AREA FRAME DESIGN FOR AGRICULTURAL SURVEYS

National Agricultural Statistics Service U.S. Department of Agriculture

Washington, D.C.

August 1987
(Reprinted June 1989)

AREA FRAME DESIGN FOR AGRICULTURAL SURVEYS. By Jim Cotter and Jack Nealon. Area Frame Section, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture.

ABSTRACT

Area frames are the backbone to the agricultural statistics program of the National Agricultural Statistics Service (NASS). The purpose of this report is to describe the procedures currently used by NASS to develop and sample area frames for agricultural surveys.

Keywords: Area frame; stratification; replicated sampling.

ACKNOWLEDGEMENT

The authors would like to thank Fred Vogel and Bob Tortora for their suggestions and support.

Washington, D.C.

August 1987
(Reprinted June 1989)

CONTENTS

PREFACE 1

CHAPTER I. AREA FRAME DEVELOPMENT

INTRODUCTION 2
- OVERVIEW 2
- BRIEF HISTORY 3
- YEAR OF IMPLEMENTATION 4
- ADVANTAGES AND DISADVANTAGES 5

PRE-CONSTRUCTION ANALYSIS 6
STRATIFICATION MATERIALS 8
LAND-USE STRATIFICATION 14
CONSTRUCTION OF PRIMARY SAMPLING UNITS 18
DIGITIZATION 21
SUBSTRATIFICATION 24
COST 28
FUTURE PROSPECTS 29
BIBLIOGRAPHY 30

CHAPTER II. AREA FRAME SAMPLING

INTRODUCTION 32
REPLICATED SAMPLING 32
SEGMENT SIZE 36
SAMPLE ALLOCATION 38
SELECTION PROBABILITIES 42
SAMPLE SELECTION 49
SAMPLE PREPARATION 53
SAMPLE MAINTENANCE 54
SAMPLE ROTATION 57
SAMPLE COST 59
SAMPLE ESTIMATION 62
BIBLIOGRAPHY 66

PREFACE

Area frames have been and will continue to be the foundation to the agricultural statistics program of the National Agricultural Statistics Service (NASS). The methodology applied by NASS to develop and sample area frames is of interest to domestic and international users of NASS's agricultural statistics. This interest has motivated the need to prepare a document that describes the current area frame procedures in NASS. Therefore, the purpose of this document is to provide a thorough presentation of the <u>present</u> area frame development and sampling procedures employed by NASS.

There are two chapters to this document. The first chapter, written by Jim Cotter from NASS, describes the procedures now used to develop new area frames. The second chapter, written by Jack Nealon from NASS, describes the area frame sampling methodology. Together, these chapters will hopefully provide the reader with a greater understanding and appreciation for the area sampling frame methods currently in use by NASS.

CHAPTER I. AREA FRAME DEVELOPMENT

BY

JIM COTTER

INTRODUCTION

OVERVIEW

The National Agricultural Statistics Service (NASS) has been developing, using and analyzing area sampling frames since 1954 as a vehicle for conducting surveys to gather information regarding crop acreage, cost of production, farm expenditures, grain yield and production, livestock inventories and other agricultural items. An area frame for a land area such as a state or country consists of a collection or listing of all parcels of land for the area of interest. These land parcels can be defined based on factors such as ownership or based simply on easily identifiable boundaries as is done by NASS.

The purpose of this document is to describe the procedures used to develop and sample area frames for agricultural surveys. The process involves many steps, which have been developed to provide statistical and cost efficiencies. Some of the key steps are:

- STRATIFICATION: The distribution of crops and livestock can vary considerably across a state in the United States. The precision of the survey estimates or statistics can be substantially improved by dividing the land in a state into homogeneous groups or strata and then optimally allocating the total sample to the strata. The basic stratification employed by NASS involves: (1) dividing the land into about six to eight land-use strata such as intensively cultivated land, urban areas and range land, and (2) further dividing each land-use stratum into substrata by grouping areas that are agriculturally similar.

- MULTI-STEP SAMPLING: Within each stratum, the land could be divided into all the sampling units or segments and then a sample of segments selected for a survey. This would be a very time-consuming endeavor. The time spent developing and sampling a frame can be greatly reduced by: (1) dividing the land into larger sampling units called first-step or primary sampling units (PSU's), (2) selecting a sample of PSU's and then delineating the segments only for these PSU's, and (3) selecting a sample of segments from the selected PSU's.

- ANALYSES: Several decisions are made that can have an appreciable impact on the statistical and cost efficiency. These include decisions such as the land-use strata definitions, the number of substrata, the size of the sampling units, the allocation of the sample to the strata and the method of selecting the sample. Statistical analyses are necessary to guide us in these decisions.

- QUALITY ASSURANCE: Care must be taken to ensure that no land is omitted from the frame (unless by design), that no land area is included more than once and that the land is properly stratified into the land-use strata. Also, accurate measurements of the area in each PSU is needed for sampling purposes. Procedures should be set up to ensure that the manual selection of segments and preparation of segment materials follow sound statistical rules to reduce sampling errors and minimize nonsampling errors. For example, it is essential that segments have boundaries that are easily identifiable to the interviewer and respondent to minimize nonsampling errors during the data collection phase.

The major area frame survey conducted by NASS is the Quarterly Agricultural Survey (QAS) in June. This mid-year survey provides area frame estimates primarily for crop acreages and livestock inventories. During the survey, the interviewers visit each randomly selected segment, which has been accurately identified on aerial photography, and interview each person who operates land in the selected segments. With the respondent's assistance, field boundaries are identified on the photography and the acreage and crop type reported for each field in the segment. Counts of livestock within each sample segment are also obtained. This area frame information is then used to provide state, regional and national estimates for crop acreages, livestock inventories and other agricultural items. Naturally, the procedures used to develop and sample area frames affect the precision and accuracy of the survey statistics.

This chapter of the document will begin by briefly outlining the history of area frames in NASS and then will move into the analysis carried out in the pre-construction stage. The source materials used in the stratification process will be discussed and then the stratification process itself. Other topics covered will be the construction of primary sampling units, the digitization process, the use of substratification and the costs incurred in designing and building an area frame. Finally, the future prospects of area frame design will be discussed.

BRIEF HISTORY

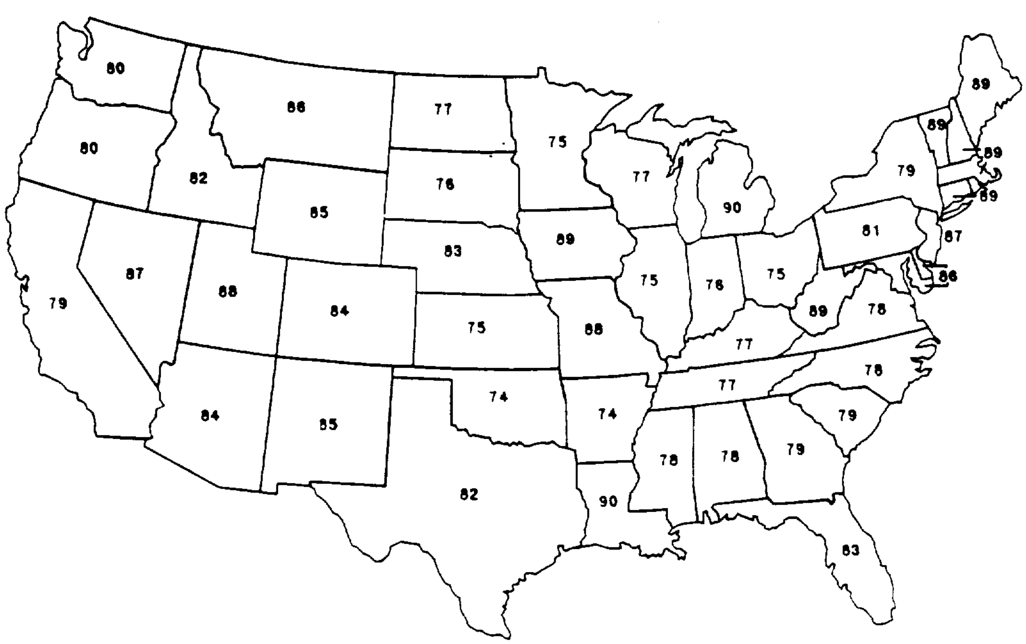
Iowa State University began construction of area frames for use in agricultural surveys in 1938. NASS began research into the use of area sampling frames in the mid-1950's to provide the foundation for conducting probability surveys based on complete coverage of the farm sector. In 1954, area frame surveys were begun on a research basis in ten states, 100 counties with 703 ultimate sampling units or segments. These surveys were then expanded over the years and made operational in 1965 in all states. Changes made to the area frame methodology during the sixties and early seventies were mainly associated with sampling methods such as land-use stratification and replicated sampling (described in detail in the second chapter of this report). Technological changes were incorporated during the seventies and eighties in the form of increased computerization, use of satellite imagery, use of analytical software and development of an area frame sample management system among others. Research has recently begun on automated frame construction techniques. The area frame program has grown over the past 33 years and is now conducted in 48 states with approximately 16,000 segments being visited by data collection personnel for the major agricultural survey conducted during June of each year.

Readers interested in a more detailed overview of the historic developments of area sampling frames in NASS are referred to a paper by Fecso, Tortora and Vogel (see Bibliography at the end of this chapter).

YEAR OF IMPLEMENTATION

Figure 1.1 displays the year of implementation of NASS's currently-used area frames for each of the states. The number shown is the year that the current frame was first used in the area frame survey program, e.g. 78 means 1978. Frames with an implementation year of 90 are currently in the planning or development stages and will be replacing area frames developed in the 1970's.

Figure 1.1 Year of Implementation of NASS's Current Area Frames



ADVANTAGES AND DISADVANTAGES

The major advantages of area frame sample surveys follow:

(A1) Versatility: Possible uses of area frame sampling are unlimited. An area frame can be used to collect varied data using one survey, for example, crop acreage, livestock, grain production and stocks, and economic data. Since reporting units can be associated with an area of land (a sampling unit), the inherent versatility is obvious. NASS randomly selects specifically cropped fields with probability proportional to size to participate in yield forecasting and estimation surveys conducted during the growing season. These yield surveys involve making counts, measurements and weighings of selected crops. The fields are selected based on data collected during the area frame survey conducted each June. NASS also utilizes the data collected from the area frame surveys to verify and classify digital satellite data. Once the satellite data has been classified for a state, an acreage estimation can be made for various crops grown in that state.

(A2) Complete Coverage: This is achieved provided there are no omissions of land area. This is a tremendous advantage since it provides the vehicle to generate unbiased survey estimates. Complete coverage is also useful in multiple frame (area and list) surveys where the area frame is used to measure the degree of incompleteness of the list frame.

(A3) Statistically Sound: The advantage of complete coverage combined with a random selection of sampling units is that it can provide unbiased estimates with measurable precision.

(A4) Nonsampling Errors Reduced: Face-to-face interviews generally result in better quality data being gathered. The interviewers have aerial photographs showing the location of the sample segment to facilitate with the collection of data such as crop acreages within the segment. Also, if the respondent refuses to participate in the survey, the interviewer has a chance to observe some of the data and to make notes which are helpful when making nonresponse adjustments.

(A5) Longevity: The area frame can be used over a period of years without having to update the sampling units.

Some disadvantages of area frame sampling are:

(D1) Can be Less Efficient than a List Frame: If a list of farm operators can be stratified by a variable related to the survey items, it will provide greater sampling efficiency than an area frame that is stratified by land-use. For example, a list frame stratified by peak number of cattle and calves will provide greater sampling efficiency than the area frame when estimating cattle inventory. Unfortunately, list frames of farm operators suffer from incompleteness, e.g. 40 percent of farm operators are not on the U.S. list frame.

(D2) Cost: An area frame can be very expensive to build and sample. Face-to-face interviews conducted by a trained staff are also very costly.

(D3) Lack of Good Boundaries: Although this is not a problem for most areas in the United States, it can be when building a frame in a foreign country. The importance of quality boundaries will be discussed later.

(D4) Sensitive to Outliers: Area frame surveys are sometimes plagued by a few "extremely large" operations that are in sample segments. These operations can greatly distort the survey estimates. A solution to this problem is to identify all very large operations prior to the survey (special list frame) and sample them with certainty.

PRE-CONSTRUCTION ANALYSIS

Prior to building a new frame (or updating an old frame), analysis is conducted to determine which states are most in need of a new frame. Data collected from approximately 16,000 segments during the QAS in June are used to evaluate the existing frames. The analysis focuses on determining the extent to which the land-use stratification has deteriorated for a frame. This involves comparing the coefficients of variation for the survey estimates of major items over the life of the frame. This also involves comparing the percentage of sample segments in each stratum satisfying the stratum definition over the life of the frame.

Another important factor in determining which state will receive a new frame is the age of the materials to be used in the stratification process. This is especially true when there are several states needing new frames. If the materials for a state being considered are not reasonably current, then it may be a waste of time and money to construct a new frame. Since these frames will be in use for 15-20 years, it would be wise to postpone that state until new photography or satellite imagery is available.

Once a state has been selected to receive a new frame, analysis is performed to determine the most appropriate stratification scheme to be used. Several land-use strata are common to all frames. These are ag-urban, urban, cultivated land and non-agricultural land. The cultivated land is divided into several strata based on the distribution of cultivation in the state. Also, if necessary, crop-specific strata such as a vegetable stratum are created. Previous years' survey data are analyzed to provide information such as the percent of cultivated land in the sample segments so that the distribution of cultivated land can be ascertained. This will help decide on the number of and definition of the cultivated strata. Also, the distributions for major crop acreages will decide whether crop-specific strata will be worthwhile. Table 1.1 presents the land-use stratification scheme generally followed along with the codes to be used during the stratification process.

Table 1.1 Land-Use Strata Codes and Definitions

STRATUM CODE	DEFINITION
11	General Cropland, 75% or more cultivated.
12	General Cropland, 50-74% cultivated.
20	General Cropland, 15-49% cultivated.
31	Ag-Urban, less than 15% cultivated, more than 20 dwellings per square mile, residential mixed with agriculture.
32	Residential/Commercial, no cultivation, more than 20 dwellings per square mile.
40	Range and Pasture, less than 15% cultivated.
50	Non-agricultural, variable size segments.
62	Water.

In many states, strata 11 and 12 are collapsed into one stratum. In the western states, a stratum may be added for Indian reservation land. The range and pasture stratum 40 is often broken into two strata in the western states, one for privately owned grazing land and one for public grazing land. The public grazing land is generally administered by the Bureau of Land Management or by the Forest Service. Other adjustments may be made to the design depending on the state involved. Crop-specific strata are also used in several states to allow the opportunity to channel a sample either into, or away from, a certain area. Some strata are defined to promote sampling efficiency and others for data collection convenience such as the public grazing stratum. The design is initially proposed by the Area Frame Section and is then sent to the State Statistical Office (SSO) for comments and suggestions.

STRATIFICATION MATERIALS

This section will describe the different types of materials which are available to perform the stratification process. Several examples of the more important materials are exhibited in this section.

Satellite Imagery: Satellite imagery is derived from digital data collected by scanners aboard the satellites. Presently, the imagery product from the LANDSAT satellite is used. A scanner mounted on the satellite collects the reflected and emitted energy from the ground. Two types of scanners are used: a multispectral scanner (MSS) and a thematic mapper (TM). The optics of the scanner separate this energy into bands - four for MSS and seven for TM. The spatial resolution is 60 meters for MSS and 30 meters for TM. The increased number of bands coupled with the much greater resolution makes TM the preferred product for stratification. TM, of course, is more costly, but keeping in mind that the frame produced using these materials will be in use for 15-20 years, the benefits far outweigh the cost.

