NASS/USDA Area Frame Sample Allocation for Estimation of Number of Farms Not on the 2007 Ag Census Mailing List

Raj S. Chhikara[1], Floyd M. Spears[2], Charles R. Perry[3], Phil S. Kott[3]
University of Houston-Clear Lake[1]
Harding University[2]
USDA National Agricultural Statistics Service[3]

Abstract

USDA/NASS needs to estimate the number of farms that are not on the census mailing list (NML) for the 2007 Ag Census. Additional sample segments are planned during its 2007 annual ag survey to determine estimates for several NML items (number of farms, and subdomains corresponding to several minority and specialty farms). Stratum standard deviations from the 2002 area frame sample data are modeled in terms of certain NML farm characteristics. The NASS multivariate allocation procedure is applied to determine sample allocations for different sets of ag items, which include NML farms and the regular annual ag survey items. Various combinations of the actual 2007 design allocation and those obtained using the NML items are investigated. The allocation that most closely meets the sample size and precision goals is developed to determine the set of supplemental samples for the USDA 2007 area frame sample allocation.

Keywords: 2007 Ag Census; Area Frame Sample Allocation; NML Farms and its Sub-domains; Stratum Standard Deviation; Multivariate Allocation; Board CV.

1 Introduction

The United States Department of Agriculture (USDA) needs to estimate the number of farms not-on-the census mailing list (NML) for the agriculture census to be conducted in 2007. It also wants to estimate the NML farms counts for the following targeted domains: Black-operated farms, Asian-operated farms, Native American-operated farms, Hispanic-operated farms, female-operated farms, vegetable farms, fruit farms, nursery farms and Christmas-tree farms.

A reliable estimation of the total number of NML farms and its breakdown for the different targeted domains is possible if the NML farms count and the specified NML domains are included as items of interest in the USDA multivariate sample allocation process. For an optimum allocation, an input of stratum standard deviations for these items is required to carry out the multivariate allocation procedure. So, we consider here the estimation of stratum variances for the NML farms count and use the 2002 USDA survey data since these have additional data from the 2547 Agricultural Coverage Evaluation Survey (ACES) segments available for estimation of the NML items. The approach involves modeling and prediction of stratum standard deviation for the total number of NML farms and for each of the targeted domain counts as described in Section 2.

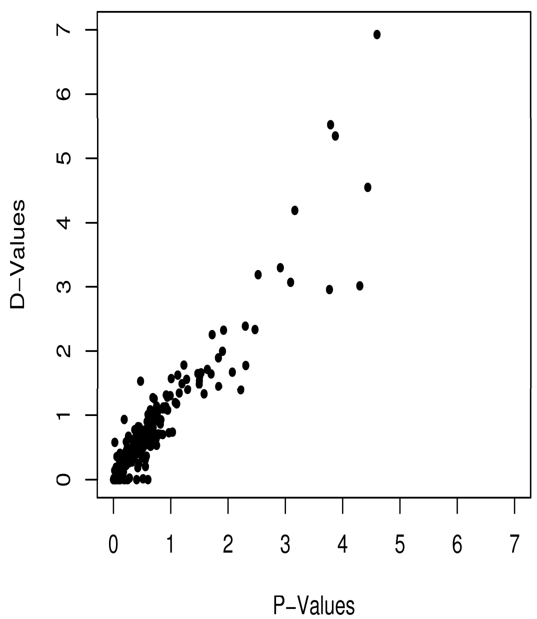
The objective of the 2007 area frame sample survey is to achieve a "Board CV" of 0.5 percent at the national level for the estimates of total number of NML farms, and that of 5.0 percent at the national level for each targeted domain of NML farms. (Here "Board CV" for a NML domain total is defined as its standard error divided by the total number of domain farms, which includes both the NML farms and the census list farms in the domain.)

The current multivariate allocation is made using the stratum standard deviations for the regular eight items that National Agriculture Statistics Service (NASS) surveys each year. This NASS multivariate allocation procedure will be carried out by replacing these eight items with the NML farms and the NML domain farms. The result of this allocation will then be combined with the sample allocation obtained for the regular eight items; and subsequently, a supplement to the regular NASS 2007 area-frame samples will be determined.

