sample farms with headquarters outside the segment (needed for the weighted segment) is where most of the difference (increase) in cost occurs.

The need to minimize coverage error requires very careful application of rules for associating farms with segments, and thus determining which farms are in the sample. To apply the open-segment procedures effectively, it will probably be necessary to contact some operators of farms that have headquarters outside the segment. This might be needed to resolve any uncertainties about the land in a farm and the location of the farm headquarters. Suppose f_w is the number of farms in a sample of n segments using the weighted-segment method, and suppose f_o is the number of farms in the same sample of segments using the open-segment method. Since f_o is approximately $(1/2)f_w$, it seems clear that C_o must be considerably larger than $(1/2)C_w$ for two reasons: (1) The costs of dividing a segment into tracts and of identifying all farms in or partly in the sample segments is common to both methods (this cost is a part of C_o and C_w, not C_f), and (2) some farms in f_w that are not in f_o would need to be contacted under careful application of the open segment method. It is not possible to make an accurate prior judgment of how C_o compares with C_w for every survey situation. However, even if the inequality does not hold, it appears that C_o in relation to C_w is large enough to justify testing and comparing the two methods, particularly when the need to minimize coverage error is considered.

6.2.2 Coverage error. It is convenient to divide coverage errors into two categories: (1) Incorrect determinations of the composition of individual farms and (2) incorrect association of farms with segments in the sample. These two kinds of error are not independent.

With the weighted segment, correct coverage depends on accurately accounting for all land within a segment and not overlooking any farms that are located partly within the segment. Field procedures, survey materials, and instructions need to be developed with that in mind. Each interviewer must have full knowledge of what a farm is and the ability to determine its location geographically. Data for the entire farm must be collected for every farm that has any land within a sample segment.

With the open segment, but not the weighted, an interviewer should know how to determine a farm's headquarters and its location. The development of specifications that define headquarters and the training of interviewers so they acquire a clear understanding of how to handle all situations is difficult and complex. Avoidance of the problems of defining headquarters and the associated coverage errors is a major reason why statisticians often look for an alternative to the

open segment. The weighted segment avoids the problems of identifying and locating headquarters but that does not necessarily mean that the coverage errors will be less.

To develop more fully the concepts of how the open and weighted segments compare, a few different situations could be considered. For example, suppose there is a small tract within a segment which shows no evidence of any farming activity on it. Assume this tract by definition is part of a farm and that the remainder of the farm is outside the segment. Since the tract inside the segment does not have the appearance of being part of a farm, it could easily be classified as not part of a farm--particularly by an interviewer who is not giving full attention to detail or who does not fully understand the survey concepts as they pertain to his job. However, suppose the tract is misclassified as not being part of a farm. This would result in an omission under the weighted-segment method, but the omission would amount to a fraction (proportion within the segment) of the farm, not the entire farm. With the open segment this misclassification would incorrectly omit the entire farm only if the headquarters of the farm happened to be in the segment. Incidentally, this is a good case that partly illustrates why the closed segment has low coverage error. If a tract within a segment has no agricultural activity that should be included in the survey, it does not matter (with the closed segment) whether the tract was correctly or incorrectly classified as part of a farm. Consideration of how coverage error might occur in various other cases might be a useful exercise, but there is no substitute for experience and testing alternatives under actual operating conditions.

Survey statisticians with experience in area sampling have different views on the potential of the weighted-segment method. The writer happens to be among those who believe the weighted segment should be fully explored and developed. It is easy to describe circumstances (perhaps hypothetical) where the open segment would clearly be preferred, especially if much of the data to be collected are characteristics of operators' households and other farm people rather than to farms. However, it was operating problems in the application of the open-segment method that led to the development of the closed segment. The writer does not expect the coverage error for the weighted segment to be as low as for the closed segment, but, as stated earlier, there are characteristics where coverage and response error combined could be lower for the weighted segment than for the closed segment. Moreover, a better method than the open segment is needed when reporting units must be farms.

Incidentally, experience has shown that coverage error varies considerably from one characteristic to another within the same survey and sample. That is to be expected if, for example, small farms are overlooked more frequently than large ones. Coverage error could be quite low for estimated totals of some items such as crop acreages but high for estimates of numbers of farms which happen to be very sensitive to how a farm is defined and to ambiguities in the application of the definition of a farm. It follows that estimates of averages per farm are also sensitive. With the open segment the number of farms per segment, as found by interviewers, has varied from one survey to another even though the definition of a farm and the sample design remained unchanged. Differences in (1) the purposes of surveys, (2) the survey materials, (3) the operating procedures, and (4) emphasis on finding all farms that should be in the

sample have a bearing on the amount of coverage error. Whether results using the weighted segment would be more consistent is unknown because of insufficient experience.

Investigations and analyses of coverage errors are urgently needed. We need to know, for example, how the average coverage error (or bias due to coverage error) in $\bar{x}_w$ compares with the average coverage error in $\bar{x}_o$, where $\bar{x}_w$ and $\bar{x}_o$ are the sample averages per segment for the weighted- and open-segment methods. With the weighted-segment method the farm data for a segment get weighted (multiplied by fractions) which gives a segment total that is equivalent to the sum of the land areas of the tracts in the segment. This means that the composition of bias due to coverage error differs from the open-segment method.

A view of one aspect of coverage error can be expressed briefly as follows: With a random sample of n segments from a population of N segments the theoretical sampling fraction is $\frac{n}{N}$. The actual sampling fraction that is realized in a survey is likely to differ somewhat from $\frac{n}{N}$ because of coverage error. As stated in preceding discussions, survey experience with the open segment indicates great difficulty in achieving an actual sampling fraction that is close to $\frac{n}{N}$. Perhaps operations with the weighted segment can be more successfully controlled in the sense that the realized sampling fraction will be closer to $\frac{n}{N}$. Conceptually, with the weighted segment, the value of the total of X for a segment (see equation (2) in 4.4.1) should be on a level that is equivalent to the sum of the land areas of the farm tracts within the segment. Remember that, with the closed segment method, a segment total of a characteristic is also on a level that is equivalent to the sum of the land areas of the farm tracts within the segment. We need an answer to the question, Does the weighted-segment method offer more potential than the open-segment method for minimizing bias due to coverage errors?

Considering the experience now acquired, greater dependence on area sampling and improved materials for area sampling, the time has come for a full exploration of the weighted-segment method, especially in situations where the open-segment method is least workable. Survey methods employed should not overlook the possibilities of a combination of methods as discussed in the next section.

6.2.3 Combination of methods. In surveys where only part of the data are amenable to being collected by the closed-segment method, either the open-segment or the weighted-segment method may be used in combination with the closed segment. Which combination of methods is better, closed and open or closed and weighted?

It appears that in all situations a well-designed sample employing the weighted segment would also be well designed for the closed segment. With reference to the open-closed combination, the principles for defining segments differ between the open and closed. In some situations the same sample of segments cannot be well suited to both closed and open. Consider the situation

where nearly all farm operators live in villages. In this case, an efficient sample for the closed-segment data would not be at all similar to an efficient sample for the open-segment data, assuming that a farm's headquarters is defined as the operator's residence. (If the difference is not clear, observe that a high proportion of the segments in an efficient closed-segment sample would be found in the open country where farmland is located. With an open-segment sample, we want to equalize the number of farms "in" the segments which would put a high proportion of the sample segments in the villages.) Moreover, in a given sample of segments under the circumstances described, very few operators would be interviewed both for the tract data (closed segment) and farm data (open segment). That is, very few of the farms and tracts involved would be in common. Considering sampling variances per dollar, it might be better to have two surveys employing different samples. One might be designed efficiently for the closed-segment method (data) and the other for the open. For the situation just described the closed-weighted combination seems clearly superior to the closed-open combination with regard to matters of sample design and the fact that the same farms are involved in the collection of tract data and farm data.

In planning a survey, consider carefully the costs per segment for the closed-weighted and closed-open combinations. The difference in costs might be small in relation to the smaller sampling variance for weighted-segment estimates.

Finally, there is an important point to be considered regarding coverage error and response error, which has not been discussed and is often overlooked. The complexity of the interviewer's job and its relation to the frequency of error is a key factor. That is, additional increments of refinement for the purpose of reducing error might actually result in a net increase in the overall number of errors. Which combination (closed-open or closed-weighted) is easier for an interviewer to understand? The closed and weighted have much in common and it is not necessary to get involved in the headquarters problems. Farms corresponding to tracts in the closed segment are in the weighted segment. Thus, the same operators are interviewed for tract data and for farm data. The concepts in the closed-open combination are generally more difficult for interviewers to understand fully.

7. Construction of Area Sampling Frames

7.1 Background

The construction of an area sampling frame is viewed herein as a major investment to be amortized over a long period and many surveys. After initial construction of the frame is completed, a staff should probably be maintained to make revisions or improvements in the frame and to select and prepare samples as needed. An adequate continuing program of maintenance and improvement could reduce or eliminate the need for finding resources for a complete reconstruction of the frame after several years have lapsed. Monroe and Finkner[7/] have discussed the construction of an area sampling frame for sampling dwellings.

7/ Monroe, John, and Finkner, A.L., "Handbook of Area Sampling," Chilton Company, 1959.

There are numerous ways of constructing an area frame depending on the available resources and the purposes involved. Hence, only general principles and some illustrations will be presented. Persons who are responsible for the construction of a sampling frame ought to try to make the best joint use of expertise on sample design and knowledge of the local conditions involved in application. Small-scale tests of alternatives should be made before determining the final specifications for a major investment.

7.2 Frame-Unit Specifications

For economy in the design and selection of area samples, a "frame unit"8/ is an integral part of an area sampling frame. A frame unit is an area of land that is larger than a segment but usually smaller than the smallest political subdivision.

The essence of an area sampling frame is (1) a set of maps on which the frame units are defined, (2) a list of the frame units, and (3) information about the frame units, such as land area or number of households, which is used for purposes of sample design and assigning numbers of segments to frame units. A number of segments (sampling units) must be assigned to each frame unit. The number assigned could vary with the purpose of the survey, whether the closed, open, or weighted segment is to be used, the topographic detail shown on maps, and information available about the land use or agriculture within the frame unit. After numbers of segments have been assigned to the frame units and specifications of the sample design have been formulated, a sample of frame units is selected with probabilities proportional to the assigned numbers of segments. Each selected frame unit is then divided into as many segments as it was assigned and one segment in the frame unit is selected at random.

There are two major questions to be considered in the development of specifications for a frame: (1) How should frame units be defined? (2) What information should be compiled about each frame unit? The two questions are not independent but will be discussed separately.

Factors having a bearing on the specifications for frame units include:

(1) The boundaries of frame units should be permanent, positively recognizable landmarks. Boundaries of minor political subdivisions (especially if they change frequently or do not follow visible landmarks) often do not make good boundaries. Frame units are the most permanent part of an area frame and should be defined by boundaries that are relatively permanent. Data pertaining to frame units, such as number of dwellings or land use, can be easily updated or revised as new information becomes available. If there are areas undergoing rapid change in land use, updating of information about frame units in such areas might be sufficient.

8/ In the first area frames that were developed in the United States, "count unit" was used. A count unit was larger than a sampling unit and it was called a count unit because farms indicated on highway maps were counted for each "count unit." Although the term "count unit" has become widely used, the writer believes it should be discarded in favor of a more general term, such as "frame unit."

(2) Frame units should be large enough to accommodate alternative specifications of segments that are appropriate for various surveys.

(3) Frame units should provide economy in the selection of area samples. A frame unit need not be divided into segments unless a sample segment is to be selected from it for a particular sample. In general, the amount of work required to select a sample is least when the number of frame units is much larger than the number of segments needed for a sample. The total number of frame units is inversely related to the average size of frame units. There is a trade-off between the cost of defining a large number of small frame units (rather than a smaller number of larger frame units) and the costs of selecting samples after a frame has been constructed.

The use of frame units also provides in some cases, a possibility of a saving in the cost of maps or photographs. Suppose relatively inexpensive maps are available and adequate for delineating frame units, as well as providing an office record of the boundaries of frame units. Such maps might not provide sufficient detail for doing a satisfactory job of dividing a frame unit into segments. More detailed maps or photographs for dividing frame units into segments might be available but costly. It might be sufficient to limit the purchase of the more costly maps or photographs to coverage of the frame units in which a segment is to be selected.

(4) Consideration should be given to various kinds of information that might be available and assembled by frame units for use in the design of samples. This could have a bearing on the frame-unit specifications. For example, to use data from a census of agriculture, one might want the frame units to coincide with the enumeration districts for the census.

(5) Populations and subpopulations to be surveyed are usually defined in terms of geographic coverage as well as reporting units. There might be some advantages to having frame units defined with regard to geographic boundaries that might be used in the specifications of survey populations.

(6) There are two general approaches (and combinations thereof) to setting specifications for frame units: (a) One is to set the specifications primarily with reference to size (land area) and topographic landmarks that are suitable for boundaries. In this case the work of defining frame units is minimal. After the frame units are defined, appropriate information would be compiled for the frame units with regard to the kind of populations to be sampled and how segments are to be defined. (b) In the second approach, the specifications for the frame units would include factors such as land use to achieve greater homogeneity within the frame units. If the variation within frame units is small, stratification of frame units for sampling purposes should be effective. Also, different procedures might be applied to different classes of frame units which could have a bearing on how frame units are defined. For example, frame units covering residential areas might be treated quite differently from frame units that include only open country. In any event, regardless of what the frame unit specifications are, the end result is a defined set of frame units and some information about each frame unit that is useful and available for sampling purposes. The two approaches involve differences in the physical boundaries of the frame units and differences in the way auxiliary information is used. However, the objectives are clear. We want permanent, visible

landmarks for boundaries of frame units and economical, effective use of auxiliary information to reduce sampling variance. The compromises involved will be clarified to some extent in the sections that follow.

