OPTIMUM STRATIFICATION AND ALLOCATION BY ARITHMETIC RESP. GEOMETRIC SEQUENCES AND ITERATIVE REFINEMENT

Hans-Dieter Heike and Wolfgang Jaspers, Technical University of Darmstadt
Hans-Dieter Heike, Institute of Statistics and Econometrics, Residenzschloß, D-64283 Darmstadt, Germany

KEY WORDS: Optimum Stratification and Allocation.

1. Introduction

Methods of segmentation and allocation are of great practical importance in conducting surveys of businesses, farms and institutions, stock-taking in enterprises and in other areas. Referring to stock-taking in enterprises we introduce a method of optimum stratification and allocation that allows to work with smaller sample sizes than traditional methods do.

Several alternatives of sample design may be used to support stock-taking, e. g. simple random sampling, stratified sampling, systematic sampling, probability proportional to size sampling and estimation without and with auxiliary informations. The general objective of sample design is to choose that sampling and estimation procedure that results in highest precision at a given cost resp. in minimum cost at a specified level of precision. Insofar a sample design that exploits additional information will be preferred in view of the enhanced precision of the estimate. Of course there are further factors to be considered when choosing the sample design for stock-taking in enterprises, f. i. the actual inventory and storing system. In this paper we take as given a traditional storing system and a fixed date inventory system. The book values of the stock accounting are serving as auxiliary information. Since this additional information can be exploited efficiently by an optimum stratification procedure and there is great variability of the article values we decided to work with a stratified sampling design.

Standard methods of optimum stratification are solving the optimization problem as a function of strata boundaries and sample allocation only. The method proposed in this paper is an extension of standard methods since it allows approximate optimization with respect to the number of strata, strata boundaries and allocation of sample size. It consequently demands a sample size significantly lower than those of traditional methods.

2. Optimum Stratification and Allocation by Arithmetic resp. Geometric Sequences and Iterative Refinement

The proposed method is basically a grid search procedure where the grid is formed by elements of sequences and iterative refinement. This approximate optimization method can be written as a constrained nonlinear program:

$$n(\bar{x}, \text{Ney}) = \frac{\left(\sum_{i=1}^{I} N_i S_i\right)^2}{N^2 \frac{D^2}{q^2_{1-\frac{\alpha}{2}}}} \Rightarrow \text{MIN!}$$

Constraints:

$a = x_0 \le x_1 \le \ldots \le x_i \le \ldots \le x_I = b$

$\frac{n_i}{N} > 0; \; \frac{n_i}{N} \le \frac{N_i}{N}; \; \sum n_i = n$

$n(\bar{x}, \text{Ney})$ sample size with optimum allocation according to Neyman

S_i^2 variance

N_i size of stratum i

$N = \sum_i N_i$

n_i sample size of stratum i

D maximum absolute sampling error of $\bar{x}$ at a specified confidence level

$q_{1-\frac{\alpha}{2}}$ quantile corresponding to the specified confidence level

The above nonlinear optimization problem cannot be solved analytically, but local extreme values can be found by numerical methods working with or without derivations. We decided to take a grid search method because the optimum solutions of numerical methods primarily depend on local characteristics of the objective function, further assumptions regarding f. i. the distribution of attributes are not necessary and costs of calculation are within reasonable limits. The dimensions of the grid are "number of strata" and "location of strata boundaries". The search is performed in two stages. In the course of the first stage the grid points are calculated by means of arithmetic or geometric sequences:

$$a_k = a_1 + (k-1)\cdot d \quad \text{resp.}$$
$$a_k = a_1 \cdot q^{k-1}$$

The single steps within the first stage are as follows:

1. Files of a_1 and d resp. q are defined.
2. For each combination of a_1 and d resp. q a sequence is calculated.
3. The elements of the sequences are taken as strata boundaries.
4. For each strata model the Neyman allocation of sample sizes is calculated.
5. Having determined all sequences the a_1 and q resp. d combination with smallest sample size is found.
6. The procedure is then started again in the neighborhood of the best a_1 and q resp. d combination a. s. o.

By way of this procedure the number of strata is changed through variation of d resp. q and the location by variation of a_1. The first stage of the optimization comes to an end when a further reduction of the sample size by sequence value calculation is not possible.

In the second stage the final grid points of the first stage that means the final strata are virtually changed by systematic movement of strata boundaries and insertion / deletion of strata boundaries in order to check whether the sample size is reduced or not. If the sample size is reduced the modification is retained otherwise it is dismissed. The procedure stops when this source of sample size reduction is also exhausted.

3. Illustration by a Numerical Example

The method described will be illustrated by means of a generated example population containing 100 articles, the smallest value of an article being DM 0.78, the largest DM 87.15. The maximum relative sampling error is defined to be 0.01, the confidence level is chosen to be 0.95.

The parameters of the geometric sequence algorithm which was taken as an example are as follows:

a 1		q	
start	stop	start	stop
1.00	100	1.01	3.00

Table 1: Parameters of the Geometric Sequence Algorithm.

The results of step 1 of the first stage of the grid search are shown in the following table, the smallest sample size being N=52.

a q	**1.00**	10.90	20.80	30.70	40.60	50.50	60.40	70.30	80.20	90.10	100.00
1.010	84	78	78	83	88	88	91	91	99	99	99
1.209	**52**	53	53	55	65	67	70	81	91	99	99
1.408	69	67	72	79	74	73	78	81	91	99	99
1.607	71	66	82	82	82	79	81	81	91	99	99
1.806	79	77	80	90	83	90	81	81	91	99	99
2.005	72	74	78	90	91	90	81	81	91	99	99
2.204	86	84	84	84	96	90	81	81	91	99	99
2.403	73	74	89	78	96	90	81	81	91	99	99
2.602	82	67	92	80	96	90	81	81	91	99	99
2.801	79	79	86	87	96	90	81	81	91	99	99
3.000	77	79	81	87	96	90	81	81	91	99	99

Table 2: Optimization: First Stage, Step 1.

Already with step 3 the smallest sample size of the first stage (N=41) is reached.

a q	2.98	3.17	3.37	3.57	**3.77**	3.97	4.16	4.36	4.56	4.76	4.96
1.090	49	45	46	47	43	47	47	47	47	44	45
1.098	45	47	48	44	46	43	45	44	45	44	47
1.106	50	43	44	48	47	47	47	48	45	41	45
1.114	48	46	46	44	46	43	48	44	44	47	43
1.122	46	48	44	48	41	49	44	48	45	42	47
1.130	46	48	43	49	45	42	47	44	48	47	44
1.138	50	46	48	47	45	46	47	46	46	46	46
1.146	50	46	46	48	46	48	45	45	47	46	46
1.154	48	50	48	47	48	45	47	48	45	45	47
1.162	50	52	50	48	48	48	48	48	47	46	50
1.170	50	43	48	49	51	47	50	51	46	48	49

Table 3: Optimization: First Stage, Step 3.

The protocol of the optimization procedure is shown next:

opt.	a1		q		actual		sample
step	start	stop	start	stop	a1	q	size
1	1,000	100,000	1,010	3,000	1,000	1,209	52
2	1,000	10,900	1,010	1,408	3,970	1,130	42
3	2,980	4,960	1,090	1,170	3,772	1,122	41
4	3,574	3,970	1,114	1,130	3,774	1,122	41
5	3,734	3,814	1,120	1,124	3,766	1,122	41
6	3,758	3,774	1,121	1,123	3,764	1,122	41
7	3,762	3,766	1,121	1,123	3,763	1,122	41
8	3,762	3,764	1,121	1,123			

Table 4: Optimization: First Stage, Protocol.

The optimization procedure of the first stage finally results in the following strata model:

stratum	upper stratum boundary	number of elements of stratum sample
1	3.77	2
2	9.47	2
3	13.38	2
4	18.90	2
5	21.21	2
6	26.70	2
7	29.95	2
8	33.61	2
9	37.71	2
10	42.31	2
11	47.74	2
12	53.26	2
13	59.76	2
14	67.05	2
15	75.23	6
16	84.41	5
17	94.70	2

Table 5: Optimization: First Stage, Step 3, Strata Model, Size of this strata model: N=41.

By means of iterative refinement the sample size can be reduced to N=28 and the corresponding strata model is shown in the following table:

stratum	upper stratum boundary	number of elements of stratum sample
1	4.92	2
2	15.17	2
3	21.06	2
4	29.55	3
5	35.49	2
6	40.39	2
7	47.64	2
8	57.82	3
9	64.24	2
10	70.59	2
11	75.37	2
12	81.47	2
13	87.15	2

Table 6: Optimization: Second Stage, Final Strata Model, Size of this strata model: N=28.

First Monte Carlo comparisons with standard procedures such as the method of Dalenius and Hodges, the Dalenius equations, the method of Dalenius and Gurney, the method of Ekman and that of Mahalanobis and the optimization algorithm of Powell show that the proposed procedure of optimum stratification and allocation is the most efficient one. This result could be reached although the calculations were performed with a set of numbers of strata which includes the optimal one determined by our method. Nevertheless the results are tentative and should be confirmed by further experiments.

4. Application

In 1992 several practical applications have been carried through at German enterprises one of which that has been performed on November 1st 1992 at the fixed date October 30th 1992 will be reported in this section.

In the course of preparing the stock-taking the orderliness of the store had been checked throughout the year. The store population had been properly defined. The 20th substore was chosen for sampling. The stock-taking was performed during a normal business day. All store inputs and outputs during this time were registered.