Satellite imagery is currently not used for boundary identification. Instead, it is used when delineating agricultural strata since it provides a very recent picture of the degree of cultivation. Different land uses will show different color signatures on the imagery. The use of satellite imagery will be further discussed in the land-use stratification section of this chapter. Figure 1.2 presents a photo of a TM image.

National Aerial Photography Program (NAPP): NAPP is the product of a consortium of federal agencies, each of whom need and use aerial photography. NASS began using NAPP during the early 1980's to aid in the stratification process. Contact prints are used which are nine inches square and are scaled at 1:58000 (roughly one inch equals one mile). This scale is changing to 1:40000 in 1987. NAPP is a primary stratification tool. Over the time frame of 1980-86, approximately 95 percent of the U.S. had been photographed through the NAPP program. Figure 1.3 exhibits a NAPP photo. Note the resolution and identifiable ground features which can be seen in the photo.

Photo Index (PI): A PI is a mosaic made from individual frames of low level aerial photography which depicts the land area for a location. The PI is a photo of those individual frames arranged in their logical order. PI's were provided by the Agricultural Stabilization and Conservation Service (ASCS) in the past until the NAPP program was begun. PI's are no longer being produced. NAPP is replacing the PI's as NAPP is much more current. PI's are only used for stratification in areas where no NAPP exists.

Figure 1.2 Thematic Mapper Image

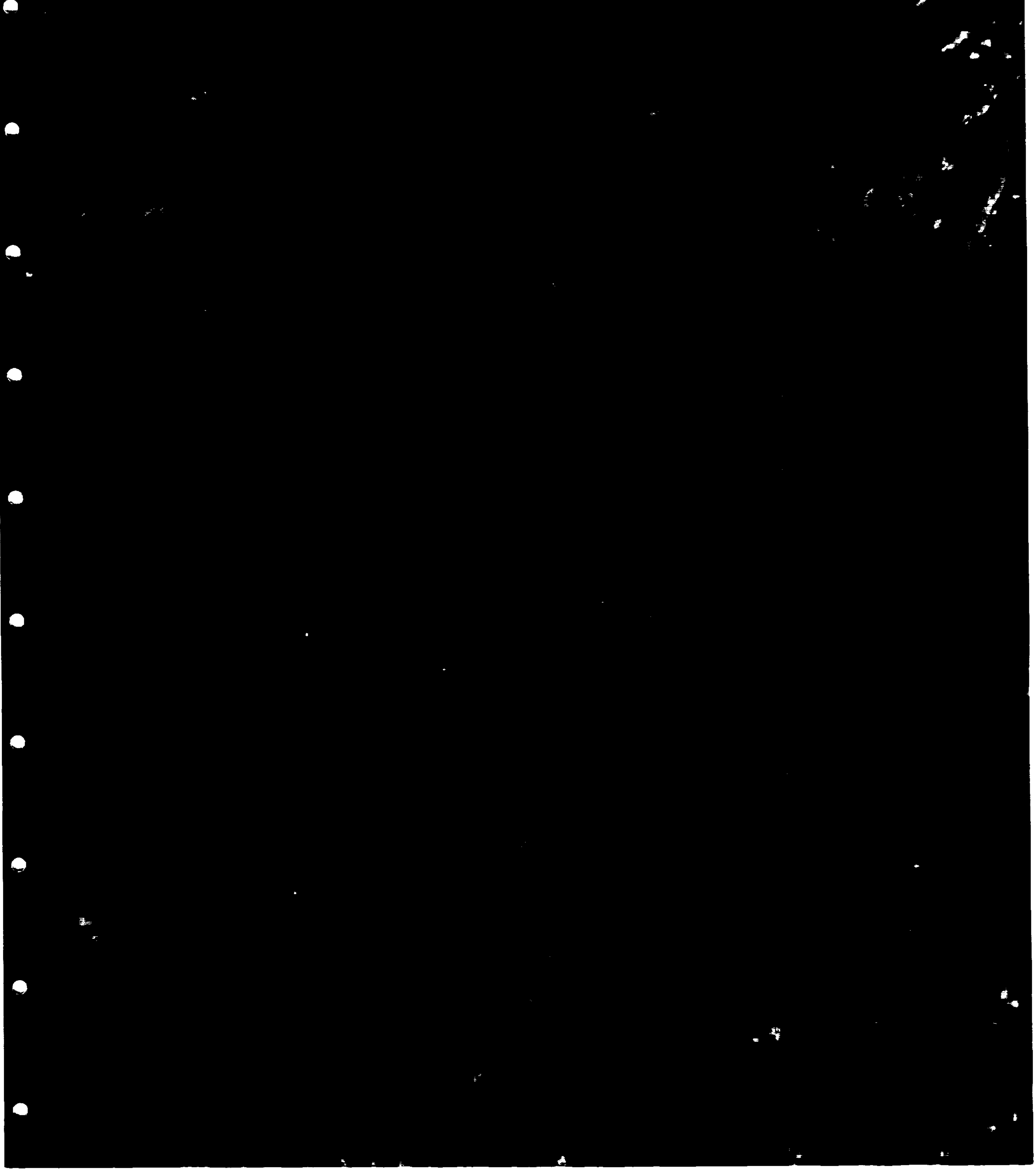


Figure 1.3 National Aerial Photography

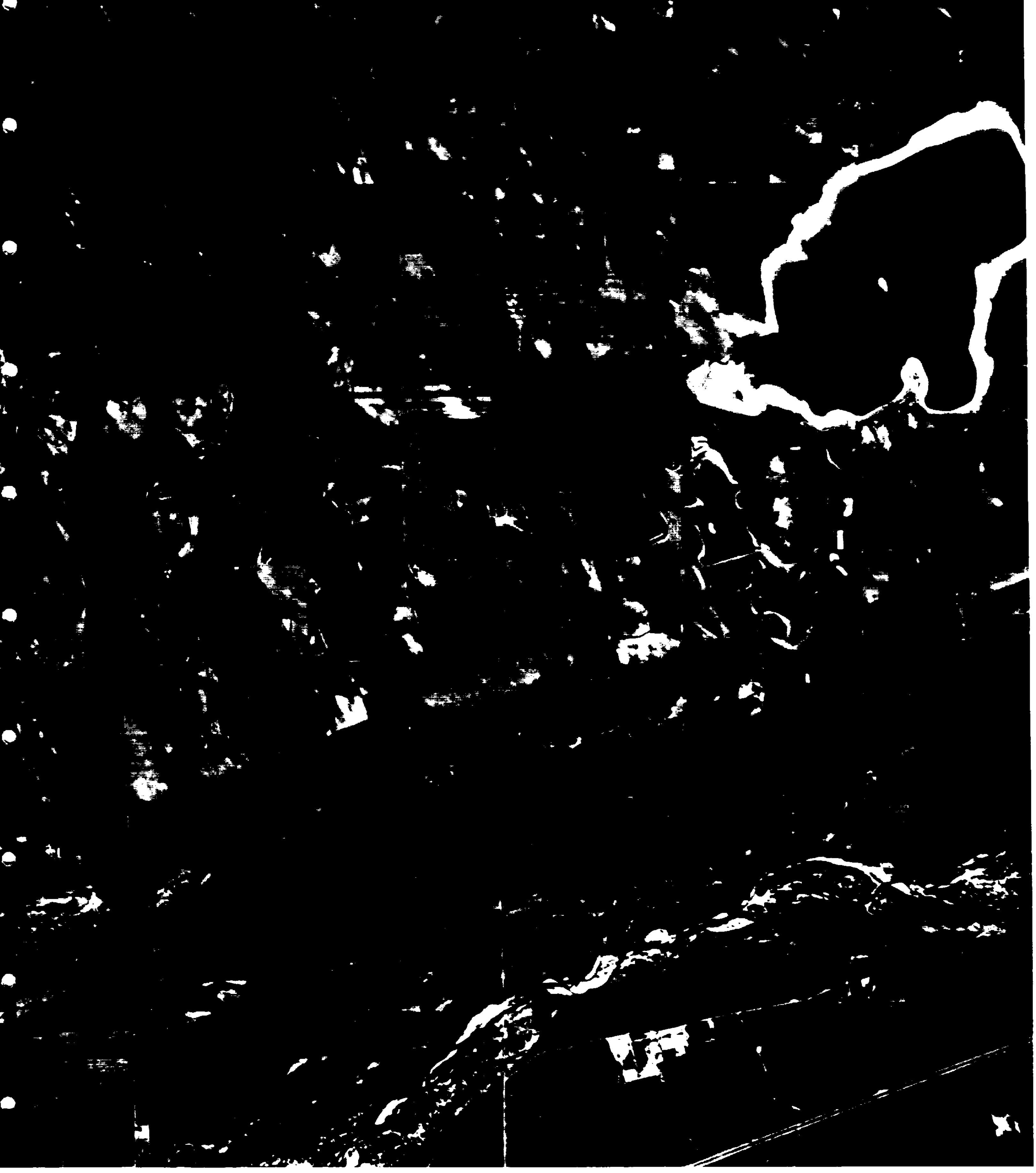


Figure 1.4 Topographic Quadrangle Map

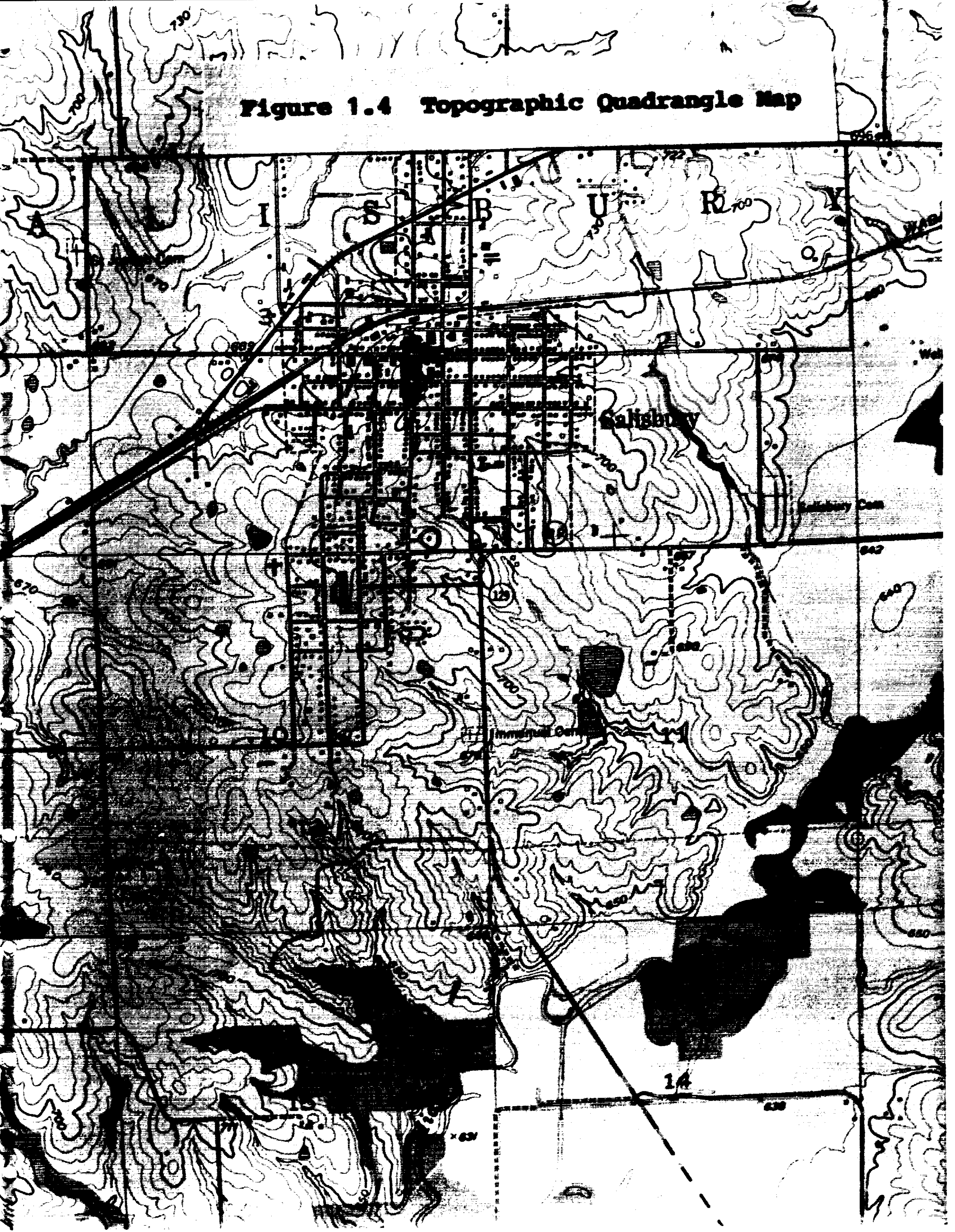


Figure 1.5 Bureau of Land Management Map

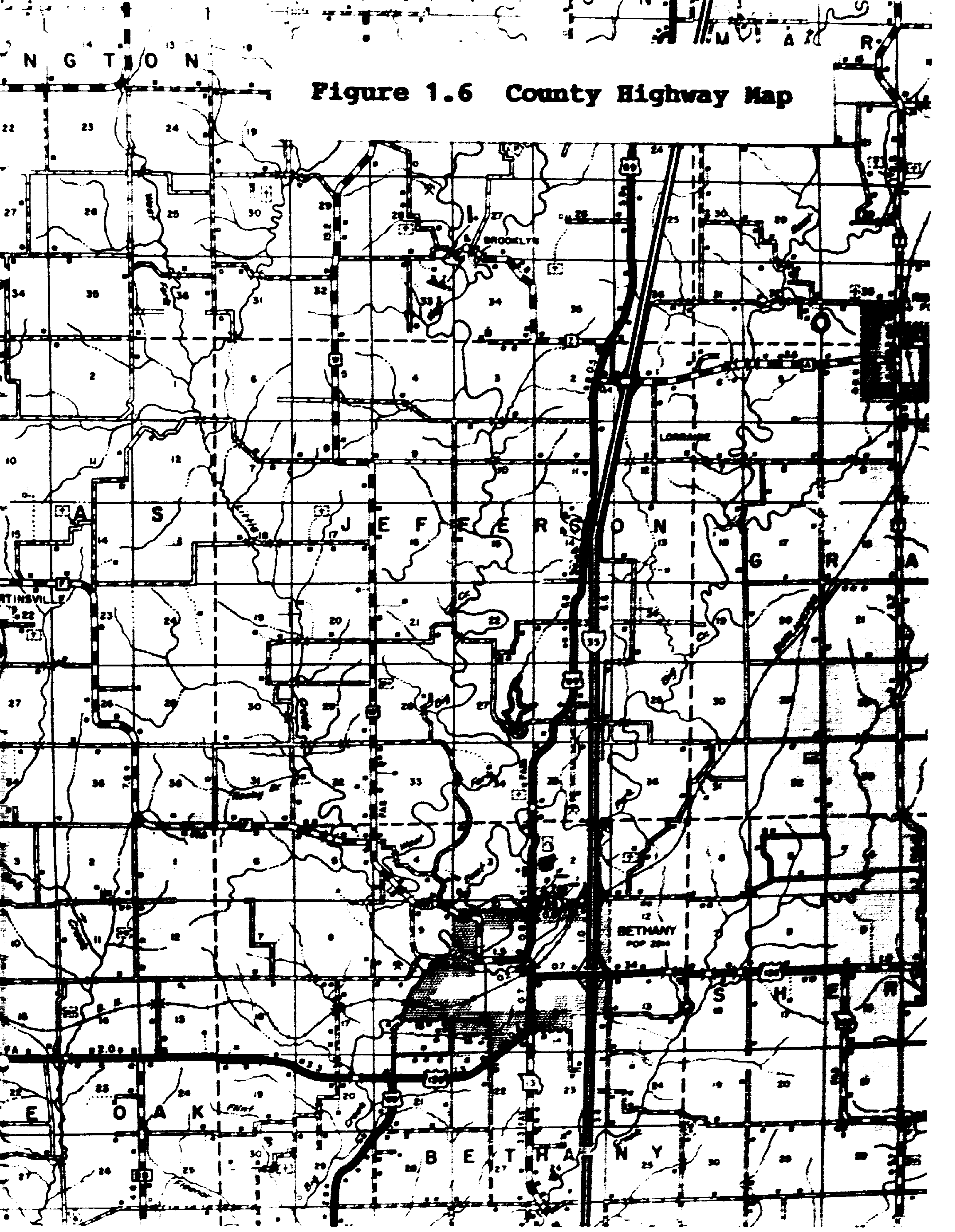


Figure 1.6 County Highway Map

Topographic Quadrangle Map (Quad): Produced by the U.S. Geological Survey (USGS), the preferred scale is 1:24000 (7.5 minute series - 2.6 inches to a mile) which makes them useful for urban and ag-urban stratification and sampling. NASS generally uses the 7.5 minute series and occasionally the 15 minute series (scale 1:62500) where the former is not available. Figure 1.4 displays a portion of a 7.5 minute quad.