7.3 Auxiliary Information and Its Use

Information or data that are available for use in the design of samples will be referred to as "auxiliary information" or "auxiliary data". There is a wide variety of auxiliary information and there are many ways of using such information in the design of samples, the general objective being to achieve maximum accuracy, assuming a fixed cost of the survey. At this point, perhaps a brief review of the key principles involved in the application of single-stage stratified random sampling, as they relate to area sampling, will be useful.

To minimize sampling variance, the sample designer wants to define strata and area sampling units (segments) so that variation among sampling units within strata is as small as practical. That is, a sample designer is concerned with (1) the choice of criteria for stratification and the allocation of the sample among strata, and (2) the definition of sampling units, including the control of variation in size of the sampling units. Within strata, variation among sampling units will be relatively small when the sampling units are nearly equal in "size" and have similar characteristics. The designer also seeks an average size of sampling unit that is efficient with regard to mean square error for a given cost. These matters of sample design are related to the purpose of the survey.

As just indicated, there are typically two ways of using auxiliary data in the design of an area sample: One is for stratification, the objective being to achieve homogeneity within strata; and the second is the use of an auxiliary variable as a "measure of size", the purpose being to achieve segments of equal "size" where the measure of size is a variable that is correlated with the variables to be included in the survey. Some kinds of information are useful for purposes of stratification but are not useful as measures of size for controlling the size of segments. (Examples are geographic location, soil types, or maps, showing broad types of farming areas.) There are characteristics (e.g., acres of cropland) that can be used either as a measure of size or as a basis for stratification. Generally, the same auxiliary variable would not be used as both a measure of size and a criterion for stratification.

Theoretically, the choice of criteria for stratification of frame units and the choice of a measure of size, which is used for assigning numbers of segments to frame units and controlling variation in size of segment, are not independent choices. When the options permit, the author generally prefers to give first priority to the choice of a measure of size to control segment size and second priority to the criteria for stratification with due regard to the measure of size, the estimator, and the survey objectives. However, opportunity to consider design alternatives is limited by the degree to which the area sampling frame is developed (including auxiliary data by frame units) to accommodate various survey objectives.

7.3.1 Control of segment size. In theory, ways of controlling (reducing) sampling variance associated with variation in the size of sampling units include stratification of the sampling units by size, selecting sampling units

with pps (probability proportional to size), ratio or regression estimators, and equalizing the size of the sampling units. In the discussion and illustrations that follow, attention will be on equalizing segment size. However, mapping detail and topographic landmarks, as well as the kind of auxiliary information that might be available, often severely limit the degree to which equalization of segment size can be achieved. If relevant information exists for controlling variation among segments, but topography severely limits equalization of segment size, the other methods listed could be considered. With regard to ratio estimation, remember the precaution stated in section 4.4.3.

The selection of individual segments with pps has generally not been used, and it involves technical considerations beyond the scope of this publication. Incidentally, selecting segments with pps is not the same as selecting frame units with probability proportional to assigned numbers of segments and then dividing each selected frame unit into its assigned number of segments, etc. The latter is a method that gives each segment an equal probability of being in the sample. Stratification applies to frame units, rather than individual segments. Under some circumstances stratification can be a useful aid in controlling segment size. This situation will be illustrated later.

The choice of a measure of size of segments depends on the purpose of the survey and whether the open-, closed-, or weighted-segment method is to be used. Controlling the size of segment involves the assignment of a suitable number of segments to each frame unit and the appropriate division of frame units into segments. For example, consider a survey of fruit crops. Suppose the closed-segment method is to be used and that an approximate measure of the amount of land used for fruit crops is available by frame units. In this case, the number of segments assigned to frame units would be proportional to the approximate amount of land used for fruit crops. The goal would be to divide a frame unit into the assigned number of segments so each segment has approximately the same amount of land under fruit crops. This principle is used in the illustrations presented later.

7.3.2 Stratification and the definition of frame units. As stated in paragraph (6) of 7.2, auxiliary information might be used in ways that have a bearing on how frame units are defined. A leading example of this is classification of all land area according to land use and then delineating frame units within each of the land use classes. An alternative is to delineate frame units with very little, if any, regard for land use and then stratify the frame units by land use for sampling purposes. The question of whether to take land use information into account before or after the frame units have been delineated is one of the first questions to be answered. When comparing the alternatives and making a choice, it is important to distinguish between procedural advantages and other matters such as sampling efficiency or potential for bias. Situations can be described where, for practical purposes, the choice would be a matter of procedure rather than sampling efficiency.

Land use classes might be delineated, prior to the delineation of frame units, with such purposes in mind as (1) stratification to achieve homogeneity within strata, (2) having frame unit boundaries coincide with areas that might be used as domains of study, or (3) forming classes of frame units so the frame units within a class would be treated alike but one class might be treated differently from another. The land use pattern, topography, and the anticipated

purposes of the sampling frame have an important bearing on the choice of specifications for frame units. Perhaps a brief discussion of two hypothetical cases involving very different land use and topographic patterns will be helpful.

Case 1. Suppose the total land area for which an area frame is to be developed, can be readily divided into four land use areas (classes): Tree crops, cultivated crops, grazing land, and nonagricultural land. Assume the land use patterns and topography are such that (1) the land classes can be delineated so the boundaries of the classes are suitable as frame-unit boundaries, and (2) the land use within a class conforms to the class except for rather small widely scattered parcels of land which do not account for more than 10 or 15 percent of the total land area of the class.

In this case, delineating land use classes and frame units within the land use classes is probably advantageous. The frame units within a class would be relatively alike and the land area of the frame units could serve as a useful measure of size for a number of sampling purposes. That is, a list of frame units by land use class and the land area of each frame unit provides a basis that is reasonably satisfactory for general purpose sampling; and, it gives a basis that can be refined or further developed as needed.

As an illustration, suppose a sample for a survey of cultivated crops is to be designed and selected. One of the first decisions to be made is the geographical extent of the population to be sampled. Let us assume that the two land use classes, nonagricultural and grazing, may be omitted, but the tree-crop land use class has too much land in cultivated crops to be ignored. The two land use classes, cultivated and tree crops, would be sampled differently as follows.

Assuming that either the closed segment or the weighted-segment method is to be used, an appropriate measure of the size of a segment is the amount of cultivated land. That is, the goal is to define segments so they all have approximately equal amounts of cultivated land. In the cultivated land use class, a very high proportion of the land is cultivated. Therefore the total land area of the frame units is a suitable measure of size in lieu of estimates of the amount of cultivated land by frame units. Thus, under the circumstances, making the numbers of segments assigned to the frame units proportional to total land area of the frame units will probably lead to segments that are about as equal in size as would be the case if the assigned numbers were proportional to the amount of cultivated land in the frame units. Converting the land areas of frame units to numbers of segments is a simple matter after a decision on the average size of segment is made. For example, suppose the average size of segment is set at 300 acres. A frame unit with an estimated 1,400 acres would be assigned five segments. For sampling purposes, the frame units in the cultivated land class could be stratified geographically, or according to any other appropriate criteria that might be available. The selection of frame units and the division of the selected frame units into segments would be in accord with principles that have already been discussed.

In the tree-crop land use class, consideration should be given to how the cultivated land is geographically distributed. If the cultivated land is uniformly distributed among the frame units, the assignment of numbers of segments to frame units could proceed in the same way except that the average size (land area) of segment would be larger. For example, if about 10 percent of the land

is cultivated and a decision has been made to have the average segment contain 100 acres of cultivated land, the total land area of the average segment would be 1,000 acres. Hence, a frame unit with a total land area of 5,000 acres would be assigned five segments. If the proportion of cultivated land varies widely among frame units, the method just described could be used, but consideration should be given to an alternative that would have lower sampling variance. For example, it might be feasible to examine photographs and assign segments to frame units approximately in proportion to the apparent amount of cultivated land in each.

If the open-segment method had been chosen for this survey, attention to the density of farm headquarters would be needed, instead of the amount of cultivated land, when assigning numbers of segments to frame units.

Case 2. In contrast to Case 1, suppose that the area for which a frame is to be constructed has a pattern of land use and topography such that it is not possible to delineate land use classes, within which frame units would be alike, unless frame unit boundaries are allowed to be tenuous. An example is an area where most of the land is not cultivated because of soil or topographic conditions, and the land that is cultivated is mostly small, widely scattered, irregularly shaped parcels of land. If one is to delineate broad land use classes, within which frame units would be delineated, a major compromise must be made. Either homogeneity of land use within a class or the quality of frame unit boundaries must be sacrificed. Moreover the task of delineating land use classes prior to delineating frame units could be time consuming and difficult under some circumstances.

At relatively low cost, frame units could be delineated with very little, if any, regard to land use. Approximations of the amount of land under various uses could be compiled for each frame unit and used either as (1) measures of size for the assignment of numbers of segments to frame units or (2) criteria for stratifying frame units for sampling purposes. Thus it is possible to make effective use of land use information without using it in the delineation of frame units and without introducing tenuous frame unit boundaries. Section 9 presents an illustration of this kind of situation.

The writer regards the choice of frame-unit boundaries as critical. A part of the boundary of many segments will be a frame-unit boundary. Experience has shown that tenuous frame-unit boundaries are very troublesome in the application of area sampling, especially after a few years have passed since the frame was constructed. As stated earlier, the frame units should be regarded as the most permanent aspect of an area sampling frame. Flexibility to serve various kinds of surveys is not necessarily restricted by how the frame units are defined. Regardless of definition, frame units may be stratified in various ways and they may be divided into segments in various ways for various purposes. Also, auxiliary information about frame units may be updated or supplemented at any time. Achievement of efficiency in the design of a sample depends on the relevance and accuracy of information pertaining to individual frame units. That is, it is the range of relevant information about individual frame units that provides adaptability of the frame for various survey purposes.

The delineation of land use or other classifications prior to delineating frame units is, in effect, one way of compiling information about frame units.

Compared with the simple approach of delineating frame units with minimum regard to land use, it should be justifiable on the basis of (1) more effective use of the auxiliary information involved (which, in general, seems doubtful to the writer), or (2) economy in the operations of constructing a frame and selecting samples. In any case, land use probably should not be ignored completely when delineating frame units. For example, urban and other nonagricultural areas might require special consideration. But consider the alternatives carefully before making a large investment in the delineation of land use classes prior to defining frame units, especially if the quality of the boundaries of frame units is sacrificed.

7.3.3 Selection of auxiliary data about frame units. The availability of auxiliary data varies among countries and applications from almost none to information that is highly relevant and effective in the design of samples to minimize sampling variance. The sample designer is constantly confronted with making choices among alternatives that have a bearing on sampling efficiency and bias. Also, it is often his responsibility to make recommendations or decisions about auxiliary data that seem to be worth acquiring for future use in sample designs. For continued improvement of sampling plans and operations, there should be a continuing program of investigation and analysis of various components of error and components of cost in surveys that are conducted.

Total land area is likely to be near the top of any list of auxiliary information that is to be compiled for frame units. It can be approximated quite easily from scaled maps and will probably be used in many sampling plans. Estimates of the amount of land in each frame unit by land use classes might be important, depending on the kind of surveys that are expected and the circumstances as discussed in 7.3.2. The amount of land in each frame unit by land use classes is generally more useful (effective in reducing sampling variance) for the closed- and weighted-segment methods than for the open segment.

Possible sources of information about frame units include: (1) Census data if frame units correspond to enumeration districts, (2) land use maps if sufficiently detailed, (3) aerial photographs, and (4) visual estimates from field observations of the frame units. Visual estimates of the proportions of land in the various uses for each frame unit could be multiplied by the estimated land area of the frame unit to obtain measures of the amounts of land under various uses, which might be useful for sampling purposes. The land area of a frame unit can be estimated by using a planimeter or a grid overlay, if scaled photographs or maps are available.

If the open-segment method is to be used for surveys of all farms, an indication of the number of farms "in" each frame unit would be useful, assuming it contributes to the objective of equalizing the number of farms "in" segments. For surveys of households, information on the number of households by frame units would be important.

Information about frame units should not be obtained, especially if much cost is involved, unless there are good prospects that it will be used in an effective manner to reduce sampling variance. The cost of obtaining auxiliary data needs to be considered with regard to the reduction in sampling variances that might be achieved through improved sample design. How does the cost compare with the cost of achieving comparable reductions in sampling variance by

increasing sample size? An investment in auxiliary data to improve the sample for one survey might not be advisable. But if surveys involving the same commodities (or subjects) are conducted periodically, a substantial investment in auxiliary data might be fully justified.

Special important needs should be considered very carefully. For example, suppose a particular tree or vine crop is commercially very important to the economy of a country. Information about the exact location of the crop, or approximations of the amount of the crop in each frame unit, might be critical to obtaining a satisfactory degree of sampling efficiency. Field work to acquire auxiliary information about frame units might seem too expensive, but the cost of low sampling efficiency might be greater. It is of interest to note that census counts of fruit trees have sometimes been justified mostly on the need for a good basis for sampling for current forecasts or estimates of production. Information about frame units that is very effective in designing samples for current, special-purpose surveys can sometimes be obtained at a much lower cost than a census.

The capability for designing efficient area samples in agriculture (especially special-purpose sampling) is heavily dependent on information about where various crops or commodities are produced. If no auxiliary information is available for designing efficient samples and if such information is too expensive to obtain, consider the possibility of a double sampling plan. That is, select a large sample and collect data on the characteristics of farms in the sample. This would provide a basis for selecting subsamples that are efficient for various specific needs. Also, do not overlook any possibilities for linking data from censuses with an area frame. A census utilizing a short questionnaire might be planned for two purposes: (1) Provide statistics about key items for publication, and (2) supply auxiliary data to be associated with an area sampling frame that would enable more efficient sampling and estimation from current surveys.

