ALTERNATIVE SAMPLING FRAME CONSTRUCTION FOR PRICE SURVEYS

Carol C. House, USDA, Nancy J. Carter, California State University, Chico,
and Douglas C. Bond, USDA

ABSTRACT

A method of building sampling frames for estimating prices paid by farmers was studied, as an alternative to the operational USDA program. Questions were added to an annual multiple frame survey of farmers to ask them where they bought certain items and to estimate how much they spent at each firm. The names of the firms made up frames that could be sampled with probability proportional to sales to farmers.

Major findings included: the research method was feasible; the operational frames were seriously incomplete; differences were indicated in prices estimated by the two methods; and the research pps estimators were not very robust. Problems that arose, which are addressed in this report, include small sample sizes and rarity of questionnaire items.

INTRODUCTION

The purposes of this research were to study several actual or suspected problems with USDA's surveys for estimating prices paid by farmers, and to try a different way of building sampling frames for these surveys.

At the time of this research, USDA estimated prices paid by farmers for farm production items with two stage samples of firms that were thought to sell these items. The sampling frames were lists based mainly on telephone yellow pages, and they were updated by using other sources of information. This method had several real or potential problems: the frames were probably incomplete; they probably included firms that did not sell to farmers; there was no way to estimate firm size for probability proportional to size (pps) sampling; it was costly to build and maintain the frames; the frames quickly became outdated; and a price per item was desired, but not estimated. Instead, a price per firm was estimated. There is a detailed discussion of these problems in another report (2).

To study some of the problems of the operational method, and to offer some alternatives to it, a different method of building frames was tried. Questions were added to an annual USDA multiple frame survey of farmers to find out where they bought groups of commodities. Frames of businesses selling to farmers were constructed directly from this information. To make it possible to sample these frames with probability proportional to sales to farmers, the farmers were asked to tell what part of their expenditures went to each business. Samples were drawn pps from these frames, and prices paid by farmers were estimated.

The feasibility and advantages of the new method and the incompleteness of the operational frames were studied. Price estimates from the operational and research methods were compared. Finally, the robustness of the research estimators was studied.

In the next two sections, results are presented. Problems of small sample size and rare items arose during the study, and these are discussed, too. Results are followed by a description of the methods used. Then there is a discussion of the background for this research. Finally, closing remarks are made.

HIGHLIGHTS OF RESULTS AND RECOMMENDATIONS

The research method, gathering names of firms on an annual survey of farmers, was a feasible way to build frames of businesses selling products to farmers. Also, use of this method solved several problems of the USDA operational method: frame incompleteness, inclusion of firms not selling to farmers, and cost and aging of the frames. The operational frames were found to be, on average, about 50 percent incomplete. Significant, though small, differences were found in prices estimated by the two methods; this further suggested problems with the operational method.

The research estimators were not very robust. A simulation study showed that inaccurate estimates of firm size could significantly change price estimates. Since it was not possible to accurately estimate firm size for specific items, pps sampling is not considered feasible for the research method.

Small sample sizes and item rarity were problems with both the research and operational methods. Item rarity occurred when a firm completed a questionnaire for only a portion of the items. These problems could lead to large biases in estimates.

On the basis of the results, it is recommended that the operational method be replaced with a modified version of the research method. One possibility is to build frames from the annual survey of farmers, as was done in the research, but to retain these frames for several years and draw simple random samples (without pps sampling). Other possibilities are presented in a technical report by the authors (4). In all of these designs, a large enough sample size must be ensured.

DETAILED FINDINGS

Feasibility of Procedures

In general, the experience collecting names and addresses of firms and percent of expenditures at each was encouraging. It is felt that the feasibility of this method of frame construction was demonstrated.

Several problems were anticipated with collecting names and addresses of businesses on the annual survey of farmers. First, a farmer might simply forget where he bought something. Second, he might have trouble giving complete or correct names and addresses. For example, he might refer to a business called "A & A Feed" as "Jones Feed," since Mr. Jones owns the store. Or he might say that a business is in a certain community when in fact it should be listed under the postal designation of a nearby town. Third, it might be hard for the farmer to recall or estimate what percent of his expenditures went to each firm.

