United States
Department of
Agriculture

Statistical
Reporting
Service

Statistical
Research
Division

SRS Staff Report
Number 86

May 1985

Sample Design for the 1985 ISP/JES

James Bethel

SAMPLE DESIGN FOR THE 1985 ISP/JES. By James Bethel, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250. April, 1985. SF&SRB Staff Report No. 86.

ABSTRACT

The ISP consolidates several SRS surveys into one multipurpose survey. In this report we redesign the list stratification for the three ISP states and obtain optimal sample allocations for both the list and area frames for each state. The choice of stratification variables, the effectiveness of the sample design, the allocation model, and methods for optimizing stratum boundaries and sample allocation are discussed.

Table of Contents

1. Introduction

The Integrated Survey Program (ISP) attempts to unify some of the many surveys carried out by the USDA Statistical Reporting Service into one global, multivariate survey program from which estimates on many variables can be generated over the course of the year. The heart of this program is the June Enumerative Survey (JES), a multivariate, multiple frame survey from which subsamples can be drawn for the ensuing periodic surveys (eg., the quarterly hog, chicken and grain stocks surveys). The ISP was introduced in Illinois, Tennessee and Arizona in 1984 and will be used again in these states in 1985.

The sample design for the JES is rather simple. There are two frames, a list frame and an area frame. Both are stratified: the list frame is stratified by a number of variables (which will presently be examined in detail); the area frame is stratified by land use (50%-80% agricultural, 20%-50% agricultural, urban, etc.) and by geographical groupings, usually referred to as "paper strata". Within the list frame strata systematic samples of farming operations are selected. Technically the area frame sampling is two-stage cluster sampling, but only one secondary sampling unit is selected per primary sampling unit, thus eliminating intra-class correlations. For our purposes--that of analysing variances within strata--this sampling method is equivalent to the selection of simple random samples of area segments within paper strata. Farming operations in the area frame sample which also appear on the list frame are deleted (from the sample), then estimated totals are generated from both frames and combined to form the multiple frame estimates.

The purpose of this study is to evaluate and attempt to simplify the list frame stratification used in these states in 1984 and to determine optimal sample allocations. Our approach to this can be outlined as follows: we first assessed the performance of the current list stratification; then we redesigned the stratification and evaluated the results, comparing them with those for the current design. The new stratification design compares favorably with the 1984 design; while maintaining virtually the same level of sampling efficiency, the number of list strata has been reduced, in each case, by about 40%. Using 1984 JES survey and cost information, the optimal sample allocations were then determined for a range of sampling requirements and, after considerable experimentation, specific recommendations were made for each state.

The recommended sample allocations--which cover both list and area frames--do not differ greatly from the current allocations in either the estimated cost or the projected level of sampling efficiency. In each case, an attempt was made to maintain or slightly improve the coefficients of variation of key variables while keeping the cost slightly below 1984 levels.

Sections 2 and 3 of this paper examine and compare the current and proposed stratificatiion schemes and section 4 discusses the sample allocation. While some technical detail is given in these sections, most of it is reserved for the appendices in section 6.

Before moving on we should address the question of how this survey fits in with the rest of the ISP. This report focusses on the June survey period and establishes sample sizes which guarantee certain levels of accuracy. On subsequent surveys this master sample will be subsampled and, possibly, augmented and/or partially replaced by independent replicates. The sampling variability will be about the same for subsequent surveys provided that the same sample sizes are maintained. Many other issues are involved here: subsampling versus rotation, double sampling, and post-stratification, to name a few. These are certainly worthy of attention but they are, unfortunately, beyond the scope of this paper.

2. The Current Stratification

The current list strata are given in Tables 1A-1B, along with the corresponding population and 1984 JES sample sizes. (The sample size given here is the number of useable data, not the original target sample size.) The strata are created starting with the last category and working upward. That is, any farming operation in Arizona with at least 1600 hens is placed in the HPLA 1600+ stratum regardless of any other characteristics. Simililarly, if a farm has at least 200 hogs (and less than 1600 hens) it is classified into the Hogs 200+ stratum regardless of other characteristics.

In each state, the current stratification is quite detailed, encompassing all agricultural items of interest, with several stratification categories of each. Each state is stratified by all major livestock commodities (dairy, hogs, cattle), as well as some less common ones (sheep, chickens). Specific crops are not used in stratification, although total cropland (or total land) is used in each state. In addition, Illinois uses storage capacity (to stratify grain stocks) and Tennessee uses some geographic strata (crop reporting districts).

Table 2 gives estimates for the standard deviation and coefficient of variation (CV) of the sample mean under simple random sampling and under the current stratification*. From this table it is clear that the strata are effective in Arizona, somewhat less so in Tennessee, and relatively ineffective in Illinois. While there some truly impressive gains (eg., cattle in Arizona, hogs in Illinois), generally speaking these results are somewhat disappointing. As we will see, however, the control information available to implement this (or any other) design has serious shortcomings and it seems unlikely that the design itself is at fault.

---

*Variance calculations are discussed in section 6.3.

Table 1A. Current Stratification

Arizona Stratum	Population	Sample	Tennessee Stratum	Population	Sample
Cropland 1-24	120	10	CRD* 10 or 20	12128	353
Cattle 1-49	353	28	CRD* 30, 40 or 50	37436	875
Hogs 1-199	20	4	CRD* 60	28823	534
Cropland 25-499	508	78	Cattle 50-99	5972	237
Cropland 500-999	231	61	Cattle 100-499	2209	129
Cropland 1000-1999	128	43	Dairy 50-199	1357	81
Cropland 2000-4999	64	30	Hogs 50-99	2969	184
Cattle 50-499	731	119	Hogs 100-499	2073	206
Dairy 50-199	20	8	Cropland 500-1999	463	77
Cattle 500-999	159	38	Sheep 1-39	188	39
Cattle 1000-3999	76	14	Cattle 500-1499	72	6
Dairy 200-999	101	28	Hogs 500-1999	303	29
Cropland 5000+	25	25	Cropland 2000+	36	36
Cattle 4000+	4	4	Cattle 1500+	7	7
COF* 300+	25	25	Dairy 200-499	97	12
Dairy 1000+	16	16	Dairy 500+	4	4
Sheep 1+	73	37	Sheep 40+	65	16
Hogs 200+	44	22	Hogs 2000+	29	29
HPLA* 1600+	5	2	HPLA* 3000+	26	7