7.4 Maps for Frame Construction

It might be helpful to recognize two broad categories of maps: (1) Maps that provide useful topographic detail for delineating frame units and segments; and (2) maps that provide useful auxiliary information for the design of samples. Some examples are maps that show land use, irrigated areas, soil types, or other information that might be used for stratification or for assigning numbers of segments to frame units.

In the first category, the maps most commonly used are road maps, aerial photographs, and topographic maps. The map requirements with regard to scale and detail differ considerably for (1) purposes of delineating frame units and of providing an office record of the boundaries of frame units, and for (2) purposes of dividing frame units into segments and showing the boundaries of sample segments for use in the field. For the first purpose, road maps (or topographic maps which show roads) are generally used. File space and cost considerations might dictate that the frame units be defined or recorded on relatively inexpensive maps (and perhaps transferred to microfilm). For the second purpose, the frame maps (maps on which the frame units are defined) are not always adequate. Photographs or more detailed maps might be used or it might be necessary to adopt techniques like those described in the next section. Incidentally, when segments are delineated on aerial photographs for use in the field by

interviewers, the photographs are a valuable aid to achieving complete and accurate coverage of the sample segments, as well as providing positive identification of segment boundaries.

7.5 Division of Frame Units into Segments

The division of frame units into segments often presents a wide range of problems. It might be feasible to divide some frame units using the frame maps, but aerial photographs or more detailed maps are generally very useful and often necessary. When available mapping detail does not enable satisfactory division of a frame unit into its assigned number of segments, there are a number of techniques that might be helpful. Some alternative techniques are:

(1) Have the interviewer enumerate the frame unit completely. That is, treat the frame unit as a sample segment and fill out a questionnaire for all reporting units in the frame unit. Suppose k is the number of segments assigned to the frame unit. For purposes of tabulation, a subsample (using $\frac{1}{k}$ as the subsampling fraction) of the reporting units enumerated might be used. If all reporting units are included in tabulation, remember to use the probability p as a basis for weighting, where p is the probability which the frame unit had of being in the sample.

(2) Before the survey begins, have a list of reporting units in the frame units prepared and select a subsample of reporting units, using a sampling fraction of $\frac{1}{k}$. In this case an interviewer would be given a sample of reporting units rather than a segment.

(3) Travel to the frame and divide it into k segments on the basis of observed landmarks. Make sure that sketches and notes provide adequate description of the segments. Select one segment at random in the office.

The first alternative is most practical when k is small, say 2 or 3. Generally speaking, the third alternative appears preferable to the second, when the closed or weighted segment is being applied or when the same sample is to be used repeatedly.

It is often feasible, using the maps on hand, to partly divide (but not completely divide) a frame unit. For example, assume a frame unit is to be divided into five segments. It might be feasible to divide it into two parts and to assign three segments to the first part and two segments to the second part. One of the two parts would be selected at random giving the first part a probability of 3/5 and the second part a probability of 2/5. The selected part could then be handled in accordance with one of the above alternatives. The value of k would be 3 or 2, depending upon which part was selected. This technique of partly dividing a frame unit might reduce the number of maps or photographs that are needed. For example, it might be feasible to partially divide a frame unit using only a road map. Then, to complete the job of dividing the frame unit into segments, photographs would be needed for only part of the total frame unit.

Sometimes one finds that dividing a frame unit into the assigned number of k segments is possible only if undesirable boundaries are accepted. However, the landmarks might be such that the frame unit will divide very satisfactorily into k-1 segments. This situation presents a choice between "forcing" a division of the frame unit into k segments or dividing it into only k-1 parts. If the division into k-1 parts is accepted, two alternatives are open: (1) Treat the k-1 parts as segments, select one at random, and for purposes of estimation, change the probability of selection from $(\frac{1}{k})$ to $p(\frac{1}{k-1})$, where p is the probability which the frame unit had of being selected. (2) Number the parts 1 thru k-1. Suppose part 1 is the largest. Assign it two segments and assign one segment to the remaining k-2 parts. Then select one part with probability proportional to its assigned number of segments. If one of the parts 2 thru k-1 is selected, use it as a segment. It had a probability of selection equal to $p(\frac{1}{k})$. If the first part is selected, one of the three techniques described at the beginning of this section could be applied to it. The value of k would be 2.

In the processes of delineating and selecting segments, always be on the alert to specify procedural detail that eliminates the possibility of bias. For example, it is very important that the process of dividing frame units into segments be separated from (that is, be completely independent of) the process of making random selections. To illustrate how bias can be introduced, suppose the instruction to the clerical staff is to divide a frame unit into segments and to select one at random before proceeding with the next frame unit. When a random number is selected it might be possible, unless special precautions are taken, to see the next random number on the list. Knowledge of the next random number could seriously bias the work of delineating and numbering segments in the next frame unit.

Another illustration of potential bias is changing a segment boundary after the segment has been selected. There might be a strong inclination to do this when one finds that a better boundary is needed for an interviewer to follow. If changes are allowed, changes should be held to a minimum and strict rules for making any changes in boundaries should be specified, which are believed to be unbiased for practical purposes. Such practices always introduce a potential for bias and a degree of uncertainty about the magnitude of any bias in the results. On the other hand, some adjustments in boundaries might involve less risk of bias than letting interviewers enumerate segments that have ambiguous boundaries. The best policy is to avoid this situation to the fullest extent feasible. Be as sure as possible that boundaries are satisfactory before random selections are made. This gives emphasis to the point made earlier, namely that frame unit boundaries should coincide with permanent, well-defined landmarks.

Sometimes a difference in detail seems unimportant and a decision is made on the basis of convenience. Do not take unnecessary risks with procedural detail that could introduce bias.

Thoroughly test feasible alternatives before setting final specifications for a sampling frame. Testing is needed to determine costs, to evaluate alternatives, and to debug procedures.

8. Frame Construction--Illustration No. 1

Two areas representing different topographic and land use situations were selected for illustration of area sampling frames and sample selection. The first area for illustration is a part of Mills County, Iowa. Nearly 95 percent of all land in Mills County is in farms. About 85 percent of the land in farms is cropland, and the average size of a farm is more than 300 acres (or 121 hectares). Approximately 85 percent of the farm operators live on their farms. The density of farms is about two per square mile.

In a large part of the United States, including Mills County, the Public Land Survey divided land into sections (square miles). The standard section has 640 acres of land (nearly 260 hectares). On the county road map (see figure 2a) each section is shown as a square (1/2 x 1/2 inches) and identified by a number. A landmark of some kind (a road, a fence, or the edge of a field) follows most section lines; but, as farm practices have changed and fields and farms have become larger, landmarks that follow section lines have disappeared to some extent. In Mills County, sections can usually be identified from visual inspection of photographs, but section lines are not always satisfactory as frame-unit or segment boundaries.

The county road map, figure 2a, provides a satisfactory basis for defining frame units. In fact, in this illustration the frame units were very easy to delineate as shown in figure 2b. County lines were regarded as acceptable frame unit boundaries. Other than county lines, there was no need to consider any landmarks other than permanent roads for frame-unit boundaries. Figure 3a shows a photograph of frame unit 17. To avoid covering any detail shown in the photographs, the boundary of frame unit 17 is shown in figure 3b which is the same photograph with frame unit and segment boundaries added. Figure 3b will be discussed later. Some readers may wish to match landmarks shown on the highway map, figure 2a or 2b, with landmarks on the photograph, figure 3a.

In addition to specifications for frame unit boundaries, a specification on the minimum size of frame unit is needed. In this illustration, 4 square miles was set as the preferred minimum with 3 square miles being the absolute minimum. The maximum size of frame unit is not critical. It was about 6 or 7 square miles. Variation in size of frame unit was dictated mostly by the pattern of topographic features that were suitable for frame-unit boundaries.

The agriculture and land use pattern in Mills County is such that segments larger than 3 or 4 square miles in size are not likely to be needed for a survey. If the land was to be classified by land use and frame units defined within land use classes, specifications would have been needed regarding: (1) The land use classes, (2) the landmarks for boundaries of the classes, and (3) the minimum size of a parcel of land for each class.

Time spent on delineation of frame units could be saved by making them larger, but such savings did not appear to be important. In fact, less time is required to select samples when the frame units are small. If necessary to accommodate use of larger segments, frame units and related data can be combined to form larger frame units. The amount of auxiliary data that might be needed for frame units did not appear to be an important factor favoring larger (and hence fewer) frame units in this illustration.

The land areas of the frame units could be estimated by planimetering the frame map, figure 2b. However, by looking at the frame map one can judge the land areas with an error of not more than about 1/2 square mile, which is probably sufficiently accurate for sampling purposes. Column (2) of table 6 shows the approximate land area of each frame unit as determined by visual interpretation of the frame map. Needs for auxiliary information (other than land area) about frame units will be considered as the discussion continues. Incidentally, every frame unit should always be assigned at least one segment and have a chance of selection unless there is conclusive evidence that it contains nothing that contributes to the population being sampled.

To illustrate how the frame might be used to design and select samples, three kinds of surveys will be considered: (1) a survey of crop acreages, (2) a survey for economic data, and (3) a survey of beef cattle.

8.1 A Survey of Crop Acreages

Suppose a sample survey is to be conducted, after crops have been planted, for the purpose of estimating the acreage planted to each crop. For this purpose the closed-segment method is superior to the open- and weighted-segment methods, assuming that tracts are satisfactory as reporting units. Criteria for stratification and sample size are among the important aspects of a sampling plan, but attention will be focused primarily on illustrating the specification and delineation of segments. Also, the sampling problem will be considered in the context of a general-purpose sample of all crops rather than a sample designed for one or two specific crops.

With reference to the purposes and conditions that have been outlined, an appropriate goal in delineating the closed segments is equalization of the sizes of the segments with regard to amount of cropland. The first step is to assign a number of segments to each frame unit. If the segments are to contain equal amounts of cropland, the assigned numbers of segments should be in proportion to the amounts of cropland in the frame units. In Mills County, a very high proportion of all land is cropland. Thus the land area of the frame units, after making any feasible deductions for nonfarmland, is a very good measure of size.

Since photographs are available for dividing the frame units, it is feasible to set the average size of segment at one-half of one section. A smaller average size that might be considered is a quarter section, but that does not appear to be practical, and coverage error tends to increase as the segments become smaller. The fourth column of table 6 shows the number of closed segments assigned to each frame unit. The numbers assigned are two times the estimated numbers of square miles (column (2), table 6) with the exception of frame unit 24. The frame units were reviewed quickly to identify apparent areas of nonfarmland that were larger than about 1/2 of one square mile. The only such area was a town that was partly in frame unit 24. There were three square miles in frame unit 24, but it was assigned five segments rather than six because it had at least 1/2 of one square mile of area that was residential. Thus, the idea was to have the assigned numbers of segments proportional to the land areas of frame units after deduction of any nonfarm areas larger than 1/2 of one square mile. If frame unit 24 is selected for division into segments, its entire land area would be included in the five segments, even though the residential part of

Table 6.--Frame units and numbers of segments for illustration #1

Frame unit number	Approximate size of frame units in square miles	Indicated number of farms	Closed or weighted segments		Open segments	
			Assigned number	Accumulated number	Assigned number	Accumulated number
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	7	25	14	14	16	16
2	4	15	8	22	10	26
3	4	10	8	30	6	32
4	5	11	10	40	7	39
5	5	17	10	50	11	50
6	3	8	6	56	5	55
7	8	14	16	72	9	64
8	6.5	20	13	85	13	77
9	6	17	12	97	11	88
10	4	14	8	105	9	97
11	5	11	10	115	7	104
12	4	12	8	123	8	112
13	4	11	8	131	7	119
14	3	17	6	137	11	130
15	5.5	19	11	148	12	142
16	4	13	8	156	8	150
17	4	6	8	164	4	154
18	4	12	8	172	8	162
19	6	14	12	184	9	171
20	6	18	12	196	12	183
21	6	25	12	208	16	199
22	6	24	12	220	16	215
23	7	15	14	234	10	225
24	3	8	5	239	5	230
25	4.5	17	9	248	11	241
Total		373	248	248	241	241

the frame unit was not counted when the number of segments was assigned. That is, the residential part of the frame unit would be included in one or more of the five segments. Also, one would attempt to define the five segments so they contained equal amounts of cropland.

The fifth column of table 6 shows the accumulated number of segments for the closed- or weighted-segment methods. Cumulative totals are often generated as a convenient way of selecting frame units with probabilities proportional to the assigned number of segments. A discussion of alternative methods of selecting a sample of segments from the 248 assigned in column 4 of table 6 involves technical consideration beyond the scope of this publication. However, suppose one segment is to be selected at random. A random number is selected from 1 thru 248. Assume the random number is 157, which with reference to the accumulated total is more than 156 and less than 165. Thus, frame unit number 17 is selected. It had a probability of selection equal to $\frac{8}{248}$.

The next step is to divide frame unit number 17 into the assigned number of segments, which is 8. This frame unit divides very satisfactorily under the criteria of good boundaries and uniformity in size with regard to amount of cropland (see figure 3b). After numbering the 8 segments 1 through 8, one of the 8 is selected at random. Suppose segment number 7 is selected. It has an overall chance equal to $(\frac{8}{248})(\frac{1}{8}) = \frac{1}{248}$ of being selected.