There was no way of quantifying the first problem. However, it can be assumed that if a farmer did forget where he bought something, he probably had a relatively small volume of purchases from that business. In written evaluations after the first year of the study, enumerators who collected the information reported that farmers generally did not have difficulty

listing places where they made purchases. The evaluations tended to indicate that respondents could in fact remember where they made their major purchases.

Enumerators also reported that farmers generally could report addresses in enough detail to use them in a sampling frame. If records were available during the interview, the respondent generally had receipts with letterheads containing complete addresses. In many cases, when the respondent did not give a street address, enumerators supplied it with information from telephone directories, post offices, or other respondents. Also, in one state, enumerators reported that street addresses were neither used nor needed by firms in many towns.

The biggest problem that enumerators found was collecting data on percent of expenditures. They reported that respondents would generally give their best guess, but that they were often unsure.

Incompleteness of Operational Frames

In the research method, names and addresses of businesses collected during the annual survey of farmers were used to build new frames for the surveys of prices paid by farmers. Using the farmer expenditure data that were also collected for each firm, an analysis was made of the incompleteness of the operational frames. It was found that a substantial part of total farm-related expenditures for feed, fertilizer and pesticides, new farm machinery, and farm supplies was made in businesses that were not on the operational frames. This is a serious problem since it creates the potential for large biases in price estimates.

On average, the operational frames were missing businesses where an estimated 46 percent of farm production expenditures occurred (Table 1). Averaging over the four commodity groups, the state estimates of incompleteness ranged from 38 to 52 percent. Even with farm machinery, where the operational frames were expected to be the most complete, state frames were estimated to be 13 to 34 percent incomplete. Farm supply frames were the most incomplete; about 75 percent of total expenditures was estimated to have been in businesses not on the operational frames.

Comparison of Price Estimates

Because of the incompleteness of the operational frames, an analysis was made to see if there were any differences between prices estimated from the operational frames and from the research frames. Statistically significant, though small, differences were found. However, due to limitations to the test results, the comparisons should be viewed as rough indications of the relationship between the two methods, rather than as definitive results.

Paired comparisons were made for each state, month, and price item whenever operational and research data were both available. A total of 841 comparisons was made. Several tests were used to compare the two methods, but only one test is presented here. See the technical report for the other tests.

Ninety percent confidence intervals were constructed for the operational and research price estimates. The percent of paired intervals which intersected was computed. If the two methods were measuring the same price, it would be expected that the confidence intervals would intersect at least 90 percent of the time.

Over all estimates, the confidence intervals intersected only 79 percent of the time (Table 2). At the state level, all but one state showed a significant difference in price estimates. The range was 64 to 81 percent in the four significant states, and the nonsignificant state had a 93 percent intersection rate. When compared by commodity group, the prices were again significantly different, except for farm supplies. The range was 76 to 84 percent intersection for the significant differences, while farm supplies had 95 percent intersection.

Even though the price estimates were, in general, significantly different, the differences were quite small. About half the comparisons had a relative difference (research estimate minus operational estimate, divided by operational estimate) of less than five percent.

There are limits to the inferences that can be made from these tests, for two reasons. First, the validity of the price estimates and confidence intervals is suspect because of small sample sizes. Second, there is a good chance that assumptions in pps sampling were violated, which could have caused increased bias in price estimates. These topics are discussed in more detail later in this paper and in the technical report.

If the significant differences are assumed valid, there are two possible interpretations: (1) there are differences in prices between firms selling to farmers and those not selling to farmers, and/or (2) the differences are due to the fact that a "price per firm" estimator was used in the operational method and a "price per item" estimator was used in the research method. It is probable that both interpretations are correct.

It can be concluded that there is evidence of differences in price estimates from the two methods. Given the results of the incompleteness analysis, this is more evidence of deficiencies in the operational method.

Robustness of Estimators

In the research method, it is assumed that size of firm is estimated unbiasedly. To study the effects of violating this assumption, weights reflecting alternative estimates of size were generated by simulation techniques. These weights were used to estimate prices, to see if the estimates were significantly changed. This was done for two items in one state.

