United States
Department of
Agriculture

Statistical
Reporting
Service

Statistical
Research
Division

SF&SRB Staff
Report No. 79

Washington, D.C.

February 1984

Paper Stratification in SRS Area Sampling Frames

Jeff Geuder

PAPER STRATIFICATION IN SRS AREA SAMPLING FRAMES, by Jeff Geuder. Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. February 1984. SF&SRB Staff Report No. 79.

ABSTRACT

This paper explains the process of paper stratification in the area sampling frames used by the Statistical Reporting Service. Paper strata are simply substrata of larger land use strata, and are often referred to as geographic substrata. In the SRS design, count units (equivalent to primary sampling units) in a land use stratum are ordered prior to sampling, so that similar units are grouped. This ordered population of sampling units is then divided into equal sized subpopulations, which are the paper strata. Depending on how the count units are ordered prior to sampling, a paper stratum may be comprised of those sampling units in a particular geographic region of a state. Since similar count units may be located in different areas of a state, however, a paper stratum may comprise two or more disjoint areas in various parts of the state. The use of paper stratification also makes it possible for SRS to use an interpenetrating sample design. In this design, several independent replications are selected in a land use stratum, each containing one sampling unit from each paper stratum. This design disperses the sample across the state and allows for the rotation of sample units to reduce respondent burden.

ACKNOWLEDGEMENTS

The author would like to thank Paul Blackwood, Ron Fecso, Bob Hale, Dale King, Ron Radenz, Marlowe Schlegel and Bob Tortora for reviewing early drafts of this paper. Dick Coulter and Bill Pratt made several helpful suggestions during critical reviews of the final draft.

TABLE OF CONTENTS

PAPER STRATIFICATION IN SRS AREA SAMPLING FRAMES

INTRODUCTION

Area sampling frames are used by SRS for estimating a variety of agricultural characteristics (crop acreages, livestock inventories, farm labor, etc.). A separate sampling frame is constructed and maintained for each of the 48 conterminous states. These sampling frames are stratified according to land use, with five basic strata: cultivated land; range and pasture; water; nonagricultural land; and cities and towns. The cultivated land in most states is further stratified, by separating "intensively" cultivated land from "extensively" cultivated land. Since a separate frame is maintained for each state, different land use strata can be used, depending on the agricultural characteristics of the state.

Within each land use stratum, primary sampling units are defined using identifiable, physical boundaries. The population of primary sampling units is ordered to group those with similar agricultural characteristics. The ordered population is then divided into equal sized sub-populations, which are called "paper strata." Independent random samples are then selected from each of the paper strata.

The purpose of this paper is to explain the second level of stratification in the SRS area frame design, the "paper stratification." This process is not clearly understood within the agency, primarily because the term "paper strata" is not very descriptive. Stratification is usually based on some characteristic (such as land use), and "paper stratification" does not indicate what criteria are used to do the stratification. The term came into use several years ago when the sample selection process from the area frame was carried out by listing the frame units (on paper) and accumulating the number of sampling units. The population of sampling units in a land use stratum was then divided into equal sized sub-populations, called paper strata.

Research reports published by SRS and dealing with area frame sampling [1, 2, 3] have equated paper strata with "geographic" substrata. This term can be interpreted in two ways. In one sense, it implies that paper strata correspond to geographic regions within a state. This will be true whenever the ordered population of sampling units is continuous (i.e. it does not "skip" from one area of the state to another). In another sense, it implies that the substratified design disperses the sample throughout the state. This is true because of the "interpenetrating" sample design (which is described in a later section of this paper).

This paper will describe and give examples of each stage of the area frame construction process, since each stage has a bearing on the paper stratification. The examples will all be taken from the Nebraska area sampling frame, to provide consistency and allow the reader to see how each stage is related to the others. Appendix A shows the land use strata used in the 1983 June Enumerative Survey (JES) in Nebraska. These are typical of those used in most states, except that stratum 11 is sometimes combined with stratum 12.

SRS AREA FRAME PROCEDURES

A detailed explanation of the concept of paper stratification cannot be achieved without describing some of the general aspects of the SRS area frame design. Also, some of the terminology used by SRS is different from that in more general sampling applications. This section gives an overview of the frame construction procedures and definitions of the SRS terminology.