Bureau of Land Management (BLM) Map: These maps, scaled at 1:100000, show the distribution of the federal and state land. They are useful in western states for delineating the range strata and for locating the boundaries of Indian reservations. Figure 1.5 exhibits a portion of a BLM map.

USGS 1:100000 Map: These maps are of high quality and have the potential to provide NASS with an accurate map base on which to stratify and digitize (to be defined later). These maps are not produced one sheet to a county and therefore present organizational difficulty for stratification and digitization.

County Highway Map: These maps are obtained from the state highway or transportation department. They have been used as the final map base for a majority of the frames since, in the past, they were almost universally available. Many of the county maps have the advantage of being black and white with no extraneous coloring. The extraneous color can hinder stratification and digitization. (A color code is used to separate strata in the stratification process). Some are published with color. Their main disadvantage is the degree of inaccuracy which pervades these maps. Problems include physical characteristics occasionally not drawn to scale, roads and railroads drawn in the wrong place and so forth. Figure 1.6 presents a small portion of a county highway map.

Landsat Overlay: A clear film positive image of a county highway map is obtained at the scale of the satellite imagery (usually 1:250000). The overlay provides points of reference to the map base when stratifying the agricultural areas. Also, the location and identification of the NAPP photography will be marked on the overlay.

LAND-USE STRATIFICATION

The process of land-use stratification is the delineation of land areas into land-use categories on photography and a corresponding map base. The purpose of stratification is to reduce the sampling variability by creating homogeneous groups of sampling units. Although certain parts of the process are highly subjective in nature, precision work is required of the personnel stratifying the land (called stratifiers) to ensure that overlaps and omissions of land area do not occur and land is correctly stratified.

Perhaps the most important concept conveyed during the initial training of personnel is the idea of using quality boundaries. A quality boundary is a permanent or, at least, long-lasting geographic feature which is easily found and identifiable by an interviewer. If an interviewer cannot accurately locate a segment in a timely manner, there is the potential for nonsampling errors to be introduced into the survey data. If the field interviewer, unknowingly, does not collect data associated with all of the land inside the sampled area or collects data for an area outside of that selected, then survey results will be biased.

The objective of using permanent boundaries and the objective of obtaining homogeneous sampling units within a stratum often conflict in the actual practice of area frame stratification. Concessions may have to be made in marginal situations. Given that the area frame is to be used over a period of 15-20 years and represents a major investment, the best and most permanent boundaries must be used. Roads and rivers make good strata boundaries, while intermittent streams and field edges do not and should rarely be used. The following is a list of geographic features which represent strata boundaries. The list is ranked by quality from highest to lowest:

1. Paved highways.
2. Secondary all-weather roads.
3. Local farm to market roads.
4. Railroads.
5. Permanent rivers and streams.
6. Permanent drainage and irrigation canals.
7. Intermittent streams and rivers or prominent water courses that carry water during and immediately following rains.
8. Field boundaries and visible section lines.
9. Trails and internal roads.

Most of the boundaries used in the U.S. are 1 through 5. Rarely are 6 through 9 used although in some foreign countries it would be necessary to do so.

Aerial photography makes it possible to stratify an area without fieldwork. Naturally, the more recent the photography, the better. Although NAPP photography may be one to five years old for an area, the best quality boundaries will still be there.

Once all of the materials are accumulated for a state, a review is made of each county highway map and a list of non-agricultural land parcels (stratum 50) is assembled. The list includes national and state parks, wildlife refuges, recreation areas, designated wilderness areas, military installations and airports. This list is sent to the state receiving the new frame to be checked for completeness and for consideration of several questions regarding each of the areas. These questions focus on whether livestock are allowed to graze on this land and whether any agriculture or cultivation is allowed. Although these areas are originally perceived to be non-agricultural, they will be placed in a more appropriate stratum if grazing is allowed or if any agricultural activity is permitted.

The stratification is performed a county at a time for administrative purposes. Each stratifier is assigned a county and will work on that county until its completion. Each county's materials will include a tracking sheet. The tracking sheet includes several items such as Landsat imagery identification, supplemental materials available, major crop items and their area, and population areas. Also included are county characteristics, notes on boundaries, unusual situations and a place to date and initial each phase of the stratification process. Attached to the tracking sheet is a county report published by the Bureau of the Census, U.S. Department of Commerce. The census report provides the stratifier with detailed agricultural data such as the number of farms, size of farms, acreage of various land uses and acreage of all crops grown in the county.

Stratification generally begins with determining the urban and ag-urban strata for the county. All cities and towns are located on the NAPP in a county. Using the quad maps for boundary identification and the stratification design, the urban areas are delineated on the photography with a grease pencil. The urban areas are drawn off using a color code scheme which was set up before the stratification work started. The ag-urban areas, which represent a mixture of residential and agricultural activity, are then located and drawn off accordingly. The ag-urban areas are usually located in a band around a city, where the city blends into the rural area. Ag-urban areas may also include small rural towns.

Stratification of the agricultural areas is the next step. Perhaps the most useful tool used in stratification is TM satellite imagery. The imagery is used primarily to ascertain where the cultivated areas and the pasture areas are present in a county. Different crops will correspondingly produce different color signatures, depending on the growth stage. For example, as the wheat crop emerges, a reddish appearance occurs and becomes a deeper red as growth continues. A ripe wheat field will have a greenish-yellow color, becoming greener just prior to harvest. At harvest, the field will appear white or tan. Most small grain crops will follow this general pattern. Crop calendars are provided to the stratifier showing approximate planting, growing and harvesting intervals for major crops. Other geographic features and their image colors or signatures are:

- <u>Water</u> appears blue to black. The cleaner and deeper the water, the darker the blue color. Lighter colors are usually the result of high turbidity, not due to temperature.

- <u>Cities and other urban areas</u> appear light blue-gray to bright blue. Concrete is usually white or very light blue. Asphalt is very dark blue or black.

- <u>Clouds</u> are white. The shadows should be black and have the same shape as the clouds next to them. Distances from the clouds to their respective shadows should be similar for all clouds in a given area.

- <u>Bare soil</u> - Soil colors vary dramatically and can be shades of green, gray, blue or brown. Moist soils are darker than dry soils. Bare rock is often brighter than more developed soils.

* Green, growing vegetation will appear pink to red on the imagery.

These color signatures are only general guides. The photo process involved in making a paper enlargement can create variations in the signatures. Therefore, no single color key will work for all scenes.

The satellite imagery is very useful in stratification because it is so timely. Although aerial photography may be one to five years old, the Landsat imagery usually covers the most recent growing season, providing a very recent look at the area. NASS also uses multitemporal imagery which is imagery for the same area for more than one time during the year. Imagery from early in the growing season and later in the growing season may be used to accurately distinguish crops from pasture and forest. Using the Landsat imagery for locating crops and pasture and the photography for boundaries, the stratifier must make subjective decisions on placing areas in their respective strata.

Working within the agricultural strata, the intensely cultivated areas are delineated first while recognizing the importance of quality boundaries, the targeted cultivation percentage, and a minimum size. The minimum size is the segment size. A listing of the two previous years' major crop data for the QAS in June are provided and located on the satellite imagery to give the stratifier a knowledge base of what crops are grown in that county, whether any irrigation is present and to help in interpreting the satellite signatures. Areas of water greater than one square mile are drawn off at this time. Then the less intensive areas are delineated (stratum 20). Range and pasture areas which are less than 15% cultivated are common in the western states (stratum 40). These areas are drawn off along with the non-agricultural areas (stratum 50) to complete the first phase of the land-use stratification.

Quality assurance is a major concern during the stratification phase. Throughout the process, checking and rechecking is performed to ensure a high quality product and to obtain the benefits of a second opinion by someone with a more experienced "eye". When the strata boundaries have been drawn on the photography, a check is then made of the urban and ag-urban strata boundaries. As the checker examines this stratification, he or she transfers those strata boundaries to the topographic quad map. The quad will be used during the sample selection process (described in chapter 2). Once the checker and the stratifier are satisfied with their efforts, the materials are examined by a more experienced person referred to as the reviewer.

The reviewer will not only check the urban and ag-urban stratification, but will proceed to check the rest of the work. When disagreements occur, the reviewer and the stratifier will meet and resolve any differences. Surrounding counties are also checked to match up the strata boundaries which may extend across county lines. Land trade-offs do not occur between states - all strata boundaries must end on the state line.

After the stratification on the photo mosaics has been approved by the reviewer, the stratifier will then transfer the strata boundaries to the county highway map base (also called the frame sheets or frame maps). The county highway map will later be digitized (electronically measured) to determine the areas of the primary sampling units (to be defined later). Accurately digitizing a mosaic of photos with edges and slightly varying scales would be very cumbersome - thus the reason for the transfer to the map base. After the boundaries have been outlined on the map base, the inside border is lightly shaded (approximately 1/8 inch) in order to make it stand out. Transferring the strata boundaries is aided with the use of a zoom transfer scope (ZTS). The ZTS is an optical instrument which can project an image of the photography onto the corresponding map base. The use of the scope produces a very accurate transfer of boundaries to the map base. Figure 1.7 presents a portion of a map base showing the delineated strata. Once this transfer is completed, the transfer is examined by the checker and the next phase of stratification is begun...construction of primary sampling units.

CONSTRUCTION OF PRIMARY SAMPLING UNITS

The next step in the development of the area frame is to further subdivide the strata into primary sampling units (PSU's). The desired size of the PSU varies by strata. As can be seen from the example in Table 1.2, the desired PSU size will vary depending on the strata, but contains, on the average, six to eight final sampling units or segments. The minimum PSU size is generally one segment.

Table 1.2 Primary Sampling Unit Size Tolerance Guide

LAND-USE STRATUM	MINIMUM SIZE	DESIRED SIZE	MAXIMUM SIZE
	<- - - - - - - -	- square miles -	- - - - - - ->
11	1	6-8	12
12	1	6-8	12
20	1	6-8	12
31	0.25	1-2	3
32	0.10	0.5-1	1
40	4	None	None
50	1	None	None
62	1	None	None

The use of primary sampling units introduces economic saving into area frame sampling. An entire frame need not be divided into segments in order to select a sample. Only the required number of PSU's will be randomly selected and further subdivided into segments - saving a tremendous amount in labor costs. In delineating PSU's, the main focus is not homogeneity of land-use - that will have already been accomplished with the land-use stratification. The main concern is to achieve the desired size while trying to maintain that each PSU is a smaller representation of the strata as a whole.

The choice of PSU boundaries is critical. PSU boundaries should be as permanent as strata boundaries. A part of the boundary of many segments will be a PSU boundary. Therefore, these boundaries must be recognizable on a map and identifiable on the ground by an interviewer. PSU's are drawn on the frame maps, not the photography. Once all the PSU's have been delineated on the map sheet for a county, the PSU identification number is attached. The PSU's are numbered in a serpentine manner beginning in the northeast corner of the county and ending in the southeast or southwest corner. The PSU identification consists of two parts, for example, 11-1. The first number in the sequence is the stratum and the second is an incremental counting number. Figure 1.8 displays the PSU numbering scheme for Banner county, Nebraska.

Upon the completion of the county, the reviewer will examine the PSU construction and identification tagging to ensure the work has been properly performed.

In the interest of quality assurance, several more reviews are made to catch any errors which may have entered into the process. The photography is again arranged as a mosaic and the stratification is checked for consistency between overlapping photos. The frame maps are reviewed by a statistician for completeness as a final check. The polygons created by drawing the PSU's are examined to make sure they form a closed polygon. The numbering system is checked for strata identification accuracy and sequential accuracy. The frame maps are further checked to ensure that omissions and overlaps do not exist. Once these checks have been accomplished, the frame maps are ready for the next step in the process...measuring the PSU's.