Results were the same for both items, so the table for only one of these items is shown (Table 3). The table indicates whether the self-weighted estimate was in the 90 percent confidence interval or out of the 95 percent confidence interval of the average price estimated by using simulated weights. When the means of the generated weights were the same as the weights used in self-weighted estimators (mean of 1.0 for type 1 and 1.208 for type 2), the self-weighted estimators always fell in the 90 percent confidence intervals, regardless of the standard deviation of the weights. (Types 1 and 2 are defined in the "methods" section.) However, if the means of the generated weights were varied by as little as ± 0.1, the self-weighted estimators fell out of the 95 percent confidence intervals, for all standard deviations and for both types of estimators.

It can be concluded that if the assumption about estimating size of firm is violated such that the expected values of the weights mentioned above are off by as litle as ± 0.1, then the estimate of price per item will be seriously biased. Unfortunately, the analysis of

problems associated with estimating rare items indicates that size of firm cannot in general be accurately estimated for specific items, using the techniques outlined in this research study. Since the assumption of unbiased estimation of size no longer seems reasonable, pps sampling from the research frames is not recommended.

Problems of Small Sample Sizes and Rare Items

In both the operational and research prices paid surveys, most responding firms reported prices for only a small portion of the surveyed items. This probably means they did not sell these items. Since analysis showed over 50 percent of the operational estimates and almost 40 percent of the research estimates were based on less than 10 responses, many of the surveyed items are considered "rare items."

It was found that 30 percent of the operational estimates and 21 percent of the research estimates were made with 5 or less observations. The frequency with which operational estimates were based on 10 or less observations was 14 percentage points more than for the research estimates. A possible reason is that firms identified by farmers in the frame building process are more likely to carry a wide range of products used by farmers than firms identified by the operational frame building process.

Estimating prices of rare items can cause a number of problems. First, the estimates are generally more variable than those based on more observations. Second, the estimators have a bias which is negligible for large sample sizes but which could be large for sample sizes as small as in these surveys. This is discussed more fully in the technical report. Third, assumptions about size of firm can be violated. A firm is selected for the survey with probability proportional to sales of a group of commodities. If it does not sell a particular item in that group, the size of firm for that item is really zero. But the size that is used in the research method is based on sales of the entire commodity group. In the section on robustness of estimators it was shown that such a violation of an assumption can cause serious biases in price estimation.

Since there were small sample sizes and so many rare items on both the operational and research prices paid surveys, it is likely that price estimates were affected. In future uses of the research method, adequate sample sizes should be ensured. Also, pps sampling should not be used. Instead, a price per firm should be estimated (as in the operational method) to avoid biased estimates of price per item.

METHODS

For this research, sampling frames for the prices paid surveys were built in the following way. Questions were added to the annual Farm Production Expenditure Survey (FPES) conducted by USDA. The FPES is a survey of farmers used to estimate annual expenditures for farm production and to determine the relative importance of various groups of expenditures. It is a multiple frame (list and area) survey, designed to provide efficient, complete coverage of the population of farm operations.

The questions were added at the end of each section of the FPES questionnaire, where expenditures were recorded for a given group of commodities. The respondents were asked to give the names and addresses of all firms where the reported purchases were made. They were then asked, for each commodity group, to report the percent of total expenditures made at each firm. Enumerators were trained to probe until enough firms were reported to account for at least 95 percent of the total expenditures for the commodity group. Sampling frames of businesses were built directly from this information. One frame was built for each of the five commodity groups: farm supplies, fuel and motor supplies, farm machinery, feed, and fertilizer and pesticides. The study was conducted in five states in 1980 and 1981.

To measure the incompleteness of the operational frames, the names and addresses on the research frames were matched against those on the operational frames. This was done by commodity group. For example, a firm on the research frame for feed was only checked to see if it was on the operational frame for feed, regardless of whether it was listed for another commodity group, such as farm machinery.

The incompleteness was estimated in the following way. For each state, year, and commodity group, farmers' expenditures were expanded to the state level for overlap and nonoverlap firms. (Overlap firms are those that are on both the research and operational frames, and nonoverlap firms are only on the research frames.) Then percent incompleteness was estimated by

$$\text{Percent Incompleteness} = (E_{nol} / E_{nol+ol}) \times 100$$

where

E_{nol} = Expanded expenditures, nonoverlap firms

E_{nol+ol} = Expanded expenditures, nonoverlap and overlap firms.