Additional segments could be selected in the same manner. However, systematic selection as follows is often used. Suppose the sampling fraction is 2 percent or 1 out of 50. A random number from 1 thru 50 would be selected. This designates the first number in a series of numbers having an interval of 50. Suppose the random number is 12. The series is 12, 62, 112, 162, and 212, which with reference to table 6 designates frame units 1, 7, 11, 17, and 22, within which segments are to be delineated and one segment is to be selected at random. Since there are two steps, selecting frame units and then selecting a segment in each, this selection procedure is sometimes confused with two-stage sampling. In the case just described, the two steps are two selection steps in a single-stage sampling plan.

Figure 3c is a photograph of segment number 7 on an enlarged scale. It is an example of a photograph that an interviewer might take to the segment, except that the tract and field lines within the segment would not be shown. After traveling to a segment and getting oriented (that is, matching the boundaries as shown on the photograph with the actual topography) the interviewer divides the segment into tracts. In segment 7 there are only three tracts: A, B, and C. Next, in an interview with the operator of a tract the interviewer divides the tract into fields and obtains the desired information about the crops. Notice that a photograph of a segment is an important aid to minimizing coverage and measurement error.

If the photographs are scaled, the fields could be planimetered and the results used as a check on acreages reported by the operators. Even when an operator is not available for interview, an interviewer can probably obtain most of the desired information about crop acreages. He might talk with suitable informants, or by visual observation he might delineate fields and record,

to the extent possible, the crop that has been planted in each field. The field acreages can be estimated. Thus, the closed segment provides a means for getting data that are accurate and very nearly complete, compared with what is possible or feasible when some other survey methods are used.

It was stated above that, in this example, the land area of a frame unit, less nonfarmland, was a good measure of size. That is true primarily for crops that are generally grown. For minor crops (crops with relatively small acreages) an auxiliary variable such as acres of cropland or farmland is generally of less value in reducing sampling variance.

Special attention must be given to any important "minor" crops with requirements that their sampling variances be low. One approach is to select a "general" sample that is designed to be adequate only for major crops, but information about all crops would be collected. In addition, one or more supplemental samples could be designed specifically for the minor crops. Results from the general and supplemental samples would be combined, using appropriate weights. A basis for designing supplemental samples for particular minor crops is implied. Otherwise, there is no alternative to making the "general" sample larger.

Auxiliary information by frame units giving some indication of the amount (or proportion) of the land that is likely to be planted to each of the minor crops can be very useful in sample design. Assuming it is possible, the measure of size of frame units and of segments for a supplemental sample might be very different from the measure of size used in the general sample. As stated before, a major question is how much to invest in obtaining auxiliary information about frame units. The analogous question with regard to list frames (lists of farm operators for sampling purposes) is, What information should be developed and maintained about individual farms on the list? Incidentally, the production of some minor crops might shift from year to year among farms or locations so that auxiliary data on where they were grown at some time in the past might be little or no value.

Before proceeding to the next example, a comment about the value of photographs seems in order. In the absence of photographs, the requirement that boundaries of segments be identifiable from the county maps would have meant larger segments and less success with equalization of the sizes of segments. In other words, at least for the situation discussed above, some reduction in sampling variance can be attributed to use of photographs. The photographs also help reduce coverage error. Initial reaction to the cost of photographs might be that they are too expensive. Before reaching such a conclusion, consider the cost of not using photographs. That is, consider the cost of achieving an equivalent reduction in sampling variance by increasing the size of the sample. Also, consider the possibility of the same photographs being used for several surveys.

8.2 A Survey for Economic Data

For a survey to collect data about economic characteristics of all farms, it is possible to use either the open or weighted segment. Since the procedure outlined above for closed segments is also appropriate for a survey of farms using the weighted segment, the following discussion will pertain to the open-segment method. Although a survey is regarded as general purpose, there might be a need, because of analytical purposes for varying the sampling rates by,

for example, size or type of farm. This will be discussed later. In the meantime, it is assumed that all farms should have an equal chance of being in the sample.

The density of farms in Mills County is about two per square mile. Experience based on analyses of variance, costs, and coverage error suggests that the best average size of open segment is probably less than two farms for the agriculture and topography in this illustration. An average size of one farm per segment is assumed, which means that we want the number of segments assigned to a frame unit to be equal to the number of farms "in" the frame unit. There is no practical way of accomplishing this exactly.

The basis for assignment of segments should be determined with regard to how farm headquarters is defined. If the operator's residence is by definition the farm headquarters, information on where operators live is useful. In this case, the goal would be to assign numbers of segments to frame units which are in proportion to the numbers of operators living in the frame units. There might not be a good basis for doing that. On the other hand, suppose a specified point within the boundaries of each farm is the farm headquarters. If information on the location of headquarters is not available, segments might be assigned in proportion to amount of farmland or cropland.

With regard to Mills County, about 85 percent of the operators live on their farms and some of the remaining 15 percent live in the open country. Let us assume that the farm headquarters is the operator's residence if the operator lives on the farm; otherwise, it is some other defined point on the farm. Available information and the discussion in the preceding paragraph point to two alternatives. The first is to assign segments to frame units in proportion to land area. The goal was an average of one farm per open segment; and, since the density is two farms per square mile, the average size of segment would be 1/2 of one square mile. Therefore, this alternative gives an assignment of segments that happens, in this case, to be the same as the assignment of closed segments in column (4) of table 6. The division of frame units into segments would be different, however, because the objective is to equalize the number of farms in the segments.

The second alternative is to derive, as follows, an indication of the number of farms "in" each frame unit and then allocate segments in proportion to the indicated numbers of farms. In the open country, the road maps show square symbols, ■ , which indicate the location of farm dwellings (or farmsteads). These symbols are not always correct, but they are useful. At some of these indicated locations the dwelling unit might not be occupied by a farm operator. In fact a dwelling might not be found at one of the indicated locations. Moreover, some operators live at locations which are not identified on the maps. However, a count of the indicated farm dwellings shown on the county map is presented in the third column of table 6. This count (373) is, judging from the census of agriculture, about 50 percent more than the actual number of farms. A more accurate indication of the numbers of farm dwellings in the frame units can probably be obtained by examining photographs. From photographs one can identify building sites where farmers probably live, but again this does not give an accurate and complete identification. However, indicated numbers of farms have often been derived and used in the assignment of open segments to frame units. Regardless of how open segments are assigned to frame units, when

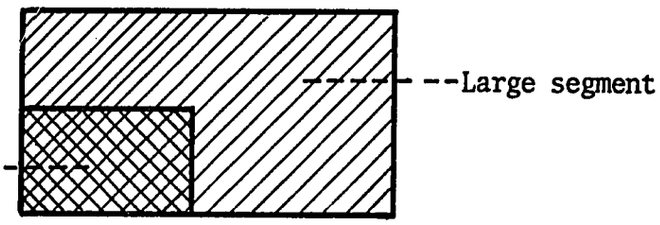
a frame unit is divided, it should be divided into the assigned number of segments with the objective of having the same number of farms in each segment.

For purposes of illustration, we will use the indicated numbers of farm dwellings in column (3), table 6, for assigning segments. Recall that these indicated numbers are about 50 percent larger than the actual number of farms. We are seeking an average of one farm per segment. Thus, the assigned numbers of open segments in column (6) are about two-thirds of the indicated number of farm dwellings shown in column (3).

As an example, frame unit number 17 will be divided into segments, since it was used previously. The number of segments assigned was 4 (see column (6), table 6). Figure 3d shows frame unit 17 divided into four segments, assuming use of the open-segment method. Incidentally, the photograph (figure 3a) shows six places where a farm operator probably resides. This happens to agree with the road map.

As a special case, suppose that a uniform sampling fraction is satisfactory except for estimates needed for a domain that is a small proportion of the population. Sampling variances of estimates for this domain are too large. Let us call the farms in this domain "type A" farms. How can the size of the sample of type A farms be increased without increasing the sample of all farms? If the type A farms are concentrated sufficiently, it might be feasible to define the area of concentration and simply increase the sampling fraction in that area only. If that technique is not appropriate, there are variations of at least two other general approaches that might be considered:

(1) The first is most applicable in situations where the type A farms are uniformly distributed among all farms. In this case, it is appropriate to make the segments to be screened for type A farms "larger" than the segments for a sample of all farms. This suggests the possibility of using a large and small segment where the small segment is also a part of the large one. For example, suppose type A farms are to be sampled using a sampling fraction that is four times larger than the sampling fraction for all farms. The first step is to design and select a sample of large segments to be screened for type A farms. Then divide each large segment into four segments and select one of the four at random. The following sketch illustrates a pair of large and small segments.



The sample of small segments gives a sample of all farms and the sample of large segments, which includes the small segments, is the sample for type A farms. An interviewer would probably be instructed to work the small segment first and treat it as though the large segment did not exist. He would then screen the remainder of the large segment for type A farms only.

A specific example of a possible use of a pair of large and small segments is a survey of the costs, amounts, and kinds of materials used in the construction of new farm buildings and in repairing and remodeling old farm buildings. Repairs are made on a very high proportion of all farms each year, but in any one year a new building is constructed on only a small proportion of all farms. New construction is important and its sampling variance per farm is relatively large, hence a larger sampling fraction is needed for new construction than for general repair and maintenance. Thus if the method of large and small segments were adopted, information would be collected on all repair, maintenance, and new construction in the small segments. The remainder of each large segment would be screened for new buildings that had been constructed and data about the new buildings would be collected.

(2) The second general approach is to design two samples: A general-purpose sample of all farms and an independent sample specifically designed for type A farms. The next section, 8.3, presents an example of special-purpose sampling. But first a word of caution is interposed.

Although, conceptually, there should be no difference, in practice there is a likelihood that farms identified as type A in the sample of small segments will differ on the average from farms identified as type A in the large segments. The same could be said for farms identified as type A in a general-purpose sample and farms identified as type A in a supplemental special-purpose sample for type A farms. Differences greater than expected from sampling error often occur when changes in survey procedures are made, even though the concepts and definitions of the parameters are the same.

8.3 A Beef Cattle Survey

If more than about one-third or one-half of the farm operators produced beef cattle and if none of the operators has extremely large numbers of cattle, a rather simple area sampling plan that did not make use of specialized auxiliary data about beef cattle might provide satisfactory sampling efficiency. But as farming becomes more specialized and larger farms develop, it becomes increasingly necessary to treat each commodity (or group of commodities) as a special sampling problem.

In Mills County there are less than 900 farms. Census data show that nearly 40 percent of the farms have no cattle and less than 50 farms account for almost half of the beef cattle. Thinking of area sampling and the possibility of using sections (areas that are one square mile) as sampling units, there would be many sections with no beef cattle and a very small number of sections with large feeding lots that might have more than 1,000 cattle. Area sampling as described in 8.1 and 8.2 would be inefficient. That is, very large sampling fractions would be required to get satisfactory results. One solution is to compile a list of large cattle enterprises and use multiple-frame sampling as mentioned in section 2.5.2. But this discussion is being limited to area sampling.

To have a basis for efficient area sampling for a cattle survey, it is essential that information be available about the location of cattle. Large feedlots, or facilities for feeding large numbers of cattle, can often be identified on recent aerial photographs. If necessary, someone could travel

over the area involved and make appropriate inquiries to identify and locate at least the large cattle enterprises. ("Large" in this context might mean enterprises that would have a selection probability greater than 0.5 if individual enterprises were selected for a sample with probability proportional to size). If medium-to-large enterprises can be identified with a moderate additional cost, that probably would be worthwhile. Incidentally, for sampling purposes, "size" of a feedlot enterprise probably should be measured in terms of capacity rather than number of cattle present on a particular date.

As a simple illustration, suppose 50 large beef-producing enterprises have been located on maps. Fifty segments would be defined, which would include the 50 enterprises, one corresponding to each enterprise. Each segment should be large enough to include all of an enterprise and the usual requirement of identifiable boundaries should be fulfilled. These 50 segments would be treated as a separate subpopulation or stratum and an appropriate sampling plan applied to it. To sample the remainder of the population, the 50 segments would be deleted from the frame units in which they are found. After this deletion, the design and selection of an area sample of the remainder would follow principles that have already been discussed. The subpopulation of 50 segments would be sampled, using a large sampling fraction relative to the remainder.

The above procedure is applicable for the closed- and weighted-segment methods. For the open segment, special attention should be given to the definition of farm headquarters. If the definition of headquarters results in any of the 50 enterprises not being included in the stratum of 50 segments, there could be serious loss in sampling efficiency.

Three examples of area sampling have been outlined briefly for an area where a high proportion of the land was cultivated and where the topography was relatively favorable for area sampling. In the next illustration, the topographic and land use patterns are different.

9. Frame Construction--Illustration No. 2

For the second illustration a part of Johnson County, in southern Illinois, was selected. Figures 4a and 4b for Johnson County correspond to 2a and 2b for Mills County. All of the county is shown except a narrow strip along the eastern edge, which was omitted to avoid having to show the map on a smaller scale. Because of the topography, the frame units are larger and more irregular in shape than the ones in the first illustration. The choice of landmarks for frame-unit boundaries is more difficult. For example, county lines are fully described and shown on official land records, but visible landmarks do not always coincide with county lines. Technically, frame units could overlap county lines. In that case, if the boundaries of the county happen to coincide with the boundaries of a population to be sampled, each frame unit overlapping the county line (boundary of the population) would be identified prior to sampling. Then, the part of each such frame unit that is within the county would be marked and treated as any other frame unit of the population. Allowing frame units to overlap county lines might provide for better frame-unit boundaries. On the other hand, many maps, photographs, and statistics are prepared by counties and there is some inconvenience in having frame units overlap county boundaries. In this illustration, figure 4b, the frame units were allowed to overlap the county lines.