Table 1B. Current Stratification

Illinois

Stratum	Population	Sample	Stratum	Population	Sample
All land 1-499	30611	618	Hogs 150-499	5579	477
Cattle 1-49	10730	312	Hogs 500-999	1869	199
Cattle 50-99	2679	87	Sheep 30-99	1020	138
Hogs 1-149	7002	254	Sheep 100-499	1040	188
All land 500-999	7519	274	Hogs 1000-1999	740	148
All land 1000-3499	2028	79	Hogs 2000-6999	215	92
Capacity 1-9999	3073	189	All land 3500+	57	57
Capacity 10000-49999	3567	290	Capacity 150000+	90	90
Capacity 50000-149999	741	99	Cattle 1000+	32	8
Sheep 1-29	1120	94	COF* 1000+	36	9
Cattle 100-199	2294	185	Dairy 200+	19	19
Cattle 200-499	982	91	Sheep 500+	171	42
Dairy 50-199	1403	140	Hogs 7000+	19	19
Cattle 500-999	172	24	HPLA* 3000+	115	28

*ABBREVIATIONS: COF: cattle on feed; CRD: crop reporting district; HPLA: hens and pullets of laying age.

Table 2. Standard Deviations and CV's:
Current Stratification vs. Simple Random Sampling

Variable	Standard Deviation		Coefficient of Variation	
	Stratified	SRS	Stratified	SRS
Arizona				
Wheat	5.17	13.79	.12	.32
Cotton	8.23	22.22	.06	.16
Barley	1.72	2.44	.11	.16
Hay	3.94	5.40	.09	.14
Cattle	13.47	88.02	.04	.27
Illinois				
Corn	2.20	2.76	.02	.02
Soybeans	1.94	2.21	.02	.02
Wheat	.69	.70	.04	.04
Hay	.44	.48	.04	.04
Hogs	1.32	3.89	.02	.07
Cattle	1.33	1.31	.06	.06
Corn Stocks	43.38	57.03	.04	.05
Soybean Stocks	30.04	27.64	.10	.09
Dairy Cattle	.14	.20	.06	.09
Tennessee				
Corn	.38	.58	.06	.09
Soybeans	1.10	1.54	.08	.12
Cotton	.50	.53	.20	.21
Tobacco	.03	.03	.06	.06
Hay	.45	.53	.04	.05
Cattle	.63	.93	.03	.05
Dairy Cattle	.18	.30	.07	.12

3. The Proposed 1985 Stratification

Stratification has two purposes: one is to increase sampling efficiency (ie., lower the variances of the estimates), the second is to guarantee sufficient data on rare items to make projects or estimates and to do analyses. The more complicated the design is, however, the greater the opportunity for non-sampling error and the more work it is to carry out the survey. In redesigning the stratification, we attempted to simplify it as much as possible, while maintaining both functions of increasing efficiency and ensuring the inclusion of rare items.

In each state, we restricted our attention to a list of important variables. This list of variables includes all major crops and items from regular agricultural surveys carried out in the state. These variables were then ordered by size of CV under simple random sampling. (These values are given in Table 2.) Regression models were fit to the variables with the largest CV's to determine the most appropriate stratification variables and boundaries were chosen using the "cumulative sqrt(f)" method. (Details are given in section 6.) The resulting stratification consists of a two- or three-way table; ie., with classifications being made on the basis of two or three variables.

Initially this basic method was applied in all states to the small and intermediate sized strata. The largest strata of the current stratification --the so-called extreme operators (EO's)--were generally left untouched, with the exception that any EO stratum for a variable not on the variable list was deleted. (For example, the EO hog stratum in Arizona was eliminated because, since Arizona does not participate in the quarterly hog survey, this was not deemed to be an important variable.)

There were additional modifications of this approach for each state. In Arizona cattle and cropland were the initial stratification variables, but the upper cattle strata were later collapsed over cropland, since these strata were very sparse. In Illinois, three variables were used: land, dairy cattle and hogs. In Tennessee the initial variables were dairy cattle and cropland, with the latter collapsed in the upper strata; this, however, was less effective than the current stratification and crop reporting district was introduced to reduce the variance of the lowest stratum.

The stratifications for each state are presented in Tables 3A-3C. The sample sizes given here result from post-stratifying the 1984 survey data. (The post-stratification was done using the control data for the elements of the 1984 survey sample.) As previously mentioned, the number of strata, when compared with the current stratification, is reduced by about 40% for each state.

Table 4 gives the estimated standard deviations and CV's under the current and proposed stratifications*. For each state the comparison is about the same: there are slight gains or losses for one or two variables but generally the level of sampling efficiency is nearly identical. As mentioned above, the sample "allocation" is the result of post-stratification and is not necessarily optimal; we will see, however, that it differs little from the optimal allocations which will be presented later.

*Variance calculations are discussed in section 6.3.

As will be apparent, we have used the population control data extensively throughout this project. This information is stored on computer tape and updated frequently. We obtained copies of the versions that were as close as possible to those which were used to classify the 1984 survey population into strata. There were however, some slight discrepancies and these will be evident from time to time. One example is the very slight difference in total population and sample sizes between Tables 1A-C and 3A-C.

At this point we will mention one error which was discovered after all the analysis for this report had been completed. Through a programming error, while reclassifying the list population and the 1984 survey data, the members of stratum 91 were misclassified into various other strata, primarily stratum 63. The effect of this error is negligble, but should be mentioned. The stratum weights used for much of this analysis were, for example, very slightly affected (by .36%, .2%, and .01%,for the three strata); also the estimated variance for stratum 63 may be slightly inflated, causing a very slight over-allocation.