2 Stratum Variance Estimation

Data from the NASS 2002 Annual Ag Survey were used to determine the stratum variances for each of the ten NML items considered. One set of stratum variances consisted of directly survey computed values, called the d-values. Another set of stratum variances was made of those computed from the predicted values, called the p-values. These values were obtained from the empirically developed NML models by Chang and Kott (2004). They utilized the 2002 area frame survey data to model the probability of an area-frame farm not being on the census mailing list and used covariates, including gender, ethnicity and farm type. A logistic regression methodology was the basis for developing model-fits for the empirical NML models.

Figure 1: Plot of the paired d-values and p-values for stratum standard deviations of the total number of NML farms across all land use strata in U.S.



When the two estimates of d-value and p-value for stratum standard deviation are compared, it exhibited a strong linear relationship for each NML item. For example, depicted in Figure 1 is a plot of the paired values for stratum standard deviation of the total number of NML farms across all land use strata in U.S. It is seen that the d-values are on the average varying more than do the p-values. This is an expected outcome since a NML model-fit is likely to predict values that have less variability than those directly computed from the survey data.

When estimated based on a small sample size, the directly survey computed stratum standard deviation is not reliable and needs to be improved upon. So we considered modeling the direct survey computed values using the NML model predicted value as a covariate. Since the covariate values are more stable

Table 1: Model Fits for NML Items

NML Item	Slope	R-Square
Total	1.12	0.931
Asian	1.84	0.934
Black	1.44	0.983
Christmas Tree	1.47	0.973
Female	1.55	0.971
Fruit	1.74	0.969
Hispanic	1.57	0.939
Indian	1.54	0.919
Nursery	1.71	0.892
Vegetable	1.66	0.918

and less varying than the response values, this modeling of stratum standard deviation should smooth any anomalies that exist in the directly computed survey estimates and the model-fits should result into more reliable predicted values. This in turn shall safeguard against the use of unreliable stratum standard deviations in the determination of sample allocation to strata.

Table 1 lists the model-fit characteristics, Slope and R-Square, for each of the NML items. It indicates a strong linear relationship with a correlation coefficient close to 1. The model-fits are used to predict its stratum standard deviations for the different NML items. These model-predicted stratum standard deviations, called the m-values, then provided a third set of stratum standard deviations for each NML item. Thus three different stratum standard deviations were employed in this investigative study of sample allocation for the 2007 area frame sampling of segments by NASS. The three sets of estimates are referred to as follows:

1. **d-values** which are directly survey computed stratum standard deviations,

2. **p-values** which are predicted from the empirical NML model developed by Chang & Kott (2004),

3. **m-values** which are predicted from the model-fits described in Table 1.

3 Sample Allocations

The following three different ways of sample allocation were considered for the NML farm items to develop the final sample allocation. The maximum (by stratum) of the regular eight items NASS 2007 sample design and the allocation from each of the following cases A - C was obtained.

Table 2: Sample allocations by stratum group for cases A, B and C using (a) d-values, (b) p-values and (c) m-values to estimate NML standard deviations.

	Case A		Case B		Case C	
Stratum Group	Max of 10 Univ. NML's (A1)	Max of NASS Design & A1 (A2)	Mult. with 10 NML's (B1)	Max of NASS Design & B1 (B2)	Mult. with 18 Items (C1)	Max of NASS Design & C1 (C2)
(a) Using d-values						
10's	3653	6545	3057	6339	6038	6630
20's	3924	4294	3336	3872	3689	3883
30's	767	844	616	700	554	634
40's	4214	4431	3619	3922	3176	3435
50's	96	104	96	104	96	104
Total	12654	16218	10724	14937	13553	14686
(b) Using p-values						
10's	1886	6067	1808	6041	6026	6575
20's	2002	3010	1885	2987	2644	3076
30's	440	521	391	472	355	436
40's	2292	2701	2229	2664	1826	2261
50's	96	104	96	104	96	104
Total	6716	12403	6409	12268	10947	12452
(c) Using m-values						
10's	2751	6314	2303	6192	6011	6559
20's	3227	3815	2728	3474	3305	3580
30's	767	834	651	728	608	687
40's	3601	3879	3122	3491	2752	3036
50's	96	104	96	104	96	104
Total	10442	14946	8900	13989	12772	13966

Case A: Univariate allocations for NML items

1. Perform univariate allocation for each of the 10 NML items.
2. Take the maximum of these 10 allocations by stratum and state to obtain the combined univariate allocation.

Case B: A single multivariate allocation for 10 NML items

Case C: A single multivariate allocation for the regular 8 ag items and all 10 NML items together.