This was estimated for every commodity group except fuel and motor supplies, since there were no operational frames for it at the state level.

Estimates of prices from the two methods were compared in the following way. Firms from the research frames were sampled with probability proportional to estimated size of firm for a particular commodity group. The same questionnaires were used as on the operational surveys. Ninety percent confidence intervals were constructed where price estimates were available from both the operational and research methods, for a given state, month, and item.

In the operational method, cluster sampling estimates of means and variances of price per firm were used to construct confidence intervals. In the research method, the mean price $\bar{Y}$ was estimated by

$$\hat{\bar{Y}} = \sum_{i=1}^{n-m} \bar{p}_i/(n-m)$$

where

$\bar{p}_i$ = average price reported by firm i
n = number of sampled firms
m = number of sampled firms that did not report a price for the item.

This is a simplified version of a more complex formula. This simplification was achieved by assuming that size of firm, in terms of number of units of an item sold, is estimated unbiasedly. For

the purpose of comparing estimated prices, the variance of the research mean was estimated by treating an item that was missing from a completed report as a zero value. The technical report gives the formula for this variance.

To see how robust the research price estimators are when the assumption of unbiased estimation of firm size is violated, a simulation study was conducted. Two types of estimators of average price were considered. With one, called the type 1 estimator in this report, nonresponse for an item was treated by assigning a value of zero for its price. With the other, the type 2 estimator, nonresponse was treated as though the firm had never been sampled for that item. In either case, if the assumption is true, the estimator of average price is given by $\bar{Y}$ above.

However, if the assumption is not true, a weighted estimator must be computed. It is of the form

$$\sum_{i=1}^{n-m} \bar{p}_i w_i/(n-m)$$

where w_i is the weight for firm i based on its size and p_i, n, and m are defined as before.

Weights were generated for the type 1 and type 2 estimators from the following distributions: normal truncated at zero with standard deviations of 0.3, 0.4, 0.5, 0.75, and 1 and means of 1 $\pm$0.1 and (n/(n-m)) $\pm$0.1, and chi-square with the same means and standard deviations. For each distribution, 50 weights were generated, and price per item $\bar{Y}_w$ was estimated as follows:

$$\hat{\bar{Y}}_w = \sum_{h=1}^{50} \hat{\bar{Y}}_{wh}/50$$

where

$$\hat{\bar{Y}}_{wh} = \sum_{i=1}^{n-m} \bar{p}_i w_{ih}/(n-m)$$

is the estimated price per item from generation h, where w_{ih} is the weight for firm i from generation h. The values of $\bar{p}_i$ came from two data sets of actual reported prices.

Confidence intervals were constructed as

$$\hat{\bar{Y}}_w \pm t_\alpha s_w/\sqrt{50}$$

where

$$s_w = \left(\sum_{h=1}^{50} (\hat{\bar{Y}}_{wh} - \hat{\bar{Y}}_w)^2/49 \right)^{\frac{1}{2}}$$

and t_α is the appropriate normal deviate value.

Then it was determined if $\bar{Y}$ fell inside 90 percent confidence intervals or outside 95 percent confidence intervals, for the different means and standard deviations. In this way, it could be seen if violating the assumption of unbiased estimation of firm size would significantly alter the estimate of price.

BACKGROUND

One of the services of USDA is to provide periodic estimates of prices paid by farmers for items used in farm production. USDA began this service in 1910, when information was collected on prices paid for 86 different items. The goal then, as now, was to estimate the price paid per item.

A nonprobability survey of firms was initially used. Each state maintained a small, geographically dispersed list of firms for each group of commodities, such as feed, farm machinery, and so forth. Questionnaires were mailed to everyone on the lists, and average prices per reporting unit were calculated. Chronic nonrespondents were eventually dropped from the lists. Replacements were added when the lists got too short.

As with most surveys of this type, there were a number of problems. There was no way to measure the precision of the estimates. Many firms selling products to farmers had no chance of selection, so large biases in the estimates were likely. The response rate was low, which increased the chances for bias. And the lack of a scheme for rotating firms in and out of the sample put excessive burden on respondents and probably further reduced the response rate.