Table 3A. Proposed Stratification: Arizona

Code	Stratum Description	Population Size	Sample Size
50	Cattle 0-250 and Cropland 1-430	1472	210
61	Cattle 0-250 and Cropland 431-1248	369	96
62	Cattle 0-250 and Cropland 1249-4999	145	53
70	Cattle 251-799	423	93
80	Cattle 800-3999	160	45
91	Cropland 5000+	25	25
92	Cattle 4000+	4	4
93	COF* 300+	25	25
95	Sheep 1+	73	37
99	HPLA* 1600+	5	2

Table 3B. Proposed Stratification: Tennessee

Code	Stratum Description	Population Size	Sample Size
51	Cropland 1-49 Dairy Cattle 1-4 and CRD 10 or 20	12263	433
52	Cropland 1-49 Dairy Cattle 1-4 and CRD 30,40 or 50	39838	1182
53	Cropland 1-49 Dairy Cattle 0-4 and CRD 60	27900	578
61	Cropland 50-280 and Dairy Cattle 0-4	5768	201
62	Cropland 280-1999 and Dairy Cattle 1-4	1109	96
70	Dairy Cattle 5-49	5608	187
80	Dairy Cattle 50-499	1629	113
90	Cropland 2000+	36	36
91	Cattle 1500+	7	7
93	Dairy 500+	4	4
94	Sheep 40+	65	16
99	HPLA* 3000+	26	7

Table 3C. Proposed Stratification: Illinois

Code	Stratum Description	Population Size	Sample Size
50	All Land 1-434 Dairy Cattle 0-22 and Hogs 0-279	58106	2138
61	All Land 1-434 Dairy Cattle 0-22 and Hogs 280-6999	2390	283
62	All Land 1-434 Dairy Cattle 23-199 and Hogs 0-279	2440	171
63	All Land 435-3499 Dairy Cattle 0-22 and Hogs 0-279	17769	950
71	All Land 1-434 Dairy Cattle 23-199 and Hogs 280-6999	87	10
72	All Land 435-3499 Dairy Cattle 0-22 and Hogs 280-6999	2806	352
73	All Land 435-3499 Dairy Cattle 23-199 and Hogs 0-279	690	66
80	All Land 435-3499 Dairy Cattle 23-199 and Hogs 280-6999	96	13
90	All land 3500+	57	57
91	Capacity 150000+	90	90
92	Cattle 1000+	32	8
93	COF* 1000+	36	9
94	Dairy 200+	19	19
95	Sheep 500+	171	42
96	Hogs 7000+	19	19
99	HPLA* 3000+	115	28

Table 4. Standard Deviations and CV's:
Proposed Stratification vs. Current Stratification

Variable	Standard Deviation		Coefficient of Variation	
	Proposed	Current	Proposed	Current
Arizona				
Wheat	4.48	5.17	.10	.12
Cotton	7.81	8.23	.06	.06
Barley	1.73	1.72	.11	.11
Hay	4.10	3.94	.10	.09
Cattle	11.65	13.47	.04	.04
Illinois				
Variable	Standard Deviation		Coefficient of Variation	
	Proposed	Current	Proposed	Current
Corn	2.06	2.20	.02	.02
Soybeans	1.73	1.94	.02	.02
Wheat	.63	.69	.04	.04
Hay	.42	.44	.04	.04
Hogs	1.81	1.32	.03	.02
Cattle	1.28	1.33	.06	.06
Corn Stocks	45.85	43.38	.04	.04
Soybean Stocks	24.82	30.04	.08	.10
Dairy cattle	.11	.14	.05	.06
Tennessee				
Variable	Standard Deviation		Coefficient of Variation	
	Proposed	Current	Proposed	Current
Corn	.45	.38	.07	.06
Soybeans	1.00	1.10	.07	.08
Cotton	.51	.50	.20	.20
Tobacco	.03	.03	.06	.06
Hay	.46	.45	.04	.04
Cattle	.72	.63	.03	.03
Dairy cattle	.17	.18	.07	.07

4. Sample Allocation

The ISP/JES, as we have noted, is a multiple frame survey and the cost of sampling is much higher from the area frame than from the list frame--not surprisingly, since virtually all area frame enumeration involves personal interviewing. Given these costs, and a required level of accuracy for the survey results, the goal of sample allocation is to distribute the sample among the strata in such a way as to minimize the cost of the survey. To do this we have used nonlinear programming techniques, a discussion of which is contained in section 6. For several reasons, the minimization program was not applied to all strata. In the list frame, the EO strata were fixed, as they were in the stratification design, and most of the nonagricultural strata in the area frame were fixed as well. In both cases it was felt that these parts of the sample design had evolved to meet contingencies of a degree of sublety which go far beyond the simple requirements of a nonlinear programming model. This was especially apparent in the latter situation, where estimates of variances (which are typically either very large or very small) are not realistic. In Arizona--because of the structure of the area frame stratification--this amounted to ignoring a large part of the design, in dollar terms about 32%. In Illinois the fixed part of the design accounted for 9% of the cost and in Tennessee it accounted for only 4%.

The cost model we employed is linear and somewhat elementary. Table 5 gives the cost per sampling unit for the list and area frames for each state. This simple structure does not allow for variable costs between strata. It seems realistic to assume that these differences would be negligible for the list frame and in the intensive agricultural strata of the area frame; since most of the nonagricultural strata were not included in the minimization program it seems safe to assume that the model is reasonably accurate. (More detailed cost information appears in section 7.)

Table 5. ISP/JES Cost Per Sampling Unit.

Arizona		Illinois		Tennessee	
List	Area	List	Area	List	Area
16.58	66.57	5.97	138.84	5.47	96.70

Table 6. Current Multiple Frame Coefficients of Variation.