In each of these cases, three sets of stratum standard deviations (d-values, p-values and m-values, as discussed in Section 2) were used to perform each of the allocations for the NML items. Table 2 lists the sample allocations obtained at the stratum group level for each of the cases A - C using (a) d-values, (b) p-values, and (c) m-values for the NML farm item stratum standard deviations. There are five stratum groups, which correspond to the land use strata numbered in 10's, 20's, 30's, 40's and 50's.

Each of these allocations meets the goal of an achievable Board CV of 0.5 percent at the national level for the total number of NML farms and of 5 percent or less at the national level for the targeted NML domains with two exceptions. The black-operated farms have the achievable CV slight above 5 percent, and the Asian-operated farms have the achievable CV close to 6 percent.

4 Proposed Sample Allocation

The univariate NML allocations vary substantially for strata across various NML items. On the other hand, a single multivariate allocation for the ten NML items may smooth these out and thus seems more appropriate. For the 2007 agricultural survey, the sample allocation for the regular eight ag items is combined with that for the ten NML items. This is done by taking the maximum of the two sample allocations for each land use stratum in each state as outlined in Case (B) above.

NASS has set a target of about 3000 additional segments for its supplemental sample in 2007 in support of its achieving reliable estimates for the various NML items. With this requirement, along with the rationale of having a more stable stratum sample allocation, the following approach to allocation is proposed and is carried out to develop the new 2007 sample allocation:

1. Carry out a model-fit of direct estimates vs NML model predicted values of stratum standard deviations, and use it to obtain model-predicted stratum standard deviations as described in Section 2.
2. Determine a single multivariate allocation for the 10 NML items using the model-predicted stratum standard deviations.
3. Finalize the allocation using the maximum (by stratum) of the allocation resulting from item 2 above and the regular 8 items NASS 2007 allocation design.

Table 3: Sample Allocations and Supplemental Samples by Stratum Group

Stratum Group	2007 Design	Max of 2007 Design Mult. With 10 NML's	Supplemental Samples
10's	5905	6192	287
20's	2775	3474	699
30's	347	728	381
40's	1767	3491	1724
50's	104	104	0
Totals:	10898	13989	3091

The current 2007 design allocation and the proposed new 2007 design allocation are tabulated. The results are given along with the supplemental samples by Stratum Group in Table 3 and by State in Table 4. For full details, see Chhikara, Spears, Perry and Kott (2006).

5 Conclusion

A total of 3091 supplemental samples are required, of which 287 are in stratum Group 1; 699 are in Group 2; 381 are in Group 3 and 1724 are in Group 4 strata. There is a 110 percent increase in stratum group 3, and a 98 percent increase in stratum group 4 when compared to the current 2007 sample design.

There is a substantial increase in sample allocations in California, Florida, Michigan, New Mexico, Oregon, and Texas. The substantial increases in these states are due to one or more NML items predominant there. Listed next to each state in Table 5 is the primary NML item(s) for which reliable estimation would require such an increase in sample allocation. Reason for increase as stated herein is determined by

Table 4: Sample Allocation and Supplemental Samples by State

State FIP	State Abbrev.	Regular 2007 Design	Proposed 2007 Design	Supp. Samp.
	Totals:	10898	13989	3091
1	AL	236	272	36
4	AZ	118	158	40
5	AR	328	438	110
6	CA	404	919	515
8	CO	267	291	24
9	CT	8	11	3
10	DE	23	29	6
12	FL	100	219	119
13	GA	290	382	92
16	ID	148	208	60
17	IL	401	401	0
18	IN	264	264	0
19	IA	452	452	0
20	KS	487	487	0
21	KY	189	227	38
22	LA	249	276	27
23	ME	32	63	31
24	MD	61	61	0
25	MA	12	37	25
26	MI	145	252	107
27	MN	393	419	26
28	MS	298	395	97
29	MO	383	458	75
30	MT	316	316	0
31	NE	473	473	0
32	NV	26	30	4
33	NH	10	39	29
34	NJ	48	59	11
35	NM	124	229	105
36	NY	96	143	47
37	NC	319	415	96
38	ND	420	420	0
39	OH	220	308	88
40	OK	335	371	36
41	OR	194	426	232
42	PA	179	225	46
44	RI	8	8	0
45	SC	119	228	109
46	SD	395	395	0
47	TN	334	334	0
48	TX	1120	1745	625
49	UT	69	102	33
50	VT	21	52	31
51	VA	179	223	44
53	WA	267	307	40
54	WV	66	77	11
55	WI	219	262	43
56	WY	53	83	30