Figure 1.7 Stratification Shown on a Map Base

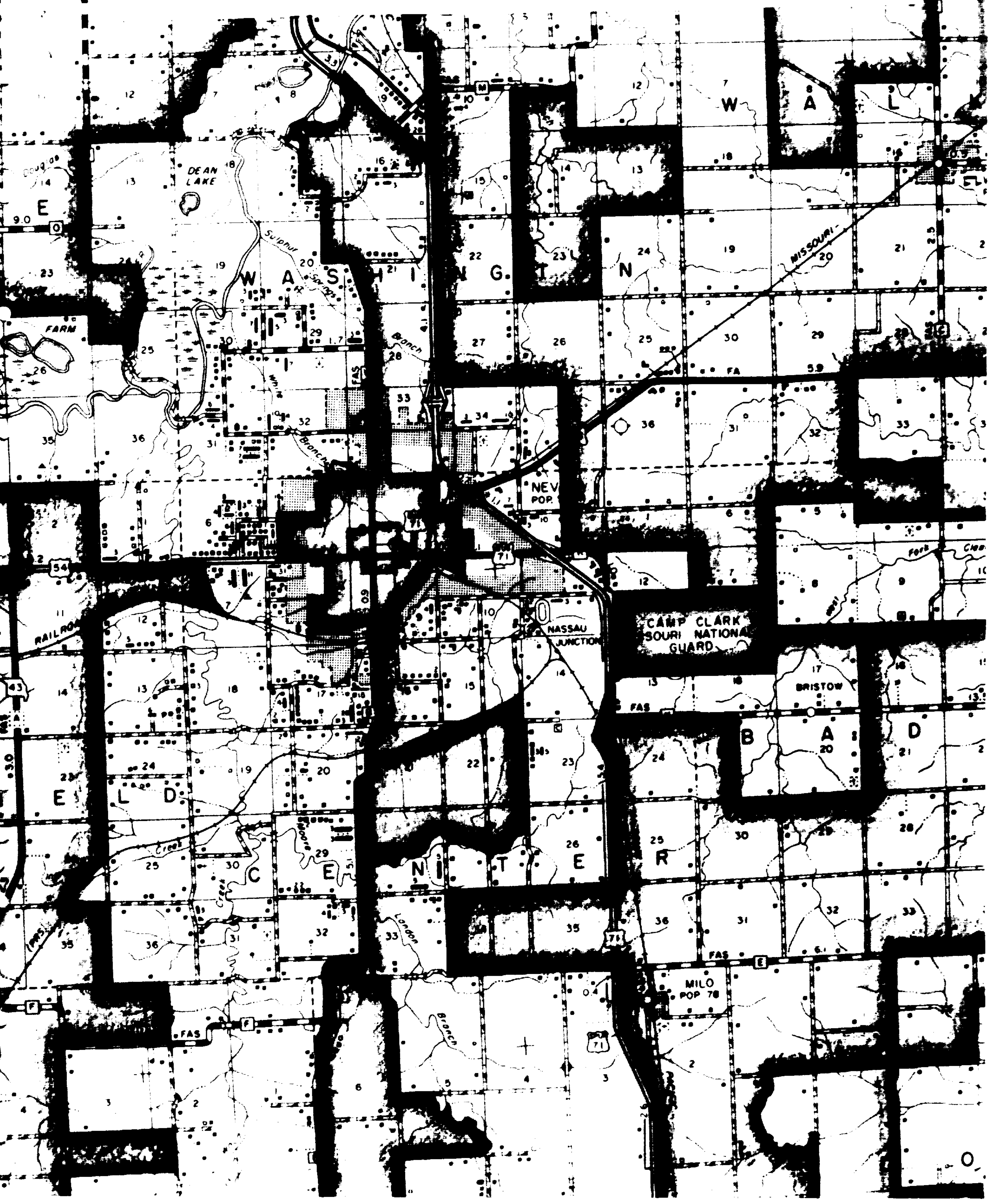
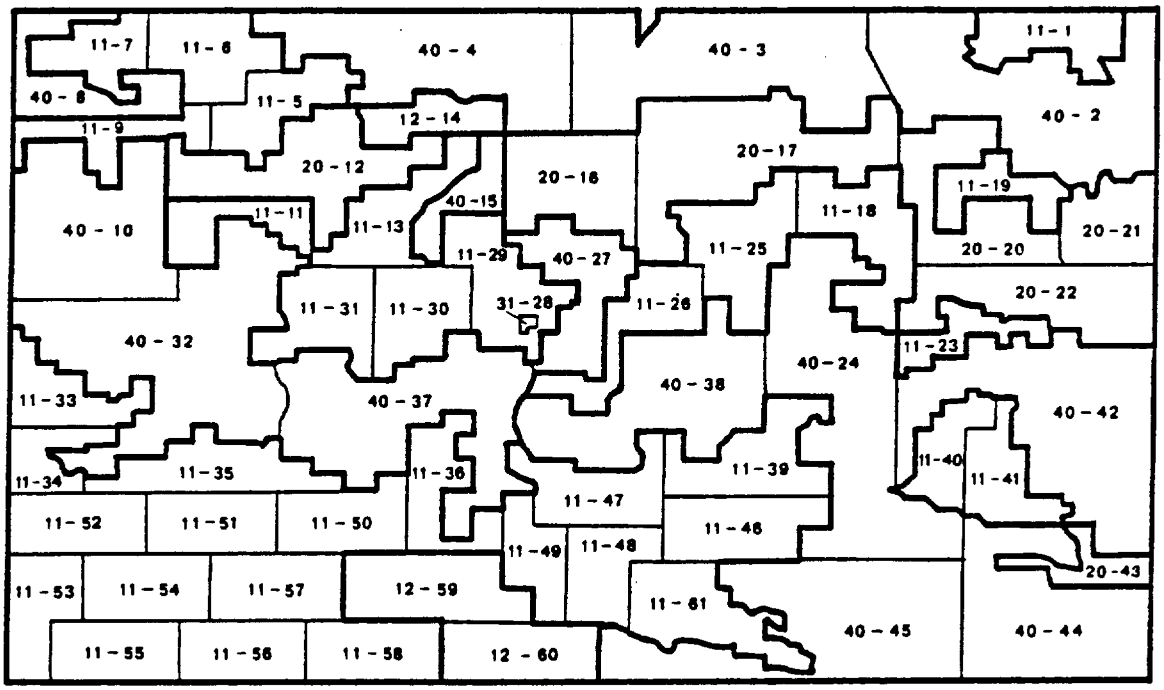


Figure 1.8 **Primary Sampling Units Numbered in a Serpentine Manner for Banner County, Nebraska**



DIGITIZATION

The conversion of map points into two-dimensional X-Y coordinates is called digitization. Digitization involves electronically measuring the area of the PSU's on the frame maps. The PSU's need to be measured to determine the number of segments per PSU for sampling purposes. Electronically recording the PSU areas allows:

- measuring the PSU accurately,
- quality assurance,
- retaining a digital backup copy of the frame map in the unlikely event that a frame map is lost.

NASS utilizes analog to digital conversion tablets (4' x 5' digitizing tables) to establish a coordinate system overlaying the frame. A reference point, known as the origin (0,0), is established for X-Y coordinates on the map. X-Y coordinates, tagged with the appropriate identification, uniquely describe the borders of a PSU and therefore create a polygon for each PSU. The digitizing software records the X-Y coordinates in a file. Using the map scale, the area of each polygon (PSU) is calculated in terms of square miles and stored in a separate file.

After each frame map has been digitized, the X-Y coordinate file is processed by an error-check program which looks for errant points and digitization mistakes. The files are stored on a multi-user minicomputer and also off-site on magnetic tape. Using the X-Y point file for input, a graphic representation of the frame map is generated. This graphic image can be sent to a plotter for hard copy output or to a graphics display terminal. Both are used for visual checking. A plotted frame can be overlaid with the original frame map on a light table and checked for inconsistencies and digitizing errors. An operator using a graphics display terminal can zoom in and check a small detailed area for any digitization errors. The person who digitized the frame map and the reviewer work together to eliminate all the discrepancies. Figure 1.9 exhibits a frame map displayed on a terminal. The colors correspond to various strata.

After frame maps for a state have been digitized and reviewed, area files for the individual frame maps are concatenated into a single PSU area file. A computer program is then executed using the PSU area file as input to detect several other types of errors which may not be detected by visual inspection of the frame or a plot. Some of these errors include invalid stratum codes, invalid county code, PSU numbering sequence errors and so on.

The PSU areas for each county are summed and compared against the official county size. The same procedure is done for the state area. County areas are allowed to vary ±3.0 percent from the published area. The accumulated state area is only allowed to vary ±0.5 percent from the published area. The county area is allowed more variance because of the smaller area involved and because primary sampling units are allowed to cross county boundaries. Since the stratification is never allowed to cross state boundary lines, only a small amount of error is allowed.

The PSU areas are then accumulated for each stratum. The area of the PSU divided by the target segment size for the stratum is equal to the total number of ultimate sampling units (segments) in that PSU. Summing the number of segments will yield the total number of segments in the stratum. Table 1.3 shows the results of that calculation for each stratum in Banner County, Nebraska and at the state level in Nebraska.

[illegible] Map on a Graphics Display Terminal

Table 1.3 **Digitized Land Area and the Number of Segments for Banner County, Nebraska and the Entire State Level. Area is in Square Miles.**

LAND-USE STRATUM	TARGET SEGMENT SIZE	BANNER COUNTY		NEBRASKA	
		DIGITIZED AREA	# OF SEGMENTS	DIGITIZED AREA	# OF SEGMENTS
11	1	290.3	289	30112.4	30202
12	1	27.3	27	8754.6	8794
20	2	80.5	40	9531.4	4785
31	0.25	0.2	1	649.1	2610
32	0.10	0.0	0	167.7	1677
40	4	346.8	88	27695.2	6915
50	1	0.0	0	179.5	184
62	1	0.0	0	132.4	133

The population of sampling units will be utilized in the sample allocation and selection processes described in detail in the second chapter of this report.

SUBSTRATIFICATION

There is a further level of stratification which is applied to the frame. This is not a process which is performed by the personnel in the stratification or digitization units, but by a statistician using cluster analysis to perform a grouping analysis. The substratification involves ordering the population of sampling units using a criteria of agricultural similarity.

The land-use stratification is based on the percent of cultivation. Therefore, while the majority of the segments within a stratum may be intensely cultivated, the agricultural makeup of the segments may differ depending on the location of the segments within the state. Ordering the population of primary sampling units according to agricultural content will yield greater precision in the estimates for individual commodities. Substratification is particularly effective in areas of intensive cultivation where cropland content varies across the state. Utilizing substrata in grazing or range strata contributes very little to reducing variance except maybe for cattle. Therefore, more substrata are used in the intensely cultivated strata as compared to the range or lightly-cultivated strata.

Ordering the population is a two-stage process. First, the primary sampling units are ordered or numbered in a serpentine manner within the county (refer back to Figure 1.8). The second stage involves the ordering of the counties. This county ordering is based on a multivariate cluster analysis of county level crop and livestock data. The purpose of cluster analysis is to group counties into clusters or groups which generally have the same overall agricultural makeup.

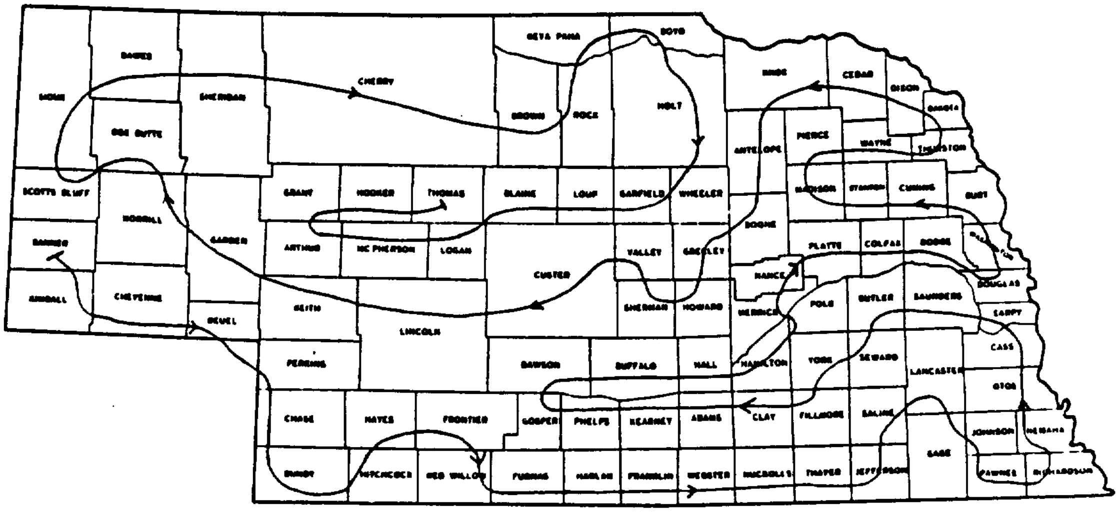
Figure 1.10 exhibits the county ordering used in the Nebraska area sampling frame. Note that in all but one instance, the ordering proceeds from one county into an adjacent county. The reason for the exception in the southwest corner of the state is that Perkins county is more similar to Deuel county, and Keith county is similar to Lincoln and Garden counties. The county ordering need not be continuous. If the counties in one corner of the state were very similar to those in another corner, the ordering could skip across several counties. The starting point of the ordering is somewhat arbitrary, so a logical starting point would be any corner of the state. However, if the cluster analysis indicates a clear distinction between two groups of counties, it may be advantageous to start in one area and end in the other. In Nebraska, the county ordering begins in an area where some wheat is grown, and proceeds through other counties in which there is a high density of cropland. The ordering proceeds into the area of the state that is primarily rangeland and pasture.

When the ordering "enters" a county from the west or the south, the order of the primary sampling units in the county is reversed. Recall that the PSU's within a county are ordered by arbitrarily starting in the northeast corner of the county. Therefore, reversing the order will ensure a fairly continuous ordering of PSU's from one county to the next.

After the population of segments has been determined for each stratum and the two-stage ordering has been accomplished, the number of substrata for each land-use stratum is established as will be discussed shortly. Several factors are considered in the determination, including experience with sampling frames in other states, the number of sample segments and replicates within each stratum and the degree of homogeneity among the sampling units within the various strata.

NASS employs a concept called replicated sampling which provides several key benefits in the estimation process which are described in the second chapter of this report. About twenty percent of the replicates are rotated out of the sample each year with new replicates taking their place. To facilitate this yearly rotation, quite often multiples of five replicates are chosen. As can be seen in Table 1.4, dividing the sample size by the number of replicates, equals the number of substrata.

Figure 1.10 County Ordering Used for the Nebraska Area Frame



Each substratum will contain the same number of sampling units, except the last, which may contain slightly more or less than the others due to rounding. The 30202 sampling units in stratum 11 are divided into 15 substrata. Fourteen of the substrata will contain 2013 units and the last will contain 2020.

Table 1.4 Nebraska Sample Design Showing the Number of Substrata and Replications

LAND-USE STRATUM	POPULATION NUMBER OF SAMPLING UNITS	SAMPLE SIZE	NUMBER OF SUBSTRATA	NUMBER OF REPLICATIONS
>75% cult.	30,202	225	15	15
50-74% cult.	8,794	70	7	10
15-49% cult.	4,785	35	7	5
Ag-Urban	2,610	10	2	5
Residential	1,677	5	1	5
Range	6,915	40	4	10
Non-ag	184	5	1	5

Table 1.5 exhibits the number of substrata in use for a variety of states. The approximate average number of substrata used within each of the general land-use strata is also given. As can been seen in Table 1.5, there are more substrata in the intensely cultivated land-use strata. As a rule of thumb, five replicates are generally used if the sample size for a stratum is less than 50 segments, five or ten replicates if the sample size for a stratum is between 50 and 100 segments. Ten replicates are used if the sample size is between 100 and 200 segments. The number of substrata is therefore simply the sample size divided by the number of replicates.

Table 1.5 Number of Substrata in Use by Stratum for a Variety of States*

STATE	STRATUM GROUP						
	CROP SPECIFIC	INTENSIVE CULT.	LESS INT. CULT.	AG URBAN	RESI-DENTIAL	RANGE	NON-AG
Alabama		9	15	4	2	15	1
Arizona		4,2	4,1	4	3	4,2	2
California	24,10	24	12	8	2	10,5,5,2	2
Colorado		8,2	4,1,2	2	2	3	2
Florida	3,1	5	5,3	8	4	6,4	4
Illinois		17,10	8	4	2	2	–
Indiana		19,11	6	5	1	3	1
Kansas		17,12	10	3	3	3	1
Michigan		11,7	4	7	2	7	1
Minnesota	2,1	12,7	5	4	1	6	1
New Mexico		10,8	6	3	1	3	1
New York		13	12	4	2	2,6	1
Ohio		14,11	6	6	2	5	1
Oklahoma		9,12	13	4	2	8,7	1
Penn.		12	12	7	2	8	1
Texas	11,7,2,9	15	18	6	7	9,3	1
Wyoming		4,4	4	3	1	5	1
Average	7	10	7	5	2	5	1

* Multiple entries denote multiple strata within a stratum group. For example, there are four crop specific strata in Texas. Those four strata utilize 11,7,2 and 9 substrata respectively.

COST

This section will discuss the estimated cost to design and build an area frame. The cost data are naturally divided into two main areas, labor and materials, plus a miscellaneous category. The miscellaneous area includes costs incurred by the State Statistical Office, travel and other small items. The cost breakdown for developing a frame for a state is shown in Table 1.6.

Table 1.6 Breakdown of Costs Incurred in Developing an Area Frame for a State

COMPONENT	DESCRIPTION	COST
Labor	Cartographic	
	- Stratification (approx. 8000 hours)	$56000
	- Digitization (approx. 3000 hours)	24000
	Statisticians	24000
	Systems Analyst	4000
	Administrative and Secretarial	5000
Materials	Satellite imagery	15000
	Aerial photography	10000
	Transfer maps	1000
	Scaled overlays	1000
	Data processing expenses	2000
Misc.	SSO support and travel	2000
	Computer and equipment maintenance	6000
Total		$150000

As can be seen from the table, labor costs are the most prominent, accounting for over seventy-five percent of the cost of developing an area frame for a typical state. These costs do not reflect the cost of sample selection and sample preparation which are discussed in the second chapter of this report.

The estimated time and cost figures to build an area frame can vary widely from one state to the next. Factors to be considered include the size of the state, the number of counties, the availability of good boundaries, the type of agricultural activity involved and the type of materials to be utilized. Total cost may vary from $50,000 to $250,000.