Arizona		Illinois		Tennessee	
Variable	CV	Variable	CV	Variable	CV
Wheat	.083	Corn	.022	Corn	.075
Cotton	.052	Soybeans	.023	Soybeans	.064
Barley	.125	Wheat	.043	Cotton	.174
Hay	.095	Hay	.046	Tobacco	.099
Cattle	.053	Hogs	.082	Hay	.050
		Cattle	.099	Cattle	.038
		Corn Stocks	.043	Dairy Cattle	.079
		Soybean Stocks	.080		
		Dairy Cattle	.097		

Our first step in attempting to find an adequate allocation was to require that the CV's obtained by using the proposed stratification be no larger than those from the current stratification. The current multiple frame CV's are given in Table 6.* The optimal allocations to obtain those CV's are given in Tables 7A-7C, along with the current multiple frame allocations. These tables allow us to compare the effects of the two stratification schemes. While there are some striking differences, the two allocations are generally quite similar. If the list frame allocation had dropped significantly while the area frame allocation remained constant, this would have indicated a substantial increase in efficiency over the current stratification. However, the similarity between the two allocations reinforces our assertion that the two designs are quite similar with respect to sampling efficiency.

The costs of the current and proposed allocations are compared in Table 8. The costs for the latter are slightly smaller, but this is to be expected, since the allocations were chosen to minimize cost. In Arizona the allocation to the list is about the same for both the current and proposed designs while the area frame allocation drops slightly for the latter; in Illinois there is a shift from the list frame to the area frame under the proposed stratification, while in Tennessee there is a somewhat more marked shift in the opposite direction. While we point out these differences, they do not seem pronounced enough to warrant any interpretation or conclusion, other than that the designs are similar but not identical.

Our next step was to obtain allocations for a range of uniform CV restrictions. We wished to see, for example, what allocations would result from requiring that all CV's be no larger than, say, .10. Some variables will still have small CV's, while others will have CV's of exactly .10--depending on the size of the population CV's and the relationships among the variables themselves. The allocations are given in Tables 9A-9C, with the accompanying CV and cost information in Tables 10A-10C.

What is perhaps most striking in these tables is the rate at which the cost accelerates as the CV requirements are tightened. It is also interesting to note that in each state there are several variables which "drive down" the others as their variances become smaller; an example of this would be barley and hay in Arizona: as the CV's for these variables are forced down, the CV's for other variables decrease at a steady rate, though they are well below the actual constraint.

Another reason for this analysis was to see which strata drew the largest allocations as more accuracy was demanded. For example, comparing the Illinois optimal allocation for current CV levels with the one with all CV's less than .08, we see that the allocation to stratum 11 in the area frame increases from 177 to 318. This increase alone accounts for the difference between the survey costs. Generally speaking, the largest increases are in the allocations to the heavily agricultural area frame strata (strata 11, 12, 13 and 14).

---

*In the area frame strata, variances for closed estimates were used for all variables except for grain stocks, for which weighted estimates were used.

After consideration of the variables involved and some experimentation, we arrived at suggested allocations for each state, which are given in Tables 11A-11C. We attempted to develope allocations for each state which cost slightly less than the current survey but which offer slight improvement on most variables.

It is important to keep in mind that these allocations are for the required sample sizes. Since there will be refusals, inaccessibles and farmers who have gone out of business, larger samples must be drawn to obtain these target figures. The cost estimates will not be affected, since these were based on cost per final sampling unit.

Table 7A. Arizona: Allocation to Proposed Strata for Current CV Levels.					
List Frame			Area Frame		
Stratum	Optimal*	Current	Stratum	Optimal*	Current
50	233	210	13	98	120
61	96	96	14	24	20
62	70	53	20	49	40
70	65	93	21	2	5
80	40	45	31	36	40
91	25	25	32	15	15
92	4	4	41	24	24
93	25	25	44	15	15
95	37	37	45	15	15
99	2	2	46	15	15
			47	15	15
			48	15	15
			49	15	15
			50	20	20

*Strata 91-99 (List Frame) and 32-50 (Area Frame) were fixed and not entered in the optimization program.

Table 7B. Illinois: Allocation to Proposed Strata for Current CV Levels.					
List Frame			Area Frame		
Stratum	Optimal*	Current	Stratum	Optimal*	Current
50	1943	2138	11	177	170
61	125	283	12	53	50
62	122	171	20	36	40
63	1163	1014	31	20	20
71	15	10	32	10	10
72	237	359	33	2	2
73	118	69	40	6	6
80	9	13	61	2	2
90	57	57			
91	90	90			
92	8	8			
93	9	9			
94	19	19			
95	42	42			
96	19	19			
99	28	28			

*Strata 90-99 (List Frame) and 32-61 (Area Frame) were fixed and not entered in the optimization program.

Table 7C. Tennessee: Allocation to Proposed Strata for Current CV Levels.

List Frame			Area Frame		
Stratum	Optimal*	Current	Stratum	Optimal*	Current
51	398	433	13	91	90
52	1238	1182	20	109	120
53	578	578	31	44	60
61	228	201	32	15	15
62	131	96	33	2	2
70	286	187	40	61	60
80	191	113	50	2	2
90	36	36			
91	7	7			
93	4	4			
94	16	16			
99	7	7			

*Strata 90-99 (List Frame) and 32, 33, and 50 (Area Frame) were fixed and not entered in the optimization program.

Table 8. Comparison of Costs to Obtain Current CV's. Current

State		List Frame	Area Frame	Total Cost
Arizona	Current	9782	24897	34679
	Proposed	9898	23832	33730
Illinois	Current	25844	41652	67496
	Proposed	23844	42485	66329
Tennessee	Current	15644	33748	49392
	Proposed	17066	31331	48397

Table 9A. Optimal Allocations for Various CV Constraints.
Arizona

Stratum	CV < .11	CV < .10	CV < .09
List Frame			
50	203	240	287
61	116	136	161
62	81	95	112
70	33	40	49
80	13	15	18
91	25	25	25
92	4	4	4
93	25	25	25
95	37	37	37
99	2	2	2
Area Frame			
13	112	132	156
14	44	51	60
20	26	32	39
21	3	4	4
31	21	25	30
32	15	15	15
41	24	24	24
44	15	15	15
45	15	15	15
46	15	15	15
47	15	15	15
48	15	15	15
49	15	15	15
50	20	20	20