The frequency distribution of the articles by quantities and values can be seen in the next table:

number of article	cumulated rel. frequency of quantitites	cumulated rel. frequency of values	value of upper limit article
1 - 520	0.07286	0.00000	0.04
521 - 720	0.10088	0.00034	6.30
721 - 1432	0.20064	0.00489	21.83
1433 - 2144	0.30041	0.01478	38.85
2145 - 2856	0.40017	0.03139	64.17
2857 - 3576	0.50105	0.05762	95.04
3577 - 4288	0.60081	0.09629	143.92
4289 - 5000	0.70057	0.15430	218.70
5001 - 5536	0.77568	0.21883	305.46
5537 - 5712	0.80034	0.24551	350.85
5713 - 6424	0.90010	0.40863	696.00
6425 - 6496	0.91019	0.43299	673.83
6497 - 6568	0.92027	0.45971	840.40
6569 - 6640	0.93036	0.48917	932.60
6641 - 6712	0.94045	0.52181	1,046.90
6713 - 6784	0.95054	0.55834	1,157.95
6785 - 6856	0.96063	0.59957	1,336.00
6857 - 6928	0.97072	0.64894	1,680.00
6929 - 7000	0.98080	0.71258	2,200.00
7001 - 7072	0.99089	0.78719	3,220.00
7073 - 7112	0.99650	0.86835	5,217.00
7113 - 7137	1.00000	1.00000	41,643.00

Table 7: Frequency Distribution.

Detailed data of the total article file are as follows:

number of articles	7137
smallest value	0.00
largest value	41,643.00
stock value	2,166,668.81
arithmetic mean quantity	134.80
variance quantity	3,685,229.04
standard deviation quantity	1,919.70
coefficient of variation quantity	14.24141
arithmetic mean value	303.58
variance value	868,040.78
standard deviation value	931.69
coefficient of variation value	3.06897
number of zero-positions	514

Table 8: Data File: Total Population.

This total population is subdivided into the sampling population and the complete stock-taking population, the last one being the subpopulation with high value articles:

number of articles	7103
smallest value	0.00
largest value	4,544.00
stock value	1,856,934.84
arithmetic mean quantity	133.79
variance quantity	3,689,285.84
standard deviation quantity	1,920.75
coefficient of variation quantity	14.35598
arithmetic mean value	261.43
variance value	223,276.55
standard deviation value	472.52
coefficient of variation value	1.80745
number of zero-positions	510

Table 9: Data File: Sampling Population.

number of articles	30
smallest value	4,601.92
largest value	41,643.00
stock value	309,733.97
arithmetic mean quantity	390.05
variance quantity	3,148,266.89
standard deviation quantity	1,774.34
coefficient of variation quantity	4.54905
arithmetic mean value	10,324.47
variance value	52,791,366.39
standard deviation value	7,265.77
coefficient of variation value	0.70374
number of zero-positions	0

Table 10: Data File: Complete Stock-taking Population.

The optimum stratification sample is now determined on the basis of the following sampling parameters:

quantile of the 95% confidence level	1.96
maximum relative sampling error	0.01

Table 11: Sampling Parameters.

The first stage of the optimization procedure is carried through by means of geometric sequences in the following way:

	a1		**q**		**sample**
step	**start**	**stop**	**start**	**stop**	**size**
1	1.000	100.000	1.010	5.000	318
2	1.000	20.800	1.010	1.808	114
3	8.920	12.880	1.090	1.250	105
4	10.108	12.900	1.122	1.154	104
5	10.108	10.187	1.143	1.149	104

Table 12: Optimization: First Stage, Protocol.

The strata model of step 4 is taken as start model for the second stage of the optimization procedure and there results a strata model with 35 strata and 73 sampling elements:

stratum no.	border	number of elements	number of sampling elements
1	8.35	806	2
2	24.00	716	3
3	41.05	692	3
4	60.74	572	3
5	79.53	449	2
6	98.80	406	2
7	121.28	349	2
8	145.95	317	2
9	171.45	310	2
10	201.58	266	2
11	236.00	230	2
12	271.99	221	2
13	308.63	216	2
14	355.35	176	2
15	409.90	145	2
16	465.79	148	2
17	525.59	137	2
18	592.89	118	2
19	665.74	103	2
20	749.97	95	2
21	835.96	83	2
22	927.32	76	2
23	1037.96	73	2
24	1147.48	67	2
25	1267.49	59	2
26	1434.76	47	2
27	1637.73	41	2
28	1838.96	31	2
29	2084.00	32	2
30	2319.12	33	2
31	2618.70	29	2
32	3086.68	19	2
33	3519.00	21	2
34	4121.95	12	2
35	4655.57	8	2
		7103	73

Table 13: Optimization: Second Stage, Final Strata Model.

On the basis of the above strata model a sample of 73 elements was drawn and the estimation results are presented in the following table:

	total	sample
number of articles	7137	7103
number of stock-taking articles	107	73
selection percentage	1.499	1.028
book value	2166668.81	
estimated value	2178364.97	1867831.00
deviation book value from estimated value	11696.16	
deviation percentage	0.539619	
confidence interval, upper limit	2197060.32	1886526.35
confidence interval, lower limit	2159669.62	1849135.65
deviation from estimated value	18695.35	18695.35
deviation percentage	0.858229	1.000912

Table 14: Estimation Results.

The estimated confidence interval shows a maximum relative sampling error of 0.858% and was well below the admissible limit of 1%. The result is accepted as orderly.

The deviations between book values and sampling values of the single articles included in the sample amounted to DM 817.19. Having taken into account this amount properly there remains a difference of DM 10,878.97 between book-value and estimated value. The balancing establishment has the right now to choose between the alternative of performing the sampling inventory as acceptance sampling or as estimation procedure. If the sampling inventory is interpreted as acceptance sampling the stock bookkeeping will be accepted as orderly, the remaining difference will not be taken into account and the articles not sampled are taken from the stock bookkeeping. If on the other side the sampling inventory is interpreted as estimation procedure the articles not sampled likewise are taken from the stock bookkeeping but the remaining difference of DM 10,878.97 has to be taken into account as returns.

5. Final Remarks

The algorithms are written in Turbo-Pascal 6.0 operating system DR-DOS 6.0 and implemented on an IBM-compatible PC, clockpulse 33MHz, processor INTEL 80486, 8MB core, 256MB cache, co-processor. A graphics card (1 MB) is used and disk memory 1GB with mean access time of 12ms. Printers are EPSON LQ 500 and HP LASERJET.

The handling of a stock population of about 11,000 articles required 30.79 min. Remarkable good results show the efficacy of the proposed optimization procedure compared to standard methods. The results should be established by way of further testing. The algorithm will be applied also in connection with official statistics and with several estimation and stratification attributes.

6. Literature

Aggarval, O.P. (1966), Bayes and Minimax Procedures for Estimating the Arithmetic Mean of a Population with Two Stage Sampling. The Annals of Mathematical Statistics, 37, 1186 - 1195.

Cochran, W.G. (1972), Stichprobenverfahren. Berlin.

Dalenius, T., Hodges,Jr. L. (1959), Minimum Variance Stratification. Journal of the Americal Statistical Association, 54, 88 - 101.

Ericson, W. A. (1965), Optimum Stratified Sampling Using Prior Information. Journal of the American Statistical Association, 60, 750 - 771.

Heike, H.-D., Danzer, K. (1982), Formulierung des Bankensektors in einem ökonometrischen Modell mit Hilfe der optimalen Kontrollrechnung. Statistische Hefte, 23, Heft 1, S. 11 - 26.

Heike, H.-D., Rossa, H. (1983), Optimal Stabilization with a Quarterly Model of the Federal Republic of Germany. In: Gruber, J. (Hrsg.), Econometric Decision Models. Berlin, Heidelberg, New York.

Kpedekpo, G.M.K. (1979), The Efficiency of Optimum Stratification Methods on Frequency Distribution with and without Discontinuities or Zero Classes, Sankya, Series B. 32, 1 - 12.

Raj, D. (1964), On Forming Strata of Equal Aggregate Size. Journal of the American Statistical Association, 59, 481 - 486.

Rao, T.J. (1977), Optimum Allocation of Sample Size and Prior Distributions: A Review. International Statistical Review, 45, 173 - 179.

Schneeberger, H. (1981), The Problem of Optimum Stratification and Allocation with q=n/N>0. Metrika, 28, 179 - 189.

Strecker, H. (1987), Statistische Erhebungen: Methoden und Ergebnisse. Ausgewählte Schriften, hrsg. von M.J. Beckmann und R. Wiegert, Göttingen.

Sukhatme, B.V., Tang, V.K.T. (1975), Allocation in Stratified Sampling, Subsequent to Preliminary Test of Significance. Journal of the American Statistical Association, 70, 175 - 179.

STRATIFICATION AND ALLOCATION METHODS FOR U.S. BUREAU OF THE CENSUS WHOLESALE TRADE SURVEYS

Ruth E. Detlefsen, Carol V. Caldwell, Timothy J. Braam,
B. Timothy Evans, U.S. Bureau of the Census
Carol V. Caldwell, Room 2653-3, Washington, D.C. 20233

KEY WORDS: Stratum boundaries, allocation, multiple variance constraints, sample size

1. INTRODUCTION

The Business Division of the U.S. Bureau of the Census provides measures of economic activity in the U.S. for wholesale businesses. We produce these measures from censuses, which we conduct every five years, and from monthly and annual sample surveys.

This paper focuses on the wholesale sample surveys. We describe the wholesale surveys and steps taken to reselect the samples that were introduced in early 1992. We also describe recent efforts to improve upon the stratification and allocation scheme that was used to reselect the wholesale samples. Improvements in the stratification should make wholesale samples selected in the future more efficient. The improvements will also enable analysis of estimates at more detailed kind-of-business levels than is currently possible. The improvements can also be applied to other Business Division sample surveys.

2. OVERVIEW OF WHOLESALE TRADE

Businesses in the wholesale trade are classified in the two-digit Standard Industrial Classification (SIC) codes 50 (Durable Goods) and 51 (Nondurable Goods) as defined in the 1987 Standard Industrial Classification Manual. Wholesale firms may be further categorized into three-digit SIC codes, 501 through 509 and 511 through 519. Within most of these three-digit industry groups, wholesale firms may also be classified by four-digit SIC codes. More detailed wholesale kind-of-business (KB) codes based on the SIC are also defined in the 1987 Industry and Product Classification Manual. Throughout this paper, we will use the terms SIC and KB interchangeably.

Wholesale establishments are primarily engaged in selling goods to retailers and to industrial, commercial, institutional, farm, construction, and professional business users. They also act as agents to bring buyers and sellers together.