Background Data

When a new area sampling frame is to be built for a state, data on the state's agricultural characteristics are evaluated to determine which land use strata should be used. There are general land use strata that are used in most states, but some states have characteristics which suggest the use of additional and/or special strata. For example, the area sampling frames in Pennsylvania, New York, and Nebraska contain only the basic land use strata, while the area frames in California, Texas, and Florida all contain one or more "crop specific" strata. The area frames in Washington, Oregon, and Idaho have strata for dryland grain. Crop specific stratification is described in [1] and [2].

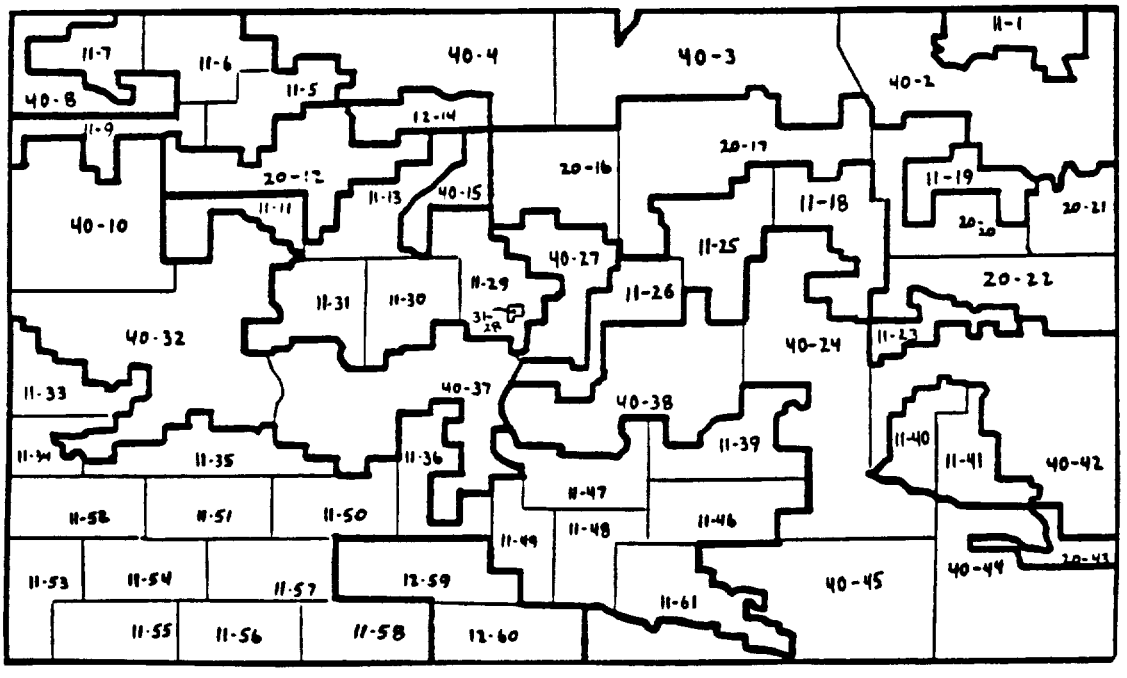
Stratification

Once the land use strata have been defined, county highway maps, aerial photography, and other cartographic material are obtained. The stratification unit of the Sampling Frame Development Section (SFDS) then begins to stratify the land in each county. Areas of land which meet the various stratum definitions are identified and delineated on the aerial photographs using permanent and identifiable boundaries. These homogeneous areas of land are called stratum blocks.

The boundaries of the stratum blocks are then transferred to county highway maps and subdivided into count units, which are equivalent to primary sampling units or PSU's. They were originally called count units because highway maps showed the number of farm units in a given block of land, and these farm units were "counted" to determine the appropriate stratum for the unit. Figure 1 shows the land use stratification for Banner county Nebraska. The heavy lines correspond to stratum boundaries, which are color-coded on the actual frame materials. The boundaries of the count units are denoted by the lighter lines.