Table 10A. CV's for Table 9A Allocations.
Arizona

Variable	CV < .11	CV < .10	CV < .09
Wheat	.081	.073	.065
Cotton	.049	.045	.040
Barley	.110	.100	.090
Hay	.110	.100	.090
Cattle	.063	.058	.052
Cost	32569	36425	41095

Table 9B. Optimal Allocations for Various CV Constraints. Illinois

Stratum	CV < .10	CV < .09	CV < .08
List Frame			
50	1421	1750	2207
61	49	60	76
62	158	195	245
63	596	729	913
71	4	5	6
72	115	141	178
73	55	67	85
80	6	7	9
90	57	57	57
91	90	90	90
92	8	8	8
93	9	9	9
94	19	19	19
95	42	42	42
96	19	19	19
99	28	28	28
Area Frame			
11	205	229	318
12	17	21	26
20	19	24	30
31	20	20	20
32	10	10	10
33	2	2	2
40	6	6	6
61	2	2	2

Table 10B. CV's for Table 9B Allocations. Illinois

Variable	CV < .10	CV < .09	CV < .08
Corn	.028	.025	.022
Soybeans	.030	.027	.024
Wheat	.059	.053	.047
Hay	.059	.053	.047
Hogs	.099	.089	.080
Cattle	.100	.090	.080
Corn Stocks	.056	.051	.045
Soybean Stocks	.100	.090	.080
Dairy cattle	.100	.090	.080
Cost	54990	62855	81308

Table 9C. Optimal Allocations for Various CV Constraints.

Tennessee

Stratum	CV < .12	CV < .11	CV < .10
List Frame			
51	1580	1851	2200
52	253	300	364
53	137	162	196
61	503	592	703
62	346	407	484
70	100	117	143
80	53	62	78
90	36	36	36
91	7	7	7
93	4	4	4
94	16	16	16
99	7	7	7
Area Frame			
13	103	121	144
20	73	86	102
31	10	12	15
32	15	15	15
33	2	2	2
40	54	64	77
50	2	2	2

Table 10C. CV's for Table 9C Allocations.

Tennessee

Variable	CV < .12	CV < .11	CV < .10
Corn	.120	.110	.099
Soybeans	.074	.068	.062
Cotton	.120	.110	.100
Tobacco	.120	.110	.100
Hay	.072	.067	.060
Cattle	.058	.053	.048
Dairy Cattle	.120	.110	.099
Cost	41849	48874	57933

Table 11A. Arizona: Suggested Allocation.			
List Frame		Area Frame	
Stratum	Sample Size	Stratum	Sample Size
50	254	13	108
61	120	14	40
62	85	20	24
70	35	21	3
80	27	31	19
91	25	32	15
92	4	41	24
93	25	44	15
95	37	45	15
99	2	46	15
		47	15
		48	15
		49	15
		50	20
	List Frame	Area Frame	Total
Cost:	10180	22834	33014
Variable	CV		
Wheat	.075		
Cotton	.048		
Barley	.110		
Hay	.110		
Cattle	.054		

Table 11B. Tennessee: Suggested Allocation.			
List Frame		Area Frame	
Stratum	Sample Size	Stratum	Sample Size
51	691	13	104
52	1060	20	103
53	565	31	40
61	304	32	15
62	162	33	2
70	252	40	61
80	126	50	2
90	36		
91	7		
93	4		
94	16		
99	7		
	List Frame	Area Frame	Total
Cost	17668	31621	49289
Variable	CV		
Corn	.075		
Soybeans	.059		
Cotton	.145		
Tobacco	.099		
Hay	.050		
Cattle	.038		
Dairy Cattle	.077		

Table 11C. Illinois: Suggested Allocation.

List Frame		Area Frame	
Stratum	Sample Size	Stratum	Sample Size
50	1465	11	231
61	121	12	37
62	177	20	21
63	1004	31	20
71	6	32	10
72	221	33	2
73	61	40	6
80	9	61	2
90	57		
91	90		
92	8		
93	9		
94	19		
95	42		
96	19		
99	28		

	List Frame	Area Frame	Total
Cost	19916	45678	65594

Variable	CV
Corn	.024
Soybeans	.025
Wheat	.049
Hay	.052
Hogs	.080
Cattle	.093
Corn Stocks	.045
Soybean Stocks	.080
Dairy Cattle	.093

5. Conclusion

In this paper we have introduced and evaluated a new list frame stratification. The new stratification is simpler, with about 40% fewer strata, and it appears to be as efficient as the one currently in use.

Using cost and variance information from the 1984 JES we have explored various sample allocations and recommended one for each state. These allocations do not depart radically from the design that was used last year; in each case, the cost is slightly lower than that of last year's survey, while the CV levels are the same or slightly improved; the balance between the list and area frames is also about the same.

While presenting the results of this study, we have attempted to outline a general approach to the problem of multivariate sample design. We hope this will be helpful to others as the ISP expands beyond the research stage.

6. Appendix

This section contains technical details on several topics: stratification methods (6.1), allocation method (6.2), variance calulations (6.3), variables used (6.4), and cost information (6.5).

6.1 Stratification Method

In principle the method of stratification is straightforward: first we decided which variables have the worst population CV's and what two or three stratification variables predict them the best. Once the stratification variables were chosen their ranges were broken into two or three ordinal categories and the population was then stratified by sorting it into all possible combinations of these categories. In application, as we shall see, it was necessary to deviate at least slightly from this format in two of the three states. (See Cochran, 1977, pp. 123-133, and Kish and Anderson, 1978, for technical discussions of these methods.)

As we have noted, attention was restricted to only certain variables in each state. (See section 6.4 for a more detailed discussion) Table A1 (adapted from Table 2) gives the CV's for all of these variables. In Arizona, all of the variables have large CV's, but cattle and wheat are the worst; cotton, since it an important crop in Arizona, also deserves special attention. In Illinois dairy cattle, soybean stocks and hogs are problematic variables; dairy cattle, cotton and soybeans stand out in Tennessee.