3. SAMPLE DESIGN FOR THE MONTHLY WHOLESALE TRADE SURVEY

Business Division conducts two surveys of **merchant** wholesalers to estimate levels and trends of sales, inventory and other data. Merchant wholesalers are wholesalers that take title to the goods they sell. We conduct a monthly survey, called the Monthly Trade Survey (MTS), and an annual survey, called the Annual Trade Survey (ATS).

Before selecting the samples introduced in 1992, we conducted studies of the universe to determine stratum bounds, sample sizes and other parameters required for selection and estimation. Our parameter studies are outlined briefly below; see Detlefsen, et. al., 1991 for more details.

3.1 Universe Creation and Analysis

We constructed a computer file containing data records for all merchant wholesale employer establishments that were tabulated in the 1987 Census of Wholesale Trade. We then performed a preliminary analysis of the establishment universe to present a general picture of the universe and to verify that the universe contained no systematic data errors. The analysis included descriptive statistics, frequency distributions and edit failure listings.

3.2 Creation of Company Summaries

We summarized establishment data within each company to form company summary records. We determined a three-digit major sampling KB for each company based on the KB within each company contributing most to total annual sales for the company.

3.3 Certainty Determination

We set initial sales and inventory certainty boundaries by visually inspecting frequency distributions for each sampling KB to determine where the companies began to thin out in number to form the "tails" of the distributions.

Any company with sales or inventories exceeding the respective certainty cutoffs for its sampling KB or any associated KB was made a certainty sampling unit.

In addition, survey analysts added as certainties companies with known unique characteristics that could potentially lead to problems with the estimates if not included as certainty.

3.4 Creation of Noncertainty Sampling Units

We re-summarized establishments of all noncertainty companies into EI sampling units in the same way as company summaries were formed. The EI sampling units were used to determine noncertainty strata, sample sizes and allocations for each sampling KB.

3.5 CV Constraints

We determined a single coefficient of variation (CV) for each three-digit and aggregate KB to use in determining sample sizes. These design CVs ensured that target CVs identified by survey analysts for monthly level and ratio sales and inventory estimates could be met simultaneously. We determined these design CVs by applying theoretical relationships to data observed over several months and on actual relationships observed during that time.

3.6 Initial Noncertainty Sample Allocation

We applied the Dalenius-Hodges cumulative $\sqrt{f}$ rule to the EI sampling units for each sampling KB to determine noncertainty stratum bounds based on annual sales. Neyman allocation, based on end-of-year inventories, was used to determine minimum sample sizes for a number of target CVs and strata.

To study the sensitivity of sample sizes to changes in CV constraints, we used as target CVs the design CVs along with slightly lower and higher CVs (90% and 110% of the design CVs, respectively). We stratified using six, nine and twelve noncertainty strata and selected one allocation for each sampling KB to achieve minimum sample size.

3.7 Final Sample Allocation

We fed information about the selected allocations into a multiple CV constraints routine developed by Dr. Beverly Causey to determine optimum sample sizes and allocations to simultaneously meet all design CVs for sampling KBs and KB aggregates.

Following the multiple CV constraint analysis, we adjusted sample counts to prevent stratum sample sizes from exceeding the universe counts, keep stratum sizes above three, and where possible, force weights in smaller sales-size strata to exceed those in larger sales-size strata.

4. STUDY OF ALTERNATIVE STRATIFICATION STRATEGIES FOR THE WHOLESALE SAMPLE

4.1 Motivation for Study

Budget restrictions limit publication of wholesale estimates to the three-digit SIC/KB levels and the broader KB aggregates. We designed the wholesale sample introduced in 1992 to accommodate the publication requirements. Specifically, as described earlier, we formed sampling units based on three-digit KBs and allocated the sample to meet the variability constraints at the three-digit and higher KB levels.

We conjectured that the wholesale sample could be made more efficient by stratifying sampling units into more detailed KB by sales size categories. Variability constraints would be imposed at the three-digit and broader aggregate KB levels as usual, to meet publication requirements. In addition, variability constraints could also be imposed at the more detailed KB levels. Stratifying in this way could result in:

- smaller sample sizes needed to meet the variance constraints at the three-digit and higher KB levels, or
- reliable estimates at the more detailed KB levels, with the same sample size determined using stratification at the three-digit KB levels.

4.2 Description of Study

We studied the above conjecture, testing several different schemes for stratifying sampling units and imposing CV constraints. These strategies are summarized in Table 1.

In each scheme, sampling units were classified according to their major three-digit or four-digit kind-of-business, as indicated in Column 2 of the table below. We set certainty cutoffs and stratified the sampling units by major KB and sales size. We then determined the sample sizes needed to meet CVs imposed at all KB levels shown in Column 3.

The CV constraints we used for this study were:

KB Level	CV Constraint	
4-digit	7.0%	
3-digit	6.0%	for most 3-digit KBs; some variations to ensure that both level and ratio CVs were met for both sales and inventory
2-digit	2.0%	for durables (KB 50)
	2.5%	for nondurables (KB 51)
1-digit	2.0%	

Scheme 1 is the strategy we applied in selecting our current sample. We classified sampling units according to their major three-digit kind-of-business, and set certainty cutoffs and stratified the sampling units by major KB and sales size. We then determined the sample sizes required to meet CV constraints for the three-digit sampling KB levels and the two-digit and one-digit (total wholesale) aggregate levels.

For the remaining tests, we used as input the same computer file of wholesale establishment data that we used in determining the sample sizes in Scheme 1. In Schemes 2, 3 and 4, we classified sampling units according

Table 1

1	2	3
Scheme	Form strata based on:	Impose CV constraints for:
1	3-digit KB by sales size	3-, 2- and 1-digit KB
2	4-digit KB by sales size	3-, 2- and 1-digit KB
3	4-digit KB by sales size	4-, 3-, 2- and 1-digit KB
4	4-digit KB by sales size	selected 4-digit KB, 3-, 2- and 1-digit KB
5	selected 4-digit KB by sales size	3-, 2- and 1-digit KB
6	selected 4-digit KB by sales size	selected 4-digit KB, 3-, 2- and 1-digit KB

to their major four-digit kind-of-business, and set certainty cutoffs and stratified the sampling units by major KB and sales size. For Scheme 2, we determined sample sizes required to meet CV constraints for the three-, two- and one-digit KB levels. For Scheme 3, we included CV constraints at all 4-digit sampling KB levels.

For Schemes 4, 5, and 6, we made use of a "selected" group of four-digit KBs. Prior to our study, we asked Business Division wholesale specialists to identify the four-digit KBs that they would most like to publish estimates for, if possible. The four-digit KBs they identified are the "selected" KBs mentioned in the table above.

For Scheme 4, we stratified sampling units based on all four-digit KBs, but determined sample sizes required to meet CV constraints only for the selected four-digit KBs and for all three-, two- and one-digit KBs.

For Schemes 5 and 6, we formed sampling units based on the selected four-digit KBs and groupings of the remaining four-digit KBs. If a sampling unit had most of its sales in one of the selected KBs, then that four-digit KB became the sampling unit's major KB. If a sampling unit had most of its sales in a non-selected KB, then we assigned a major sampling KB representing all non-selected KBs within the three-digit KB.

For example, within three-digit KB 501 (Motor Vehicles and Motor Vehicle Parts and Supplies), there are four four-digit KBs:

- 5012 — Automobiles and Other Motor Vehicles
- 5013 — Motor Vehicle Supplies and New Parts
- 5014 — Tires and Tubes
- 5015 — Motor Vehicle Parts, Used

The survey specialists identified KBs 5012 and 5013 as levels they would most like to publish estimates for within KB 501. Any sampling unit having most of its sales in KB 5012 or KB 5013 was assigned a sampling KB of 5012 or 5013, respectively. Any sampling unit having most of its sales in three-digit KB 501 but not in KB 5012 or KB 5013 was assigned a sampling KB of 501X, which represented the balance of KB 501.

The wholesale specialists chose 21 four-digit KBs for which they most wanted to publish estimates. We established 18 "balance" category sampling KBs — one for each three-digit KB in the sample — giving a total of 39 sampling KBs for Schemes 5 and 6. By contrast, Schemes 2, 3 and 4 used 69 four-digit sampling KBs.

For Schemes 5 and 6, we set certainty cutoffs and stratified sampling units based on the selected four-digit sampling KBs and the "balance" KB groups. For Scheme 5, we imposed CV constraints at the three-, two-, and one-digit KB levels. For Scheme 6, we included CV constraints for the selected four-digit KB levels.

4.3 Study Results

Table 2 (next page) provides a summary of total required sample sizes resulting from the six tests.

These results show that:

4.3.1 Forming sampling units and strata for all 69 four-digit KBs as we did in Schemes 2, 3, and 4 requires more certainty and noncertainty sampling units than forming sampling units and strata at the three-digit KB levels, as we did in Scheme 1. The larger number of certainty sampling units is due to the unique certainty cutoffs that we set for each four-digit sampling KB in order to reduce variability at the four-digit KB levels and obtain efficient noncertainty samples for each KB.

Many of the additional non-

Table 2

Scheme	Number of Required Certainty Sampling Units	Number of Required Noncertainty Sampling Units	Total Number of Required Sampling Units
1	1,669	1,471	3,140
2	2,497	2,052	4,549
3	2,497	2,447	4,944
4	2,497	2,121	4,618
5	1,314	1,714	3,028
6	1,314	1,896	3,210

certainty sampling units are needed to meet the requirement that we take a minimum of three sampling units within each sampling KB by sales size strata. With 69 four-digit KBs, and a total of 555 noncertainty strata, we need a minimum of 1,665 noncertainty sampling units to meet this requirement, no matter what the CV constraints are at the four-digit KB levels. With the 18 three-digit KBs used in Scheme 1, we needed 1,471 sampling units to meet **all** requirements for noncertainty sampling. Under existing budget conditions, we cannot afford to sample all four-digit KBs as in Schemes 2, 3 and 4.