For a perception of the land use and topography see figures 5 and 7. Figure 5 is an aerial photographic mosaic for a portion of the county that includes frame units 22, 23, 29, and most of 21. This mosaic is part of an index to individual photographs which are identified by numbers in the upper right corners, for example BGS 1 MM-42. When looking at the mosaic do not mistake the edge of a photograph for a landmark. There is a large amount of overlap among the photographs and the photographs do not match exactly because of scale differences. The scale of the mosaic in figure 5 is approximately 3/4 of an inch equals 1 mile. Each photograph covers an area approximately 2-1/4 by 2-1/4 inches. Figure 7 is a photograph of a part of frame unit 23. It is on a larger scale and shows more detail. Figure 7 will be discussed later.

Three broad classifications of land use can be recognized in the photographs: woodland, residential or built-up areas, and the remaining land which is used mostly for agricultural production and will be referred to as "farmland". This information on land use may be used in different ways. One way, mentioned earlier, is to delineate land use classes and then delineate frame units within each class. The topography in Johnson County is such that the proportion of farmland, for example, would vary widely among frame units belonging to the same land use class. That is unavoidable unless the condition that frame units must have permanent, unmistakable boundaries is relaxed to a degree that would permit frame units to have very tenuous boundaries. The frame units in figure 4b were delineated without regard to land use. That is, the idea in this illustration is to use information about land use after the frame units have been delineated. It is assumed that the land areas of the frame units have been estimated, probably by planimetering the frame maps.

9.1 A Survey of Crop Acreages

In Johnson County, the proportion of farmland varies among frame units from about 35 to 75 percent. For a survey of crop acreages assuming the closed- or weighted-segment methods, the approximate acreage of farmland in each frame unit appears to be a much better measure of size than the total land area. There are at least two feasible methods of approximating the amount of farmland in the frame units:

(1) Estimate the amount (or proportion) of farmland in each frame unit by placing a transparent grid overlay on a photograph or by planimetering. If proportions are estimated, amounts can be estimated by multiplying the proportions by the approximate land areas of the frame units. The work should be done with care, but a large amount of time spent on trying to make such measurements as accurate as possible is probably not worthwhile in terms of effect of sampling variance. A high degree of accuracy compared to rough approximation might make very little difference in the numbers of segments assigned to the frame units. Furthermore, when a frame unit is divided, it is possible to equalize the amount of farmland in the segments only to a limited degree, depending on the available landmarks for segment boundaries.

(2) The second method is less exact and consumes less time. By looking at the photographs, classify the frame units as high, medium, or low with regard to the proportion of the land that is farmland. For example, the objective might be to visually classify frame units with more than 60 percent farmland as high, 40 to 60 percent as medium, and less than 40 percent as low.

The census of agriculture shows that about one-half of the total land area of Johnson County is in farms. Crops are harvested from about one-fourth of the land in farms, almost one-fourth of the land in farms is woodland, and much of the land in farms is used for grazing. The average size of farm is approximately 200 acres and there are about 1.8 farms per square mile. Land judged to be farmland from looking at the photographs (that is, land not covered by trees or used for residential or industrial purposes) might be quite different from land in farms according to the census. However, for a crops survey, using the closed- or weighted-segment methods, farmland as interpreted from photographs is a useful and feasible measure of size for assigning segments to frame units.

The topography of Johnson County is such that the average size of segment probably should not be less than about 500 or 600 acres of farmland. Thus, one square mile (640 acres) of farmland is specified as the average size of segment for this illustration. The average segment will contain about 160 acres (1/4 of a square mile) of land from which crops are harvested. If an estimate of the amount of farmland, expressed in square miles, is available for each frame unit, the number of segments assigned to the frame units would be the estimated square miles of farmland rounded to the nearest whole number. Every frame unit should be assigned at least one segment, with the exception of any frame units that have been intentionally omitted from the population to be sampled.

Suppose that each frame unit has been classified as high, medium, or low with regard to the proportion of its total land area which is farmland. Assume that the average proportions of farmland for these three classes are 0.7, 0.5, and 0.3. The land areas in square miles of the frame units in each class would be multiplied respectively by 0.7, 0.5, or 0.3 to determine the assigned numbers of segments. A more exact assignment of segments is possible. One must judge whether a more exact method would be worthwhile. Note that the classification of frame units by land use was discussed as a device for assigning segments and not as a criterion for stratification in the sense of stratified random sampling. The frame units may be stratified in any way that is appropriate for the survey.

Frame unit 23 has been selected for illustration. By planimetering the frame map, which is scaled, an estimate of 8.6 square miles in frame unit 23 was obtained. From larger-scale photographs than shown in figure 5, it was estimated with the aid of a grid overlay that approximately 60 percent of the land in frame unit 23 was farmland. This gives 5.2 square miles (8.6 x .6) as an estimate of the amount of farmland. Thus, according to the specifications for segments, which were discussed above, frame unit 23 is assigned 5 segments. Assume that a number of segments has been assigned to all frame units in a similar manner. Further assume that, for a crop acreage survey, frame unit 23 has been selected and that it is now ready to be divided into 5 segments.

A study of photographs with detail comparable to that in figure 7 showed that frame unit 23 does not divide easily into 5 segments with nearly equal amounts of farmland. The situation presents the typical problem of trade-off between clarity of segment boundaries and equalization of segment size. However, frame unit 23 divides into four well-defined parts by the roads shown in figure 4b. The four parts are shown in figure 6.

The photographs indicate that part no. 1 has the most farmland and that it will subdivide quite satisfactorily into two parts. Thus, two alternatives

are presented: (1) Accept parts 2, 3, and 4 as segments and divide part 1 into two segments giving a total of 5; or (2) permit tenuous segment boundaries in order to equalize the amount of farmland. The first alternative does not make full use of the information on land use. The second alternative reduces sampling variance but increases the potential for bias. Under the circumstances, the writer prefers the first alternative unless tests under operating conditions show that the second alternative is operationally feasible and that bias can be avoided.

Figure 7 shows part 1 of frame unit 23 divided into two segments. A small but well-defined river was used as a boundary. For a livestock survey using the closed-segment method, the small river is a questionable boundary. Rivers often flow through grazing areas and livestock are free to cross the river. This presents a problem because the operator will not always know where his livestock are in relation to the river (the segment boundary). Notice, in figure 7, the small town and how the segment boundaries follow roads or streets into the center of the town. With the closed- or weighted-segment methods, the existence of a residential area in a segment should not, in most cases, present difficulties for an enumerator. From the viewpoint of sampling, the important part of his job is accurate delineation of tracts within the segment. With the open-segment method, residential areas present special problems.

9.2 A Survey of All Farms

For a survey of all farms using the weighted-segment method, segments would probably be defined and delineated as discussed in the preceding section. As stated earlier, the open-segment method has been used many times and many alternative ways of applying it have been tried and studied. No particular way of applying the open-segment method can be recommended as generally superior.

With regard to the application of the open-segment method to obtain a sample of all farms in Johnson County, there are no new points for discussion. To repeat, the general objective is to (1) assign numbers of segments to frame units in proportion to the numbers of farms with headquarters within the frame units and (2) divide frame units into segments so there is an equal number of farms with the headquarters in each segment. The limited means for attaining this objective leaves much to be desired. But the problem of coverage error is more serious, owing to the lack of a conceptually sound and workable definition of farm headquarters. Recall that "headquarters" is the name for a unique point that determines whether a farm is in the sample. A sampling frame that is constructed only for the application of the closed- and weighted-segment methods is simplified because it would not involve considerations of the definition of farm headquarters and the locations of headquarters. The need for full exploration of the weighted-segment method, as an alternative to the open-segment method, has become urgent.

10. Summary and a Brief Look Forward

Sampling frames should be constructed in recognition of the fact that agriculture is composed of numerous subpopulations that must be sampled. A sample designed efficiently for one subpopulation might be of little value for

another. Thus several sampling frames might be required; or, if a single sampling frame is to be constructed, it probably should be multipurpose.

In general, as agricultural enterprises become more specialized and larger, it is necessary to develop more flexible sampling frames for selecting samples for many purposes. For example, 30 years ago in some regions of the United States the same sample might have been reasonably efficient for both crops and livestock. But this is no longer the situation. To sample efficiently for a commodity such as beef cattle, it is necessary to (1) have an adequate list of cattle producers for sampling purposes, (2) use multiple-frame sampling involving area sampling and a list of at least the largest producers, or (3) develop area sampling on an efficient basis for special purposes as in 8.3. The development of improved sampling frames is called for by (1) the trend toward larger, more specialized farms, (2) the general demand for more accurate statistics, and (3) the need to keep sample sizes and costs as low as possible. Also, to some degree, sample size is inversely related to capability for controlling nonsampling error, which is another point in favor of efficient sampling to keep sample sizes as small as possible. The problem of respondent burden in answering survey questions is another factor that supports smaller, more efficient samples. These factors are calling for directing more resources to the construction and maintenance of sampling frames that will provide for higher degrees of efficiency in the design of samples.

There are numerous sources of error and ways of reducing error. Survey plans should include provision for studies of sampling variance, response errors, coverage errors, and costs. Such studies should provide a continuing basis for adjusting the allocation of resources in an effort to achieve maximum accuracy at a given cost.

Area sampling is not likely to replace sampling from lists of farm operators or vice versa. One of the most important problems in surveys of farm enterprises lies in the unclear linkage between operators and farms which results in coverage error. The linkage problems are prevalent when sampling from lists and in area sampling, especially when the open segment is used. The closed segment avoids most of the coverage error caused by obscure linkage between operators and farms. This is a major important point in favor of the closed segment. For surveys where tracts are suitable reporting units, the closed segment is likely to continue as an effective method. For surveys where farms are the reporting units, the writer believes that the weighted segment should be fully explored as an alternative to the open segment. The sampling variance per segment for the weighted segment is less than the sampling variance for the open segment. We need to know more about comparative costs and coverage error to get a clearer indication of the circumstances under which one method might be better than the other.

In a situation where the closed segment is applicable to only part of the questions, the closed segment might be used in combination with the open or weighted in order to take full advantage of the closed segment. In the writer's judgment, experience will show that the closed-weighted combination is better. If experience happens to show that the weighted segment has low coverage error, the question of whether to use the closed-weighted combination or only the weighted might come into play because the latter has the advantage of using

only one definition of a segment in the same survey. Incidentally, with modern computing equipment, the weighting of data should no longer be regarded as a major obstacle to use of the weighted segment.

In recent years, many people have become very interested in remote sensing, including the impact that it might have on area sampling and procedures for making agricultural estimates generally. This is a major subject involving a large amount of conjecture. However, perhaps a few of the author's general views are worth stating.

One short-range impact of remote sensing will be an increase in the demand for "ground" data from area samples which can be correlated with sensor recordings. That demand is already developing. In a somewhat longer range, as remote-sensing technology develops, information will probably become available which can be used to improve substantially area sampling frames and the efficiency of area sampling. This could result in major reductions in the size of area samples for some purposes, particularly for characteristics closely related to land use and physical environment.

A large fraction of all agricultural statistics involves quantities or activities that are not amenable to measurement by remote sensing. But consider crop acreages and yields. Is it possible that remote-sensing technology could completely eliminate the need for collecting data on acreage and yields by present methods?

The development of models for estimating or forecasting crop yields from sensor recordings requires accurate data on crop yields from an independent source, that is, measurements on the ground. Assuming that practical operational models are developed, a continuing need to improve the structure of the models is expected, and this will require, to some extent, continued collection of data on crop yields by present methods. Furthermore, changes in yields associated with technological advancements will change parameters in the models and require a continued effort to update the models. This means ground observations for a sample of fields representing the range of conditions that are involved.

A similar point applies to estimating crop acreages. Models for interpreting sensor recordings are required. Probably the models or parameters in the models will always be subject to change. At best this will require a minimal amount of area sampling on the ground that is concurrent with collecting sensor data. Also, to serve the analytical purposes of some farm surveys, it is necessary to have data on crop acreages and yields by farms. The only source of such data is from operators.

One major foreseeable potential for remote sensing lies in the improvement of area sampling frames, which results in a choice between estimates of greater accuracy or smaller sample sizes to achieve present levels of accuracy. This, of course, applies only to agricultural data that are at least moderately correlated with information collected by sensors. Correlations of less than about 0.6 or 0.7 are usually not high enough to be seriously considered. A second important foreseeable potential is a basis for improving some kinds of statistics for small areas, such as counties or parts of counties.

Assuming that an adequate coordinate system for representing the boundaries of frame units or segments on computer tape becomes operational, a large amount of sensor data for frame units could become available. Thus there is a foreseeable potential for maintaining area sampling frames on tape. For some purposes, such a sampling frame could be highly efficient with regard to sampling variance. The computer could be programmed to supply well-designed samples for specific purposes.