The next step was to determine the best stratification variables. The object here is to control the variables with the largest CV's, so, wherever possible, these variables themselves were used as stratifiers. For crops and grain stocks, regression models were used to determine which of the control variables predicted the dependent variable with the largest R-square. Table A2 shows the availability of control information for the three states. For each potential stratification variable, this table shows how many records have and how many are missing an entry for that variable.

In Arizona, cattle is available on 63% of the records. While this is not as high as might be hoped, it is better than any other variables in Arizona and, since cattle has the highest CV in Arizona, this was used as one of the stratification variables. Wheat is present on only 14% of the records, but a regression model showed that it was predicted rather well (R-square=.81) by cotton. Unfortunately, cotton is present on only 29% of the records. It was decided to try cropland as the second stratification variable despite its poor performance in predicting wheat and cotton (R-square negligible in each case). This decision was based on the fact that it appeared to work well in the current stratification; in fact, both cotton and cropland were tried--at different times--and cropland seemed to be slightly superior.

For Illinois, the control information for hogs and for dairy cattle is fairly complete and since both of these variables have high CV's these were used as stratification variables. In running stepwise regressions to predict soybean stocks, the best single control variable was cropland (R-square=.08--not overwhelming, but the best nonetheless). Cropland, however, is missing from 77% of the population records; since "all land" is missing from less than

1% of the records it was used instead and, judging from the slight improvement in soybean stocks, the substitution seems to have been acceptable. Thus we arrive at using dairy cattle, hogs and all land for stratification variables in Illinois.

For Tennessee, we initially tried using dairy cattle and cropland as stratification variables. The former was was chosen because control information is relatively complete; the latter predicts soybeans very well (R-square=.86) and is better at predicting cotton (R-square=.10) than anything besides soybeans (R-square=.14). This stratification design was not quite satisfactory, however, since some of the CV's were slightly above those for the current stratification, and crop reporting district was added as a stratification variable, again because it seemed to perform well in the current stratification.

Having chosen the stratification variables, it remained to set the stratum boundaries. These were set separately for each variable using the "cumulative sqrt(f)" procedure, as described in Cochran (1977, pp. 127-130). In essence, the procedure consists of constructing a frequency table and, from this, a table giving the cumulative of the square roots of the frequencies. The boundaries are constructed so that the cumulative sqrt(f) scale is divided into equal parts. In constructing the frequency tables we used 100 cells, divided into equal increments, beginning at 0 and ending with the cut-off of the appropriate EO stratum. For purposes of illustration, reduced versions (with 10 cells) are presented in Table A3.

Applying this method to Arizona, for example, the cattle stratum boundaries are chosen so that the cells are 225.2/3 = 75 units wide on the cumulative sqrt(f) scale. The is accomplished if the actual boundaries are (with rounding) 250 and 800. All other boundaries are calculated in the same manner and, for convenience, these are given in Table A4.

Once the boundaries were set, the population was sorted into the resulting categories. This was done in such a way that any records without control information were sorted into the lowest stratum.

As we noted earlier, the Arizona strata of large cattle operators with large amounts of cropland were quite small, the largest being about .5% of the population. These strata were collapsed to create the proposed Arizona stratification.

The large dairy strata in Tennessee were also collapsed over cropland values. Here the motivation was less the sparseness of the strata (although altogether these four strata comprised only about 2% of the population) than the fact that the presence of these strata contributed quite negligible reductions in variance. It turned out to be more effective to split the small-cropland/small-dairy stratum into crop reporting districts.

Table A1. Estimated Coefficients of Variation.					
Arizona		Illinois		Tennessee	
Variable	CV	Variable	CV	Variable	CV
Wheat	.32	Corn	.02	Corn	.09
Cotton	.16	Soybeans	.02	Soybeans	.12
Barley	.16	Wheat	.04	Cotton	.21
Hay	.14	Hay	.04	Tobacco	.06
Cattle	.27	Hogs	.07	Hay	.05
		Cattle	.06	Cattle	.05
		Corn Stocks	.05	Dairy Cattle	.12
		Soybean Stocks	.09		
		Dairy Cattle	.09		

Table A2. Control Information Availability.						
	Arizona		Illinois		Tennessee	
Variable	Present	Missing	Present	Missing	Present	Missing
Land	1376	1327	84443	480	35403	58854
Barley	0	2703				
Cropland	1376	1327	19390	65533	21940	72317
Capacity	0	2703	13675	71248	1465	92792
Corn			19069	65854	4055	90202
Cotton	779	1924			549	93708
Hay	548	2155	0	84923	5502	88755
Soybean			19068	65855	3109	91148
Wheat	384	2319	19068	65855		
Tobacco					5337	88920
Cattle	1694	1009	84080	843	77055	17202
Dairy	289	2414	83819	1104	78075	16182
COF	25	2678	78382	6595	0	94257
Hogs	115	2588	83289	1634	79404	14853
Sheep	73	2630	83250	1673	259	93998
HPLA	113	2590	35779	49144	1045	93212

Table A3.1 Cell Frequencies and Cumulative Sqrt(f): Arizona				
	Cattle		Cropland	
Cell	Boundary	Cum.sqrt.	Boundary	Cum.sqrt.
10	400	101.4	500	85.6
20	800	149.4	1000	135.8
30	1200	177.8	1500	166.3
40	1600	196.0	2000	187.2
50	2000	204.7	2500	203.1
60	2400	211.4	3000	208.8
70	2800	216.4	3500	216.3
80	3200	219.8	4000	220.8
90	3600	223.2	4500	227.1
100	4000	225.2	5000	229.8

Table A3.2 Cell Frequencies and Cumulative Sqrt(f): Tennessee				
	Dairy Cattle		Cropland	
Cell	Boundary	Cum.sqrt.	Boundary	Cum.sqrt.
10	50	448.0	200	333.3
20	100	553.5	400	413.7
30	150	599.2	600	456.8
40	200	629.6	800	487.9
50	250	648.4	1000	512.3
60	300	659.9	1200	529.4
70	350	663.6	1400	540.7
80	400	666.0	1600	551.3
90	450	670.8	1800	559.2
100	500	671.8	2000	564.1