Note that we ran an additional test which set the certainty cutoff for each four-digit KB to the certainty cutoff used at the three-digit KB level in Scheme 1. We imposed CV constraints for the three-, two- and one-digit KB levels as in Scheme 2. This strategy yielded the same number of certainty sampling units as Scheme 1, but required 620 more noncertainty sampling units than Scheme 1 and 40 more than the original running of Scheme 2. Setting one cutoff based on all sampling units within a three-digit KB leads to inefficient samples for some of the four-digit KBs.

4.3.2 Forming sampling units and strata based on the 21 selected four-digit KBs and 18 balance groups as we did in Schemes 5 and 6 requires fewer certainty sampling units and more noncertainty sampling units than Scheme 1. The total required sample sizes in Schemes 5 and 6 are very close to the required sample size for Scheme 1. Scheme 6 requires 70 more sampling units than Scheme 1, but 355 fewer certainty sampling units. Certainty sampling units are more expensive than noncertainty sampling units in the Monthly Trade Survey, because data for certainties are collected every month and data for noncertainties are collected only once a quarter. The total cost of using the scheme proposed in Scheme 6 is likely smaller than the cost of the existing stratification and allocation scheme. The total cost of Scheme 5 is definitely smaller, because it requires fewer certainties and a smaller total sample size than Scheme 1.

4.4 Conclusions

The results above show that we can achieve approximately the same sample size as we have in our existing wholesale sample by forming sampling units and stratifying at selected detailed KB levels, as we did in Schemes 5 and 6. By forming sampling units and sampling at these finer levels, we impose greater control on our sample, enable the survey specialists to analyze estimates for important four-digit kinds of business within each three-digit category, and realize some cost savings. If we use Scheme 6 and place CV constraints on selected 4-digit KBs as well as the three-, two- and 1-digit KB levels, we may even be able to publish estimates for some of the four-digit KBs.

4.5 Future Work

We plan to continue this study by examining several stratification and sample allocation schemes with CV constraints for two variables, sales and inventories. Since Schemes 5 and 6 work best with CV constraints on sales alone, we will focus our future study on these two schemes.

We also plan to examine whether stratification and allocation schemes similar to Schemes 5 and 6 will benefit our retail and service samples which have designs similar to our wholesale sample. We foresee implementing results of these studies in the next sample selection operations, which are scheduled to take place after results of the 1992 Economic Censuses become available.

References

Causey, B. (1972), "Optimal Allocation in Stratified Sampling with Multiple Variance Constraints," Technical Notes No. 5, U.S. Bureau of the Census, Washington, D.C., 8-13.

Cochran, William G. (1977), Sampling Techniques, John Wiley & Sons, New York.

Detlefsen, R. and Veum, C. (1991), "Design of the Retail Trade Sample Surveys at the U.S. Bureau of the Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, 214-219.

Garrett, J., Detlefsen, R., and Veum, C. (1987), "Recent Sample Revisions and Related Enhancements for Business Surveys of the U.S. Bureau of the Census," Proceedings of the Business and Economic Section, American Statistical Association, 141-149.

Isaki, C., Wolter, K., Sturdevant, T., Monsour, N., Trager, M. (1976), "Monthly Business Surveys," Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Office of Management and Budget (1987), Standard Industrial Classification Manual 1987, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1967), The X-11 Variant of the Census Method II Seasonal Adjustment Program, Technical Paper 15, Washington, D.C.

U.S. Bureau of the Census (1979), Standard Statistical Establishment List Program, Technical Paper 44, Washington, D.C.

U.S. Bureau of the Census (1987), 1987 Industry and Product Classification Manual, Washington, D.C., Department of Commerce.

Wolter, K., Isaki, C., Sturdevant, T., Monsour, N., and Mayes, F. (1976), "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys," Proceedings of the Business and Economic Section, American Statistical Association, 99-109.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

CLUSTER SAMPLING FOR PERSONAL-VISIT ESTABLISHMENT SURVEYS

David W. Chapman, Aspen Systems
AMS Division, 962 Wayne Avenue, Suite 701, Silver Spring, MD 20910

KEY WORDS: Establishment surveys, cluster sampling, in-person interviewing

1. Introduction

With large-scale national personal-visit household surveys in the U.S. (e.g., the Current Population Survey, the National Crime Survey, and the National Health Interview Survey), the sample is clustered within several hundred first-stage or primary sampling units (PSUs). These PSUs typically consist of one or more adjacent counties or county equivalents. The clustering of the sample is introduced to improve overall sample efficiency (i.e., to minimize the mean square error of one or more key survey estimates for a fixed budget).

Many of the national establishment surveys are conducted by telephone or mail. Since both commercial and government sources for establishment sampling frames exist, there is generally no need or cost advantage to clustering for such surveys. As a result, much of the optimum design research for establishment surveys relates to strata formation, sample allocation to strata, and probability proportional to size (PPS) sampling, rather than to multi-stage cluster sampling.

For those establishment surveys that are conducted by personal-visit interviews, optimum design analysis should include the issue of clustering establishments at the first stage of selection. Because of some basic differences between household and establishment surveys (discussed in the next section, optimum design analysis for personal-visit establishment surveys is more complex than for personal-visit household surveys.

The purpose of this paper is to discuss some fundamental issues regarding sample designs for national personal-visit establishment surveys. An example of a clustered design for a national personal-visit establishment survey is described in Section 4.

2. Some Differences Between National Personal-Visit Household and Establishment Surveys

For personal-visit surveys (either household or establishment), the question of clustering is a sampling efficiency one: a trade-off between cost and variance. However, for establishment surveys, there are a number of basic differences, as compared to household surveys, that effect optimum design considerations.

First, establishments vary considerably in terms of any typically-used measure of size (e.g., number of employees or assets). According to Dun and Bradstreet (1992), about 90% of U.S. establishments (individual locations) contain less than 20 employees, and about 98% contain less than 100 employees. This size skewness creates a need, in terms of optimum design, to oversample larger establishments for surveys for which size is correlated with the basic survey characteristics.

Second, there are several list sources available commercially of establishments that contain a number of characteristics that can be used to measure size (e.g., number of employees), to define strata, to assign differential sampling rates to strata, and to derive measures of size for PPS sampling. Among the many list sources are three that provide comprehensive business files: (1) Dun and Bradstreet (D&B), (2) American Business, Inc. (ABI), and Database America (DBA). The D&B list, which is perhaps the most comprehensive, is developed primarily from credit inquiries. The ABI and DBA lists are compiled mostly from yellow-page phone directories.

Third, establishments are more sparsely distributed geographically than households, especially if the survey is focussing on a specific type of establishment (e.g., hospitals, colleges, or construction companies). This makes it difficult to define geographic clusters that contain some minimum number of establishments required for subsampling and yet are not too large geographically for cost-efficient interview coverage.

Fourth, many types of establishments, especially small ones, are not easy to identify from the outside. In some cases, a small business location may appear to be an ordinary household. In cities or suburban areas, establishments may be clustered in a large, multi-story office building, making it difficult to identify each one separately. Consequently, it is difficult to conduct area establishment surveys.

Finally, many establishment surveys require that information be collected from a specific officer of the company (e.g., the CEO or head of human resources) that can require considerable effort just to make contact with the respondent, let alone gaining cooperation. This can add considerably to the cost of data collection.

3. A Basic Clustered Design for National Establishment Surveys

For some establishment surveys the ultimate sampling unit is one or more employees. In such cases there would be at least one additional stage of sampling within establishments to obtain the sample. However, since the focus of this paper is on the issue of clustering establishments, the simplifying assumption will be made that the establishment (or equivalently, an establishment spokesperson) is the ultimate sampling unit.

Under this assumption, a basic design that would apply to many clustered establishment surveys is a two-stage design, where establishments are grouped into PSUs at the first stage, and establishments are drawn from the selected PSUs at the second stage. As with household surveys, the PSUs would often be stratified (at least geographically) and selected with probability proportional to size (PPS) from strata. Within selected PSUs, the establishments would generally be stratified by type of business, size, and perhaps other characteristics. The selection of establishments from strata can be done randomly, systematically, or with PPS. If random or systematic selection is used within strata, differential sampling rates by strata are often used, especially if "size" is a stratifying variable.

The following subsections give a more detailed discussion of the characteristics of the basic two-stage design introduced above. In Section 4, a description is given of a specific two-stage establishment survey design prepared for a personal-visit establishment survey sponsored by the U.S. Postal Service (USPS).

3.1 The Choice of Primary Sampling Units

Defining PSUs that are manageable in the field can be difficult because of the uneven geographic spread of establishments. Two possible choices of PSUs for establishment surveys are counties and zip codes. There are two problems with the use of counties. First, county information is not always provided on the establishment records in the list sources. Second, many counties outside of metropolitan areas have relatively small numbers of establishments. Creating a minimum number of establishments for a PSU in rural areas might require combining several counties.

It might be possible to use the county-based PSUs defined for a large federal household survey (e.g., The Current Population Survey) for establishment survey PSUs. However, if any additional combining of counties were needed for the establishment survey, it could be difficult to identify geographically adjacent counties for combining.

In terms of the use of zip codes for PSUs, they are almost always available on list sources. There is a question as to whether to use five-digit or three-digit zip codes. Five-digit zip codes are typically smaller than counties and may require considerable combining for most applications. Also, five-digit codes are not totally contiguous. Therefore, for zip-code based PSUs, three-digit zip codes are probably the best choice.

There is a natural grouping of three-digit zip codes which generally contain one or more central cities. However, in some parts of the U.S., three-digit zip codes cover large geographic areas. In such instances, an attempt could be made to break up the three-digit zip codes into 5-digit zip code PSUs, except that often the entire three-digit zip code area is needed to provide some minimum number of establishments in the PSU. This is an inherent problem with most establishment surveys, no matter how PSUs are defined.

In other countries, a mailing code similar to the U.S. zip code is generally available to use to define PSUs. PSU definitions for establishment surveys is an area where further research is warranted.