The count units are numbered serpentinely throughout the county. The number assigned to a count unit is comprised of the stratum code and a sequence number, corresponding to the location of the count unit in the county. Count unit number one is usually in the northeast corner of the county. Also note that the size of the count units may vary by stratum. The primary reason for the use of count units is to avoid the time consuming task of delineating every individual sampling unit in the state. The count unit serves as the first-stage sampling unit, and only

Figure 1 - Land use stratum and count unit boundaries, Banner county, Nebraska.



those count units that are selected will be further subdivided into segments, which are equivalent to elementary sampling units. Experience has shown that to make this process efficient, count units should contain from two to ten segments. Notice that count unit number one in Figure 1 is much smaller than count unit number two. The reason is that the segment size in stratum 11 is one square mile, while in stratum 40 it is four square miles. Therefore, the count units in stratum 11 will generally be smaller than those in stratum 40.

Digitization

All count units are then digitized, which is a process that electronically measures the area of the count unit. The number of segments in each count unit is determined by dividing the area of the count unit by the target segment size, which is determined in advance for each stratum. Figure 2 shows portions of the list of count units in Banner county, the area of each count unit, and the number of segments it will contain.

Figure 2 - Partial listing of count unit data in Banner county Nebraska.

Count unit number	Land Use Stratum	Count unit size	Segment Size	Number of segments
		- (square miles) -		
1	11	8.9	1.0	9
2	40	29.5	4.0	7
3	40	30.6	4.0	8
4	40	27.8	4.0	7
5	11	8.7	1.0	9
6	11	8.0	1.0	8
.		.		
.		.		
.		.		
11	11	6.0	1.0	6
12	20	14.0	2.0	7
13	11	7.8	1.0	8
14	12	6.0	1.0	6
15	40	5.6	4.0	1
16	20	12.0	2.0	6
.		.		
.		.		
.		.		
27	40	17.6	4.0	4
28	31	0.2	0.25	1
29	11	7.7	1.0	8
.		.		
.		.		
.		.		
60	12	10.1	1.0	10
61	11	10.3	1.0	10

When all count units in every county have been digitized, the number of sampling units in each land use stratum is accumulated to obtain population counts. Table 1 summarizes these counts for Banner county and for the entire state.

Ordering the Population

At this point, all of the land area in a state has been stratified and digitized. The number of sampling units in each stratum has been determined. The next step is to order the population of sampling units to group units that have similar agricultural characteristics. Recall that the land use stratification was

Table 1 - Number of count units, land area, and number of sampling units by land use stratum in Banner county and the state.

Land use stratum	Banner county			State		
	Number of count units	Land area	Number of sampling units	Number of count units	Land area	Number of sampling units
		(sq. mi.)			(sq. mi.)	
11	36	290.3	289	3554	30,112.4	30,202
12	3	27.3	27	1131	8,754.6	8,794
20	7	80.5	40	982	9,531.4	4,785
31	1	.2	1	643	649.1	2,609
32	0	0	0	88	167.7	1,677
40	14	346.8	88	1161	27,695.2	6,915
50	0	0	0	43	179.5	184
62	0	0	0	22	132.4	133
Total	61	745.1	445	7624	77,222.3	55,299

based primarily on the amount of cultivation, not on specific commodities (except in the case of crop specific strata). Therefore, while most segments in a particular stratum may be highly cultivated, the content of the segments may differ depending on the location of the segment in the state. Ordering the population of sampling units to group those with similar content often leads to increased precision in the estimates of individual commodities. The sampling units in crop specific strata are ordered in the same manner as those in the general strata, although the ordering is not as critical.