Table A3.3 Cell Frequencies and Cumulative Sqrt(f): Illinois						
	Dairy Cattle		Land		Hogs	
Cell	Boundary	Cum.sqrt.	Boundary	Cum.sqrt.	Boundary	Cum.sqrt.
10	20	372.5	350	732.5	700	549.9
20	40	484.2	700	116.1	1400	659.1
30	60	586.1	1050	1400.8	2100	712.5
40	80	659.8	1400	1531.3	2800	742.1
50	100	706.7	1750	1614.0	3500	762.4
60	120	734.5	2100	1666.5	4200	774.4
70	140	754.3	2450	1698.5	4900	785.3
80	160	767.2	2800	1722.7	5600	791.1
90	180	775.1	3150	1740.7	6300	796.1
100	200	779.0	3500	1754.6	7000	798.1

Table A4. Proposed Strata Boundaries.						
Arizona		Tennessee		Illinois		
Cattle	Cropland	Dairy Cattle	Cropland	Dairy Cattle	Hogs	Land
250	430	5	49	22	279	434
800	1250	50	129	200	6000	3500
4000	5000	500	2000			

6.2 Allocation method

For variable j, let

$\hat{Y}_j^{(L)}$ = total for farms on the list frame

$\hat{Y}_j^{(A)}$ = total for farms on the area (but not list) frame.

Since these estimates are independent, we have

$$\begin{aligned} var(\hat{Y}_j) &= var(\hat{Y}_j^{(L)} + \hat{Y}_j^{(A)}) \\ &= var(\hat{Y}_j^{(L)}) + var(\hat{Y}_j^{(A)}) \\ &= \sum_{h=1}^{L_1} w_h^2 \, var(Y_{hj}) + \sum_{h=L_1+1}^{L} w_h^2 \, var(Y_{hj}) \\ &= \sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h - 1} \frac{S_{hj}^2}{n_h} \\ &\approx \sum_{h=1}^{L} N_h \frac{N_h - n_h}{n_h} S_{hj}^2 . \end{aligned}$$

Here L_1 denotes the number of list strata and L the total of all strata on both frames. Otherwise the notation is that conventionally used in survey sampling literature, eg., Cochran (1977).

For any variance constraint b_j we have

$$CV_j \leq b_j$$

if and only if

$$Y_j^{-2} \sum_{h=1}^{L} N_h^2 \frac{S_{hj}^2}{n_h} \leq b_j^2 + Y_j^{-2} \sum_{h=1}^{L} N_h \, S_{hj}^2 .$$

We used this cost function:

$$C = a_0 + \sum_{h}^{L} a_h \, n_h .$$

In this model a_0 represents overhead and a_h is the cost per sampling unit. Taking

$$x_h = n_h^{-1}$$

we wish to minimize the function

$$C = a_0 + \sum_{h}^{L} \frac{a_h}{x_h}$$

subject to the constraints

$$N_h^{-1} \leq x_h \leq 1, \quad \text{for } 1 \leq h \leq L$$

$$\sum_{h}^{L} c_{hj} \, x_h \leq d_j, \quad \text{for } 1 \leq j \leq P$$

where

P = number of variables

$$c_{hj} = Y_j^{-2} \sum_{h=1}^{L} N_h^2 \, S_{hj}^2$$

$$d_j = b_j^2 + Y_j^{-2} \sum_{h=1}^{L} N_h \, S_{hj}^2 .$$

To accomplish this, we used a search procedure adapted from methods suggested by Kokan and Khan (1967). A forthcoming article will provide the details of this technique.

6.3 Variance calculations

To estimate the variances under simple random sampling (eg., for Table 2) we used

$$V = \frac{N-n}{n(N-1)} \left(\frac{1}{N} \sum_h^L \frac{N_h}{n_h} \sum_j y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right)$$

(see Cochran, 1977, p. 136). (The notation is that conventionally used is survey sampling literature.)

To estimate the variances under the proposed stratification we first sorted the survey data into the new strata using control data, then this formula was applied with in each new stratum. Here the summation is over the intersections with the current strata. That is, to estimate the variance of the kth stratum we calculated

$$V_k = \frac{N_k - n_k}{n_k(N_k-1)} \left(\frac{1}{N_k} \sum_h^L \frac{N_{kh}}{n_{kh}} \sum_j y_{khj}^2 - \bar{y}_{k,st}^2 + v(\bar{y}_{k,st}) \right)$$

where

N_k = population number in stratum k

n_k = post-stratified sample number in stratum k

N_{kh} = population number in both strata k and h

n_{kh} = sample number in both strata k and h

y_{khj} = (khj)th element the of intersection of strata k and h

$\bar{y}_{k,st} = \sum_h^L \frac{N_{kh}}{N_k} \sum_j \frac{y_{khj}}{n_{kh}}$ = stratified sample mean of stratum k

$v(\bar{y}_{k,st}) = \sum_h^L \left[\frac{N_{kh}}{N_k}\right]^2 \frac{S_{kh}^2}{n_{kh}}$ = estimated variance of $\bar{y}_{k,st}$.

While it might be protested that this estimate ignores the randomness of the sample sizes (due to post-stratification), in fact Holt and Smith (1979) argue strongly for this kind of conditional approach. They make a distinction between planning (ie., planning to post-stratify the sample) and analysis: unconditional estimates are appropriate for the former and conditional estimates are more suitable for the latter. Here we are post-stratifying for the purpose of analysis; in the survey we are planning we will use bona fide stratification.

Once the stratum variances were estimated, the stratified variance estimate was compiled in the usual way.

The area frame non-overlap variances were calculated by simply removing the overlap operators (ie., operators who also appear on the list frame) and then forming the usual variance estimates. As we have noted, variances based on closed estimates were used for all commodities except grain stocks, where weighted estimates were used.