3.2 Optimum Sample Design

As noted in Section 1, much of the optimum sample design work that is carried out for establishment surveys focusses on strata formation and allocation of the sample to strata. These are typically survey applications for which clustering is not used. Often the optimum allocation is based on Neyman allocation, for which the optimum sampling rate for a stratum is proportional to its standard deviation (see Cochran, 1973, pp. 96-99).

For a stratified cluster design, Neyman allocation, which ignores the sample clustering, may still provide approximately optimum stratum sampling rates (assuming roughly equal unit costs rates between strata), though this hypothesis needs to be verified. If not, it may be necessary to develop more complex optimization approaches for determining stratum sampling rates that take clustering into account.

However the establishment stratum sampling rates are derived, this can be viewed as the first step in the development of an optimum design. The next step would be to determine how the total sample should be allocated in terms of the number of clusters to select and the within-cluster sample sizes. The optimum number of clusters to select depends on the cost structure, the sample design, and the within and between PSU variances.

The optimum cluster size for the simplest two-stage cluster design model is given by Hansen, Hurwitz, and Madow (1953), p. 286. Although that sample design is much simpler than the one discussed here, it may still be helpful when the better-fitting models are too complex to use. Hansen, et. al. (1953) introduce stratification of PSUs into the optimization model (Chapter 7) and PPS selection of PSUs (Chapter 8). If feasible, these more realistic models should be used to approximate the optimum design.

In deriving optimum designs, there are many field operations issues which are difficult to incorporate into a cost model. Examples of these issues are the importance of manageable interviewer workloads and the time constraints on data collection.

3.3 Certainty Selections

In a design involving PPS selection at the first stage, it should be determined whether or not any of the PSUs are large enough to be included in the sample with certainty (also referred to as "self-representing" PSUs). Consideration must also be given to the selection of any establishments with certainty. The determination of which establishments to select with certainty can be approached in two ways. With the first approach, PSUs are selected without regard to the size of individual establishments. Then, establishment certainty selections, if any, would be identified from within only those PSUs that were selected at the first stage. A problem with this approach is that it is possible to leave out of the sample very large establishments because they were not located in PSUs selected at the first stage.

With the second approach, establishments to be selected with certainty, if any, are identified at two stages: (1) prior to selecting PSUs and (2) as part of within-PSU sampling. This approach has the problem that the certainty selections identified at the first stage could be located in PSUs that are not selected into the sample, thus creating field inefficiencies. To avoid this, all PSUs that contain a certainty establishment could be made certainty PSUs for the first stage of selection.

If establishments to be selected with certainty are identified before selecting PSUs (second approach), the certainty criterion is based on a single-stage systematic PPS selection taken across all establishments in a sampling stratum, derived as though the sample were unclustered. With this procedure, the sum of the measures of size of the establishments in a stratum is divided by the stratum sample size to obtain a hypothetical selection interval. Any establishment whose measure of size exceeds about 70% of the selection interval would be included with certainty. This is a somewhat arbitrary cutoff based on variance considerations (i.e., it avoids having a relatively low probability of not selecting a large establishment).

The identification of certainty PSUs is similar to that described above for identifying certainty establishments at the first stage. If PPS sampling of establishments is used at the second stage, the method of identifying certainty establishment selections is also similar to that used at the first stage. If simple random or systematic sampling is used at the second stage, then the identification of certainty selections is based on the derived second stage sampling rate for the stratum within each PSU (discussed further in Section 3.6). If the derived within-PSU sampling rate is greater than or equal to 1.0 for a stratum, all establishments in the stratum are selected with certainty.

In choosing between the two alternative methods of identifying establishments to be selected with certainty, they should be compared in terms of their impact on survey variances and field costs.

3.4 Assigning Measures of Size to PSUs

In multi-stage surveys (either households or establishments) involving PPS selection of PSUs at the first stage, generally measures of size are assigned to PSUs that are equal to the number of ultimate sampling units they contain, if available. For example, in a national household survey, the PSU measure of size is often taken to be the best available count of the number of households in the PSU, though sometimes the number of persons is used.

With a clustered establishment survey, the best measure of size to use depends on the method of selecting establishments at the second stage. If a PPS sample of establishments is selected at the second stage, the measure of size assigned to a PSU would be the sum of the size measures of the establishments in the PSU. With a computerized establishment frame, it is not difficult to compute these measures of size.

If stratified random sampling is applied within PSUs at the second stage, one might consider using the total number of establishments in the frame in a PSU as the measure of size. However, the stratum sampling rates vary, it is better to use a composite measure of size that incorporates both the stratum sampling rates and frame counts. This method of defining measures of size was proposed by Folsom, Potter, and Williams (1987).

With their approach, the measure of size, A_i, for the i^{th} PSU is computed by multiplying the number of establishments in the PSU in stratum h, N_{hi}, by the corresponding target sampling rate for the stratum, f_h, and summing these products across all strata:

$$A_i = \sum_h f_h N_{hi} . \qquad (1)$$

This measure of size is a weighted sum of the number of establishments in each stratum, where the weights are the stratum sampling rates. The sum of the measures of size of all PSUs equals the overall target sample size.

Folsom, et. al., show that with this method of assigning measures, the overall workload (sample size) in each stratum will be constant, assuming that random samples of establishments are selected from the strata and that the sampling rates applied are those that provide the target overall stratum sampling rates.

3.5 Stratification and Selection of PSUs

The stratification and selection of PSUs would generally be conducted in a way similar to that for household surveys. Specifically, the PSUs would be stratified geographically and perhaps by degree of urbanization. If the PSUs are county-based, urbanization level could be based on MSA/non-MSA designations, and by population size of the PSU or its central city. If PSUs are zip-code based, an approximate MSA code could be assigned to use for urbanization. Also, the size of the PSU in terms of the total number of establishments it contained could be used as part of the urbanization stratification.

If m represents the number of PSUs that are to be selected for the sample, then the PSUs would be grouped into m/2 strata for a two-PSU-per-stratum design, or into m strata for a one-PSU-per-stratum design. In either case, an attempt would be made to define strata that were approximately equal in terms of the total measures of size of the PSUs they contain.

The designated number of sample PSUs would then be selected with PPS from each stratum. For a two-per-stratum design, the two PSUs could be selected using systematic PPS or by some other scheme especially designed for PPS without replacement sampling. A discussion of many of these procedures is given by Cochran (1973), pp. 258-270.

3.6 Selection of Establishments within PSUs

For most establishment surveys, larger establishments are selected with higher overall selection probabilities than are smaller establishments, in an attempt to approximate an optimum allocation of the sample to size groups. This oversampling can be accomplished using either of two approaches. With the first, establishments are selected with PPS, where the size is assigned in such a way as to approximate an optimum allocation of the sample to size groups.

The other approach, which is often preferred because it is easier to execute and simplifies survey estimation and variance estimation, is to use size class as a major stratification variable and to select establishments within strata with equal probability (either a simple or systematic random sample.) The oversampling of larger establishments is achieved by assigning higher sampling rates to strata that contain the larger establishments.

The specific sampling rate applied to each stratum in a PSU is determined by the overall "optimum" stratum sampling rate, f_h, which is derived by methods discussed in Section 3.2. If P_i represents the selection probability of the i^{th} PSU and N_{hi} is the number of establishments in PSU i in stratum h, the sampling rate, f_{hi}, applied to the N_{hi} establishments is:

$$f_{hi} = f_h / P_i . \qquad (2)$$

If PPS selection were used at both stages, the total establishment sample would first be allocated to the PSU strata in proportion to the total of the measures of size of the establishments in each stratum. This allocation would be done separately by each of the establishment-based strata. Therefore, if there were L establishment-based strata, each PSU stratum would receive L sample allocations. Next, the one or two PSUs to be selected from the stratum would be selected with PPS, where the size would be the sum of the sizes of all the establishments in a PSU. Finally the sample allocated to a PSU stratum would be selected with PPS from the one or two PSUs chosen, generally using a systematic PPS procedure. If two PSUs are selected from each PSU stratum, the L stratum allocations would be split evenly between the two PSUs. In this case the PPS selection is somewhat disrupted due to the uneven distribution across PSUs of the establishments in the various (establishment-based) strata.

4. An Example: Sample Design for MAIS III

The Marketing Analysis Inventory System (MAIS) Survey, sponsored by USPS, consists of a national probability sample of U.S. business establishments, nonprofit organizations, and government agencies. The basic purpose of MAIS, which is conducted every two or three years, is to measure the non-residential use of a variety of types of mail. Following is a summary of the sample design proposed for MAIS III, the third installation of MAIS. This design was developed in 1992 under contract to the USPS, when the author was an employee of National Analysts of Philadelphia. This summary is abstracted from a memorandum prepared by Chapman, Rothschild, and Finkbeiner (1992).

The proposed sampling plan for MAIS III was a two-stage stratified cluster sample of 5,000 U.S. establishments or agencies, with clusters (PSUs) being defined by one or more 3-digit zip codes. Differential stratum sampling rates, based on an optimum allocation analysis applied to MAIS II data, was proposed. Within each selected PSU, a stratified systematic random sample of establishments was recommended, with differential sampling rates across strata.

There was complexity in the proposed design due to the use of multiple frames, the need for clustering the sample to improve sample efficiency, and the desire to oversample various population subgroups. The proposed sampling plan is broken down into seven major steps, summarized below:

(1) Creating the multi-source sampling frame. The proposed frame for this study was a list generated by merging the following four sources:

- (a) Database America's (DBA's) main business file, which is primarily a yellow-page based source of about 9.1 million establishments.
- (b) The Postal Service's AMIS file of key and National Accounts. This file contains over 5,000 of the Postal Services's largest customers in terms of mail volumes and revenues.
- (c) The Census Bureau's Government file, which contains listings of Federal, state, and local government agencies from the most recent Census of Governments.
- (d) USPS's Nonprofit Mailer's file, which is a special file of Second and Third Class mailers.

The DBA file was the primary source for the frame. The other three sources were used to identify certainty selections (i.e., the AMIS file) or to improve the coverage of the target universe.