FUTURE PROSPECTS

The stratification process has been, and will continue to be, a very labor intensive, manual process. The Area Frame Section currently employs the equivalent of approximately 10 full time people to carry out this effort. It is also a very time-consuming process thus new frames are built in only 2-3 states per year. The efficiency of this approach as opposed to an updating process is questionable. An updating process would utilize a change detection system to determine which primary sampling units have changed their agricultural character. Updating only those PSU's which need to be changed to a different stratum would be more efficient than starting over from scratch each time.

A jointly funded three year research project between NASS and National Aeronautics and Space Administration (NASA) has begun to explore the use of digital data in the development of area frames. This system will allow the person (stratifier) doing the land use stratification to delineate PSUs on-screen, eliminating the transfer to the map base and the digitization process. This system would allow the use of previously-classified satellite data as an aid in the stratification process. The system or project is named CAS for Computer Aided Stratification.

The key ingredient to our system is the availability of digital map data. USGS is producing digital data representing the hydrography and transportation information seen on the 1:100,000 topographical map series. This product is called Digital Line Graph (DLG). The use of DLG data allows us to use computer graphics procedures to combine ground observable features with the remotely sensed data.

The other major input to the CAS system is digital Landsat thematic mapper satellite data. We are using current image processing techniques to combine the DLG data with the satellite data in order to generate (on a image display device) an aerial photograph like image with map attributes. The stratifier uses this generated image to interactively delineate PSU boundaries, thereby replacing the current use of Landsat paper products, high altitude photography, county highway maps, photo index sheets, and other data sources currently used.

BIBLIOGRAPHY

Ciancio, N.J., Rockwell, D.A., and Tortora, R.D. (1977): An Empirical Study of Area Frame Stratification. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Fecso, R. and Johnson, V. (1981): The New California Area Frame: A Statistical Study. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 22. Washington, D.C.

Fecso, R., Tortora, R.D., and Vogel, F.A. (1986): Sampling Frames for Agriculture in the United States. Journal of Official Statistics. Vol. 2, No. 3. Statistics Sweden.

Geuder, J. (1983): An Evaluation of SRS Area Sampling Frame Designs: Ordering Count Units and Creating Paper Strata. U.S. Department of Agriculture, Statistical Reporting Service Staff Report No. AGES830715. Washington, D.C.

Geuder, J. (1984): Paper Stratification in SRS Area Sampling Frames. U.S. Department of Agriculture. Statistical Reporting Service, SF&SRB Staff Report No.79. Washington, D.C.

Hanuschak, G. and Morrissey, K. (1977): Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Houseman, E. (1975): Area Frame Sampling in Agriculture. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 20. Washington, D.C.

Huddleston, H. (1976): A Training Course in Sampling Concepts for Agricultural Surveys. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 21. Washington, D.C.

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

Pratt, W. (1974): The Use of Interpenetrating Sampling in Area Frames. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

U.S. Department of Agriculture (1983): Scope and Methods of the Statistical Reporting Service. Publication No. 1308. Washington, D.C.

U.S. Department of Agriculture (1984): Area Frame Analysis Package. Statistical Reporting Service Staff Report. Washington, D.C.

U.S. Department of Agriculture (1987): International Training: Area Frame Development and Sampling. National Agricultural Statistics Service. Washington, D.C.

U.S. Department of Agriculture (1987): Supervising and Editing Manual, June Enumerative and Agricultural Surveys. National Agricultural Statistics Service. Washington, D.C.

Wigton, W. and Bormann, P. (1978): A Guide to Area Sampling Frame Construction Utilizing Satellite Imagery. U.S. Department of Agriculture. Statistical Reporting Service Staff Report. Washington, D.C.

CHAPTER II. AREA FRAME SAMPLING

BY

JACK NEALON

INTRODUCTION

The purpose of this chapter on area frame sampling is to provide an in-depth presentation of NASS's area frame sampling concepts and procedures. The chapter will begin by discussing NASS's use of replicated sampling. Next, the procedures used to determine the size of the ultimate sampling unit (segment) and the allocation of the sample across and within states are described. The probability models are then detailed for the various methods of selecting the sample segments. An overview of the office procedures applied to select and prepare the sample segments for agricultural surveys are also discussed. In addition, the sample rotation scheme and the costs associated with the sampling process are described. Finally, the estimation approaches used to estimate agricultural production from area frame surveys will be discussed briefly.

REPLICATED SAMPLING

Area frames developed since 1974 have been sampled using a replicated design. Replicated sampling is characterized by the selection of a number of independent subsamples or replicates from the same population using the same selection procedure for each replicate. Each replicate is therefore an unbiased representation of the population.

A replicate for NASS's area frame sample design is a random sample of land areas (segments) selected within a land-use stratum. The substratification within each land-use stratum, which was described in the first chapter, has been incorporated into the sampling process to improve the sampling efficiency and the sample dispersion. Therefore, a replicate is more specifically defined as a simple random sample of <u>one</u> segment from each substratum in a land-use stratum.

The first segment randomly selected in each substratum in a land-use stratum is designated as replicate 1, the second segment selected from each substratum is designated as replicate 2, and so forth. The number of replicates is the same for each substratum in a given land-use stratum. Therefore, the number of sample segments in a land-use stratum is simply the product of the number of replicates and the number of substrata in the land-use stratum. That is,

$$n_i = r_i s_i$$

where n_i = the number of segments in the sample for the i^{th} land-use stratum,

r_i = the number of replicates for each substratum in the i^{th} land-use stratum,

s_i = the number of substrata in the i^{th} land-use stratum.

Suppose, for example, we want to select a replicated sample of two replicates from a land-use stratum consisting of three substrata with ten segments in each substratum. Then the total sample size for the land-use stratum would be $n_i = r_i s_i = 2x3 = 6$ segments, as illustrated in Table 2.1. Notice that a simple random sample of one segment is selected in each substratum for a replicate so that the number of sample segments in a replicate is simply the number of substrata.

The number of replicates certainly does not have to be the same in each substratum. Sometimes it may be advantageous to vary the number of replicates in the substrata for a land-use stratum. For example, if a crop is localized to a few counties in a state and greater precision is desired for data pertaining to this crop, then the sampling variance could be reduced for this crop by increasing the number of replicates in the substrata corresponding to these counties.

There are six reasons why NASS uses replicated sampling. These reasons will now be discussed.

(1) <u>SAMPLE ROTATION</u>: A sample rotation scheme is used to reduce respondent burden caused by repeated interviewing, avoid the expense of selecting a completely new area sample each year, and provide reliable measures of change in the production of agricultural commodities from year to year through the use of the ratio estimator. Sample rotation is accomplished each year by replacing segments from specified replicates in each land-use stratum with newly selected segments. Approximately twenty percent of the replicates in each land-use stratum are replaced annually. The sample design does not rotate exactly twenty percent of the segments because the number of replicates is not always a multiple of five.

Table 2.1 Replicated Sampling Process for a Land-Use Stratum

SUBSTRATUM	SEGMENT	REPLICATE 1	REPLICATE 2
1	1		
	2		
	3		X
	4		
	5		
	6	X	
	7		
	8		
	9		
	10		
2	11		
	12		
	13		
	14		
	15		
	16		
	17		X
	18	X	
	19		
	20		
3	21		
	22		
	23	X	
	24		
	25		
	26		
	27		X
	28		
	29		
	30		

To illustrate how replicated sampling simplifies the sample rotation process, Table 2.2 shows the numbering scheme for a hypothetical land-use stratum with five replicates in each of eight substrata. The first digit in the four-digit segment number represents the year the segment rotated into the sample, e.g. 3001 entered in 1983. The remaining three digits are simply unique numbers.

The sample rotation in 1988 will be performed by replacing the segments in the 3000 series (replicate 1), which have been in the sample for five years, with segments numbered 8001, 8002,...,8008. In 1989, the segments will be replaced from replicate 2 since the 4000 series would have completed its five-year sample cycle. In 1990, the segments from replicate 3 will be replaced and so forth.

Table 2.2 Segment Numbering Scheme for Replicates in a Land-Use Stratum

SUBSTRATUM	REPLICATE				
	1	2	3	4	5
1	3001	4009	5017	6025	7033
2	3002	4010	5018	6026	7034
3	3003	4011	5019	6027	7035
4	3004	4012	5020	6028	7036
5	3005	4013	5021	6029	7037
6	3006	4014	5022	6030	7038
7	3007	4015	5023	6031	7039
8	3008	4016	5024	6032	7040

(2) METHODOLOGY RESEARCH: Replicated sampling provides the capability to test alternative survey procedures or evaluate current methodology since different replicates can be assigned to the research and operational methods. For example, if there are a total of ten replicates in a land-use stratum and there is a need to compare two approaches to asking a particular question, then five replicates could be assigned to each method. The test statistic could then be easily derived using the means or totals from each replicate for each approach. Some examples of survey procedures that might be tested are different questionnaire designs and alternative interviewing approaches.

(3) QUALITY ASSURANCE: Replication also facilitates quality assurance analysis by allowing data comparisons among years in order to determine if significant differences in survey processes exist over time. For example, segment sizes can readily be compared among replicates to determine if the average size and the variability in size differ significantly from year to year. If so, this may indicate that the manual procedures for delineating segments (to be discussed later) need to be reviewed.

(4) SAMPLE MANAGEMENT: Replication allows easy management of the sample due to the replicate numbering scheme. This simplifies the process of designating a subsample of segments for one-time or repetitive surveys, increasing or decreasing the sample size in a land-use stratum to improve sampling efficiency, and identifying segments to be rotated out of the area frame sample.

(5) VARIANCE ESTIMATION: Replicated sampling provides a simple, unbiased method for estimating the sampling variance using replicate means or totals. NASS estimates the sampling variance for agricultural surveys using the sub-stratification design rather than replicate totals. However, replicate totals are sometimes used for variance and covariance estimation to simplify multi-variate statistical analysis in research studies. The benefit of using replicate totals to estimate the sampling variance is most pronounced in under-developed countries where a computer facility or the necessary statistical software is not available.

(6) ROTATION EFFECTS: Replication readily provides NASS the vehicle for evaluating sample rotation effects. Rotation effects are defined as the impact to survey data resulting from the number of years a segment has been in the sample. NASS has a five-year rotation process which permits replicate totals to be compared for segments in the sample from one to five years.

SEGMENT SIZE

What size should the ultimate or final sampling unit (segment) be for each land-use stratum? By definition, the optimum size of the segment has traditionally been the size that provides the most precision in the survey estimates for a given cost. The optimum size depends upon a multitude of often interrelated factors such as the survey objectives, estimation method, data collection costs, data variability among segments, interview length, population density, concentration of cropland, the reporting unit, and the availability of identifiable boundaries for the segments.

The optimum size for segments is very difficult to determine in practice, especially for multipurpose and multitemporal surveys like the QAS where information for a wide variety of agricultural characteristics is obtained. This difficulty is compounded by the fact that NASS uses multiple estimation methods not only within the same survey but often for the same survey item. For example, hog inventories are estimated in the major producing states using a multiple frame estimator and three area frame estimators (open, closed and weighted segment estimators), each of which may have a different optimum value for the size of the segment. Finally, the surveys are not designed to provide the data necessary to conduct analysis of sampling variances for different segment sizes. Therefore, it is not possible to determine the optimum segment size for each land-use stratum based on sampling variances.

The discussion so far has viewed the optimization in terms of sampling error rather than the total survey error. Nonsampling errors also need to be considered when determining the target size of the segments. Two types of nonsampling errors that are of concern are:

(1) As the size of the segments decreases, the availability of suitable boundaries for the segments also decreases. This decline in good boundaries results in more reporting errors during the data collection phase.

(2) As the size of the segments decreases, so does the ability to delineate segments manually on the aerial photography (to be discussed later) that are homogeneous with respect to the amount of cultivated land. Therefore, the sampling variability among segments increases for a given sample size.

Little quantitative information is presently used to determine the target segment size for each land-use stratum in a state. Two sources of information are relied on when deciding on the segment sizes. First, survey data from the two most recent years are analyzed in terms of the number of reporting units per segment and the data distributions for major survey items. Second, the latest aerial photography is examined to determine the adequacy of boundaries throughout the state. Basically, there are two rules of thumb that guide the decisions on what the target segment sizes will be. These are:

(1) Use the smallest size that is practical in terms of providing easily identifiable boundaries for delineating segments. For example, it is very practical to use one square mile segments in the agricultural strata for states where much of the rural land is divided by roads into one square mile sections. These boundaries are readily identifiable on maps, photos and during the interview.

(2) Select a size for which data collection can be completed within one day under normal conditions. This rule of thumb is designed to control the within-segment data collection costs. An average segment in the agricultural strata in the U.S. contains about one resident farm operation and less than three farm operations.

With this in mind, NASS has historically used larger segment sizes than are considered optimal from the point of view of sampling errors to ensure clearly identifiable boundaries for the segments and thereby reduce nonsampling errors. The segment size in the intensively cultivated strata (at least 50 percent cultivated) is usually one square mile except in southern states where one-half square mile segments dominate. In the less intensively cultivated strata (15-49 percent cultivated), the target segment size is usually three-fourths square mile in the south, one square mile in the northeast, and one or two square miles in the central and western states. The target segment size for range strata is generally one square mile in the south, four square miles in central states, and varies in the northeast (one, two or four square miles depending on the state). There is generally no target segment size in western range strata where segments are allowed to vary greatly in size (from two to several hundred square miles). The target size in the non-agricultural strata has traditionally been one square mile except in western states where two square miles is often used. Nowadays, no target size is used in this stratum when new frames are constructed. The target segment sizes in the urban and ag-urban strata are always one-tenth and one-quarter square mile, respectively.

The reporting unit for the closed and weighted segment estimators is an agricultural operation and for the open segment estimator is a resident agricultural operation. (This will be discussed later). Table 2.3 provides summary information on the average number of farm operations and resident farm operations in the 48 contiguous states based on the current target segment sizes in the area frames. (A farm operation is an agricultural operation with at least $1,000 in gross sales). Notice that the averages are less than three and about one reporting unit per segment, respectively.

Table 2.3 Average Number of Farm Operations and Resident Farm Operations per Segment; 1986

REGION	NUMBER OF SEGMENTS	AVERAGE NUMBER OF	
		FARM OPERATIONS PER SEGMENT	RESIDENT FARM OPERATIONS PER SEGMENT
West	4,175	2.2	0.9
South	5,928	2.5	0.9
Northeast	1,641	2.7	1.2
Central	3,921	3.5	1.0
48 States	15,665	2.7	1.0

SAMPLE ALLOCATION

The area frame sample is used to collect data on a wide range of agricultural items such as crop acreages, livestock inventories and economic data. Therefore, the allocation of the sample across states and within states to the land-use strata is extremely important. NASS evaluates optimum allocations of the sample to obtain the most precision in the major survey estimates for a given budget. The number of sample segments allocated to each land-use stratum and state depends on factors such as the average data collection cost per segment in each stratum, the variability of the data in each stratum resulting from the intensity and diversity of agriculture, the total number of segments or land area in each stratum, and the importance of the state's agriculture relative to the national agricultural statistics program.

An optimum sample allocation to the land-use strata is generated for each of the most important agricultural survey items (univariate) and for all the important commodities considered simultaneously (multivariate). The allocations are evaluated not only from an area frame perspective but also from a multiple frame point of view where the area frame is used to measure the incompleteness in the list frame. Finally, optimum allocations are conducted at the national, regional and state levels to assess the allocations at the various inference levels.

NASS places the most importance on the multivariate optimum allocation for the area frame nonoverlap estimates at the state level since it is important to provide useful statistics at the state level. Adjustments are made to this sample allocation to improve the precision of the regional and national estimates without seriously hindering the precision levels for the states. Minor adjustments to the optimum allocation are also made to provide a multiple of five replicates in each stratum to simplify the sample rotation process and to protect against the impact of outliers by not allowing the sampling rate to be too small in a stratum, e.g. 1 in 750 segments.