6.4 Variable descriptions

As we have noted, a list of "important" variables was selected for each state. An effort was made to include all major crops, although no explicit formula for inclusion was delineated. For commodities subject to periodic surveys (eg., grain stocks, cattle, hogs, dairy cattle) these variables appear on a state's list if and only if that state participates in the survey. Some variables have been aggregated (wheat in Arizona, tobacco in Tennessee), in other cases a single variable represents a series of more detailed items (eg., hogs or cattle). It has perhaps been noted that both sheep and HPLA (hens and pullets of laying age) are represented in the EO strata but are not on any of the states' lists. In the former case sufficient information was not available to make a detailed study of the variable, while in the latter case the structure of the population is such that a single EO stratum suffices for estimation purposes; in both cases all states participate in the relevant periodic surveys.

Table A5 gives a list of all variables used, along with the position on the GE strung record. For aggregated variables the items in the "Position" column indicate how the variable was created.

Table A5. Variable Descriptions.

Variable	Position	Comments
Barley	P165	Planted for all purposes on entire farm
	P535	Planted for all purposes on tract
Cattle	P350	Total number on farm
	P250	Total number on tract
Corn	P167	Planted for all purposes on entire farm
	P530	Planted for all purposes on tract
Corn stocks	P121	Bushels stored on entire farm
Cotton	P171	Upland cotton on entire farm
	P524	Upland cotton on tract
Dairy cattle	P352	Number of milk cows on entire farm
	P252	Number of milk cows on tract
Hay	P184+P185+P186	Any hay on entire farm
	P653+P654+P656	Any hay on tract
Hogs	P300	Total number on farm
	P300	Total number on tract
Soybeans	P180	Planted for all purposes on entire farm
	P600	Planted for all purposes on tract
Soybean stocks	P125	Bushels stored on entire farm
Tobacco	P187+P189+P188	Any tobacco on entire farm
	P670	Any tobacco on tract
Wheat	P161+P162	Arizona: Any wheat planted on entire farm
	P553+P554	Arizona: Any wheat planted on tract
	P174	Illinois: Winter wheat planted on entire farm
	P540	Illinois: Winter wheat planted on tract

6.5 Cost information

The cost information used in allocating the sample is given in Table A6. It was used in a very straightforward way: all area frame sampling costs were aggregated, all list frame sampling costs were aggregated, and then the training and quality control costs were halved and added to each total. Finally these were divided by the appropriate number of reporting units in the sample (number of segments for the area frame, number of useable interviews for the list frame). Thus, for example, the Arizona area frame cost per segment is

$$\frac{8761 + 10749 + (10028/2)}{374} = 65.57.$$

It was felt that all of these costs are variable (as opposed to fixed) costs. While it might be argued that training is a fixed cost, many of these expenditures (travel expenses, hotel accomodations, rental of meetings facilities) fluctuate with the number of interviewers, which of course depends on the sample size.

Table A6. Cost Information.*

	Arizona	Illinois	Tennessee
Area Frame			
Between segment	8761	10054	13043
Within segment	10749	22914	15983
List Frame			
Telephone Interview	796	8682	3985
Personal Interview	3414	8018	6948
Training and quality control	10028	17366	9443

*This information was compiled by Douglas Kleweno.

6.6 Variance Information

Tables A7-A10 give within-stratum standard deviations and totals used in the allocation program. These are given only for the strata that were used in the optimization program.

Table A7. Totals and Within-Stratum Standard Deviations.

Arizona

Stratum	Wheat	Cotton	Barley	Hay	Cattle
List Frame					
50	92	135	28	79	82
61	134	267	71	123	80
62	250	605	126	222	10
70	9	17	6	76	294
80	34	13	14	58	1050
Area Frame					
13	30	81	23	42	98
14	1	95	75	16	631
20	1	20	1	60	17
21	1	37	16	15	14
31	1	1	1	9	18
Total (000)	147	485	62	174	983

Table A8. Totals and Within-Stratum Standard Deviations.

Tennessee

Stratum	Corn	Soybean	Cotton	Tobacco	Hay	Cattle	Dairy Cattle
List Frame							
51	27	110	68	1	16	32	1
52	24	31	1	1	23	36	6
53	6	11	1	1	20	27	1
61	20	46	46	2	38	47	3
62	86	320	165	6	58	136	7
70	52	94	7	2	35	50	22
80	80	95	1	3	65	145	70
Area Frame							
13	25	44	16	1	25	27	7
20	8	32	8	1	22	32	4
31	14	9	1	1	10	13	3
40	6	3	1	4	16	19	1
Total (000)	846	1984	314	84	1554	2740	260

Table A9. Totals and Within-Stratum Standard Deviations.
Illinois

Stratum	Corn	Soybean	Wheat	Hay	Hogs
List					
50	80	75	27	22	59
61	111	86	28	22	480
62	58	17	12	32	556
63	95	74	45	38	43
71	242	195	72	33	88
72	252	183	73	31	690
73	211	148	65	152	111
80	256	113	78	59	804
Area					
11	58	56	14	12	188
12	46	62	24	13	158
20	67	34	23	14	38
Total (000)	11450	9354	1849	1152	6171

Table A10. Totals and Within-Stratum Standard Deviations.
Illinois

Stratum	Cattle	Corn Stocks	Soybean Stocks	Dairy Cattle
List				
50	78	1528	543	4
61	51	3696	787	5
62	68	3057	665	23
63	59	2381	1869	35
71	73	5433	3462	6
72	124	8600	1530	3
73	98	4051	2264	41
80	91	4603	527	28
Area				
11	99	936	529	9
12	21	789	367	2
20	13	207	72	5
Total (000)	2508	105133	30427	245

REFERENCES:

Cochran, W.G. (1977). Sampling Techniques. John Wiley and Sons, New York.

Holt, D., and Smith, T.M.F. (1979). Post Stratification. J. R. Statist. Soc. A, 142, 33-46.

Kish, L., and Anderson, D. (1978). Multivariate and multipurpose stratification. Jour. Amer. Stat. Assoc., 73, 24-34.

Kokan, A.R., and Khan, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. J. R. Statist. Soc. B, 29, 115-125.