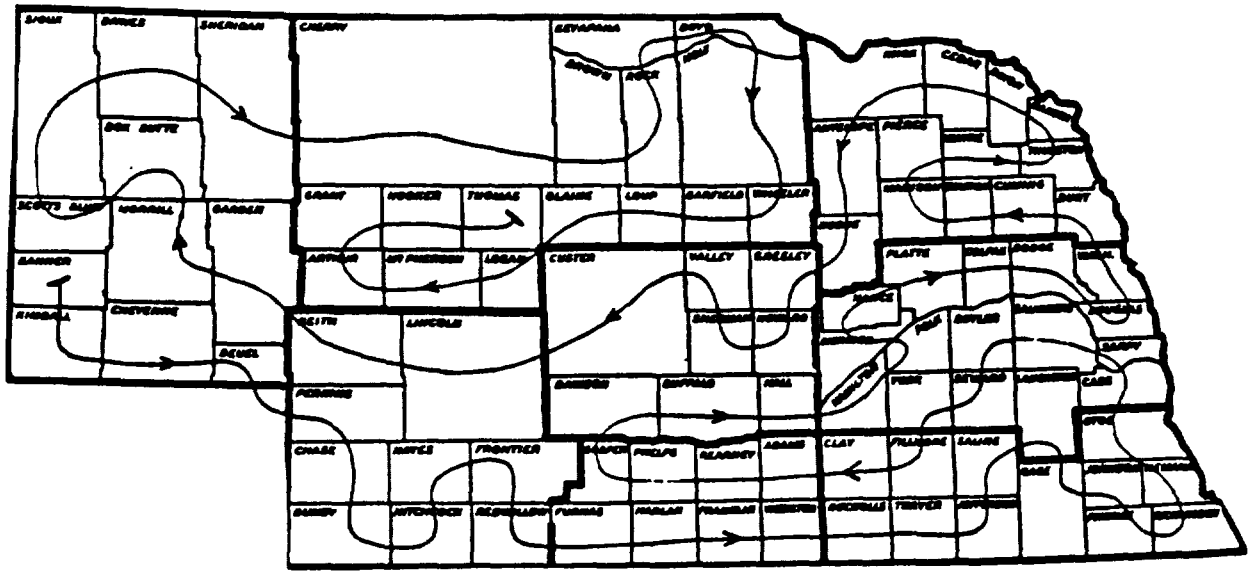
Ordering the population is a two-stage process. First, count units within a county are ordered in a serpentine fashion, as shown in Figure 1. Next, the counties are ordered in some logical manner. In the frames constructed in recent years, the counties have been ordered based on a multivariate cluster analysis of county crop and livestock data. The analysis identifies counties that have roughly the same overall agricultural characteristics, usually identifying four to seven groups of counties. For example, one group might consist of those counties that have a high proportion of land area in corn and sorghum, while another might be those that are primarily hay and pasture. In general, gains can be achieved by ordering the counties in some logical manner. However, it has been shown [2] that there is usually little difference between two "logical" orderings.

Figure 3 shows 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 more 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 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, and finally proceeds into the area of the state that is primarily rangeland and pasture.

When the ordering "enters" a county from the west or the north, the order of the count units in the county is reversed. Recall that the count units within a county are ordered by arbitrarily starting in the northeast corner of the county. Therefore, reversing the order will insure a fairly continuous ordering of count units from one county to the next.

Figure 3 -- County ordering used in the Nebraska area sampling frame.



DEFINING THE PAPER STRATA

At this point the paper strata are defined. The number of sampling units in each stratum has been determined and the population has been ordered. After the total sample size and allocation to strata is determined, the number of paper strata that will be used in each land use stratum must be determined. This determination is based on several factors, including experience with frames in other states, the number of sampling units in each stratum, and the degree of homogeneity among the segments in the stratum. A general cropland stratum, in which segments contain several commodities, will be divided into more paper strata than a crop specific stratum, in which segments contain primarily one commodity [2]. Also, since the paper strata are, in effect, geographic substrata, the need for geographic control over the sample is a consideration in determining the number of paper strata.

The ordered population of sampling units is then divided into the prescribed number of paper strata. Each paper stratum will contain the same number of population sampling units, except the last paper stratum, which may contain slightly more or less than the others, due to rounding. Table 2 shows the sample design used in Nebraska. The 30,202 sampling units in stratum 11 are divided into 15 paper strata. (The replications are simply independent random samples selected from the paper strata. These will be explained in more detail in a later section.) In this case, the first 14 paper strata in stratum 11 will each contain 2,013 sampling units and the last paper stratum will contain 2,020. In stratum 12, the 8,794 sampling units are to be divided into seven paper strata. The first six will each contain 1,256 and the last will contain 1,258.