The optimum allocation of a sample for multipurpose surveys can be viewed as a problem in convex programming. An iterative, nonlinear programming algorithm is used to provide the univariate and multivariate optimum sample allocations for the area frames. The algorithm is guaranteed to converge to the optimum solution.

A brief description of the multivariate sample allocation model follows. Suppose each of the j survey items, $1 \leq j \leq p$, from the p selected survey items must satisfy the following constraint:

$$V(\hat{Y}_j) \leq v_j$$

where $V(\hat{Y}_j)$ = the estimated sampling variance for the j^{th} survey total,

v_j = the desired or target sampling variance for the j^{th} survey total.

Assume the following cost function:

$$C(x) = \sum_{i=1}^{l} c_i n_i = \sum_{i=1}^{l} c_i / x_i$$

where c_i = the average cost per segment in the i^{th} land-use stratum,

n_i = the number of sample segments in the i^{th} land-use stratum,

l = the number of land-use strata,

$x_i = 1/n_i \; ; \; n_i \geq 1.$

The problem then reduces to minimizing the cost function subject to the constraints that :

$$\sum_{i=1}^{l} a_{ij} \, x_i \leq 1 \; ; \; 1 \leq j \leq p$$

where $$a_{ij} = \frac{N^2{}_i \; S^2{}_{ij}}{v_j + \sum_{i=1}^{l} N_i \; S^2{}_{ij}}$$

$S^2{}_{ij}$ = the square of the standard deviation for the j^{th} survey item in the i^{th} land-use stratum,

N_i = the number of segments in the i^{th} land-use stratum.

The nonlinear algorithm iteratively finds the intersection between $A_k = \{x\colon C(x) = k\}$, for fixed values of k, and $F = \{x\colon a'_j x \leq 1\}$. The intersection is the optimal solution. Experience has shown that the program converges rapidly to the optimal solution.

Given this allocation model, the input for the model is generated as follows:

(A) The average cost per segment for each land-use stratum, c_i, is estimated by having the interviewers keep time records during the field work.

(B) The desired sampling variance for the estimated total of each item, v_j, is established by the Area Frame Section after consultation with others in NASS.

(C) The square of the standard deviation, S^2_{ij}, for the j^{th} item in the i^{th} land-use stratum is estimated using the previous two years' survey data.

The area frame sample allocations among and within states are evaluated periodically to determine if a reallocation of the sample is worthwhile. The sample allocations among the 48 states for 1987 are shown in Table 2.4.

Table 2.4 Number of Segments in the Area Frame Sample; 1987

STATE	NUMBER OF SEGMENTS	STATE	NUMBER OF SEGMENTS
Alabama	359	Nebraska	390
Arizona	374	Nevada	104
Arkansas	400	New Hampshire	30
California	911	New Jersey	247
Colorado	457	New Mexico	292
Connecticut	48	New York	380
Delaware	72	North Carolina	391
Florida	425	North Dakota	376
Georgia	436	Ohio	324
Idaho	362	Oklahoma	360
Illinois	300	Oregon	372
Indiana	324	Pennsylvania	330
Iowa	298	Rhode Island	14
Kansas	435	South Carolina	335
Kentucky	338	South Dakota	352
Louisiana	376	Tennessee	349
Maine	150	Texas	840
Maryland	252	Utah	324
Massachusetts	48	Vermont	70
Michigan	343	Virginia	343
Minnesota	343	Washington	360
Mississippi	402	West Virginia	250
Missouri	450	Wisconsin	310
Montana	362	Wyoming	257

There were 15,665 segments in the national area frame sample. This represents approximately 0.6 percent of the total segments in the 48 states. The sample size is smallest in the New England states, Delaware and Nevada and largest in California and Texas. Excluding these states, the average sample size is about 300 segments for northeastern states, 350 segments for central and western states, and 375 segments for southern states.

The allocation of the sample across land-use strata naturally concentrates the majority of the sample segments in the cultivated strata. Approximately 46 percent of the national sample is in the intensively cultivated strata, 24 percent in the less intensively cultivated strata, 19 percent in the range or pasture strata, 7 percent in the ag-urban strata, 3 percent in the urban and resort strata, and 1 percent in the non-agricultural strata.

SELECTION PROBABILITIES

There are two methods for selecting the ultimate sampling unit or segment--equal and unequal selection. Which method is used depends on the availability of adequate boundaries for segments. If good boundaries are plentiful so that segments can be made approximately the same size within a land-use stratum, then segments are selected with equal probability. If adequate boundaries are not available, then unequal probability of selection is used since segment sizes are allowed to vary greatly in order to ensure easily identifiable segment boundaries.

The use of unequal selection probabilities is restricted to the non-agricultural stratum in area frames developed since 1985 and to most range strata in twelve western states. In all other land-use strata in the U.S., equal probability of selection is used. About 95 percent of the approximately 16,000 segments in the area frame sample are selected based on the equal probability of selection method.

The probability expressions for equal and unequal probability of selection will now be derived in the context of NASS's area frame design. These expressions provide the statistical foundation for area frame sampling.

<u>EQUAL PROBABILITY OF SELECTION:</u> Either a two-step or three-step procedure is used to select a sample segment depending upon the size of the PSU. If the PSU contains ten or fewer segments, then a two-step process is employed. If the PSU contains more than ten segments, then a three-step process is used.

Recall from the first chapter that once the size of a PSU has been accurately determined using the digitization process, the number of segments in the PSU can be derived. The number of segments in a PSU is simply the total area of the PSU divided by the target (desired) segment size for the land-use stratum in which the PSU has been stratified. This quotient is rounded to the nearest integer since fractional segments are not allowed.

For example, if a PSU in an intensively cultivated stratum is 4.1 square miles and the target segment size is 0.5 square mile, then the number of segments for the PSU is eight. Since eight segments are required for the PSU, the two-step selection process is used.

(A) Two-Step Procedure: The first step involves the selection of a sample of PSU's within each substratum in a given land-use stratum. Selection is done randomly, with replacement, with probability proportional to the number of segments in the PSU. That is, the probability of selecting the k^{th} PSU in the j^{th} substratum from the i^{th} land-use stratum is:

$$P(A_{ijk}) = \frac{N_{ijk}}{N_{ij}}$$

where A_{ijk} = the k^{th} PSU in the j^{th} substratum from the i^{th} land-use stratum,

N_{ijk} = the number of required or potential segments in the k^{th} PSU from the j^{th} substratum in the i^{th} land-use stratum,

N_{ij} = the number of required segments in the j^{th} substratum from the i^{th} land-use stratum.

After the sample of PSU's is selected, each "selected" PSU is divided into the required number of segments. The second step of the two-step sampling process involves randomly selecting a segment with equal probability from the selected PSU. That is, the probability of selecting the m^{th} segment given the k^{th} PSU was selected from the j^{th} substratum in the i^{th} land-use stratum is:

$$P(B_{ijkm} \mid A_{ijk}) = \frac{1}{N_{ijk}}$$

where B_{ijkm} = the m^{th} segment in the k^{th} PSU from the j^{th} substratum and i^{th} land-use stratum,

A_{ijk} and N_{ijk} are as previously defined.

Therefore, the unconditional probability of selecting the m^{th} segment in the k^{th} PSU from the j^{th} substratum in the i^{th} land-use stratum is:

$$P(B_{ijkm}) = P(A_{ijk})\, P(B_{ijkm} \mid A_{ijk})$$

$$= \frac{N_{ijk}}{N_{ij}} \cdot \frac{1}{N_{ijk}}$$

$$= \frac{1}{N_{ij}}$$

where N_{ij} is as previously defined.

Therefore, all segments within a given substratum in a land-use stratum have an "equal" probability of selection using the two-step selection procedure. This fact is illustrated in Table 2.5 for a hypothetical substratum with seven PSU's. This table shows the number of required segments in each PSU, the probability of selecting each PSU, $P(A_{ijk})$, the probability of selecting a segment given the PSU was selected, $P(B_{ijkm} \mid A_{ijk})$, and the unconditional probability of selecting a segment in the PSU, $P(B_{ijkm})$. Notice that the unconditional selection probability is the same for all segments, as previously stated.

Table 2.5 Selection Probabilities for the Two-Step Procedure

PSU	NUMBER OF SEGMENTS	$P(A_{ijk})$	$P(B_{ijkm} \mid A_{ijk})$	$P(B_{ijkm})$
1	2	2/40	1/2	1/40
2	3	3/40	1/3	1/40
3	5	5/40	1/5	1/40
4	6	6/40	1/6	1/40
5	7	7/40	1/7	1/40
6	8	8/40	1/8	1/40
7	9	9/40	1/9	1/40

(B) Three-Step Procedure: If there are more than ten segments in a selected PSU, then a three-step selection procedure is used to reduce the number of segments that have to be delineated on the aerial photography. (This will be explained in more detail later.) The first step in the selection process is identical to the first step of the two-step process, namely, a sample of PSU's is selected with replacement, with probability proportional to the number of segments in the PSU. That is,

$$P(A_{ijk}) = \frac{N_{ijk}}{N_{ij}}$$

In the second step of the three-step selection process, the "selected" PSU is divided into secondary sampling units (SSU's), each with approximately the same number of potential segments. Then, an SSU is selected with probability proportional to the number of segments in the SSU. That is,

$$P(B_{ijkl} \mid A_{ijk}) = \frac{N_{ijkl}}{N_{ijk}}$$

where B_{ijkl} = the l^{th} SSU from the k^{th} PSU in the j^{th} substratum and i^{th} land-use stratum,

N_{ijkl} = the number of required segments in the l^{th} SSU of the k^{th} PSU from the j^{th} substratum and i^{th} land-use stratum,

N_{ijk} is as defined earlier.

The unconditional probability of selecting the l^{th} SSU in the k^{th} PSU from the j^{th} substratum and i^{th} land-use stratum is therefore:

$$\begin{aligned} P(B_{ijkl}) &= P(A_{ijk})\, P(B_{ijkl} \mid A_{ijk}) \\ &= \frac{N_{ijk}}{N_{ij}} \; \frac{N_{ijkl}}{N_{ijk}} \\ &= \frac{N_{ijkl}}{N_{ij}} \end{aligned}$$

where N_{ij} and N_{ijkl} are as previously defined.

The third and final step simply involves selecting a segment with equal probability from the selected SSU. In formula notation,

$$P(C_{ijklm} \mid B_{ijkl}) = \frac{1}{N_{ijkl}}$$

where C_{ijklm} = the m^{th} segment from the l^{th} SSU in the k^{th} PSU for the j^{th} substratum and i^{th} land-use stratum.

Therefore, the unconditional probability of selecting the m^{th} segment in the l^{th} SSU from the k^{th} PSU in the j^{th} substratum and i^{th} land-use stratum is:

$$
\begin{aligned}
P(C_{ijklm}) &= P(A_{ijk})P(B_{ijkl} \mid A_{ijk})P(C_{ijklm} \mid B_{ijkl}) \\
&= P(B_{ijkl})P(C_{ijklm} \mid B_{ijkl}) \\
&= \frac{N_{ijkl}}{N_{ij}} \; \frac{1}{N_{ijkl}} \\
&= \frac{1}{N_{ij}}
\end{aligned}
$$

Therefore, as with the two-step procedure, the three-step selection procedure selects segments within a given substratum from a land-use stratum with "equal" probability.

UNEQUAL PROBABILITY OF SELECTION: Segments are selected with unequal probability in most range strata for twelve western states and in the non-agricultural stratum for states receiving a new area frame since 1985. This type of selection is performed because adequate boundaries are not available in these areas to draw off segments of approximately the same size. Either a one or two-step selection process is used.

(A) One-Step Procedure: The federal and state governments still retain ownership of large areas of land in western states. This land is generally administered by the Bureau of Land Management or Forest Service. Some of this land is suitable for cattle grazing and is divided into grazing allotments and leased for long periods of time. As mentioned in the first chapter, NASS stratifies this land into public range strata. The samples from these strata in South Dakota, Oregon and Washington are selected using a one-step selection process. In South Dakota, all grazing allotments were identified and drawn off as PSU's at the time the area frame was developed. A sample of PSU's is then selected with probability proportional to the "acres" +in the allotment. In Oregon and Washington, the grazing allotments were also drawn off as PSU's when the frames were developed. The only difference in these two states is that the PSU's are sampled with probability proportional to the "AUM's" (animal units per month) in the allotment. Land areas in the non-agricultural strata for frames developed since 1985 are also selected with probability proportional to the size or acreage of the land area or PSU.

For the one-step selection process, the PSU and segment are synonymous. The probability of selecting the k^{th} PSU (grazing allotment or non-agricultural area) in the j^{th} substratum from the i^{th} land-use stratum is:

$$P(A_{ijk}) = \frac{S_{ijk}}{S_{ij}}$$

where A_{ijk} = the k^{th} PSU from the j^{th} substratum in the i^{th} land-use stratum,

S_{ijk} = the size (acres or AUM's) of the k^{th} PSU in the j^{th} substratum from the i^{th} land-use stratum,

S_{ij} = the size of the j^{th} substratum in the i^{th} land-use stratum.

(B) Two-Step Procedure: A two-step selection process is used in virtually all range strata (public and private) for eight western states (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah and Wyoming) and in the public range stratum in California. In the first step, a sample of PSU's is selected with replacement, with probability proportional to the size or acres in the PSU. The probability of selecting the k^{th} PSU in the j^{th} substratum from the i^{th} land-use stratum is again defined as:

$$P(A_{ijk}) = \frac{S_{ijk}}{S_{ij}}$$

The second step makes use of point sampling, which will be described in more detail in the next section. Briefly, a point is randomly selected within the selected PSU and then boundaries are identified by state office personnel that encompass the selected point, thereby defining the segment. The boundaries for the segment can extend beyond the PSU boundaries but not beyond the substratum boundaries. The probability of selecting the m^{th} segment given the k^{th} PSU is selected in the j^{th} substratum from the i^{th} land-use stratum is:

$$P(B_{ijkm} \mid A_{ijk}) = \frac{S_{ijkm}}{S_{ijk}}$$

where B_{ijkm} = the m^{th} segment in the k^{th} PSU from the j^{th} substratum and i^{th} land-use stratum,

S_{ijkm} = the size (acres) of the m^{th} segment within the k^{th} PSU from the j^{th} substratum and i^{th} land-use stratum.

Therefore, since the boundaries for a segment can extend into multiple PSU's in the substratum, the unconditional probability of selecting the m^{th} segment from the j^{th} substratum in the i^{th} land-use stratum is:

$$
\begin{aligned}
P(B_{ij.m}) &= \sum_{k=1}^{M_{ij}} P(A_{ijk})\, P(B_{ijkm} \mid A_{ijk}) \\
&= \sum_{k=1}^{M_{ij}} \frac{S_{ijk}}{S_{ij}} \; \frac{S_{ijkm}}{S_{ijk}} \\
&= \sum_{k=1}^{M_{ij}} \frac{S_{ijkm}}{S_{ij}} \\
&= \frac{S_{ij.m}}{S_{ij}}
\end{aligned}
$$

where $B_{ij.m}$ = the m^{th} segment in the j^{th} substratum from the i^{th} land-use stratum,

M_{ij} = the number of PSU's in the j^{th} substratum from the i^{th} land-use stratum,

$S_{ij.m}$ = the size of the m^{th} segment from the j^{th} substratum and i^{th} land-use stratum,

S_{ij} is as previously defined.

That is, the probability of selection is proportional to the size of the segment for the one and two-step procedures.