(2) Identification of establishments to be selected with certainty. Because of the existence of very large USPS customers (i.e., the National Account customers), it was planned to identify initial certainty selections before selecting PSUs. These consisted of all National Accounts plus other accounts in the AMIS file that would have more than a 50% chance of being selected in an unclustered, systematic PPS sample of establishments, where size is the annual mail volume. In addition, it was proposed to select with certainty any establishment which accounted for some minimum percent (perhaps 5%) of the total mail volume for a specific type of mail in the MAIS II survey.

(3) Stratification of establishments and the derivation of stratum sampling rates. It was proposed to construct one or two separate strata out of the noncertainty listings in the AMIS file. It was planned to stratify the remainder of the establishments in the frame by 19 industry types (based on SIC code) and by four employee size classes (1-9, 10-19, 20-99, and 100+). Crossing the 19 SIC groups with the four employee size classes would yield a total of 76 strata.

It was planned to allocate the sample to these 76 strata using Neyman allocation, based on estimated stratum standard deviations of total mail volumes, computed from MAIS II data. If all 76 strata were retained for the sample, there would be many empty or low-count strata in some of the selected PSUs. Therefore, the proposal recommended grouping the strata on the basis of the estimated optimum sampling rate and, as a result, collapsing the 76 strata to between 4 and 7 sampling strata.

(4) Definition of zip-code based PSUs and assignment of measures of size to the PSUs. It was proposed that PSUs be defined by three-digit zip codes, using groupings of three-digit zip codes defined as "Sectional Areas" by Rand McNally in their Three-Digit Zip Code Atlas. (There are about 600 Sectional Areas defined across the U.S., most of them containing one or more central cities.) It was recognized that some of the Sectional Areas might have to be subdivided because of their geographic size. This subdividing could be done at the PSU level or at a second stage of sampling.

Once these PSUs were defined, frame counts of establishments by sampling strata would be made for each PSU. These counts would be used to create the composite measure of size for each noncertainty PSU using Equation (1) in Section 3.4.

(5) Determination of the Number of PSUs to Select. It was anticipated that the number of PSUs selected would be determined primarily on the basis of practical factors: survey costs and workload size. As an initial proposal, it was suggested that 100 PSUs be selected. Deriving an approximate optimum number of PSUs to select for the survey using standard formulas was not proposed primarily because required cost and correlation data would be difficult to assemble. Also, the optimization models do not take into account some of the practical considerations, like desirable workload sizes.

(6) Stratification and Selection of PSUs. It was proposed to allocate the 100 sample PSUs to ten Postal Customer Service Areas in proportion to the total of the measures of size. The first step proposed in selecting PSUs from a Service Area was to identify certainty selections. It was suggested that this be done by calculating the PSU selection interval used for systematic PPS selection: the total of the measures of size of all the PSUs in the Service Area divided by the number of PSUs to be selected from that Service Area. It was recommended that a PSU whose size exceeded 70% of the selection interval be selected with certainty.

It was proposed to select the non-certainty PSUs using a systematic PPS procedure. It was recommended that the PSUs be sorted by geography and urbanization level before selection. This would provide some "implicit" stratification in terms of the sort variables.

(7) Selection of Establishments from Each PSU. It was proposed that within each PSU the establishments be grouped into the 4-7 sampling strata referred to in step (3). It was planned to select a systematic random

sample from each of the sampling strata. The sampling rate used in each stratum would be computed in such a way that, within round-off error, the overall selection probability of the establishments in the sampling stratum would equal the target figure, f_h. To achieve this, the appropriate within-stratum sampling rate, f_{hi}, would be computed using Equation (2) in Section 3.6.

It was recommended that oversampling be applied to all the sampling strata to allow for bad listings, non-contacts, and non-interviews. It was proposed to select six times as many establishments (i.e., 30,000 in total) to be sure that enough sample cases would be available.

5. Conclusions

Optimum sample design for personal-visit establishment surveys is more complex than it is for personal-visit household surveys primarily because of the size variation among establishments, the availability of list sources, and the uneven geographic distribution of establishments. A fundamental question is whether or not to introduce clustering. For a large government establishment survey there may be a companion household survey that is already being conducted with a clustered design. In such a case, using the same PSUs for the establishment survey would be efficient. This is indeed the circumstances that apply to NCHS's National Health Care Survey, described in Appendix A in the document edited by Wunderlich (1992).

For an isolated personal-visit establishment survey, it is possible, though it seems unlikely, that the optimum design does not call for clustering due to the large geographic spread of establishments within PSUs. To check this, optimum cluster sizes can be approximated from available data. For example, consider the simplest formula for the optimum cluster size, n, for a simple two-stage design given by Hansen, et. al. (1953), p. 286:

$$n = \sqrt{\frac{C_1}{C_2} \cdot \frac{1-\delta}{\delta}} \qquad (3)$$

where

C_1 = the marginal cost associated with each PSU in the sample,

C_2 = the marginal cost associated with each establishment in the sample,

δ = the measure of homogeneity between units selected from the same PSU.

This is an oversimplified model for the two-stage establishment survey design that has been addressed in this paper, but the basic ideas are still relevant. Namely, a small value of δ, which may exist for many establishment surveys, would suggest a larger within-cluster sample size. However, a relatively high marginal cost, C_2, associated with the enumeration of units within a PSU, compared to the per-PSU survey costs, C_1, would suggest a lower optimum within-cluster sample size. With large, sparsely-populated establishment PSUs, C_2 could be relatively high.

Such considerations, and other optimum allocation issues, need to be investigated in the design of personal-visit establishment surveys.

ACKNOWLEDGEMENTS

Most of the work for the research reported in this document was conducted while the author was employed by National Analysts of Philadelphia, under contract to the United States Postal Service (USPS). The contract officer for the USPS was Fred Lesnett. The project director for National Analysts was Beth Rothschild. Both Ms. Rothschild and Carl Finkbeiner provided general guidance and direct input into the sample design research. Molly Lambert of also made major contributions to the research.

REFERENCES

Chapman, David, Beth Rothschild, and Carl Finkbeiner (1992), "Sampling Plan for MAIS III." A National Analysts memorandum to Fred Lesnett, USPS, September 24.

Cochran, William G. (1973), Sampling Techniques, 3rd ed. John Wiley and Sons, New York, NY.

Dun and Bradstreet (1992), "The New Dun & Bradstreet Catalog of Sales and Marketing Information." A brochure prepared by Dun and Bradstreet, Cherry Hill, NJ.

Folsom, R.E., F.J. Potter, and S.K. Williams (1987), "Notes on a Composite Measure for Self-Weighting Samples in Multiple Domains." Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 792-796.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953), Sample Survey Methods and Theory, Vol. 1. John Wiley and Sons, New York, NY.

Wunderlich, Gooloo S., ed. (1992), "Toward a National Health Care Survey." Report prepared by the Panel on the National Health Care Survey, Committee on National Statistics, National Academy of Sciences. National Academic Press, Washington, DC.

SAMPLING FARMS USING AREA FRAMES IN EUROPE

F. J. Gallego, J. Delincé, Joint Research Centre of the E.C.
F.J. Gallego, JRC tp 440, 21020 Ispra (Varese), Italy

Key words: Area Sampling Frame, Weighted Segments

SUMMARY

In the MARS Project (Monitoring Agriculture with Remote Sensing) of the E.C., area frames based on a squared grid are used for area estimation through ground survey and satellite images. The sample elements (segments) of the area survey are used as well for sampling farms with a template of points overlaid on the segment. Most often we use a fixed number of points (agricultural or not) per segment. Farmers are asked to provide global data for the farm (weighted segment approach), and estimates are computed with a Horvitz-Thompson approach. In Rumania different sampling rates are used for private owners or state farms. Major problems include locating farmers and checking for misunderstanding of instructions.

Possible bias can be partially checked by comparing area estimates of the farm survey with area estimates from the segment survey. Good results are obtained for area and production of the main extensive crops. Area frames need to be complemented with list frames (multiple frames) to give reliable estimates for cattle.

1. AREA SAMPLING FRAME BASED ON A SQUARED GRID.

In the regional crop inventories of the MARS Project (Monitoring Agriculture with Remote Sensing) of the E.C., area frames are being used primarily for area estimation through ground survey and satellite images. Unlike the area frames used by the USDA (Allen, 90, Cotter, 87), we generally use frames based on a squared grid (Gallego 93). Fig. 1 illustrates a small example of this kind of sample with a very simple stratification and segments of 25 Ha. Sampling is systematic repeating a pattern in squared blocks. In this case the blocks have a size of 10 Km. × 10 Km., and the pattern has 4 replicas in the most agricultural stratum (plain), 2 replicas in the hills, and one in the mountains.

The pattern is drawn at random with a restriction on the distance between segments in order to avoid segments that are too close to each other. The size of the segments varies from region to region depending on the agricultural landscape, especially on the size of fields. In the Czech Republic, the segment size was 400 Ha. For the area survey, enumerators locate the segments, draw fields on a transparent sheet, and write down their land use.

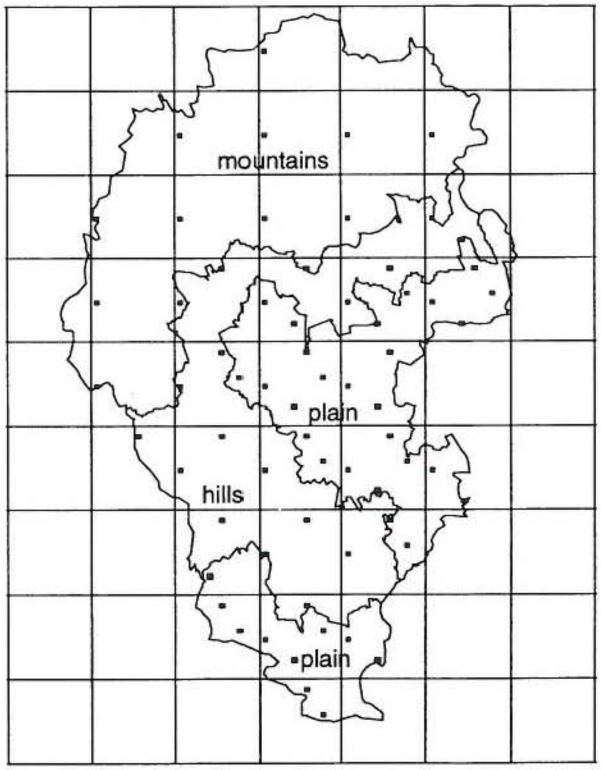

Fig. 1: Example of area frame sample with squared segments and squared blocks

2. SAMPLING FARMS BY POINTS

For agricultural surveys in the European Community, farms are traditionally sampled on a list frame (Eurostat, 91). The list is a census of farms that reach a certain size threshold. In many countries censuses are made every 10 years. Hence there may be a substantial difference between the sampling frame and the actual population at the date of the survey. The situation is worse in the countries of the former Eastern Block, where the change of land property structure is so fast that there may be no census at all. Area frames on squared segments can be easily defined when the geographic borders of the region are known.