These problems, and renewed interest in agricultural prices, as evidenced by the emergence of terms like "target prices," "deficiency payments," and "100 percent parity," led to the creation in 1980 of an improved survey design. This design was in operation at the time of the study described in this paper. Lists of businesses were purchased, to create frames in each state for each commodity group. The groups were feed, fertilizer and pesticides, fuel and motor supplies, farm machinery, autos and trucks, and general farm supplies. The lists were built mainly from telephone yellow pages, and they were updated and supplemented by USDA office workers. A two stage sample of firms was selected for each commodity group.

Although this new design used probability survey methods, a number of problems were known or suspected. The frames were believed to cover the target population incompletely. (The target population is made up of all firms selling any of a particular list of farm production items to farmers.) Firms from outside this target population were known to be in the frames. There was no information on size of firms, so pps sampling could not be done. It was costly to build and maintain the frames, and they quickly became outdated. Finally, the surveys were designed only to estimate price per firm, rather than the desired price per item. To study these problems and look into an alternative method, the research method of building frames described in this paper was introduced and studied.

FINAL REMARKS

An alternative way of building sampling frames for the USDA surveys of prices paid by farmers was studied. Real and suspected problems with the operational surveys led to this study. The feasibility of the alternative method was shown, and several problems with the operational method were demonstrated, especially frame incompleteness. The research method was an improvement in several areas, but there were unsolved problems, such as infeasibility of pps sampling, small sample sizes, and item rarity. Consequently, it is recommended that a modified version of the research method be implemented. There are a number of possible modifications, several of which are suggested in the technical report, but each has some drawbacks. These alternative methods should be further investigated by USDA. In any future research, adequate sample sizes should be ensured.

REFERENCES

1. Cochran, William G. Sampling Techniques, 3rd Edition. New York: John Wiley & Sons, Inc., 1977.

2. House, Carol C. Evaluating New Reporter Sources for a USDA Price Survey. U.S. Department of Agriculture, Economics, Statistics and Cooperatives Service, Statistical Research Division, 1980.

3. House, Carol, Nancy Carter, and Doug Bond. Alternate Sampling Frame Construction for Prices Paid by Farmers Surveys -- An Overview. U.S. Department of Agriculture, Statistical Reporting Service, Statistical Research Division, 1983.

4. House, Carol, Nancy Carter, and Doug Bond. Alternative Frame Construction For the Prices Paid by Farmers Survey -- Technical Report. Unpublished staff report. U.S. Department of Agriculture, Statistical Reporting Service, Statistical Research Division, 1983.

5. Kish, Leslie. Survey Sampling. New York: John Wiley & Sons, Inc., 1965.

6. SAS Institute Inc. SAS User's Guide, 1979 Edition. Raleigh, NC: SAS Institute Inc., 1979.

7. Wolter, Kirk M., and others. Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. U.S. Bureau of the Census.

Table 1--Estimated percent incompleteness of the operational sampling frames.

State	Feed		Fertilizer and Pesticides		Farm Supplies		Farm Machinery		State Average
	1980	1981	1980	1981	1980	1981	1980	1981	
1	60	27	54	59	81	72	28	34	52
2	40	--	48	--	78	--	29	--	49
3	28	55	38	52	56	78	13	23	43
4	--	7	--	47	--	74	--	23	38
5	--	26	--	58	--	86	--	20	48
Average	43	29	47	54	72	77	23	25	46

Table 2--Comparison of 90 percent confidence intervals for price estimates.

Data Class	Number of Comparisons	Percent of Comparisons With Intersecting 90% Confidence Intervals
	--By Commodity Group--	
Farm Supplies	39	95
Fuel	82	80
Machinery	91	84
Feed	566	78
Fertilizer	63	76
	--By State--	
State 1	48	64
State 2	236	93
State 3	137	79
State 4	319	70
State 5	101	81
Total	841	79

Table 3--Results of simulation study--indication of whether self-weighted estimator was in or out of generated confidence intervals.

Standard Deviation of Generated Weights	Mean of Generated Weights: Type 1 Estimator			Type 2 Estimator		
	0.9	1.0	1.1	1.108	1.208	1.308
0.3	out	in	out	out	in	out
0.4	out	in	out	out	in	out
0.5	out	in	out	out	in	out
0.75	out	in	out	out	in	out
1.0	out	in	out	out	in	out