The selection probabilities for all situations encountered during the sampling process have now been formulated. The expansion factor or raising factor or weight assigned to each segment to expand the survey data to population totals is derived from these selection probabilities. The expansion factor for a segment in a substratum is simply the inverse of the product of the probability of selection for the segment and the number of segments in the sample for the substratum. That is,

$$e_{ijm} = \frac{1}{p_{ijm}\, n_{ij}}$$

where e_{ijm} = the expansion factor for the m^{th} segment in the j^{th} substratum and i^{th} land-use stratum,

p_{ijm} = the probability of selecting the m^{th} segment in the j^{th} substratum from the i^{th} land-use stratum,

n_{ij} = the number of segments or replicates in the sample for the j^{th} substratum in the i^{th} land-use stratum.

SAMPLE SELECTION

The office procedures used to select the area frame samples will be described in this section for the equal and unequal probability of selection methods.

EQUAL PROBABILITY OF SELECTION: Recall that either a two or three-step selection procedure is followed when segments are selected with equal probability. The first step of both procedures (PSU selection) has been automated since 1976. However, the rest of the selection process is not automated and requires a considerable amount of manual input.

After the sample of PSU's is selected, each selected PSU is identified on the appropriate frame map. Frame maps show the PSU's boundaries from when the area frame was developed. The boundaries for the selected PSU are then transferred from the frame map to black and white aerial photography at a scale of approximately one inch equals one mile. Aerial photography is used because it provides valuable detail in terms of land use and availability of boundaries.

If the number of required segments for the selected PSU is not more than ten, then the two-step selection process is appropriate. This requires that the PSU be divided into the required number of segments on the aerial photography. If the number of segments in the PSU exceeds ten, then the PSU is first divided on the photography into SSU's of approximately the same number of segments. SSU's cannot have less than five segments or more than ten segments. Since the procedures followed do not permit a PSU to have as many as twenty segments, the number of SSU's cannot exceed three for a PSU. SSU's are introduced to reduce the number of segments that have to be delineated on the aerial photography. One of the SSU's is then randomly selected using a random number table and divided into the required number of segments.

Three criteria are followed when delineating SSU's and segments on the aerial photography in order to control the total survey error (nonsampling errors and sampling variability). These criteria are:

(1) Use the most permanent boundaries available for each SSU and segment so that reporting problems during the data collection phase caused by ambiguous boundaries will be minimized.

(2) Create SSU's and segments that are as homogeneous as possible with respect to agricultural content. Since crop types are generally not distinguishable on the photography, homogeneity is usually based on the amount of cultivated land. This criterion reduces the sampling variability among segments in a given substratum.

(3) Make the size of each segment as close to the target segment size as practical. Deviations from the target size as large as 25 percent are permitted to satisfy the first two criteria. This criterion, like the second criterion, helps control sampling variability.

After the required number of segments has been delineated for a selected PSU or selected SSU from a PSU, the segments are manually numbered on the photography in a serpentine order starting in the northeast corner. The work done so far (location of PSU on the frame map, boundary transfer to the photography, delineation of segments in the PSU or SSU, and numbering the segments) is then checked by another person to identify and resolve any errors. The person doing the checking then randomly selects the sample segment. The actual selection is performed by selecting a number from a random number table between one and the number of segments in the PSU or SSU. The segment corresponding to the random number is the selected segment. The random selection is performed by the person checking the work rather than the original person to keep the delineation of segments and the random selection of a segment independent so that the random number cannot be previewed. Later, another person (called a reviewer) reviews all the work again as part of the quality assurance process.

As an example, assume that a selected PSU requires twelve segments to be created. The PSU would first be split into two SSU's each of approximately six segments. The PSU might be split in half as shown in Figure 2.1. A random number between one and twelve is then selected. If the number selected is between one and six, then the top SSU is selected. Next, this SSU is divided into the six required segments again following the three criteria described earlier. The segments are then numbered as shown in Figure 2.1 and a random number between one and six selected. If the random number selected is five, then the segment labeled with the number five is the sample segment.

<u>UNEQUAL PROBABILITY OF SELECTION</u>: As mentioned earlier, the first step of any of the selection procedures (PSU selection) is automated. Therefore, sample selection for the public range strata in South Dakota, Oregon and Washington and for non-agricultural strata in area frames developed since 1985 is completely automated since a one-step selection procedure

is used. The location of the selected PSU is identified simply by referring to the appropriate frame map.

The two-step selection procedure used for most of the range strata in the other nine western states relies on point sampling within the selected PSU as the second step. All possible x,y coordinates within the selected PSU represent the possible sampling points. The digitizing equipment is used to identify the exact location of the random x,y coordinates selected from a random number table. If the x,y coordinates fall within the PSU, then the point is marked on the frame map as the selected point. For example, if the coordinates (40,60) were selected from the random number table for PSU 44-5 shown in Figure 2.2, then (40,60) would be the selected point.

Figure 2.2 Point Sampling in the Selected PSU

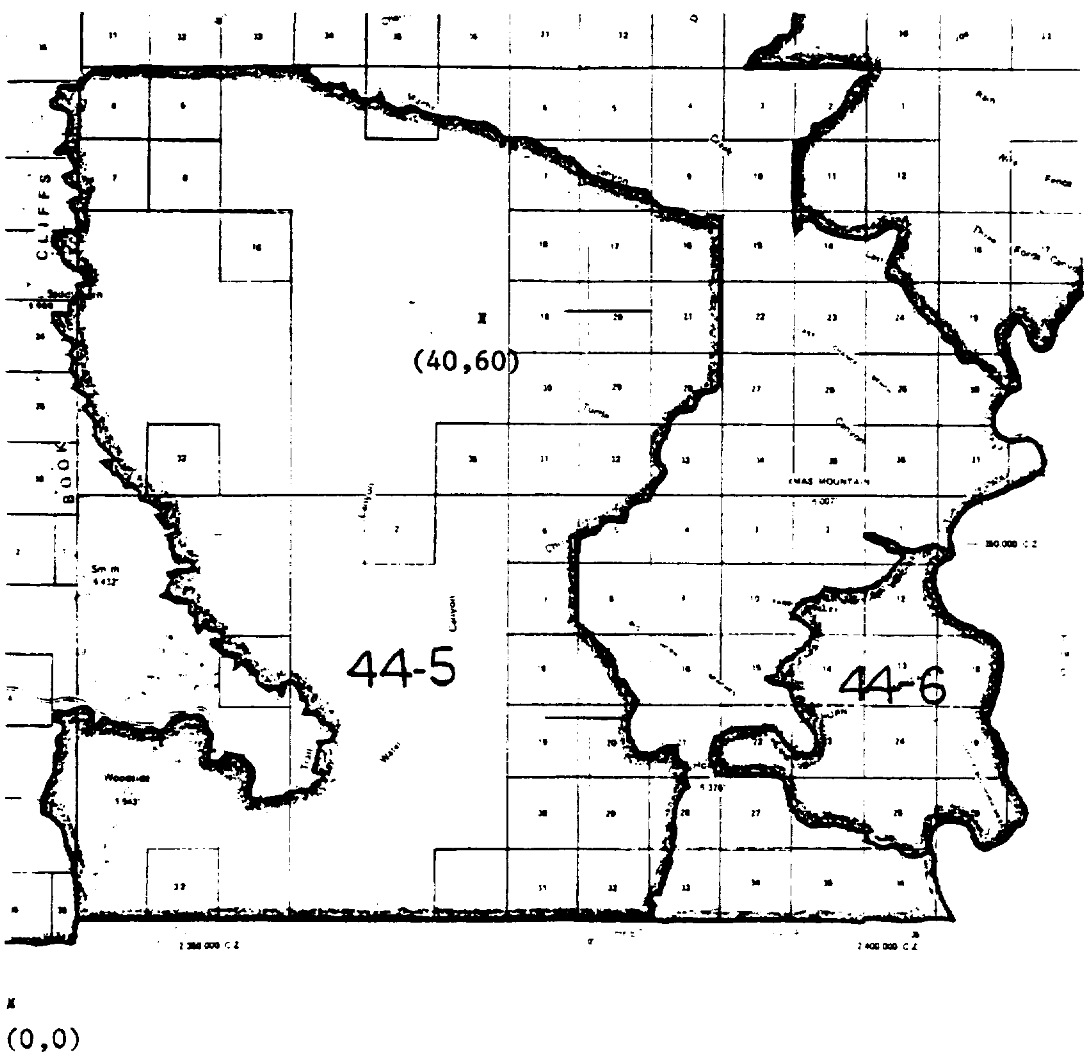


Figure 2.1 Delineating and Numbering the Segments

A copy of the frame map showing the location of the selected point and strata boundaries is then sent to the state office after the work is reviewed. The selected point is located in the field and boundaries are drawn around the selected point, thereby defining the segment. Strict rules are followed in the field to determine what land should and should not be included in the segment. Frequently, the adopted boundaries correspond to those of an allotment or grazing unit. The state office then reviews the segment's boundaries and returns the map to the Area Frame Section for a final review.

SAMPLE PREPARATION

The activities performed to prepare the samples for the survey work will be discussed in this section; again for the equal and unequal selection procedures. As was the case with the sample selection activities, the work to prepare the samples is carefully reviewed upon completion so that high quality materials are made available to the interviewers.

<u>EQUAL PROBABILITY OF SELECTION</u>: After the segment has been randomly selected, the location of the segment is identified on a county highway map for use by the interviewer when locating the segment during the survey. Then, the most recent photo coverage (at a scale of 5.28 inches equal 1 mile) is ordered for each selected segment from the Agricultural Stabilization and Conservation Service (ASCS), U.S. Department of Agriculture in Salt Lake City, Utah. The 17 inch by 17 inch photo enlargement is obtained to facilitate data collection activities such as delineating crop fields and locating farmsteads. Information concerning each photo enlargement is entered into a data file as part of an automated sample management system for keeping track of pertinent information on each selected segment, e.g. number of photos ordered, photo identification information, year the photography was flown, and date the photography was ordered.

When the photo enlargement is received from ASCS approximately six weeks later, the segment's boundaries are transferred from the small-scale aerial photograph (one inch equals one mile) to the photo enlargement. The scale of the photo enlargement is checked using reference points on the enlargement and county map to detect possible distortion in the scale. If the scale is in error, the actual scale is noted on the enlargement. Next, unique identification information such as the county name and segment number is entered on the enlargement using a computerized electromechanical device. A 12 inch by 12 inch photostat of the photo enlargement is then made on a photo reproduction camera. The photostat serves as a backup copy for the state office. The periphery of the photo enlargement is then taped to protect it from ripping during the data collection work. Finally, an accurate measure of the acreage in the segment is obtained using the digitizing equipment. That is, two independent measurements are made and the average value is used as the digitized acreage if the second measurement is within 1.5 percent of the first measurement. If the discrepancy exceeds 1.5 percent, additional measurements are taken to resolve the difference. The digitized acreage is obtained for the following two reasons:

(1) Provide a final quality assurance check before mailing the enlargements to the state offices. An automated check is made between the digitized and target acres for each segment to identify large discrepancies that might be the result of an error.

(2) Provide the interviewer with an accurate acreage measurement for the segment as a guide during data collection.

UNEQUAL PROBABILITY OF SELECTION: The sample preparation activities for segments in the range and non-agricultural strata selected with unequal probability are straightforward. A copy of the frame map showing the sample segment is provided to the state for use during the survey. Sometimes a topographic map or aerial photograph is also obtained if it would be helpful. Each segment on the maps is then labeled with its unique segment identification. The location of the segment is also marked on the county map maintained by the Area Frame Section. Finally, an accurate measure of the acreage in each segment is obtained using two or more digitized measurements as described in the previous paragraph. Using the map, an independent grid measurement is provided for the point samples in addition to the digitized values as a final check to detect acreage errors. Such care is taken with the point samples because the probability of selection and expansion factor for each segment are derived based on the digitized acreage.

The final activity performed in the sample preparation process for all segments regardless of how they were selected (equal or unequal probability) is the creation of a segment-level data file called the area frame master. The master contains pertinent sampling information on each segment such as the land-use stratum, substratum, segment and replicate numbers, the expansion factor and digitized acreage. This information is later used along with the survey data for computing the survey statistics and standard errors.

SAMPLE MAINTENANCE

NASS has an active program for reviewing the area frame sample to uncover any problems that might affect the quality of the survey results. This review consists of four components which will now be described.

(1) PROBLEM SEGMENTS: Two situations sometimes surface for which corrective measures are taken to control nonsampling errors during the data collection phase. Segments that fit either one of the following two situations are referred to as "problem" segments:

(A) The segment's boundaries on the photo enlargement are not well defined (cut through crop fields or a farmstead) which causes reporting errors.

(B) The segment is extremely difficult to canvass and enumerate due to its large size (some range segments in western states) or due to an unusually large number of interviews (twenty or more interviews of which at least ten are with farm operators). These segments can be burdensome to the interviewer, especially if the interviewer has several of these large segments, and are costly since they require two or more days to complete. Also, these segments often do not provide enough additional information from the point of view of sampling efficiency to justify the excessive contacts required.

The Area Frame Section can take corrective action for these situations provided strict statistical standards are followed so that potential biases resulting from the actions are negligible.

The first situation (poor boundaries) is identified either when originally transferring the segment's boundaries to the photo enlargement in the Area Frame Section or during the field work. Action is taken only if the boundaries can be adjusted without introducing bias. Small adjustments to the boundaries (less than five percent of the land area in the segment) can only be made if:

(A1) Neither a land-use stratum boundary nor a PSU boundary is affected.

(A2) Minor acreage adjustments resulting from changing a boundary are offsetting to the extent possible. All changes are documented and analyzed for each state.

The second situation (burdensome segments) surfaces during the field work. An allowable solution to this problem performed annually by the Area Frame Section is to divide the segment into p parts (usually two parts) given that:

(B1) Good boundaries are used to split the segment.

(B2) Each part has approximately the same amount of total land, cultivated land, agricultural operations, and agricultural operations with headquarters inside the segment.

One of the p parts is then randomly selected with equal probability of selection for use in future surveys. The selection probability for the partial segment is then multiplied by 1/p. That is, the expansion factor is multiplied by p.

(2) PHOTO REPLACEMENT: Each year, state office personnel have the opportunity to request the latest photo coverage for segments having out-of-date photography that causes interviewing problems due to major land use changes such as new housing developments. State offices also request new photo enlargements for segments with damaged or lost photography.

(3) POST-SURVEY ANALYSES: Software is available that provides graphical and statistical analyses of each area frame and area frame sample using sur-

vey data. The major graphical and statistical analyses provided for a state are:

(A) Percentage of the segments in each land-use stratum satisfying the stratum definition. This information is used to evaluate the area frame stratification.

(B) Comparison of the reported, digitized and target segment sizes in each land-use stratum. This information is used to evaluate the area frame digitizing and sampling procedures.

(C) Descriptive statistics for major agricultural commodities for each land-use stratum.

(D) Distribution of the survey estimates in each land-use stratum for each of the five sample rotation groups.

(E) Survey estimates and coefficients of variation at the substratum, land-use stratum and state levels.

These analyses sometimes uncover nonsampling errors, or suggest improved sample allocations and design alternatives which could benefit future surveys and area frame development and sampling activities.

(4) HISTORIC INFORMATION: Pertinent historic information about each area frame and area frame sample is maintained in an automated system. Basically there are three types of information for each state. These are:

(A) General information is provided for each land-use stratum such as the stratum definition, the number of substrata, the total number of segments in the stratum, and the number of segments in the sample.

(B) Survey estimates and coefficients of variation since 1979 at the state and land-use stratum levels are available for key survey items. Also, information on the number of operations, agricultural operations and resident farm operations is provided for each land-use stratum. Finally, nonresponse counts and survey estimates since 1979 are given for each of the five sample rotation groups.

(C) Historic information is provided on the number of problem segments submitted each year by each state and the number for which corrective action was taken.