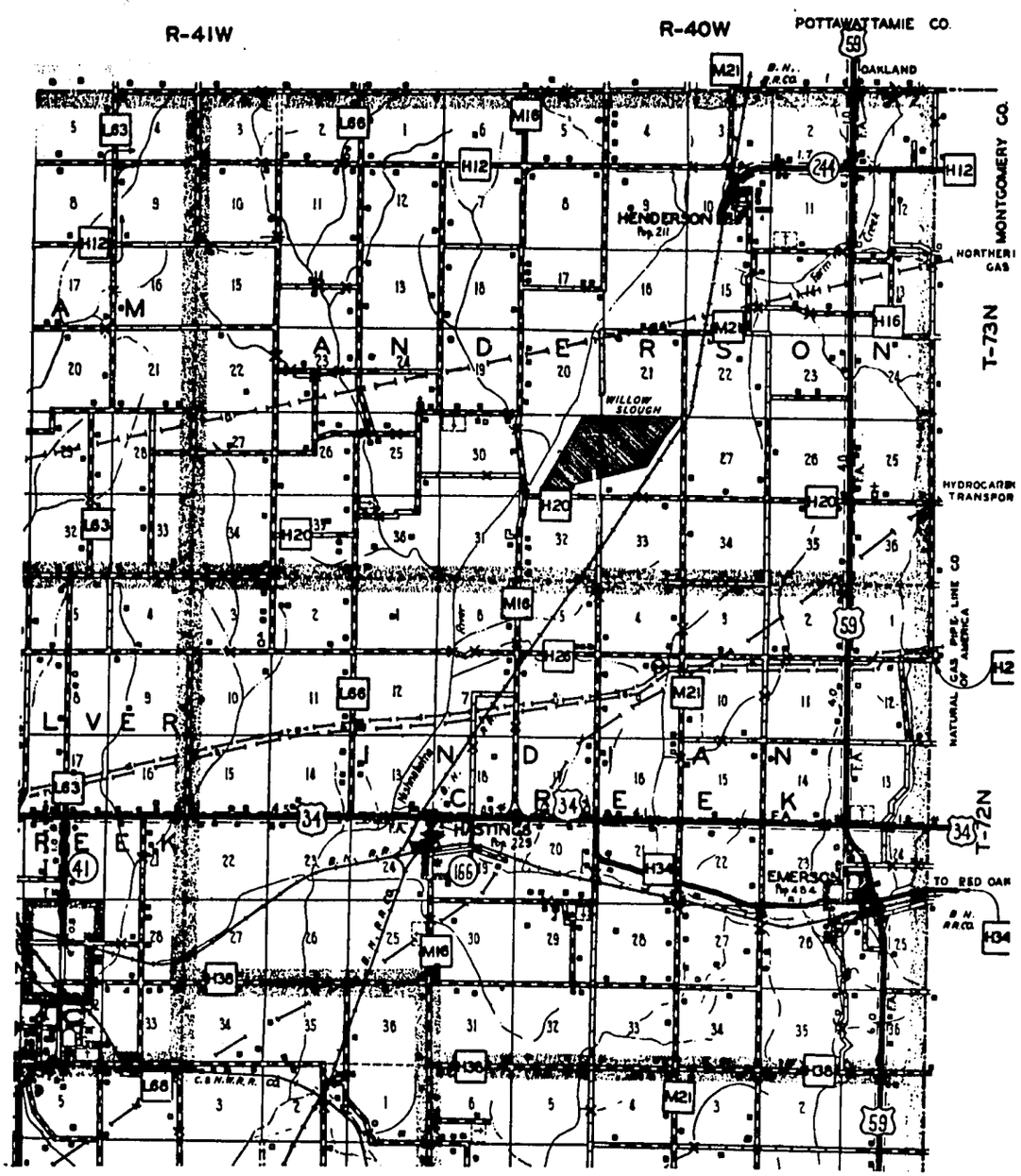


Figure 2a.--Mills County Road Map (Northeast part of County)

Scale: 1/2 inch = 1 mile

Legend (incomplete)

Roads		Other	
Unimproved		Farm unit in use	
Graded and drained		Farm unit not in use	
Soil surfaced		Business establishment	
Bituminous surfaced		Group of dwelling units	
Paved		Section line	

