United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

Staff Report
Number SSB-88-01

January 1988

A Preliminary Look At A National Area Sample Allocation

James W. Mergerson

A PRELIMINARY LOOK AT A NATIONAL AREA SAMPLE ALLOCATION. James W. Mergerson, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C., 20250, January 1988, Staff Report No. SSB8801.

ABSTRACT

This paper presents the methodology and results of a multivariate sample reallocation analysis of the June area frame sample. Results indicate a 25% reduction in the number of sampling units can be obtained without loss of national-level precision but with some substantial changes in precision at state levels. Cost savings would be in excess of $400,000.

Keywords: Multivariate allocation, stratified sampling, survey design.

ACKNOWLEDGMENTS

Thanks are expressed to Stan Mason who wrote a C program which greatly facilitated the creation of input data and to Robert DeGeus who provided many valuable suggestions and input.

CONTENTS

A PRELIMINARY LOOK AT A NATIONAL AREA SAMPLE ALLOCATION

James W. Mergerson

INTRODUCTION

The purpose of this paper is to present the results of a preliminary sample reallocation analysis for the area frame. The primary purpose of the analysis was to determine the reduction in sample size if we were only interested in preserving current national precision levels. Results are summarized by state. The reallocation was performed using a multivariate optimal allocation algorithm and is based on the constraint that current national precision levels (not state levels) for major items be maintained. Results indicate the total number of area sampling units in the June Survey can be reduced by 25% without loss of precision at the national level. However, this would result in some substantial loss in precision for some items in some states.

BACKGROUND

Sample reallocation analysis is computationally intensive. Bethel [2] provided the Agency with a multivariate allocation algorithm which was implemented on a Zilog System 8000 microcomputer. However, Bethel's implementation was cumbersome to use since many hand calculations were required to create the input. Mergerson and others [7] improved this algorithm and also implemented an improved version on an IBM-PC/AT. However, due to compiler limitations on both microcomputers, it was not possible to perform national-level sample allocations on either microcomputer. Additional modifications were made to a version of the Zilog-based allocation program and a procedure was developed to upload the program for execution on a mainframe computer using a remote job entry facility.

The program, written in PASCAL, requires the following inputs: a page heading to accompany the output of the program, the number of strata, the number of survey items to be included in the allocation analysis, a convergence criterion, the average data collection cost per sampling unit for each stratum, total number of sampling units in each stratum, estimated standard deviations by stratum and survey item, the maximum acceptable coefficient of variation (CV), and an estimate of the population total for each item.

OVERVIEW OF ALLOCATION METHOD

A linear cost function is minimized subject to a nonlinear constraint for each item included in the allocation analysis. The program executes an iterative algorithm which converges rapidly. Less than 20 iterations is typical. The convergence criterion is the maximum relative constraint violation which is based on the maximum acceptable CV for each survey item. For example, if a required CV is C, setting a convergence criterion of epsilon (ε) would mean that CV must be no larger than $C * (1 + \varepsilon)$.

The allocation model used in this analysis is as follows:

$$\text{minimize} \quad C = \sum_{i=1}^{430} c_i / x_i$$

$$\text{subject to:} \quad \sum_{i=1}^{430} a_{ji} * x_i \leq 1 \ , \quad 1 \leq j \leq 10$$

where

$$a_{ji} = \{w_i^2 * s_{ij}^2\} / \{v_j + \sum_{i=1}^{430} w_i^2 * s_{ij}^2 / N_i\}$$

and

$$x_i = \{ 1/n_i \text{ if } n_i \geq 1 \ , \text{ infinity otherwise } \}$$

$$w_i = N_i / N$$

$$N = \sum_{i=1}^{430} N_i$$

s_{ij} = standard deviation of survey item j in strata i

$v_j = ((CV_j * est_j) / N)^2$

CV_j = desired coefficient of variation for survey item j

est_j = estimated population total of survey item j

c_i = average data collection cost in stratum i

N_i = total number of sampling units in stratum i

n_i = optimum sample size for stratum i

This model will provide an allocation which provides the desired CV levels simultaneously for all the survey items analyzed. More in depth technical details concerning multivariate optimal allocation can be found in [1], [3], [4], [5], and [6].