This information is useful for providing a statistical history to assess if a state needs a new or updated area frame stratification and to provide information on the sample for evaluation purposes.

SAMPLE ROTATION

As mentioned earlier, NASS uses a five-year rotation scheme for the sample segments. Rotation is accomplished by replacing segments from specified replicates within a land-use stratum with newly selected segments. Preferably, the number of replicates is a multiple of five to provide a constant workload for sample selection and preparation activities in the Area Frame Section and for data collection work in the state offices. Naturally, instances occur when the number of replicates is not a multiple of five, especially for area frames developed in the 1970's. Table 2.6 illustrates how the replicates are rotated over a five-year cycle (1987 - 1991) for different numbers of replicates. For example, if a land-use stratum has twelve replicates, Table 2.6 shows that two different replicates are rotated during 1987, 1988 and 1991 while three distinct replicates are rotated during 1989 and 1990.

Table 2.6 Rotation of Replicates Depending Upon the Number of Replicates

NUMBER OF REPLICATES	YEAR				
	1987	1988	1989	1990	1991
2			1	2	
3			1	2	3
4	4	1		2	3
5	4	5	1	2	3
6	4	5,6	1	2	3
7	4,7	5,6	1	2	3
8	4	5	1,6	2,7	3,8
9	4,9	5	1,6	2,7	3,8
10	4,9	5,10	1,6	2,7	3,8
11	4,9	5,10	1,6,11	2,7	3,8
12	4,9	5,10	1,6,11	2,7,12	3,8
13	4,9	5,10	1,6,11	2,7,12	3,8,13
14	4,9,14	5,10	1,6,11	2,7,12	3,8,13
15	4,9,14	5,10,15	1,6,11	2,7,12	3,8,13

All segments are not in the sample exactly five years as has been implied. Segments from the first and last four years of an area frame's life are not in the sample exactly five years as shown in Table 2.7. This table presents the rotation cycle for an area frame assuming a twenty-year life and, for simplicity, five replicates in each land-use stratum. Segments that rotate out after the first and second years of a new frame are reentered six years later (see Table 2.7) to get more use out of the segments. Segments that rotate out after the third and fourth years of a new frame are not reentered, however. Finally, segments that rotate into the sample during the final four years of an area frame are used less than five years since a completely new area frame sample is currently introduced with the new area frame. An alternative would be to rotate into a new area frame to get more use from the segments selected during the final years of the old frame and to provide a measure of change in the production of agricultural commodities as the new frame is implemented.

Table 2.7 Rotation Cycle for a Twenty-Year Period Assuming Five Replicates in the Stratum

YEAR	REPLICATE																					
1989	1	2	3	4	5																	
1990		2	3	4	5	1																
1991			3	4	5	1	2															
1992				4	5	1	2	3														
1993					5	1	2	3	4													
1994						1	2	3	4	5												
1995	1						2	3	4	5												
1996	1	2						3	4	5												
1997	1	2							4	5	3											
1998	1	2								5	3	4										
1999	1	2									3	4	5									
2000		2									3	4	5	1								
2001											3	4	5	1	2							
2002												4	5	1	2	3						
2003													5	1	2	3	4					
2004														1	2	3	4	5				
2005															2	3	4	5	1			
2006																3	4	5	1	2		
2007																	4	5	1	2	3	
2008																		5	1	2	3	4
NUMBER OF YEARS IN SAMPLE	6	7	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	3	2	1

The national area frame sample size is approximately 16,000 segments. The total number of segments rotated each year is almost 4,000. This results from slightly less than 1,000 segments being selected for new area frames and about 3,000 segments being selected based on a twenty percent rotation of the remaining 15,000 or so segments. Therefore, almost twenty five percent of the national area frame sample is based on newly selected segments each year.

SAMPLE COST

This section will discuss the cost to select and prepare a segment for survey use based on salaries and materials during 1987. The cost estimates do not include any data collection or state office expenses, but only costs incurred by the sampling process. The costs will now be presented for the equal and unequal selection methods.

<u>EQUAL PROBABILITY OF SELECTION</u>: Recall that approximately 95 percent of the segments (3,800 of 4,000 segments on the average) are selected each year using equal selection probabilities within each substratum. Therefore, virtually all the sample selection and preparation costs will be associated with this selection procedure. Table 2.8 details the costs for labor and materials to select and prepare a segment. Over three-fourths of the costs ($92.75 per segment) are associated with labor expenses, which include not only cartographic support but also supervisory, technical and administrative support. Costs for photography and other materials average $22.25 per segment. The estimated total cost is $115.00 per segment or about $437,000 for the total sample each year.

<u>UNEQUAL PROBABILITY OF SELECTION</u>: The costs for labor and materials are considerably less for this selection method since the selection process is much less time consuming and since aerial photography is usually not purchased. The average cost for the selection and preparation of a segment is $66.00 as shown in Table 2.9. Therefore, the total annual cost for the approximately 200 segments selected by this method is about $13,200.

In summary, the annual cost for area frame sampling activities for about 4,000 segments is approximately $450,000.

Table 2.8 Cost for Selecting and Preparing a Segment (Using Equal Probability of Selection)

COST PER SEGMENT		
COMPONENT	DESCRIPTION	COST
Labor	Location of selected PSU on frame map, transfer of PSU boundaries to aerial photograph, delineation of segments, selection of sample segment, transfer of segment to county map, identification of photography needed, ordering photography and quality assurance reviews.	
	(5 hours per segment x $7.25 per hour)	$36.25
	Transfer of boundaries to photo enlargement, labeling the photo, make photostat copy, digitize the segment, tape periphery of photo, quality assurance reviews and mail to state office.	
	(2 hours per segment x $7.25 per hour)	14.50
	Supervision - - Cartographic Technicians	12.00
	Statisticians	9.00
	Technical Support -- Statisticians	3.00
	Systems Analyst	2.50
	Office Support -- Administrative & Secretarial	3.50
	Annual & Sick Leave (about 15 percent of labor)	12.00
Materials	17"x17" aerial photography (An average of 1.5 photos x $10.00 per photo)	15.00
	12"x12" copy of photograph (An average of 1.5 photos x $3.00 per photo)	4.50
	Office supplies, computer expenses, mailing costs, microfiche and so forth	2.75
TOTAL	Labor and Materials	$115.00

Table 2.9 Cost for Selecting and Preparing a Segment (Using Unequal Probability of Selection)

COST PER SEGMENT		
COMPONENT	DESCRIPTION	COST
Labor	Location of selected PSU on frame map, selection of random point, make a copy of frame map for state office, quality assurance review and mail to state office.	
	(1.5 hours per segment x $7.25 per hour)	$10.88
	Transfer segment boundaries to county map, label segment, digitize segment, quality assurance review and mail to state office.	
	(1.5 hours per segment x $7.25 per hour)	10.87
	Supervision - - Cartographic Technicians	13.00
	Statisticians	10.00
	Technical Support -- Statisticians	4.00
	Systems Analyst	2.50
	Office Support -- Administrative & Secretarial	3.50
	Annual & Sick Leave (about 15 percent of labor)	8.25
Materials	Office supplies, computer expenses, mailing costs and so forth	3.00
TOTAL	Labor and Materials	$66.00

SAMPLE ESTIMATION

This final section will briefly discuss the approaches used to estimate agricultural production with an area frame sample of segments. NASS uses three area frame estimators, namely, the closed, open, and weighted segment estimators. All three require that the interviewer establish what farms are related to each segment. (A farm is defined to be all land under one operating arrangement with gross farm sales of at least $1,000 a year.) The interviewer finds out what portion of the segment is under the operation of each farm. This portion is called a tract, and the interviewer draws the boundaries of each tract on the aerial photograph, accounting for all land in the segment.

When an interviewer contacts a farmer, the closed segment approach requires that the interviewer obtain data only for that part of the farm within the tract. For example, the interviewer might ask about the total number of hogs on the land in the tract. The most common uses of the closed segment estimator is to estimate crop acreages and livestock inventories. An interviewer accounts for all land in each tract by type of crop or use and for all livestock in the tract. The main disadvantage of the closed segment estimator arises when the farmer can only report values for the farm rather than for a tract which is a subset of the farm. For example, "How many tractors do you own?" can only be answered on a farm basis. Thus, the closed segment estimator is not applicable for many agricultural items. Economic items and crop production are two major examples which farmers find difficult or impossible to report on a tract basis.

The open and weighted segment estimators, by contrast, do not have this limitation. They can be used to estimate all agricultural characteristics. This broad applicability is a major advantage for both estimators. The open segment and weighted segment approaches require that the interviewer obtain data on the entire farm. For example, the interviewer would ask about the total number of hogs on all land in the farm. The open segment approach uses these data only when the headquarters of the farm is within the segment boundaries. (Thus, the headquarters is used to identify each farm uniquely with one segment.) Using the weighted segment approach, the interviewer obtains farm data for each tract, but these farm data are weighted; the current weight used by NASS is the ratio of tract acres to farm acres. A weight based not on total acres but on the acreage of the largest crop is presently being evaluated.

Suppose the following situation occurs for a specific farm: tract acres = 10, farm acres = 100, hogs on the tract = 20, and hogs on the farm = 40. The closed segment value of number of hogs would be 20; the weighted segment value would be 40 x (10/100) = 4; and the open segment value would be 40 (if the headquarters is in the segment) or 0 (if the headquarters is not in the segment).

When estimating survey totals and variances for these estimators, segments can be treated as a stratified sample with random selection within each substratum. The formulas for each of the three estimators can be described by the following notation. For some characteristic, Y, of the farm population, the sample estimate of the total for the closed segment estimator is:

$$\hat{Y}_C = \sum_{i=1}^{l} \sum_{j=1}^{s_i} e_{ij} \sum_{k=1}^{n_{ij}} y_{ijk}$$

where l = the number of land-use strata,

s_i = the number of substrata in the i^{th} land-use stratum,

n_{ij} = the number of segments sampled in the j^{th} substratum in the i^{th} land-use stratum,

e_{ij} = the expansion factor or inverse of the probability of selection for each segment in the j^{th} substratum in the i^{th} land-use stratum,

$$y_{ijk} = \begin{cases} \sum_{m=1}^{f_{ijk}} t_{ijkm} & \text{if } f_{ijk} > 0 \\ 0 & \text{if } f_{ijk} = 0 \end{cases}$$

f_{ijk} = the number of tracts in the k^{th} segment, j^{th} substratum, and i^{th} land-use stratum,

t_{ijkm} = the tract value of the characteristic, Y, for the m^{th} tract in the k^{th} segment, j^{th} substratum, and i^{th} land-use stratum.

For the open segment estimator, the sample estimate would be the same form as the closed segment estimator:

$$\hat{Y}_O = \sum_{i=1}^{l} \sum_{j=1}^{s_i} e_{ij} \sum_{k=1}^{n_{ij}} y_{ijk}$$

except that

$$y_{ijk} = \begin{cases} \sum_{m=1}^{f_{ijk}} b_{ijkm}\, y_{ijkm} & \text{if } f_{ijk} > 0 \\ 0 & \text{if } f_{ijk} = 0 \end{cases}$$

$b_{ijkm} = $ 1 if the farm headquarters is within the segment,

0 if the farm headquarters is not within the segment.

y_{ijkm} = the value of the entire farm for the m^{th} tract in the k^{th} segment, j^{th} substratum, and i^{th} land-use stratum.

The weighted segment estimator would also be of the same form:

$$\hat{Y}_w = \sum_{i=1}^{l} \sum_{j=1}^{s_i} e_{ij} \sum_{k=1}^{n_{ij}} y_{ijk}$$

except that

$$y_{ijk} = \begin{cases} \sum_{m=1}^{f_{ijk}} a_{ijkm}\, y_{ijkm} & \text{if } f_{ijk} > 0 \\ 0 & \text{if } f_{ijk} = 0 \end{cases}$$

a_{ijkm} = the weight for the m^{th} tract in the k^{th} segment, j^{th} substratum, and i^{th} land-use stratum.

The following weight is currently in use:

$$a_{ijkm} = \frac{\text{tract acres for the } m^{th} \text{ tract}}{\text{farm acres for the } m^{th} \text{ tract}}$$

The precision of an estimate can be measured by the standard error of the estimate. An estimate becomes less precise as the standard error increases. Given the same number of segments to make each estimate, weighted segment estimates are usually more precise than closed segment estimates, and closed segment estimates are usually more precise than open segment estimates.

For all three estimators, the formula for the sampling variance can be written as:

$$V(\hat{Y}) = \sum_{i=1}^{l} \sum_{j=1}^{s_i} \frac{(1-1/e_{ij})}{(1-1/n_{ij})} \sum_{k=1}^{n_{ij}} (y'_{ijk} - y'_{ij\cdot})^2$$

where $y'_{ijk} = e_{ij} \, y_{ijk}$

$$y'_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y'_{ijk}$$

The standard error is then:

$$SE(\hat{Y}) = \{V(\hat{Y})\}^{\frac{1}{2}} .$$

In closing, research into nonsampling errors associated with the three estimators has shown that the closed estimator, when applicable, is generally the least susceptible to nonsampling errors while the open estimator is the most susceptible. Therefore, based on sampling and nonsampling error considerations, the open segment estimator is the least preferred estimator. The closed segment estimator is much relied-on for NASS's area frame surveys, and the weighted segment estimator is the most used of the three estimators for multiple frame surveys where the area frame is only used to measure the incompleteness in the list frame.

BIBLIOGRAPHY

Bethel, J. (1986): An Optimum Allocation Algorithm for Multivariate Surveys. U.S. Department of Agriculture, Statistical Reporting Service, SF&SRB Staff Report No. 89. Washington, D.C.

Fecso, R. and Johnson, V. (1981): The New California Area Frame: A Statistical Study. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 22. Washington, D.C.

Ford, B.L. (1975): Rotation Group Effects in SRS Surveys. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Ford, B.L., Nealon, J.P., and Tortora, R.D. (1986): Area Frame Estimators in Agricultural Surveys: Sampling Versus Nonsampling Errors. Agricultural Economics Research. Vol. 38, No. 2. Journal of the U.S. Department of Agriculture. Washington, D.C.

Geuder, J. (1981): An Evaluation of the New York Area Sampling Frame. U.S. Department of Agriculture, Statistical Reporting Service, Staff Report No. AGES810929. Washington, D.C.

Geuder, J. (1984): Paper Stratification in SRS Area Sampling Frames. U.S. Department of Agriculture. Statistical Reporting Service, SF&SRB Staff Report No.79. Washington, D.C.

Houseman, E. (1975): Area Frame Sampling in Agriculture. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 20. Washington, D.C.

Huddleston, H. (1976): A Training Course in Sampling Concepts for Agricultural Surveys. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS No. 21. Washington, D.C.

Nealon, J.P. (1984): Review of the Multiple and Area Frame Estimators. U.S. Department of Agriculture, Statistical Reporting Service, SF&SRB Staff Report No. 80. Washington, D.C.

Pratt, W. (1974): The Use of Interpenetrating Sampling in Area Frames. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

U.S. Department of Agriculture (1983): Scope and Methods of the Statistical Reporting Service. Publication No. 1308. Washington, D.C.

U.S. Department of Agriculture (1984): Area Frame Analysis Package. Statistical Reporting Service Staff Report. Washington, D.C.

U.S. Department of Agriculture (1987): International Training: Area Frame Development and Sampling. National Agricultural Statistics Service. Washington, D.C.

U.S. Department of Agriculture (1987): Supervising and Editing Manual, June Enumerative and Agricultural Surveys. National Agricultural Statistics Service. Washington, D.C.

Wigton, W. and Bormann, P. (1978): A Guide to Area Sampling Frame Construction Utilizing Satellite Imagery. U.S. Department of Agriculture. Statistical Reporting Service Staff Report. Washington, D.C.