Table 5: Primary NML Items Causing Substantial Increases in Sample by State

St.	Inc.	(%)	NML Items
AR	110	(34)	Total, Asian, Black
CA	515	(128)	Asian
FL	119	(119)	Total, Black, Hispanic
GA	92	(32)	Total, Black
MA	25	(208)	Total, Nursery, Chr. Tree
ME	31	(97)	Total, Chr. Tree
MI	107	(74)	Total
MS	97	(33)	Black
NH	29	(290)	Total, Vegetable
NM	105	(85)	Total, Hispanic
NC	96	(30)	Total, Asian, Black
OH	88	(40)	Total, Chr. Tree
OR	232	(120)	Total, Asian, Chr. Tree
SC	109	(92)	Total, Black
TX	625	(56)	Total, Black
VT	31	(148)	Total, Hispanic

identifying the NML item for which the univariate allocation in a state were substantially higher compared to the univariate allocations for other NML items.

Listed in Table 6 are the specific strata that have substantial supplemental samples when compared to the regular 2007 design allocation. Reason for increase as stated herein is determined by identifying the NML item for which the univariate allocation in a stratum was substantially higher compared to the univariate allocation for any other NML item. NML farm total is the primary cause of many increases, especially in the stratum group of 40's strata. Asian-operated farms are the cause of large increases in California for several strata across all stratum groups.

The actual NASS sample allocation to be used for the 2007 Agricultural Coverage Evaluation Survey (ACES) samples takes into account targeted sampling when a large increase in sample size is needed to estimate Asian-operated or black-operated NML farms. The allocation also looks more closely at state-level Board CV's for NML farms. Thus the actual NASS allocation varies slightly from what was recommended in this report.

Acknowledgements

Work on this project was supported in part under a cooperative research program of the United States Department of Agriculture at the University of Houston-Clear Lake. The authors would like to thank George Hanuschak and Roberta Pense for their keen interest and support for the project. The authors would also like to thank Bill Wigton and Herb Eldridge for their suggestions and making the research available.

References

Chang, Theodore, and Kott, Phillip S. (2004). "Modeling NML Using the Area Frame Survey" Technical Manuscript. August 27, 2004, USDA/NASS, Washington, D.C.

Chhikara, Raj S., Spears, Floyd M. and Perry, Charles R. (2002), "Sample Allocation for Estimation of the Number of *Not on Mail List* (NML) Farms for the 2002 Census of Agriculture," USDA-NASS RDD Research Report No. RDD-02-01, June, 2002.

Chhikara, Raj S., Spears, Floyd M., Perry, Charles R. and Kott, Phil (2006), "Supplemental Sample Allocation for 2007 Area Frame Design for Estimation of NML Farms," USDA-NASS RDD Research Report No. RDD-06-03, November, 2006.

NML/ACES Sample Design Team (2005), "Decision Memorandum for the 2007 Agriculture Coverage Evaluation Survey (ACES) Sample Size", Technical Memo, February 11, 2005, NASS/USDA, Washington, D.C.

Table 6: Reason for Supplemental Increases by Strata

St.	Strat.	Suppl. Samp.	Increase due to
AR	21	67	Asian
	42	43	Total, Indian
CA	17	56	Asian
	21	99	Asian
	27	115	Asian
	31	162	Asian
	41	81	Asian
FL	22	20	Total, Black
	40	28	Total
	42	32	Total
GA	40	92	Total, Black
KY	40	24	Total
LA	40	27	Total, Black, Vegetable
ME	40	29	Chr. Trees
MA	40	25	Total, Chr. Trees, Nursery
MI	20	43	Total, Chr. Tree Black, Indian
	40	46	Total, Hispanic
MS	20	46	Black
	40	51	Total, Black
MO	40	75	Total
NH	40	29	Total, Vegetable
NM	13	82	Total, Hispanic
NY	40	41	Total, Chr. Trees
NC	40	88	Asian
OH	40	57	Total, Chr. Trees
OK	40	36	Total, Indian
OR	31	122	Asian
	41	65	Total, Chr. Trees
SC	40	71	Total, Black, Indian
TX	26	34	Total
	27	47	Total, Chr. Trees
	42	482	Total, Black, Chr. Trees
VA	40	44	Total, Black, Chr. Trees
VT	40	31	Total, Hispanic
WA	31	15	Total, Vegetable
WI	12	42	Chr. trees
WY	42	19	Total