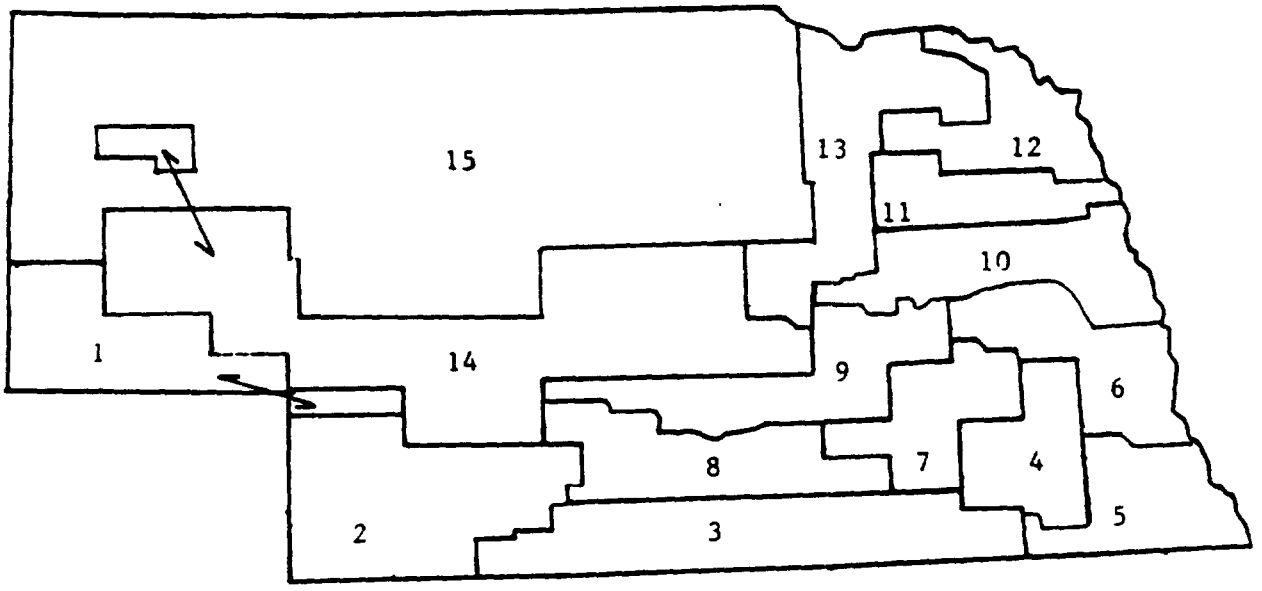
Table 2 -- Sample design used in Nebraska, 1983 JES.

Land use stratum	Population number of sampling units	Sample size	Paper strata	Number of Replications
11	30,202	210	15	14
12	8,794	70	7	10
20	4,785	35	7	5
31	2,610	10	2	5
32	1,677	5	1	5
40	6,915	40	4	10
50	184	5	1	5

The boundaries of the 15 paper strata that will be created for stratum 11 are shown in Figure 4. It is emphasized that these boundaries apply only to those sampling units in stratum 11. Since the boundaries are determined by the number of sampling units in the stratum and the number of paper strata used, they are not the same for all strata. The reader can compare this figure with Figure 3 to see how the county ordering determines the paper stratum boundaries.

The discontinuity in the first paper stratum is due to the "skip" in the county ordering described earlier. The discontinuity in paper stratum 14 is due to the fact that the count unit ordering was not reversed in Box Butte county. As a result, the "first" 184 sampling units (in the northern part of the county) were included in paper stratum 14, rather than the "last" 184 sampling units (in the southern part of the county). Notice from Figure 3 that the ordering proceeded from Morrill county into Box Butte, and then into Scotts Bluff. If the count units had been reversed in Box Butte, there would have been a "skip" from the northeast corner of Box Butte to the northeast corner of Scotts Bluff. So in either case, there would have been the potential for a minor discontinuity in the ordering.

Figure 4 -- Paper stratum boundaries in stratum 11, Nebraska.