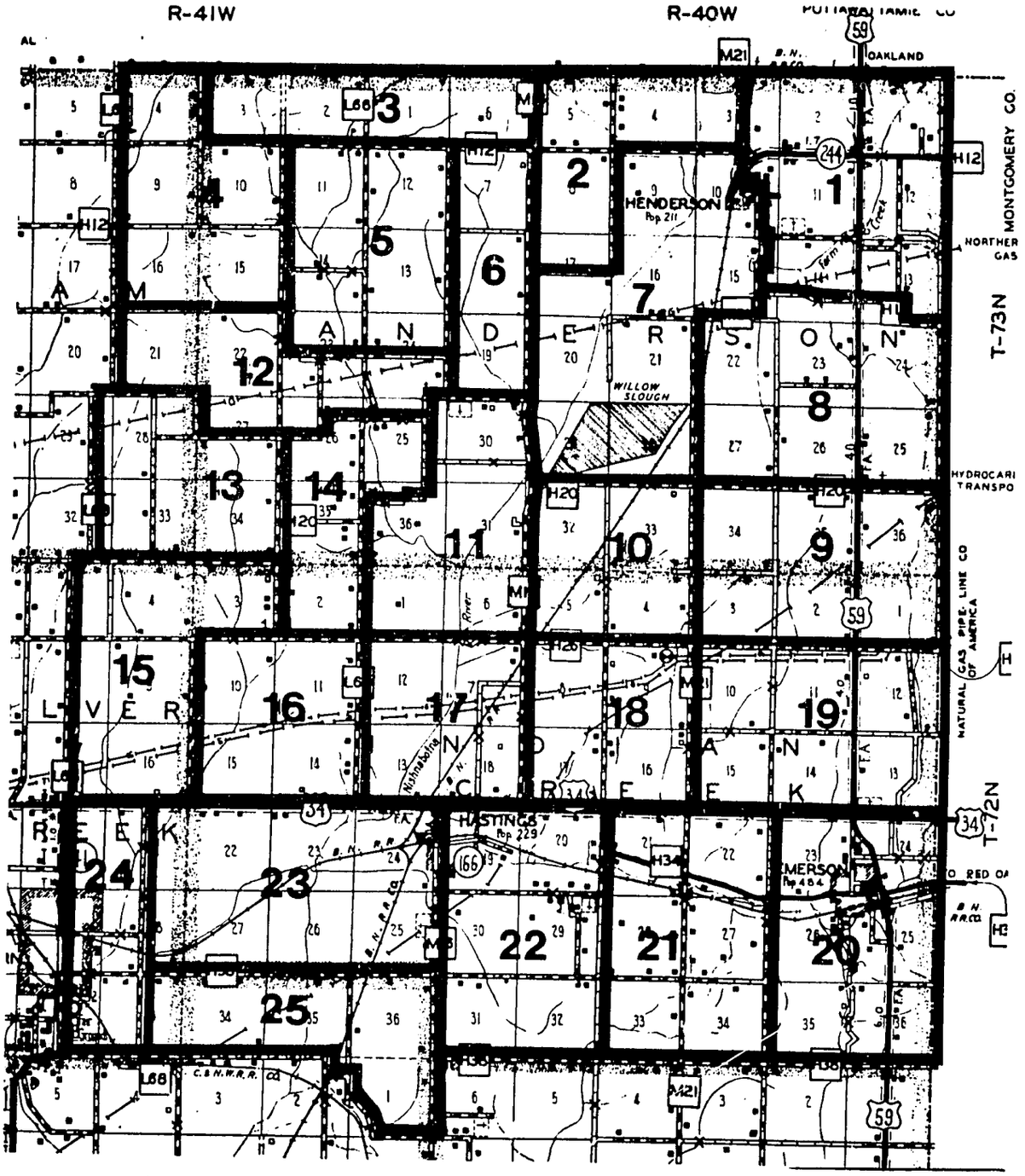


Figure 2b.--Frame Units for Part of Mills County

Legend

Frame unit boundary
12 Frame unit number

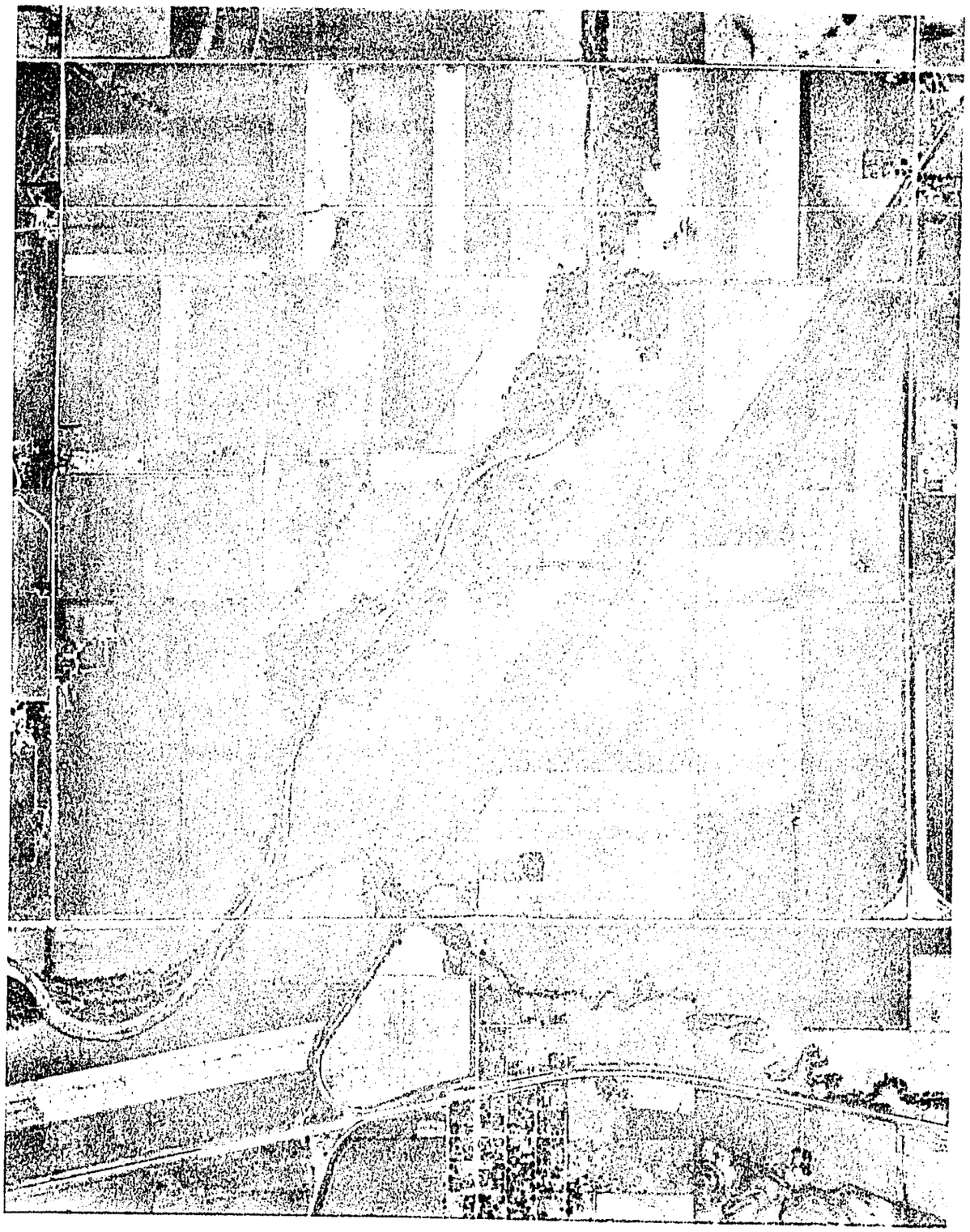


Figure 3a.--Photograph of Frame Unit No. 17

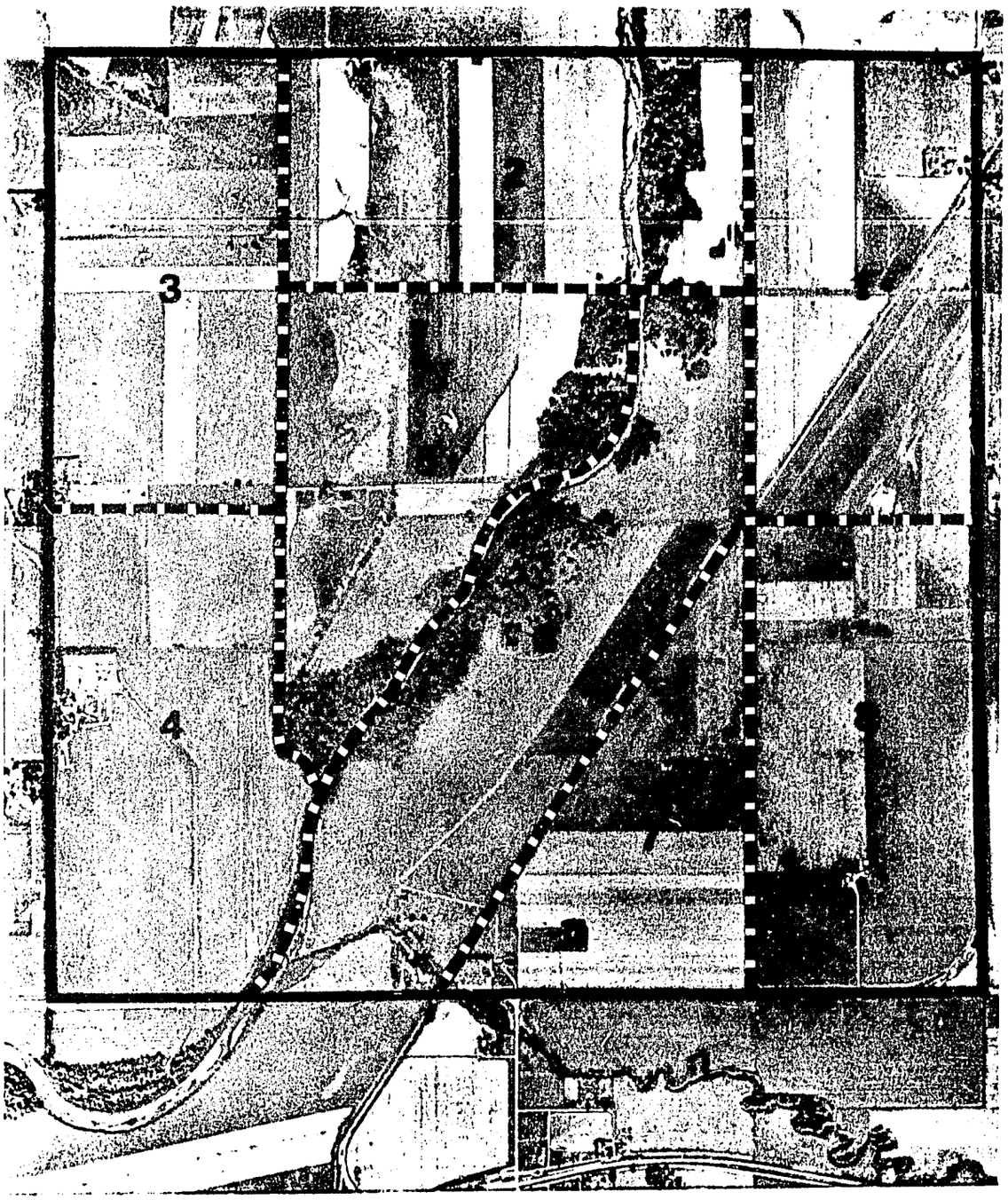
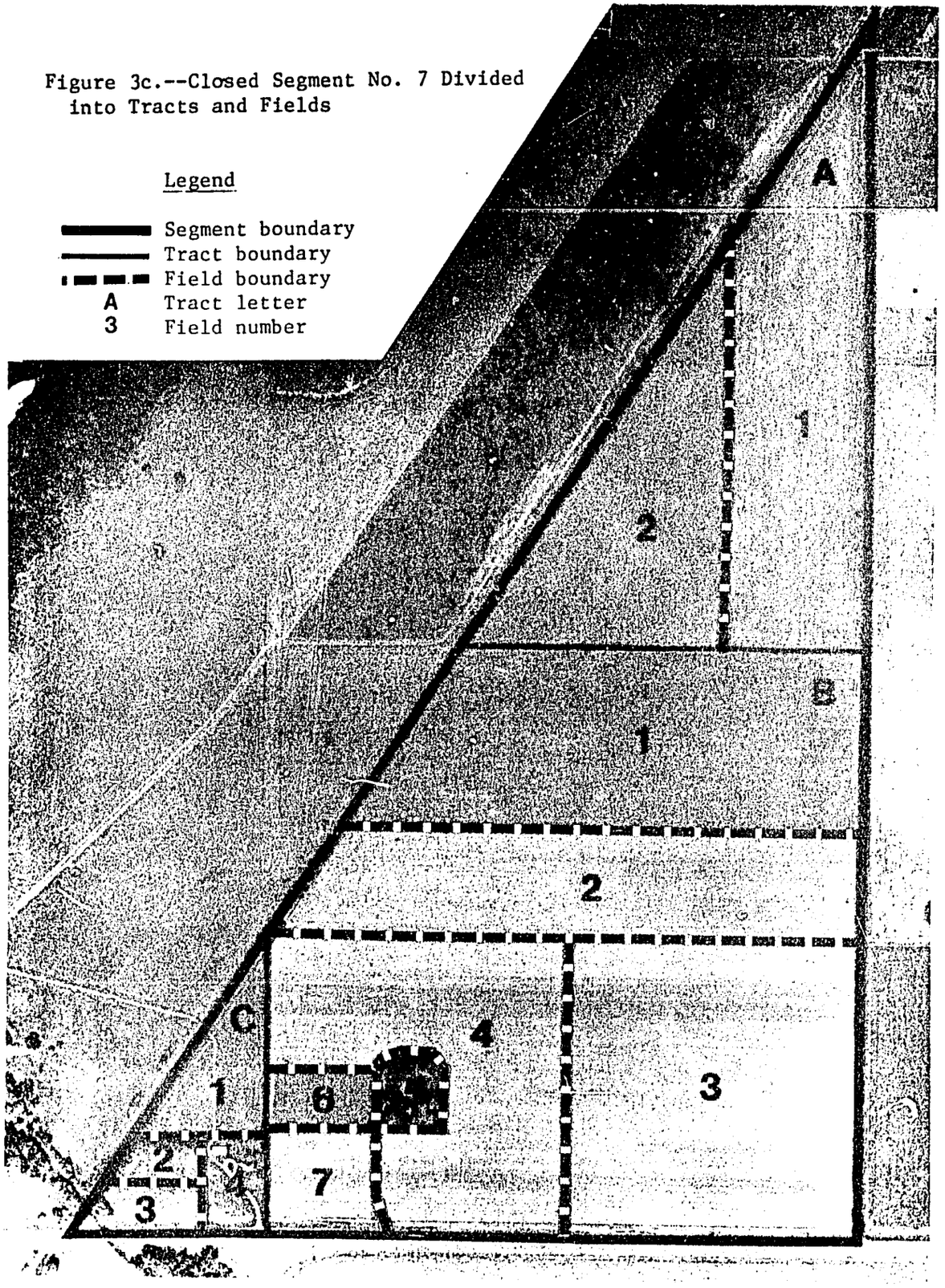