LIMITATIONS

Some limitations of this analysis are as follows: (1) the analysis pertains only to the June survey and does not consider follow-on surveys, (2) current state-level CV's are not maintained, (3) only ten items were included in the analysis, and (4) economic items were not considered. Data from a follow-on survey and economic items will be considered in future analysis. Ability to simultaneously consider both state-level and national-level CV constraints require additional research and development.

Another limitation of this analysis is that it is not based completely on the survey design. That is, the area frame sub-stratification was ignored due to complexity of inserting constraints at this time that the sample size be equal for all substrata in a given stratum. Standard deviations were computed at the stratum rather than the sub-stratum level. However, this limitation should have very little impact on the results. Target CV's were based on national level CV's which were recomputed ignoring the sub-stratification. At the national level, CV's computed ignoring sub-stratification versus CV's computed considering sub-stratification differ only slightly.

Also, the analysis was performed relative to the area frame closed estimator (full frame and non-overlap component). The weighted non-overlap (NOL) components for livestock, grain stocks and crops were not considered at this time since they were not available for all states in 1986. These components will be considered in future analysis using 1987 June and September data.

DATA SOURCE

Analysis was performed using 1986 June Survey area frame data. The ten survey items included in the analysis are as follows: total hogs and pigs (NOL component), total cattle and calves (NOL component), upland cotton, corn, winter wheat, sorghum, soybeans, oats, barley and rye. Input standard deviations by stratum and item, population sizes by stratum, and national-level estimates and CV's were obtained from the 1986 June Survey. The estimates and CV's used as input to the analysis are shown in Table 3.

DISCUSSION

Multivariate sample reallocation analysis was performed based on 1986 June Survey estimates and precision levels. The performance measure is the reduction in sample size which results in the same national levels of precision for each of the items considered. Analysis results are shown in Table 1. Current allocations versus optimal allocations are listed by state.

The allocations for seven of the states are greatly influenced by the inclusion of oats, barley and rye in the analysis. If these items were excluded from the analysis, the allocations for these states would be substantially smaller. However, these states are some of the top producers of oats, barley or rye.

Considering the contributions to the national estimates for each state for items included in the analysis, the allocations in a relative sense appear to be very reasonable. Some may question the large reductions in California and Florida. However, considering the contribution of these states to the national-level estimates for the items considered, the stated allocations are very reasonable.

At the national level, the total sample size is reduced from 15,663 sampling units to 11,622 units. Total enumeration cost could be reduced from about $1,900,000 to about $1,500,000 based on the average cost per sampling unit in each state. This is a savings of about $400,000 for the June Survey. Additional dollars would be saved due to a reduction in area frame maintenance activities resulting from a reduction in the number of sample units for rotation.

Current and expected state level CV's under an optimal national-level sample reallocation for selected items in five states are shown in Table 2. As expected, states with decreased sample sizes have larger CV's and states with increased sample sizes have some smaller CV's. The increase or decrease in some state-level CV's is substantial.

CONCLUSION

Based solely on the present national precision levels relative to the area frame closed estimator for ten selected items, many state-level allocations are far from optimum. Dollars could be saved by a more efficient national sample allocation of fewer sampling units or greater precision could be obtained by a more efficient allocation of the current 16,000 units. Additional analysis will be conducted using 1987 June and September data before making definite recommendations concerning the current area frame sample allocations.

REFERENCES

[1] Bethel, J.W. (1985). Sample Design for the 1985 ISP/JES. USDA-SRS Staff Report, SF&SRB-86.

[2] Bethel, J.W. (1986). An Optimum Allocation Algorithm for Multivariate Surveys, USDA-SRS Staff Report, SF&SRB-89.