At this point the sampling frame is complete. Before describing the interpenetrating sampling design, the paper stratification process will be summarized. There are six basic stages in the process:

1. Within each county, subdivide the stratum blocks into count units.
2. Order the count units serpentinely.
3. Order the counties in some logical manner.
4. Reverse the count unit ordering in counties when necessary to maintain continuity.
5. Determine the total number of segments in the land use stratum.
6. Divide the ordered list of count units equally into the prescribed number of paper strata.

The computer program used to carry out the sample selection produces a listing which contains all of the count units, their area, their order in the population and many of the other factors that enter into the paper stratification process. A portion of this listing is shown in Figure 5, as a summary of the preceding examples. The items highlighted with blocks are examples of some of the particular characteristics of the paper stratification process. For example, block number 1 shows that count units 1, 5, 6, and 7 in Banner County are in land use stratum 11. Count units 2, 3, 4, etc., are in other strata, and are therefore not included in this portion of the frame.

Block 2 illustrates the point at which the ordering leaves the first county and enters the second (Kimball). Due to the way the count units are ordered within a county, the last count unit (61) in Banner County should be geographically close to the first count unit (1) in Kimball county. Block 3 shows the point where the first paper stratum ends and the second begins. Since the first paper stratum in this stratum is defined to contain 2,013 segments, count unit 40 in Perkins county must be split. The first 5 segments in the count unit will be assigned to the first paper stratum and the last 4 will be assigned to the second paper stratum.

Block 4 illustrates the use of the "switch" to reverse the order of the count units in a county. The way the counties are ordered, Hitchcock county will be "entered" from the west. Therefore the continuity of the ordering will be maintained by reversing the count unit ordering. Block 5 shows the progression from count units in stratum 11 to count units in stratum 12.

The sample selection process is then repeated using the parameters for stratum 12 (number of paper strata, number of replications, accumulated number of sampling units, etc.). The relationship between the number of paper strata and the number of replications is explained in the next section.

Figure 5 - Selected portions of the sampling frame illustrating various components of the paper stratification process.

Land use stratum	County Code	Order	Switch	County Name	Count Unit	Area	No. of Segments	Accumulated Segments	Paper Stratum
11	7	1	0	Banner	1	8.9	9	9	1
11	7	1	0	Banner	5	8.7	9	18	1
11	7	1	0	Banner	6	8.0	8	26	1
11	7	1	0	Banner	7	6.8	7	33	1
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.
11	7	1	0	Banner	58	8.1	8	279	1
11	7	1	0	Banner	61	10.3	10	289	1
11	105	2	0	Kimball	1	8.2	8	297	1
11	105	2	0	Kimball	2	7.6	8	305	1
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.
11	135	5	0	Perkins	39	9.0	9	2008	1
11	135	5	0	Perkins	40	9.3	5	2013	1
11	135	5	0	Perkins	40	9.3	4	2017	2
11	135	5	0	Perkins	41	9.6	10	2027	2
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.
11	57	7	0	Dundy	85	4.8	5	3190	2
11	57	7	0	Dundy	98	9.8	10	3200	2
11	87	8	1	Hitchcock	86	4.1	4	3204	2
11	87	8	1	Hitchcock	83	7.6	8	3212	2
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.
11	47	39	1	Dawson	32	9.1	9	16099	8
11	47	39	1	Dawson	31	9.1	5	16104	8
11	47	39	1	Dawson	31	9.1	4	16108	9
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.
11	5	90	1	Arthur	32	1.0	1	30202	15
12	7	1	0	Banner	14	6.0	6	6	1
.	.		.		.		.		.
.	.		.		.		.		.
.	.		.		.		.		.

INTERPENETRATING SAMPLING

In interpenetrating sampling, the population of sampling units (segments) is divided into equal sized sub-populations (paper strata), and several independent random samples (called "replications") are selected. The sample selection process is such that a replication is comprised of one segment from each paper stratum. Therefore, since the paper strata correspond to geographic regions within the state, this design insures that the sample will be dispersed throughout the state. This illustrates a second reason for the use of paper strata (the first being increased precision of the estimates). If paper strata were not used, the possibility would exist of having the sample clustered in one or two areas of the state, due to the random selection process. The use of replications allows SRS to use a rotation scheme to reduce respondent burden. After each survey, approximately 20 percent of the sample is rotated, and replaced by new replications. The flexibility offered by using replications also makes it possible to reallocate the sample without having to select an entirely new sample.