Figure 3b.--Frame Unit No. 17 Divided into
Eight Closed Segments

Legend

Frame unit boundary
Segment boundary
3 Segment number

Figure 3c.--Closed Segment No. 7 Divided into Tracts and Fields



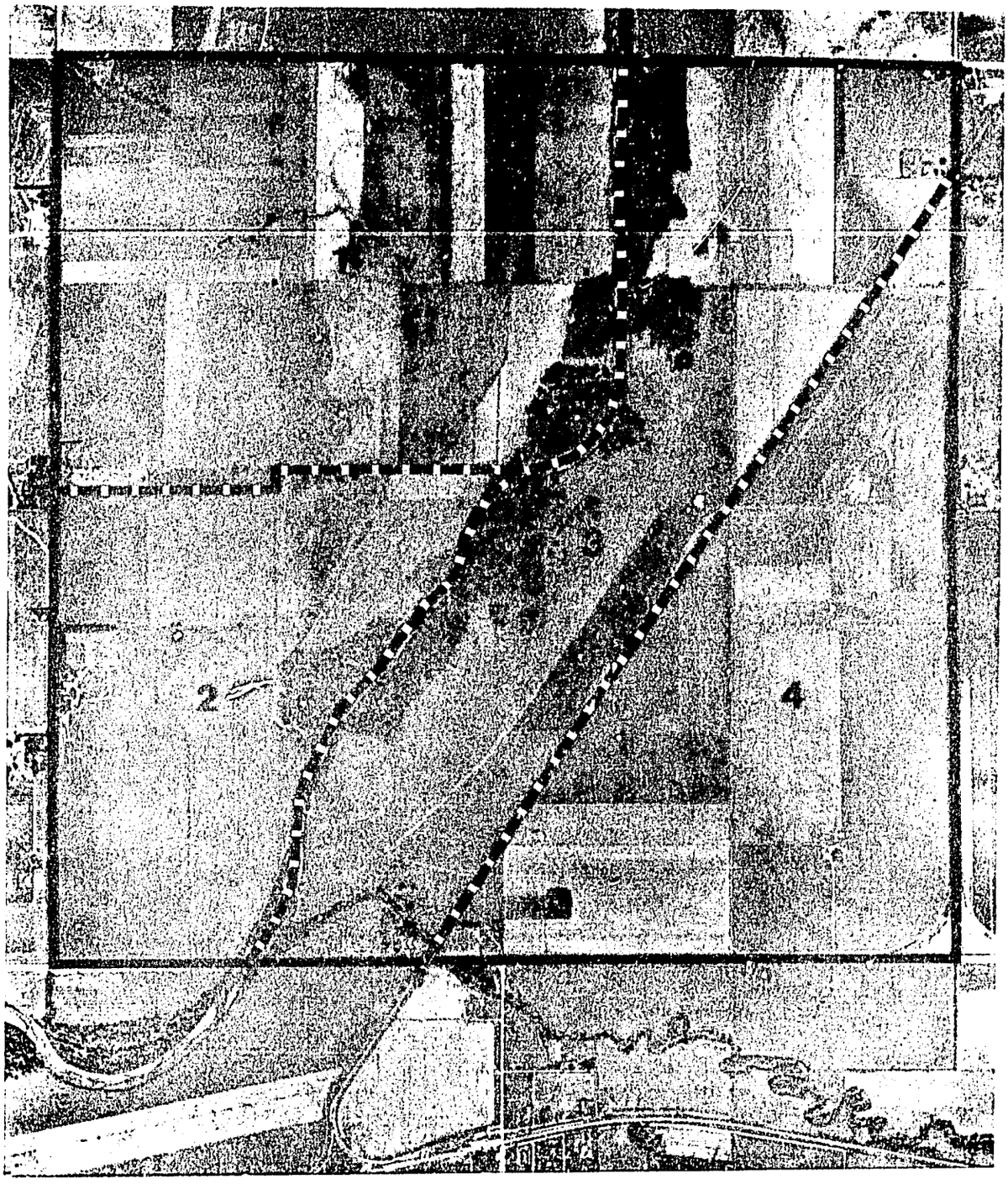


Figure 3d.--Frame Unit 17 Divided into Four Open Segments

Legend

Frame unit boundary
Segment boundary
2 Segment number

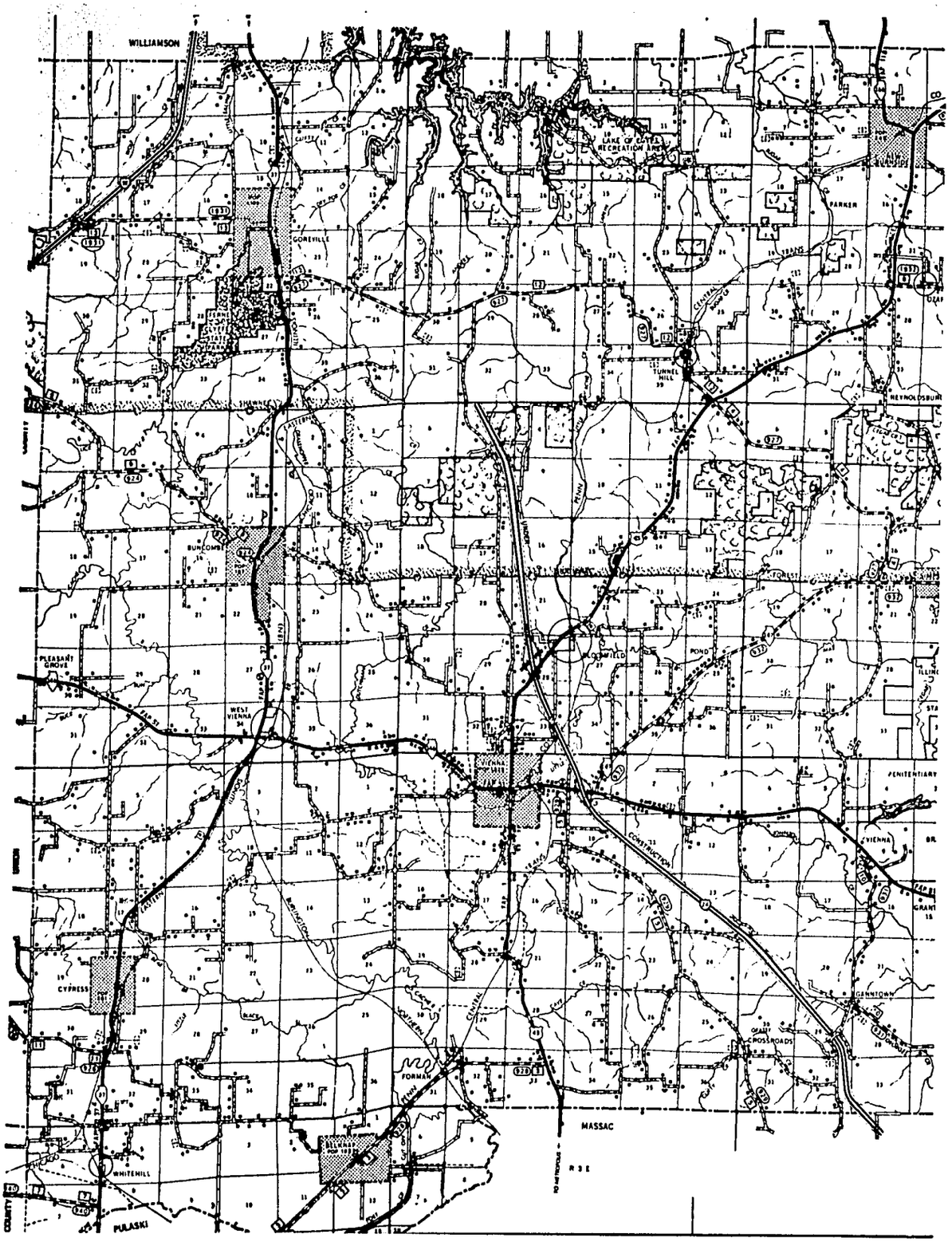


Figure 4a.--Johnson County Road Map

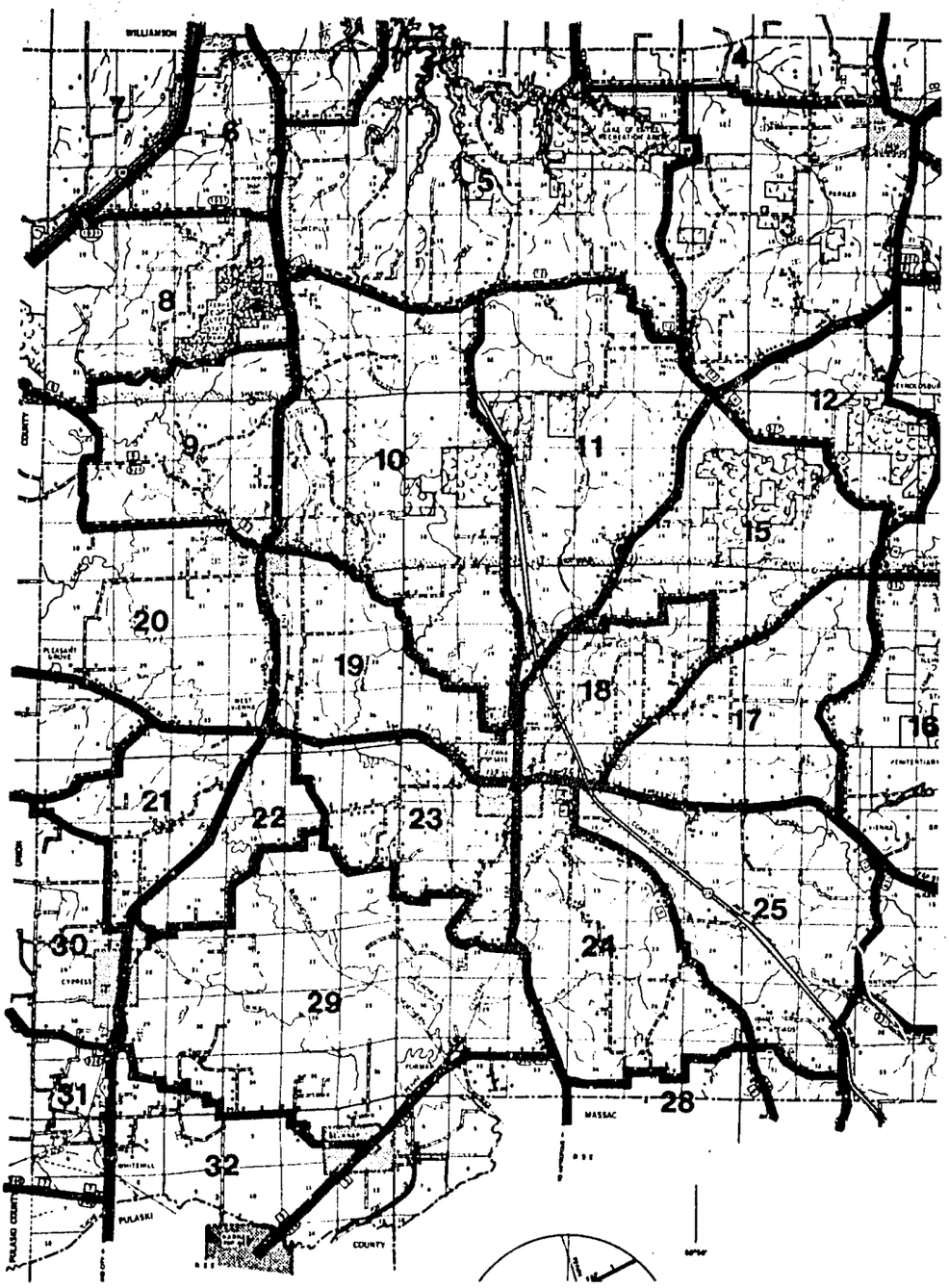


Figure 4b.--Frame Units for Johnson County

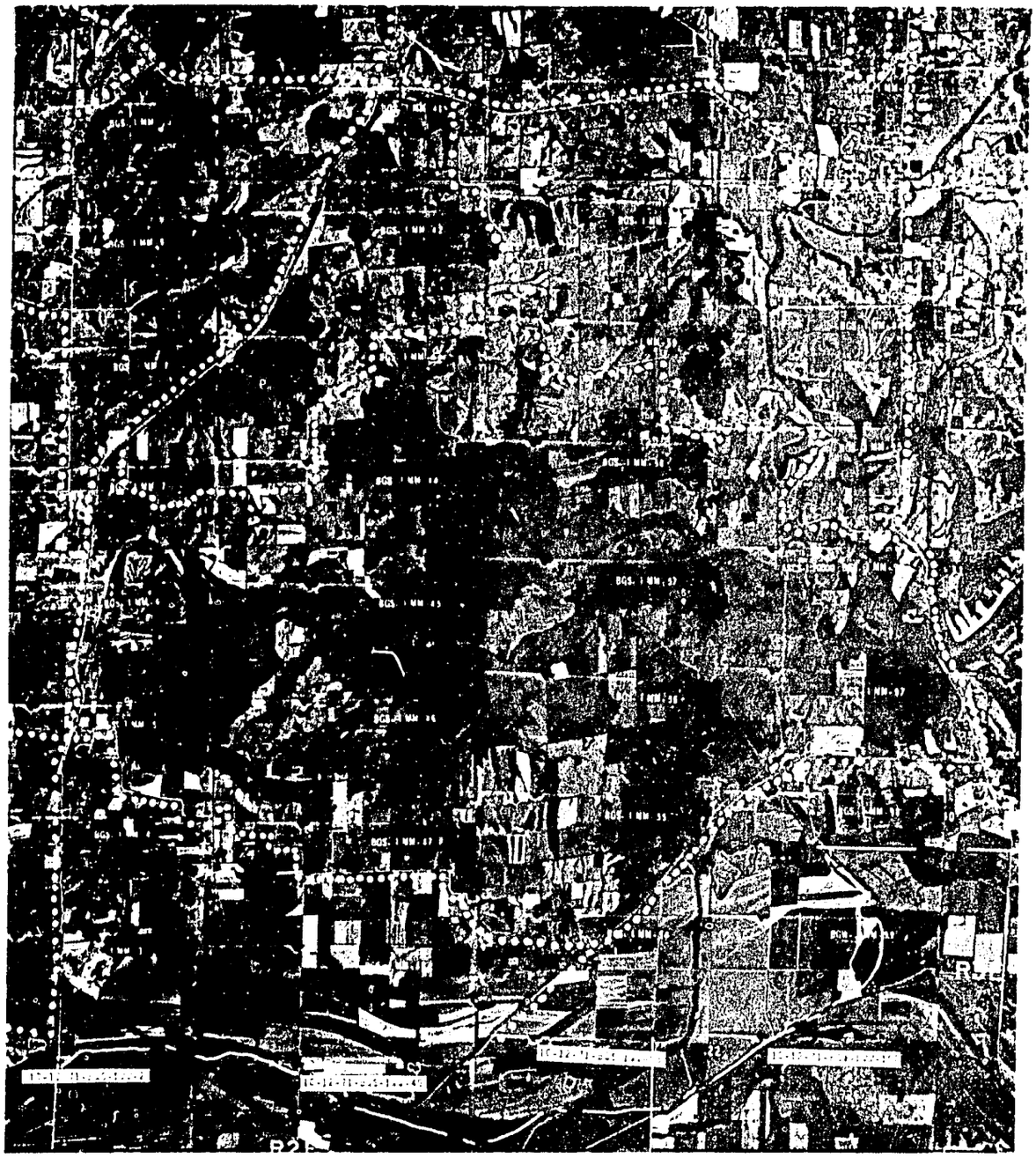


Figure 5.--Photo Index for Part of Johnson County

Legend

	Frame Unit Boundary
22	Frame Unit Number
BGS-1-MM-57	Photo Number

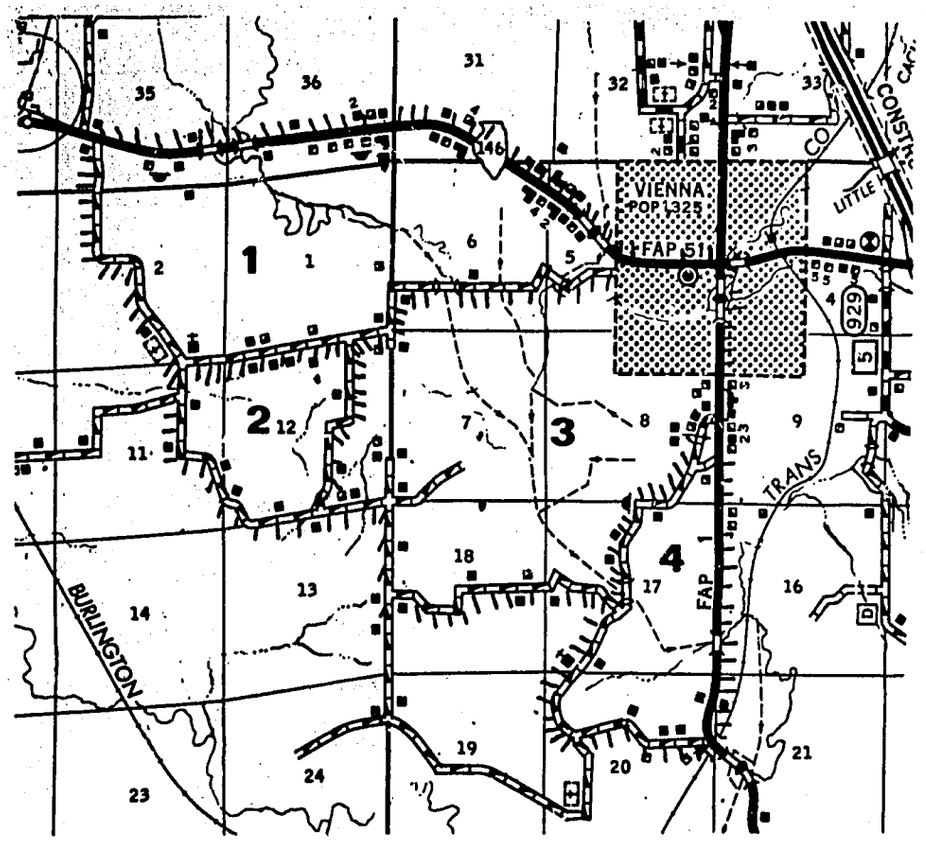


Figure 6.--Frame Unit 23 Divided into Four Parts

Legend

Part Boundary

2 Part Number

a

b

c

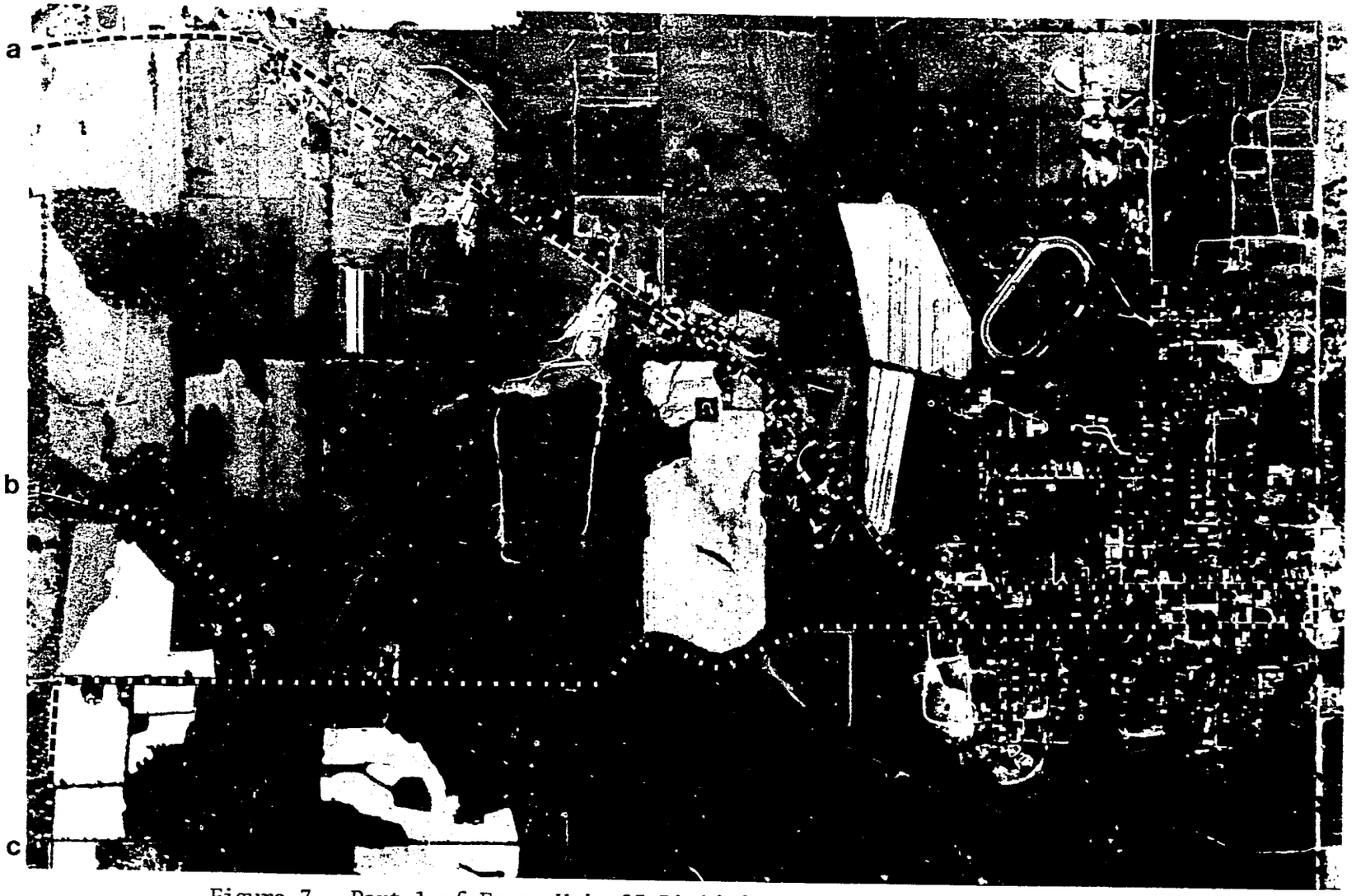


Figure 7.--Part 1 of Frame Unit 23 Divided into Two Segments

▬ ▬ ▬ ▪ Segment Boundary