[3] Hartley, H.H., and Hocking, R.R. (1963). Convex Programming by Tangential Approximation. Management Science, 9, 600-612.

[4] Huddleston, H.F., Claypool, P.L., and Hocking, R.R. (1970). Optimum Sample Allocation to Strata Using Convex Programming. Journal of the Royal Statistical Society C, 19, 273-278.

[5] Kokan, A.R. (1963). Optimum Allocation in Multivariate Surveys. Journal of the Royal Statistical Society A, 126, 557-565.

[6] Kokan, A.R., and Khan, S. (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution. Journal of the Royal Statistical Society B, 29, 115-125.

[7] Mergerson, J.W., Clark, M., Fenley, B. (1986), Optimum Allocation for Multivariate Surveys: An Improved Implementation, Technical Note: USDA-NASS-AFS-86-01.

Table 1 -- National multivariate optimal reallocation - based on area frame closed estimator - using 1986 June survey data

State	Current Sample Size	Optimal Sample Size	State	Current Sample Size	Optimal Sample Size
Alabama	359	273	Nevada - *	100	87
Arizona	374	121	New Hampshire	30	23
Arkansas	400	256	New Jersey	250	80
California	911	434	New Mexico	292	142
Colorado	457	289	New York	380	161
Connecticut	48	35	North Carolina	391	271
Delaware	72	27	North Dakota - *	376	479
Florida	425	132	Ohio	324	226
Georgia - *	436	416	Oklahoma	360	386
Idaho - *	362	192	Oregon	372	214
Illinois	300	379	Pennsylvania	330	124
Indiana	324	266	Rhode Island	14	8
Iowa	298	397	South Carolina	335	166
Kansas	435	441	South Dakota - *	352	466
Kentucky	338	166	Tennessee	349	176
Louisiana	376	284	Texas	840	1045
Maine	150	40	Utah	324	80
Maryland	252	59	Vermont	70	23
Massachusetts	48	35	Virginia	343	211
Michigan	343	226	Washington	360	229
Minnesota - *	343	594	West Virginia	250	85
Mississippi	402	284	Wisconsin	310	201
Missouri	450	345	Wyoming	256	129
Montana - *	362	432			
Nebraska	390	487	Total	15,663	11,622

* - The allocation for these states would be much less if oats, barley, and rye were not included in the analysis.

Table 2 -- Current and expected coefficients of variation (CV's) for selected survey items in five states.

State	Survey Item	Current CV (%)	Expected CV (%)
California	Hogs (NOL)	31.0	54.4
	Cotton	9.3	13.4
	Corn	11.5	16.7
	Wheat	12.1	19.1
	Barley	20.0	23.9
Montana	Hogs (NOL)	85.5	80.1
	Wheat	10.8	10.7
	Oats	24.8	25.2
	Barley	9.5	7.8
Nebraska	Corn	5.1	4.5
	Wheat	8.5	7.7
	Sorghum	10.5	9.4
	Soybeans	7.8	7.0
	Oats	11.3	10.5
North Carolina	Cattle (NOL)	13.5	16.5
	Cotton	32.2	33.2
	Corn	7.6	8.8
	Wheat	11.9	13.1
	Soybeans	8.0	9.4
Maryland	Cattle (NOL)	14.1	28.3
	Corn	5.7	11.7
	Wheat	10.4	20.6
	Soybeans	8.0	15.8

Table 3 -- Input estimates and coefficients of variation (CV's) ignoring sub-stratification

Item	Estimate (000)	CV's
Hogs (NOL)	10,650	0.087
Cattle (NOL)	32,300	0.027
Cotton	9,950	0.038
Corn	76,650	0.012
Wheat	53,450	0.019
Sorghum	14,900	0.036
Soybeans	62,250	0.015
Oats	14,600	0.022
Barley	13,550	0.030
Rye	1,900	0.070