The U.S. Department of Agriculture has extensive experience using area frames, that are currently built with sampling units that follow physical limits, such as roads, canals, etc. (Cotter, 87). In Europe this approach is very cumbersome because of the complex agricultural landscape, with farms often made up of a large number of small, scattered fields. Hence we have preferred frames based on squared segments. These segments are used as well for sampling farms in several countries with the help of a template of points overlaid on the segment. This has been tested in Germany, Portugal, Spain, Greece, Italy (Carfagna, 91), Rumania and Czechoslovakia.

The template is the same for all the segments in a stratum. Data are obtained only for farms corresponding to points falling on agricultural land. In the example of Fig. 2, point 3 fell on woodland and point 2 on a built area. They will generate two zero-valued records in the farm file. The enumerator will have to locate the farmers for the other three points. The farm corresponding to point 1 has other fields in the segment, that will be implicitly included in the survey, but the enumerator will not need to find out if these fields exist. Points 4 and 5 belong to the same farm, that will appear twice in the farm file.

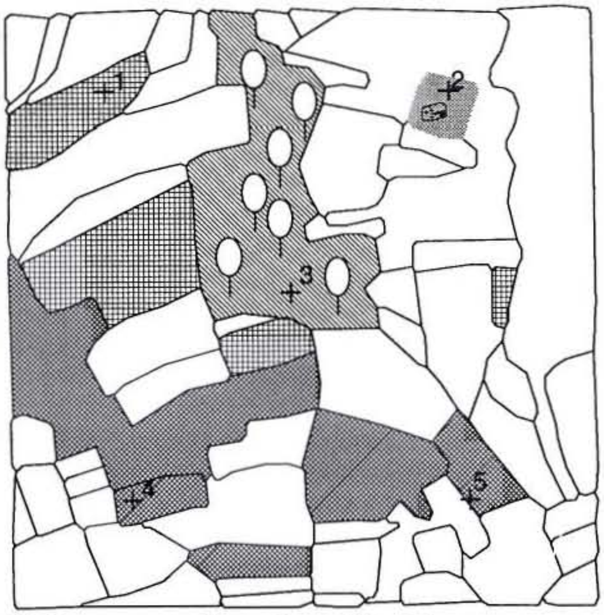

Fig 2.: segment with a pattern of 5 points

Farmers are located and asked to provide global data for the farm, including total area and production of each target crop. No question is asked about the production of each field or the set of fields inside the segment.

Major problems include: a) locating farmers in countries where farmer dwellings tend to be concentrated in urban nuclei, b) checking that the instructions have not been misunderstood in any of the steps: JRC → National Administration → Regional co-ordinators → Enumerators. In some countries, linguistic limitations are a serious barrier to direct contact with enumerators. One important problem is the definition of UAA (Utilised Agricultural Area) that is applied in the field work. It can be adapted to each region, but it must be consistent with the definition used for computation.

In Rumania different sampling rates are used depending on the type of land owner (private farms or state farms) since the area under private cultivation is small, but the production is not negligible and the need for information is still high. Hence a higher sampling rate is used for private farms. The double sampling rate can be performed because the enumerator can decide easily if the field is private by the size of the plot.

3. ESTIMATES BASED ON FARMS SAMPLED BY POINTS.

We assume that the population Ω is divided into strata Ω_h , $h=1...H,$ the total population size is N segments (N_h for stratum Ω_h) and the sample size is n segments (n_h). The size of our sample of points in each segment will be K_i, previously fixed; in general we have $K_i=K$, constant, out of which F_i correspond to the farms on which these points fall. Each segment i has a total UAA surface U_i.

We have a two-step sampling. In the first step the segment i is selected with probability $p_i=1/N_h$ in each of the n_h trials. In the second step the unit is the tract (set of agricultural fields, or pieces of fields in a segment, that belong to the same farm). The tract k of segment i has an area T_{ik}. The total UAA of the farm is A_{ik}. U_i is the sum of the tracts T_{ik} in segment i.

3.1. Estimates based on Farm And Non-Farm Points.

There will be K-F_i observations (fictitious farms) with value 0 corresponding to points outside the UAA.

Farms are sampled in segment i with a probability p_{ik} proportional to T_{ik}/D_i , where D_i is the size of the segment . The sampling is made with replacement: a farm can be selected more than once, which gives easier formulae. The crossed probabilities $p_{ikk'}$ that farms k and k' are in the sample are not exactly the same as if the different points of the template were drawn independently, since there is usually a relatively large distance between them. We will disregard this fact for now.

W_{ik} will be an additive quantity for a farm, most often the production or the area of a particular crop. The estimates are also possible for cattle, but the results will be presumably bad if there are a substantial number of farms without any UAA, that will not be sampled. It is obvious that yield is not an additive variable.

Since we have no information about how W_{ik} is distributed inside the farm, we create a fictitious variable X that is uniformly distributed, and that has, by definition, the same total as W for each farm:

$$X_{ik} = \frac{T_{ik}}{A_{ik}} W_{ik} \qquad (1)$$

Estimating the totals of X and W are equivalent problems.

The two-stage version of the Horvitz-Thompson estimator for the total of X in the stratum Ω_h gives :

$$\hat{X}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{1}{K_i}\sum_{k=1}^{K_i}\frac{X_{ik}}{p_{ik}} = \frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{D_i}{K_i}\sum_{k=1}^{K_i}\frac{W_{ik}}{A_{ik}} \qquad (2)$$

This means that, even if the second stage sampling unit is the tract, we do not need to know its area nor X_{ik}., but just the global information about the farm.

The estimator is a linear function of the estimates on the selected segments. Its variance in stratum Ω_h can be estimated as (Cochran, 1977, section 11.6):

$$\hat{V}(\hat{X}_h) = \begin{array}{l} \frac{N_h^2}{n_h}(1-\frac{n_h}{N_h})\sum_{i=1}^{n_h}\frac{(\hat{X}_i-\overline{X}_h)^2}{n_h-1} + \\ +\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{1}{K_i(K_i-1)}\sum_{k=1}^{K_i}(\frac{W_{ik}D_i}{A_{ik}}-\hat{X}_i)^2 \end{array} \qquad (3)$$

The estimates for the total are:

$$\hat{X} = \sum_{h=1}^{H}\hat{X}_h \qquad \hat{V}(\hat{X}) = \sum_{h=1}^{H}\hat{V}(\hat{X}_h) \qquad (4)$$

Crop areas can be estimated both from the segment survey and from the farm survey. Comparing both area estimates can be useful to check for possible bias on the production estimate based on the farm survey.

3.2. Estimation Based Only on Farm Points.

We shall mention another possible option, that consists in using only points that fall in the UAA. In this case, we previously fix F_i, the number of points that fall in UAA (often $F_i=F_h$, constant in each stratum). In segment i we observe as many points as necessary to have F_i points in the UAA. If the segment i has no UAA, one observation (fictitious farm) is added with 0 values. In this case (2) and (3) are to be adapted substituting K_i by F_i and D_i by U_i:

$$\hat{X}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h}\sum_{h=1}^{n_h}\frac{1}{F_i}\sum_{k=1}^{K_i}U_i\frac{W_{ik}}{A_{ik}} \qquad (5)$$

$$\hat{V}(\hat{X}_h) = \begin{array}{l} \frac{N_h^2}{n_h}(1-\frac{n_h}{N_h})\sum_{i=1}^{n_h}\frac{(\hat{X}_i-\overline{X}_h)^2}{n_h-1} + \\ +\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{1}{F_i(F_i-1)}\sum_{k=1}^{K_i}(\frac{W_{ik}U_i}{A_{ik}}-\hat{X}_i)^2 \end{array} \qquad (6)$$

the second term of (6) is null for segments with no UAA. This term cannot be computed if $F_i=1$. A value 0 can be attributed, though this will lead to a slight underestimation of the variance.

3.3. Farms with fields in different strata

At first sight, the estimator (2) seems to assume that a farm k that has been selected through a point in stratum Ω_h is completely included in this stratum. It is obvious that a farm can have fields in different strata, and the question arises as to whether this fact disturbs the reliability of the results.

We should stress again that the variable used is not really W_{ik}, but X_{ik} defined for each individual tract. The total production W does not coincide with the total of X in each stratum, but it does in the whole region if the farms are assumed to be entirely inside the region.

3.4. Non Response

If a farmer does not co-operate or cannot be found, the farm is not be included in the sample size K_i. There is a risk of bias if farmers who cannot be located or refuse to cooperate have a peculiar behaviour.

If in the second stage (sampling farms inside the segment) we consider farm and non-farm points, and give value 0 to the points that fall in non agricultural land, the exclusion of non respondents can produce a serious bias, because the zero values corresponding to non-agricultural land are never missing. We can think of two different ways of overcoming this problem: substituting missing values with a kind of "average farm" values, or eliminating a proportional number of 0 values corresponding to non-UAA points. Both give the same estimate for the total.