The interpenetrating sample design is described schematically in Figure 6. This figure shows the relationship between paper strata and replications in stratum 11 in Nebraska. In this stratum, there are fifteen paper strata and 14 replications for a total of 210 segments. Note that segments numbered 3001 through 3015 comprise the first replication. Since each replication is an independent random sample from the population of segments in the stratum, the data from these fifteen segments could be used to estimate stratum totals. Also, note that segments 3001, 3016, 3031, ..., through 3196 comprise the sample from the first paper stratum. These segments are used to estimate totals for the area defined by the paper stratum boundaries for the first paper stratum (shown in Figure 4).

Figure 6 -- Schematic diagram of interpenetrating sampling--numbers in the table correspond to segment identification numbers.

Paper Stratum	Replication									
	1	2	3	4	.	.	.	.	13	14
1	3001	3016	3031	3046	.	.	.	.	3181	3196
2	3002	3017	3032	3047	.	.	.	.	3182	3197
3	3003	3018	3033	3048	.	.	.	.	3183	3198
4	3004	3019	3034	3049	.	.	.	.	3184	3199
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
14	3014	3029	3044	3059	.	.	.	.	3194	3209
15	3015	3030	3045	3060	.	.	.	.	3195	3210

STATISTICAL GAINS DUE TO PAPER STRATIFICATION

It was pointed out earlier that one of the advantages of using paper strata was that it insured that the sample would be dispersed throughout the state. Paper stratification also produces some statistical gains. Sampling error is reduced by creating substrata that are more homogeneous than the larger land use strata. The amount of reduction depends on how effective the process is in grouping similar count units. If there are a few small "pockets" in a state where a particular commodity is grown, and the paper stratification isolates these areas, gains may be substantial. If a crop is grown throughout a state, the gains will probably be minimal. The point to be made here is that any gains achieved through the use of paper stratification are "bonuses", since the main purpose of the process is to allow the use of interpenetrating sampling for sample dispersion and rotation.

SUMMARY

Paper stratification is the term used to describe the second level of stratification in the SRS area sampling frames. The term came about because of the "artificial" way in which the substrata were created. The sampling units from similar counties were grouped (on paper) prior to selecting the sample.

Paper strata are used in the SRS design for two reasons. First, they insure that the sample will be distributed throughout the state rather than possibly being clustered in a few small areas. They also have the potential to improve the precision of the estimates by creating substrata which are more homogeneous than a land use stratum as a whole.

The paper strata correspond to geographic regions in the state. These regions may or may not be comprised of a continuous area. This depends on how the count units in the population are ordered or arranged prior to selecting the sample. If the ordering of the counties in the state is not continuous (i.e. it skips from one county to a nonadjacent county), a paper stratum may be comprised of two or more disjoint areas. Whether the paper strata are continuous or disjoint, however, is not the main concern. If paper stratification is effective, gains will be achieved in the precision of the estimates made from the frame.

REFERENCES

[1] Fecso, Pon, and Van Johnson, The New California Area Frame: A Statistical Study. U.S. Dept. of Agr., Stat. Rept. Serv. September 1981.

[2] Geuder, Jeff. An Evaluation of SRS Area Sampling Frame Designs: Ordering Count Units and Creating Paper Strata. U.S. Dept. Agr., Stat. Rept. Serv. July 1983.

[3] Pratt, William L. The Use of Interpenetrating Sampling in Area Frames. U.S. Dept. Agr., Stat. Rept. Serv. May 1974.

APPENDIX A

Land Use Stratification in Nebraska, 1983 JES

Stratum Code		Stratum Definition
11	-	General cropland, 80% or more cultivated
12	-	General cropland, 50% to 79% cultivated
20	-	General cropland, 15% to 49% cultivated
31	-	Agri-urban, residential mixed with agriculture, more than 20 dwellings per square mile
32	-	Residential/commercial, no agricultural land
40	-	Pasture and range, less than 15% cultivated
50	-	Nonagricultural land (parks, military installations, etc.)
62	-	Water, more than 1 square mile in area