3.5. Two types of farms with different sampling rate.

We may wish to improve the precision for crops mainly produced by a certain type of farms that can be identified in the area survey. This is the situation in Rumania, where private farms can be identified by the size of the field. We wish a sampling rate B times higher (B=2 in Rumania) for private farms (type 1) than for state farms (type 0).

The sampling strategy is as follows: For segment i select K_{ia} points. The points outside the UAA generate observations with value 0. The other points generate records with areas and productions as in the general case (any type of farm). Then we select $(B-1)K_{ia}$ extra points, keeping the points in farms of type 1. We shall have a part of the segment (farms type 0 and non UAA) with area D_{i0} with the basic sampling rate, where K_{i0} points have been selected. K_{i1} points are selected in an area D_{i1} (farms of type 1).

Another sampling strategy is fixing K_{i0} and K_{i1} previously and selecting as many points as necessary in the segment to have the wished subsample sizes.

We assume now that K_{i1} is fixed for each segment (second strategy). For the first strategy (K_{i1} random) we would need a supplementary term in the variance, that can be significant if $K_i=K_{i0}+K_{i1}$ is small. We use X defined as in (1).

If we make a random permutation of the K_i points of the sample, the probability of tract k in each trial is:

$$p_{ik} = \frac{K_{i0}}{K_i}\frac{T_{ik}}{D_{i0}} \qquad\qquad p_{ik} = \frac{K_{i1}}{K_i}\frac{T_{ik}}{D_{i1}} \qquad (7)$$

for farms of type 0 and 1; and D_{i1} the total area of the segment for farms of type 1.

The Horvitz-Thompson estimator for the total of X in Ω_h becomes:

$$\hat{X}_h = \frac{N_h}{n_h}\sum_{i=1}^{n_h}\hat{X}_{i0} + \hat{X}_{i1} = \qquad (8)$$
$$\frac{N_h}{n_h}\sum_{h=1}^{n_h}[(\frac{1}{K_{i0}}\sum_{k0=1}^{K_{i0}} D_{i0}\frac{W_{ik0}}{A_{ik0}}) + (\frac{1}{K_{i1}}\sum_{k1=1}^{K_{i1}} D_{i1}\frac{W_{ik1}}{A_{ik1}})]$$

$$\hat{V}(\hat{X}_h) = \frac{N_h^2}{n_h}(1-\frac{n_h}{N_h})\sum_{i=1}^{n_h}\frac{(\hat{X}_i-\overline{X}_h)^2}{n_h-1} + \frac{N_h}{n_h}\sum_{i=1}^{n_h}\psi_{h0}+\psi_{h1} \qquad (9)$$

$$\text{with} \quad \psi_{hl} = \frac{1}{K_{il}(K_{il}-1)}\sum_{kl=1}^{K_{il}}(\frac{W_{ikl}D_{il}}{A_l} - \hat{X}_{il})^2 \qquad l = 0,1$$

3.7 Software to compute estimates: AIS STIM

A program in C for PC has been written (Dicorato, 93) with the name of AIS_STIM to compute estimates using the described methods. The main part of the program was first written to compute estimates on a segment survey

4. RESULTS: SOME EXAMPLES.

We present here some results from regions that represent some kind of particular behaviour or in which some specific test has been made.

4.1 Emilia Romagna 1990

In Emilia Romagna, locating the farmer was only a moderate problem:

Segments sampled:	285	
Points sampled:	1565	
Points on non UAA:	326	
Address not found:	206	
Farmer not found:	38	
Refusals:	32	
Completed forms:	617	(963 points)

As suggested in section 3.1, we can have an idea of an eventual bias in the farm survey by comparing with the area estimates of the segment survey, more objective and complete. Estimates match well for cereals and permanent crops. Official statistics for durum wheat might be very debatable. Some problems in sugar beet might be related with misunderstandings on how to declare second crops in the same field, or a bias due to missing values.

The coefficients of variation in the farm survey have a reasonable behaviour for cereals, but become more difficult to understand for sugar beet and soybeans. The high C.V. (Coefficient of variation) for the production can be due to higher yields in larger, more specialised farms.

A correction of the production estimate can be made using the difference of area estimates between the segment survey and the farm survey. A regression estimator approach might be a good solution.

Livestock is seriously underestimated since many livestock owners do not have agricultural land. A mixed approach was used for cattle and pigs with an exhaustive survey on list frame for the 50 largest farms and point sampling for the rest. The procedure works for pigs, but CV are not yet satisfactory.

Tab. 1: Results of the Segment Survey and the Farm Survey for main crops in Emilia Romagna (1990)

Emilia Romagna	Segment Survey		Farm survey				ISTAT
	Area*1000 Ha		Area*1000 Ha.		Prod* 1000 Tm.		Area
	Estim.	C.V. %	Estim.	C.V. %	Estim.	CV%	
Soft Wheat	212	5.7	208	6.9	1177	8	212
Durum Wh.	46	14.9	48	15.2	260	14	72
Barley	43	11.2	50	17.7	184	17	38
Rice	-	-	4	59	23	61	6
Sugar beet	111	7.1*	96	9.6	5474	28	119
Soybeans	76	6.0*	55	11.6	321	39	47
Vineyards	78	13.3*	76	18.7			75
Orchards	91	13.1*	96	19.7			85

*: Estimate corrected by regression on classified satellite image.

ISTAT: Official statistics. No precision provided, unpublished methodology.

Tab. 2: Results of the Farm Survey on Area Frame and Mixed Frame for Livestock in Emilia Romagna (1990)

Emilia Romagna	Census	Area frame		Mixed frame	
	* 1000	Estim.	C.V.%	Estim.	C.V.%
Cattle	869	829	14	894	13
Pigs	1876	1312	37	1818	27
Sheep	90	38	74		

4.2 Other regions 1990.

The same procedure was tested in 1990 in three other regions: Oberpfalz-Niederbayern (Germany), Makedonia Kentriki-Dytiki (Greece), and Valladolid-Zamora (Spain). In Spain, unidentified farms or non respondents were substituted using further points in the template.

Reasonable results are obtained for extensive crops, but the precision of the estimates should be improved by enlarging the sample. The intra-segment variance turns out to be generally negligible (Carfagna, 92), which means that raising the number of segments is more efficient than sampling more points per segment.

Tab. 3: Results of the Segment Survey and the Farm Survey in several regions(1990)

	Segment survey		Farm survey			
	Area	CV %	Area	CV %	Prod	CV %
Germany						
Wheat	153	1.9	161	8.7	1042	10.3
Barley	71	5.6	56	12.1	226	11.7
Rape	45	9.9	40	18.4	142	18.6
Sug.beet	32	3.0	33	20.6	647	28.4
Spain						
Cereals	639	2.8	663	5.0	1207	5.8
Wheat	106	7.8	123	12.2	220	13.4
Barley	501	3.3	502	6.1	943	6.8
Greece						
Soft wheat	160	8.2	181	15.9	301	16.1
Durum wheat	233	5.5	190	14.5	258	19.1
Cotton	46	10.4	42	26.6	93	31.0

Tab. 4: Sample size of the Segment Survey and the Farm Survey in several regions(1990)

	Segment survey	Farm survey		
	segments	segments	valid data	missing
Germany	449	155	358	59
Spain	460	152	381	-
Greece	445	134	486	138

4.3 Czech Republic 1992.

Area frames seem especially useful in the former communist countries in Europe because of the rapid change of property structure. Agricultural statistics are mainly produced with no sampling error by adding the data reported by each state farm or co-operative. This procedure will collapse in the next years. It will be extremely difficult to have an idea of the number of existing farms, and an agricultural census will be out of date before the data are elaborated. Area frames might be the best alternative.

In 1992, a survey was made with segments of 400 Ha.. and a grid of 5 points in each segment, giving 2085 points: 858 non-agricultural, and the rest from 458 farms. No missing data were recorded.

Tab. 5: Results of the Segment Survey and the Farm Survey in the Czech Republic(1992)

	Segment surv.		Farm survey				CSO	
	Area	CV	Area	CV	Prod	CV	Area	Prod
Wheat	824	5.4	757	3.7	3412	4.9	780	3413
Barley	655	5.1	630	3.8	2521	4.3	640	2512
Rapeseed	140	11.6	137	6.8	310	7.5	136	296
Sugar beet	119	11.5	127	8.1	4172	11.0	125	3874
Maize	361	7.5	326	4.8	8884	4.3	361	8904
Potatoes	109	13.6	92	7.9	1706	8.7	111	1969

CSO: Czech Statistical Office

ACKNOWLEDGMENTS

We are grateful to the various National and Regional Administrations that have collaborated for this work, and to E. Carfagna and A. Burrill for their kind revision of the paper.

REFERENCES

Allen, J.D., 1990, A Look at the Remote Sensing Applications Program of the National Agricultural Statistics Service. Jour. of Official Stat. Vol 6, n. 4, pp. 393-409.

Carfagna E., Delincé J., 1992, Farm survey based on area frame sampling. The case of Emilia Romagna in 1990.Conf. Appl. of Remote Sensing to Agric. Stat.(Belgirate). Off. Publ. of the E.C. Luxembourg.

Carfagna, E. Ragni, P., Rossi, L., Terpessi C., 1991, Area Frame: un Nuovo Istrumento per la Realizzazione delle Statistiche Agricole in Italia. Contributi alla Statistica Spaziale. Univ. Parma.

Cochran W., 1977, Sampling Techniques. New York: John Wiley and Sons

Cotter, J. Nealon J. 1987, Area Frame Design for Agricultural Surveys. U.S. Dept. of Agriculture. Nat. Agr. Stat. Serv.

Dicorato F. , 1993, AIS Estimation Programs. User Documentation. JRC Ispra.

EUROSTAT 1991. Working Party "Crops Products Statistics". Methodological reports. Doc. AGRI/PE/333. Luxembourg.

Gallego F.J., Rueda C., Delincé J., 1993, Estimating Land Use Area through Remote Sensing: Stability of Regression Correction. Int. J. of Remote Sensing. (In print)

Gallego, F.J., Delincé, J., 1993. Area estimation by segment sampling. In Euro-Courses: Remote Sensing applied to Agricultural Statistics.